

<https://helda.helsinki.fi>

---

## Semantic Domains in Akkadian Text

Svärd, Saana Sofia

Brill

2018-08-07

---

Svärd , S S , Jauhiainen , H A , Linden , B K J & Sahala , A J A 2018 , Semantic Domains in Akkadian Text . in V B Juloux , A R Gansell & A di Ludovico (eds) , CyberResearch on the Ancient Near East and Neighboring Regions : Case Studies on Archaeological Data, Objects, Texts, and Digital Archiving . Digital Biblical Studies , no. 2 , Brill , Leiden , pp. 224-256 .

---

<http://hdl.handle.net/10138/241805>

---

cc\_by\_nc

publishedVersion

---

*Downloaded from Helda, University of Helsinki institutional repository.*

*This is an electronic reprint of the original article.*

*This reprint may differ from the original in pagination and typographic detail.*

*Please cite the original version.*

## CyberResearch on the Ancient Near East and Neighboring Regions

# Digital Biblical Studies

## *Series Editors*

Claire Clivaz (*Swiss Institute of Bioinformatics*)  
David Hamidović (*University of Lausanne*)

## *Editorial Board*

Heike Behlmer (*University of Göttingen*)  
Paul Dilley (*University of Iowa*)  
Laurence Mellerin (*Institute of Christian Sources, Lyon*)  
Sarah Savant (*Aga Khan University, London*)

## VOLUME 2

The titles published in this series are listed at *brill.com/dbs*

# CyberResearch on the Ancient Near East and Neighboring Regions

*Case Studies on Archaeological Data, Objects,  
Texts, and Digital Archiving*

*Edited by*

Vanessa Bigot Juloux  
Amy Rebecca Gansell  
Alessandro di Ludovico



BRILL

LEIDEN | BOSTON



This is an open access title distributed under the terms of the prevailing CC-BY-NC License at the time of publication, which permits any non-commercial use, and distribution, and reproduction in any medium, provided the original author(s) and source are credited.



An electronic version of this book is freely available, thanks to the support of libraries working with Knowledge Unlatched (KU). KU is a collaborative initiative designed to make high quality content Open Access for the public good. More information about the initiative and links to the Open Access version can be found at [www.knowledgeunlatched.org](http://www.knowledgeunlatched.org)

The Library of Congress Cataloging-in-Publication Data is available online at <http://catalog.loc.gov>

Typeface for the Latin, Greek, and Cyrillic scripts: "Brill". See and download: [brill.com/brill-typeface](http://brill.com/brill-typeface).

ISSN 2452-0586

ISBN 978-90-04-34674-1 (hardback)

ISBN 978-90-04-37508-6 (e-book)

Copyright 2018 by the Editors and Authors.

This work is published by Koninklijke Brill nv. Koninklijke Brill nv incorporates the imprints Brill, Brill Hes & De Graaf, Brill Nijhoff, Brill Rodopi, Brill Sense and Hotei Publishing.

All rights reserved. No part of this publication may be reproduced, translated, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without prior written permission from the publisher.

Authorization to photocopy items for internal or personal use is granted by Koninklijke Brill nv provided that the appropriate fees are paid directly to The Copyright Clearance Center, 222 Rosewood Drive, Suite 910, Danvers, MA 01923, USA. Fees are subject to change.

This book is printed on acid-free paper and produced in a sustainable manner.

*Digitus Dei est hic!*  
(Translation: The finger of God is here!)  
Roberto A. Busa, 2004





# Contents

Acknowledgments	XI
Editors' Note	XII
Phonology	XIII
Shortened Forms	XIV
List of Figures, Tables, and Appendices	XIX
Notes on Contributors	XXIV

## Introduction to CyberResearch on the Ancient Near East and Neighboring Regions 1

*Vanessa Bigot Juloux, Amy Rebecca Gansell, and Alessandro di Ludovico*

### PART 1

#### *Archaeology*

- 1 A Conceptual Framework for Archaeological Data Encoding 25  
*Sveta Matskevich and Ilan Sharon*
- 2 Landscape Archaeology and Artificial Intelligence: the Neural  
Hypersurface of the Mesopotamian Urban Revolution 60  
*Marco Ramazzotti*  
In collaboration with *Paolo Massimo Buscema & Giulia Massini*

### PART 2

#### *Objects*

- 3 Data Description and the Integrated Study of Ancient Near Eastern  
Works of Art: The Potential of Cylinder Seals 85  
*Alessandro di Ludovico*
- 4 A Quantitative Method for the Creation of Typologies for  
Qualitatively Described Objects 111  
*Shannon Martino and Matthew Martino*



**PART 3*****Texts***

- 5    **A Qualitative Approach Using Digital Analyses for the Study of Action  
in Narrative Texts: KTU 1.1-6 from the Scribe 'Ilimilku of Ugarit as a  
Case Study**    151  
      *Vanessa Bigot Juloux*
  
- 6    **Network Analysis for Reproducible Research on Large Administrative  
Cuneiform Corpora**    194  
      *Émilie Pagé-Perron*
  
- 7    **Semantic Domains in Akkadian Texts**    224  
      *Saana Svärd, Heidi Jauhiainen, Aleksi Sahala, and Krister Lindén*
  
- 8    **Using Quantitative Methods for Measuring Inter-Textual Relations in  
Cuneiform**    257  
      *M. Willis Monroe*

**PART 4*****Online Publishing, Digital Archiving, and Preservation***

- 9    **On the Problem of the Epigraphic Interoperability of Digitized Texts  
of the Mediterranean and Near Eastern Regions from the First  
Millennium BCE**    283  
      *Doğu Kaan Eraslan*
  
- 10   **Digital Philology in the Ras Shamra Tablet Inventory Project: Text  
Curation through Computational Intelligence**    314  
      *Miller C. Prosser*
  
- 11   **Publishing Sumerian Literature on the Semantic Web**    336  
      *Terhi Nurmikko-Fuller*

**Maps**    365

**Toponyms Related to Ancient Settlements or Regions Mentioned in  
the Volume**    370

**Glossaries**    373

<b>Index of Authors and Researchers (including authors in bibliographies, scholars, and historical figures)</b>	<b>433</b>
<b>Index of CyberResearch (including computer science, digital practice, mathematical, and technological terms)</b>	<b>442</b>
<b>General Index (including terms associated with Archaeology, History, Geography, Literature, Philology, and their methods)</b>	<b>454</b>



## Acknowledgments

We would like to thank the organizers of the annual conferences of the American Schools of Oriental Research (San Antonio, Texas, November 2016), the American Oriental Society (Los Angeles, California, March 2017), and Computer Applications and Quantitative Methods in Archaeology (Atlanta, Georgia, March 2017) for enabling many of the authors to introduce the work included in this volume.

Our gratitude also goes to our anonymous external peer reviewers for Brill and to our internal peer reviewers: Adam Anderson, Marine Béranger, Sergio Camiz, Jean-Baptiste Camps, Hanan Charaf Mullins, Francesca Cioè, Helen Dixon, Aaron Gidding, Nathan Morello, Terhi Nurmikko-Fuller, Andrew Pottorf, Anne-Caroline Rendu Loisel, Glenn Roe, Daniel Stockholm, Matthew L. Vincent, and Juan Pablo Vita, as well as our anonymous readers. We are also grateful to other contributors: Sarah Witcher Kansa, Brigitte Lion, Massimo Maiocchi, Palmiro Notizia, and Elena Pierazzo. We thank Eve Levavi Feinstein, Adina Yoffie, and Joshua J. Friedman for their assistance with editing and Elizabeth DiPippo and Susan Shapiro for their help with proofreading. In addition, we thank the Brill editors, and especially the editors of the Digital Biblical Studies series, Claire Clivaz and David Hamidović, who patiently supported us in this wonderful project.

Finally, the co-editors, Vanessa Bigot Juloux, Amy Rebecca Gansell, and Alessandro di Ludovico, sincerely thank the authors for their close collaboration and for having agreed to join us in this adventure toward the advancement of the humanities for both educational and scientific purposes.

More than anything else, this project would never have been successful without the strong support of our families (including Vanessa's numerous cats and dogs), friends, and peers, of whom we are deeply appreciative.

Special support for this volume was provided by St. John's University (New York).

## Editors' Note

We look forward to hosting future conference sessions showcasing Cyber-Research on the ancient Near East and neighboring regions. Furthermore, we hope that these sessions will generate content for additional edited volumes relevant to a wide audience of both digital and non-digital specialists as well as digital neophytes and students.

Through this project, we are proud to contribute to the preservation of cultural heritage by sharing techniques and results for the digital knowledge-discovery and documentation of ancient archaeological contexts, objects, and texts, as well as modern archives and data repositories. Above all, this volume is dedicated to our friends and colleagues, and to the many students and scholars, whose work and lives suffer on account of the ongoing military conflicts and political crises in regions of the Middle East.

# Phonology

(*vocalization*)

Consonants with diacritical marks in Akkadian, Sumerian and Ugaritic:

- ʾ (aleph), a voiced laryngeal with a glottal stop that is used with vowels /a/, /i/, /u/ for ʾa, ʾi, ʾu.
- ʿ (ayin), a voiced pharyngeal.
- ḏ (ḏal), a voiced interdental, as “th” in “that.”
- ġ (ġain), a voiced velar fricative, as “r” in French.
- ḡ a velar nasal, as “ng.”
- ḥ (ḥa), a voiceless velar fricative, as “ch” in German.
- ḥ (ḥota), a voiceless pharyngeal with a stricture of the pharynx.
- š (šin), a voiceless palate-alveolar fricative, as “sh” in shame.
- š (tsadé), an emphatic sibilant, as “ts.”
- ṭ (ṭet), an emphatic voiceless dental, as “t.”
- ṭ (taana), a voiceless interdental, as “th” in thin.
- z (zu), an emphatic voiced fricative dental, as “ts.”

# Shortened Forms

## Series, Periodicals, and Publishers

<i>Acalc</i>	<i>Archeologia e Calcolatori</i>
ACL	Association for Computational Linguistics
ACM	Association for Computing Machinery
<i>AJA</i>	<i>American Journal of Archaeology</i>
<i>AmA</i>	<i>American Anthropologist</i>
<i>AmAnt</i>	<i>American Antiquity</i>
<i>AnatS</i>	<i>Anatolian Studies</i>
AOAT	Alter Orient und Altes Testament
<i>AoF</i>	<i>Altorientalische Forschungen</i>
ARA	<i>Annual Review of Anthropology</i>
ASOR	American Schools of Oriental Research
<i>BaM</i>	<i>Baghdader Mitteilungen</i>
BAR-IS	British Archaeological Reports International Series
<i>BASOR</i>	<i>Bulletin of the American Schools of Oriental Research</i>
BSAJ	British School of Archaeology in Jerusalem
<i>CAD</i>	<i>Chicago Assyrian Dictionary</i>
<i>CDA</i>	<i>A Concise Dictionary of Akkadian</i>
<i>CH</i>	<i>Computers and the Humanities</i>
<i>CISE</i>	<i>Computing in Science &amp; Engineering</i>
<i>CMAO</i>	<i>Contributi e Materiali di Archeologia Orientale</i>
CNRS	Centre National de la Recherche Scientifique
CUSAS	Cornell University Studies in Assyriology and Sumerology
DANS	Data Archiving and Networked Services
<i>DHQ</i>	<i>Digital Humanities Quarterly</i>
HdO	Handbuch der Orientalistik
HSS	Harvard Semitic Studies
ICE	Initiative for Cuneiform Encoding
<i>IEJ</i>	<i>Israel Exploration Journal</i>
<i>IM</i>	<i>Istanbuler Mitteilungen</i>
<i>JAMT</i>	<i>Journal of Archaeological Method and Theory</i>
<i>JAOS</i>	<i>Journal of the American Oriental Society</i>
<i>JAR</i>	<i>Journal of Anthropological Research</i>
<i>JASC</i>	<i>Journal of Archaeological Science</i>
<i>JCS</i>	<i>Journal of Cuneiform Studies</i>
<i>JCSSS</i>	<i>Journal of Cuneiform Studies Supplemental Series</i>

<i>JEurArch</i>	<i>Journal of European Archaeology</i>
<i>JGRSC</i>	<i>Journal of Global Research in Computer Science</i>
<i>JNES</i>	<i>Journal of Near Eastern Studies</i>
<i>JTEI</i>	<i>Journal of the Text Encoding Initiative</i>
MDOG	Mitteilungen der Deutschen Orient-Gesellschaft
OBO	Orbis Biblicus Orientalis
OIP	Oriental Institute Publications
<i>OxfJA</i>	<i>Oxford Journal of Archaeology</i>
<i>PEQ</i>	<i>Palestine Exploration Quarterly</i>
<i>QDAP</i>	<i>Quarterly of the Department of Antiquities in Palestine</i>
RA	<i>Revue d'assyriologie et d'archéologie orientale</i>
<i>RANT</i>	<i>Res Antiquae</i>
SAA	State Archives of Assyria
SAOC	Studies in Ancient Oriental Civilization
<i>ScAnt</i>	<i>Scienze dell'Antichità</i>
<i>SEL</i>	<i>Studi Epigrafici e Linguistici sul Vicino Oriente Antico</i>
<i>SJIS</i>	<i>Scandinavian Journal of Information Systems</i>
StCh	Studia Chaburensia
<i>StTr</i>	<i>Studia Troica</i>
<i>UF</i>	<i>Ugarit-Forschungen</i>
<i>ZfSem</i>	<i>Zeitschrift für Semiotik</i>

### Additional Shortened Forms

AOS	American Oriental Society
AIUCD	Associazione per l'Informatica Umanistica e la Cultura Digitale
CAA	Computer Applications and Quantitative Methods in Archaeology
CIIT	Coimbatore Institute of Information Technology
CDLI	Cuneiform Digital Library Initiative
DARIAH	Digital Research Infrastructure for the Arts and Humanities
DH	Digital Humanities
ePSD	Pennsylvania Sumerian Dictionary
ETCSL	Electronic Text Corpus of Sumerian Literature
GLAM	Galleries, Libraries, Archives, and Museums
ICOMOS	International Council on Monuments and Sites
IFLA	International Federation of Library Associations and Institutions
Oracc	Open Richly Annotated Cuneiform Corpus
RAI	Rencontre Assyriologique Internationale
TEI-C	Text Encoding Initiative Consortium



## CyberResearch

AAS	Artificial Adaptative System
AI	Artificial Intelligence
ANN	Artificial Neural Network
API	Application Programming Interface
AS	Artificial Sciences
ASCII	American Standard Code for Information Interchange
ATDM	A-Temporal Diffusion Model
ATF	ASCII Transliteration Format
Auto-CM	Auto-Contractive Map
C-ATF	c(anonical) ASCII Transliteration Format
CAL	Comprehensive Aramaic Lexicon
CBOW	Continuous Bag-of-Words
CDLI	Cuneiform Digital Library Initiative
CIDOC CRM	CIDOC Conceptual Reference Model
CLTK	Classical Language Toolkit
CNNs	Convolutional Neural Networks
CRM	Conceptual Reference Model
CSV	Comma Separated Value
DANA	Digital Archaeology and National Archive
DBMS	Database Management System
DTM	Document Term Matrix
FRBR	Functional Requirements for Bibliographic Records
GUI	Graphical User Interface
HTML	HyperText Language Markup
HTTP	HyperText Transfer Protocol
HTTP URIS	HyperText Transfer Protocol Universal Resource Identifiers
ID	Identifier
ISO	International Organization for Standardization
JSON	JavaScript Object Notation
LD	Linked Data
LOD	Linked Open Data
log	logarithm
MdC	Manuel de Codage
MRG	Maximally Regular Graph
MST	Minimum Spanning Tree
NC	Natural Computing
NLP	Natural Language Processing
NPMI	Normalized Pointwise Mutual Information

OCHRE	Online Cultural and Historical Research Environment
OCR	Optical Character Recognition
OD	Open Data
OM	OntoMedia
OWL	Web Ontology Language
PMI	Pointwise Mutual Information
RDF	Resource Description Framework
RDFS	Resource Description Framework Schema
RSTI	Ras Shamra Tablet Inventory
SKOS	Simple Knowledge Organization System
SOM	Self-Organizing Map
SQL	Structured Query Language
SPAD	Système Portable pour l'Analyse des Données
SVG	Scalable Vector Graphics
SW	Semantic Web
TEI	Text Encoding Initiative
TF-IDF	Term Frequency–Inverse Document Frequency
TWC	Topological Weighted Centroid
UPGMA	Unweighted Pair Group Method with Arithmetic Mean
URIS	Universal Resource Identifiers
UTF	Unicode Transformation Format
W3C	World Wide Web Consortium
XDR	External Data Representation
XML	Extensible Markup Language

### Corpora of Texts and Seals

Adab	Sargonic Inscriptions from Adab
BM	British Museum
CUSAS	Cornell University Studies in Assyriology and Sumerology
KTU	Keilalphabetischen Text(e) aus Ugarit
LBAT	Late Babylonian Astronomical and Related Texts
MAMA	Monumenta Asiae Minoris Antiqua
PF	Persepolis Fortification
RS	Ras Šamra

Other

BVU	Basic Volume Unit
EDM	Electronic Distance Measurement
MURL	Mesopotamian Urban Revolution Landscape

# List of Figures, Tables, and Appendices

## Figures

### *Archaeology*

- 1.1 (a) A photograph from the archives of the Palestine Exploration Fund (PEF-P-1660), from a collection pertaining to Garstang's excavations at Dor (photographer unknown, 1922, courtesy of the Palestine Exploration Fund). Compare to (b) a photograph of W8424 according to the current project's recording system, identified as the north wall of the temple; Wall 2 on Garstang's plans (photograph by I. Hirshberg, 1986, courtesy of the Tel Dor Project) 33
- 1.2 Base of a Cypriot Monochrome vessel from the Israel Antiquities Authority's National Treasures, purportedly from Emanuel Anati's excavations at Tell Abu-Hawam in 1963 (photograph by Sveta Matskevich, courtesy of the Israel Antiquities Authority and the Digital Archaeology Lab) 37
- 1.3 Flint finds from a spit (photograph by Sveta Matskevich) 40
- 1.4 Megiddo, Area CC, part of the Stratum VIIB and VIIA plan (Loud 1948, fig. 409) 41
- 1.5 (a) Attic red-figure sherds from Garstang's excavation (Iliffe 1933, pl. 7b:5)  
(b) The refitted sherds (Stewart and Martin 2005, fig. 4; photograph by Gabi Laron, courtesy of the Israel Antiquities Authority) 44
- 1.6 Modeling ambiguities 50
- 1.7 Modeling multiple interpretations of one find 53
- 1.8 Modeling uncertainty of the stratigraphic definition of a feature 54
- 2.1a Auto-CM ANN (by Paolo M. Buscema © Semeion) 68
- 2.1b Tree-Visualizer (by Guilia Massini © Semeion) 68
- 2.2a Auto-CM – MST-MRG graph. The presence of the clay sickles (green points) and clay cones (yellow points) is uniformly distributed on the MST-MRG tree-graph. The distribution on the tree branches could indicate a homogeneous spatial relationship between food-production activities and symbolic/religious functions. 69
- 2.2b Auto-CM – MST-MRG graph. Testing the same neural hypersurface on the distribution of the two variables together (clay sickles and the clay cones), we observed a sensible reduction of the number of occurrences (purple points). The coexistence of the two variables in specific regions of the graph-tree could thus implicate a possible nuclear organization. 70
- 2.3a-d The four period TWC Gamma ( $\gamma$ ) maps of the MURL. Each map describes the extent to which each point of the space is activated by its closeness to any of

- the points belonging to any of the nonlinear trajectories connecting each point to the center of mass. 72
- 2.4 Machine-learning system for the neural simulation of the MURL (© Semeion – Guilia Massini) 73
- 2.5 TWC algorithms (© Semeion – Paolo M. Buscema) 75

### *Objects*

- 3.1 Encoding strategies adopted for the earliest (alphanumeric, above) and latest (textual, below) analyses. The position of the reference points and the ideal partition of the cylindrical surface are indicated in the example above and in that below to the right. The example of textual encoding shown here can be “translated” as follows: “([A goddess flounced robe headgear type-c5 hairstyle type-n1 turned right right-hand before-her-face left-hand before-her-face] [a man fringed mantle hairstyle type-c, turned right, right-hand by-the-waist left-hand by-the-waist has offering-o1] [astral symbol type-l2 above]) ([a god flounced robe headgear type-c5 hairstyle type-b2 turned left left-hand by-the-waist right-hand brought-forward has beard has bracelets] sits upon [throne type-Q1c dais-1c] [legend three-framed-cases content-e4]).” 93
- 3.2 Procedure followed within SPAD 5.5. The boxes connected vertically represent the successive steps that have been followed, while those connected to them horizontally represent the relevant outcomes. 98
- 3.3 Graphs showing the outcomes of the binary correspondence analyses on the toponyms, with (left) and without (right) the specimens of unknown origins 99
- 3.4 Total elements (forms) of the typical vocabulary that are shared by the different toponyms. Positive and negative forms have been considered together; the thickness of each edge is directly proportional to the total number of shared forms. 104
- 4.1 *Fliegende Blätter*, October 23, 1892 113
- 4.2 Example of core defined for Demircihüyük bottom section (images adapted from Obladen-Kauder 1996: pls. 114.2–3, 114.5, 114.7, 115.3–5, 115.7–8) 129
- 4.3 Example of core for Demircihüyük middle section (images adapted from Obladen-Kauder 1996: pls. 115.10, 116.3–2) 129
- 4.4 All forms captured by analysis (images adapted from Seeher 1988) 130
- 4.5 Variations of form 6 (images adapted from Seeher 1988) 131

### *Texts*

- 6.1 Information processing flowchart 198
- 6.2 A triple 203

- 6.3 The Adab network graph 211
- 7.1 Syntagmatic (abCde) and paradigmatic (xCy) relationships between words 231
- 7.2 Syntagmatic and paradigmatic relationships, using “orange” as an example 232
- 7.3 Formula for PMI score 239
- 7.4 Example of the formula for the PMI score 239
- 7.5 Formula for NPMI 240
- 7.6 Formula for the final score 240
- 8.1 Counts of ingredients in the Micro-zodiac 269
- 8.2 Cosine similarities greater than zero 273
- 8.3 Detail from dendrogram of Micro-zodiac cells 276
- 8.4 Detail from dendrogram of Micro-zodiac ingredients 277

### *Online Publishing, Digital Archiving, and Preservation*

- 9.1 C-ATF: Cropped detail of Tablet PF404 (CDLI) 286
- 9.2 Bavant-XML Standard: Cropped detail of DNA (Schmidt 1970, pl. 32) 286
- 9.3 mdc-88: Cropped detail of Jansen-Winkel 57103 designed with Jsesh software by Serge Rosmorduc (Jansen-Winkel 2014, 460, no. 57103) 288
- 10.1 Drawing of RS 3.320 324
- 10.2 Text, recomposed view in RSTI 326
- 10.3 The Lexicography Wizard 330
- 10.4 Dictionary lemma in RSTI 331
- 10.5 The Prosopography Wizard 332

### *Maps of Toponyms Mentioned in the Volume*

- 1 General map 365
- 2 Toponyms for Assyria, Babylonia, Dilmun, Elam, and the Zagros Mountains 366
- 3 Toponyms for the Levant and Syria 367
- 4 Toponyms for Anatolia 368
- 5 Location of Vindolanda (England) 369

### **Tables**

#### *Objects*

- 3.1 Corpus of specimens analyzed 96
- 3.2 Scenes from seals and impressions 97
- 3.3 Features of toponyms 101

- 4.1 Groups unlike Seeher's forms 131
- 4.2 Groups that best fit with Seeher's forms 132

### *Texts*

- 6.1 From text to glossary 205
- 6.2 Excerpt from a MySQL table storing information on the occurrence of tokens in the corpus 207
- 6.3 Excerpt from an edge list 208
- 6.4 Sample of entity frequency in maximal cliques 215
- 7.1 Three prototypical contexts for *sisû*, "horse" (syntagmatic relationships) 235
- 7.2 Three prototypical contexts for *qabû*, "to speak" (syntagmatic relationships) 236
- 7.3 Five prototypical contexts for *danānu*, "to be strong" (syntagmatic relationships) 237
- 7.4 Merging the score tables 241
- 7.5 NPMI results for the top fifteen suggestions for *sisû*, "horse" 241
- 7.6 NPMI results for the top fifteen suggestions for *qabû*, "to speak" 243
- 7.7 NPMI results for the top fifteen suggestions for *qabû*, "to speak," when using a window size of 5 244
- 7.8 NPMI results for the top fifteen suggestions for *danānu*, "to be strong" 245
- 7.9 Word2vec results for the top ten suggestions for *sisû*, "horse" 249
- 7.10 Word2vec results for the top ten suggestions for *qabû*, "speak" 250
- 7.11 Word2vec results for the top ten suggestions for *danānu* (*dnn*), "to be strong, powerful" 251
- 8.1 Traditional Babylonian zodiac 258
- 8.2 List of ingredients 264
- 8.3 Example row from the input data 268
- 8.4 Bag-of-words model for three cells 271
- 8.5 Document Term Matrix for example dataset 272
- 8.6 Adjacency matrix for example dataset 273

### *Online Publishing, Digital Archiving, and Preservation*

- 9.1 Elamite text in C-ATF 286
- 9.2 Elamite text in Bavant-XML 287
- 9.3 Ancient Egyptian text in mdc-88 288
- 9.4 Imperial Aramaic text in CAL code 289
- 9.5 Ancient Greek text in EpiDoc-XML 290

## Appendices

### *Archaeology*

- 2.1 By Giulia Massini: The Experimental and Simulation Procedures 72
- 2.2 By Paolo Massimo Buscema: The Topological Weighted Centroid Basic Definitions 74

### *Objects*

- 4.1 List of Figurine Attributes 137
- 4.2 Attribute Lists by “TopMiddleBottom” 142
- 4.3 List of Ceramic Attributes 145

### *Texts*

- 5.1 An Overview of ‘Anatu’s Actions 188



## Notes on Contributors

### *Vanessa Bigot Juloux*

holds an advanced degree in Ugaritic (École des Langues et des Civilisations de l'Orient Ancien). She is a PhD candidate at the École Pratique des Hautes Études (EPHE) and Paris Sciences et Lettres (PSL). Her research focuses on Ugaritic narrative texts and violent behavior for political purposes. She is involved in several committees at the American Schools of Oriental Research (ASOR). Since 2016, she has been co-organizing and co-chairing sessions on violence and digital humanities at ASOR and Computer Applications and Quantitative Methods in Archaeology (CAA) annual meetings, and she is also co-organizing and co-chairing an ASOR/EPHE European symposium (2018). She has recently developed open-access guidelines for analyzing actions in TEI-XML. She is currently co-editing a volume on violence in ancient cultures.

### *Doğu Kaan Eraslan*

has a BA in Philosophy from Galatasaray University (Turkey) and an MA in Egyptology from the Centre de Recherches Égyptologiques de la Sorbonne (CRES). He is a PhD candidate at the École Pratique des Hautes Études (EPHE) in Paris, where he combines his study of ancient civilizations with computer science, while being employed by EPHE's Cognition Humaine et Artificielle (CHART) laboratory. His research interests include cross-language information retrieval for ancient languages, the application of data-science methodologies to ancient history, the encoding of ancient languages for machine-learning and computer-vision purposes, and Late Period Mediterranean history and international relations.

### *Amy Rebecca Gansell*

holds a BA in Anthropology/Archaeology from Barnard College of Columbia University and an MA and PhD in Art History from Harvard University. She is an Associate Professor of Art History in the Art and Design Department at St. John's University in New York. Her research has been supported by grants from the U.S. National Endowment for the Humanities (NEH) and The American Academic Research Institute in Iraq (TAARI). Her collaborative data-mining projects on Levantine ivory sculptures have been published in the *Journal of Archaeological Science* and the proceedings of the Computer Applications and Quantitative Methods in Archaeology (CAA) annual meeting. Her most recent work, presenting a 3D model of a Neo-Assyrian queen, appears in the *American Journal of Archaeology*.

*Heidi Jauhiainen*

received her PhD in Egyptology in 2009 from the University of Helsinki. Since then, she has earned a bachelor's degree in Computer Science and has worked on a project identifying and gathering pages on the internet written in Finno-Ugric languages. At the moment, she is part of the team of the Academy of Finland project working on finding semantic domains in Akkadian texts.

*Krister Lindén*

is Research Director of Language Technology at the University of Helsinki and National Coordinator of FIN-CLARIN. He has published more than 90 scientific papers on developing resources for Language Technology and tools for Corpus Linguistics.

*Alessandro di Ludovico*

has a PhD in Archaeology and Art History of the Ancient Near East from La Sapienza University of Rome. He is currently a research fellow on projects concerning the historical geography of pre-classical Western Asia. His research interests mainly focus on figurative languages, material production, and the cultural history of third-millennium BCE Syria and Mesopotamia, all fields that he has investigated for many years using both quantitative and qualitative methods. He has participated in several international meetings on Near Eastern Archaeology, Assyriology, and Digital Humanities, and he has worked as an archaeologist on field projects in Syria and Palestine.

*Matthew Martino*

received his BA in Physics and Mathematics from the University of Chicago in 2003. He graduated with his PhD in Physics and Astronomy from the University of Pennsylvania in 2009 with a dissertation on cosmology, specifically large-scale structure formation and modified gravity. He currently teaches at the University of Chicago Laboratory Schools.

*Shannon Martino*

received her BA in Anthropology from the University of Chicago in 2003 and her PhD in Art History from the University of Pennsylvania in 2012. She was a postdoctoral fellow at the Field Museum in Chicago during the 2012–2013 academic year and has participated in excavations in Bulgaria, Turkey, and Syria. She currently teaches at the School of the Art Institute in Chicago, Illinois, and at Morton College in Cicero, Illinois. Her work focuses on Chalcolithic connections between Southeast Europe and Anatolia, particularly in terms of their pottery and anthropomorphic figurines.

*Sveta Matskevich*

is a postdoctoral fellow at the Zinman Institute of Archaeology at the University of Haifa and an associated research fellow in the Institute of Archaeology at the Hebrew University of Jerusalem where she earned her PhD. Trained as a field archaeologist, she has participated in excavation projects in Israel, Greece, and Turkey, mainly as a surveyor, draughtsperson, and Database Management System architect and administrator. Her research interests include archaeological data management, field methods, and the history of archaeology. She is currently starting her own excavation of the site of Tel Mevorakh (Israel) as part of a regional project investigating the natural environment and transportation networks in southern Phoenicia during the Middle-to-Late Bronze Age.

*M. Willis Monroe*

works on cuneiform documents from the Achaemenid and Hellenistic periods in Mesopotamia. His principle interest is in the structure and format of types of knowledge and the way in which ancient scribes transmitted their expertise between texts. His work has focused primarily on Babylonian astrology and multi-modal documents. He holds a PhD from Brown University and is currently based in Vancouver, Canada, as a postdoctoral researcher at the University of British Columbia.

*Terhi Nurmikko-Fuller*

holds a PhD from the University of Southampton. She is currently a lecturer in Digital Humanities at the Australian National University. Her research focuses on examining the potential of harnessing Semantic Web technologies to support and diversify scholarship in the humanities. To date, she has published on the use of Linked Open Data with musicological and library metadata; on the ontological representation of the narrative structure, philological, bibliographical, and museological data of ancient Mesopotamian literary compositions; and on the roles gamification and informal online environments can play in facilitating the learning process. Terhi was selected as a Fellow of the Software Sustainability Institute in 2016 and has visiting scholar status at the University of Oxford e-Research Centre.

*Émilie Pagé-Perron*

is a PhD candidate in Assyriology at the University of Toronto. She researches third-millennium BCE Mesopotamian social history, exploring administrative sources using traditional and computational approaches. She holds a Master's degree in Mesopotamian Studies from the University of Geneva, where she wrote a thesis, supervised by Antoine Cavigneaux, on the division of labor in

the fish industry of Early Dynastic Girsu. Previously a Jackman Junior Fellow and a Canadian Social Sciences and Humanities Research Council (SSHRC) doctoral awardee, she is co-principal investigator at the Cuneiform Digital Library Initiative (CDLI), where she manages the international Machine Translation and Automated Analysis of Cuneiform Languages (MTAAC) project and the CDLI Framework Update.

*Miller C. Prosser*

is a research database specialist for the OCHRE Data Service of the Oriental Institute of the University of Chicago. He has a PhD in Northwest Semitic Philology from the University of Chicago, where his doctoral research focused on the economic texts from the Late Bronze Age site of Ras Šamra-Ugarit. Miller has worked on various research projects at the Oriental Institute. He is currently the data and photography lab manager for the Persepolis Fortification Archive Project. He has also joined various archaeological excavations as a field and object photographer. As co-director of the Ras Shamra Tablet Inventory (RSTI), he is interested in pursuing database solutions to archaeological and philology data problems.

*Marco Ramazzotti*

has been, since 2007, a researcher and adjunct professor of Ancient Near Eastern Art and Archaeology in the Department of Classics at La Sapienza University of Rome. He combines historical-artistic and cognitive analysis of the ancient Near Eastern cultural milieu and researches the relationships among Analytical Archaeology, Artificial Intelligence, and Natural Computing. Since 1989, he has participated in many archaeological excavations, surveys, and restoration field projects in Palestine, Turkey, Libya, Yemen, Jordan, Iraq, and, especially, Syria. In Syria, he has been the field director of excavations and coordinated different interdisciplinary scientific teams for the planning and opening of the Ebla Archaeological Park.

*Aleksi Sahala*

received his MA in the discipline of Language Technology (2014) from the University of Helsinki, where he is currently a doctoral student writing his thesis under the supervision of Krister Lindén and Saana Svärd.

*Ilan Sharon*

is the Nachman Avigad Professor in the Institute of Archaeology at the Hebrew University of Jerusalem. For almost his entire career, he has worked at the site of Tel Dor, starting as a beginning graduate student in 1980 and assuming

co-directorship of the project in 2002. His other abiding interest is Mathematics in Archaeology. As such, he has taught and done research on statistics in archaeology, GIS, 3D modeling, computerized stratigraphic modeling, and analytic typology, and he is one of the founders of a computational archaeology laboratory at the Institute of Archaeology at the Hebrew University of Jerusalem.

*Saana Svärd*

has a PhD from the University of Helsinki, where she is now a docent of Assyriology and Cultural History of the Near East. She is one of the leading experts in the study of women and gender in Mesopotamia, especially regarding the Neo-Assyrian Empire. Currently, she is the principal investigator of the project “Deep Learning and Semantic Domains in Akkadian Texts” (2017–2020). The project engages with cross-disciplinary endeavors to use language technology to gain a more nuanced understanding of semantics in Akkadian. This goal is also one of the focal points of the National Centre of Excellence in Ancient Near Eastern Empires (University of Helsinki, 2018–2025), which she directs.

# Introduction to CyberResearch on the Ancient Near East and Neighboring Regions

*Vanessa Bigot Juloux, Amy Rebecca Gansell, and  
Alessandro di Ludovico*

No longer does digital research impact only scientific and technical studies. It is now established as a powerful approach for revealing new information and supporting new interpretations in the humanities. Disciplines such as Archaeology, Anthropology, Art History, History, Philology, and Literary Studies have begun to benefit. Today more and more researchers of the ancient Near East, neighboring regions, and other ancient-world domains extend their traditional investigations by using digital technologies as standard and experimental tools. However, because digital research is a relatively new approach, the space dedicated to cyber-research in the humanities, especially in the study of the ancient world, is still limited and in many cases relegated to quite isolated contexts, such as specialized research centers, congresses, and conference sessions. Furthermore, until recently, critical discussions have been lacking among practitioners across projects and disciplines, and communication between cyber-researchers and traditional humanists remains inadequate. As a result, the many significant cyber-research projects that have been conducted are little known outside of specialized circles, even when their impact is relevant to the traditional humanities. Thus, a true debate about cyber-research projects and their results, impact, and potential is still missing.

Because of the great value and potential of cyber-research to the humanities overall, we hope to see its further integration into the mainstream. But if the *digital* humanities (DH) continue to grow along a specialized, independent trajectory, the humanities could suffer a division between digital and non-digital projects and practitioners. As a result, DH contributions to the study of human culture could become more and more inaccessible to non-DH scholars. The use of jargon and specialized digital tools (such as software and programming languages) in DH research can impede the accessibility of DH projects to non-DH scholars. However, this volume is dedicated to broadening the understanding and accessibility of DH research tools, methodologies, and results.

In an effort to build bridges among DH projects, methods, and researchers, and to invite non-digital practitioners to follow the processes and results of DH research, we have attempted to make all papers in this volume meaningful and

accessible to non-DH researchers as well as to DH practitioners across methodological and disciplinary silos. For instance, we hope that a computer scientist would be able to appreciate archaeological discussions and issues, an archaeologist would be able to follow philological analyses, and a humanities student would understand the logic and mechanisms of computer science techniques. To support such cross-disciplinary understanding, all papers have been peer reviewed not only for technical and factual soundness, but also for accessibility.<sup>1</sup> In addition, we provide definitions for technical and field-specific words in glossaries at the end of the volume.<sup>2</sup> We believe that our communicating to both extended and specialized audiences could allow the papers in this volume to inspire new ideas, research innovations, and collaborations among humanities researchers, whether or not they are “digital” specialists. We also hope that the essays could be used pedagogically, both in classroom contexts and by scholars who wish to develop or expand their own digital research toolkits.<sup>3</sup>

### What are Cyber-research and the Digital Humanities?

The first step in demystifying cyber-research for the uninitiated and affirming its broad value and applications to the seasoned practitioner is a semantic one. What exactly is “cyber-research”? What is its unique value, and how is it distinguished from other forms of investigation? “Cyber” can be understood today as a generic term meaning “of, relating to, or involving computers or computer networks.”<sup>4</sup> In this sense, “cyber” can refer to many things, such as online

- 
- 1 Our internal peer review committee included digital and non-digital practitioners, with specialization in the ancient Near East as well as in other historical fields and disciplines.
  - 2 We thank our authors, especially Terhi Nurmikko-Fuller and Émilie Pagé-Perron, for their support of and assistance with the glossaries, as well as peer reviewers Massimo Maiocchi and Nathan Morello for reinforcing their significance to the volume.
  - 3 The use of digital techniques is sometimes recommended, solicited, or even rewarded by academic institutions. Scholars who are not digital practitioners may feel ill-equipped to participate in the digital arena. However, with only a basic understanding of the logistics and problem-solving potential of DH procedures and instruments, it is possible for beginners to conceive of their own DH projects for which they can then seek expert collaborators and/or develop the necessary technical skills themselves.
  - 4 We use the term “CyberResearch” in the titles of this volume and chapter as a generic, umbrella term referring to the research methods described here in their broadest sense. We propose to introduce CyberResearch as a new, formally recognized methodology that can be integrated into Ancient Near Eastern Studies as well as across the humanities (see Bigot Juloux and Ludovico 2018). *Merriam Webster*, s.v., “cyber,” last modified January 30, 2018, <<https://www.merriam-webster.com/dictionary/cyber>> (accessed February 21, 2018). “Cyber” derives from the Greek stem κυβερνήτης (*kybernētēs*), “cybernetic,” which was used by Plato in *Alcibiades*

shopping (Amazon.com), digital communication (email and cell phone),<sup>5</sup> and even futuristic punk fashion. With regard to research, “cyber” indicates the use of innovative computer technologies, digital media, and algorithmic data processing to facilitate research. Computer technologies used in cyber-research go beyond the standard “front-end” use of application software such as spreadsheets, text editors, and databases;<sup>6</sup> they require “back-end” coding in special languages to manipulate, analyze, and visualize data,<sup>7</sup> as well as to display content on a browser (e.g., Firefox, Chrome, Safari). Algorithmic data processing involves a sequence of mathematical actions to perform tasks dedicated to solving a set of problems. When customized to solve problems that entail the processing of large amounts of data, algorithms are mostly calculated by computers, rather than by humans.<sup>8</sup> Overall, then, cyber-research refers to investigations that rely on programming languages and/or mathematical formulas to relay data and instructions to a computer. The computer generates output that can be interpreted by researchers as results that may answer their questions,

---

1 to describe the governance of people. See *Merriam Webster* (s.v., “cybernetics,” second definition, <<https://www.merriam-webster.com/dictionary/cybernetics>> [accessed February 21, 2018]): “the scientific study of how people, animals, and machines control and communicate information.” The theme of control is etymologically central to the idea of cybernetics (note also the ancient Greek verb κυβερνάω, “to govern, to drive,” the etymology of which is related to the Sanskrit *kubara*, “rudder”), while the common use and meaning of this and other terms belonging to the same family are due to the research and theories of the American mathematician Norbert Wiener (1894–1964), who founded the discipline now known as “Cybernetics” (Wiener 1948).

- 5 Digital communication is exemplified by texting on a mobile device, instant messaging (such as on Messenger and WeChat), and communication via social networking media (such as Twitter, Facebook, and LinkedIn).
- 6 A software application (also known as an “application program”) enables the user to execute tasks on a computer. Examples of software applications used by our authors include the FileMaker (Martino and Martino) and MySQL (Pagé-Perron) databases, as well as the following text editors: Atom (Pagé-Perron), Oxygen XML Editor (Bigot Juloux), Stylus Studio (Prosser), SPAD (Ludovico), and xCode (Martino and Martino).
- 7 Examples of programming languages used by our authors include C++, Java, PHP, Python, SQL, XML, and XML-TEI. For an example of a graph visualization, see in this volume Pagé-Perron, 211, Fig. 6.3.
- 8 The word “algorithm” derives from the latinized version of the name of the Persian scholar Al-Khwārizmī (c. 780–850 CE), who brought the concept of algebra to Europe (*Encyclopedia Britannica*, s.v., “Al-Khwārizmī,” <<https://www.britannica.com/biography/al-Khwarizmi>> [accessed June 8, 2017]). The immense amount of complex calculations that are often required would be impractical, if not impossible, for even the most skilled mathematician to carry out. Digital algorithmic data-processing, however, is efficient, and it can compare, manage, and integrate huge amounts of data on a very large scale.



support or contradict their hypotheses, and/or illuminate new topics and paths of inquiry.

When cyber-research is employed to further our understanding of the humanities, it can reveal new information that is difficult or impossible to discern through the use of traditional methods. These investigations are often classified as “digital humanities” projects. According to the U.S. National Foundation on the Arts and the Humanities Act, 1965, as amended, the humanities, as a set of disciplines, includes “but is not limited to, the study and interpretation of the following: language, both modern and classical; linguistics; literature; history; jurisprudence; philosophy; archaeology; comparative religion; ethics; the history, criticism and theory of the arts; those aspects of social sciences which have humanistic content and employ humanistic methods; and the study and application of the humanities to the human environment with particular attention to reflecting our diverse heritage, traditions, and history and to the relevance of the humanities to the current conditions of national life.”<sup>9</sup> The digital humanities are concerned with the same topics as the humanities; where they are distinguished from the humanities is specifically in their methodologies of cyber-research. Basically, in the term “digital humanities,” “digital” stands in for “cyber,” referring to the computer technologies and algorithmic data processing employed as research methodologies. However, a more in-depth definition of “digital humanities” is actually more ambiguous.

We are in the midst of a global technological revolution, with digital technologies appearing everywhere.<sup>10</sup> The worldwide impact of DH is currently more prevalent in some disciplines and in some countries.<sup>11</sup> In general, DH research has tended to begin in disciplines that intersect with the natural sciences and/or are innately rich in quantitative or text-based data; only subsequently have digital methodologies been adopted for research on more abstract material, such as art. This is currently seen, for example, in Turkey, where digital humanities is still a nascent field, and DH methodologies are mostly used in the disciplines of Comparative Literature, Linguistics, Sociology, and Political Science.<sup>12</sup> In Finland, although the concept of DH (*digitaaliset*

9 <<https://www.neh.gov/about>> (accessed June 9, 2017). For additional definitions of “the Humanities,” see Liu (2014, <<http://4humanities.org/2014/12/what-are-the-humanities/>> [accessed June 9, 2017]).

10 “The digital revolution entered a new phase, giving rise to a vastly expanded, globalized public sphere” (Burdick et al. 2012, <[https://mitpress.mit.edu/sites/default/files/titles/content/9780262018470\\_Open\\_Access\\_Edition.pdf](https://mitpress.mit.edu/sites/default/files/titles/content/9780262018470_Open_Access_Edition.pdf)> [accessed June 11, 2017], 2). Also see Evans and Rees 2012, 37.

11 The following discussion of the digital humanities in an international context provides examples from our authors’ countries.

12 The Turkish expressions *sayısal beşeri bilimler* and *dijital beşeri bilimler* both refer to the digital humanities. The first expression is similar to the French terminology *humanités*

*ihmistieteet*, literally, “digital human-sciences”) is known, *digitaaliset ihmistieteet* is also a developing field and not widespread in its application across disciplines. As in Turkey, Finland lacks any DH centers to support research, networking, collaboration, and communication within the field.<sup>13</sup> Through these small examples, it is evident that international opportunities are essential to the global growth of DH research.

Especially since our volume includes papers from scholars working in many different countries, where they speak many different languages, we would like to acknowledge the challenge of defining “digital humanities” for a multilingual audience by considering some examples of the complexities of international semantics. For example, “Hebrew chooses to speak about *rouach digitalit*, the ‘digital spirit’.”<sup>14</sup> However, Italians avoid the term “digital” by saying either *informatica per le scienze umane*,<sup>15</sup> which means “computer science applied in/to humanities” or “*informatica umanistica*,” meaning “humanities computing”—an expression that was actually used in English prior to the general acceptance of the term “digital humanities.”<sup>16</sup> Through the word *informatica* (referring to “computing” or “computer science”), one understands the point of view of the Italian scholar Enrica Salvatori, who asserts that the digital humanities are specifically a transdisciplinary approach that produces, rather than simply applies, computing tools to serve a specific purpose.<sup>17</sup> In contrast, the French use

---

*numériques* and is mostly used in everyday speech, while the second expression refers to digital humanities for academic studies (Salah 2015, <<http://sam.sehir.edu.tr/tr/series-of-talks-on-cities-buyuk-veri-caginda-dijitallesen-beseri-bilimler/#sthash.HDPesAD1.dpuf>> [accessed June 10, 2017]). We thank Doğu Kaan Eraslan for this information.

13 “In Finland the concept of digital humanities is known but there is still fairly little of formal contribution” (Viiri, 2014, <<https://talkarttalksociety.wordpress.com/category/archives/>> [accessed June 9, 2017]). For a map and list of Digital Humanities Centers worldwide, see <<http://dhcenter.net/org/centers>> (accessed June 10, 2017).

14 Clivaz 2017, <<http://digiHubb.centre.ubbcluj.ro/journal/index.php/digitalia/article/view/4/18>> (accessed June 10, 2017), 28.

15 See, for example, Ciraci (2012).

16 Ciotti 2014, <<https://infouma.hypotheses.org/244>> (accessed June 9, 2017).

17 “Una prima visione percepisce l’informatica—all’interno dell’endiadi Informatica Umanistica—sostanzialmente come uno strumento che non deve semplicemente essere ben applicato all’ambito delle scienze umane, ma può e deve essere dedicato ad esso, espressamente studiato per servire a uno scopo preciso ... L’informatico umanista—in questa visione punta a ottenere una rilevante specializzazione su uno o più specifici *tool*, su cui acquisire tutte le necessarie competenze per un loro uso corretto e funzionale all’interno di un ambito specifico di impiego” (Salvatori 2015, <<https://esalvatori.hypotheses.org/204>> [accessed June 10, 2017]). Translation by Bigot Juloux and Ludovico: “A first perspective perceives information technology—inside the hendiadys Digital Humanities—

the terms *humanités numériques* and, more rarely, *humanités digitales*, which is the literal translation of the English term “digital humanities.”<sup>18</sup> Both expressions are translated as “digital humanities,” although *humanités numériques* preserves reference to the “numerical,” i.e., computational, nature of the research. However, if one differentiates between *numérique* and *digitale* from the perspective of French semantics, according to the semiologist Anthony Mathé, the word *numérique* is most typically used to refer to “digital” technologies and media, such as films and music, while the term *digitale* describes the use of “digital” (in the English-language sense) technology.<sup>19</sup> The French thus

---

essentially as a tool that should not simply be applied accurately to the field of the humanities, but that can and must be devoted to it, specifically designed to serve a specific purpose ... The digital humanist—from this perspective—aims to obtain a relevant specialization in one or more specific tools, in order to capture all of the necessary skills for their correct and functional use within a specific scope of applications.” See also a more detailed paper on this topic: Angiolini et al. (2015).

- 18 *Humanités digitales* is the literal translation of “digital humanities,” following the Anglicism “digital” (*Merriam Webster*, s.v., “digital,” updated February 15, 2018, <<https://www.merriam-webster.com/dictionary/digital>> [accessed February 21, 2018]). See also Centre National de Ressources Textuelles et Lexicales (s.v., “digital,” first definition, meaning B, 2012, <<http://www.cnrtl.fr/definition/digital>> [accessed June 10, 2017]), which follows the definition provided in *Académie française* (s.v., “digital,” updated November 7, 2013, <<http://www.academie-francaise.fr/digital>> [accessed June 10, 2017]). Valérie Carayol and her colleagues from the University of Bordeaux first used “humanités digitales” in 2008 (Clivaz, 2017, 29). Regarding the meaning of “humanités” and its semantic development, see Clivaz (2017, 31–33, 36): “As astonishing as it is, while many scholars emphasize the corporeal elements related to the use of the word ‘digital/e’ in its relation to the word ‘humanités’, they all ignore that such potential traces belong even more so to the history of the word ‘humanité’ itself”...[And with regard to the academic understanding,] “the French word *humanités* signals an humanist interdisciplinary perspective: it includes History, Letters and Political Sciences.”

- 19 Ropars (2017, <<http://www.blogdumoderateur.com/numerique-ou-digital/>> [accessed June 8, 2017]) quotes Anthony Mathé, saying, “On parle d’industrie numérique et de pratiques digitales ... Numérique tend à renvoyer de fait au technologique, à la dimension discrète de la technologie, celle que manipulent les ingénieurs et qui restent intangible. Digital semblerait concerner plutôt l’usager dans son expérience de cette technologie numérique.” Bigot Juloux’s translation: “One speaks of numerical industry and digital practices ... In fact, numerical refers to technology, to the discrete scope of the technology, the one that engineers manipulate and that remains intangible. Digital would rather appear to concern the user in his experience of this numerical technology.” For additional discussion, see *Histoires de Digital Makers* blog post (February 27, 2015), “‘Digital’ or ‘Numerique’? A New Linguistic Debate Rages in France” (<<https://www.digitalforallnow.com/en/digital-numerique-linguistic-france/>> [accessed June 8, 2017]).

speak of *pratiques digitales* (digital practices), *technologies numériques* (digital technologies), and, whether they are *digitales* (in either the French or English sense) or *numériques*, the French scholar Aurélien Berra has definitively recognized *humanités digitales* as a research practice.<sup>20</sup> Today, in France and internationally, *numérique(s)* or *digitale(s)* research is extended from science into humanities practice.<sup>21</sup>

In any language, the term “digital humanities” (and its equivalents and near-equivalents) is quite new. Its meaning and scope are not yet agreed upon, even within English usages. Although a pioneering wave of digital humanities scholarship took place in the late 1990s,<sup>22</sup> the term “digital humanities” was first introduced to the academic community by John Unsworth (from the University of Virginia) in 2001, when he proposed a new Master’s degree program called “Digital Humanities.”<sup>23</sup> The same year, in collaboration with the editorial and marketing teams at Wiley-Blackwell, he also developed the title *A Companion to Digital Humanities*, for a volume on what, at the time, was commonly called

20 In the words of Aurélien Berra (2012, <<http://books.openedition.org/oep/238>> [accessed June 9, 2017], 25–43): “Les humanités numériques sont donc principalement une pratique de recherche.” Translation by Bigot Juloux: “Digital humanities are therefore primarily a research practice.”

21 For further discussion of the French translation, see Clivaz (2017, 29–33).

22 Presner 2010, <<https://cnx.org/contents/JoK7N3xH@6/Digital-Humanities-20-A-Report>> (accessed June 9, 2017). But the premise of the digital humanities can be dated to 1949 when the Jesuit scholar Roberto A. Busa was preparing materials on St. Thomas Aquinas for the *Index Thomisticus* project, a computer-generated concordance produced in collaboration with IBM (Busa 2004, xvi–xvii; Burdick et al., 2012, 123). See also the University College London project “Hidden Histories: Computing and the Humanities c. 1949/1980” (<<http://www.ucl.ac.uk/infostudies/research/hiddenhistories/>> [accessed June 8, 2017]).

23 Unsworth considered but intentionally avoided the terms “Humanities Informatics” and “Humanities Computing.” For the draft proposal for the degree, see <<http://www.people.virginia.edu/~jmu2m/laval.html>> (accessed June 8, 2017). The degree, however, was never realized (Kirschenbaum 2012, <<http://dhdebates.gc.cuny.edu/debates/text/48>> [accessed June 8, 2017], 418). For the emergence of DH, see Bernard (2012, <<http://books.openedition.org/oep/242>> [accessed June 7, 2017], 45–58) and Hockey (2004). To quote *The Digital Humanities Manifesto 2.0* (<<http://manifesto.humanities.ucla.edu/2009/05/29/the-digital-humanities-manifesto-20/>> [accessed June 7, 2017]): “Digital Humanities is not a unified field but an array of convergent practices that explore a universe in which: a) print is no longer the exclusive or the normative medium in which knowledge is produced and/or disseminated; instead, print finds itself absorbed into new, multimedia configurations; and b) digital tools, techniques, and media have altered the production and dissemination of knowledge in the arts, human and social sciences.”

“humanities computing.”<sup>24</sup> While Berra has asked the question of whether DH constitutes a type of practice or methodology, or whether it is a field,<sup>25</sup> Christine Borgman maintains that it addresses research problems in humanities disciplines using new sets of technologies.<sup>26</sup> Frederic Darbellay has suggested that DH refers to the junction between new information technologies and humanities’ disciplines.<sup>27</sup> Matthew Kirschenbaum has raised the question “What is (or are) the ‘digital humanities,’ aka ‘humanities computing?’”<sup>28</sup> and for Eileen Gardiner and Ronald Musto “digital humanities” refers to a methodology, while “humanities computing” is a field.<sup>29</sup> Meanwhile, Patrik Svensson has proposed that “humanities computing ... is largely a common interest in methods, methodology, tools and technology.”<sup>30</sup> Amidst this debate, referring to the initial goal of the “computing humanities,” David Berry has suggested that we should consider “digital humanities” as “a technical support to the work of the ‘real’ humanities scholars, who would drive the projects.”<sup>31</sup> Indeed, there are so many apparent “definitions” of “digital humanities” that Jason Heppler has established a website dedicated to compiling how different people have defined the term.<sup>32</sup> Also, there is an annual international “Day of DH” (established in 2009), during which anyone can electronically submit questions to the DH

- 
- 24 Kirschenbaum 2010, <<https://www.adelphi.edu/bulletin/article/adelphi.150.55>> (accessed June 7, 2017), 56–57. The volume bearing the seminal title (coined in 2001) *A Companion to Digital Humanities* was not published until three years later (Schreibman 2004).
- 25 Berra 2012.
- 26 “The digital humanities is a new set of practices, using new sets of technologies, to address research problems of the discipline” (Borgman 2009, <<http://digitalhumanities.org/dhq/vol/3/4/000077/000077.html>> [accessed June 10, 2017]).
- 27 “‘Digital humanities’ désignent la rencontre entre les nouvelles technologies de l’information et de la communication et les disciplines des sciences humaines et sociales, des arts et des lettres” (Darbellay 2012, <<https://www.nss-journal.org/articles/nss/abs/2012/03/nss120033/nss120033.html>> [accessed June 7, 2017], 269). Bigot Juloux’s translation: “Digital humanities means the meeting between new information and communication technologies and the disciplines of the humanities and social sciences, arts and literature.”
- 28 “Humanities computing as a whole maintains a very instrumental approach to technology in the Humanities” (Kirschenbaum 2010, 55).
- 29 Gardiner and Musto 2015, 4.
- 30 Svensson 2009, <<http://digitalhumanities.org/dhq/vol/3/3/000065/000065.html>> (accessed June 10, 2017).
- 31 Berry 2011, <<http://www.culturemachine.net/index.php/cm/article/viewarticle/440>> (accessed June 7, 2017), 2.
- 32 See: <<http://www.whatisdigitalhumanities.com>> (accessed February 22, 2018). A new definition is displayed with each refresh of the screen.

community.<sup>33</sup> In the early years, it was asked, “How do you define Humanities Computing/Digital Humanities?”<sup>34</sup> Since 2013, a new question has arisen: “Just what do digital humanists do?”<sup>35</sup> A major point of debate concerns which cyber-research methods fall under the umbrella of DH.<sup>36</sup> From some perspectives, DH incorporates all aspects of cyber-research, including any techniques borrowed from Natural Sciences, Engineering, Applied Mathematics, and Computer Science. Other points of view exclude computer-science techniques from the domain of the digital humanities and prefer to say that computer science can be integrated as a tool in the process of DH analyses; others specifically see coding but not algorithmic formulas as DH tools. Perhaps a compromise would be to follow the Stanford Humanities Center’s broad definition of digital humanities as a “hybrid domain ... at the crossroads of computer science and the humanities.”<sup>37</sup>

Because the very definition of DH is unstable, research falling under the broadest definition of digital humanities is not necessarily classified as such according to narrower definitions (for example, a project may be described as a Computer Science, rather than a DH, investigation). The threshold for classifying humanities research as “digital” is also ambiguous. For instance, research using front-end technologies, such as automated functions in off-the-shelf software, is sometimes—but not typically—accepted as technically rigorous enough to be described as digital humanities versus humanities research. For example, virtual-reality ancient world simulations, projects that produce 3D scans of sites, objects, and texts, as well as archaeological investigations that collect and disseminate data through digital rather than manual methods, are

---

33 *A Day in the Life of the Digital Humanities* is an open community publication (Rockwell et al. 2012, <<http://www.digitalhumanities.org/dhq/vol/6/2/000123/000123.html>> [accessed February 22, 2018]). For Day of DH 2017, see <<http://dayofdh2017.linhd.es/>> (accessed February 22, 2018).

34 See “How do you define Humanities Computing/Digital Humanities?” (wiki page of the University of Alberta, last updated March 16, 2011, <[http://www.artsmn.ualberta.ca/tapor/wiki/index.php/How\\_do\\_you\\_define\\_Humanities\\_Computing/\\_/Digital\\_Humanities%3F](http://www.artsmn.ualberta.ca/tapor/wiki/index.php/How_do_you_define_Humanities_Computing/_/Digital_Humanities%3F)> [accessed, August 17, 2017]).

35 See the welcome page on the Day of DH 2014 website, <<http://dayofdh2014.matrix.msu.edu/>> (accessed, August 17, 2017).

36 Presner 2010; Alvarado 2012, 50.

37 See <<http://shc.stanford.edu/digital-humanities>> (accessed June 8, 2017). Also consider: “Digital humanities is by its nature a hybrid domain, crossing disciplinary boundaries and also traditional barriers between theory and practice, technological implementation and scholarly reflection” (Flanders, Piez, and Terras, 2007, <<http://digitalhumanities.org/dhq/vol/1/1/000007/000007.html>> [accessed June 10, 2017]).

sometimes described as DH enterprises. Even more “traditional” uses of digital products could be considered to fall within the scope of DH research, since the cognitive and sensorial perception of digital objects is so different from the experience of working with concrete, material sources of information. Finally, some views only refer to research conducted in specific branches of the humanities as DH research. In particular, there is a notion that DH research is specific to textual and literary corpora.<sup>38</sup> Clearly, the boundaries of the digital humanities are not yet solidified. In order to avoid debates over what is or is not a DH project, and so that we do not perpetuate confusion, this volume equates all of its projects with cyber-research. That is, all of the papers presented here exemplify cyber-research methodologies for humanities investigation.<sup>39</sup>

### Overview of the Volume

With this volume, we specifically hope to intervene in a scholarly landscape in which DH research has had a less-than-optimal impact upon the general narratives of ancient history.

Here we present some of the latest interdisciplinary research projects in which cyber-methodologies (including computational and computer science techniques) play a central role in the investigation of humanities-based questions about ancient Near Eastern and surrounding cultures from the Chalcolithic period through the Iron Age. The 11 contributions collected here represent the work of archaeologists, anthropologists, art historians, and philologists,

---

38 Regarding digital humanities and text analysis, see Tüfekçi (2015, <<http://monografjournal.com/sayilar/4/yazinsal-calismalarda-dijital-yonelimler-monograf-sayi-4.pdf>> [accessed June 7, 2017], 92): “İnsani bilimlerde dijital yönelimler çerçevesinde geliştirilen çözümleme yöntemleri belirli bir izlek, bakış açısı ve ileti kaygısı sınırlaması taşımadan metnin iskeletini ve buna bağlı tüm verileri olduğu gibi görselleştirerek yeni çalışmaların kullanımına sunup belirli bir görüşü dayatmayan yeni okumaların, yeni metinlerin oluşmasını sağlar.” Eraslan’s translation: “Analytical methods with digital orientations that are developed in the humanities provide new readings—new texts that do not impose a predefined view, by visualizing the skeleton of the text and all the data related to it without limiting itself to predefined notions, points of view, and to a worry for a message.” By saying “worry for a message,” Tüfekçi means that prior biases do not impact one’s statements.

39 “Digital Humanities is defined by the opportunities and challenges that arise from the conjunction of the term digital with the term humanities to form a new collective singular” (Burdick et al., 2012, 122).



many of whom also engage methods and theories derived from other disciplines, such as Philosophy and Geography. Most of the papers were introduced at the 2016 and 2017 annual meetings of the American Oriental Society (AOS), the American Society for Oriental Research (ASOR), and Computer Applications and Quantitative Methods in Archaeology (CAA).<sup>40</sup> To round out the volume and facilitate cross-fertilization among topics and methodologies, a few additional papers were independently solicited. All essays, however, have been written or revised (from their conference sessions) to fit the mission and framework of this volume.

The book is divided into four parts: “Archaeology,” “Objects,” “Texts,” and “Online Publishing, Digital Archiving, and Preservation.” The first part, “Archaeology,” presents papers dealing with data related to fieldwork activities, entire sites, and landscape archaeology, while part II, “Objects,” focuses on specific corpora of material culture. The papers in part III, “Texts,” employ written documents as primary sources for information discovery. Then, through three text-based case studies, part IV, “Online Publishing, Digital Archiving, and Preservation,” addresses the value of sharing and publishing data online. Such data sharing and publishing not only creates an electronic record of ancient evidence, but it also facilitates collaboration and provides an efficient basis for future research.<sup>41</sup>

Across the four parts of this volume, the reader will find connections among the projects presented in the chapters. For example, in part III (“Texts”), chapter 5 includes a section, “Analytical Taxonomies in TEI,” that relates to methods employed in projects discussed in parts I (“Archaeology”), II (“Objects”), and IV (“Online Publishing, Digital Archiving, and Preservation”). To facilitate the study of methods and tools used across projects, cross-references are provided in the chapters. Although the authors of each chapter employ unique approaches to examine specific topics, all participate in the domain of cyber-research and complement one another with humanities contributions to the study of the ancient Near East and neighboring regions.

---

40 ASOR: <<http://www.asor.org/>>; CAA: <<http://caa-international.org/>>; AOS: <<https://www.americanorientalsociety.org/>> (all accessed June 9, 2017).

41 “Digital technology can enhance the preservation of artifacts by providing superlative surrogates of original sources while at the same time protecting the artifact from overuse” (Smith, 2004, <[http://digitalhumanities.org:3030/companion/view?docId=blackwell/9781405103213/9781405103213.xml&chunk.id=ss1-5-7&toc.depth=1&toc.id=ss1-5-7&brand=9781405103213\\_brand](http://digitalhumanities.org:3030/companion/view?docId=blackwell/9781405103213/9781405103213.xml&chunk.id=ss1-5-7&toc.depth=1&toc.id=ss1-5-7&brand=9781405103213_brand)> [accessed August 18, 2017], 589). Also, according to Paul Conway (1996, <<https://www.clir.org/pubs/reports/conway2/index.html#gen8>> [accessed August 18, 2017]), digital preservation is made through digital imagery technologies “to protect original items.”



In the first part, “Archaeology,” two papers explore new approaches to excavation data and data recording. In chapter 1, Sveta Matskevich and Ilan Sharon investigate the excavation records of Tel Dor in Israel and reflect upon the conceptual basis of archaeological data recording. They address questions such as: How can data be best digitized and curated for the long-term? And can interoperability among different registration systems be achieved? In response, they seek a common denominator and logical structure for all excavation-recording systems and “present the new idea of [a] graph database” as a means of organizing disparate records.<sup>42</sup> Moving from matters of recording archaeological data to interpreting it, chapter 2, by Marco Ramazzotti (in collaboration with Paolo Massimo Buscema and Giulia Massini), presents the innovative approach of applying Artificial Intelligence models to landscape archaeology records in order to test Mesopotamian Urban Revolution theories. This research “enhances [our] understanding of complex cultural processes in ancient anthropic contexts,”<sup>43</sup> as it reveals previously unidentified interrelations during the four periods (Ubaid, Uruk, Jemdet Nasr, and Early Dynastic) of the Mesopotamian Urban Revolution through an analysis of archaeological settlement data. The results move us beyond traditional perspectives by providing “new insight into the current knowledge on the settlement processes in Lower Mesopotamia.”<sup>44</sup>

Part II, “Objects,” also contains two chapters. In chapter 3, Alessandro di Ludovico considers a corpus of late third-millennium BCE (Ur III period) Mesopotamian cylinder seal imagery. Using encoding and software-assisted correspondence analysis, he develops descriptions, classifications, interpretations, and comparisons among presentation scenes depicted on seals and seal impressions. The results of his “analysis both add to current scholarly reconstructions of the role that geographic distribution (among other factors) plays in glyptic traditions and effectively demonstrates the application of quantitative and automated methods in the study of the ancient Near East.”<sup>45</sup> Chapter 4 shifts in focus from images to artifacts. Here, Shannon Martino and Matthew Martino work with third-millennium BCE Early Bronze Age figurines and ceramics from Turkey. They question subjectivity and objectivity, then propose a

---

42 We asked our reviewers to indicate how our authors’ methodologies contribute to various scholarly domains and may help to motivate and support new investigations. In the following summaries of the chapters, we quote, with their kind permission, some of the reviewers’ statements. Here we quote reviewer Sergio Camiz.

43 Here we quote reviewer Francesca Cioè.

44 Here we quote reviewer Francesca Cioè.

45 Here we quote reviewer Andrew Pottorf.

methodology for classifying figurines and ceramics to build a typology based on common attributes. Their method implies a strong collaboration between programmers and archaeologists, thus showing the vital benefits of interdisciplinarity. The Martinos' project presents "a very interesting possibility to deal with typological analyses in a somewhat more objective fashion,"<sup>46</sup> and it offers a new model that is "potentially useful for a preliminary classification of a large variety of archaeological artefacts."<sup>47</sup>

Part III, "Texts," contains four chapters covering three millennia of writing in Ugaritic, Sumerian, and Akkadian. Chapter 5, by Vanessa Bigot Juloux, explores actions between characters in a second-millennium BCE Ugaritic myth ('Anatu in the Ba'lu and 'Anatu Cycle [KTU 1.1–6]) in order to set up a "hermeneutics of action," as well as to investigate gender-related roles. She encodes the text using "markup language to allow an ... automatic analysis,"<sup>48</sup> and her results "really show how actions within verbs are complex."<sup>49</sup> Next, in chapter 6, Émilie Pagé-Perron situates us in Mesopotamia with a corpus of third-millennium BCE Old Akkadian texts. She employs network analysis to reveal labor structures and relationships among individuals mentioned across the corpus. Both her methods and results "may help to formulate new scientific questions and create new fields of investigations."<sup>50</sup> Chapter 7, by Saana Svärd, Heidi Jauhiainen, Aleksi Sahala, and Krister Lindén, uses statistical methods to analyze semantic fields in Akkadian vocabulary and thereby understand cultural meaning from ancient, emic perspectives. "The paper proves to be the first step of a potentially huge contribution to the philological and historical research of the ancient Near East."<sup>51</sup> Its methodology could confront "many important open questions in our field,"<sup>52</sup> and "the use of text to understand the cultural mores of ancient cultures [also] has interesting implications for archeology."<sup>53</sup> In the final paper (chapter 8) of part III, M. Willis Monroe employs computer-aided textual analysis to discover patterns of associations among astrological cuneiform documents created under Achaemenid and Hellenistic rule in Babylonia from the late sixth to second century BCE. While bringing an intriguing "Micro-zodiac"

---

46 Here we quote reviewer Matthew Vincent.

47 Here we quote an anonymous reviewer.

48 Here we quote reviewer Sergio Camiz.

49 Here we quote reviewer Anne-Caroline Rendu Loisel. Her original comment was in French: "des différentes catégories qui montrent bien la complexité des actions dans les verbes" (translation by Bigot Juloux).

50 Here we quote reviewer Marine Béranger.

51 Here we quote reviewer Nathan Morello.

52 Here we quote reviewer Nathan Morello.

53 Here we quote reviewer Aaron Gidding.

text to light, his results contribute to the understanding of ancient scholarly practices behind the composition and editing of these astrological records. Monroe's project "shows the feasibility of applying digital humanities methods to a fragmentary, specialized corpus ... [and the benefits of digital analysis in] removing the idiosyncratic ... biases an Assyriologist might bring to the analysis of lists of technical vocabulary."<sup>54</sup>

Our volume closes with part IV, "Online Publishing, Digital Archiving, and Preservation," which celebrates the value and future prospects of making cyber-research data, methodologies (including algorithms and code, for example), and results freely accessible online to worldwide audiences. The three chapters in this part present exemplars of text-based research, but they also offer methods and standards applicable to any sort of cyber-project, including those in different fields and disciplines across and beyond the humanities. Chapter 9, by Doğu Kaan Eraslan, considers the problem of the multilingual interoperability of encoding schemes through a case study of the quadrilingual vase of the ancient Persian king Darius I (r. 522–486 BCE). The vase bears four inscriptions, which are written with Elamite cuneiform, Akkadian cuneiform, Egyptian hieroglyphs, and Old Persian cuneiform. Eraslan explores how to best encode these ancient texts across languages, keeping in mind epigraphic accuracy as well as the importance of conserving information about the physical features of the inscribed object and the writing itself. "The potential of the approaches described in the paper is huge, as it may open up new avenues of research, crossing data from multilingual corpora—a desideratum in the Humanities."<sup>55</sup> In chapter 10, Miller C. Prosser introduces the Ras Shamra Tablet Inventory (RSTI), a project aimed at creating reliable digital editions of the texts in the Ras Šamra-Ugarit corpus within a research database environment that also seeks to integrate archaeological data from the excavations at Ras Šamra. The Online Cultural and Historical Research Environment (OCHRE), an online database environment, is used to add new data to RSTI, allowing project members to have immediate, live access to the data. OCHRE "allows researchers to edit their data so that it is 'Linked Data' ready ... and it enables researchers to do this themselves, without necessarily having to ask a programmer to help them."<sup>56</sup> Finally, in Chapter 11, through a case study of the Electronic Text Corpus of Sumerian Literature (ETCSL), Terhi Nurmikko-Fuller demonstrates the benefits of Linked Data and online publication for philological, museological, curatorial, and archaeological data. Nurmikko-Fuller "introduces a novel

54 Here we quote reviewer Helen Dixon.

55 Here we quote an anonymous reviewer.

56 Here we quote author and reviewer Terhi Nurmikko-Fuller.

method ... to create linked ontologies for digitized cuneiform texts [that] ... will be a valuable application to additional cuneiform datasets online.”<sup>57</sup> This paper “is of broader interest than the simple publishing of a single corpus, because it shows clearly the impact that the data and their methodological treatment can have for the evolution and progress of ... [cyber-]technologies and their usability by researchers and curators alike.”<sup>58</sup> Nurmikko-Fuller’s chapter thus represents the spirit and potential of sharing data, knowledge, and knowledge-discovery techniques that this volume and all of its contributions embrace and aim to inspire.

### What Does That Mean? Check Our Glossaries!

It is necessary to speak the same language in order to communicate and collaborate productively, but the accessibility of both cyber-research and ancient world studies is hampered by specialized terminologies. This volume, however, aims to overcome this hurdle. To accommodate the diverse experience and expertise of our readers, jargon from all domains, ranging from Mathematics and Computer Science to Archaeology and Ancient Near Eastern Studies, is either defined in footnotes or, when the terminology has a broad usage across the volume, is defined in our customized CyberResearch and General (ancient world) Glossaries.

The CyberResearch Glossary includes computer science, mathematical, and technological terms. It contains definitions and practical examples quoted from the essays. In some cases, a term has specific meanings in the contexts of different methodologies. For example, in addition to its generic definition, the term “node” has a specific function within the approaches of four of our essays (see the contributions of Bigot Juloux, Pagé-Perron, Prosser, and Nurmikko-Fuller). On account of such specific usages, the CyberResearch Glossary aims to clarify the meaning of technical terminology both for neophytes and for specialists who may not recognize the specialized usage of a familiar term in a new domain.

The General Glossary includes terms from Archaeology, History, Geography, Literature, and Philology. Its goal is to support readers who are not ancient world specialists, but who wish to study the application of digital methods to humanities research. Among these readers, we hope, would be self-educators as well as instructors and students from diverse disciplines and fields in

---

<sup>57</sup> Here we quote reviewer Adam Anderson.

<sup>58</sup> Here we quote reviewer Jean-Baptiste Camps.

humanities and DH programs who participate in cyber-research courses on subjects such as Python,<sup>59</sup> the analysis and visualization of data with R, digital philology and quantitative methods, web semantics (RDF and SPARQL), and XML data using XPath and XQuery. To date, such courses lack broadly accessible case studies that provide practical examples of the application of cyber-research methods in non-scientific fields. Such material would be of particular pedagogical value, since the digital humanities, as an interdisciplinary field of teaching, fosters the “theoretical consideration and practical progress of interdisciplinarity with the new mode of production and dissemination of knowledge.”<sup>60</sup> We hope the projects presented in this volume may help fill this gap in pedagogical materials, since diverse, concrete examples of DH projects can be very useful training resources that would broaden the sample of methodologies to be studied and compared across disciplines.

---

59 Some international examples of digital humanities programs include Stanford University's undergraduate minor in Digital Humanities (<<https://dhminor.stanford.edu/>> [accessed August 18, 2017]), the Australian National University's undergraduate major and minor in Digital Humanities, as well as its Master's degree in Digital Humanities and Public Culture (<<http://cdhr.anu.edu.au/>> [accessed August 18, 2017]), and King's College London's MPhil and PhD in Digital Humanities (<<https://www.kcl.ac.uk/study/postgraduate/research-courses/digital-humanities-research-mphil-phd.aspx>> [accessed August 18, 2017]). Also, Paris Sciences et Lettres (PSL) offers an interdisciplinary Master's degree in Digital Humanities with an especially broad pedagogical reach for students with a bachelor's degree in a humanities field. The courses are also open to PhD candidates, postdoctoral fellows, and established researchers from several institutions and research centers that are members of PSL, including Collège de France, École Nationale des Chartes, École des Hautes Études en Sciences Sociales, École Normale Supérieure, École Pratique des Hautes Études, École Française d'Extrême Orient, Le Centre national de la recherche scientifique (CNRS, the French National Center for Scientific Research), and INRIA (the French National Institute for Computer Science and Applied Mathematics). These French universities and centers include several humanities fields covering prehistoric through contemporary periods.

60 “Les Digital Humanities se présentent comme un champ interdisciplinaire d'enseignement et de recherche qui a pour ambition de connecter les réflexions théoriques et les avancées pratiques de l'interdisciplinarité avec les nouveaux modes de production et de diffusion des connaissances” (Darbellay, 2012, 270). English translation by Bigot Juloux: “Digital humanities can be described as an interdisciplinary field of teaching and research that aims to interlink theoretical reflections and advanced interdisciplinary practices with new modes of knowledge production and dissemination.”

## Conclusion: The Path Forward

Three central, critical topics represent the milestones along the path we would like to begin to walk with this volume. First is the question of the concrete role of digital technologies and their applications in the investigation of artifacts and languages that fall into historical and cultural frames (for instance, pre- and early historic periods, Sumerian culture, and Ugaritic literature) that present, for modern scholars, large gaps of knowledge. A second fundamental issue to which this volume responds is the need for a wide coordination of efforts in the use and acceptance of digital technologies and their results. This is especially important for establishing collaborations and methodological debate within and beyond the DH community. We particularly aim to bring together scholarly contributions using cyber-technologies for the study of antiquity, in order to give the ancient world greater visibility and to allow for the comparison of approaches across the many disciplines and subfields of ancient world studies. Furthermore, we hope that non-DH researchers will critically engage with and potentially benefit from the outcome of DH projects, such as those presented here. Finally, we have begun to confront the complexities and challenges of functionally embedding the use of digital tools into the study of the ancient Near Eastern and other ancient world cultures. We must consider solutions for managing the rapid obsolescence of digital tools and for maintaining control of them without becoming dependent on their peculiar modes of operation—or on information technology (IT) specialists who, in the manner of modern scribes, sometimes hold the exclusive knowledge of their logistics and mechanisms. It is also necessary to establish models for preserving digital products, such as databases and interactive visualizations, in order to keep them active, usable, and open for users in the long term.

Digital technology has changed our world. The ways we read, write, learn, communicate, and play have fundamentally changed due to the advent of networked digital technologies.<sup>61</sup>

Even though today the “digital” and “traditional” humanities are still different enterprises,<sup>62</sup> in a near future it is likely that “digital,” including “cyber”

---

61 Here we quote the U.S. National Endowment for the Humanities, see <<https://www.neh.gov/divisions/odh/about>> (accessed June 10, 2017).

62 By “traditional” we understand the “practice” as “a way of thinking, behaving, or doing something that has been used by the people in a particular group, family, society, etc., or a long time” (*Merriam Webster*, s.v., “traditional,” updated February 15, 2018,

methodologies, will become integral to humanities research.<sup>63</sup> In 2013, William Pannacker even predicted that “we’ll lose the ‘digital’ [from ‘digital’ humanities] within a few years, once practices that seem innovative today become the ordinary methods of scholarship.” During this exciting time, “digital” and “traditional” communities of scholars are beginning to communicate, collaborate, and restructure the methods, theories, and expectations of the humanities.<sup>64</sup> Contributing to this shift, this volume eagerly supports the assimilation of “cyber” methodologies into general humanities practices.

In conclusion, differentiating our efforts from many previous stand-alone DH projects, we assume the responsibility of disseminating DH methodologies and research to a broad and diverse audience for the purposes of supporting scientific cyber innovation, digital and non-digital humanities research, and classroom education. We aim to increase expert communication among highly specialized projects that are sometimes impenetrable to outsiders and to make these projects accessible to researchers who themselves are not trained in cyber-research techniques. We welcome all readers and encourage those from across disciplines to read the chapters that are both within and outside of their areas of expertise for the sake of finding new approaches to managing, studying, analyzing, visualizing, and interpreting their own data.

---

<<https://www.merriam-webster.com/dictionary/traditional>> [accessed February 22, 2018]). Differences also exist among communities within the humanities. Still relevant today, although stated over a decade ago, is this statement by Flanders, Piez, and Terras (2007): “We need to work hard to explain our work and ideas and to make them visible to those outside our community who may find them useful.”

63 Pannacker 2013, <<http://www.chronicle.com/article/Stop-Calling-It-Digital/137325/>> (accessed June 10, 2017).

64 According to Thomas Kuhn (1996, 84–85), “All crises begin with the blurring of a paradigm and the consequent loosening of the rules for normal research ... The transition from a paradigm in crisis to a new one from which a new tradition of normal science can emerge is far from a cumulative process, one achieved by an articulation or extension of the old paradigm. Rather it is a reconstruction of the field from new fundamentals, a reconstruction that changes some of the field’s most elementary theoretical generalizations as well as many of its paradigm methods and applications. During the transition period, there will be a large but never complete overlap between the problems that can be solved by the old and the new paradigms. But there will also be a decisive difference in the modes of solution. When the transition is complete, the profession will have changed its view of the field, its methods, and its goals.”



## References

- Alvarado, Rafael C. 2012. "The Digital Humanities Situation." In *Debates in the Digital Humanities*, edited by Matthew K. Gold, 50–55. Minneapolis: University of Minnesota Press.
- Angiolini, Andrea, Francesca Di Donato, Luca Rosati, Federica Rossi, Enrica Salvatori, and Stefano Vitali. 2015. "Digital Humanities: 'Crafts and Occupations.'" In *AIUCD '14 Proceedings of the Third AIUCD Annual Conference on Humanities and Their Methods in the Digital Ecosystem*, article (5), edited by Francesca Tomasi, Roberto Rosselli Del Turco, and Anna Maria Tammaro. New York: Association for Computing Machinery.
- Bernard, Lou. 2012. "Du literary and linguistic computing aux digital humanities: retour sur 40 ans de relations entre sciences humaines et informatique." In *Read/Write Book 2: Une introduction aux humanités numériques*, edited by Pierre Mounier, 45–58. Marseille: OpenEdition Press. <<http://books.openedition.org/oep/242>>.
- Berra, Aurélien. 2012. "Faire des humanités numériques." In *Read/Write Book 2: Une introduction aux humanités numériques*, edited by Pierre Mounier, 25–43. Marseille: OpenEdition Press. <<http://books.openedition.org/oep/238>>.
- Berry, David M. 2011. "The Computational Turn: Thinking about the Digital Humanities." *Culture Machine* 12. <<http://www.culturemachine.net/index.php/cm/article/viewarticle/440>>.
- Bigot Juloux, Vanessa, Alessandro di Ludovico. 2018. "Digital Practices vs. Digital Humanities: Reflections to Bridge the Gap in Order to Improve Research Methods and Collaboration." Paper presented at CAA annual meeting, Tübingen, Germany.
- Borgman, Christine. 2009. "The Digital Future is Now: A Call to Action for the Humanities." *DHQ* 3 (4). <<http://digitalhumanities.org/dhq/vol/3/4/000077/000077.html>>.
- Burdick, Anne, Johanna Drucker, Peter Lunenfeld, Todd Presner, and Jeffrey Schnapp, eds. 2012. *Digital Humanities*. Cambridge, MA: MIT Press. <[https://mitpress.mit.edu/sites/default/files/titles/content/9780262018470\\_Open\\_Access\\_Edition.pdf](https://mitpress.mit.edu/sites/default/files/titles/content/9780262018470_Open_Access_Edition.pdf)>.
- Busa, Roberto A. 2004. "Foreword: Perspectives on the Digital Humanities." In *A Companion to Digital Humanities*, edited by Susan Schreibman, Ray Siemens, and John Unsworth, xi–xxvii. Malden, MA: Wiley-Blackwell.
- Ciotti, Fabio. 2014. "Digital Humanities in Italy and their role in DARIAH research infrastructure." *Leggere, Scrivere e Far di Conto* (blog), September 14, 2014. <<https://infouma.hypotheses.org/244>>.
- Ciraci, Fabio. 2012. *Informatica per le scienze umane. Fonti scientifiche e strumenti per la ricerca storico-filosofica in ambiente digitale*. Milan: McGraw-Hill.
- Clivaz, Claire. 2017. "Lost in Translation? The Odyssey of 'Digital Humanities' in French." *Studia UBB Digitalia* 62 (1). <<http://digihubb.centre.ubbcluj.ro/journal/index.php/digitalia/article/view/4/18>>.



- Conway, Paul. 1996. "Preservation Management in the Digital World." Council on Library and Information Resources. <<https://www.clir.org/pubs/reports/conway2/index.html#gen8>>.
- Darbellay, Frédéric. 2012. "Les Digital Humanities: vers une interdisciplinarité 2.0?" *Natures Sciences Sociétés* 20: 269–270. <<https://www.nss-journal.org/articles/nss/abs/2012/03/nssi20033/nssi20033.html>>.
- Evans, Leighton, and Sian Rees. 2012. "An Interpretation of Digital Humanities." In *Understanding of Digital Humanities*, edited by David M. Berry, 21–41. Basingstoke: Palgrave Macmillan.
- Flanders, Julia, Wendell Piez, and Melissa Terras. 2007. "Welcome to Digital Humanities Quarterly." *DHQ* 1 (1). <<http://digitalhumanities.org/dhq/vol/1/1/000007/000007.html>>.
- Gardiner, Eileen, and Ronald G. Musto. 2015. *The Digital Humanities: A Primer for Students and Scholars*. New York: Cambridge University Press.
- Hockey, Susan. 2004. "The History of Humanities Computing." In *A Companion to Digital Humanities*, edited by Susan Schreibman, Ray Siemens, and John Unsworth, 3–19. Malden, MA: Wiley-Blackwell.
- Kirschenbaum, Matthew. 2010. "What Is Digital Humanities and What's It Doing in English Departments?" *ADE Bulletin* 150: 55–61. <<https://www.ade.mla.org/bulletin/article/ade.150.55>>.
- Kirschenbaum, Matthew. 2012. "Digital Humanities As/Is a Tactical Term." In *Debates in the Digital Humanities*, edited by Matthew K. Gold, 415–428. Minneapolis: University of Minnesota Press. <<http://dhdebates.gc.cuny.edu/debates/text/48>>.
- Kuhn, Thomas S. 1996. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.
- Liu, Alan. 2014. "What are the Humanities?" *4Humanities* (blog), December 21, 2014. <<http://4humanities.org/2014/12/what-are-the-humanities/>>.
- Pannapacker, William. 2013. "Stop Calling It 'Digital Humanities.'" *The Chronicle of Higher Education*, February 18, 2013. <<http://www.chronicle.com/article/Stop-Calling-It-Digital/137325>>.
- Presner, Todd. 2010. "Digital Humanities 2.0: A Report on Knowledge." *OpenStax CNX*, June 8, 2010. <<https://cnx.org/contents/JoK7N3xH@6/Digital-Humanities-20-A-Report>>.
- Rockwell, Geoffrey, Peter Organisciak, Megan Meredith-Lobay, Kamal Ranaweera, Stan Ruecker, and Julianne Nyhan. 2012. "The Design of an International Social Media Event: A Day in the Life of the Digital Humanities." *DHQ* 6 (2). <<http://www.digitalhumanities.org/dhq/vol/6/2/000123/000123.html>>.
- Ropars, Fabian. 2015. "Faut-il dire numérique ou digital?" *blog du modérateur*, February 11, 2015. <<http://www.blogdumoderateur.com/numerique-ou-digital/>>.

- Salah, Almıla Akdağ. 2015. "Büyük Veri Çağında Dijitalleşeri Bilimler." Lecture announcement, Istanbul Sehir University. <<http://sam.sehir.edu.tr/tr/series-of-talks-on-cities-buyuk-veri-caginda-dijitallesen-beseri-bilimler/#sthash.HDPesAD1.dpuf>>.
- Salvatori, Enrica. 2015. "L'identità dell'informatico umanista e la visione sistemica." *Hypotheses Academic Blogs*, January 23, 2015. <<https://esalvatori.hypotheses.org/204>>.
- Schreibman, Susan, Ray Siemens, and John Unsworth, eds. 2004. *A Companion to Digital Humanities*. Malden, MA: Wiley-Blackwell.
- Smith, Abby. 2004. "Preservation." In *A Companion to Digital Humanities*, edited by Susan Schreibman, Ray Siemens, and John Unsworth, 576–590. Malden, MA: Wiley-Blackwell.
- Svensson, Patrik. 2009. "Humanities Computing as Digital Humanities." *DHQ* 3 (3). <<http://digitalhumanities.org/dhq/vol/3/3/000065/000065.html>>.
- Tüfekçi, Şevket. 2015. "Yazınsal Çalışmalarda Dijital Yönelimler." *Monograf* 4: 91–130. <<http://monografjournal.com/sayilar/4/yazinsal-calismalarda-dijital-yonelimler-monograf-sayi-4.pdf>>.
- Viiri, Sampo. 2014. "Digital Humanities and Future Archives." *The Finnish Institute in London* (blog), June 6, 2014. <<https://talkarttalksociety.wordpress.com/category/archives/>>.
- Wiener, Norbert. 1948. *Cybernetics: Or, Control and Communication in the Animal and the Machine*. New York: J. Wiley.



**PART 1**

*Archaeology*





# A Conceptual Framework for Archaeological Data Encoding

*Sveta Matskevich and Ilan Sharon\**

## Introduction

The work presented here evolved within the framework of the Tel Dor project. Dor is a major port town on the Mediterranean, between present-day Tel Aviv and Haifa in Israel.<sup>1</sup> It was active from the Middle Bronze Age to the Roman era. The present Dor expedition is the custodian of a long history of excavations and archaeological research that started in the early twentieth century and still continues.<sup>2</sup> Such a “long durée” excavation poses particular challenges for data curation. The following remarks describe the different excavation projects conducted at Dor and the datasets they produced.

In 1922–1923, John Garstang, then director of the British School of Archaeology in Jerusalem, conducted two excavation seasons on the tell,<sup>3</sup> digging the temple compounds in the western part of the mound and placing a probe trench through the southern slope, down to bedrock at the shore of the Southern Bay. Our search for Garstang’s original records at the Rockefeller Museum and the Palestine Exploration Fund uncovered some plans (essentially the same ones published in the preliminary reports),<sup>4</sup> photographs, and partial lists of finds. Nothing remained of the site grid he established, but it can be approximated from the extant remains of the temples. We do not know what kind of registration system he used.

In the 1950s, the regional antiquities inspector, Joseph Leibowitz, conducted salvage excavations around the tell. He was the first to locate and excavate the

---

\* The research was supported by a postdoctoral fellowship funded by the Zinman Institute of Archaeology at the University of Haifa, Israel and Tel Dor Archaeological Project.

1 geo:32.617428, 34.916364.

2 For an up-to-date bibliography on Tel Dor, see <<http://dor.huji.ac.il/bibliography.html>> (accessed February 22, 2018).

3 Tell: an artificial mound formed as a result of successive cycles of construction, occupation, and destruction, sometimes separated by periods of abandonment.

4 Garstang 1924a; 1924b.

Roman theater to the north and the church to the southeast. Some years ago, Leibowitz's heirs gave us some of his notebooks containing field notes and the beginning of an unfinished site report. These constitute the only known records of his excavations, except for a brief published note.<sup>5</sup>

The church was excavated again by Claudine Dauphin in 1979–1983.<sup>6</sup> Dauphin's expedition produced several articles but no final report. We do not possess any of the original field records. Shimon Gibson, who worked with Dauphin, also conducted some surveys on and around the tell.<sup>7</sup>

Underwater exploration, as well as probes and surveys of maritime installations along the beach, began in the mid-1970s and continued in the 1980s, coinciding with the first years of large-scale excavation on the tell. These investigations were conducted by different people, including Avner Raban and Michal Artzi from the University of Haifa's Center for Maritime Civilizations, and Shelley Wachsman, Sean Kingsley, and Kurt Raveh, who were working for the Israel Department of Antiquities at the time. The expeditions were loosely collaborative,<sup>8</sup> but the records of each were managed independently. Even within the maritime exploration, each subproject was conducted independently, often with a local grid and datum. The Haifa excavators used the classic "Israeli" recording system with the locus as a primary spatial unit,<sup>9</sup> and they managed their records through a field diary, a daily top-plan, and a set of preprinted forms borrowed from the Tel Akko excavation.<sup>10</sup> We do not have the records of the other teams, nor do we know what kind of registration systems they used. Since these various expeditions never shared their raw data (with each other or with the tell excavations), the compatibility of their field

---

5 Leibowitz 1950.

6 Dauphin 1984.

7 Dauphin and Gibson 1993.

8 Stern et al. 1995.

9 About the recording system, see Aharoni et al. (1973). Locus (pl. loci): the basic spatial unit in most recording systems in Near Eastern Archaeology. The exact denotation of this term (i.e., the archaeological entity that is modeled by the abstract unit) can vary among different systems. Spatial unit: refers to a locational unit of excavation defined in absolute coordinates, relatively to the site grid, or to architectural features and other units in its vicinity.

10 Daily top-plan: a daily graphical record of an excavation area or part of it. The background plan shows all architectural features extant in the area on that particular day. The foreground information reflects how the excavation proceeded during the day. It comprises several types of data, such as outlines of excavation units, elevations, and the registration numbers and coordinates of finds. For the Tel Akko excavation, see Raban (1993).

records was never questioned. However, the recently launched joint land-sea excavation (discussed below) necessitates the integration of these datasets.

The next extensive excavation of the tell was initiated in 1980 by Ephraim Stern of the Hebrew University of Jerusalem. The recording system employed by the team at the very beginning was a slightly different version of the “Israeli” system, borrowed from the Tel Batash excavation.<sup>11</sup> It was altered considerably during the first ten campaigns<sup>12</sup> and somewhat less during the next decade.<sup>13</sup> Starting from the mid-1980s, recording procedures were written down for the purpose of instructing new staff members, and in 1991 the recording system was described in the first edition of the Dor Staff Manual.<sup>14</sup>

Even though one of the present authors was involved with this project since its initiation, from the point of view of recording technology, 1980 might as well be antediluvian. Naturally, no computers were available (the first Dor database, which was alphanumeric, was established in 1989). All record-keeping was of the pen-and-paper type. There were no photocopiers, either: all sketch plans (e.g., daily top-plans) and detail drawings were manually retraced (blueprints of the main phase plans were produced after the season). All records were one-off; if a record was lost or destroyed, it was gone. Digital photography was two decades into the future, and even fast developing and printing of film was not locally available. Photographs were taken once a week, and it was necessary to wait until the next week to see them developed. The contact prints were then pasted into albums. The first site grid (actually, not a full grid but a master N–S line) was established with surveying rods aligned with a compass and distanced with a tape measure. It had to be redone as early as 1982 with a theodolite, but it was not until the mid-1990s that an Electronic Distance Measurement (EDM) instrument was used on site (mounted on an optical theodolite).<sup>15</sup> For reasons of backward compatibility, our current Dor expedition still uses the arbitrary site grid (and site datum) established in 1980.

---

<sup>11</sup> Mazar 1997, 15–19.

<sup>12</sup> Sharon 1995b.

<sup>13</sup> Zorn, Sharon, and Gilboa, in press.

<sup>14</sup> Zorn 1991.

<sup>15</sup> Electronic Distance Measurement (EDM): an instrument that uses reflected light (infrared, laser, microwave) to calculate distances to remote objects; the models produced in the 1970s through the 1990s were mounted on top and externally connected to a theodolite. Newer models, in which a digital theodolite and an EDM are integrated within the same instrument, are known as “Total Stations.” Theodolite: a surveying instrument with a rotating telescope for measuring horizontal and vertical angles.



1980 was still the Age of Innocence in other respects, Although David Clarke had already proclaimed Archaeology's loss of innocence in 1973,<sup>16</sup> Near Eastern Archaeology was still blissfully normative in 1980. The impact of the "New Archaeology" would only be fully felt later in the decade, followed closely by post-processualism.<sup>17</sup> These powerful swings of the theoretical pendulum should have had—but, we will argue below, did not have—an effect on the way data was observed and recorded in Levantine Archaeology.<sup>18</sup>

The present expedition was launched in 2003, at which time the recording system was again revamped (by one of the present authors) to accommodate all the above technologies.<sup>19</sup> As part of our mission, we set out to locate all the available records of prior excavations. Fully digitizing them proved infeasible, however. Handwritten and hand-drawn records were scanned and attached as auxiliary documents to the appropriate database entities.

Nor are the problems of record consolidation limited to legacy data. In 2016, the land and maritime operations were reunited as Assaf Yasur-Landau joined our team to relaunch Avner Raban's operation. This not only necessitated the incorporation of Raban's data but also introduced problematics relating to (shallow) undersea excavation. For example, while working with loci is in principle possible underwater, keeping them bounded may prove difficult, as the sands shift in the bottom of the bay from day to day.

In 2017, working with the Israel Nature and Park Authority and the Israel Antiquities Authority (IAA) in anticipation of the opening of the Dor National Park to the public, we conducted a small test excavation north of the tell. For this excavation, we decided to use the IAA's DANA registration system.<sup>20</sup> DANA is in many respects a more advanced database than ours, but a facile data import is not one of its merits. Therefore, it was decided it would be easier to export the data from DANA to the Dor system than vice versa. This introduced new

---

16 "The loss of disciplinary innocence is the price of expanding consciousness; certainly the price is high but the loss is irreversible and the prize substantial" (Clarke 1973, 6).

17 New Archaeology: an approach, first advocated in the 1960s, that argued for an explicitly scientific framework of archaeological method and theory, in which hypotheses are rigorously tested rather than simply described (Renfrew and Bahn 2008, 582). Post-processualism: a number of developments in archaeological theory of the late 1980s and the 1990s; they are characterized by the denial of all or part of the positivistic movement of the 1960s–1970s and pursuing interpretational approaches (Renfrew and Bahn 2008, 491).

18 Levy and Holl 1995, 4–5, and references therein.

19 An up-to-date version of the Dor staff manual is available at <[http://dor.huji.ac.il/staff\\_manual.html](http://dor.huji.ac.il/staff_manual.html)> (accessed June 12, 2017).

20 Digital Archaeology and National Archive (DANA): software for archaeological field recording developed by the Israel Antiquities Authority.

compatibility issues that should be resolved systematically if similar enterprises are to continue in the future.

Currently, we are also initiating wider collaborations. Several projects working in the northern coastal region are forming the “Southern Phoenicia Initiative” to jointly pursue goals broader than those of the individual sites. Some of the Dor staff, including one of the present authors, are pursuing spinoff projects under this umbrella. These not only introduce further data-transfer and foreign-database-access issues, but they also require that typologies, chronological schemes, and other working protocols (e.g., sifting protocols) of the participating expeditions be at least commensurable. In the latter case, ontologies or thesauri enabling the translation of one project’s terminology into another’s need to be constructed.

Underlying the issues particular to the Dor project are challenges facing the digital-humanities community as a whole: How can digital data be curated for the long term? Can interoperability among different registration systems be achieved? Can meta-databases (for Archaeology as well as for other disciplines and interdisciplinary projects) be constructed on a national or transnational scale?

Because our work at Dor must tackle so many of the abovementioned issues, it is an appropriate starting point for thinking about the conceptual basis of archaeological data recording. What, if any, are the common denominators of *all* excavation-recording systems, or at least those of roughly compatible complexes, such as Near Eastern tell excavations? What should be the logical structure of a database that might accommodate all the permutations of such systems? Such an inquiry is quite technical, but it touches the very theoretical foundations of our discipline.

### Theoretical Background

The notion of a hierarchy of theoretical concepts, with observations at its base and “grand” or “unified” theories at its apex, is part of the so-called “received view” of the logical-positivist school of philosophy of science of the early- to mid-twentieth century.<sup>21</sup> Each level of this hierarchy forms an interpretation of the level[s] below and introduces increasingly abstract theoretical terms. Logical positivists used this notion in their (unsuccessful, many would say today) quest to rid science of metaphysics and to anchor “non-observables” such as “energy,” “gene,” or “society” solely in empirical facts in a rigorously logical

---

21 Losee 2001, 159–160, 171–172.

manner.<sup>22</sup> Low-level theory, in this layered model of the structure of science, involves the initial mapping of data onto primitive theoretical entities (e.g., “*length* is the property measured by a ruler”) as well as protocols and procedures of data retrieval (“length should be measured from one end of an entity to the opposite end”) and summaries of data (e.g., statistical analysis).

A parallel and probably dependent model, positing a hierarchy of theories, developed in the field of social theory, closely associated with the work of sociologist Robert K. Merton. Just like the positivist philosophers (who mainly used physics as the model of a “mature” science), Merton was bothered by the disjunction between “low” and “high” theory.<sup>23</sup> He argued that no amount of “low theory”—descriptive, empiricist, or statistical studies of individual societies or social attributes—can by itself produce “grand” theory—the various “isms” of social thought. These are mostly a product of social philosophy rather than of empirical study and are often produced with very little (or even erroneous) factual support. To counter this deficiency, Merton advocated a program of “middle range” theories. These should be closely tied to empirical studies and produce testable propositions. They should not be total theoretical systems that attempt to cover all aspects of social life with a limited set of “universal laws.” In time, though, they might develop into a unified theory by a process of inclusion, whereby a “higher” theory is proposed that subsumes several “lower” theories as special cases.

Just how much each of these conceptions of hierarchy within theoretical constructs affected Lewis Binford’s concept of “middle-range theory” is not quite clear.<sup>24</sup> In his initial article, he refers to neither. Nonetheless, in view of his disciplinary commitment to anthropology and his abiding loyalty to the tenets of logical positivism,<sup>25</sup> there can be no doubt that he was well aware of both. Binford did give middle-range theory a uniquely archaeological twist: the particular challenge of archaeology is reconstructing the behavioral patterns of ancient societies from material remains (in Binford’s language “inferring dynamics from statics”). Thus, for him, the ultimate aim—general, or “high,” theory—is providing an explanation for cultural change (others would say this is simply anthropological theory). Systematics of the material record *per se* (e.g. stratigraphy, typology, etc.) is for him apparently the domain of “low” theory.

---

22 For an example of a logical positivist, see Hempel (1952).

23 Merton 1949.

24 Binford 1977.

25 “Archaeology is Anthropology or it is nothing” (Binford 1962, 217, quoting Willey and Phillips 1958, 2). “The accuracy of our knowledge about the past can be *measured*” (Binford 1968, 17).

We say “apparently” here, because Binford, significantly, never quite defines it. His proposed middle-range theory should be the linchpin connecting the two: a set of “laws” specifying how human dynamics can be “read” in the static archaeological record.<sup>26</sup> For Binford, that meant primarily ethnoarchaeology. Only in the present, he argues, can one observe how dynamics are statically encoded. However, certain other types of research (such as site formation processes and experimental archaeology) are also within the scope of middle-range theory.

These connected, but not identical, definitions of “middle range” have been a source of some confusion.<sup>27</sup> Raab and Goodyear assert founders’ rights for the use of “middle range” in Archaeology, and hence that any definition different from theirs (i.e., Merton’s) is mistaken. Because of this “confusion” (and due to the lack of any explicit definition of what constitutes “low-range”) the scope of the latter is also vague. There is no doubt that the subject of our discussion here, registration systems—how practitioners view, denote, and record their data—should be considered “low” under almost any definition. Note, however, that even a simple label like “Wall X” (rather than “Unit X: consolidated stone depositional feature, the nature, function, and meaning of which is to be determined”) assumes human builders and human intent. It thus cuts right through Binford’s hierarchy without stopping at the middle-range level. Likewise, the definition of some category of artifacts as a “type” involves all kinds of assumptions of what types of attributes and variability people might find relevant.<sup>28</sup> Another source of ambiguity is the overlap between “low level theory” and “methodology.” A method is a way of doing. It requires some active, manipulative aspect, which not every theoretical construct needs to possess. Also, there are methodological issues with little or no theoretical baggage (e.g. the way in which the site grid is labeled, or how to balance a dumpy-level). Nevertheless, the two fields largely intersect and there can be no clear dividing line where method ends and theory begins. Yet, the fact that “method” and “theory” are put into different compartments in the archaeologist’s toolbox often leads to dissonance between one’s views of [high] theory and the field methodology one uses.

---

26 Static archaeological record: one of the keywords of the middle-range theory in Archaeology; it refers to all findings and the fact that, although uncovered almost simultaneously and in close physical proximity, they represent some diachronic reality.

27 Raab and Goodyear 1984.

28 Artifact: any portable object used, modified, or made by humans (Renfrew and Bahn 2008, 578).

One of the present authors surveyed the historical development of registration systems in Levantine Archaeology.<sup>29</sup> She noted that the development of the method and theory did not follow the same path. The “high” theory professed by practitioners tended to reflect—usually with some time lag—the changing fashions in general archaeological thought. These often vary between intellectual generations, with students taking stances that oppose those of their professors. With registration systems, on the other hand, practitioners tend to adopt whatever schemes they were first exposed to. Change, when it does occur, occurs by accretion. New technologies or techniques are absorbed into existing systems *ad hoc*, in whatever manner would least disrupt the extant framework.

The “New Archaeology” of the 1960s to 1980s vaulted scientific methods and exact measurements over impressionistic description and interpretation. A highly structured registration system, in which each observation is compartmentalized to a mandatory field and reduced to a measurement or a code, would fit such a positivistic vision much better than a rambling journal. However, by the time technologies of data collection and processing (such as digital measuring devices and relational databases) were readily available in the field, “high” theory was switching to post-processualism.

At the same time, the debates between processual and post-processual ideas were waged primarily over the high- and middle-range theory,<sup>30</sup> and they rarely reached the basics of method.<sup>31</sup> This is odd, considering that one widely recognized flaw of the positivist program is in the realm of “low” theory. According to many, and not only rank postmodernists, the fact: theory dichotomy, upon which the “layered” model of scientific language is based, is a fallacy. Positivism failed in establishing formal criteria to distinguish between observation and interpretation (or “fact” and “theory”): “All our language is thoroughly theory-infected ... The way we talk, and scientists talk, is guided by the pictures provided by previously accepted theories ... Hygienic reconstructions of language such as the positivists envisaged are simply not on.”<sup>32</sup> Kaplan fa-

---

29 Matskevich 2015.

30 Processual (in Archaeology): during the 1960s and 1970s, an approach to interpreting the past via generalized cultural processes; it is based on a belief that the dynamics of the development of societies have regularities that can be applied to chronologically and geographically remote instances.

31 See Hodder (1997, 691–693) for an exception that proves the rule, with a few further references therein.

32 Van Fraassen 1980, 14. See also Kelley and Hanen (1988, 8–16) and Losee (2001, 178–180).

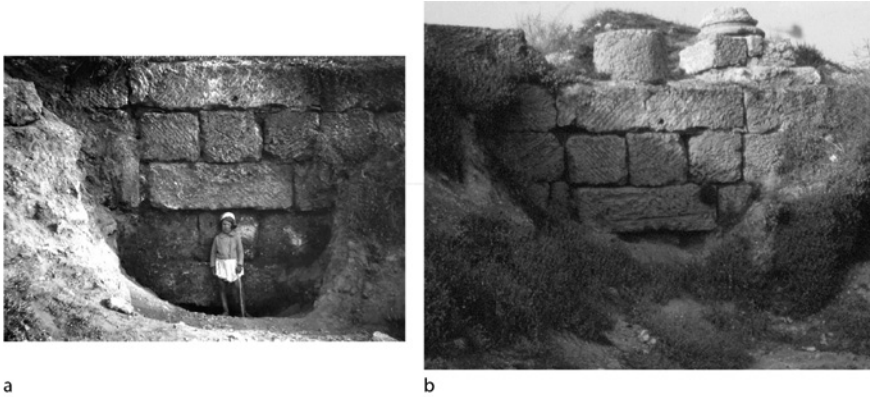


FIGURE 1.1 (a) A photograph from the archives of the Palestine Exploration Fund (PEF-P-1660), from a collection pertaining to Garstang's excavations at Dor (photographer unknown, 1922, courtesy of the Palestine Exploration Fund). Compare to (b) a photograph of W8424 according to the current project's recording system, identified as the north wall of the temple; Wall 2 on Garstang's plans (PHOTOGRAPH BY ISRAEL HIRSCHBERG, 1986, COURTESY OF THE TEL DOR PROJECT)

mously called the positivist view of “data” as being objective and theory-independent “the dogma of immaculate perception.”<sup>33</sup>

The analog for field recording systems is that if one rejects “the dogma of immaculate perception,”<sup>34</sup> one need not be overly hygienic about calling a spade (artifact type) a spade, or a wall a wall. To paraphrase Binford,<sup>35</sup> we have no way of characterizing the statics without imagining the dynamics that produced them.

This ties in with another central concern for post-processualists: the denial of a single privileged (scientific) point of view.<sup>36</sup> If one uses a form-based recording system—such as single context recording or a locus card based system, one is implicitly limiting the recording of any attribute to preplanned fields,<sup>37</sup> each accommodating a single value. In most cases, the form would not even identify who made the observation, when, and under what circumstances. If one's persuasions are of the positivist kind, this is par for the course. There is but one correct observation for every field and—assuming it was

33 Kaplan 1998, 131–136.

34 Slaatte 1979.

35 Binford 1983, 19–30.

36 Hodder 2001, 3–5.

37 Attribute (in Archaeology): “a minimal characteristic of an artifact such that it cannot be further subdivided” (Renfrew and Bahn 2008, 578).

made by an objective observer—the who and when do not matter. In a journal-based system, on the other hand, different log-keepers can record the same attribute or relation differently in their notebooks; and they get to elect which attributes they deem worthy of recording. The same person can even record the same observation differently in different journal entries. This should be grist for the post-processualist mill but anathema to the processualists. And yet our survey has shown that there is no correlation between the type of recording system practitioners use and the paradigm they profess. There seems to be a general trajectory toward form-based systems due mainly to the fact that there is readily available technology to support them.<sup>38</sup>

Most textbooks on methodology, such as Barker for British Archaeology,<sup>39</sup> Joukowsky for the Near East,<sup>40</sup> and Hole and Heizer for American Archaeology,<sup>41</sup> seem to be explicitly or implicitly positivist/processualist in outlook. In the matter that concerns us here, they share the conviction that interpretation can and should be separated from observation.<sup>42</sup> The field archaeologist's primary task is to objectively observe and meticulously register "the archaeological record." Interpretation has its place, but it should operate at a higher level than the primary observations. Indeed, it is only the integrity of the observation and recording that ensures the testability of propositions in whatever interpretive framework one chooses. This view is underpinned by an implicit axiom of intrinsicity—i.e., that archaeological entities (e.g., finds, contexts, deposits) have intrinsic qualities that define what they are, independent of who digs them up.

After more than twenty years of post-processualist archaeology, we have yet to see a post-processual field manual. Interpretative archaeology refers to method on some occasions, while talking about "interpretation at the trowel's edge"<sup>43</sup> or emphasizing the importance of the area supervisor's diaries for achieving multivocality in the record.<sup>44</sup> Yet these glimpses are no replacement for textbooks or comprehensive field manuals instructing how to plan a site, what units to use, what artifacts to collect, and how to sort and record them in the view of new theories.

---

38 See also Hodder (2001, 4).

39 Barker 1982.

40 Joukowsky 1980.

41 Hole and Heizer 1969.

42 Barker 1982, 146.

43 Hodder 1997, 92–98; Bender, Hamilton, and Tilley 1997.

44 For example, see Hodder (1997) and Hodder (1999, 121–123).



The positivist interest in the genre of methodological manuals is natural. At the bottom line, their view is prescriptive: there is a right way and a wrong way to do science.<sup>45</sup> Instructing acolytes on the correct way is therefore the duty of those who have mastered it. Post-processual reluctance to play that game reflects an existential dilemma: “any notion of a general methodology ... could conflict with approaches which emphasize critique, interpretation and multivocality.”<sup>46</sup> Carried to its logical end, this approach means each worker gets the freedom to choose what to excavate, how to excavate it, what to keep, and how to tag it.

Few, if any, actual archaeological field projects go the full monty with their postmodern convictions. Hodder paints an idyllic picture,<sup>47</sup> in which a laboratory specialist tours the field once or twice daily, “empowering and informing” the field team;<sup>48</sup> a video crew is always on call to film these encounters; and a staff anthropologist critiques the interaction in real time. This is arguably possible in the rarified atmosphere of a top-percentile project of a celebrated site, in which world experts vie with each other to be included. The implementation of “methodological freedom” no doubt depends on the fact that all of these experts and most of the field staff are highly qualified professionals who have learned their craft in other—most probably positivistically oriented—projects. They thus possess a deep understanding of what is reliable and useful data and what might be the requirements of various other team members in terms of data retrieval and recording. Significantly, while Neopagans, local communities, and the world at large are invited to join the *discourse* about Çatalhöyük via virtual reality modeling or other so-called “experiential” techniques, it is only members of the “core group” who are allowed to “undertake ... fieldwork, research and publication.”<sup>49</sup> Moreover, their reflexive musings are recorded on top of, and not instead of, a standard codified single-context system.<sup>50</sup>

In the following sections, we propose a meta-model that supports an “agnostic” approach to field recording.<sup>51</sup> The data recorded via this model can be successfully used by practitioners of all theoretical schools. It also is meant to

---

45 See Losee (2001, 265–276) on prescriptive versus descriptive philosophies of science.

46 Hodder 1997, 691.

47 Hodder 1997, 691.

48 Hodder 1997, 695.

49 Hodder 1997, 698.

50 Hodder 1997, 696.

51 Meta-model: a model that describes a model. More specifically, it is a high-level abstraction that uses a modelling language to describe a model, which, itself, is defined as an external and explicit representation of a system or a part of reality.



suit legacy data of all sorts and to ensure its compatibility with new, digitally born records. In order to do this, we have had to strip archaeological recording systems to their basics and examine their elementary components. We are also purposefully nonjudgmental. We are not asking what components a good registration system should have (a futile quest in the case of legacy systems) but what are the common components that all archaeological registration systems possess.

## Methodology

### *Definition and Basic Components of Archaeological Recording Systems*

Archaeological recording is a domain of various mixed-and-matched methods, tools, and documents. In seeking the atomic elements that all such systems share, we must first distinguish between syntactic elements and semantic entities. Syntactic elements are the components of the record itself, irrespective of what is being recorded. The most basic of these is the individual recording event. This event is a piece of content (visual, textual, numeric measurement, for example) recorded by a particular observer at a particular time and place that conveys information concerning one or more semantic entities.

That a recording system is made out of records is a bit of a “duh?” statement. Consider, however, Figure 1.1a. By itself, it is a picture of a child and a wall, and it is archaeologically meaningless. If we would add information culled from where the picture was found and (our identification of) where it was taken—metadata for this image—it becomes an observation or a recording. Note that the time and the observer are not precisely known in this case, but even their approximation has given meaning to the record.

Consider further Figure 1.2: the base of a Cypriot Monochrome vessel from the Israel Antiquities Authority’s National Treasures. It is purportedly from Emmanuel Anati’s excavations at Tell Abu-Hawam in 1963. On the sherd are several notations, in different ink and different handwritings. The first, probably written during or immediately after the excavation, reads “N wall | s square, SE corner | 1.25 to 1.80m ↓ surface.” Note that architectural features (N. Wall) and grid squares are used to roughly locate the find-spot, but there is no attempt to use loci as strictly bounded units. Also, the object itself lacks an identifier (ID).<sup>52</sup> The next tagging event (in black ink) was probably the assignment

52 The identifier (ID) of an object is expressed in a label that consists of a combination of any alphanumeric characters attached to that object. Most data-management systems (and most archaeological recording systems, be they manual or computerized) require

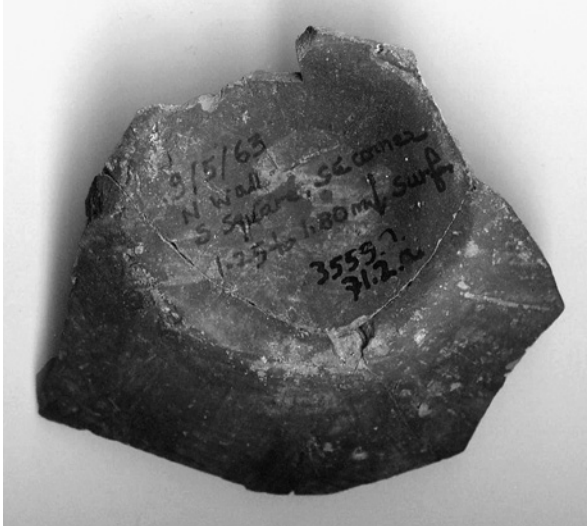


FIGURE 1.2

*Base of a Cypriot Mono-chrome vessel from the Israel Antiquities Authority's National Treasures, purportedly from Emanuel Anati's excavations at Tell Abu-Hawam in 1963*

(PHOTOGRAPH BY SVETA MATSKEVICH, COURTESY OF THE ISRAEL ANTIQUITIES AUTHORITY AND THE DIGITAL ARCHAEOLOGY LAB)

of an inventory or collection number. The appellation 71.2 fits the format that Jacqueline Balensi used when re-cataloging the finds from Tell Abu Hawam for her PhD.<sup>53</sup> The number 3555 (or 3559?) may be an IAA inventory number. These identifiers, too, were amended at some point with lighter additions. The most recent addition is the red dot that marks that the sherd was 3D-scanned in the Digital Archaeology Lab in the Hebrew University, from which we obtained the image.

That someone (or at least two people, in this case) re-identified and marked the artifact may or may not be significant for our own assessment. Note, however, that in most computer-based systems today, the whole history of recording events is lost. In most systems, it is only possible to record one observation per field and correcting that field would delete the previous entry in it.

The distinction between the “recording event” and the “record” accords with the post-processual preception that an “observation” is always an interpretation made in a specific context.<sup>54</sup> The implementation of the meta-system pro-

---

that any entity has a unique ID. In this volume, see a unique “ID” applied to a) texts analyzes in TEI-XML, Bigot Juloux, 165n70, and b) a database, Pagé-Perron, 204.

53 Balensi 1985.

54 “Context” here should be taken as both the scholarly context in which the interpretation is made and the ancient context that it is made about. Calling a feature made of consolidated stone “Wall X” should be more or less automatic if one is excavating a Roman site. If the site is Paleolithic, caution would be warranted, and one would be well advised to reserve judgement.

posed herein onto a positivistically oriented recording system may choose not to fill in the metadata (who, where, when), thus “objectifying”<sup>55</sup> the observation into a generic record.

### Basic Semantic Entities in a Recording System

The semantics of a recording system is what is being recorded, irrespective of syntax. Only two types of archaeological entities are common across the whole spectrum of existing recording methods, systems, and schools. The two mandatory pieces of information that must be recorded by an excavator in every system are a find designator and the spatial unit from which that find originated. Many systems have more than one unit of either type.

In trying to define the semantics of entities in a given system, we must consider both formal and denotational aspects of these entities. The formal facets of a spatial entity are its geometric or topological properties. The denotational facet is concerned with what it is that the abstract unit attempts to model.

### The Topology of Spatial Entities

Is the unit bounded? In all contemporary registration systems that we know of, the basic spatial unit (locus, context, spit) has—at least in principle—well defined edges. So it is hard to imagine that this is not a necessary condition, or that it has not always been the case.

Like most Levantine excavators between the two World Wars, Garstang used artifact provenience designations such as “below the sand layer” and “near Wall X.”<sup>56</sup> Certain features were named or numbered, but these served as locational “hotspots” rather than as polygons with borders between them. As Figure 1.2 shows, such “impressionistic” recording systems persevered, in some cases, into the second half of the twentieth century. Indeed, even in contemporary systems non-primary spatial entities might be fuzzily defined. When referring to “Temple F” in Figure 1.1, do we mean the entire precinct or only the

---

55 Meta-system: data description system external to a data-management system; a system that describes a system. For positivists, the state of objectivity is the ultimate, achievable goal. They would argue that if rigorously recorded, an “observation” turns into “data.” Objectifying would then mean rendering it objective. Since objective data is by definition timeless and free of point of view, the meta-data (who recorded? when? under what circumstances?) is redundant. Post-processualists either despair of ever reaching objectivity or do not believe it is desirable in the first place. For them, stripping the interpretation from its context (the metadata) “objectifies” it in the sense that it degrades the subject of interpretation to an object.

56 For example, see Garstang (1924b).

temple podium? The answer will probably depend on the context of the reference. As an entity, “Temple F” might not be bounded.

Are the units mutually exclusive and/or exhaustive? If this is so, then every point in (the excavated) space must belong to one and only one spatial unit. This is the case, for instance, with the set of grid squares. Note that units can be defined as exclusive but not exhaustive (for example, a coin found in the dump may legitimately have no locus).

Again, a look at higher-order units might prove revealing to those of us working in single-context or locus systems where the primary units are by default exclusive: in an attempt to clarify the term “Temple F,” I might define “Precinct F” and “Podium F.” Both of these are second-order or aggregate units (i.e., they can be explicitly specified as a list of [primary] loci). However, “Precinct F” would contain “Podium F.” The intersection between two features in this case is not empty, and thus “feature” is a non-exclusive spatial unit.

Are units by definition contiguous? A curious feature of Kenyon’s excavation at Jericho is that her layers need not be.<sup>57</sup> If two disparate deposits, even far away from each other, were judged to originate from the same depositional event, they were given the same number, and apparently finds from them were placed in the same container (the user of the report, at least, cannot spatially differentiate between them).

#### Denotational Classification of Basic Spatial Entities

A spatial unit can be arbitrary, depositional, or behavioral.<sup>58</sup> Arbitrary units are used as primary spatial designators mostly on sites that lack any architectural remains or other clear spatial features that would allow for horizontal subdivision of an excavation area and defining vertical layers.<sup>59</sup> The standard spatial unit in most prehistoric excavations in the Levant today is the spit: a  $.5 \times .5 \times .05$  m volume, consisting of a quarter of a  $1 \text{ m}^2$  grid square, carried down for 5 cm. Flint finds from a spit (Fig. 1.3) receive the unique identifier (here P28b/525-520) of the basic volume unit: (P28 is the grid square, quadrant b, elevations 525-520). We put these chips and chunks together in one bag because, at least initially, we wish to study them as a group. We might weigh the bag, count the number of stones in it, etc. None of these actions requires that the pieces in the bag be individuated, and it is a waste of time and effort to do so. Arbitrary units (grid squares) are often used as secondary, or higher-order, spatial designations in sites with architecture, too.

<sup>57</sup> Kenyon 1981.

<sup>58</sup> Matskevich and Sharon 2016.

<sup>59</sup> Hole and Heizer 1969, 100–111; Roskams 2001, 212–216.

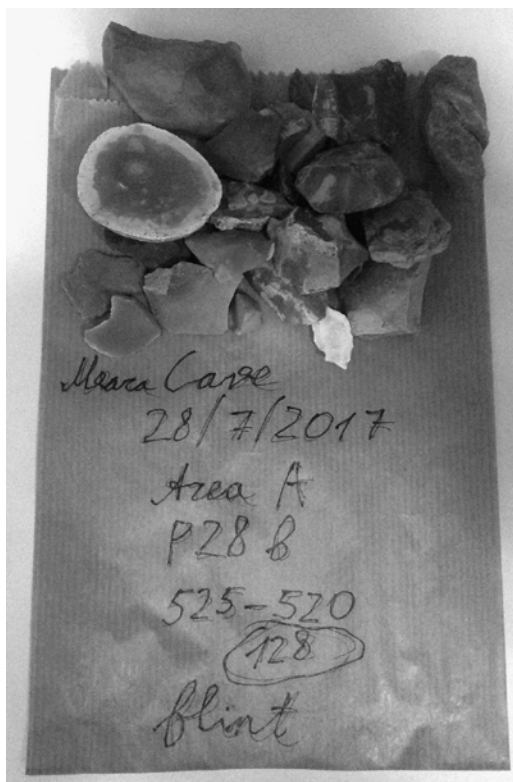


FIGURE 1.3  
*Flint finds from a spit* (PHOTOGRAPH  
 BY SVETA MATSKEVICH)

The de facto standard definition of the “locus” in Levantine Archaeology today on sites with architecture is a contiguous segment of the site volume interpreted as being the result of a single depositional event. This is what we call the depositional denotation of the locus.<sup>60</sup>

This was not always the case, however. The classic “Locus to Stratum” model initially defined the “locus” as an architectural space (Fig. 1.4).<sup>61</sup> Features other than rooms (e.g., tombs, pits, and installations) often obtained “locus” designations as well; but the denotation of a “locus” under this system is always a space within which some specific human activity took place. It is thus a unit of (ancient) behavior, according to the interpretation of the excavator. Among the

60 Sharon 1995a, 22–23.

61 Stratum: a single construction–destruction cycle within a stratified sequence of deposits that form a multi-layered site (e.g., tell). About the “Locus to Stratum” model, see Lamon and Shipton (1939, xxiii–xxiv). Some rooms on this plan have a locus number, but open spaces and areas where the architecture is not well preserved typically do not. Walls, floors, pits, and installations are drawn but not numbered.

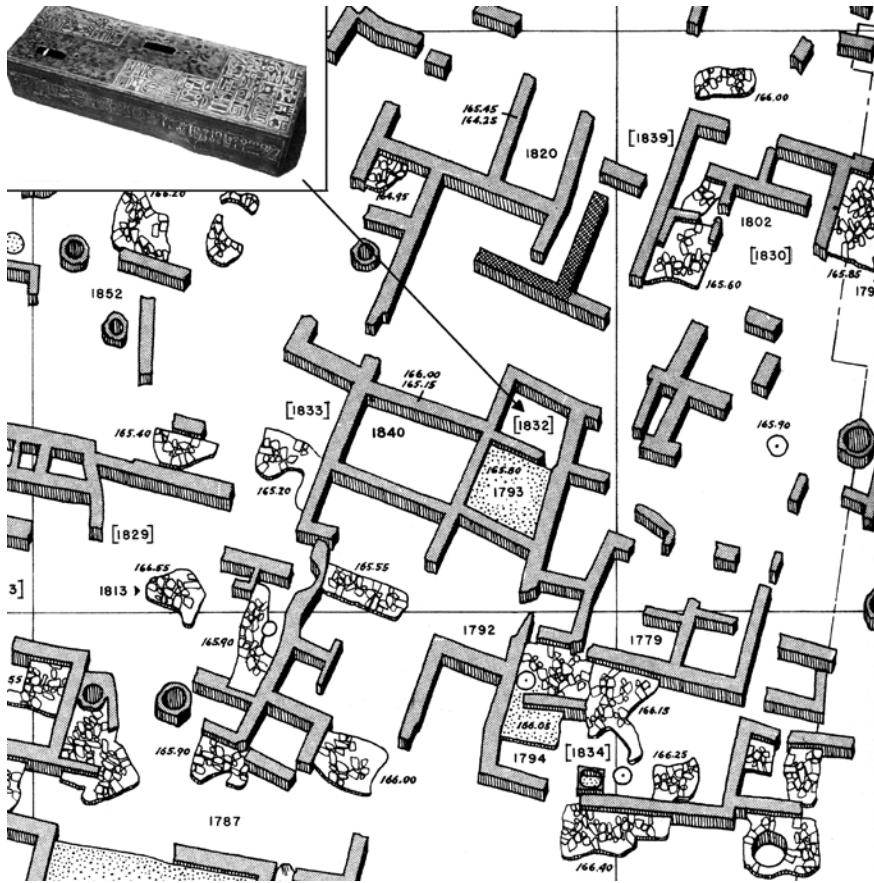


FIGURE 1.4 *Megiddo, Area CC, part of the Stratum VII B and VII A plan (Loud 1948, fig. 409)*

most famous examples of employing these units are excavations at Beersheba,<sup>62</sup> Tell Jemmeh,<sup>63</sup> and to some extent Megiddo of Yigael Yadin.<sup>64</sup>

Tracing the paths by which the “locus” evolved from a behavioral unit to a depositional one in the archaeology of the Near East is complicated by the fact that excavators rarely explicitly defined their units, or even acknowledged that their “loci” are different from those of their predecessors. This issue, however, is outside of our scope. Suffice it to say that there is no direct influence between this evolution and that of the “context” in British and European archaeology, although the result is similar. It has been claimed that depositional units

62 Aharoni et al. 1973, 119–120.

63 Van Beek 1988, 158.

64 Zarzecki-Peleg 2016.



are observational,<sup>65</sup> while behavioral ones are interpretational, and hence that the former are “scientific” while the latter are not. This is patently untrue. What constitutes an “event” within the depositional process is strictly in the eyes of the beholder. Indeed, our continued involvement with micromorphologists on the tell teaches us that each depositional “event” can be broken into sub-events,<sup>66</sup> all the way down to the molecular level.

There is, however, one way in which depositional units can be seen as more basic than behavioral ones. An implementation where the basic volume unit is depositional can define features (behavioral units) as higher-order entities. In as much as the construction and use of a feature will result in the deposition of at least one (positive or negative) layer,<sup>67</sup> it can be defined as an aggregate of at least one but usually more primary depositional units.

### Semantics of the “Find”

It seems self-evident that any archaeological registration system needs an entity to model, identify, and tag the artifacts and ecofacts that are found.<sup>68</sup> However, as more and more classes of items are being studied in Archaeology, defining what constitutes a “find” proves to be less straightforward than it might seem.

In addition to the individual find, many systems also have ways of tagging batches of objects. Such is the “basket” in many tell excavations (reminiscent of the old days when potsherds from the locus were collected in wicker baskets—nowadays usually plastic buckets). Similarly, most prehistoric excavations in the Levant would individually save flint tools, tagging each of them with its exact coordinates, but keep all the debitage from the spit together in one bag (Fig. 1.3).

In our quest for primitives, however, we ask: which of these entities can be considered a special case of the other? Somewhat counterintuitively, we contend that it is the batch and not the artifact that constitutes the basic unit of “finding.”<sup>69</sup> An (individual) object is merely a sample of size one. Note also that

65 For example, Chapman 1986.

66 Shahack-Gross et al. 2005; Shahack-Gross 2011.

67 Layer (in Archaeology): a unit of sediments in a stratified archaeological site, created as a result of one of the site-formation cycles. “Negative” layers are the result of the removal of sediment (for example, the outlines of a pit would represent a negative layer, while the sediment within the pit would be called a layer).

68 Ecofact: non-artifactual organic (botanical or biological) or inorganic (resulting from geological processes) object recovered from an archaeological context.

69 Batch: a group of items dealt with at the same time or considered similar in type

for many types of finds (such as a sediment sample) the “count” attribute is simply irrelevant. The sample size in such cases might be weight, or some other attribute.

There are often cases in the archaeological workflow that require the re-identification of, for example, an artifact, ecofact, or batch. The potsherd in Figure 1.2 was apparently not given a unique identifier in the excavation (unless the whole story written on it is considered its ID). Later events in its life history (including cataloging and storage) had—literally, in this case—left their mark by giving it different numbers.

Another usual case is sub-sampling from a batch. Consider the bag of debitage in Figure 1.3. Suppose we happen to choose two chunks from this bag for a provenience study of a random sample of the raw materials on site. As these two particular pieces will now each be subject to a battery of tests, they need to be individually named. A similar case might occur if we split an olive pit and sent the two halves to two different radiocarbon laboratories for an inter-calibration study. What was once an ecofact is now two ecofacts, which will have two different (hopefully similar) records in our registration system.

The opposite case can also be found: a new aggregate entity that is different from the set of its parts. Mending or refitting, for instance, can create a “find” that will need to be recorded in its own right. This new object, however, may be made of several different objects, possibly from different find-spots. A somewhat extreme case is illustrated in Figure 1.5. In 2004, S. Rebecca Martin managed to refit a fifth-century red-figure krater sherd found in Stern’s excavation at Dor in 1993 (Dor Area F, L8605, Reg. no. 85234) with another piece found in Garstang’s excavation in 1923–1924 (Reg. no. P 2875). Note the mending hole that indicates the pot was already broken in antiquity. Martin also identified the scene: Heracles in the Garden of the Hesperides. This would not have been possible based on either piece alone.<sup>70</sup> Thus the aggregate entity has additional attributes, over and above those of its individual constituents.

Even the differentiation between a spatial unit and a “find” is less trivial than it might seem at first glance. Consider the case of a rock-cut olive press versus a portable one. The rock-cut installation would probably be recorded as a spatial unit, because it has fixed coordinates and it is the context for olive pits found on its surface. A portable press, on the other hand, could be taken out of its context. Therefore, it would probably be treated as an artifact. It would get a find ID and possibly be moved to a museum or artifact storage space.

---

(*Cambridge Dictionary*, s.v. “batch,” <<http://dictionary.cambridge.org/dictionary/english/batch>> [accessed June 20, 2017]).

70 Stewart and Martin 2005.



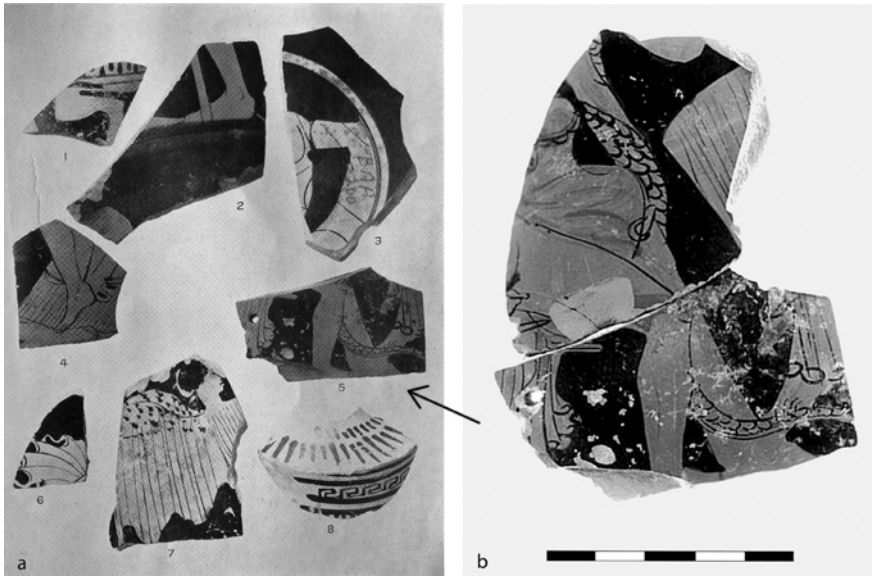


FIGURE 1.5 (a) Attic red-figure sherds from Garstang's excavation (Iliffe 1933, pl. 7b:5) (b) The refitted sherds (Stewart and Martin 2005, fig. 4; PHOTOGRAPH BY GABI LARON, COURTESY OF THE ISRAEL ANTIQUITIES AUTHORITY)

To deal with these conceptual difficulties, we propose that the ontological order between signifier (the tag) and signified (the find) be reversed. The basic entity in the recording system is not the “find” but the “sampling event” (or perhaps more specifically, the “tagging event”). This definition has considerable conceptual impact. There is no intrinsic property that defines a “find.” Rather, it is the fact that the archaeologist picked out some “stuff” from the archaeological record and tagged it for further study. All other permutations—the sub-sampling of a sample, the retagging of an artifact with a different identifier, etc.—can be described as further tagging events. These can be connected (and possibly inherit attributes, such as the find's original context) from previous events in the workflow; but they remain distinct. The tag should be thought of not as a permanent identifier to a persistent object but as a bill of lading that facilitates the archaeological workflow.

That the most basic ingredients of archaeological registration are “events”—the “recording event” and the “tagging event”—emphasizes how early in the archaeological process interpretation starts: not during report-writing or discussions of the dig but the moment something is picked out of the ground. The raw data of an excavation is a collection of interpretations fixed in pre-printed forms, diaries, and tables. The beginning of the existence of an archaeological record is not the freezing of “cultural dynamics” into fossil “statics” but the act

of recording “stuff” on the excavation. This is not to say, of course, that the ancients had no impact on the formation of an “archaeological record” but that it is a co-production of ancient societies and modern archaeologists.

### *Semantic Modeling and Mapping of the Recording Systems*

Conceptual Reference Models for Archaeological Field Recording  
The CIDOC Conceptual Reference Model (CIDOC CRM)<sup>71</sup> is a leading standard (ISO 21127/2006)<sup>72</sup> for conceptual referencing of cultural heritage data.<sup>73</sup> Developed in the last decade by one of the special-interest groups of ICOMOS for museums and object collections,<sup>74</sup> it gained popularity in other domains of the humanities and cultural heritage.

The CIDOC CRM was first adopted into the archaeological domain by English Heritage. After some adaptations and additions (the modified version was dubbed CRM-EH), they successfully mapped their field recording systems.<sup>75</sup> Due to the standard’s growing popularity, the initial group of researchers developed several extensions of the basic CRM, aiming at incorporating additional information about the objects, such as bibliographic references, provenience, and other scientific analyses, as well as spatio-temporal models. The archaeological fieldwork extension, CRMarchaeo,<sup>76</sup> accommodates stratigraphy-related concepts that were missing from CIDOC CRM.

A Conceptual Reference Model is a *lingua franca* of recording systems, and the main purpose of ontological mapping is to ensure compatibility of the terminologies used by various data-management systems, which is one of the necessary conditions for creating Linked Open Data clouds.<sup>77</sup>

71 Conceptual Reference Model (CRM): provides definitions and a formal structure for describing the implicit and explicit concepts and relationships within a system (*IGI Global Dictionary*, s.v. “Conceptual Reference Model,” <<https://www.igi-global.com/dictionary/conceptual-reference-model-crm/5242>> [accessed June 20, 2017]). For further discussion, see in this volume, a) Eraslan (299–300), who suggests using CIDOC CRM for features’ vector and semantic schemes; b) Nurmikko-Fuller (353–354), who suggests improving online publishing of cuneiform and photography with CIDOC CRM classes.

72 ISO stands for International Organization for Standardization, which is a non-governmental body that develops and publishes international standards of products, services, and systems.

73 Doerr 2003; Doerr, Schaller, and Theodoridou 2010.

74 International Council on Monuments and Sites (ICOMOS), <<https://www.icomos.org/en/>> (accessed June 20, 2017).

75 Cripps et al. 2004.

76 CRMarchaeo, CIDOC CRM website, <<http://www.cidoc-crm.org/crmarchaeo/home-3>> (accessed June 18, 2017).

77 Isaksen et al. 2010; May et al. 2012. For further information on Linked Open Data, see in this volume, Nurmikko-Fuller, 344.

### SKOS Format for Controlled Vocabularies

The Simple Knowledge Organization System (SKOS)<sup>78</sup> is a model for expressing vocabularies, thesauri, and taxonomies in a machine-readable format, namely the Resource Description Framework (RDF).<sup>79</sup> SKOS is a World Wide Web Consortium (W3C) standard employed for over a decade for linking data in the Semantic Web.

Controlled vocabularies and taxonomies are essential for efficient data retrieval in archaeological practice in general, and in field recording in particular. Presented in SKOS format, they can be integrated into the Linked Data cloud to serve all participant datasets.

### Graph Databases

In a graph model, the data are structured so that each entity (subject),<sup>80</sup> its attribute (object), and the relationships between them (predicate) create a triple.<sup>81</sup> The triples are linked via entities and attributes into a triple store.<sup>82</sup> The main advantage of the graph model is its flexibility, which allows for querying large datasets by setting relations as query criteria. The second and third generations of the WWW intensively use triple storage (RDF format) as a data representation model.

- 
- 78 Simple Knowledge Organization System (SKOS): "an area of work developing specifications and standards to support the use of knowledge organization systems (KOS), such as thesauri, classification schemes, subject heading systems, and taxonomies within the framework of the Semantic Web" (SKOS home page, W3C, <<https://www.w3.org/2004/02/skos/>> [accessed June 20, 2017]). On other languages that are easily machine-readable (or machine-actionable), see in this volume, Bigot Juloux, 163–164; Pagé-Perron, 200; Nurmikko-Fuller, 336, 339–340.
- 79 For further information in this volume, especially for online publishing, see Nurmikko-Fuller, 338–340, 343–344, 352, 360.
- 80 Graph model: a model that describes relations between objects as a collection of nodes connected by edges; graphs are studied and explained by graph theory ("Graphs and Networks," Mathigon, <<https://mathigon.org/course/graphs-and-networks>> [accessed June 20, 2017]). For further information on Graph Theory, see in this volume, Ramazzotti, 66n24, 67n28, 73.
- 81 Triple: a semantic statement that consists of three parts: subject, predicate (property), and object. The statement "Floor A reaches Wall B" is a (stratigraphic) triple. For further explanation of the triple in this volume, see a) Nurmikko-Fuller, 345–347, for online publishing, and b) Pagé-Perron (202–203), who uses triples for network graph.
- 82 Triple store: a database for the storage and retrieval of triples; see Curé and Blin (2015).

In a graph database any entity may be connected to any other.<sup>83</sup> It is up to the individual application to determine if such a connection makes sense. This is not usually the case in conventional relational databases, where relationships between entities are predetermined by the schema.<sup>84</sup>

Following other domains, Linked Data (LD) initiatives in archaeology worldwide use interlinked RDF data stores to create cross-searchable platforms of geographical, historical, and archaeological data.<sup>85</sup> Among the most successful projects are Europeana, Pelagios, and Nomisma.<sup>86</sup>

The major challenges of making archaeological data available on the web in a format that can be referenced from other resources are mapping the datasets to an ontology and converting data from various databases to RDF format.<sup>87</sup> These issues were addressed in a step-by-step process by the team of scholars from the Hypermedia Research Unit at the University of Glamorgan.<sup>88</sup> The STAR and STELLAR projects developed a set of tools that simplify the process of preparing data to be semantically linked.<sup>89</sup> We used the STELLAR.Console util-

- 
- 83 Graph database: see graph model. For a graph data structure in the context of text analysis, see in this volume, Pagé-Perron, 196–197, 198 fig. 6.1, 202, 219.
- 84 Relational database: a computer database in which data are organized in tables, where each table contains all the instances of one entity. Each “tuple” (meaning an ordered list or row) in a table represents one instance of the entity and must have a unique identifier as one of its attributes; the columns (fields) of the table represent all other attributes of the entity, including keys of related entities. The relationships between tables are defined as links between keys. The database schema describes the structure of the database (tables, attributes, and keys) in a formal language supported by the Database Management System (DBMS), software that enables an administrator to create and manage databases and to monitor, modify, and analyze their interactions with users and other applications. See also in this volume, Pagé-Perron, 196, 202, 204.
- 85 For further information on LD, see in this volume, Nurmikko-Fuller, 340, 344.
- 86 Europeana: [www.europeana.eu/portal/en](http://www.europeana.eu/portal/en); Pelagios: <http://commons.pelagios.org/>; Nomisma: <http://nomisma.org/>. All accessed June 18, 2017.
- 87 Ontology (in data sharing): the formal conceptualization of a particular domain that is shared by a group of practitioners of that domain (Maedche 2002, xv). For additional information, see in this volume, a) in the philological field, Bigot Juloux, 165–181, and Prosser, 320–322; b) applied to online publishing, Nurmikko-Fuller, 343, 348–350, 353–360.
- 88 Now part of the University of South Wales. Hypermedia Research Unit, University of South Wales, <http://hypermedia.research.southwales.ac.uk/> (accessed June 20, 2017).
- 89 Binding et al. 2008; May et al. 2012; “Semantic Technologies for Archaeological Resources” (STAR), Hypermedia Research Unit, University of South Wales, <http://hypermedia.research.southwales.ac.uk/kos/star/> (accessed June 18, 2017); “Semantic Technologies Enhancing Links and Linked data for Archaeological Resources” (STELLAR), Hypermedia Research Unit, University of South Wales, <http://hypermedia.research.southwales.ac.uk/kos/stellar/> (accessed June 18, 2017).

ity at all stages of the presented study and wish to express our gratitude to its authors.<sup>90</sup>

### *Selected Issues in Data Integration*

#### Modeling and Linking Legacy Data

Samples of the Dor datasets described above, from the excavation of the tell and from underwater explorations around the tell, were linked using a two-step procedure: SKOSifying their terminologies and mapping the data models to CIDOC CRM.<sup>91</sup>

The first stage is meant to create a common platform for the terminologies used by the two teams. The problem was not only the differences in the terms used to define excavation and interpretation units, but also the fact that the terminologies overlapped. The same terms denote different concepts in the two systems. For instance, on the terrestrial excavation at Dor, and most excavations in the Levant today, an archaeological site is an accumulation of archaeological remains spatially related to each other, and an area is a subdivision of the site (usually several contiguous grid squares, encompassing at least one, possibly several, architectural features). In Raban's terminology, "site" is an extent of excavation that encompasses previously identified large-scale architectural features, exposed (at least partly) by erosion of the tell slopes and shifting sands at the beach or bottom of the bay (for example, "Southern Quay" or "Purple-dye Factory"). Conceptually, this term is equivalent to "area" on the tell. Moreover, rather than any sort of open-area excavation, Raban usually dug small-scale probes in specific points of these large-scale features, such as the corner between two walls, straightening a sea-cut section, etc. Each of these was called an "area." There is no equivalent unit in the terrestrial excavation.

The term "locus" presents a more challenging problem in the data-integration effort. In both recording systems, locus is the basic excavation unit. In the terrestrial excavation, the locus is contiguous, bounded, exclusive, and exhaustive and models a depositional unit.<sup>92</sup> Raban never defined what he calls a "locus," but from his records it seems that the same term was used to describe

90 "STELLAR Applications," Hypermedia Research Unit, University of South Wales, <<http://hypermedia.research.southwales.ac.uk/resources/STELLAR-applications/>> (accessed June 18, 2017).

91 SKOSifying: a process of mapping user-defined terminology (thesaurus or vocabulary) to SKOS. See also in this volume, Nurmikko-Fuller, 353–354. Data model: an abstract model that describes the data structure and relations within a system, where the data is stored, and how it is processed. See also in this volume, Prosser 319–320.

92 Zorn 1991, 3.

a vaguely defined location within an area, often (but not always) bounded by architectural remains or natural features, sometimes identified as loci and sometimes not (i.e., not exhaustive). Therefore, several “terrestrial loci” might comprise one “maritime locus”. Nevertheless, in both cases these are the basic unit of excavation. This means that in practical terms there is little we can do about their incompatibility. For instance, comparing ceramic assemblages from loci from the different recording systems is potentially misleading, but since no smaller subdivisions exist, there is little we can do except note the problem. The conceptual cross-referencing of these terms as equivalent would be incorrect. Mapping both types of loci to the Basic Volume Unit (BVU)<sup>93</sup> entity of the reference model and expressing their difference through their properties can at least make the user aware of their characteristics.

### *Mapping Ambiguities*

While we have enough trouble with various terms used to define spatial units of excavation, there are more complex cases to handle.

Let us return to the case of the rock-cut vs. portable olive presses. In a table-based database,<sup>94</sup> these objects will end up in different tables, the first one as a spatial unit, and the second as a find. Each will be recorded in a different set of forms, and there is (usually) no single query, in which both olive presses can be retrieved. The combined flexibility of an event-driven (rather than a persistent-item) registration system and a graph database can provide a partial solution. Both BVU-s and the finds will be parts of the data cloud,<sup>95</sup> therefore they could be queried together. The other part of the solution is using one united thesaurus of terms for various value lists within the database. A search for the “olive press” entry anywhere in the database will return both contexts and finds that contain this value in one of the description fields (Fig. 1.6).

### *Modeling Multivocality*

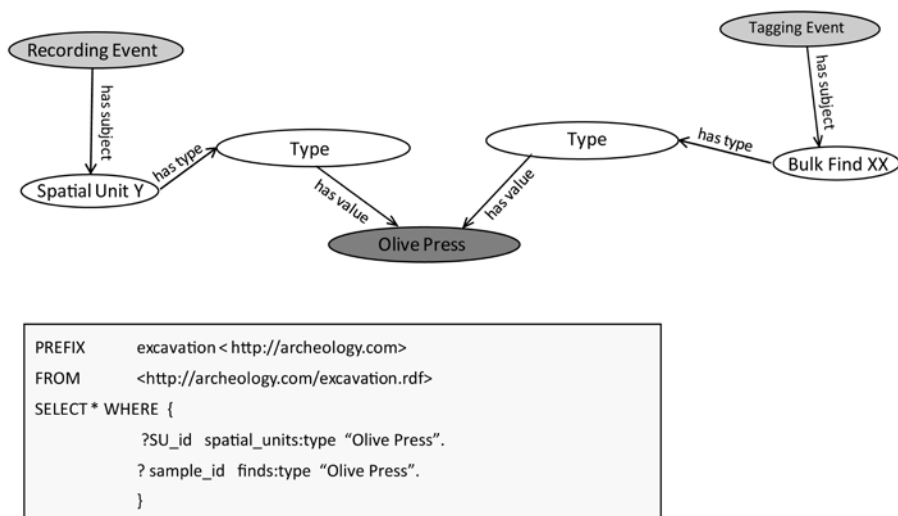
This leads us to another feature that can be handled in a graph data model, namely the multivocality of the archaeological record. The term “multivocality” is used in different theoretical approaches within the discipline or as various layers of perception in communicating archaeology to the public.<sup>96</sup> In this

93 Basic Volume Unit (BVU): the primary excavation unit of a recording system, defined arbitrarily (grid square), stratigraphically (locus, context), or behaviorally (feature).

94 Table-based database: see relational database (note 84, above).

95 Data cloud: visualization diagram for a triple store or multiple graph datasets in the Linked Open Data initiative.

96 See, for example, Habu, Fawcett, and Matsunaga (2008).

FIGURE 1.6 *Modeling ambiguities*

context, however, multivocality is meant as pluralism in the very basis of archaeological process, the initial interpretation of the excavated matter.

Multiple opinions at this level result both from the fact that today archaeology is a team effort, not the one-man shows of nineteenth- and early-twentieth-century projects, and from the constantly growing timespan between the excavation and the publication, which allows for repeated analysis of the finds, using various methods that bring inconsistent results. In the proposed meta-system, there can be (and usually are) multiple recording events to any entity. This means that every attribute assigned to every entity within excavation dataset (except the entity ID) has a timestamp and authorship. This approach allows values to be added to any of these attributes by different people at different times. If today a find dated to the Persian period turns out to be Hellenistic,<sup>97</sup> the first dating need not be deleted and overwritten (although it would be connected to the subsequent “dating event” so a timeline of evolving interpretation can be established). The new interpretation would be added, and the old one would be kept, because it is part of the knowledge-creation process.

In a graph database, the implementation of this approach is as simple as adding two properties to each attribute, without any changes in the data

97 Persian period (in the Levant): 539–334 BCE. Hellenistic period (in the Levant): 333 BCE–63/57 BCE.



structure or alterations in the existing records. Since we do not propose here the actual replacement of all relational databases with graphs but suggest using a graph database as a conceptual model only, the solution for pluralistic records can be only a partial one. By default, relational databases can only support one attribute per field and one-to-many relationships between tables. So multiple relationships between attributes can only be implemented where the underlying schema has been specifically restructured to allow for multiple values and many-to-many relationships (either for all attributes in all tables or for specific attributes in some tables). Other NoSQL solutions,<sup>98</sup> such as document-oriented databases,<sup>99</sup> have a more abstract structure that allows for flexibility in the implementation of the underlying storage for graph databases. They also often enable more efficient data retrieval and easier partitioning of large volumes of data.

Keeping track of various interpretations can be important, because in a domino effect they influence our understanding of similar objects/features found later, as well as the next stages of interpretation of related findings. An example is shown in Figure 1.7. Fragments of large terracotta masks were found in several different Persian period loci in Area D at Dor in the 1990's. Ephraim Stern had several of these fragments reconstructed as a Gorgon antefix—a common architectural element in Archaic Greek temples.<sup>100</sup> Based on this reconstruction, he argued for the existence of a Greek-style temple in Area D or nearby, and for the existence of a Greek community at Dor at that time. Contrary to this reconstruction and interpretation, Martin maintains:<sup>101</sup> (a) This tile is reconstructed from several non-contiguous fragments. We do not know that it would have looked like that, or even that all of these fragments were from the same original object. (b) The tile lacks the peg that fixes the antefix to the architecture behind it. (c) Terracotta antefixes form (a small) part of an assemblage of simpler terracotta tiling elements in the roofs of Greek temples. The latter are missing at Dor. Had there been a Greek-style temple at Dor, we would have found many simple tiles per each decorated one. (d) There is a long tradition of Phoenician cultic masks (albeit in a different style from these). She

98 NoSQL: Database models that utilize a non-tabular or not only tabular data structure. Examples include a document store (such as XML), triple store (RDF), graph databases, and object-oriented databases.

99 Document-oriented database: a data store, the main concept of which is the document. Documents can be organized hierarchically, grouped into collections by some criteria, or tagged. Entities in a document store do not necessarily share a structure; differently structured documents can be stored in the same database.

100 Stern 2001; Stern 2010, 27–30, pl. 19, fig. 32.

101 Martin 2014.



sees this mask as the appropriation of a well-known Greek style by the local (Phoenician) population, for use on a local artifact type. There is minimal evidence, according to Martin, for any Greeks actually residing at Dor. Without endorsing either of these views, we maintain that such debates are fruitful and indeed essential to the evolution of the discipline. Registration systems should encourage such multivocality.

In other cases, an excavator might base his or her decision about dating, function, and other characteristics of a context on a single especially indicative object, while others might claim that the object's relation to the context was doubtful. Note locus [1832] in Figure 1.4: (the square brackets denote that it belongs to VIIB, according to the excavators). This is the find-spot of the bronze base of statue Ramesses VI, a main chronological peg for the dating the end of the Bronze Age in the Levant. According to the excavator, it was found "under a wall in stratum VIIB room 1832 as if deliberately buried there and therefore intrusive"<sup>102</sup> (note that the context has no locus identifier; it is spatially defined relative to features in the stratum above). However, the excavators of Megiddo believed, along with most of the archaeological community at the time, that the Late Bronze Age ended with Merneptah, half a century before, and that Stratum VII was still the Bronze Age. They were well aware that this object undermined the chronological scheme that they favored, but they did not really provide any evidence for its intrusiveness. Later scholars have used this particular find, and its findplace, to argue that the end of the Bronze Age in the Levant should be lowered by about 50 to 75 years.<sup>103</sup> Our contention here is that while data-consistency checks built into the registration system might be used to flag such anomalies as a find whose intrinsic date conflicts with that of the context in which it was found, they should nevertheless allow these anomalies to be retained. Erroneous (or at least seemingly discordant) observations can be as conducive to research as concordant ones.

### *Modeling Uncertainties*

Even more common than multiple opinions are the cases in which there is a single indecisive perspective. A wall is possibly dovetailed with another wall, but maybe one abuts the other; a ceramic vessel rim fragment might belong to a jar or to a jug; a ceramic petrographical sample, the provenience of which could be Cyprus or Lebanon, and so on. Uncertainties are present at all stages of interpretations and in all attributes of the entities and relations between them. On the data-management level, this means that question marks should

<sup>102</sup> Loud 1948, 135.

<sup>103</sup> Finkelstein 1995, 213–239.

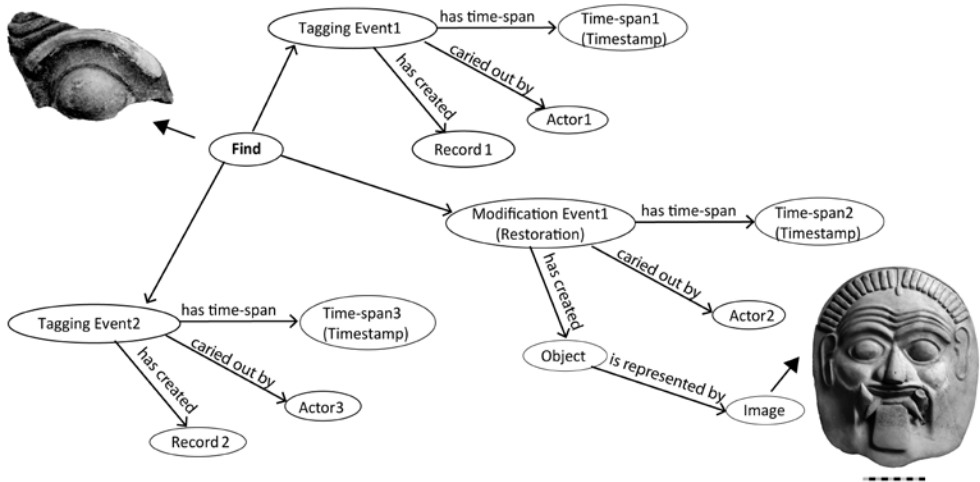


FIGURE 1.7 Modeling multiple interpretations of one find

be allowed in every piece of recorded data to register the fact that there were doubts.

It often happens that even when no new data confirms an interpretation, the initial question marks disappear on the way from the raw data to the report. According to Barker,<sup>104</sup> as part of a “high-level” interpretation process, doubts should be removed from the final report in order to avoid presenting too much information.<sup>105</sup> On the contrary, it can be argued that revealing all doubts to a report reader is a matter of professional ethics and no less important than compiling a convincing story. The history of this debate goes back to the late nineteenth century in England and two major figures on the archaeological scene: General Pitt-Rivers and Sir Flinders Petrie. Pitt-Rivers claimed that all information is objective; therefore, he believed everything should be recorded and published.<sup>106</sup> Petrie accepted the presence of uncertainties, but he was convinced that doubts and queries were “worth nothing in themselves”<sup>107</sup> and that an excavator should decide in the field what information should be recorded as reliable *facts*. The rest should be omitted completely (including discarding finds).

<sup>104</sup> Barker 1982.

<sup>105</sup> See the discussion of Barker’s ideas in Hodder (1997, 692–693).

<sup>106</sup> Pitt-Rivers 1887, xvii.

<sup>107</sup> Petrie 1904, 50.

Locus Index (fragment)

Number	Sq	Phase	PoC	Comments	Context	Chapter
L 9206	AJ/32	7	u	Makeup of floor L9154 including the oven.	--	14
<i>Loci 9207 – 9208 are late (Phase 1 – 4)</i>						
L 9209	AI/33	6?/5?	u	Mudbrick debris under Roman remains.	--	10, 13
<i>Locus 9210 was canceled</i>						
W 9211	AI-AJ/34	6–8	--	Fieldstone north wall of the Phases 9–8 house.	--	5, 10, 11, 12, 13, 15
W 9211b	AI-AJ/34	9–10	--	Lower portion of W9211, west of W18481=W18516.	--	5, 10, 11, 12, 13, 15
W 9212	AJ/34	6b?/6a?/7??	--	Fieldstone wall.	--	12
L 9213	AI/32	6a?	d	Phase 6a fill, contaminated by Pit L9168?	6 and later	9
L 9214	AJ/34	6b?/7?	u	Mudbrick debris.	Pit 9181 - 6 and later	12
L 9215	AJ/34	6?/7?	d	Fill, contaminated by Pit L9181.	--	12
W 9216	AJ/34	6b?/6a?/7?? -7?/6b?/8??	--	Mudbrick wall.	--	12
W 9217	AK/32	5?/4??	--	Fieldstone wall forming a corner with W9162.	--	14

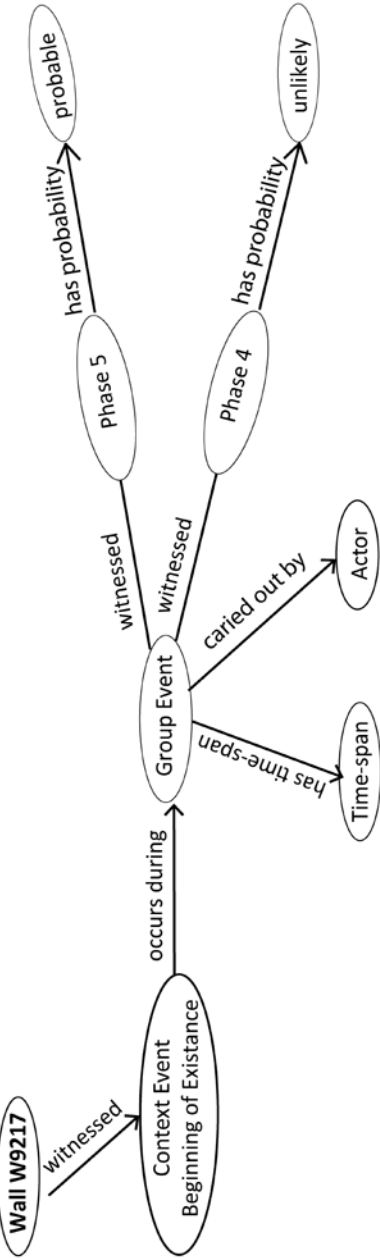


FIGURE 1.8 Modeling uncertainty of the stratigraphic definition of a feature

Back to field records, very often it would be extremely valuable to know how certain the author was about each of his or her assumptions, i.e., could the level of uncertainty be expressed in one, two, three, or more question marks or expressions indicating “almost sure,” “likely,” “very unlikely,” etc. The problem of data uncertainty in archaeology is dealt with in the context of computerized data modeling,<sup>108</sup> GIS,<sup>109</sup> and virtual reconstructions,<sup>110</sup> but it is barely addressed in the Linked Data projects.<sup>111</sup>

In a graph database, the grading of uncertainty can be implemented as a property of the relation between entities. Thus, everything in the data cloud (again, except of the entity ID) can be “tagged” to tell a user about the author’s attitude toward her own theory (Fig. 1.8).

Recording the levels of uncertainty allows for per-record or even per-property evaluation of the data. In stratigraphic analysis, which typically operates only with the records, because the excavation units are gone, such qualifications could be invaluable. A dataset can be filtered by this property to show “probable” or only “very likely” scenarios. Finally, a methodological study might find it interesting to trace the dynamics of the levels of uncertainties during the excavation campaigns, between the team members, or in relation to specific topics.

## References

- Aharoni, Yohanan, Zeev Herzog, Moshe Kochavi, Shmuel Moshkovitz, and Anson F. Rainey. 1973. “Methods of Recording and Documenting.” In *Beer-Sheba I: Excavations at Tel Beer-Sheba, 1969–1971 Seasons*, edited by Yohanan Aharoni, 119–132. Publications of the Institute of Archaeology. Tel Aviv: Tel Aviv University.
- Balensi, Jacqueline. 1985. “Revising Tell Abu Hawam.” *BASOR* 257: 65–75.
- Barker, Philip. 1982. *Techniques of Archaeological Excavation*. 2nd ed. London: B.T. Batsford.
- Bender, Barbara, Sue Hamilton, and Christopher Tilley. 1997. “Leskernick: Stone Worlds; Alternative Narratives; Nested Landscapes.” *Proceedings of the Prehistoric Society* 63: 147–178.

<sup>108</sup> Computerized data modeling: the process of creating a digital data model; see Brouwer Burg, Peeters, and Livis, eds. (2016).

<sup>109</sup> Lawrence, Bradbury, and Dunford 2012.

<sup>110</sup> Danielová, Kumke, and Peters 2016.

<sup>111</sup> But see Isaksen et al. 2010.

- Binding, Ceri, Keith May, and Doug Tudhope. 2008. "Semantic Interoperability in Archaeological Datasets: Data Mapping and Extraction via the CIDOC CRM." In *Proceedings (ECDL 2008) 12th European Conference on Research and Advanced Technology for Digital Libraries, Aarhus*, edited by Birte Christensen-Dalsgaard, Donatella Castelli, Bolette Ammitzbøll Jurik, and Joan Lippincott, 280–290. Berlin: Springer.
- Binford, Lewis R. 1962. "Archeology as Anthropology." *AmAnt* 28: 217–225.
- Binford, Lewis R. 1968. "Archeological Perspectives." In *New Perspectives in Archeology*, edited by Sally R. Binford and Lewis R. Binford, 5–32. Chicago: Aldine.
- Binford, Lewis R. 1977. "General Introduction." In *For Theory Building in Archaeology: Essays on Faunal Remains, Aquatic Resources, Spatial Analysis, and Systemic Modeling*, edited by Lewis R. Binford, 1–10. New York: Academic Press.
- Binford, Lewis R. 1983. *In Pursuit of the Past: Decoding the Archaeological Record*. London: Thames & Hudson.
- Brouwer Burg, Marieka, Hans Peeters, and William A. Lovis, eds. 2016. *Uncertainty and Sensitivity Analysis in Archaeological Computational Modeling*. Interdisciplinary Contributions to Archaeology. Cham: Springer International.
- Chapman, Rupert L. 1986. "Excavation Techniques and Recording Systems: A Theoretical Study." *PEQ* 118: 5–26.
- Clarke, David. 1973. "Archaeology: The Loss of Innocence." *Antiquity* 185: 6–18.
- Cripps, Paul, Anne Greenhalgh, Dave Fellows, Keith May, and David Robinson. 2004. "Ontological Modelling of the Work of the Centre for Archaeology." CIDOC CRM Technical Paper. <<https://www.semanticscholar.org/paper/Ontological-Modelling-of-the-work-of-the-Centre-for-Cripps-Greenhalgh/0245f5f4ca0ba450d5c568314b9bf5641b60371>>.
- Curé, Olivier, and Guillaume Blin. 2015. *RDF Database Systems: Triples Storage and SPARQL Query Processing*. Amsterdam: Morgan Kaufmann.
- Danielová, Mariana, Holger Kumke, and Stefan Peters. 2016. "3D Reconstruction and Uncertainty Modelling Using Fuzzy Logic of Archaeological Structures: Applied to the Temple of Diana in Nemi, Italy." *Cartographica: The International Journal for Geographic Information and Geovisualization* 51 (3): 137–146.
- Dauphin, Claudine. 1984. "Dor, Byzantine Church." *IEJ* 34: 271–274.
- Dauphin, Claudine, and Shimon Gibson. 1993. "Dor-Dora: A Station for Pilgrims in the Byzantine Period on their Way to Jerusalem." In *Ancient Churches Revealed*, edited by Yoram Tzafrir, 90–97. Jerusalem: Israel Exploration Society.
- Doerr, Martin. 2003. "The CIDOC CRM—an Ontological Approach to Semantic Interoperability of Metadata." *AI Magazine* 24 (3): 75–92.
- Doerr, Martin, Kurt Schaller, and Maria Theodoridou. 2010. "Integration of Complementary Archaeological Sources." In *Beyond the Artifact: Digital Interpretation of the Past: Proceedings of CAA2004, Prato 13-17 April 2004, Italy*, edited by Franco Niccolucci, and Sorin Hermon, 64–69. Budapest: Archaeolingua.

- Finkelstein, Israel. 1995. "The Date of the Settlement of the Philistines in Canaan." *Tel Aviv* 22: 213–239.
- Fraassen, Bas C. van. 1980. *The Scientific Image*. Oxford: Clarendon.
- Garstang, John. 1924a. "Tanturah (Dora)." *Bulletin of the BSAJ* 6: 65–73.
- Garstang, John. 1924b. "Tanturah (Dora)." *Bulletin of the BSAJ* 7: 80–85.
- Habu, Junko, Clare Fawcett, and John M. Matsunaga. 2008. *Evaluating Multiple Narratives: Beyond Nationalist, Colonialist, Imperialist Archaeologies*. New York: Springer.
- Hempel, Carl G. 1952. *Fundamentals of Concept Formation in Empirical Science*. Chicago: University of Chicago Press.
- Hodder, Ian. 1997. "'Always Momentary, Fluid and Flexible:' Towards a Reflexive Excavation Methodology." *Antiquity* 273: 691–700.
- Hodder, Ian. 1999. *The Archaeological Process: An Introduction*. Oxford: Blackwell.
- Hodder, Ian, ed. 2001. "Introduction: A Review of Contemporary Theoretical Debates in Archaeology." In *Archaeological Theory Today*, edited by Ian Hodder, 1–13. Cambridge: Polity.
- Hole, Frank, and Robert F. Heizer. 1969. *An Introduction to Prehistoric Archaeology*. New York: Holt, Rinehart & Winston.
- Iliffe, John H. 1933. "Pre-Hellenistic Pottery in Palestine." *QDAP* 2: 15–26.
- Isaksen, Leif, Kirk Martinez, Nicholas Gibbins, Graeme Earl, and Simon Keay. 2010. "Linking Archaeological Data." In *Making History Interactive. Computer Applications and Quantitative Methods in Archaeology. Proceedings of the 37th International Conference, Williamsburg, Virginia, United States of America, March 22–26*, edited by Bernard Frischer, Jane Webb Crawford, and David Koller, 130–136. BAR–IS 2079. Oxford: Archaeopress.
- Joukowsky, Martha. 1980. *A Complete Manual of Field Archaeology: Tools and Techniques of Field Work for Archaeologists*. Englewood Cliffs: Prentice-Hall.
- Kaplan, Abraham. 1998. *The Conduct of Inquiry: Methodology for Behavioral Science*. New Brunswick: Transaction.
- Kelley, Jane H., and Marsha P. Hanen. 1988. *Archaeology and the Methodology of Science*. Albuquerque: University of New Mexico Press.
- Kenyon, Kathleen M. 1981. *Excavations at Jericho III: The Architecture and Stratigraphy of the Tell*. London: The British School of Archaeology in Jerusalem.
- Lamon, Robert S., and Geoffrey M. Shipton. 1939. *Megiddo I*. Chicago: The University of Chicago Press.
- Lawrence, Dan, Jennie Bradbury, and Robert Dunford. 2012. "Chronology, Uncertainty and GIS: A Methodology for Characterising and Understanding Landscapes of the Ancient Near East." In *Landscape Archaeology. Conference (LAC 2012)*, edited by Wiebke Bebermeier, Robert Hebenstreit, Elke Kaiser, and Jan Krause, 353–359. *ETO-POI, Special Volume 3*.
- Leibowitz, Joseph. 1950. "Dor." *IEJ* 1: 249.

- Levy, Thomas E., and Austin F. C. Holl. 1995. "Social Change and the Archaeology of the Holy Land." In *The Archaeology of Society in the Holy Land*, edited by Thomas E. Levy, 4–5. New York: Facts On File.
- Losee, John. 2001. *A Historical Introduction to the Philosophy of Science*. 4th ed. Oxford: Oxford University Press.
- Loud, Gordon. 1948. *Megiddo 2: Seasons of 1935–39; Text*. Chicago: University of Chicago Press.
- Maedche, Alexander. 2002. *Ontology Learning for the Semantic Web*. Boston: Kluwer Academic.
- Martin, S. Rebecca. 2014. "From the East to Greece and Back Again: Terracotta Gorgon Masks in a Phoenician Context." In *Phéniciens d'Orient et d'Occident: Mélanges Josette Elayi*, edited by Andre Lemaire, 289–299. Paris: Maisonneuve.
- Matskevich, Sveta. 2015. "'Off the Record.' Recording Systems for Archaeological Excavations in the Levant: Past and Future." PhD diss., The Hebrew University of Jerusalem.
- Matskevich, Sveta, and Ilan Sharon. 2016. "Modelling the Archaeological Record: A Look from the Levant. Past and Future Approaches." In *CAA 2015. Keep the Revolution Going. Proceedings of the 43rd Annual Conference on Computer Applications and Quantitative Methods in Archaeology*, edited by Stefano Campana, Roberto Scopigno, Gabriella Carpentiero, and Marianna Cirillo, 103–115. Oxford: Archaeopress.
- May, Keith, Ceri Binding, Doug Tudhope, and Stewart Jeffrey. 2012. "Semantic Technologies Enhancing Links and Linked Data for Archaeological Resources." In *Revive the Past: Proceedings of the 39th Conference in Computer Applications and Quantitative Methods in Archaeology, Beijing, China, 12–16 April 2011*, edited by Mingquan Zhou, Iza Romanowska, Zhongke Wu, Pengfei Xu, and Philip Verhagen, 261–272. Amsterdam: Pallas.
- Mazar, Amihai. 1997. *Timnah (Tel Batash) I: Stratigraphy and Architecture*. Qedem 37. Jerusalem: The Hebrew University of Jerusalem.
- Merton, Robert K. 1949. *Social Theory and Social Structure: Toward the Codification of Theory and Research*. Glencoe, IL: Free Press.
- Petrie, William M. Flinders. 1904. *Methods and Aims in Archaeology*. London: MacMillan and Co.
- Pitt-Rivers, Augustus H. L.-F. 1887. *Excavations in Cranborne Chase, near Rushmore, on the Borders of Dorset and Wilts. 1880–1896*. Vol. 1. London: Harrison and Sons.
- Raab, L. Mark, and Albert C. Goodyear. 1984. "Middle-Range Theory in Archaeology: A Critical Review of Origins and Applications." *AmAnt* 49 (2): 255–268.
- Raban, Avner. 1993. "Acco: Maritime Acco." In *New Encyclopedia of Archaeological Excavations in the Holy Land*, edited by Ephraim Stern, 29–31. Vol. 1. Jerusalem: Israel Exploration Society & Carta.
- Renfrew, Colin, and Paul Bahn. 2008. *Archaeology: Theory, Methods, and Practice*. 5th ed. London: Thames & Hudson.



- Roskams, Steve. 2001. *Excavation*. Cambridge Manuals in Archaeology. Cambridge: Cambridge University Press.
- Shahack-Gross, Ruth. 2011. "Herbivorous Livestock Dung: Formation, Taphonomy, Methods for Identification, and Archaeological Significance." *JASC* 38 (2): 205–218.
- Shahack-Gross, Ruth, Rosa-Maria Albert, Ayelet Gilboa, Orna Nagar-Hilman, Ilan Sharon, and Steve Weiner. 2005. "Geoarchaeology in an Urban Context: The Uses of Space in a Phoenician Monumental Building at Tel Dor (Israel)." *JASC* 32: 1417–1431.
- Sharon, Ilan. 1995a. "Models for Stratigraphic Analysis of Tell Sites." PhD diss., The Hebrew University of Jerusalem.
- Sharon, Ilan. 1995b. "The Registration System." In *Excavations at Dor, Final Report 1A: Areas A and C*, edited by Ephraim Stern, John Berg, Ayelet Gilboa, Bracha Guzz-Zilberstein, Avner Raban, Renate Rosenthal-Heginbottom, and Ilan Sharon, 13–20. Qedem Reports 1. Jerusalem: The Hebrew University of Jerusalem.
- Slaatte, Howard A. 1979. *The Dogma of Immaculate Perception: A Critique of Positivist Thought*. Lanham, MD: Rowman & Littlefield.
- Stern, Ephraim. 2001. "A Gorgon's Head and the Building of the First Greek Temples at Dor and along the coast." *Qadmoniot* 34: 44–48 (in Hebrew).
- Stern, Ephraim. 2010. *Excavations at Dor: Figurines, Cult objects and Amulets, 1980–2000 Seasons*. Jerusalem: Israel Exploration Society.
- Stern, Ephraim, John Berg, Ayelet Gilboa, Bracha Guzz-Zilberstein, Avner Raban, Renate Rosenthal-Heginbottom, and Ilan Sharon. 1995. *Excavations at Dor, Final Report 1A: Areas A and C*. Qedem Reports 1. Jerusalem: The Hebrew University of Jerusalem.
- Stewart, Andrew, and S. Rebecca Martin. 2005. "Attic Import Pottery at Tel Dor, Israel: An Overview." *BASOR* 337: 79–94.
- Van Beek, Gus W. 1988. "Excavation of Tells." In *Benchmarks in Time and Culture. An introduction to Palestinian Archaeology*, edited by Joel F. Drinkard, Gerald L. Mattingly, James Maxwell Miller, and Joseph A. Callaway, 131–167. Atlanta: Scholars Press.
- Wiley, Gordon R., and Philip Phillips. 1958. *Method and Theory in American Archaeology*. Chicago: University of Chicago Press.
- Zarzecki-Peleg, Anabel. 2016. *Yadin's Expedition to Megiddo: Final Report of the Archaeological Excavations (1960, 1966, 1967 and 1971/2 Seasons)*. Qedem 56. Jerusalem: The Hebrew University of Jerusalem.
- Zorn, Jeff R., ed. 1991. "Tel Dor Excavations—Staff Manual." <[http://dor.huji.ac.il/Download/Tel\\_Dor\\_Staff\\_Manual/Tel\\_Dor\\_Staff\\_Manual\\_1991.pdf](http://dor.huji.ac.il/Download/Tel_Dor_Staff_Manual/Tel_Dor_Staff_Manual_1991.pdf)>.
- Zorn, Jeff R., Ilan Sharon, and Ayelet Gilboa. In press. "Introduction: History of the Area G Excavations (1986–2004); Post Excavation Analysis (1993–2010) and Introductory Remarks on Excavation, Documentation and Methods." In *Excavations at Dor, Final Report, Volume 2, Area G*, edited by Ayelet Gilboa, Ilan Sharon, and Jeff R. Zorn. Qedem Reports. Jerusalem: The Hebrew University of Jerusalem.



# Landscape Archaeology and Artificial Intelligence: the Neural Hypersurface of the Mesopotamian Urban Revolution

*Marco Ramazzotti*

In collaboration with *Paolo Massimo Buscema and Giulia Massini\**

What people call “meanings” do not usually correspond to particular and definite structures, but to connections among and across fragments of the great interlocking networks of connections and constraints among our agencies.

MARVIN MINSKY, 1986, 131.



## Introduction

Since the 1990s, there has been an unceasing debate over computer semiotics as an autonomous discipline aimed at establishing the function of the logical operators of programming and computing.<sup>1</sup> In fact, the structural and semantic encoding of the analytical object also comprise one of the main trends in Natural Computing (NC) and in the fast-moving field of computer science.<sup>2</sup>

---

\* From the Semeion Research Center of Rome.

- 1 Computer semiotics: any empirical approach mainly interested in the ways that humans and machines may communicate with each other using written languages, texts, and/or codes (Andersen 1991a, 3–30; 1991b, 1–32; Figge 1991, 321–330).
- 2 Natural Computing (NC): a nonlinear dynamics and pattern formation computing inspired by concepts, principles, and mechanisms underlying natural systems (Anderson and Rosenfeld 1988; McClelland and Rumelhart 1988). NC algorithms are inspired by natural and biological phenomena and include evolutionary algorithms, neural networks, molecular computing, and quantum computing (Brabazon, O’Neill, and McGarraghy 2015, 221–280). Computer Science brings together disciplines including Mathematics, Engineering, the Natural Sciences, Psychology, and Linguistics (Beckerman 1997; Miller and Pages 2007).

Moreover, the possible encoding of the historical, archeological, anthropological, aesthetic, and linguistic records and contexts as a minimum unit of meaning (*sema*) defined a new epistemic perspective in ancient world studies.<sup>3</sup> This new epistemic perspective, which currently also can be considered a recent branch of the digital humanities, is interested in translating the complex systems of the past into systems of signs and in turning each sign into a point (site) and node (or cell) composing many artificial formal networks.<sup>4</sup>

The present contribution introduces the theoretical and experimental approaches in encoding the complexity of the Mesopotamian Urban Revolution as an artificial network and in simulating the multifactorial relationships of this network with the biological computing of the Artificial Adaptive Systems (AAS).<sup>5</sup>

### Methodology: Encoding, Translating, and Modeling

The artificial formal network of a complex system can be considered a synthetic and formalized representation of observed reality. It is both an abstract model and an algebraic generalization of the reality that can be explored and simulated using statistical, mathematical, and physical models.<sup>6</sup> Exploring and simulating both abstract models and algebraic generalizations of complex configurations, the mathematical and physical models of NC equal tracking down, selecting, and (separately) recreating: 1) a wide variety of functions associating variables, 2) a wide variety of inferences controlling their conceptual structure, and 3) an equally wide variety of rules producing their transformations.<sup>7</sup>

---

3 A new epistemic perspective recently advanced in theoretical (Ramazzotti 2010), historical (Ramazzotti 2013d), and empirical (Ramazzotti 2014a, 15–52) approaches to the study of the past.

4 For network analysis in Archaeology, see Bentley and Shennan (2003, 459–485), Brughmans (2010, 277–303), Barthélemy (2011, 1–101), Knappet (2011), Kohler (2012, 93–123), Ramazzotti (2013b, 283–303), Leeuw (2013, 335–349), and Collar et al. (2015, 1–32).

5 Artificial Adaptive Systems (AAS) are biological computing methods, techniques, and algorithms forming part of the vast world of Natural Computation/Natural Computing, which is itself a subset of the Artificial Sciences (AS). AS are those sciences for which an understanding of natural and/or cultural processes is achieved by the recreation of those processes through automatic models (Buscema 2014, 53–84).

6 Ramazzotti 2014a, 15–42.

7 For a general introduction and overview of NC—both of abstract models and of algebraic generalizations of archaeological, anthropological, and visual complex configurations—see

Since the end of the 1980s, a matrix encoding of the Mesopotamian Urban Revolution has been in the process of development for the purpose of tracking down,<sup>8</sup> selecting, and recreating the functions, inferences, and rules that produced the political, cultural, and economic transformations of the most ancient Urban Revolution.

The artificial formal networks obtained by such structural and semantic matrix encoding were thus massively described, analyzed, and simulated through the quantitative, qualitative, and symbolic methods of Artificial Intelligence (AI).<sup>9</sup> After some 30 years of such theoretical and experimental research on the origin of the city-state and urbanism in the ancient Near East, the simulations of the Mesopotamian Urban Revolution Landscape (MURL)<sup>10</sup> through the AI symbolic paradigm seem to maintain a distinct value. This distinct value is most evident in archaeological thought, where the MURL simulations serve both as a new analytical paradigm for computational modeling in Archaeology and as a new theoretical approach for the study of urbanism and urbanization.<sup>11</sup>

---

Ramazzotti (1997, 495–522), Reeler (1999, 3–10), Zubrow (2003), Bintliff (2005), Barceló (2008, 154–184), and Ludovico and Ramazzotti (2008, 263–280); Ramazzotti (2014a, 15–52).

- 8 On matrix encoding and combinatorial analysis in Archaeology, see Kendall (1971, 215–252) and Shuchat (1984, 3–14).
- 9 For the so-called symbolic paradigm, AI has been conceived as the development of models using symbol manipulation. The computation in the models is based on explicit representations that contain symbols organized in some specific structures. The connectionist paradigm aims at massively parallel models that consist of a large number of simple and uniform processing elements interconnected with extensive links, the Artificial Neural Networks (Smolensky 1987, 95–109).
- 10 We use the acronym “MURL” to identify the historical period and the economic concept of the Urban Revolution as they were delineated first by Vere Gordon Childe (1950, 3–17), as well as the area of the world’s first urbanism—central-southern Iraq. On the geographical, archeological, and historical features of the MURL, see Ramazzotti (2002, 651–752).
- 11 One of the most significant turning points in the history of computational modelling in Archaeology is strictly related to the theoretical and empirical work of David Leonard Clarke on the System Theory paradigms and principles (Clarke 1968; 1972; 1977). This turning point cannot be summarized here, but for a recent critical historiography of computational modelling in Archaeology, see Ramazzotti (2013c, 23–56) and Ramazzotti (2014a, 15–52). Indeed, the most recent empirical theories on urbanism and urbanization are mainly based on settlement multifactorial complexity modelling through system theory and/or social theory (Smith 2011, 167–192); an “ecological approach,” we could say, started with many works of systematic research on Mesopotamian urban civilization at the end of 1970s (Redman 1978a; 1978b, 329–347). For a recent historiography of Babylonian urbanism, urbanization, and city-state ideal-types, and for the role played also by the neu-

The application of AAS to the MURL can elaborate new analytical hypersurfaces that can exhibit nuances and complex interrelations and, furthermore, can help the researcher to discover other, unforeseen (or even hidden) interrelationships.<sup>12</sup>

### The Neural Modeling of the MURL

One of the first applications of AAS to the MURL (encoded in different matrixes) was advanced and tested at the Semeion Research Center (Rome) and at La Sapienza University of Rome and published in 1999.<sup>13</sup>

At that time, the main problem was how to properly encode and simulate the systemic settlement complexity of the most ancient urbanism, integrating data from excavated sites, surface surveys, and a mixture of each, in order to reveal and analyze the cultural, political, and economic interrelationships of the Mesopotamian Urban Revolution.<sup>14</sup> In this respect, the land of Sumer's settlement system was first encoded in a matrix composed of records (208 archaeological sites) and variables (95 geographical, morphological, and cultural

---

ral approach to the systemic complexity of the Mesopotamian settlement systems, see Liverani (2013).

- 12 AAS are AI deep-learning models and algorithms that simulate physical or cultural phenomena and features that have been previously encoded, so that from processing such models previously unknown relationships can be revealed through the features. The application of such types of models and algorithms to the MURL natural and cultural phenomena encoded in matrixes can be considered a new combinatory analysis of an artificial formal network, and the NC of the MURL can recreate a possible world of other associations of meaning from the body of incomplete sources and scattered information (Ramazzotti 1997, 495–522; 1999; 2000, 9–38; 2002, 651–752; 2003, 15–71; 2009, 193–202; 2013a, 10–34; 2013b, 283–303; 2013c, 23–56; 2014a, 15–52; 2014b, 53–84; 2016a, 17–26; 2016b, 183–194).
- 13 Ramazzotti 1999. These were the first applications of biological computing to the MURL (encoded in different *n*-dimensional matrixes); they were tested at the Semeion Research Center and discussed in the Department of Historical, Archaeological, and Anthropological Sciences of Antiquity at La Sapienza University in 1996.
- 14 This data mainly resulted from the pioneering field, satellite, spatial, and analytical surveys directed by Robert McCormick Adams, Henry T. Wright, and Hans J. Nissen in central-southern Iraq (Adams 1955, 6–18; 1956, 227–232; 1972a, 735–749; 1972b, 60–62; 1978, 329–335; 1981; 2008, 1–23, <[http://cdli.ucla.edu/pubs/cdlj/2008/cdlj2008\\_001.html](http://cdli.ucla.edu/pubs/cdlj/2008/cdlj2008_001.html)> [accessed June 9, 2017]; Nissen 1980, 285–290; 1983a, 91–106; 1983b, 287–294; 2001, 149–179; Adams and Nissen 1972; Wright 1977, 379–397; 1978, 49–68; 1981, 297–362; 1984, 41–78; 1986, 141–155; 1998, 173–192; 2001, 123–148).

variables).<sup>15</sup> Then the matrix was analyzed by simulating any possible relationships between archaeological sites and variables through different types of Artificial Neural Networks (ANNs).<sup>16</sup> This neural modeling of the MURL empirically approached many cultural, settlement, and economic arguments of the world's first urbanism, revealing many connections from the body of incomplete sources and exhibiting the nuances and complex interrelations.

The specific use of ANNs as a classification method improved for Southern Mesopotamian protohistoric pottery technically confuted the existence of a Middle Uruk sub-typology in Sumer.<sup>17</sup>

The specific use of ANNs as a spatial analysis method on the central-southern Mesopotamian settlement distribution specifically focused on the peculiar

- 
- 15 The analysis was founded on the archeological material from the Uruk, Akkad, Kiš, and Nippur surveys (Ramazzotti 1999, 259–269). The 95 parameters describing each archaeological deposit were mostly sub-articulated in six macro-groups: a) settlement and site morphology, b) environmental and adaptive context, c) architectural and visible remains, d) cultural and technological milieu, e) pottery typologies, and e) occupation phases. The occupational sequence of each settlement from the second half of the fifth to the second half of the third millennium BCE was subdivided according to the classical periodization of ancient Mesopotamia (Porada et al. 1992, 77–121) into the main seven periods of the southern Urban Revolution process (Ubaid IV: UBIV; Early Uruk: EU; Middle Uruk: MU; Late Uruk: LU; Jemdet Nasr: JN; Early Dynastic I: EDI; Early Dynastic II – III: EDII – II), and the most conspicuous occupation phases were selected for each site based on both the quantities and qualities of the dating materials.
- 16 Artificial Neural Networks (ANNs): learning systems; these are algorithms for processing information that allow for the reconstruction, in a particularly effective way, of the approximate rules relating to a set of “explanatory” data concerning the considered problem (the input), with a set of data (the output) for which it is requested to make a correct forecast or reproduction in conditions of incomplete information (Buscema 2014, 53–84). See also in this volume, applied a) to art, Ludovico, 92–94 and b) to semantics, Svård, Jauhiainen, Sahala, and Lindén, 228–229, 246–249.
- 17 The typological subdivision of archaeological pottery also can be considered a Constraint Satisfaction (CS) problem, in which the recognition of a class or a type must satisfy a number of constraints. Most of the conventional approaches to solving CS problems improved the algorithms by processing the constraints. The ANNs’ approach was completely different, since, in a Neural Network, the constraints are encoded in a net topology (biases and connections), and when a network sets in a stable configuration, the pattern of firing neurons represents the solution. See Mézard and Mora (2009, 107–113). In particular, applying the ANNs classification method, the existence of the Middle Uruk pottery sub-typology was refuted. This absence revealed the functional role played by this latent horizon in the explanation of the southern Urban Revolution as an orthogenetic process, inspired by market-oriented economic categories (Ramazzotti 1997, 495–522).

relationship between settlement morphology and spatial organization in the hinterland of the site of Uruk (Warka).<sup>18</sup>

Since the end of the 1990s, moreover, the neural modeling approach has also analytically challenged the most ambitious demographic hypothesis on the origin of the Mesopotamian city-state.<sup>19</sup>

Indeed, to explore the possible relationships between the high spatial variability of southern Mesopotamian settlement organization and human mobility, unsupervised networks were tested to draw inferences between settlement morphologies,<sup>20</sup> settlement distributions, and settlement dynamics.<sup>21</sup>

---

18 For a specific discussion on the recognition of the Jemdet Nasr period (c. 3100–2900 BCE) as either a distinct archeological period of the Mesopotamian Urban Revolution or as a regional pottery style, see Finkbeiner (1986, 33–56) and Ramazzotti (2000, 9–38). In this case, the neural-spatial analysis showed the Jemdet Nasr settlement pattern not as a phase of collapse rendering a hard break in culture at the end of the Urban Revolution, but rather as a different and peculiar regional adaptation to manage central-southern Mesopotamian hydrological resources (Ramazzotti 2002, 651–752).

19 Adams 1972a; 1972b; Young 1972, 827–842; Gibson 1973, 447–463; Weiss 1977, 347–369; Brinkman 1984, 169–180.

20 In this case the Kohonen's features maps/Self-Organizing Maps (SOMs) have been used. The SOM is a type of learning algorithm used to draw inferences from datasets and to order high-dimensional statistical data so that neighboring nodes on the map represent similar inputs. Often the SOM is applied to numerical data in application areas such as pattern recognition, signal processing, and multivariate statistical analysis (Kohonen 1982, 59–69). In other words, the SOM is an unsupervised type of network that offers a classification of the input vectors, creating a prototype of the classes and a projection of the prototypes on a map having two dimensions (but *n*-dimensional maps are also possible) that is able to record the relative proximity (or neighborhood) between the classes (Kohonen 1996, 281–291; Massini 2010, 313–348). Regarding multivariate analysis, see in this volume, Martino and Martino (118–119), who discuss the method, in particular, for making object-typologies.

21 Since the key advantage of the SOM is the clustering, which reduces the input space into representative features of a map, comparing different output maps' strong correlations between the morphology of the archaeological sites, assessments of the hydraulic landscape and anthropic mobility trends were observed. During the Late Uruk period (c. 3500–3100 BCE), the local southern population was predominantly settled in stable anthropic deposits close to the canals, but during the subsequent, Jemdet Nasr period (c. 3100–2900 BCE), the population settled in more dispersed patterns, that fit what Adams and Nissen (1972, 24, fig. 11) describe as composite and multiple occupations. Then settlement adaptation patterns drastically changed during the subsequent Early Dynastic period (c. 2900–2350 BCE), when, on account of a process of small-settlement aggregation (synoecism), the dispersed population probably shifted northernmost and converged onto more isolated sites in the region of Nippur (Ramazzotti 2003, 15–71; 2009, 193–202).

## The Biological Modeling of the MURL

A turning point in the neural modeling of the MURL was the investigation of an empirical analogy between the complexity of the archaeological settlement systems and the complexity of the biological system.<sup>22</sup> Considering the spatial relations between points (sites) as the nodes and/or cells of a highly interconnected net, we translated the spatial-temporal segments (sites) of the MURL into a network, intended as a membrane dynamically activated by different actions, causes, and/or events.<sup>23</sup>

Each point of the membrane was thus conceived as a geo-referenced archaeological site, and the adaptive network was trained first through the most advanced generation of AAS. Then, the highly sophisticated outputs of the training were optimized, formalized, and displayed through data-mining algorithms in tree-graphs.<sup>24</sup> The graph analysis of the deep learning of the membrane/network was thus tested as a predictive spatial tool for locating the possible position of undiscovered monuments and/or sites.<sup>25</sup>

In particular, graph-algorithms were applied on a matrix generated by the last ANNs generation to predict the possible spatial localization of the Ebla Royal Mausoleums in the hinterland of Tell Mardih–Ebla, in northern Syria.<sup>26</sup>

Moreover, the same integrated method (ANNs training and graph analysis) has been applied to identify correlations between settlement distributions and cultural, technological, and economic variables between the Ubaid (c. 6500–3800 BCE) and Uruk periods (c. 3800–3100 BCE) in south-central Babylonia.<sup>27</sup>

22 For this assumption, see Ramazzotti (2014a, 15–52).

23 Ramazzotti 2013b, 283–303; 2013c, 23–56.

24 “For the purposes of archaeological analysis, graph theory can be defined as a mathematical method to describe the kind, direction and magnitude of interconnections between individuals or groups at differing nodes of activity and to analyze the underlying systematic structure that the whole set of interconnections implies” (Rothman 1987, 73).

25 A computerized learning machine takes precise input and produces definite output as “true” or “false,” which is equivalent to a human’s “yes” or “no.” Artificial neural learning of complex configurations can be understood to emulate human learning based on fuzzy logic, as it operates on degrees of possibilities of input to achieve definite output.

26 Regarding the last ANNs generation, see also footnotes 30 and 36. Auto-CMs (which stands for Auto-Contractive Maps) “spatialize’ the correlation among variables by constructing a suitable embedding space where a visually transparent and cognitively natural notion, such as ‘closeness’ among variables, reflects accurately their associations” (Buscema et al. 2009, 8; Ramazzotti 2013a, 10–34).

27 In particular, the A-Temporal Diffusion Model (ATDM) equation and Auto-CM Artificial Neural Network has been recently sketched, tested, and presented as a learning machine method for the training of a large matrix composed by 496 archaeological sites (records)



The hypersurface outputs of neural learning have been explored by integrating the traditional Minimum Spanning Tree (MST) graph analysis with the Maximally Regular Graph (MRG) new algorithms.<sup>28</sup> This integrated approach was thus adopted to recognize the logic of the spatial distribution of two of the most significant cultural-material records in the region: the Ubaid clay sickles and the Uruk clay cones (Fig. 2.1a–b).<sup>29</sup>

The uniformly distributed presence of clay sickles and clay cones in the MST-MRG tree-graph, and the distribution of the branches of the tree-graph,

---

described by 106 parameters (variables). ATDM is a recently proposed algorithm that has been developed to detect the dependencies among pairs of variables in a large dataset, while also taking approximate account of their higher-order relationships with other variables (Buscema et al. 2013b, 231–275).

- 28 From a conceptual point of view, the MST represents the energy-minimization state of a structure. In fact, if we consider the atomic elements of a structure as the vertices of a graph and the strength among them as the weight of each edge linking a pair of vertices, the MST represents the minimum of energy needed so that all the elements of the structure preserve their mutual coherence. In a closed system, all the components tend to minimize the overall energy. So the MST, in specific situations, can represent the most probable state for the system. For the classical formulation of the MST (Kruskal 1956, 48–50), see Buscema and Sacco (2016, 726–746). Despite other seriation techniques, the MST data-mining method permits branching structures that also reveal clustering in archaeological datasets (Hage, Harary, and Brent 1996, 149–155); this method has been applied as a model for spatial analysis in different ways. The MRG algorithm can be considered a new type of semantic graph that uses a new index for detecting structural/topological complexity information in undirected graphs (H function). In fact, from an MST, generated from any metric, the MRG reshapes the links among nodes in order to maximize the fundamental and most regular structures implicated in any dataset. The MRG algorithm generates, starting from the MST, the graph presenting the highest number of regular microstructures that make use of the dataset's most important connections. Compared to the MST, therefore, the MRG adds all (and only) those extra features that are really useful in understanding the prototypes that are hidden in the database; in other words, it adds the optimal amount of complexity that is necessary to read the phenomenon. In terms of our specific object of study, the MRG represents an alternative graph-theory representation of the underlying power network with respect to the one obtained by simply reconstructing the geography of common board affiliation (Buscema and Sacco 2010, 227–276; Buscema et al. 2016, 355–378).

- 29 The Ubaid clay sickles (fifth to fourth millennium BCE) are usually associated with the first intensive agricultural activities in the alluvial plain (Benco 1992, 119–134; Stein 2010, 23–44), and the Uruk clay cones (fourth millennium BCE) are linked to the ideographical and symbolical representations of the most important religious buildings of the first southern Mesopotamian cities (Brandes 1971; Wright and Johnson 1975, 267–289; Cooper 1985, 97–114).



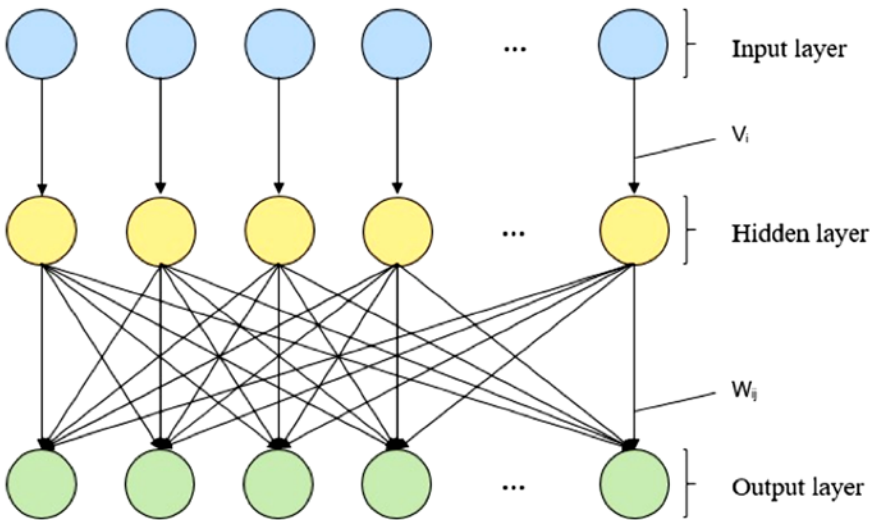


FIGURE 2.1a *Auto-CM ANN* (BY PAOLO M. BUSCEMA © SEMEION)

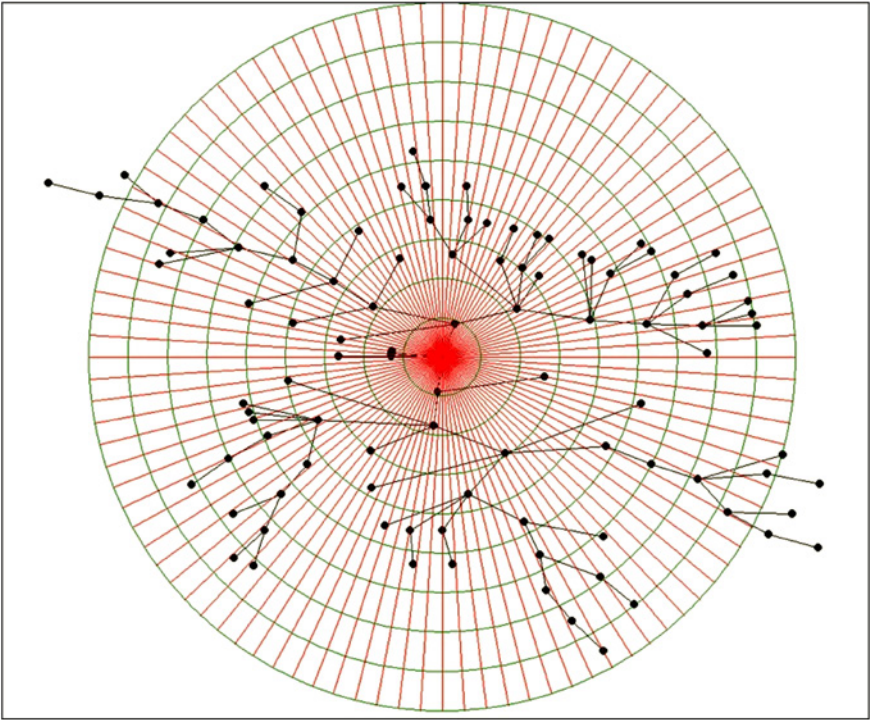


FIGURE 2.1b *Tree-Visualizer* (BY GIULIA MASSINI © SEMEION)

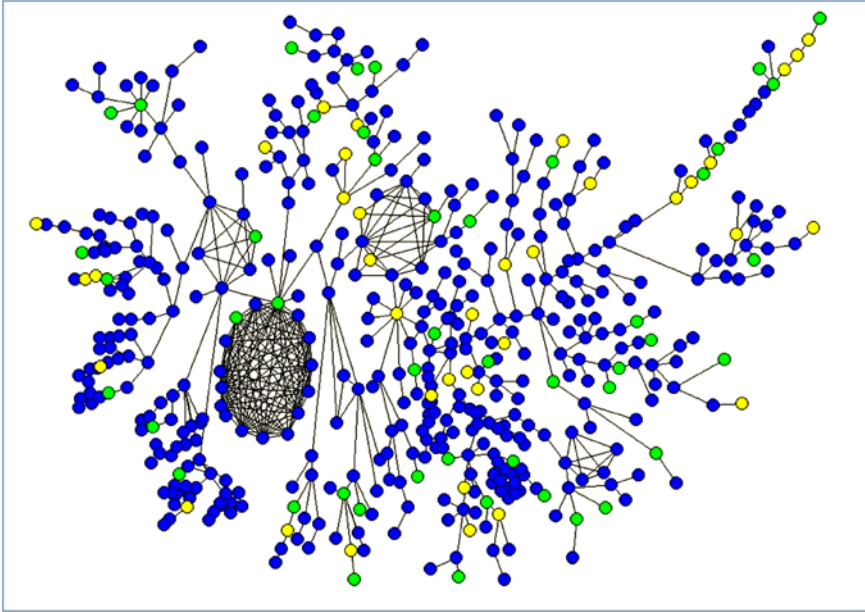


FIGURE 2.2a *Auto-CM – MST-MRG graph. The presence of the clay sickles (green points) and clay cones (yellow points) is uniformly distributed on the MST-MRG tree-graph. The distribution on the tree branches could indicate a homogenous spatial relationship between food-production activities and symbolic/religious functions.*

could indicate a homogenous spatial relationship between food production activities and symbolic and/or religious functions (Fig. 2.2a).

On the contrary, by testing the same neural training outputs on the distribution of the two variables together (clay sickles and the clay cones) we observed a sensible reduction in the number of occurrences (Fig. 2.2b).

The co-existence of the two variables in specific regions of the graph-tree could thus implicate a possible nuclear settlement organization and could suggest a pilot role, on both economic and symbolic levels, played by some specific sites across the whole settled area between the end of the fifth millennium BCE and the end of the fourth millennium BCE.

### The Topological Modeling of the MURL

Recently, research on the MURL has been advanced by the exploration, through a new data-mining experimental procedure (see Appendix 2.1), of the spatial semantics of the settlement distributions in the region between Ur

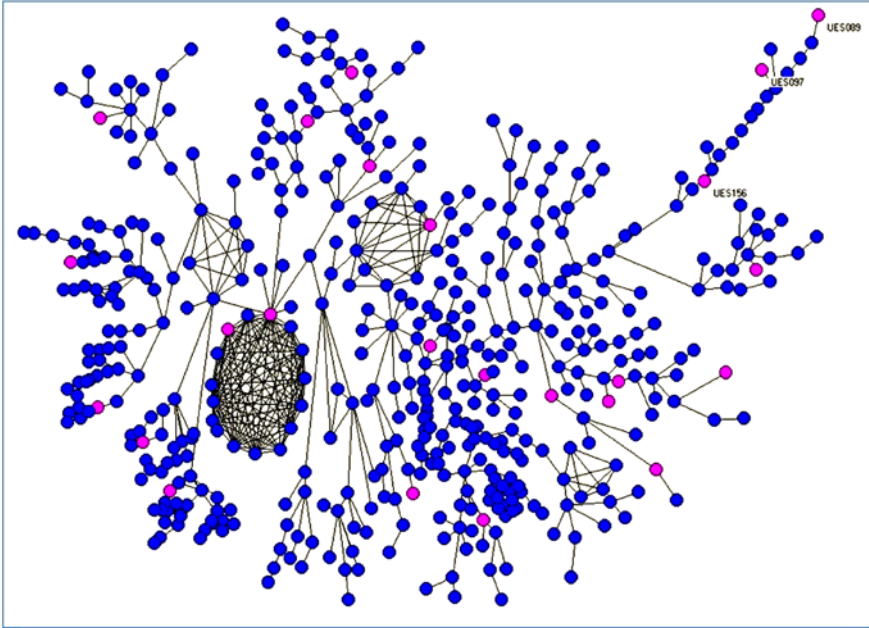


FIGURE 2.2b *Auto-CM – MST-MRG graph. Testing the same neural hypersurface on the distribution of the two variables together (clay sickles and the clay cones), we observed a sensible reduction of the number of occurrences (purple points). The coexistence of the two variables in specific regions of the graph-tree could thus implicate a possible nuclear organization.*

(Tell Muqayyar) and Uruk (Warka). In particular, the spatial semantic of the MST-MRG graph-trees resulting from the neural learning of the Ur and Uruk settlement network through the new algorithm of the Topological Weighted Centroid (TWC) has been discussed (see Appendix 2.2).<sup>30</sup>

The aims of this applied research are to focus on the possible topology of the MURL and to investigate the topography of movement of these shapes and forms through time.<sup>31</sup> Following this new space-analysis to the geographical profiling, we first have selected the fifth, fourth, and third millennium BCE settlements of the region between Ur and Uruk.

30 For a detailed explanation of the TWC, see Appendix 2.2 by Paolo Massimo Buscema. The TWC mathematical approach explores the natural and anthropic landscape with respect to certain quantities (entropy), spatializing the optimal solutions in the centers of the masses (Buscema, Breda, and Catzola 2009; Buscema, Grossi, and Jefferson 2009; Buscema 2014; Buscema et al. 2015, 532–567).

31 To approach movement from a computational perspective by creating tools that can help to cope with its fluid and emergent nature, see Mlekuž (2014, 5).

Second, we tested the TWC algorithm on the Auto-CM and the ATDM neural hypersurfaces of a new dataset (224 sites for 106 variables). Lastly, we generated four TWC maps for each individual period of the Urban Revolution process (the Ubaid, c. 6500–3800 BCE; the Uruk, c. 3800–3100 BCE; the Jemdet Nasr, c. 3100–2900 BCE; and the Early Dynastic, c. 2900–2350 BCE), superimposing the results on a satellite photo of the region between Ur and Uruk.

Since we define the TWC map of  $n$ -entities in a two-dimensional space as the *center of mass* of these entities, weighted by the proximity of each entity to the others,<sup>32</sup> we have produced TWC distribution maps for each chronological sequence of the MURL (period 1, period 2, period 3, and period 4).

The four maps selected here describe a diachronic trajectory (periods 1, 2, 3, 4) across the same space (the region between Ur and Uruk) and thus display a semantic topology of the most significant geographical, morphological, and cultural features of the MURL (Fig. 2.3).

At the present state of our research, this diachronic trajectory can only be intended as the spatial-temporal diffusion of a centroid, where the centroid here simulated is the *center of mass* of the variables (106 geographical, morphological, and cultural variables) defining each individual archaeological site in the region (224 sites).

The diachronic changes of these centers of mass on the four maps seem to demonstrate an ancient spatial history of the Mesopotamian Urban Revolution akin to a mitosis process, a biological process of cell duplication, or reproduction.<sup>33</sup>

Turning *back to the future*,<sup>34</sup> the morphogenesis of this topology opens another fascinating frontier for the analysis of the world's first urbanism: the topological modeling of the settlement networks intended as adaptive and dynamic biological membranes.

32 According to the TWC mathematical paradigm, if we take the center of mass of the distribution as the natural reference, we can therefore consider how the distance between the center of mass and any given point is influenced by the relative distance of all other points in the distribution (Buscema et al. 2013a, 75–139).

33 In fact, the first TWC map related to the Ubaid period shows two separated centroids, in the southern and northern parts of the selected area; the second and the third TWC maps related to the central periods of the Urban Revolution display only one centroid (in the northern sector of the selected area), and the fourth map repeats the same two-centroid distribution of the first map.

34 Ramazzotti 2016b, 183–194.

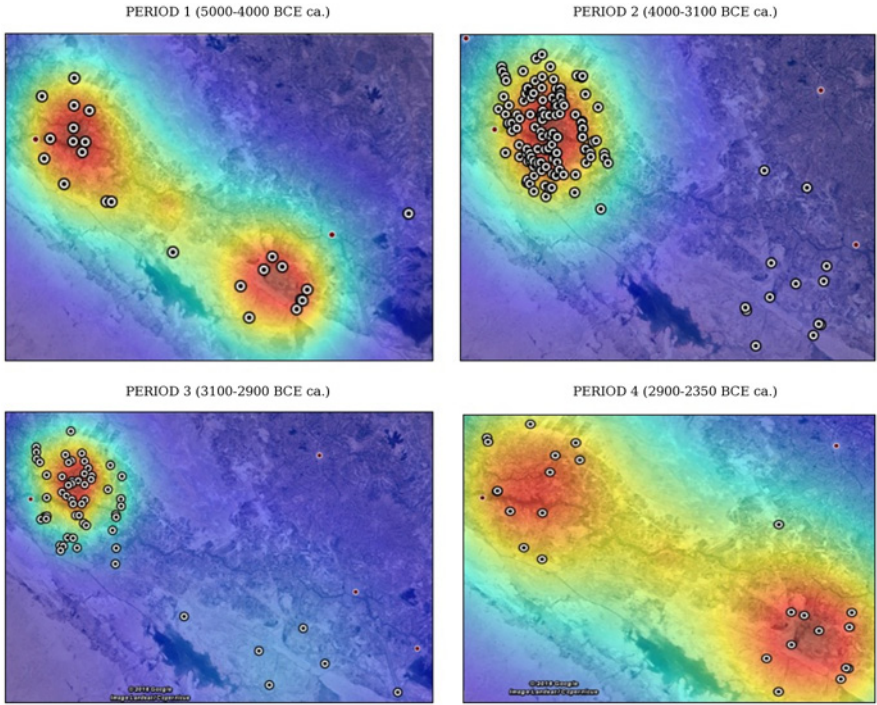


FIGURE 2.3a-d *The four period TWC Gamma ( $\gamma$ ) maps of the MURL. Each map describes the extent to which each point of the space is activated by its closeness to any of the points belonging to any of the nonlinear trajectories connecting each point to the center of mass.*

## Appendix 2.1. The Experimental and Simulation Procedures<sup>35</sup>

A complex machine-learning system (Fig. 2.4) has been adopted to process the MURL encoded data and transpose it in an  $n$ -dimensional matrix (224 sites for 106 geographical, morphological, and cultural variables).

The experimental process has been subdivided into four steps:

1. The archaeological records divided into the four main periods (Ubaid, Uruk, Jemdet Nasr, and Early Dynastic) of the Urban Revolution were first formalized in an  $n$ -dimensional matrix in which each site has been described by geographical, morphological, and cultural variables. The  $n$ -dimensional matrix, thus formalized, has been transposed in a Boolean matrix to constitute the Input of the Auto-CM (Auto-Contractive Map) Artificial Neural Network. The Auto-CM can be

<sup>35</sup> By Giulia Massini (Semeion Research Center).



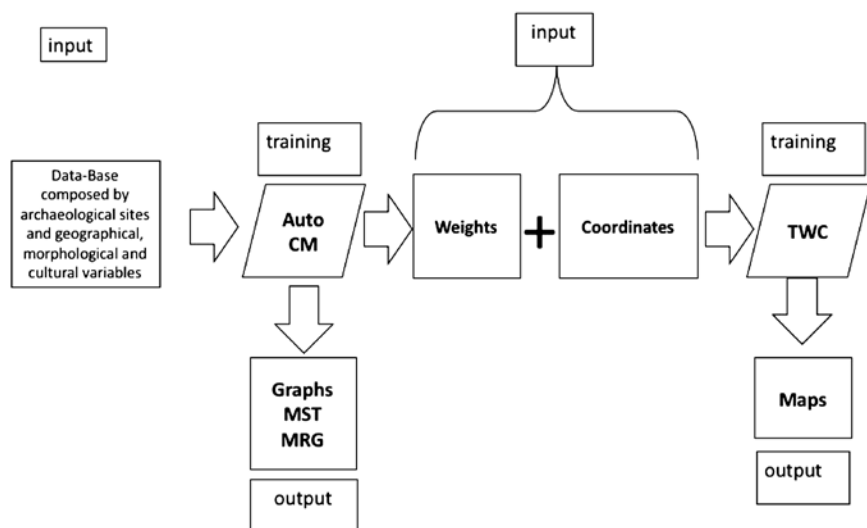


FIGURE 2.4 *Machine-learning system for the neural simulation of the MURL*  
(© SEMEION – GIULIA MASSINI)

viewed as a neural network with three layers of units, with each unit being connected to the next unit as shown in Fig. 2.1a.<sup>36</sup>

2. After the training phase, the weights developed by Auto-CM are proportional to the strength of the associations of all records to each other. The weights  $W$  are then transformed into physical distances and saved on a symmetric matrix of relationships between records, with a null diagonal. The matrix of distances derived from the Auto-CM matrix of similarities has been explored through graph theory. A graph is a mathematical abstraction that basically consists of a set of vertices and a set of edges, where an edge represents the connection between two vertices on a graph.
3. A distance matrix among vertices  $V$  represents an undirected graph, where each vertex relates to all the others except for itself. Consequently, each distance between a pair of nodes (archaeological sites) becomes the weighted edge between this pair of nodes. A simple mathematical filter represented by MST is applied to

<sup>36</sup> The Auto-CM neural network was designed by Buscema in 1999 and is characterized by a three-layer architecture consisting of 1) an input layer, where the signal is captured from the environment; 2) a hidden layer, where the signal is modulated inside the Auto-CM; and 3) an output layer, through which the Auto-CM feeds back upon the environment on the basis of the stimuli previously received and processed. The top layer is the input layer, and these are equivalent to the variables among which we seek to identify relationships.

the distances matrix, and a graph is generated, but to obtain major complexity in information, MRG was used.

4. The MRG describes other high-value relationships between records that are obtained from the Auto-CM analysis and are not shown by the MST. Such MRG relationships are represented as arcs creating circuits. Based on the values of the proximity relations between the archaeological settlements obtained by the Auto-CM, ANN has been connected to the value of the physical proximities. In other words, the values of proximity thus obtained has been linked to the spatial coordinates (latitude and longitude) to display the neural morphology of the settlement distributions (see TWC maps).

### Appendix 2.2. The Topological Weighted Centroid Basic Definitions<sup>37</sup>

The basic intuition behind the Topological Weighted Centroid (TWC) approach is that every distribution of a point in space has an implicit semantics, provided the following conditions are met: 1) each point of the distribution represents a discrete occurrence of the same process, 2) the distribution of points is statistically representative of the process to be analyzed, and 3) TWC is a set of ordered mathematical quantities that transforms the discrete dataset into different scalar fields: Alpha, Beta, Gamma, Theta, and Iota (Fig. 2.5).

The TWC Alpha ( $\alpha$ ) represents a spatial estimate of the hidden (outbreak) point, or area, where the process under study originated. The ideal candidate for an outbreak is a region that represents the portion of space from which the information needed to code and retrieve the relative position of all the other events is optimized—that is, the location from which the global entropy of the distances from all the other points attains its minimum level, and the correlated Free Energy is maximum. TWC Alpha represents the more rational past of the process.

The TWC Beta ( $\beta$ ) is the spatial estimation of the relevance of the areas probabilistically linked to the process represented by the distribution of the points to be analyzed. The closer the points are to each other, the stronger the attraction they exert. The TWC Beta represents the actual state of the process.

The TWC Gamma ( $\gamma$ ) will describe the extent to which each point of the space is activated by its closeness to any of the points belonging to any of the nonlinear trajectories connecting each point to the center of the mass. The Gamma scalar field may therefore be thought of as a further qualification of the Beta field, which transforms attraction strengths into intensities of network interaction and therefore highlights

---

<sup>37</sup> By Paolo Massimo Buscema (Semeion Research Center).

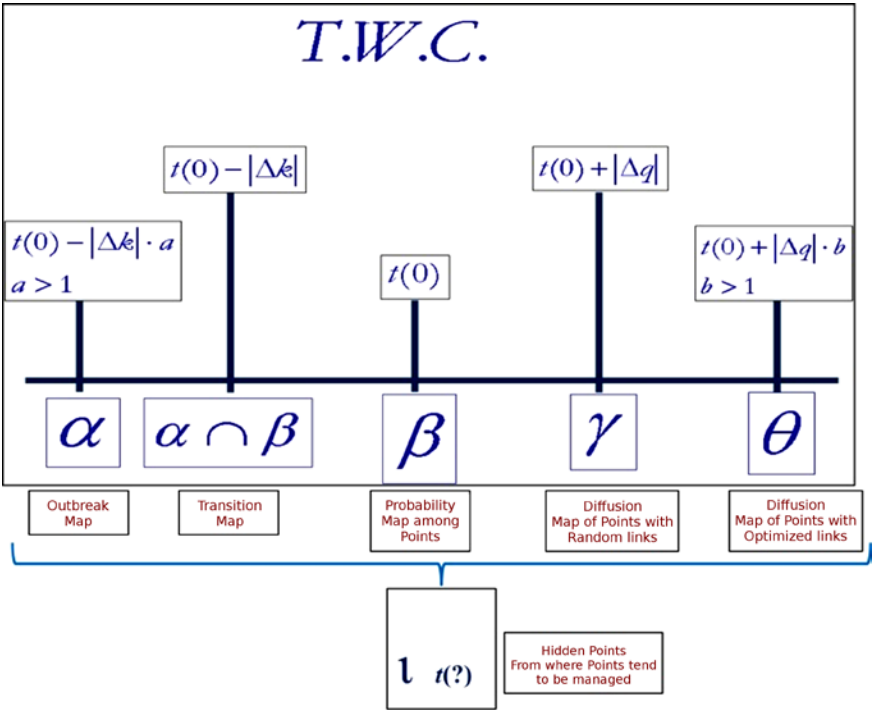


FIGURE 2.5 *TWC algorithms* (© SEMEION – PAOLO M. BUSCEMA)

longer-term activation patterns within the spatial distribution. The TWC Gamma represents the next evolution of the process.

The TWC Theta ( $\theta$ ) allows one to build the nonlinear MST that links together the points of  $e$  spatial distribution in terms of a minimal network of influence. The nonlinear MST generated represents the ideal network of intercommunication among points if the contextual conditions remain stable. The TWC Theta represents the final evolution of the process.

The TWC Iota ( $\iota$ ) allows one to estimate the vanishing points of the actual distribution of points. It indicates hidden areas where the assigned points could be clustered in an optimal way. From an empirical viewpoint, these areas may indicate the points from which the process is managed.

References

Adams, Robert McCormick. 1955. "Developmental Stages in Ancient Mesopotamia." In *Irrigation Civilizations: A Comparative Study*, edited by Julian H. Seward, 6–18.



- Washington, DC: Social Science Section, Department of Cultural Affairs, Pan American Union.
- Adams, Robert McCormick. 1956. "Some Hypotheses on the Development of Early Civilization." *AmAnt* 21 (3): 227–232.
- Adams, Robert McCormick. 1972a. "Pattern of Urbanization in Early Southern Mesopotamia". In *Man, Settlement and Urbanism*, edited by Peter J. Ucko, Ruth Tringham, and Geoffrey W. Dimbleby, 735–749. London: Duckworth.
- Adams, Robert McCormick. 1972b. "The Urban Revolution in Lowland Mesopotamia." In *Population Growth: Anthropological Implication*, edited by Brian Spooner, 60–62. Cambridge, MA: MIT Press.
- Adams, Robert McCormick. 1978. "Strategies of Maximization, Stability, and Resilience in Mesopotamian Society, Settlement and Agriculture." *Proceedings of the American Philosophical Society* 122: 329–335.
- Adams, Robert McCormick. 1981. *Heartland of Cities, Surveys of Ancient Settlement and Land Use on the Central Floodplain of the Euphrates*. Chicago: The Oriental Institute of the University of Chicago.
- Adams, Robert McCormick. 2008. "An Interdisciplinary Overview of a Mesopotamian City and its Hinterlands." *Cuneiform Digital Library Journal* 1. <[http://cdli.ucla.edu/pubs/cdlj/2008/cdlj2008\\_001.html](http://cdli.ucla.edu/pubs/cdlj/2008/cdlj2008_001.html)>.
- Adams, Robert McCormick, and Hans J. Nissen. 1972. *The Uruk Countryside. The Natural Setting of Urban Societies*. Chicago: The Oriental Institute of the University of Chicago.
- Andersen, Peter B. 1991a. "Computer Semiotics." *SJIS* 3: 3–30.
- Andersen, Peter B. 1991b. *A Theory of Computer Semiotics*. Cambridge: Cambridge University Press.
- Anderson, James A., and Edward Rosenfeld, eds. 1988. *Neurocomputing Foundations of Research*. Cambridge, MA: MIT Press.
- Barceló, Juan A. 2008. *Computational Intelligence in Archaeology. Investigations at the Interface between Theory, Technique and Technology in Anthropology, History and the GeoSciences*. London: IGI Global.
- Barthélemy, Marc. 2011. "Spatial Networks." *Physics Reports* 499: 1–101.
- Beckerman, Martin. 1997. *Adaptive Cooperative Systems*. New York: Wiley-Interscience.
- Benco, Nancy L. 1992. "Manufacture and Use of Clay Sickles from the Uruk Mound, Abu Salabikh." *Paléorient* 18: 119–134.
- Bentley, R. Alexander, and Stephen J. Shennan. 2003. "Cultural Transmission and Stochastic Network Growth." *AmAnt* 68: 459–485.
- Bintliff, John 2005. "Being in the (Past) World: Vermeer, Neural Networks and Archaeological Theory." In *Die Dinge als Zeichen: Kulturelles Wissen und materielle Kultur*, edited by Tobias L. Kienlin, 125–131. Bonn: Habelt.

- Brabazon, Anthony, Michael O'Neill, and Seán McGarraghy. 2015. *Natural Computing Algorithms*. Berlin – Heidelberg: Springer.
- Brandes, Mark A. 1971. *Untersuchungen zur Komposition der Stiftmosaiken an der Pfeilerhalle der Schicht iva in Uruk-Warka*. Baghdader Mitteilungen Beihefte 6. Berlin: Mann.
- Brinkman, John A. 1984. "Settlement Surveys and Documentary Evidence: Regional Variation and Secular Trend in Mesopotamian Demography." *JNES* 43 (3): 169–180.
- Bughmans, Tom. 2010. "Connecting the Dots. Towards Archaeological Network Analysis." *OxfJA* 29 (3): 277–303.
- Buscema, Paolo M. 2014. "The General Philosophy of Artificial Adaptive Systems." In *Archeosema. Artificial Adaptive Systems for the Analysis of Complex Phenomena. Collected Papers in Honour of David Leonard Clarke*, edited by Marco Ramazzotti, 53–84. ACalc: Supplemento 6. Florence: All'Insegna del Giglio.
- Buscema, Paolo M., Masoud Asadi-Zeydabadi, Weldon A. Lodwick, and Marco Breda. 2016. "The H Function: A New Index for Detecting Structural/Topological Complexity Information in Undirected Graphs." *Physica A* 447: 355–378.
- Buscema, Paolo M., Marco Breda, and Luigi Catzola. 2009. "The Topological Weighted Centroid, and the Semantic of the Physical Space-Theory." In *Artificial Adaptive Systems in Medicine*, edited by Paolo M. Buscema and Enzo Grossi, 68–78. London: Bentham.
- Buscema, Paolo M., Marco Breda, Enzo Grossi, Luigi Catzola, and Pierluigi Sacco. 2013. "Semantics of Point Spaces through the Topological Weighted Centroid and Other Mathematical Quantities: Theory and Applications." In *Data Mining Applications Using Artificial Adaptive Systems*, edited by William J. Tastle, 75–139. New York-London: Springer.
- Buscema, Paolo M., Enzo Grossi, and Tom Jefferson. 2009. "The Topological Weighted Centroid, and the Semantic of the Physical Space – Application." In *Artificial Adaptive Systems in Medicine*, edited by Paolo M. Buscema and Enzo Grossi, 79–89. London: Bentham.
- Buscema, Paolo M., Riccardo Petritoli, Giovanni Pieri, and Pierluigi Sacco. 2009. *Auto-Contractive Maps*. Semeion – Technical Paper 32. Roma: Aracne.
- Buscema, Paolo M., and Pierluigi Sacco. 2010. "Auto-Contractive Maps, the H Function, and the Maximally Regular Graph (MRG): A New Methodology for Data Mining." In *Applications of Mathematics in Models, Artificial Neural Networks and Arts*, edited by Vittorio Capecchi, 227–276. New York-London: Springer.
- Buscema, Paolo M., and Pierluigi Sacco. 2016. "MST Fitness Index and Implicit Data Narratives: A Comparative Test on Alternative Unsupervised Algorithms." *Physica A* 461: 726–746.
- Buscema, Paolo M., Pierluigi Sacco, Guido Ferilli, Marco Breda, and Enzo Grossi. 2015. "Analyzing the Semantics of Point Spaces Through the Topological Weighted Centroid

- and Other Mathematical Quantities: The Hidden Geometry of the Global Economic Order." *Computational Intelligence* 31 (3): 532–567.
- Buscema, Paolo M., Pierluigi Sacco, Enzo Grossi, and Weldon A. Lodwick. 2013. "Spatiotemporal Mining: A Systematic Approach to Discrete Diffusion Models for Time and Space Extrapolation." In *Data Mining Applications Using Artificial Adaptive Systems*, edited by William J. Tastle, 231–275. New York-London: Springer.
- Childe, Vere Gordon. 1950. "The Urban Revolution." *Town Planning Review* 21: 3–17.
- Clarke, David L. 1968. *Analytical Archaeology*. London: Methuen and Co.
- Clarke, David L. 1972. *Models in Archaeology*. London: Methuen.
- Clarke, David L. 1977. *Spatial Archaeology*. Boston: Academic Press.
- Collar, Anna, Fiona Coward, Tom Brughmans, and Barbara J. Mills. 2015. "Networks in Archaeology: Phenomena, Abstraction, Representation." *JAMT* 22: 1–32.
- Cooper, Jerrold. 1985. "Medium and Message: Inscribed Clay Cones and Vessels from Presargonic Sumer." *RA* 79: 97–114.
- Figge, Udo L. 1991. "Computersemiotik." *Zsem* 13 (3-4): 321–330.
- Finkbeiner, Üwe. 1986. "Uruk-Warka: Evidence of the Ġamdat Nasr-Period." In *Ġamdet Nasr. Period or Regional Style? Papers Given at a Symposium Held in Tübingen, November 1983*, edited by Üwe Finkbeiner and Wolfgang Röllig, 33–56. Beihefte zum Tübingen Atlas der Vorderen Orients, Reihe B 62. Wiesbaden: Reichert.
- Gibson, McGuire. 1973. "Population Shift and the Rise of Mesopotamian Civilization." In *The Explanation of Culture Change: Models in Prehistory*, edited by Colin Renfrew, 447–463. Pittsburgh: University of Pittsburgh Press.
- Hage, Per, Frank Harary, and James Brent. 1996. "The Minimum Spanning Tree Problem in Archaeology." *AmAnt* 61: 149–155.
- Kendall, David G. 1971. "Seriation from Abundance Matrices." In *Mathematics in the Archaeological and Historical Sciences*, edited by Frank R. Hodson, David G. Kendall, and Petre Tăutu, 215–252. Edinburgh: Edinburgh University Press.
- Knappett, Carl. 2011. *An Archaeology of Interaction: Network Perspectives on Material Culture and Society*. Oxford: Oxford University Press.
- Kohler, Timothy. 2012. "Complex Systems and Archaeology." In *Archaeological Theory Today*, edited by Ian Hodder, 93–123. Cambridge: Polity Press.
- Kohonen, Teuvo. 1982. "Self-Organized Formation of Topologically Correct Feature Maps." *Biological Cybernetics* 43: 59–69.
- Kohonen, Teuvo. 1996. "Emergence of Invariant-Feature Detectors in the Adaptive-Subspace Self-Organizing Map." *Biological Cybernetics* 75: 281–291.
- Kruskal, Joseph B. 1956. "On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem." *Proceedings of the American Mathematical Society* 7 (1): 48–50.
- Leeuw, Sander van der. 2013. "Archaeology, Networks, Information Processing, and beyond." In *Network Analysis in Archaeology. New Approaches to Regional Interaction*, edited by Karl Knappett, 335–349. Oxford: Oxford University Press.

- Liverani, Mario. 2013. *Immaginare Babele. Due secoli di studi sulla città orientale antica*. Rome-Bari: Laterza.
- Ludovico, Alessandro di, and Marco Ramazzotti. 2008. "Reconstructing Lexicography in Glyptic Art: Structural Relations between the Akkadian age and the Ur III Period." In *Proceedings of the 51st Rencontre Assyriologique Internationale, Held at the Oriental Institute of the University of Chicago, July 18–22 2005*, edited by Robert D. Biggs, Jennie Myers, and Martha T. Roth, 263–280. SAOC 62. Chicago: The Oriental Institute of the University of Chicago.
- Massini, Giulia. 2010. "Multi-Meta-Som." In *Applications of Mathematics in Models, Artificial Neural Networks and Arts*, edited by Vittorio Capecchi, 313–348. New York-London: Springer.
- McClelland, James L., and David E. Rumelhart. 1988. *Explorations in Parallel Distributed Processing*. Cambridge, MA: MIT Press.
- Mézard, Marc, and Thierry Mora. 2009. "Constraint Satisfaction Problems and Neural Networks: A Statistical Physics Perspective." *Journal of Physiology* 103 (1-2): 107–113.
- Miller, John H., and Scott E. Pages. 2007. *Complex Adaptive Systems. An Introduction to Computational Models of Social Life*. Princeton Studies in Complexity. Princeton: Princeton University Press.
- Minsky, Marvin. 1986. *The Society of Mind*. New York: Simon & Schuster.
- Mlekuž, Dimitrij. 2014. "Exploring the Topography of Movement." In *Computational Approaches to the Study of Movement in Archaeology*, edited by Silvia Polla and Philip Verhagen, 5–22. Berlin: De Gruyter.
- Nissen, Hans J. 1980. "The Mobility between Settled and Non-Settled in Early Babylonia: Theory and Evidence." In *L'Archéologie de l'Iraq du début de l'époque néolithique à 333 avant notre ère*, edited by Marie-Thérèse Barrelet, 285–290. Paris: Éditions du CNRS.
- Nissen, Hans J. 1983a. "Settlement Patterns and Material Culture of the Akkadian Period: Continuity and Discontinuity." In *Akkad, the First World Empire. Structure, Ideology, Traditions*, edited by Mario Liverani, 91–106. Padova: Sargon.
- Nissen, Hans J. 1983b. "The Urban Revolution of Mesopotamia Reconsidered." In *Studies in the Neolithic and Urban Revolutions. The V. Gordon Childe Colloquium, Mexico 1986*, edited by Linda Manzanilla, 287–294. BAR-IS 349. Oxford: Archaeopress.
- Nissen, Hans J. 2001. "Cultural and Political Networks in the Ancient Near East during the Fourth and Third Millennia BCE." In *Mesopotamia and its Neighbors: Cross-cultural Interactions and Their Consequences in the Era of State Formation*, edited by Mitchell S. Rothman, 149–179. Santa Fe: SAR.
- Porada, Edith, Donald P. Hansen, Sally Dunham, Sidney H. Babcock. 1992. "The Chronology of Mesopotamia, ca. 7000–1600 B.C." In *Chronologies in Old World Archaeology, vol. 1.*, 3rd ed., edited by R.W. Enrich, 77–121. Chicago: University of Chicago Press.
- Ramazzotti, Marco. 1997. "La fase 'Middle Uruk': studio tramite Reti Neurali Artificiali su un orizzonte latente nella protostoria della Bassa Mesopotamia." In *Studi in me-*

- morìa di Henri Frankfort (1897-1954) presentati dalla scuola romana di Archeologia Orientale, edited by Paolo Matthiae, 495–522. CMAO VII. Roma: La Sapienza.
- Ramazzotti, Marco. 1999. *La Bassa Mesopotamia come laboratorio storico in età proto-storica. Le Reti Neurali Artificiali come strumento di ausilio alle ricerche di archeologia territoriale*. CMAO VIII. Roma: La Sapienza.
- Ramazzotti, Marco. 2000. "Dall'analisi diacronica all'analisi sincronica: indagine sulle dinamiche insediamentali del periodo Jemdet Nasr nella regione di Warka." *ScAnt* X: 9–38.
- Ramazzotti, Marco. 2002. "La 'Rivoluzione Urbana' nella Mesopotamia meridionale. Replica 'versus' processo." *Accademia Nazionale dei Lincei. Classe delle Scienze Morali Storiche e Filologiche, Rendiconti IX* 13: 651–752.
- Ramazzotti, Marco. 2003. "Modelli insediamentali alle soglie del Protodinastico in Mesopotamia meridionale, centrale e nord-orientale. Appunti per una critica alla formazione 'secondaria' degli stati nel III Millennio a.C." *CMAO IX*: 15–71.
- Ramazzotti, Marco. 2009. "Lineamenti di archeologia del paesaggio mesopotamico. Descrizioni statistiche e simulazioni artificiali adattive per un'analisi critica della demografia sumerica e accadica." In *Geografia del popolamento*, edited by Giancarlo Macchi Jánica, 193–202. Siena: Fieravecchia.
- Ramazzotti, Marco. 2010. *Archeologia e Semiotica. Linguaggi, codici, logiche e modelli*. Turin: Bollati Boringhieri.
- Ramazzotti, Marco. 2013a. "Where Were the Early Syrian Kings of Ebla Buried? The Uridu Survey Neural Model as an Artificial Adaptive System for the Probabilistic Localization of the Ebla Royal è madím." *ScAnt XIX*: 10–34.
- Ramazzotti, Marco. 2013b. "Archeosema. Sistemi Artificiali Adattivi per un'Archeologia Analitica e Cognitiva dei Fenomeni Complessi." *ACalc* 24: 283–303.
- Ramazzotti, Marco. 2013c. "Logic and Semantics of Computational Models for the Analysis of Complex Phenomena. Analytical Archaeology of Artificial Adaptive Systems." In *Urban Coastal Area Conflicts Analysis Methodology: Human Mobility, Climate Change and Local Sustainable Development*, edited by Armando Montanari, 23–56. Rome: La Sapienza.
- Ramazzotti, Marco. 2013d. *Mesopotamia antica. Archeologia del pensiero creatore di miti nel Paese di Sumer e di Accad*. Rome: Editoriale Artemide.
- Ramazzotti, Marco. 2014a. "Analytical Archaeology and Artificial Adaptive Systems." In *Archeosema. Artificial Adaptive Systems for the Analysis of Complex Phenomena. Collected Papers in Honour of David Leonard Clarke*, edited by Marco Ramazzotti, 15–52. ACalc: Supplemento 6. Florence: All'Insegna del Giglio.
- Ramazzotti, Marco. 2014b. "Analytical Archaeology and Artificial Adaptive Systems Laboratory (LAA&AAS)." In *Archeosema. Artificial Adaptive Systems for the Analysis of Complex Phenomena. Collected Papers in Honour of David Leonard Clarke*, edited by Marco Ramazzotti, 53–84. ACalc: Supplemento 6. Florence: All'Insegna del Giglio.

- Ramazzotti, Marco. 2016a. "Archeologia e traduzione. Prolegomena alla meccanografia e alla simulazione artificiale del sema." In *Il segno tradotto. Idee, immagini, parole in transito*, edited by Marco Ramazzotti, Simone Celani, and Francesco Fava, 17–26. Milan: Marcos y Marcos.
- Ramazzotti, Marco. 2016b. "Back to the Future. Structuring an Analytical Model for the Mesopotamian Urbanism: a View from the South." In *Trajectories of Complexity. Socio-economic Dynamics in Upper Mesopotamia in the Neolithic and Chalcolithic Periods*, edited by Marco Iamoni and Salam al Quntar, 183–194. StTr 6. Wiesbaden: Harrassowitz.
- Redman, Charles L. 1978a. *The Rise of Civilization. From Early Farmers to Urban Society in the Ancient Near East*. San Francisco: W. H. Freeman and Co.
- Redman, Charles L. 1978b. "Mesopotamian Urban Ecology: The Systemic Context of the Emergence of Urbanism." In *Social Archaeology. Beyond Subsistence and Dating*, edited by Charles L. Redman, 329–348. New York: Academic Press.
- Reeler, Claire. 1999. "Neural Networks and Fuzzy Logic Analysis in Archaeology." In *Archaeology in the Age of the Internet. Proceedings of the 25th Anniversary Conference, University of Birmingham, April 1997*, edited by Lucie Dingwall, Sally Exon, Vince Gaffney, Sue Laflin, and Martijn van Leusen, 3–10. BAR-IS 750. Oxford: Archaeopress.
- Rothman, Mitchel S. 1987. "Graph Theory and the Interpretation of Regional Survey Data." *Paléorient* 13 (2): 73–91.
- Shuchat, Alan H. 1984. "Matrix and Network Models in Archaeology." *Mathematics Magazine* 57 (1): 3–14.
- Smith, Michael E. 2011. "Empirical Urban Theory for Archaeologists." *JAMT* 18: 167–192.
- Smolensky, Paul. 1987. "Connectionsit AI, Symbolic AI, and the Brain." *Artificial Intelligence Review* 1: 95–109.
- Stein, Gil. 2010. "Local Identities and Interaction Spheres: Modeling Regional Variation in the 'Ubaid Horizon.'" In *Beyond the Ubaid: Transformation and Integration in the Late Prehistoric Societies of the Middle East*, edited by Robert A. Carter and Graham Philip, 23–44. SAOC 63. Chicago: The Oriental Institute of the University of Chicago.
- Weiss, Harvey. 1977. "Periodization, Population and Early State Formation in Khuzestan." In *Mountains and Lowlands: Essays in the Archaeology of Greater Mesopotamia*, edited by Louis Levine and Cuyler Young, 347–369. Malibu: Undena Publications.
- Wright, Henry T. 1977. "Recent Research on the Origin of the State." *ARA* 6: 379–397.
- Wright, Henry T. 1978. "Toward an Explanation of the Origin of the State." In *Origins of the State: The Anthropology of Political Evolution*, edited by Ronald Cohen and Elman R. Service, 49–68. New York: Institute for the Study of Human Issues.
- Wright, Henry T. 1981. "The Southern Margins of Sumer: Archaeological Survey of the Area of Eridu and Ur." In *Heartland of Cities*, edited by Robert McCormick Adams, 297–362. Chicago-London: University of Chicago Press.

- Wright, Henry T. 1984. "Prestate Political Formations." In *On the Evolution of Complex Societies: Essays in Honor of Harry Hoijer*, edited by William T. Sanders, Henry T. Wright, and Robert McCormick Adams, 41–78. Malibu: Undena.
- Wright, Henry T. 1986. "The Susiana Hinterlands during the Era of Primary State Formation." In *The Archaeology of Western Iran. Settlement and Society from Prehistory to the Islamic Conquest*, edited by Frank Hole, 141–155. Washington, DC: Smithsonian Institution Press.
- Wright, Henry T. 1998. "Uruk States in Southwestern Iran." In *Archaic States*, edited by Gary Feinman, and Joyce Marcus, 173–192. Santa Fe: SAR.
- Wright, Henry T. 2001. "Cultural Action in the Uruk World". In *Mesopotamia and its Neighbors: Cross-cultural Interactions and their Consequences in the Era of State Formation*, edited by Mitchell S. Rothman, 123–148. Santa Fe: SAR.
- Wright, Henry T., and Gregory A. Johnson. 1975. "Population, Exchange, and Early State Formation in Southwestern Iran." *AmA* 77: 267–289.
- Young, Timothy C. 1972. "Population Densities and Early Mesopotamian Urbanism." In *Man, Settlement and Urbanism*, edited by Peter J. Ucko, Ruth Tringham, and Geoffrey W. Dimbleby, 827–842. London: Duckworth.
- Zubrow, Ezra B. W. 2003. "The Archaeologist, the Neural Network, and the Random Pattern: Problems in Spatial and Cultural Cognition." In *The Reconstruction of Archaeological Landscapes through Digital Technologies. Italy-United States Workshop, Boston, Massachusetts, USA, November, 1-3, 2001*, edited by Maurizio Forte and Patrick R. Williams, 173–180. British Archaeological Reports 1151. Oxford: Oxbow.



## PART 2

### *Objects*







## Data Description and the Integrated Study of Ancient Near Eastern Works of Art: The Potential of Cylinder Seals

*Alessandro di Ludovico\**

... Roboti si všechno pamatují, ale nic víc.

Dokonce se ani nesmějí tomu, co lidé říkají.<sup>1</sup>

KAREL ČAPEK, *R.U.R. Rossum's Universal Robots* (1920)



For several decades, the investigation of ancient Near Eastern cultures has needed to operate within a complex situation of modern conflicts and political instability. In some areas of great importance, such as Mesopotamia (situated in present-day Iraq and Syria), direct contact between scholars and archaeological sites and museums has been almost completely interrupted for decades, and the integrity of cultural heritage is now very precarious across Western Asia. Furthermore, since the beginning of the age of modern archaeological exploration in the region, both authorized excavations and (especially) illicit digs and the flourishing market for antiquities have led to the international dispersion of ancient artifacts into public and private collections. This removal of material products of ancient cultures from their regions of origin brings with it multiple complications affecting the work of the researcher. Most significantly, it limits the opportunity for a clear interpretation of the artifacts and the correct historiography of the civilizations to which they pertain. This is true in particular for most ancient Western Asiatic cultures, of which the largest part remained forgotten for many centuries, and a number

---

\* The author of this contribution wants to express his gratitude to Sergio Camiz for all of his help and precious suggestions regarding computer applications and statistical algorithms, tools, and reasoning, which were necessary in carrying out the analyses.

1 "The robots remember everything, but that's all they do. They don't even laugh at what people tell them." (Translated from the original Czech by David Wyllie for Penguin's edition of the drama, London-New York, 2004).

are still very poorly known. The lack of continuity through the ages of a historical memory of these peoples, including testimony relating to most of their contacts with other cultures (especially Mediterranean Europe) and the transmission of their own traditions, puts the scholar at risk for very easy misunderstandings and involuntary prejudices. Visual language or written documents, for example, can frequently become sources for a large number of ambiguities and misinterpretations, since their contents can be seemingly familiar, but the comprehension of their contexts of origin, and especially of their relative cultural values,<sup>2</sup> is very often insufficient.

The traditional art-historical approach finds in the artifacts themselves enough information that the loss of knowledge about their find contexts does not present a serious hurdle to their interpretation. One consequence of this approach is that interpretations that could seem obvious, evident, or taken for granted can hardly be incorporated with other evidence resulting from the same (or a closely related) context. Therefore, this paper advocates for the contextualized interpretation of materials, since the lives and possible meanings of ancient artifacts are very dynamic, and their interpretation is equally dynamic and complex.

### Approaching Artifacts

In addition to find contexts (when such information is available), cross-comparison and experimental approaches open the way for new perspectives on ancient materials, including promising re-studies of long-known corpora. New studies of old corpora are needed especially when the material is of unknown or not homogeneously documented provenance. For those who are accustomed to the use of digital tools, it is not surprising that quantitative and computer-aided methods and procedures can play a central role in this research,

---

2 The issue of relative cultural values pertains to the proper contextualization of the values of symbols and concepts that are expressed through images or written words. The scholar needs to face not only the deciphering of their possible meanings and roles in the specific context of use, but also the factors that played a role in their production, perception, and interpretation in the original cultures. Furthermore, it is of great importance to locate the logical context in which such production and perception acted, since it is fundamental for the identification of the correct historical and cultural contexts in which these witnesses acted. For example, modes of communication related to an oral and rural context need to be considered according to general logical principles that are very different from those coming from official written sources (on this theme see, for example, Goody 1986).

for which it is necessary to build and populate data archives and to develop strategies for cross-comparison and analysis.

The ideal following step would be to open the archives and share the data, but this still seems to be a chimeric goal, in particular because of an entrenched institutional and scholarly culture of keeping exclusive access to primary files. In this vein, a fairly pessimistic view was expressed many years ago by Jean-Claude Gardin.<sup>3</sup> Within the frame of the French school of *analyse logiciste*,<sup>4</sup> but considering theoretical developments at the international level (especially originating from Great Britain),<sup>5</sup> already in the 1950s, Gardin outlined the basic principles for computerized systems that could automatically categorize artifacts and support quantitative studies.<sup>6</sup> His research aimed to devise universal codes that could describe and represent with suitable precision any kind of archaeological material. Such codes were meant to be used on a large scale, would have permitted scholars to carry out comparisons quickly, and would have enabled the homogeneous publication and classification of materials.

The further ambition of Gardin was to extend this system so that it could translate the full content of excavation reports into a rigorously formalized language.<sup>7</sup> This approach—largely inspired by linguistics—considered the importance of using a limited number of elements to describe all aspects of available information.<sup>8</sup> His encoding systems were, of course, different, for each of his main categories of operations and artifacts (such as glyptic, pottery, and simple tools), but the final target was their universal application.

Within the field of Western Asiatic studies, Gardin's work was among the first long-term experiments on the application of quantitative and automated methods. His earliest tests used punch cards and computer machines, and the first concrete outcome of this methodology was the *Répertoire analytique des cylindres orientaux*,<sup>9</sup> a large open catalogue of cylinder seals. It was designed to overcome the problems of the dispersion and heterogeneous documentation (through diverse, isolated publications, mainly in journals) of this category of artifact by systematizing the seals' descriptions and therefore facilitating focused queries.

---

3 Gardin 1955, 107–115.

4 *Analyse logiciste*: an analytical procedure developed in France focusing on outlining and highlighting the symbolic elements (and their relevant functions) that can describe the structure of an observed phenomenon. The logicist procedure is also a critical review of the methodologies and the epistemological perspectives adopted by different scholars (Gardin 1997).

5 Moscati 2013, 9.

6 Gardin 1958, 335–336; Gardin, 1967, 13.

7 Gardin 1958; Gardin 2002, 19–21.

8 Gardin 1966; Gardin 1967, 18–26.

9 Digard 1975. See also Gardin 1967, 21–26.

It was only possible to produce the *Répertoire* with the generous collaboration of many French scholars who joined the initial efforts in collecting data on seals according to an established, uniform procedure. Gardin himself, however, foresaw that the project would never be enlarged, although it was widely appreciated by the scientific community. The basic hurdle for the future use and expansion of the catalogue, the value of which is evident from the methodological and practical points of view, was the unlikely availability of scholars willing to dedicate their own time and energies to a project that would not provide them with a personally fruitful outcome.

### The Peculiarities of Cylinder Seals

Since Gardin, the field of ancient Near Eastern studies has pursued some encoding projects whose logics and structures were comparable to those of the *Répertoire*. The best-known project is the Cuneiform Digital Library Initiative (CDLI), which is still active, thanks both to the care of its coordinators and to modern technology.<sup>10</sup> Some other developments also followed the French *logiciel* example, but they did not succeed in acquiring a large following and have all but fallen into oblivion.<sup>11</sup>

It is most notable, however, that in the field of quantitative applications to ancient Near Eastern visual languages, the majority of research has been dedicated to cylinder seals.<sup>12</sup> It is therefore worth taking a look at the relevant features of these artifacts.

Cylinder seals have a number of features that give this class of artifacts a central role in the study of the pre-classic historical peoples of the Western Asiatic regions, since they are connected to the expressions of many different aspects of culture and social life. As far as we can currently understand, cylinder seals developed parallel to writing technologies in their earliest forms, and in later historical periods their use seems to have changed parallel to meaningful changes in writing systems and logics.<sup>13</sup> Furthermore, cylinder seals not only had an official administrative use as guarantees for the reliability of docu-

10 CDLI: <<http://www.cdli.ucla.edu>> (accessed February 12, 2017). For further information on CDLI, see in this volume, Pagé-Perron, 198–200; Eraslan, 285–286.

11 For a short overview of the main experiments, see Ludovico and Camiz 2015a, 31; Ludovico 2016, 119–122.

12 Ludovico and Camiz 2015a, 29–32; Ludovico 2016, 118–122.

13 A research project related to this topic is currently being developed by the author of this chapter, but also see the relevant observations by Ross (2014).

ments in the same way as a modern signature or a rubber stamp, but they also served as personal ornaments, amulets, and status symbols.<sup>14</sup> Other remarkable features that can be attached to seals by inference are concerned with their peculiar materiality and intense relationship with the body of the owners or users. Seals usually bore their representations in a cyclical structure carved in the negative onto a curved surface in order to produce a continuous narrative when rolled onto a clay surface. The relationship of the seal with the iconography carved on its surface and that of both the cylinder and its iconography (and its replicated impressions) with the body of the person using it were thus probably complex and deserve to be part of the subject matter of the study of glyptic. The presence of cylinder seals in ancient daily life was based on their concrete materiality: on their colors, materials, and weight; on the manner in which they were worn and their possible roles as adornments; and on their concrete contact with the body of the person physically holding, carrying, or wearing them.<sup>15</sup> With this in mind, the scholar should pay due attention to an implicit corollary warning deriving from what has been said here: the traditional laboratory seal impressions used to publish and study cylinder seals can reveal quite a lot of the iconographic aspects of these artifacts, but they are completely abstracted from the seal as a cultural and handcrafted product used in daily life, as well as in official and sacred contexts. Furthermore, the typically published “aseptic” modern representations of seals’ carvings find few parallels in actual concrete ancient seal impressions, which were clearly carried out according to different logics and handling procedures than are used by today’s archaeologists and museum curators.<sup>16</sup> In other words, since modern illustrations and the “roll-outs” reflecting modern aesthetics that are photographed and displayed in galleries set aside the whole physicality of seals, except for practical information about a few features (such as dimensions and type of stone), most publications of cylinder seals are able neither to communicate, nor to pay the necessary attention to, a huge part of the potential information and qualities that these artifacts contain.

The need for alternative means of representing cylinder seals, voiced by Julia M. Asher-Greve and Willem B. Stern in the 1980s,<sup>17</sup> and demonstrated in the more recent and very promising proposal by Martine de Vries-Melein and Paul

<sup>14</sup> Cassin 1960; Collon 1987, 131–137; Haussperger 1991, 62–67; Porada 1993, 563; Klengel-Brandt and Marzahn 1997, 229–230, pls. 16, 22–24; Nam 2008; Winter 2001.

<sup>15</sup> Collon 1987, 108–112; 2001, 22–23; Haussperger 1991, 45–48.

<sup>16</sup> Bahrani 2014, 129–130.

<sup>17</sup> Asher-Greve and Stern 1983; 1986.

Boon for a low-cost optimized rendering and publication of cylinder seals,<sup>18</sup> has been largely ignored by most scholars. The latter experiment is also particularly important because it raises an issue that is central to the application of digital methods in the humanities: first, the needs of scholars dealing with antiquities can be met by computer tools, which are subject to quick obsolescence, especially when these tools are designed with the contribution and feedback of the humanities' scientist. Second, the computer scientist cannot avoid taking into account quick and numerous updates in hardware and software systems. This makes each combined multi-disciplinary approach fleeting and so frustrates the need for a long-duration tool, such as an open catalogue. Such a catalogue, inspired by Gardin's early efforts, is fundamental if one considers the need for wide and homogeneous documentation, which would permit true comparisons and progressive enlargements of the datasets on different scales.

### Digital Humanities and the Study of Cylinder Seals: The Case of Presentation Scenes

The cyber-approaches that will be introduced here began in response to the lack of comprehensive information provided by traditional catalogues of cylinder seals. In the course of some experiments I conducted on the iconography of a corpus of Mesopotamian cylinder seals dating from the Akkadian to the Ur III periods,<sup>19</sup> I discovered just how little data primary sources usually provide about seals as physical objects. As has been mentioned, the iconographic content of cylinders tends to be shown in a modern abstract representation, omitting elements that would allow for the appreciation of the concrete relationships between images and the material structure of the artifact. Technical problems that complicate an ideal representation of seals in catalogues—especially printed ones—are many and can be well understood and justified, es-

18 Boon and de Vries-Melein 2009, <<https://doi.org/10.17026/dans-x2d-cexg>> (accessed April 25, 2017); 2013.

19 Ludovico 2005; 2008. There are still some uncertainties in defining the absolute chronology of the mentioned periods; however, following the Middle Chronology, they can be dated as follows: Akkadian period, c. 2340-2155 BCE; Post-Akkadian (or Early Ur III) period, c. 2155-2110 BCE; Ur III (or the Third Dynasty of Ur) period, c. 2110-2005 BCE. In this volume, for short philological and epigraphical observations concerning these periods, see Pagé-Perron, especially 196n10, and 201n32. The Ur III period is also known (mainly from the epigraphic and philological point of view) as the "Neo-Sumerian Period" (see in this volume, Nurmikko-Fuller, 349, 351).

pecially because of the venerable age of some important and famous volumes. Today's primary publications of cylinder seals, however, still follow the original conventions, although they could certainly be more generous with their information, especially regarding the nature of the seals.

The investigation presented here is the continuation of my research on presentation scenes in Mesopotamian glyptic, which has been based, since its first approaches and tests, on linguistic logics.<sup>20</sup> It is therefore fairly obvious that the most suitable means adopted for these inquiries should have a quantitative and digital nature, especially because of the need for finding correspondences and repetitions of combinations of minimal iconographic elements.

Different types of algorithms and data-mining strategies could match the approach and the theoretical background of this research well.<sup>21</sup> For this reason, a number of applications were carried out in the past, in order both to test the models and to refine and adjust the encoding procedures. Some trials are of course necessary at the very beginning to oversee and check the full process and its phases, from encoding principles to outcomes, and to exert the necessary critical and self-critical assessments, but unforeseen responses or issues might emerge later, during the actual analysis.<sup>22</sup> Furthermore, this way of con-

---

20 The first explanation of this is presented in Ludovico 2005. To summarize briefly: the scenes were subdivided into a number of minimal iconographic units, which, in the manner of the phonemes of a language, had no assumed meaning on their own, but rather built a meaning when they were combined with other basic units, forming a composition comparable to a visual sentence. In such a frame, the relative spatial relations of each basic iconographic unit in the field (in the manner of the relative position of letters in a written word) are of fundamental importance to encode and analyze the scenes. For these reasons, it is essential to consider the materiality of the seals and the fact that the field on which the scenes were carved is cylindrical.

21 The acquisition of information from a (usually) large dataset is also useful for additional purposes (such as a research path, a concrete application, etc.). Such a task can be accomplished through a large number of different types of algorithms and strategies. Typical of data mining is the search for correlations among features, like variables, within a certain dataset, so that the results of an investigation of that dataset can give clues to predict the figures related to those features in similar cases, or the possible development of some features in case some other features change in a certain way. Just as resources can be found and extracted from a mine through proper types of investigation and extraction techniques, useful information that was not evident at a first look can be extracted from a dataset through data-mining techniques.

22 Similar issues have been stressed in the research illustrated in this volume by Martino and Martino, 136–137. However, in the case of Martino and Martino, the different attempts aimed to adjust the algorithm, rather than the encoding, as it is the case here.



sidering cylinder seals and their iconography has few, if any, precedents. As a result, the experimental phase of this research project was long.

### Algorithms and the Dataset

The scenes have always been considered here as a complex message within which all components acquired their meaning mainly from their physical, syntactic connection with the others. Some issues were automatically raised by this approach, primarily the cyclical nature of the surface on which the representations were carved. As has been noted above, cylinder seals in three-dimensional space are much different from those on a plain surface. This issue is no small matter, since the people who produced and used such artifacts chiefly perceived their cylindrical materiality, not their abstract iconographic projections.

The encoding of the presentation scenes for the quantitative investigation was based on the subdivision of the cylindrical surface into areas organized through fixed reference points that were strictly related to the surface treatment at the moment in which the carving was planned.<sup>23</sup> The “origin” of the scene, i.e., the area of the surface in which the presentation scene begins and ends, was then chosen as a reference point, together with the first figure of the scene. The first figure is identified as the one receiving the presentation and can be always identified without ambiguity. All elements composing the scene were thus described both for their nature and appearance and for their relative positions and spatial relationships to each other.

In the earliest steps, only basic statistical observations were recorded for a dataset collecting alphanumeric encodings of relatively heterogeneous presentation scene specimens from the Akkadian, Post-Akkadian, and Ur III periods (Fig. 3.1).<sup>24</sup>

The same corpus was then re-encoded into a presence/absence dataset in which the imagery was further split into smaller elementary units.<sup>25</sup> This form fits much better for investigations with algorithms pertaining to the cat-

23 Ludovico 2005, 72–78; Ludovico and Ramazzotti 2008, 268–270; Ludovico 2011, 135–137; Ludovico, Camiz, and Pieri, 2013, 495–496; Ludovico and Camiz 2014, 9–13. For quantitative approaches applied to texts analysis, see in this volume, Bigot Juloux (162–163), Pagé-Perron (212, 216–217), Svärd, Jauhiainen, Sahala, and Lindén (226), and Monroe (257–259).

24 Ludovico 2005.

25 See in this volume, Martino and Martino (119) for a presence/absence dataset structured and designed according to the same logic.



egory of Artificial Neural Networks (ANNS, such as Self-Organizing Maps and others),<sup>26</sup> which allowed the establishment of scene classifications, interpretations concerning their logical connections, and an outline of their diachronic developments.<sup>27</sup> A relatively simple algorithm even permitted the graphic representation of virtual scenes that correspond to the average figures of each class thus outlined.<sup>28</sup>

The following phase of this research, which is still in progress, is based on the use of statistical algorithms that were originally developed to investigate textual data (Fig. 3.1). They are primarily textual correspondence analysis and hierarchical classification algorithms.<sup>29</sup> Compared to the ANNS, the latter

---

26 Artificial Neural Networks (ANNS) are a class of algorithms for which the design is inspired by the observation of the mechanisms of the human brain. One of the most interesting features of ANNS is their ability to learn, which implies an attitude to flexibility, and to adapt to specific situations and datasets. Self-Organizing Maps are a type of ANN devised by Teuvo Kohonen (1997); they have had great success and a number of concrete applications in many different fields. The logic underlying this type of ANN is the progressive distribution of data in an abstract map to form groups (potentially classes or subclasses), which can show, based on their measurable distances, the degree of similarity that has been recognized among them. Such a map is *self-organizing*, since it is built on the corpus of data; that is, the records of the dataset are observed and distributed on the map by the algorithm, so that their relative similarities build the map itself. The final map is thus the result of a progressive rearrangement of the abstract space and the group of records distributed throughout it. The abstract space is based on what the machine learns after each period (a “period” is a complete analysis of the whole dataset). The training phase, during which the machine learns the structure of the dataset, consists of a previously established number (often thousands) of periods. See also in this volume, ANNS applied a) to Archaeology, in Ramazzotti in particular, 64, 66–69; b) to Semantics, Svård, Jauhiainen, Sahala, and Lindén, 229, 246.

27 Ludovico and Ramazzotti 2008; Ludovico 2011; Ludovico and Pieri 2011b; Ludovico, Camiz, and Pieri 2013.

28 Ludovico and Pieri 2011a. The map obtained through a Self-Organizing Map process shows groups of records that are more or less similar to each other. Besides the table of distances expressing the degree of similarity among groups, the algorithm also generates a chart for each group that represents a sort of average profile for it in terms of the composition of the average values calculated among the members of the group. The algorithm SEME, devised by Giovanni Pieri, can associate stylized iconographic basic units with the features (including the description of the relative position in the scene of each iconographic unit) that are well represented in the profile of a group, so that a group can be summarized and described graphically by an abstract presentation scene.

29 For the algorithms used here, see Lebart and Salem (1988); Lebart and Salem (1994). See also Camiz and Rova 2001; Camiz 2004; Ludovico and Camiz 2014. Correspondence analysis is a quantitative method used to describe a corpus of qualitative data. It falls within

methods allow a clearer view of the entire mathematical process, giving the opportunity for complete control of the dynamics underlying the detection of similarities and differences. However, unlike ANNs, these techniques cannot work on extremely large datasets, and their outcomes suggest a smaller range of hypotheses useful for filling in gaps in cases of fragmentary documentation. The datasets that have been used so far in this research are, in any case, not vast and are still made up of reasonably well-preserved specimens. In addition, a further comparison between the two types of investigation (ANNs as well as textual correspondence analysis and hierarchical classification algorithms) confirmed that both systems give very interesting outcomes and can be used productively to complement one another.<sup>30</sup>

Approaches employing correspondence analysis have given sound preliminary results and proved that the corpus is fairly representative, which means that it is quite well suited to reflect the (much larger) whole original production of presentation scenes.<sup>31</sup> At least in some experiments, the relationships of the different iconographic elements and their documented compositional strategies with the toponyms to which the scenes are referred left some doubts. These uncertainties were mainly due to the high number of specimens that cannot be connected with any places of origin. They were therefore given the toponym “Unknown.”<sup>32</sup> To verify these observations, the dataset was recently enlarged, adding only specimens of known origins, and then two identical correspondence analyses were carried out: the first on the whole corpus and the second on a subgroup from which scenes of unknown origin were removed.

---

the statistical realm of “factor methods” and can be described as the analysis of the dependent relationships existing among the different features of the observed data. When the data are collected in a table, for example, the rows can be referred to as the observed specimens (in this case the seals’ scenes), while the columns refer to the different features (here, the different possible iconographic elements that appear in the scenes) that can pertain to the specimens. Correspondence analysis permits the investigation of the associations between the rows and columns of the table and enables associations to be displayed in a graphic form. Hierarchical classification is based on similarity and proximity. It classifies groups and shows hierarchical dependences among them. In the case mentioned here, a bottom-up approach was adopted, with records that were originally isolated and then progressively grouped, at higher levels, into larger groups.

30 Ludovico, Camiz, and Pieri 2013.

31 Ludovico and Camiz 2014; 2015b.

32 Ludovico and Camiz 2014; a more detailed study of this phenomenon is currently in preparation by the same authors.

The whole corpus includes 425 specimens; of these, only 142 are of unknown origin, so the second analysis was carried out on a corpus of 283 verbal encodings of presentation scenes.

The geographic origins of these specimens are as follows (Table 3.1):<sup>33</sup>

TABLE 3.1 *Corpus of specimens analyzed*

Origin	Number of scenes
Ur	45
Telloh	43
Susa	43
Umma	75
Lagaš	35
North-Bab (Babylon+Nippur+Drehem+Sippar)	26
South (Failaka+Larsa+Adab+Uruk+Garšana)	11
North-East (Subartu+Ešnunna)	5
Unknown	142

Although uneven, this distribution of specimens across different sites and regions has the potential to give a good general picture of the similarities and differences that exist among the iconographies used by different workshops or local administrations. In any case, there is another feature in the dataset that needs to be properly considered: the presence of a large number, although still a minority, of original seal impressions, which is also unequally distributed. This is a consequence of the historical accidents and circumstances that brought these artifacts to our knowledge. For instance, all examples from Umma and Lagaš are seal impressions, as demonstrated in Table 3.2, below. This may bring a potential imbalance to the analyses, but it can be controlled, since it can be followed through the analysis. It is, however, a means to investigate the possibility that at least two types of seals would have been distinguished through specific iconographic features or surface treatments: those

33 Here “geographic origin” means the settlement in which a seal was found or the area or site to which it can be attributed by inference. For the geographic location of Garšana, I have followed the suggestions of Heimpel (2011) and Steinkeller (2011).

intended to be used on administrative documents and those designed for different uses or that had been deprived of their official value by the addition or erasing of visual elements.<sup>34</sup>

TABLE 3.2 *Scenes from seals and impressions*

Origin	Scenes from seals	Scenes from impressions
Ur	39	6
Telloh	43	0
Susa	41	2
Umma	0	75
Lagaš	0	35
North-Bab (Babylon+Nippur+Drehem+Sippar)	21	5
South (Failaka+Larsa+Adab+Uruk+Garšana)	9	2
North-East (Subartu+Ešnunna)	5	0
Unknown	140	2

### The Procedure

After being completed, the dataset made of the textual encodings of all 425 presentation scenes was imported into the SPAD 5.5 software package.<sup>35</sup> SPAD

34 Examples of presentation scenes that have been meaningfully altered—probably to deprive the seal of its administrative function without affecting its other values—can be found in Collon (1982, no. 371 [BM 130707], no. 403 [BM 115418], no. 440 [BM 103232], and the like); exceptional examples of altered presentations that kept being used in the administrative sphere are shown in Fischer (1997, 100, Tf. 6, Abb. 1 [BM 13032A]; Tf. 7, Abb. 1 [BM 13080A]).

35 Lebart and Salem 1994. *Système Portable pour l'Analyse des Données (SPAD)* is a software for statistical analysis that has a Graphical User Interface (GUI) and includes a large series of different packages and functions. With the tools included in SPAD, one can perform different types of analyses and also manage data and whole datasets, with the aim, for example, of investigating their structure or preparing them to be processed. It is a proprietary software offering user interfaces in French. SPAD was first conceived as academic software, but after 1987 it became a commercial product developed by CISIA/DECISIA. The

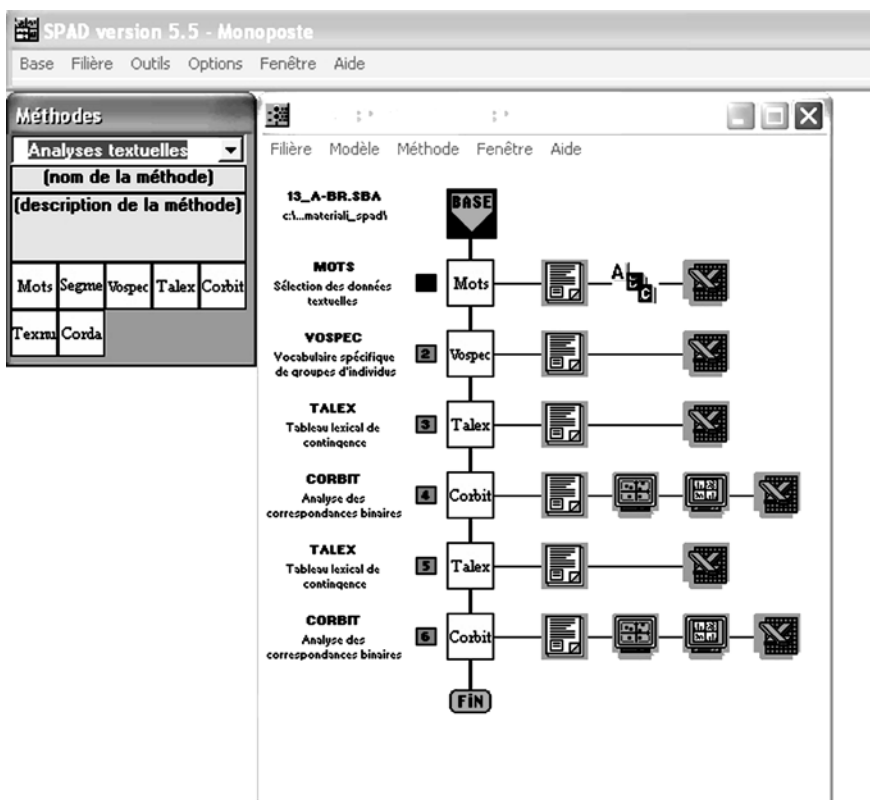


FIGURE 3.2 Procedure followed within SPAD 5.5. The boxes connected vertically represent the successive steps that have been followed, while those connected to them horizontally represent the relevant outcomes.

5.5 offers a large number of algorithms that can perform different types of statistical analyses; the ones used for this investigation belong to the category of textual analyses (*analyses textuelles*). The dataset thus needed to be specially prepared and adapted before being imported into SPAD. Since this software allows the researcher to develop each step of the analyses within a “chain” (*fil-ière*), the data necessary for each step are directly available in the format required.

The chain used to perform this analysis was structured as follows (Fig. 3.2):

Step 1: Mots. The machine acquires the dataset and analyzes its structure. The vocabulary related to the dataset is drafted, together with the charts

version 5.5, used to perform the analyses discussed here, was released in 2002. For additional information regarding a GUI, see in this volume Eraslan 303n77.

concerning the frequency of each form (i.e., the *word* of the vocabulary), their lengths, and other criteria.

Step 2: Vospec. The machine drafts the vocabulary that is specific and typical of each group of specimens (or records), i.e., the forms that are especially related to each category, a category being identifiable through its connection with a specific variable. In this investigation, the places of origin have been chosen as variables to identify categories. Through an analogous process, the machine identifies the records that can be especially associated with a variable.

Step 3: Talex. The machine outlines the contingency table that shows the type of connection between the variables and the forms of the vocabulary. This is preparatory to the binary correspondence analysis.

Step 4: Corbit. The machine performs binary correspondence analysis. The outcome is a chart with the coordinates and the rate of participation of each variable and each form to the formation of the axes. These results are also represented in the form of a two-dimensional graph showing the position of the forms and the variables in relation to the axes, which are taken into account two by two.

The process was repeated for the dataset deprived of the specimens with unknown origin.

### Analyses

The comparison between the outcomes of the binary correspondence immediately gave a clear representation of the impact caused by a meaningful presence of the records with unknown origin. As expected, the relative distances of toponyms are more easily readable along the main axes if each specimen has a

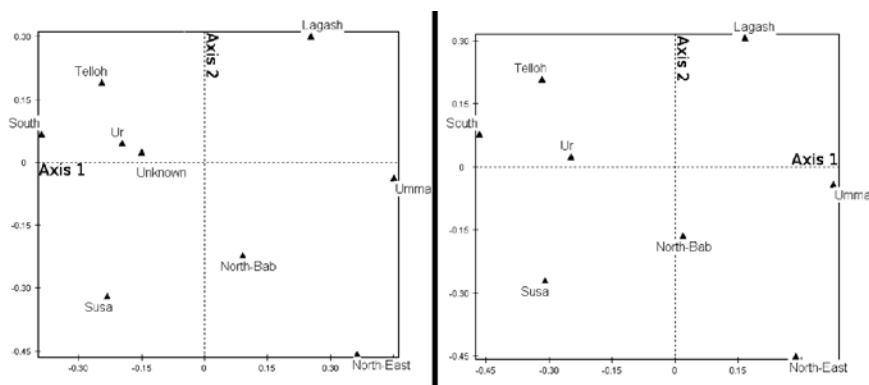


FIGURE 3.3 *Graphs showing the outcomes of the binary correspondence analyses on the toponyms, with (left) and without (right) the specimens of unknown origins*



known origin. However, it also appears that the scenes of unknown origin do not cause a great change in the opposition and similarity relations among the other toponyms, although they do influence the general balance. This could mean that the uncontrolled mix of specimens classified as having an unknown origin is so heavily internally diversified that it represents an average of all types of scenes almost homogeneously. In the results of the analysis that excluded the scenes of unknown origin, the center of the graph is nevertheless more distant from Ur, Susa, and the southern sites along the first axis, which may mean that the relevant dataset is better balanced (Fig. 3.3).

The part of the analysis that focused on the vocabulary specific to the scenes from each toponym (see above, Step 2: Vospec) is essential for understanding in greater depth the similarity and opposition relations that are expressed by these graphs. Such typical vocabulary was filled out according to the criteria of being either typically present or typically absent. A 5% threshold was chosen to define the typicality. This means that a feature (expressed through an encoded element, that is, through a “word” or “form” used to describe formally the records of the dataset) was considered positively or negatively typical of the scenes from a toponym if there was a possibility of less than 5% that its presence (positively typical) or absence (negatively typical) in a specimen related to that toponym is random.

In short, the situation emerging from the analysis is the following (Table 3.3): The specific vocabulary suggests that a proper interpretation of the oppositions of toponyms along the axes would connect them with iconographic features that probably have to do with chronological differences (second axis), on the one hand, and with the treatment of the scenes, on the other hand (first axis). The two factors, represented here by the two axes, also have a partial direct relation with each other.

To be more specific without lingering over the smallest details, it is worth taking into account the amount and type of forms that the toponyms share in their typical vocabularies. The scenes from Ur, South, and Telloh have a tendency to show earlier iconographic elements in common, which recall those present in the very early Ur III (or post-Akkadian) presentation scenes.<sup>36</sup> Comparatively frequent in such a category of presentations are the female figures; some hairstyles that tend to disappear in later years (such as the one called here “woman’s hairstyle with the hair gathered on the top back of the head”); and relatively simple thrones and divine headgear. Less present or totally absent are elements that can be connected to the presentation before a royal figure (such as the padded stool and the skull-cap), as well as personal ornaments

---

36 Ludovico 2008.

TABLE 3.3 *Features of toponyms*

Toponym	Typical features	
	Positive	Negative
Ur	presence of a goddess in the scene; presence of a crescent moon; plain robes; divine headgear with one or multiple pairs of horns; woman's hairstyle with the hair gathered on the top back of the head; simple framed throne; throne with central support and frame; simple dais under the throne; one-line legend; erased legend; legend containing a divine name	bearded character; presence of a male character; presence of characters who have material attributes; character (female) bearing bracelets or multiple necklaces; skull-cap; complex headgear with multiple pairs of horns; padded stool (royal); dais under the throne and footstool; three-lines framed legend; legend's content – type «PN1, profession, son of PN2»
North-Bab	arm depicted vertically, along the body; simple necklace; long end of necklace or robe falling along the back; little amphora; skull-cap; hairstyle with a hair gathered and going upwards; scorpion in the field; padded stool (royal); erased legend; legend's content not legible or erased; legend's content of the type i-na-ba (gift-seal)	multiple necklace; legend's content – type «PN1, profession, son of PN2»
South	presence of a woman; plain robes; headgear with one couple of horns; simple framed throne; two-lines framed legend; legend's content – type «PN1, son of PN2»	bearded character; presence of characters who have material attributes; complex headgear with multiple pairs of horns
Telloh	presence of a goddess in the scene; presence of a male character; a couple of characters hand in hand; plain robes; headgear with one pair of horns; hairstyle with double curl (female); woman's hairstyle with the hair gathered on the top back of the head; woman's hairstyle with the hair gathered in a long braid; mace in the field; simple square throne; throne	presence of a goddess; presence of a god; presence of a male character; presence of characters who have material attributes; bearded character; character having simple or double bracelets; skull-cap; complex headgear with multiple pairs of horns; hairstyle with hair gathered by the neck (male divine); little amphora; little cup; sun disc with a crescent moon in the field; padded stool (royal)

TABLE 3.3 *Features of toponyms (cont.)*

Toponym	Typical features	
	Positive	Negative
	<p>with central support and frame;  two-lines framed legend; legend's  content – type «PN1, wife of PN2»;  legend's content – type «PN1, son of PN2,  profession»; legend's content – type «PN,  profession»</p>	
North-East	<p>presence of a small goddess; presence of  characters who have material attributes;  divine headgear with multiple pairs of  horns; striped headgear; rod-and-ring;  sun disc with crescent moon in the field;  bird with long legs in the field; legend's  content not legible or erased</p>	<p>a couple of characters hand in hand</p>
Susa	<p>plain robes; fringed robes; simple man's  hairstyle; striped headgear; small flat  headgear; simple “striped” man's  headgear; woman's hairstyle with one  big curl by the neck; standard with  crescent in the field; crescent moon in  the field; bird (goose) in the field;  monkey in the field; erased legend;  simple framed throne; two-lines framed  legend; legend's content – type «PN1, son  of PN2»; legend's content – type «PN1,  servant of PN2»</p>	<p>presence of characters who have material  attributes; bracelet; simple necklace;  multiple necklaces; complex headgear  with multiple couples of horns; hairstyle  with hair gathered by the neck (male  divine); legend in 2 rows of framed lines;  bird of prey with open wings; legend's  content – type «PN1, profession, son of  PN2»</p>
Umma	<p>presence of a female character; bearded  character; presence of male deity;  presence of male figures; presence of  characters who have material attributes;  skull-cap; complex headgear with  multiple pairs of horns; simple necklace;  multiple necklaces; little cup; bracelet;  double bracelets;</p>	<p>presence of women; presence of male  figures; presence of goddesses; fringed  robe; plain robe; divine headgear with one  or multiple pairs of horns; simple man's  hairstyle; hairstyle with double curl  (female); woman's hairstyle with the hair  gathered on the top back of the head;  elements placed in the field between two</p>

Toponym	Typical features	
	Positive	Negative
	hairstyle with hair gathered by the neck (male divine); standard with bull in the field; sun disc with crescent moon in the field; padded stool (royal); simple throne with support; two-steps dais under the throne; legend smaller than the height of the field; three-lines framed legend; legend in 2 rows of framed lines; legend's content – type «PN1, profession, son of PN2»; legend's content – type «RN, royal epithets, PN, his servant»	characters; elements placed in the top part of the field; elements placed in the field in the middle; elements placed in the field, one above the other; bird of prey with open wings in the field; crescent moon in the field; scorpion in the field; goose in the field; simple square throne; simple framed throne; two-lines framed legend; legend's content – type «PN1, son of PN2, profession»
Lagaš	bird of prey with open wings; tree in the field; elements placed in the field in the middle; elements placed in the field, one above the other; small vase in the field; goose in the field; throne with multiple profiles and multiple support; throne in the form of a dragon; throne with seat-back; dais under the throne and footstool; four-lines framed legend; legend in a row of framed lines; legend's content – type «PN1, profession, son of PN2, profession»; legend's content – type «RN, royal epithets, PN1, profession, son of PN2, his servant»	simple necklace; simple man's hairstyle; plain robe; moon sickle in the field; two-lines framed legend; simple framed throne; throne with central support and frame;

of the characters, male divine figures, and (typically later) complex types of divine headgear. On the opposite side of the first axis, Lagaš and Umma have much in common, with the former seemingly more isolated, but also richer in older or peculiar iconographies, such as the tree or the dragon-throne. The toponym “Umma” seems to be rather connected with male and royal characters and with elements that probably began to appear in the scenes sometime later, such as some body ornaments, while a reduced variety of the elements in the field is also observed. This general picture is made more complicated by

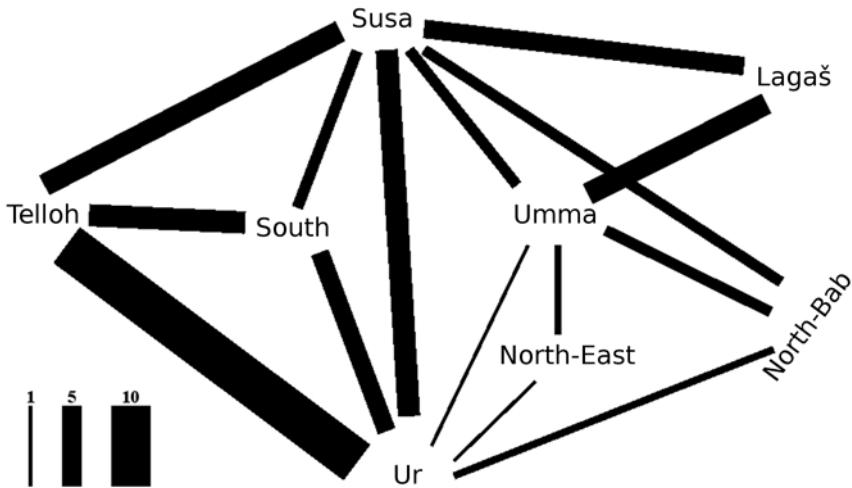


FIGURE 3.4 *Total elements (forms) of the typical vocabulary that are shared by the different toponyms. Positive and negative forms have been considered together; the thickness of each edge is directly proportional to the total number of shared forms.*

the relationship of the above-mentioned toponyms “Susa,” “North-Bab,” and “North-East,” with their specific vocabulary. Such names of origin placed in the lower region of the graph seem to share some features (including changes in the iconography and in the legend, which often have been erased) that result from the activity of secondary carving. In particular, “North-East” is related to scenes containing the later addition of a small female deity in the field, with changes in the legend, or with combinations of archaic features with secondary ones. This may explain its peripheral position in the graph. Several ancient iconographic features are typical of the Susa corpus, which has many specific forms in common with the southern sites, such as fringed or plain robes and some hairstyles, but has also signs of re-worked legends. “North-Bab,” placed in the middle-lower part of the graph, has in its typical vocabulary some features that express the presence of royal figures in the scene but also signal secondary changes (especially in the legend).

The overview of the shared vocabulary can be synthesized fairly simply (Fig. 3.4).

### *Commentary*

A feature that emerges from this investigation is the remarkable difference between the scenes known from original—i.e., ancient—seal impressions and

those recorded only from modern roll-outs of the cylinders. Furthermore, a number of cylinders were evidently re-carved during their time of use, and such secondary cuts quite probably affected the meaning of the whole representation, even when they were not particularly showy.

The two toponyms, Umma and Lagaš, that are only related to original impressions, are separated in the graph by a meaningful distance. This clearly means, if one looks at their typical vocabulary, that the specimens from Lagaš are more varied and show iconographic features typical of an earlier tradition (such as some plain robes, the small vase, or the tree in the field) and, probably, of the symbolic language of the specific site. Umma's vocabulary converges with that of Lagaš, especially in the field of the seals' legends (shape and content) and in the absence of elements (such as the crescent moon being represented alone) that began to appear in earlier phases. A distinctive feature for Lagaš is the bird of prey, usually represented with open wings in the top part of the field; this is an element that is, on the contrary, typically absent at Umma. The latter toponym is rather explicitly connected with forms that describe the presentations before a royal figure (for instance, the padded stool).

At the other end of axis 1, the toponyms share a typical vocabulary that tends to exclude royal features and incorporate female figures; typical earlier features here especially include the women's hairstyles and the plain robes. As far as the shape and content of the legends is concerned, the difference among the toponyms placed by the positive side of the same axis is evident. In Telloh, Ur, and South, they are usually shorter and with little information, and sometimes they are erased. This should have a relationship with the administrative role of the seal. As has emerged from earlier research,<sup>37</sup> the shape, structure, and dimensions of some categories of Ur III administrative documents seem to have progressively changed in tandem with the dimensions and type of the legends of the seals used on them.

Finally, North-East, North-Bab, and Susa seem in particular to be related to the negative part of axis 2, with elements that are characteristic of later phases (such as the rod-and-ring) or of re-cuts (such as the small goddess). It is probably meaningful that Susa's typical vocabulary contains forms that represent legends that are structurally similar to those of Ur, Telloh, and South.

Correspondence analysis seems to provide very clear outcomes in the investigation of the iconography of late third-millennium BCE presentation scenes in glyptic. The characterization of the traits that are typical of the different toponyms is unambiguously revealed. These results highlight some issues that are closely connected to further aspects: besides purely iconographic features, it is

37 Ludovico 2010; 2012; 2013.

worth taking into account more concrete ones, such as the specific function for which the seal was designed. Indeed, the life of a cylinder seal may have been quite short or very long, and its use may have changed. These changes were mirrored by the (primary and/or secondary) treatment of the seals' surfaces. A next step for this investigation, therefore, should be the enlargement of the dataset, or rather the creation of a completely new dataset containing only original seal impressions of known origin, and with the specification of the type of document on which they were impressed. This phase is currently under development, but it will probably be complicated because of the lack of truly shared data (in the sense of open archives) regarding original seal impressions. The latter are rather difficult to study if one has not had the chance to get into direct physical contact with the object, since they are rarely published, and, when they are, it is usually in formats that do not indicate their physical and iconographic features.

Another issue that deserves to be considered is the research on the different possible compositional structures of the presentation scenes and their main parts. Another correspondence analysis algorithm that has been tested on presentation scenes in recent years can focus on identifying structures and sub-structures of the scenes' description.<sup>38</sup> This step will be carried out again in the near future after a suitable adaptation of the dataset.

## References

- Asher-Greve, Julia M., and Willem B. Stern. 1983. "A New Analytical Method and its Applications to Cylinder Seals." *Iraq* 45: 157–162.
- Asher-Greve, Julia M., and Willem B. Stern. 1986. "Practical Advices for Collecting Data on Cylinder Seals." *Akkadica* 49: 17–19.
- Bahrani, Zainab. 2014. *The Infinite Image: Art, Time and the Aesthetic Dimension in Antiquity*. London: Reaktion Books.
- Boon, Paul, and Martine de Vries-Melein. 2009. *Mesopotamian Cylinder Seals: Description and Photographs of Five Seals from the NINO Collection*. DANS. <<https://doi.org/10.17026/dans-x2d-cexg>>.
- Boon, Paul, and Martine de Vries-Melein. 2013. "Cylinder Seals Revealed." In *Fusion of Cultures. Proceedings of the 38th Annual Conference on Computer Applications and Quantitative Methods in Archaeology, Granada, April 6-9 2010*, edited by Francisco Contreras, Mercedes Farjas, and Francisco J. Melero, 511–517. BAR-IS 2494. Oxford: Archaeopress.

---

38 Ludovico and Camiz 2015b.

- Camiz, Sergio. 2004. "On the Coding of Archaeological Data." *ACalc* 15: 201–218.
- Camiz, Sergio, and Elena Rova. 2001. "Exploratory Analyses of Structured Images: A Test on Different Coding Procedures and Analysis Methods." *ACalc* 12: 7–45.
- Cassin, Elena. 1960. "Le sceau: un fait de civilisation dans la Mésopotamie ancienne." *Annales. Economies, Sociétés, Civilisations* 4: 742–751.
- Collon, Dominique. 1982. *Catalogue of the Western Asiatic Seals in the British Museum. Cylinder Seals II. Akkadian, Post Akkadian, Ur III Periods*. London: The British Museum.
- Collon, Dominique. 1987. *First Impressions: Cylinder Seals in the Ancient Near East*. London: The British Museum.
- Collon, Dominique. 2001. "How Seals Were Worn and Carried: The Archaeological and Iconographical Evidence." In *Proceedings of the XLVe RAI: Part II: Yale University: Seals and Seal Impressions*, edited by William W. Hallo and Irene J. Winter, 15–30. Bethesda, MD: CDL Press.
- Digard, François, ed. 1975. *Répertoire analytique des cylindres orientaux publiés dans des sources bibliographiques éparses (sur ordinateur)*. Paris: Centre National de la Recherche Scientifique.
- Fischer, Claudia. 1997. "Siegelabrollungen im British Museum auf Ur III-zeitlichen Texten aus der Provinz Lagaš." *BaM* 28: 97–183.
- Gardin, Jean-Claude. 1955. "Problèmes de la documentation." *Diogène* 11: 107–124.
- Gardin, Jean-Claude. 1958. "Four Codes for the Description of Artifacts: An Essay in Archeological Technique and Theory." *AmA* 60 (2): 335–357.
- Gardin, Jean-Claude. 1966. "Éléments d'un modèle pour la description des lexiques documentaires." *Bulletin des bibliothèques de France* 5: 171–182.
- Gardin, Jean-Claude. 1967. "Methods for the Descriptive Analysis of Archeological Material." *AmAnt* 32: 13–30.
- Gardin, Jean-Claude. 1997. "Le questionnement logiciste et les conflits d'interprétation." *Enquête, anthropologie, histoire, sociologie* 5: 35–54.
- Gardin, Jean-Claude. 2002. "Les modèles logico-discursif en Archéologie." *ACalc* 13: 19–30.
- Goody, Jack. 1986. *The Logic of Writing and the Organization of Society*. Cambridge: Cambridge University Press.
- Haussperger, Martha. 1991. *Die Einführungsszene. Entwicklung eines mesopotamischen Motivs von der altakkadischen bis zum Ende der altbabylonischen Zeit*. Münchener Vorderasiatische Studien 9. München-Wien: Profil.
- Heimpel, Wolfgang. 2011. "On the Location of the Forests of Garšana." In *Garšana Studies*, edited by David I. Owen, 153–159. CUSAS 6. Bethesda, MD: CDL Press.
- Klengel-Brandt, Evelyn, and Joachim Marzahn. 1997. "Ein Hortfund mit Kreuzen aus Assur." *BaM* 28: 209–238.
- Kohonen, Teuvo. 1997. *Self-Organizing Maps*. New York: Springer.



- Lebart, Ludovic, and André Salem. 1988. *Analyse statistique des données textuelles*. Paris: Dunod.
- Lebart, Ludovic, and André Salem. 1994. *Statistique textuelle*. Paris: Dunod.
- Ludovico, Alessandro di. 2005. "Scene-in-frammenti: una proposta di analisi delle 'scene di presentazione' dei sigilli a cilindro mesopotamici orientata all'elaborazione statistica ed informatica dei dati." In *Studi in onore di Paolo Matthiae presentati in occasione del suo sessantacinquesimo compleanno*, CMAO X – Special Issue, edited by Alessandro di Ludovico and Davide Nadali, 57–95. Rome: Sapienza University Press.
- Ludovico, Alessandro di. 2008. "Between Akkad and Ur III: Observations on a 'Short Century' from the Point of View of Glyptic." In *Proceedings of the 4th International Congress on the Archaeology of the Ancient Near East 29 March – 3 April 2004, Freie Universität Berlin*, Vol. 1, edited by Hartmut Kühne, Rainer M. Czichon, and Florian J. Kreppner, 321–341. Wiesbaden: Harrassowitz.
- Ludovico, Alessandro di. 2010. "La glittica della fine del Terzo Millennio e il sentimento di immortalità del potere in Mesopotamia." In *Quale Oriente? Omaggio a un Maestro. Studi di Arte e di Archeologia del Vicino Oriente in memoria di A. Moortgat a trenta anni dalla sua morte*, edited by Rita Dolce, 241–261. Palermo: Flaccovio.
- Ludovico, Alessandro di. 2011. "Experimental Approaches to Glyptic Art Using Artificial Neural Networks: An Investigation into the Ur III Iconological Context." In *On the Road to Reconstructing the Past. Proceedings of the 36th International Conference on Computer Applications and Quantitative Methods in Archaeology (CAA), Budapest, April 2–6, 2008*, edited by Erzsébet Jerem, Redő Ferenc, and Vajk Szeverényi, 135–146. Budapest: Archaeolingua.
- Ludovico, Alessandro di. 2012. "The Uses of the Cylinder Seal as Clues of Mental Structuring Processes inside Ur III State Machinery." In *Organization, Representation, and Symbols of Power in the Ancient Near East. Proceedings of the 54th Rencontre Assyriologique Internationale at Würzburg 20–25 July 2008*, edited by Gernot Wilhelm, 275–289. Winona Lake, IN: Eisenbrauns.
- Ludovico, Alessandro di. 2013. "Symbols and Bureaucratic Performances in Ur III Administrative Sphere: An Interpretation through Data Mining." In *From the 21st Century B.C. to the 21st Century A.D. Proceedings of the International Conference on Sumerian Studies Held in Madrid, 22–24 July 2010*, edited by Steven J. Garfinkle and Manuel Molina, 125–151. Winona Lake, IN: Eisenbrauns.
- Ludovico, Alessandro di. 2016. "L'uso dei metodi quantitativi nell'indagine sui linguaggi figurativi del Vicino Oriente preclassico: Una disamina globale." In *(Digital) Humanities: nuovi strumenti per vecchi problemi (Sette casi di studio)*, edited by Simone Celani, 114–124. Status Quaestionis 10. Rome: Sapienza University Press.
- Ludovico, Alessandro di, and Sergio Camiz. 2014. "A Quantitative Approach to Ur III Mesopotamian Figurative Languages: Reflections, Results, and New Proposals." *ACalc* 25: 7–32.

- Ludovico, Alessandro di, and Sergio Camiz. 2015a. "Art History of the Ancient Near East and Mathematical Models: An Overview." In *21st century Archaeology: Concepts, Methods and Tools. Proceedings of the 42nd Annual Conference on Computer Applications and Quantitative Methods in Archaeology, Paris, April 22nd – 25th 2014*, edited by François Giligny, François Djindjian, Laurent Costa, Paola Moscati, and Sandrine Robert, 29–34. Oxford: Archaeopress.
- Ludovico, Alessandro di, and Sergio Camiz. 2015b. "Ancient Mesopotamian Glyptic Products, Statistics and Data Mining: A Research Proposal." In *21st century Archaeology. Concepts, Methods and Tools. Proceedings of the 42nd Annual Conference on Computer Applications and Quantitative Methods in Archaeology, Paris, April 22nd – 25th 2014*, edited by François Giligny, François Djindjian, Laurent Costa, Paola Moscati, and Sandrine Robert, 489–496. Oxford: Archaeopress.
- Ludovico, Alessandro di, Sergio Camiz, and Giovanni Pieri. 2013. "Comparative Use of Mathematical Models in an Investigation on Mesopotamian Cylinder Seals." In *Fusion of Cultures. Proceedings of the 38th Annual Conference on Computer Applications and Quantitative Methods in Archaeology, Granada, April 6-9 2010*, edited by Francisco Contreras, Mercedes Farjas, and Francisco J. Melero, 495–498. BAR-IS 2494. Oxford: Archaeopress.
- Ludovico, Alessandro di, and Giovanni Pieri. 2011a. "How to Facilitate Interpretation of Natural Computation Results by Converting Binary Codes of Images back to Images." *CuT International Journal of Artificial Intelligent Systems and Machine Learning* 3 (7): 437–446.
- Ludovico, Alessandro di, and Giovanni Pieri. 2011b. "Artificial Neural Networks and Ancient Artifacts: Justifications for a Multifunction Integrated Approach Using PST and Auto-CM Models." *ACalc* 22: 99–128.
- Ludovico, Alessandro di, and Marco Ramazzotti. 2008. "Reconstructing Lexicography in Glyptic Art: Structural Relations between the Akkadian Age and the Ur III Period." In *Proceedings of the 51st Rencontre Assyriologique Internationale, Held at the Oriental Institute of the University of Chicago, July 18–22 2005*, edited by Robert D. Biggs, Jennie Myers, and Martha T. Roth, 263–280. SAOC 62. Chicago: The Oriental Institute of the University of Chicago.
- Moscati, Paola. 2013. "Jean-Claude Gardin (Parigi 1925–2013). Dalla meccanografia all'informatica archeologica." *ACalc* 24: 7–24.
- Nam, Roger S. 2008. "A Different Kind of Impression: The Decorative Aspects of Cylinder Seals in Ugarit." *UF* 40: 523–531.
- Porada, Edith. 1993. "Why Cylinder Seals? Engraved Cylindrical Seal Stones of the Ancient Near East, Fourth to First Millennium B.C." *The Art Bulletin* 75: 563–582.
- Ross, Jennifer C. 2014. "Art's Role in the Origins of Writing: The Seal Carver, the Scribe, and the Earliest Lexical Texts." In *Critical Approaches in Ancient Near Eastern Art*, edited by Brian A. Brown and Marian H. Feldman, 295–317. Berlin: De Gruyter.

- Steinkeller, Piotr. 2011. "On the Location of the Town of GARšana and Related Matters." In *Garšana Studies*, edited by David I. Owen, 373–390. CUSAS 6. Bethesda, MD: CDL Press.
- Winter, Irene J. 2001. "Introduction: Glyptic, History, and Historiography." In *Proceedings of the XLVI RAI: Part II: Yale University: Seals and Seal Impressions*, edited by William W. Hallo, and Irene J. Winter, 1–14. Bethesda, MD: CDL Press.

## A Quantitative Method for the Creation of Typologies for Qualitatively Described Objects

*Shannon Martino and Matthew Martino*

In our natural attempts to determine patterns and similarity, certain attributes come to be considered more important than others. We attempt to find patterns even when they may not exist. This is well illustrated by the reaction to a famous question posed by Lewis Carroll's Mad Hatter in *Alice's Adventures in Wonderland*: "Why is a raven like a writing desk?" In the preface to the 1896 edition of the book Carroll says:

Enquiries have been so often addressed to me, as to whether any answer to the Hatter's riddle can be imagined, that I may as well put on record here what seems to be a fairly appropriate answer, viz: "Because it can produce few notes, tho [sic] they are very flat; and it is never put with the wrong end in front!" This, however, is merely an afterthought; the Riddle, as originally invented, had no answer at all.

While Carroll eventually gave into popular demand and provided an answer to this riddle, the continued fascination of his readers with this one particular phrase speaks to the human urge to find similarities between any two given objects. It is this urge, along with the oft-noted urge to detect differences, that makes classification such a natural part of the human character and one so prone to subjective conclusions.

Typologies are used by archaeologists to easily identify objects that are similar to each other by classifying similar artifacts into groups so that one can speak of a "type" rather than an individual object. And, initially, differences between these types were seen to be indications of cultural difference. This practice of classification, a key concept in Western scholarly thought, which is its own academic field,<sup>1</sup> is the underpinning of much archaeological dating.<sup>2</sup>

---

1 Gordon 1999.

2 Roy and Reason 1979.

Yet subjectivity in typologies is unavoidable given the essentializing nature of typologies as well as their inherently sociopolitical nature.<sup>3</sup> As Sorin Hermon and Franco Niccolucci have stated, “One of the major ‘weaknesses’ of typological research is the nature itself of types – being defined partly intuitively and partly rationally, [they are] partly essential and partly instrumental, most typological lists are polythetic, so there are no fixed criteria of ‘typehood.’”<sup>4</sup> Historically, the types defined in archaeological typologies have been viewed as distinct entities that have meaning within the culture to which they belong, and, in the first half of the 1900s, they were thought to be the most objective way to study ancient peoples.<sup>5</sup> By the late 1960s, however, it was recognized that archaeological types are defined by the archaeologist in ways that may not reflect the types defined by the prehistoric peoples.<sup>6</sup>

Some, such as William and Ernest Adams, have entirely dismissed the necessity of complete objectivity in the creation of typologies and even the discussion of objectivity when it comes to typologies, saying “useful typologies require intersubjective agreement,” i.e., a relationship between the work of all scholars on the subject.<sup>7</sup> They further note that the “development of type concepts” is contingent on the collection of objects being studied.<sup>8</sup> Moreover, they suggest that while initial types are often created through an intuitive recognition of similarity, later types are defined by difference from the very beginning of their recognition.<sup>9</sup>

When scholars look to create a typology, however, they try to distance themselves from assumptions of similarity and difference. Those assumptions are unavoidably colored by the invaluable knowledge of the times in which the object was created, as well as by previous research on the subject—but also by one’s possibly flawed understanding of which particular aspects of the object indicate human agency, and which were necessitated by the functionality and material composition of the object. One must ask oneself not what a defining attribute of that object is, but what it could be, and then must be brutally honest about what one sees. If we are not careful, what might be viewed a significant factor to us but not to the producer of an object can be heavily influential

---

3 Gnecco and Langebaek 2014, v.

4 Hermon and Niccolucci 2002, 217.

5 Hermon and Niccolucci 2002, 218.

6 Tixier 1967.

7 Adams and Adams 1991, 4.

8 Adams and Adams, 1991, 53.

9 Adams and Adams, 1991, 54–55.



FIGURE 4.1  
*Fliegende Blätter*, October 23,  
 1892

in the creation of a typology. In other words, the aspects of an object's production could be unnecessarily sidelined in favor of decoration, form, and presumed use. More than this, though, were typologies simply to be viewed as abstractions of data with no historical relevance to creation of the artifact itself, each independently created typology would be different. Thus each typology's utility in every case would have to be rigorously justified by the scientific question posed by the researcher.<sup>10</sup>

People's susceptibility to getting stuck seeing one thing when another is possible was satirized long ago by the rabbit/duck image, which seems to have first been published in *Fliegende Blätter*, a German humor magazine in Munich,<sup>11</sup> and was picked up that same year by *Harper's Weekly* (Fig. 4.1).<sup>12</sup>

The original text asked, "Which animals are the most similar to each other?" The answer was "rabbit and duck." But studies by the Bruggers in the 1990s found that what animal people see in the image can be influenced by the date on which they see the image. For example, children who see the image near Easter are more likely to see a rabbit.<sup>13</sup>

<sup>10</sup> Thank you to Elena Rova for presenting this insight.

<sup>11</sup> A facsimile of this image in *Fliegende Blätter* is available from the University of Heidelberg historical literature collection: <<https://doi.org/10.11588/diglit.2137#0147>> (accessed May 20, 2017).

<sup>12</sup> Brugger 1999.

<sup>13</sup> Brugger and Brugger 1993.

Though such extraneous happenings can affect what we see, the methodology for the creation of archaeological typologies is little discussed by their creators.<sup>14</sup> In this sense, it is much like the creation of sausage, for as much as one would like to ignore how the sausages are made, the methodology behind one's typology should be as clear as the linkages that eventually arise from its completion. Without a clear system of analysis, all typologies are leaps of faith and intuition. Adams and Adams even point to intuition as the way in which initial types are often created from a collection of objects.<sup>15</sup> While intuition often leads to discovery, a scientific argument builds "on such an intuition, shows how the linkage is possible, and therefore [is] arguable" rather than entirely subjective.<sup>16</sup> Because the reasoning from points A to Z is clear, the whole is clear. By presenting the reader with the details of this methodology in two separate case studies, we hope to make our own typological creation as transparent as possible. Admittedly, as with many ideas, intuition led to the project's creation.

### Case Studies Background

Using the method proposed here, the first case study presents the creation of a typology for a database of figurines, and the second case study creates a typology for a set of pottery. The first study compares Late Chalcolithic and Early Bronze Age (c. 5000–2500 BCE) clay anthropomorphic figurines from Bulgaria, Romania and Turkey by creating a FileMaker database of almost 2000 figurines.<sup>17</sup> Such a large and seemingly disparate corpus was chosen for analysis in order to elucidate some of the contemporaneity of figurine production and decoration trends in the analyzed areas. Though many scholars have previously noted the stylistic similarity of these figurines, definitive chronological relationships eluded them.<sup>18</sup> Those who noted the contemporaneity of figurines from the areas of Bulgaria, Romania, and Turkey tended to focus on earlier periods such as the Neolithic. In that period, presumably the figurines of both Greece and Turkey were similar not due to continuous interaction but, rather, to migrations of farmers bringing with them the various components of

---

<sup>14</sup> Adams and Adams, 1991, 60.

<sup>15</sup> Adams and Adams, 1991, 54.

<sup>16</sup> Buccellati 2007, 37.

<sup>17</sup> Martino 2012.

<sup>18</sup> Kökten, Özgüç, and Özgüç 1945; Alkım, Alkım, and Bilgi 1988; Thissen 1993; Bilgi 2001.

the “Neolithic package”—a collection of technologies and beliefs that was once thought to be wholeheartedly and without exception adopted along with agriculture.

Svend Hansen’s study *Bilder vom Menschen der Steinzeit* is the most comprehensive of all recent diachronic studies of figurines and includes a discussion of all the known figurine types from western Europe to the Near East, from the Paleolithic to the Early Bronze Age.<sup>19</sup> The figurine attributes listed in Hansen’s typology are numerous and comprehensive, but they are not additionally utilized in a systematic analysis. Many of them, however, were used to create the attribute list compiled for the figurines in this study.

### Methodological Background

In the interest of reducing subjectivity in their analyses, archaeologists and ancient art historians have used various methods from the statistical sciences. They have done so since the 1950s, but particularly since the 1970s, with the addition of the aid of computers.<sup>20</sup> Archaeologists, however, first approached statistical methods through the lens of factor analysis in the late 1960s.<sup>21</sup> This type of analysis, which defined objects as a descriptive list of qualities, or “factors,” allowed objects to be classified “automatically” rather than subjectively.<sup>22</sup> Statistical analyses closely followed such factor analyses and will be described in more detail below.

We might then ask, why have these approaches not been used more widely? One explanation might be found in the battle over the use of statistical techniques to determine type in archaeology. These battles marked the arrival of “New or Processual Archaeology” and its staunch belief in the objectivity of its methods. This conflict is epitomized in the heated back-and-forth regarding the use of statistical methods between the archaeologists James Ford and Albert Spaulding in the early 1950s. The debate resulted from Spaulding’s review of Ford’s typological schema and Spaulding’s promotion of statistical methods

---

19 Hansen 2007.

20 Spaulding 1953; Guralnick 1973; Bartel 1981; Hermon and Niccolucci 2002; Gansell et al. 2014, 194–205.

21 Factor analysis: a type of analysis that defines objects as a descriptive list of qualities, i.e., factors.

22 Hermon and Niccolucci 2002, 220.



for the creation of archaeological types.<sup>23</sup> While we will not put forward a new analysis of this general debate, as this has already been done elsewhere, below are a few of the most important quotes from this discussion:

Ford Response to Spaulding Article, February 1954:

Spaulding and some of his associates have been trained to prefer a stable finite world provided with tangible facts which, together with the logical truths that connect them, are all waiting to be discovered. This may make sense as a scientific ideal; I'm not too certain. However, "reality" is frequently confused with the ability of man to contact the phenomena with his five senses—a stone is real. Artifacts and observable situations are the "real" experienceable facts of archeology. Many cautious students have concentrated on these realities, thus hoping to avoid the risky business of stacking hypotheses into what may be a shaky structure...

The indiscriminate application of statistical formulas to archeological problems is not an activity of a cultural scientist.

Spaulding Response to Ford, April 1954:

Ford's objections to the ideas advanced in [my article] appear to revolve around (1) the notion that use of such techniques somehow constitutes a denial of continuous variation of culture in time and space and (2) certain implicit definitions of such terms as "artifact type" and "historical usefulness" which in effect make their use the exclusive prerogative of the archaeologist engaged in inferring chronology by ranking sites or components of sites in order of likeness as judged by relative frequency of attribute combinations. I shall attempt to show that the first objection is a gratuitous error and that the second is no more than a semantic quagmire.

Ford Response to Spaulding, April 1954:

First let me say that I am thoroughly sympathetic to all efforts toward development of more accurate methodology. But the application of statistics and other techniques to our problems, without regard for basic culture theory, cannot be regarded as an advance in technique...

Spaulding's suggestion that statistical analysis of the patterning to be found in a collection from a village site will establish pottery types useful in the study of culture history is amazingly naïve.

---

23 Ford and Steward 1954.

Spaulding Response to Ford – April 1954:

The concept of reality, which Ford persists in rejecting, does not imply tangibility, stability, finiteness, or “a world filled with packaged facts and truths that may be discovered and digested like Easter eggs hidden on a lawn,” as Ford asserts, but it does imply that the proper procedure for testing the truth is an appeal to the data of this external world.

As is clear from the quoted passages, the focus of the debate is on the real and experiential vs. the hypothetical and statistical. These are arguments that inevitably appear in relationship to the use of new technologies in archaeology. The arguments are suspicious of technology that could seemingly divorce oneself from the data to such an extent that the context of the data, and the knowledge that precedes the analysis, is superseded by the results of the analysis.

Although, to many of us, Ford’s rejection of Spaulding’s methods might seem out of place today, he is correct to remind archaeologists not to divorce their cultural knowledge from statistical analyses. It is thus unsurprising that Adams and Adams, who dismiss complete objectivity in typology creation, would also side with Ford.<sup>24</sup> While the analytical techniques of statistics can hinder us due to the assumptions they make, they also help inform us of real-world complexities. That is why it is so essential that archaeologists with knowledge of the material either learn to do the analyses themselves or work closely in developing the technology. They must also reflect on the assumptions inherent in their work. Such reflection is admittedly not always readily available, nor is it always broadcast.<sup>25</sup> That is why a volume such as this, with the intent to provide analyses and the background necessary to understand the methodologies, is so important.

As long as one remains careful about how one treats the data, statistical methods can be effective tools. Consciousness of the inherent subjectivity of one’s typology can make the analysis of objects grouped in a typology all the more meaningful. Accepting a typology blindly, as some have done, without considering other outcomes for one’s data, will not reveal the complexities of one’s dataset. This has led some to emphasize the entanglement of artifacts and Marian Feldman in particular to move away from geographically defined styles to the notion of “communities of styles.”<sup>26</sup> As she has put it, “It is not just that these objects have resisted any straightforward or singular definition of

---

<sup>24</sup> Adams and Adams 1991, 59.

<sup>25</sup> Adams and Adams 1991, 60.

<sup>26</sup> Entanglement is defined in *Archaeology* as the way in which all artifacts are interdependent and owe their existence—and, here especially, their visual appearance—to this interdependency and relatedness (Hodder 2012; Feldman 2014).

meaning; it is also my sense that this very enchantment itself exists (and existed in the past) as an integral part of [an object's] material purpose."<sup>27</sup> A typology then ought to reflect the mutable nature of an object.

While the ability of the method proposed here to capture this mutability is emphasized (see discussion section below), tried-and-true methods—particularly multivariate analysis—are heavily relied upon to provide the backbone of the code. Multivariate analysis is one method often utilized by archaeologists to create typologies, but it is not simply one tool that is easily employed in the same way by everyone regardless of the purpose. It is, rather, a general description for a type of analysis that takes many forms. One must always adapt the method to the material that one wishes to analyze. Using a multivariate analysis, archaeologists can determine the types that define a typology through the analysis of multiple variables and their interdependency. Furthermore, such statistical methods were created for quantitative analyses and have traditionally placed more value on certain attributes of the archaeological artifact. Cluster analysis,<sup>28</sup> a particular form of multivariate analysis that divides a dataset into groups based on how near the data points are to each other, gives one the ability to analyze subgroups of the dataset using an algorithm.<sup>29</sup>

Analyses of ancient figurines have often utilized multivariate analysis to elucidate connections between types of figurines. Saul Weinberg was the first scholar to use an attribute analysis in order to compare Neolithic figurines, though he did not use a computer.<sup>30</sup> In 1968, however, Peter Ucko became one of the first archaeologists and art historians to utilize a computer-assisted numerical-attribute analysis. Ucko's study examined Neolithic anthropomorphic figurines from Egypt, Anatolia, and Greece, beginning with an analysis of the numbers of regionally affiliated attributes and proceeding to cross-cultural ones.<sup>31</sup> He argued against isolating specific figurine attributes for analysis and against considering unprovenanced material.<sup>32</sup>

Brad Bartel's 1981 article introduced a multivariate-attribute-analysis approach to the study of figurines that additionally correlated and measured the

---

27 Feldman 2014, 175.

28 Cluster analysis: a type of analysis that divides a set of objects into groups (clusters) so that objects of one group are similar to each other, whereas objects within groups are dissimilar from the objects within other groups. See also in this volume, Monroe, 274–275, for cluster analysis applied to text analysis.

29 Algorithm: a step-by-step procedure for solving a problem or accomplishing some end, especially by a computer.

30 Weinberg 1951.

31 Ucko 1968, 427–444.

32 Ucko 1968, 390.

strength of relationships among attributes. The attributes he included in his factor analysis were mostly stylistic, but they also included clay and posture. Logically, attributes found in all figurines or only in one were not included in his study because neither would illuminate relations between cultures. His list of attributes and his methodology constituted a starting point for the one used in Shannon Martino's study. Unfortunately, the program Bartel used in his study was not done on a modern computer and used punch cards, which are lost to us today.

In 1995 Peter Biehl used an attribute-analysis approach in his study of Gradešnitsa figurines c. 5000–4500 BCE in northwest Bulgaria in order to determine whether designs were chosen based on their location on the body of a figure.<sup>33</sup> His interest in combining an analysis of the location of attributes with a description of those attributes informed much of the creation of Shannon Martino's database. What characterizes her study, and all these previous ones, is the large amount of data. Those of us who deal with the archaeological equivalent of "big data" are drawn to methods that naturally involve mathematical approaches necessitating computing power.

For the figurine analysis, Shannon Martino recorded the presence or absence of about 300 independent attributes that characterized the technological as well as the iconographic features of the figurines (Appendix 4.1). For example, under the heading "arm," there are 32 attributes that describe decoration, position, and modeling. Techniques of manufacture include features of surface treatment (e.g., slip, vertical burnishing, paint, or even fingerprints). By including so many attributes, Shannon Martino was able to consider a wide range of the possible traditions within figurine production and create a more nuanced typology that allowed significant clusters of features, both formal and technical, to emerge.

Rather than imposing any preconceived cluster of attributes, we developed a series of independent attributes. In this way, what might be viewed as a significant factor to us but not to the producer of the artifact would not be as heavily influential in the creation of a class.<sup>34</sup> The subjectivity of this analysis will still be the subject of debate for anyone who disagrees. For example, someone might ask what might be considered a nose, or arms, but as long as the observer is consistent, there will be consistency in the assessment of what it means to have arms. Thus the name "arms" might not always be perceived as accurate, but what is identified as "arms" will be.

---

33 Biehl 1996.

34 Class here refers to an identified typological group, though group is the term used by the program for all possible classes that are output.

## Project Methodology

Matthew Martino was responsible for writing the code of the program in C++,<sup>35</sup> and in the midst of collaboration, we discovered many difficulties in analyzing the qualitative data usually examined in regards to figurines.<sup>36</sup> The program, therefore, was written in an iterative process.<sup>37</sup> The goal was to create groups based on similarity, but determining how to gauge similarity required us to test the program several times to see what results various algorithms would produce. So a large portion of the development of the program consisted of refining and modifying the algorithms that it used to define groups. Eventually we settled on the procedure outlined below.

We used a hierarchical clustering algorithm that defined clusters (types) by considering how similar objects were to each other (in a quantitative fashion) and grouping “close” objects into the same type. The method that we used to determine how similar objects are is called metric scaling. This is done by considering each item as a point in a multidimensional space and calculating the “distance” between the points using a metric, just as one finds the distance between points on a map. As Torsten Madsen has noted, there is a problem with this method, namely that “the connection between objects and variables is broken...It is not possible to see the contribution of the individual variables to the analysis”.<sup>38</sup> This problem can, however, be overcome by an examination of the objects when the set of objects/the data is/are confined to a single item, such as figurines, pottery, or lithics. We also address this concern with the final output of the program, which makes the variables clearer by constructing a “representative” item based on the attributes of the items in a given cluster. Beside each attribute a number is given to indicate the likelihood that an object has a particular attribute in the identified group.

The creation of the typologies used in this study follows a consistent protocol. First, the two most similar objects are determined by the program based on the number of shared attributes. For the sake of the computer analysis, each qualitative attribute is given a number determining its presence or absence: 1 for present, 0 for not present and .5 for unclear. Then one sets a difference parameter, that is, a number which defines the number of possible ways one

---

35 C++ is a popular object-oriented programming language based on the older C programming language.

36 In this volume see also Bigot Juloux (162–163), who proposes a qualitative approach for text analysis.

37 For iterative process applied to text analysis, see in this volume, Monroe, 257, 266n25, 270.

38 Madsen 2007, <<http://www.archaeoinfo.dk/>> (accessed May 20, 2017), 2.

object can differ from another. This difference parameter will be set somewhat arbitrarily at first, and then the program will be run multiple times with different values of the difference parameter to see how it affects the groups that are formed. Comparing quantitative information such as diameter is done by subtracting numbers and taking the absolute value of the difference, so that numbers like 12 cm and 12.75 cm are seen as .75 cm different from each other, so  $\frac{3}{4}$  of a difference parameter unit. This can lead to significant differences depending on the units of the measurement (12 cm and 12.75 cm are .75 cm apart, but also 7.5 mm apart), so an appropriate scaling weight can be used to adjust for this. An alternative way to deal with quantitative data like this would be to calculate the mean and standard deviation and replace one's measurements with the number of standard deviations they are from the mean, but this might introduce other problems. These problems would appear particularly when the underlying measurement data is not well fit by a Gaussian distribution (a bell curve) that has a central peak, is evenly distributed about the mean, and can be quantified with just the mean and the standard deviation (related to the width of the bell curve). Therefore we would recommend using an appropriate weight chosen based on the amount of variation in the data.

The two most similar objects are determined by finding how distant each pair of objects is from one another in the multi-dimensional attribute space using a metric. In this case we used the Manhattan metric for Boolean (for example, to indicate whether an attribute is present or not present) and qualitative data, which, as mentioned above, calculates distance by adding the absolute values of the differences of each attribute, as well as the familiar Euclidean metric for quantitative data,<sup>39</sup> which takes the square root of the sum of the squares of the differences. In each case the value is scaled by the weight that one assigned the attribute. One could choose a different metric, but one advantage of a hierarchical clustering algorithm is that the choice of metric does not particularly affect the results. At most it will change the difference parameter necessary for different groups to appear. For example, imagine a very simple dataset with one qualitative attribute, color, and two quantitative, length and width. If we think the three attributes are equally important, the weights would be equal. If we had three objects, one (red, 2 cm long, 3 cm wide), another (blue, 5 cm long, 7 cm wide), and a third (red, 2 cm long, 7 cm wide), the distance between the first two would be  $1 + \sqrt{(3^2 + 4^2)} = 6$ , and the distance between the first and the third would be  $0 + \sqrt{(0^2 + 4^2)} = 4$ .

---

39 Euclidean metric: the square root of the sum of the squares of a difference, the distance formula used in geometry.

The two most similar objects form an initial core, and then one adds objects to the first two to create a type's core by adding all objects that differ from the two original objects by no more than half the difference parameter. This forms a distinct core of very similar objects for a group. Which objects are in the core of a group is tracked, so that an object that is in the core of one group is not in the core of any other group. We did this for two main reasons; one is purely practical, but the other is much more important. The purely practical reason for forcing the different cores to be distinct is that we wanted to make sure that the program did not just identify the same group over and over again. Secondly, and more importantly, this step reflects the belief that there are in fact distinct related groups in any given dataset where a bit of overlap is understandable (as will be shown in the next step), but the core defining aspects (and objects) of a group should be distinct.

The group is then filled out by adding all objects that differ from the core by no more than the total difference parameter, which can lead to overlap between different groups. After this pass, we have one more pass that adds objects that are within half of the difference parameter from any of the objects in the group. These last two passes add objects to the group as members, but not to the core, and so can be included in multiple groups. Then the group is complete, and the entire process is repeated until no two objects remain that are similar enough to create a core. Once no more objects are left to form a core, a final step is to get rid of groups that contain three or fewer objects, and then see if the objects from those cores fit into any other groups. This was done just to set a minimum size for a group, as a group that contains only two objects is not very illustrative. That said, the choice that a group needs to have more than three members is an arbitrary one that is easy to change.

In the end there remained some figurines which were too difficult to place into any single class, and thus they appeared in many groups. These are unavoidable outliers that appear clearly in the program output. This is actually an advantage of the program that is not typically found in typological analyses, which often restrict an object to its appearance in only one group.<sup>40</sup> In this algorithm the only objects required to be in just one group are those that are part of a group's core. An obvious extension of the developed program that would help, in particular, to analyze the figures that fit into more than one group, would determine how well any given object fits into any given group, much in the way that Hermon and Niccolucci used referees and "fuzzy set the-

---

40 Hermon and Niccolucci 2002, 217.



ory”—the idea that an object may be “in between” belonging and not belonging to a set—to examine the subjectivity of typologies.<sup>41</sup>

The order of appearance of a group in the output of the program indicates the cohesiveness of a group, because the most similar pairs of objects are the first to be identified. Thus every successive group is more diverse than the first, giving the program user an indication of diversity as well as allowing them to quickly examine various possible typological outcomes.

The number of groups created when following this protocol is directly dependent upon the difference parameter. Objects in a group cannot differ from at least one other object by more than the difference parameter. This does not mean that all objects have the same shared attributes, but, rather, that a cluster is formed by finding objects that are sufficiently close to the core in the space of all possible attribute values. When the difference parameter is decreased, objects will be more similar; when the difference parameter is increased, less similar items will be placed together. It is the degree of similarity that creates the groups, and, remarkably, for different values, there are distinct defined groups. This is remarkable in that one might assume that this method would result in the collection of all objects into the same group, and, for a high enough value, this would happen. Even for a difference parameter as high as 12, as was used with the figurines, however, not all fell into the same group. Another interesting consequence of looking at the groups formed with different difference parameters is that subgroups can be identified inside larger groups, while with a low difference parameter several groups are distinct; as the difference parameter is increased, objects merge to form larger clusters. Finally, for figurines, especially ones as fragmentary as we had, we had to run three separate sets in order to determine the best classes, one for the top, one for the middle, and one for the bottom sections of each figurine.

The results were designed to be output from the program using unique identifiers (ID) for each object, so that those identifiers could then be searched for in one's database to easily identify groups,<sup>42</sup> especially when the database includes images or other additional identifying information. Along with the list of objects found in each group, a list of average attributes is compiled so as to better define and refine that group. Since we do not prohibit objects from appearing in more than one group, using these average descriptors also has the advantage of helping one to identify points of comparison with other classes

---

41 Hermon and Niccolucci 2002, 225.

42 Identifiers (ID) are simply unique numbers that do not have a specific meaning. Their quality resides in their uniqueness. See also in this volume, Pagé-Perron (196, 206–208), who has developed a database in Structured Query Language (SQL).



when one object fits into two classes. In addition, these average attributes can be created from either the complete class or just the core group. These average attributes can also be used to determine which group any given object would best fit into if it did not get assigned to any group originally. This can be done by finding the distance between the representative objects and the objects of interest; the distance between the average and the object is an indicator of how well it would fit into that class. An alternative way of doing this is to calculate the distance between the object and each of the objects in the core of each group, then divide that total distance by the number of objects in the core. This method, however, is functionally equivalent to the other method.

This list of average attributes also helps to validate the clustering procedure by allowing one to compare the figurines that supposedly fit the cluster/group to the list and see how closely they fit. Additionally, the percentage of figurines in the cluster that can be associated with each attribute can be obtained, and a threshold can be set for the number of attributes an object must have in order to be included. The cluster can further be checked by the visualization of each figurine within the cluster. We were able to do this easily because FileMaker allows one to associate images with data and output them together. Such essential evaluative procedures were outlined by Mark Aldenderfer in his review of cluster analyses as used by archaeologists. Those procedures were essential to understanding the results of the analysis outlined below.<sup>43</sup>

### Smaller Figurine Case Study

For the comprehensive results of the program, one may look to Shannon Martino's dissertation,<sup>44</sup> but for a more detailed analysis of the program's application and the possibilities it affords through multiple analysis, this work must focus on a small portion of the larger dataset. Because one of the most robust local figurine studies is the work of Julia Obladen-Kauder at the site of Demircihüyük,<sup>45</sup> these figurines were chosen as a case study.

Demircihüyük is located in northwest Turkey near modern-day Eskişehir, and although it has remains from the Neolithic to the second millennium BCE, the study detailed here was focused on the Early Bronze Age figurines from the

---

43 Aldenderfer 1982.

44 Martino 2012. For another approach of cluster analysis applied to quantitative methods for inter-textual relations on cuneiform texts, see in this volume, Monroe, especially 274.

45 Obladen-Kauder 1996, 209–314.

site, which are dated c. 3000–2500 BCE based on four radiocarbon dates.<sup>46</sup> The site is about 80m in diameter and was first excavated in 1937 by Kurt Bittel. Manfred Korfmann directed excavations at the site between 1975 and 1978. During the Early Bronze Age, the site had a cemetery as well as a settlement and was radially planned with a circular layout of abutting, single-storied, trapezoidal buildings and a fortification wall with four gates.<sup>47</sup>

To determine the meaning of the Demircihüyük figurines, Obladen-Kauder analyzed their form, details, and material. She believed that the figurines were not obviously part of a cult or religion, stating that it is not compelling to conclude either that the sexual characteristics of figurines were meant to be life-giving, or that they were assigned to a goddess or ritual.<sup>48</sup> She insisted, rather, that they must somehow be associated with femininity.

Unfortunately, the contexts of the figurines are not entirely clear. Of the approximately 200 figurines found in the settlement, 116 come from secure contexts, but little is known about those contexts except that 78 came from interior courtyards, 23 come from the foremost rooms in a structure, and ten come from the back room of one building presumed to be a domestic structure.<sup>49</sup> Additional figurines were found in the associated cemetery of Demircihüyük-Sarıket, in the graves of both adults and children. Some of the graves contained up to three figurines.<sup>50</sup>

Obladen-Kauder argues that the Demircihüyük figurines are a local development relying on three observations: there are a large numbers of figurines; their stylistic development seems to take place locally without big jumps in either form or decoration; and several stylistic types from surrounding areas are combined in them.<sup>51</sup> The results of the large-scale analysis confirms her theory of the local development of the type, given that almost all of the figurines from the site fit into the same types. In addition, the subtypes within those types are almost exclusively filled with figurines from Demircihüyük, rather than from any of the more than 30 other sites examined in the study.<sup>52</sup>

When analyzing all the figurines in the corpus, the program divided the figurines from Demircihüyük into four major types. Most fit into what Shannon Martino called “Class I” and “Class XIII” in her broader analysis of figurines

46 Korfmann and Kromer 1993, 139–140.

47 Korfmann 1983, 216–217, 242.

48 Obladen-Kauder 1996, 257–258.

49 Obladen-Kauder, 1996, 273.

50 Seeher 1992: fig. 7.3–4; Aydingün 1999.

51 Obladen-Kauder 1996, 279.

52 Martino 2012.

from the region. It is interesting to note that these figurines were distinguished primarily by posture and the elaborateness of headdress. Class I will be the focus of the analysis in this article. This is by far the largest of the classes, and it is also the class about which more has been written than had been written about almost any other, particularly those that are considered of the *Violin Idol* type.<sup>53</sup> Figurines in this class are characterized by a circular head, stump arms, and either a semicircular bottom or articulated legs. The faces of each figurine are distinguished by a nose shaped through pinching or modeling. Slightly over 70% of the figurines from Demircihüyük are made from a fine clay, and only 10% are unfired.<sup>54</sup> Fifteen of the heads assigned to Class I come from good contexts in layer H of Demircihüyük, placing them in the Early Bronze I phase of the site.<sup>55</sup> They are also concentrated in layers K<sub>1</sub> and K<sub>2</sub> and L and M, which gives a rough end-of-use date of the beginning of Early Bronze II.<sup>56</sup> The attributes held most in common by figurines in this class are as follows:

- Arms—Horizontal stumps
- Back—Flat
- Back—Incision on back
- Back—X on back
- Breasts and chest—Intentionally absent
- Breasts and chest—X crossing torso
- Decorations—White paste
- Ears—Intentionally absent
- Eyes—Almond shaped
- Eyes—Horizontal slit
- Eyes—Dash eyebrows
- Eyes—Pupils
- Exterior color—Grey
- Face—Flat face
- Genitalia—Incised pubic area trapezoidal or square
- Genitalia—Pubic area
- Genitalia—Indented pubic area
- Hands—Intentionally absent

53 Makowski 2005, 14–15; Renfrew 1969, 9.

54 Obladen-Kauder 1996, 272, 274.

55 The only in situ complete figurine of this class was found in room 5 of phase F (Obladen-Kauder 1996, 272).

56 Obladen-Kauder 1996, 272.

Head—Almost flat vertically  
 Head—Flat in back  
 Mouth—Intentionally absent  
 Nose—Intentionally absent nose  
 Neck—Inward curve indication  
 Neck—Round modeled  
 Shoulder—Sloping  
 Techniques and tools—Incision  
 Techniques and tools—Indentation  
 Torso—With curves  
 Torso—Other incised decoration  
 Techniques and tools—Incision  
 Techniques and tools—Indentation  
 Waist—Modeled waist

The following is a description of the results of the computer analysis when just the figurines from Demircihüyük were examined. Using a difference parameter of two on the figurines from the site of Demircihüyük, three bottom groups are defined. Figure 4.2 indicates the members of one of the cores for the bottoms that were defined.

When the difference parameter is increased to four, middle groups begin to appear among the Demircihüyük assemblage (Fig. 4.3).

Two top groups also appear at four, but these groups have only two figurines in their core groups due to the increased diversity of the facial features. For example, it is not until a difference parameter of eight that heads with a head-dress appear, and then only one version of them. This analysis shows that the top sections are more diverse than the bottoms or middles in their design, especially given that the number of attributes for the top, middle, and bottom are similar at 85, 98, and 59, respectively. This means that the number of attributes vary by less than a factor of two, but the difference parameter required to have a group varies by a factor of three (See Appendix 4.2 for the attributes run for this site listed by section of the body). Moreover, the number of possible attributes for the middle is slightly larger. Lastly, the number of members in a group grows as one increases the difference parameter, allowing one to define subclasses based on classes and to see how some classes blend into each other and are more closely related than others.

One possible conclusion is that greater attention was given to, or license allowed, for the heads at Demircihüyük, especially for the facial features. The results of the analysis suggest that the primary marker of this class for the bottom section is posture, for this distinguishes the classes from each other.

Though many subclasses appear as the difference parameter is decreased, as the difference parameter is increased, these attributes quickly begin to become irrelevant, while seated figurines remain in an entirely separate class.

### Ceramics Case Study

Just as with the figurines, the ceramic data from the western Anatolian site of Demircihüyük, we thought, was worth pursuing because it had been robustly published.<sup>57</sup> In order to work with a smaller dataset, around 50 examples, the case study was limited to the forms of cups and bowls from Early Bronze Age levels H-M (Fig. 4.4).

These vessels varied in color from red to brown, both in terms of their slips as well as their internal fabric colors.<sup>58</sup> Their inclusions varied from fine to large,<sup>59</sup> from slate to fine clay with no inclusions to mention. The surface treatment varied from highly polished to unpolished. While such variety could have afforded one with 100 attributes, the 11 attributes which were chosen for analysis were based not on a new assessment of the material, but, rather, on an analysis of how Jürgen Seeher organized his cups and bowls into Forms. This was done even though archaeologists sometimes classify pottery by the components of the clay firstly and by the form second (Appendix 4.3). Based on analysis of the vessels' descriptions within the text, as well as in the catalog, Seeher seems to have created types based solely on form, to the exclusion of handle types; decoration and fabric were not considered. Therefore, in the new analysis, information about the presence of a handle was weighted less than all the other attributes. Our goal was to come up with the exact same groups, the major difficulty being that the Demircihüyük publication does not have entirely consistent descriptions of the forms, so determining the attributes of each form was difficult.

For example, the images in Figure 4.5 are to scale, and what distinguishes bowl from cup is not entirely clear, except that perhaps handles of the type depicted may preclude pieces from being cups. Therefore, form 6 can have such a horizontal handle, or it may not. Other problems include interpretations

---

57 Seeher 1988.

58 Slip: a mixture of clay and water poured or brushed onto the surface of pottery before firing, used to create a smooth coating of possibly another color than that of the clay that comprises the body of the vessel.

59 Inclusions: usually pieces of rock found deliberately or unintentionally mixed into the clay of a vessel.

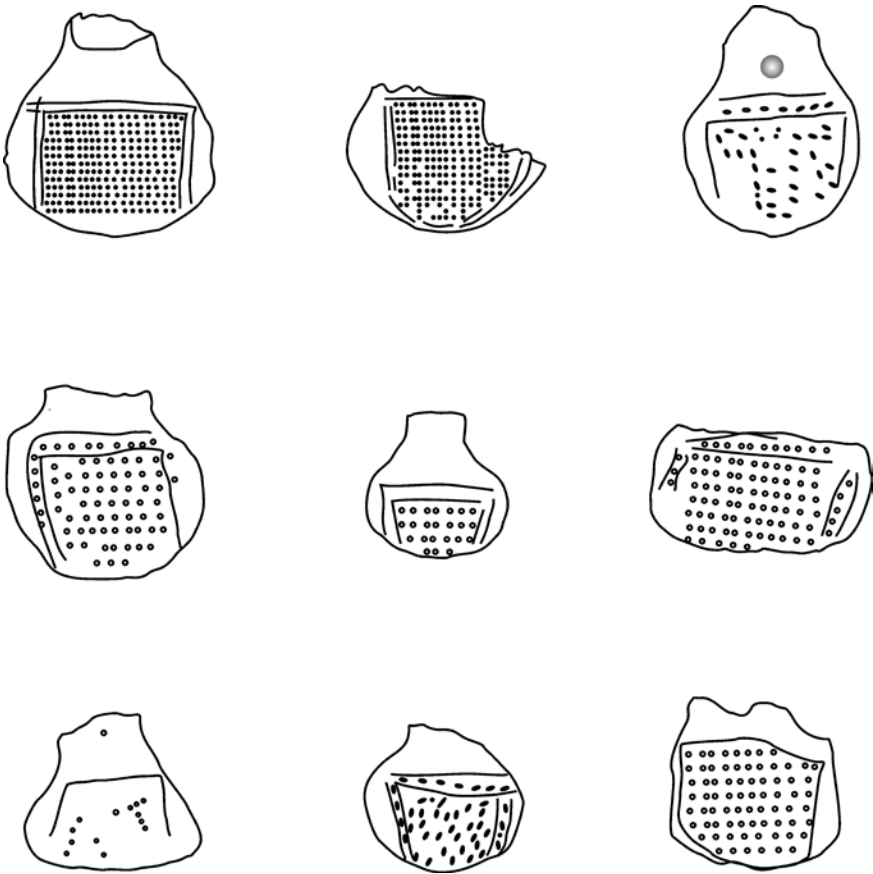


FIGURE 4.2 Example of core defined for Demircihüyük bottom section (images adapted from Obladen-Kauder 1996: pls. 114.2–3, 114.5, 114.7, 115.3–5, 115.7–8)

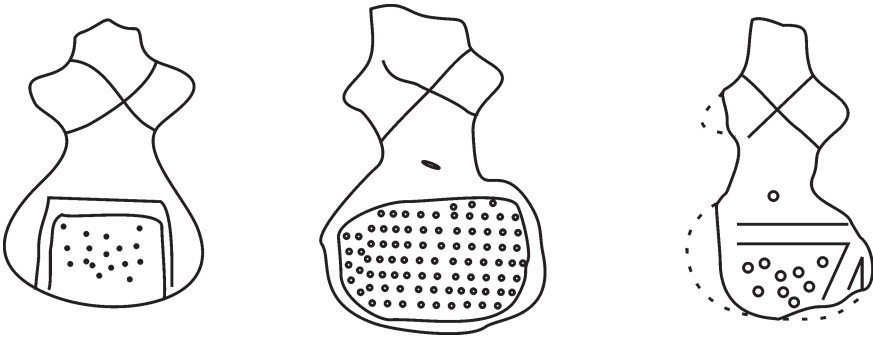


FIGURE 4.3 Example of core for Demircihüyük middle section (images adapted from Obladen-Kauder 1996: pls. 115.10, 116.3–2)



Form 3



Form 4



Form 5



Form 6



Form 7

FIGURE 4.4

*All forms captured by analysis (images adapted from Seeher 1988)*

of the rim angle, which were difficult to make based only on drawings in a book. The rim angles might be very slight, and it is not clear in those instances what designation the author might have chosen to give the vessel, as the designation is not spelled out in every vessel description. Given that, all the groups were fairly consistent in the output.

At a difference parameter of four, the program output 11 valid groups. Six groups were eliminated for being too small, so the group numbers go from 0 to 16. Table 4.1 represents groups that had some of the least correlation with Seeher's forms.

The first group is composed of vessels of form 5–7, with the average description being an everted rim, a height-to-diameter ratio of .415, a height of 5.8 cm, a diameter of 13.5 cm, with roughly 75% having a handle. This is consistent with the fact that, for Seeher, all three of these forms could be considered cups

TABLE 4.1    *Groups unlike Seeher's forms.*

First group	Sixth group
§Group number: 0 with 4 members	§Group number: 5 with 4 members
§Core: pl.2.5 pl.2.12 pl.2.14 pl.18.11	§Core: pl.18.7 pl.28.5
§Members:	§Members: pl.1.10 pl.18.8
§Average member:	§Average member:
§Height: 5.8125	§Height: 6
§Diameter: 13.5	§Diameter: 12.75
§Handle: 0.75	§Handle: 0.5
§HeightDiameterRatio: 0.430556	§HeightDiameterRatio: 0.470588
§WidthCenterBody: 12	§WidthCenterBody: 12
§Vertical Rim: 0	§Vertical Rim: 1
§Inverted Rim: 0	§Inverted Rim: 0
§Everted Rim: 1	§Everted Rim: 0
§Angled Rim: 0	§Angled Rim: 0
§S-Shaped Profile: 0	§S-Shaped Profile: 0
§Semicircular Handle: 0	§Semicircular Handle: 0
§Spout: 0	§Spout: 0

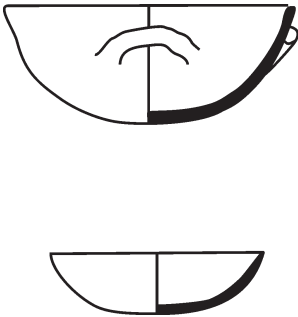


FIGURE 4.5  
*Variations of form 6 (images adapted from Seeher 1988)*

or bowls and for the most part had everted rims. The sixth group that appears is composed of vessels in cup form 3, 4, and 6, with an average height of 6 cm, diameter of 12.75 cm, 50% handles, height-to-diameter ratio of .47, and a vertical rim (Fig. 4.5).

All of these forms have a diameter consistent with one another, which may explain their overlap here.



TABLE 4.2 *Groups that best fit with Seeher's forms.*

Fifth group	Eleventh group
§Group number: 4, with 4 members	§Group number: 10 with 4 members
§Core: pl.19.7 pl.19.10	§Core: pl.2.10 pl.19.8
§Members: pl.2.6 pl.19.6	§Members: pl.2.20 pl.19.21
§Average member:	§Average member:
§Height: 6	§Height: 6
§Diameter: 15	§Diameter: 17.25
§Handle: o	§Handle: o
§HeightDiameterRatio: 0.4	§HeightDiameterRatio: 0.347826
§WidthCenterBody: 12.75	§WidthCenterBody: 14.625
§Vertical Rim: 1	§Vertical Rim: o
§Inverted Rim: o	§Inverted Rim: o
§Everted Rim: o	§Everted Rim: 1
§Angled Rim: o	§Angled Rim: o
§S-Shaped Profile: o	§S-Shaped Profile: o
§Semicircular Handle: o	§Semicircular Handle: o
§Spout: o	§Spout: o

Table 4.2 represents the groups that fit best with Seeher's forms. Of the 12 defined groups, five of them had at most only one extra member, i.e., all but one vessel belonged to the same vessel form. The fifth and fifteenth groups were composed of vessels of cup form 6 only. For the fifth group, vessels had an average height of 6 cm, diameter of 15 cm, a height-to-diameter ratio of .4, and a vertical rim and no handles. This too is consistent with Seeher's description of form 6. For the fifteenth group, the average height was 7.1 cm, diameter 18 cm, height-to-diameter ratio of .40, and an everted rim and no handles. The eleventh group was composed of vessels mostly of cup form 7, with one of form 6 with an average height of 6 cm, diameter of 17.25 cm, height-to-diameter ratio of 0.35, and an everted rim. Therefore, this group, too, is consistent with Seeher's analysis of form 7, but with a slightly larger average diameter than the second group.

Despite the fact that there is some overlap in the forms, the average descriptors of the forms as indicated by the program show that they are consistent with the descriptions of the forms given by Seeher. It seems that where the results do not exactly match Seeher's forms, something has been missed from

Seeher's methodology to define forms, or too much or too little weight has been given to a particular attribute. And, of course, as mentioned above, the rim angle is difficult to determine from a book. A last note, none of the "bowls" fit into a group until a difference parameter of 8, which seems to confirm the validity of Seeher's typological systems that placed no more than four bowls into any one type and made them a more diverse lot. Ideally this study could also be compared to an extensive reanalysis of the pottery from Demircihüyük.

### Why This and Not Another?

This program uses an agglomerative hierarchical clustering algorithm. It is agglomerative because all objects start in their own cluster, then get grouped together as the program progresses. Another way that this could be done would have been to put all objects into the same cluster to start with, then make cuts and divide it into smaller and smaller clusters. In some sense, when the program is run with different difference parameters, one gets some of the aspects of both methods; with high difference parameters one gets a few, large clusters, and with low difference parameters, one gets many, small clusters. This is also the sense of the word "hierarchical" here, as with varied difference parameters, one gets a sense of how the different clusters are related to each other. Some clusters are subclusters of other clusters, as is apparent when the difference parameter is varied and the results analyzed.

There are several advantages to the algorithm developed here. One is that no assumptions need to be made about the distributions of the underlying dataset. Using probabilistic methods, one must make assumptions (priors) about the distributions of the different attributes amongst all the possible objects; for example, a common distribution is the familiar Gaussian distribution. This is a reasonable assumption, as many things tend to follow a Gaussian distribution; for example cooking pots in a kitchen have an average size and tend to vary consistently around the average size, with some above and some below. If we considered all the plates in a house, though, the distribution would be multimodal, with several peaks in frequency, for dessert plates, dinner plates, and serving plates, each with its own corresponding bell curve. Different priors can be considered and optimized for the creation of groups, but no such assumptions need to be made to use this method. The most similar concern this method introduces is in the choice of weights and attributes, but these assumptions are explicit in the program.

As mentioned above, another advantage of this hierarchical method is that one can see how clusters are related to each other by looking at how they

change when the difference parameter is changed. Clusters formed with a higher difference parameter tend to be more heterogeneous, and they fragment as the difference parameter is decreased. For particularly diverse datasets, those with many unique or uncommon attributes, the difference parameter may need to be quite high before all the objects end up in the same cluster, but this natural breakdown of cluster and subcluster—defined by lower effective difference parameters—is one advantage of this algorithm.

One more advantage of this method is that it is similar to the process that would be followed by a person creating a typology, just significantly faster (as an example, with a dataset of 2000 objects and 285 attributes, runtimes were in the several-minutes range). Some multivariate analyses can be very opaque, producing an output that does not explain how that output was achieved and giving no direct understanding of what the output means. One example is the work done by Bartel on figurines (1981) that made little mention of how they program he developed actually worked. With the algorithm explained here, one can easily output the cores of each cluster so that they can see the set of objects that formed the basis of the cluster. In addition, because one can generate representative objects for each cluster, one can determine, with some certainty, why each object was placed in each cluster. The method is objective in that it depends only on the attributes that have been chosen and the weights that have been assigned. Varying the difference parameter and looking at the resulting subclusters makes it more evident how each cluster was formed and where it stands in relation to its subclusters and superclusters; the latter is defined by the largest effective difference parameter. This is information that would be explicit to the scholar creating a traditional typology, though perhaps not explicitly formulated in the publication of the data, and it is nearly as explicit in this method.

Although it is an advantage to have complete personal control of these aspects of the process, a major disadvantage relative to a traditional typology is that while the program creates clusters and representative objects, and the core of each cluster can be output, the actual importance of each attribute to the cluster can be difficult to determine. A frequency analysis of the attributes of objects in a cluster clarifies how many objects in a given cluster have an attribute, for example, but it cannot explain which attribute(s) were the necessary components of the cluster. One can infer this by looking at subclusters and superclusters, as well as their representative objects, but, in a handmade typology, one would have a complete understanding of why each type is the way it is.

The following is a pseudocode snippet of the program advocated here:

best = threshold; //<sup>60</sup> The threshold is the difference parameter for this run

for(i=0;i<number of items;i++) //This loop runs through all of the items in the dataset

for(j=i+1;j<number of items;j++)

if (items i and j are not in a core of a group)

and (distance between

item i and item j)<best then {

best = distance between item i and item j;

store i and j; };

//Now i and j are the two closest objects that are not already in the core of a group

set core of new group equal to items i and j;

for(k=0;k<number of items;k++)

if (distance between item k and items i and

j<threshold)

add k to core of group;

//Then the core of the group is complete, now add items that can be in multiple groups

for(k=0;k<number of items;k++)

if(distance between k and any member of core

<threshold) AND (k is not in the core of another group)

add k to group;

// Finally, one more pass of adding items to the group; these can be similar to any member of the group, but they must be more similar than in the previous step (the difference threshold is cut in half)

for(k=0;k<number of items;k++)

if(distance between k and any member of

group<threshold/2) AND (k is not

in the core of another group)

add k to group;

//Now the group is complete

---

60 “//” indicates a descriptive comment and is not in the program itself.

## Discussion

The groups defined by the computer analysis are cohesive, but, as mentioned earlier, there are outlier objects that tend to appear in all the groups, due to the simplicity of their form or their uniqueness (often fewer than three examples in the database). This situation is allowed in the analysis so as not to exclude any object from a group simply because it has already appeared in another group. This was made particularly clear in the large-scale analysis of the figurines. Such objects have to be dealt with on a case- by-case basis, and the results of the analysis must be examined for clarity.

Some objects can have too few attributes to be easily found by the program, as extremely incomplete fragments will be distant from all other objects. Really simple objects, even when complete, tend to be less distant from other objects when the number of attributes is large. This is because when the number of attributes is very large, most objects will not have most of the possible attributes, so simple objects tend to appear in many groups. These, therefore, ought not to be recognized as part of any group except the one in which they are the predominant type.

If one does not collect consistent data, then the groups produced by the program will not be consistent. That is to say, if one's definition of "arm" varies from object to object, then the results from the program will not be very useful.

Some missing information is unavoidable when working with fragmented artifacts, but a lot of missing information means that one will have to do more interpretation of the results. Others have addressed this problem of missing information by limiting their variables.<sup>61</sup> While this is a perfectly acceptable solution, it does risk not recognizing significant distinctions between objects. One way in which we tried to overcome this problem was by indicating when it was unclear whether an attribute was present or absent, and also by indicating positively when an attribute was not present. As stated above, we also analyzed figurines in groups that accommodated the usual parts in which the fragments are found, that is, head, torso, and lower body. Obviously, for different analyses different accommodations will have to be made.

The types defined may not be the ones expected. This is when one finds out if one "told" the program what was important, or if one left something out, or if one told the program information that was not important, but that information defined groups anyway. A program's ability to be flexible and to be run quickly is thus essential to any typological analysis.

---

61 Gansell et al. 2014, 198.

To use the program, all one needs is a set of attributes, a database of their objects (that can be exported to csv format),<sup>62</sup> and a set of weights that describe how important each of the attributes will be in the formation of groups. One must specify along with the weights whether the attributes are Boolean, quantitative or qualitative, and one needs to specify an initial difference parameter. All of these options are easy to change, and the program runs in minutes for datasets of several thousand objects, so there is little risk in choosing somewhat arbitrarily and then refining one's parameters. Such options need to exist because although this is a tool meant to remove subjectivity and ease the understanding of large amounts of data, that data is best and properly understood by those who study it. You as a scholar understand the significance of certain aspects of an object and the unavoidability or universality of others, and you are the one to gather the data and therefore understand the meaning of your descriptive terms. It is the responsibility of all scholars to gather data well, both comprehensively and responsibly.

#### Appendix 4.1. List of Figurine Attributes

Arms	Down at sides
Horizontal stumps	Raised straight above head
Rounded stumps	Bends elbow and reach down out to sides
Downward pointed stumps	Bends elbow and reach up out to sides
Upward pointed stumps	Folded
Freely modeled	Crossed
Applique arms	Curves/Bow down to torso
Relief modeled	Reach forward
Created by incision	Intentionally absent
On chest below breast	Holes at end
On chest beside breast	Pierced through upper arm
In back and front center	Indentation on back of elbows
To the side of abdomen	Incised decoration
Between breasts and touching	Indentation
On abdomen	
Meet at center of torso	Aprons
One up, one down	Incised
On thighs	Indented

62 csv stands for Comma Separated Values. It is a format that most programs can interact with given the simplicity of its formatting and coding.

Back	Chin
Flat	Modeled chin
Rounded	
Bump on back	Ceramic matrix
Incised spine	Coarse
Modeled spine	Fine
Spine indicated	Mica in surface
Cavity/Hood	Visible organic temper
X on back	
Indentations	Digits
Incision across top	One incision for toes
Indentations across top	Three incisions for toes
Incision on back	Four incisions for toes
Ribs	Five incisions for toes
	Modeled toes
Breasts and Chest	Two incisions for fingers
Round breasts	Three incisions for fingers
Modeled breasts	Four incisions for fingers
Pendulous breasts	Five incisions for fingers
Triangular breasts	Modeled fingers
Applique breasts	Incised wedge-shaped fingers
Indented breasts	
Asymmetrical	Decorations
Unlevel breasts	Obsidian inlay
Indented nipple	Red paint
Intentionally absent	White paint
Indented decoration across top of chest	Yellow paint
Incised decoration across top of chest	Red paste
X crossing torso	White paste
Bolero vest	
	Ears
Buttocks	Incised ears
Moderate protusion	Pinched ears
Vertical line as division	Applique ears
Modeled	Modeled ears
Horizontal line below	Indented holes
Incised circles	One piercing
Indents	Two piercings
Horizontal hole in hip	Three piercings
Hollow	Four piercings

Five piercings	Comes to point at nose
Six piercings	
Seven or more piercings	Feet
Earrings	Incised decoration of footwear
Intentionally absent	Roughly foot shaped
	Flat disc shaped
Exterior Color	
Pink	Genitalia
Brown	Modeled penis
Red	Modeled female pubic area
Orange	Incised pubic triangle
Beige	Incised pubic area trapezoidal or square
Yellow	Pubic area
Grey	Indented pubic area
Speckled	Circular pubic area
Black	Vagina indent
	Vagine incised
Eyes	Hole under rear end
Socket area impressed	
Almond shaped	Head
Semicircular	Incisions to indicate hair or headdress
Diagonal slit	Hole in top of head
Horizontal slit	Hole vertical through neck
Applied or relief oval	Cylindrical
Applied circle	Roughly spherical
Circular incised	Triangular
Indented	Upside down conical head
Pierced through	Rectangular
Modeled eyes	Circular
Modeled eyebrows	Pinched stump
Dash eyebrows	Stump with rounded point
Straight eyelids	Almost flat vertically
Rounded eyelids	Hollow
Pupils	Flat in back
Intentionally absent eyes	Inclined back
	Concave head
Face	Modeled forehead
Flat face	Flat headdress
Modeled and round	Floppy headdress
Intentionally absent	Rounded headdress or back of head



Conical or slightly pointy headdress  
 Hait, hairbun, or pigtail  
 Intentionally absent

Hands  
 Both on breast  
 One on breast  
 One between breasts  
 Neither on breast  
 On belly  
 Intentionally absent  
 Made through incision

Interior color  
 Brown  
 Dark brown  
 Dark grey  
 Orange  
 Pink  
 Grey

Legs  
 Semicircular base  
 Ovoid base  
 Cubical base  
 Indicated with modeling  
 Indicated with incision  
 Circular base  
 Conical base  
 Cylindrical base  
 Concave base  
 Square base  
 End in point  
 T-shaped base  
 Incised decoration  
 Indented decoration  
 Formed separately  
 Hole vertical through center of leg  
 Separated  
 Bent at knee

Bent at hip  
 Modeled knee  
 Applique knee  
 Knee indicated with incision  
 Modeled shin  
 Ankle area protusion  
 Incision above knee  
 Intentionally absent legs

Mouth  
 Line of indentations to indicate  
 Single indentation  
 Slit  
 Semicircular mouth  
 Circular incised mouth  
 Modeled mouth  
 Indentations below  
 Dashes below  
 Intentionally absent

Nose  
 Pinched  
 Incised  
 Modeled non-beak  
 Modeled beak  
 Indented nostril hole(s)  
 Intentionally absent nose  
 Pierced nostril hole  
 Modeled from top of head  
 Horizontal hole in nose

Neck  
 Inward curve indication  
 Elongated neck  
 Round modeled  
 Dot decoration  
 Horizontal hole through  
 Applique decoration  
 Incised decoration  
 Non-centrally placed

Navel	Torso
Vertical hole through belly	Cylindrical torso
Horizontal hole through belly	Roughly rectangular
Indentend navel	With curves
Fingernail navel	Conical
Incised navel	Incision or indentation in middle of chest
Protruding navel	Diagonal incised band with dots
Protruding circle around navel	Other incised decoration
	Applique decoration
Position	Hole through center of top
Seated	Hole through center of bottom
Standing	Belly
Leaning backward	Cavity in area
Leaning forward	One side bias
Quality of Finish	Techniques and tools
High burnish	Fingernail impressions
Burnished	Evidence of copper
Stripe burnish	Incision
Wiped	Indentation
Polished	Impression with wedge shaped tool
Evened out/smooth	Short stabs
Red slip	Drill marks
Cream slip	Hollow body
Brown slip	Incisions/Piercing after firing
Black slip	Core first construction
Beige slip	Fold construction
Slip	Finger impressions
Shoulder	Waist
Sloping	Modeled waist
Straight	Love handles
Square	Incision at waist distinct from pubic area
Indented decoration	

## Appendix 4.2. Attribute Lists by “TopMiddleBottom”

Top	Middle	Bottom
C_Modeled Chin, 1, 1	A_Applique Arms, 1, 1	AP_Incised, 1, 1
E_1xPierced, 1, 1	A_B/t breasts and touching, 1, 1	AP_Indented, 1, 1
E_2xPierced, 1, 1	A_Bends elbows & reach down out to sides, 1, 1	Bu_Hollow, 1, 1
E_3xPierced, 1, 1	A_Bends elbows & reach up out to sides, 1, 1	Bu_Horizontal Hole in hip, 1, 1
E_4xPierced, 1, 1	A_Created by Incision, 1, 1	Bu_Horizontal Line Below, 1, 1
E_5xPierced, 1, 1	A_Crossed, 1, 1	Bu_Incised Circles, 1, 1
E_6xPierced, 1, 1	A_Curve/bow down to torso, 1, 1	Bu_Indents, 1, 1
E_7x or more Pierced, 1, 1	A_Down at Sides, 1, 1	Bu_Modeled Buttocks, 1, 1
E_Applique, 1, 1	A_Downward Pointed Stumps, 1, 1	Bu_Moderate Protrusion, 1, 1
E_Earring(s), 1, 1	A_Folded, 1, 1	Bu_Vertical Line as Division, 1, 1
E_Incised Ears, 1, 1	A_Freely Modeled, 1, 1	D_1 incision for toes, 1, 1
E_Indented holes, 1, 1	A_Holes at end, 1, 1	D_3 incisions for toes, 1, 1
E_Intentionally Absent Ears, 1, 1	A_Horizontal Stumps, 1, 1	D_4 incisions for toes, 1, 1
E_Modeled, 1, 1	A_In back and front center, 1, 1	D_5 incisions for toes, 1, 1
E_Pinched Ears, 1, 1	A_Incised decoration on arms, 1, 1	D_Modeled toes, 1, 1
Ey_Almond Shaped, 1, 1	A_Indentation on Arms, 1, 1	Ft_Flat disc-shaped, 1, 1
Ey_Applied circle, 1, 1	A_Indentation on back of elbows, 1, 1	Ft_Incised decoration or footwear, 1, 1
Ey_Applied or Relief Oval, 1, 1	A_Intentionally absent arms, 1, 1	Ft_Roughly foot-shaped, 1, 1
Ey_Circular Incised, 1, 1	A_Meet at Center of Torso, 1, 1	G_Circular Pubic Area, 1, 1
Ey_Dash Eyebrows, 1, 1	A_On abdomen, 1, 1	G_Hole Under Rearend Text, 1, 1
Ey_Diagonal Slit, 1, 1	A_On chest below breast, 1, 1	G_Incised pubic area trapezoidal or square, 1, 1
Ey_Horizontal slit, 1, 1	A_On chest beside breasts, 1, 1	G_Incised Pubic Triangle, 1, 1
Ey_Indented, 1, 1	A_On Thighs, 1, 1	G_Indented Pubic Area, 1, 1
Ey_Intentionally Absent Eyes, 1, 1	A_One up one down, 1, 1	G_Modeled female pubic area, 1, 1
Ey_Modeled Eyebrows, 1, 1	A_Pierced through upper arm, 1, 1	G_Modeled Penis
Ey_Modeled Eyes, 1, 1	A_Raised straight Above Head, 1, 1	G_Pubic Area, 1, 1

Top	Middle	Bottom
Ey_Pierced through, 1, 1	A_Reach forward, 1, 1	G_Vagina Incised, 1, 1
Ey_Pupils, 1, 1	A_Relief Modeled, 1, 1	G_Vagina Indent, 1, 1
Ey_Rounded Eyelids, 1, 1	A_Rounded Stumps, 1, 1	L_Ankle Area Protrusion, 1, 1
Ey_Semi-circular Eyes, 1, 1	A_To the side of abdomen, 1, 1	L_Applique knees, 1, 1
Ey_Socket Area impressed, 1, 1	A_Upward Pointed Stumps, 1, 1	L_Bent at hip, 1, 1
Ey_Straight Eyelids, 1, 1	B_Bump on Back, 1, 1	L_Bent at knee, 1, 1
F_Comes to point at Nose, 1, 1	B_Cavity/Hood, 1, 1	L_Circular base, 1, 1
F_Flat Face, 1, 1	B_Flat, 1, 1	L_Concave Base, 1, 1
F_Intentionally Absent Face, 1, 1	B_Incised Spine, 1, 1	L_Conical Base, 1, 1
F_Modeled in Round, 1, 1	B_Incision across top, 1, 1	L_Cubical base, 1, 1
H_Almost flat vertically, 1, 1	B_Incision on Back, 1, 1	L_Cylindrical Base, 1, 1
H_Circular, 1, 1	B_Indentations, 1, 1	L_End in a Point, 1, 1
H_Concave Head, 1, 1	B_Indentations across top, 1, 1	L_Formed Separately, 1, 1
H_Conical or Slightly Pointy, 1, 1	B_Modeled Spine, 1, 1	L_Hole vertical through center of leg, 1, 1
H_Cylindrical, 1, 1	B_Ribs, 1, 1	L_Incised decoration, 1, 1
H_Flat Headdress, 1, 1	B_Rounded, 1, 1	L_Incision above knee, 1, 1
H_Flat in Back, 1, 1	B_Spine Indicated, 1, 1	L_Indented decoration, 1, 1
H_Floppy Headdress, 1, 1	B_X On back, 1, 1	L_Indicated with incision, 1, 1
H_Hat, hair bun or pigtail, 1, 1	BC_Applique Breasts, 1, 1	L_Indicated with modeling, 1, 1
H_Hole in Top of Head, 1, 1	BC_Asymmetrical, 1, 1	L_Intentionally Absent Legs, 1, 1
H_Hole Vertical Through, 1, 1	BC_Bolero Vest, 1, 1	L_Knee indicated with incision, 1, 1
H_Hollow, 1, 1	BC_Incised Decoration Across Top of Chest, 1, 1	L_Modeled knee, 1, 1
H_Incisions to Indicate Hair or Headdress, 1, 1	BC_Indented Breasts, 1, 1	L_Modeled Shin, 1, 1
H_Inclined Back, 1, 1	BC_Indented Decoration Across Top of Chest, 1, 1	L_Ovoid Base, 1, 1
H_Intentionally Absent, 1, 1	BC_Indented Nipple, 1, 1	L_Semicircular base no legs, 1, 1
H_Modeled Forehead, 1, 1	BC_Intentionally Absent Breasts, 1, 1	L_Separated, 1, 1
H_Pinched stump, 1, 1	BC_Modeled Breasts, 1, 1	L_Square Base, 1, 1
H_Rectangular, 1, 1	BC_Pendulous Breasts, 1, 1	L_T-shaped Base, 1, 1

(cont.)

Top	Middle	Bottom
H_Roughly Spherical, 1, 1	BC_Round Breasts, 1, 1	P_Seated
H_Rounded Headdress or Back of Head, 1, 1	BC_Triangular Breasts, 1, 1	P_Standing
H_Stump With Rounded Point, 1, 1	BC_Unlevel Breasts, 1, 1	P_Leaning backward
H_Triangular, 1, 1	BC_X Crossing torso, 1, 1	P_Leaning forward
H_Upsidedown Conical Head, 1, 1	D_2 incisions for fingers, 1, 1	
M_Circular incised Mouth, 1, 1	D_3 incisions for fingers, 1, 1	
M_Dashes below, 1, 1	D_4 incisions for fingers, 1, 1	
M_Indentations below, 1, 1	D_5 incisions for fingers, 1, 1	
M_Intentionally Absent Mouth Text, 1, 1	D_Incised wedge-shaped fingers, 1, 1	
M_Line of Indentations to indicate, 1, 1	D_Modeled Fingers, 1, 1	
M_Modeled Mouth, 1, 1	HA_Both on Breasts, 1, 1	
M_Semi-circular Mouth, 1, 1	HA_Intentionally absent hands, 1, 1	
M_Single Indentation, 1, 1	HA_Made Through Incision, 1, 1	
M_Slit, 1, 1	HA_Neither on Breast, 1, 1	
N_Horizontal hole in nose, 1, 1	HA_On Belly, 1, 1	
N_Incised, 1, 1	HA_One b/t Breasts, 1, 1	
N_Indented Nostril Hole(s), 1, 1	HA_One on Breast, 1, 1	
N_Intentionally Absent Nose, 1, 1	Nv_Fingernail Navel, 1, 1	
N_Modeled Beak, 1, 1	Nv_Horizontal hole through belly, 1, 1	
N_Modeled from top of head, 1, 1	Nv_Incised navel, 1, 1	
N_Modeled Non-beak, 1, 1	Nv_Indented navel, 1, 1	
N_Pierced Nostril Hole, 1, 1	Nv_Protruding circle around navel, 1, 1	
N_Pinched, 1, 1	Nv_Protruding navel, 1, 1	
Nk_Appliqué Decoration, 1, 1	Nv_Vertical hole through belly, 1, 1	
Nk_Dot Decoration, 1, 1	Sh_Indented Decoration, 1, 1	
Nk_Elongated, 1, 1	Sh_Sloping, 1, 1	
NK_Horizontal hole through, 1, 1	Sh_Square, 1, 1	
Nk_Incised Decoration, 1, 1	Sh_Straight, 1, 1	
Nk_Inward Curve Indication, 1, 1	T_Appliqué Decoration, 1, 1	
Nk_Non_centrally placed, 1, 1	T_Belly, 1, 1	

Top	Middle	Bottom
Nk_Round Modeled, 1, 1	T_Cavity in Area, 1, 1	
	T_Conical, 1, 1	
	T_Cylindrical Torso, 1, 1	
	T_Diagonal incised band with dots, 1,	
	1	
	T_Hole Through Center of Bottom, 1, 1	
	T_Hole Through Center of Top, 1, 1	
	T_Incision or indentation middle of	
	chest, 1, 1	
	T_One side bias, 1, 1	
	T_Other incised decoration, 1, 1	
	T_Roughly rectangular, 1, 1	
	T_With curves, 1, 1	
	W_Incision at Waist Distinct from	
	Pubic Area, 1, 1	
	W_Love Handles, 1, 1	
	W_Modeled Waist, 1, 1	

Appendix 4.3. List of Ceramic Attributes

Ceramic Attributes	Weights
Height	2
Diameter	2
Handle	1
HeightDiameterRatio	2
WidthCenterBody	2
Vertical Rim	2
Inverted Rim	2
Everted Rim	2
Angled Rim	2
S-Shaped Profile	2
Semicircular Handle	1
Spout	2

## References

- Adams, William Y., and Ernest W. Adams. 1991. *Archaeological Typology and Practical Reality: A Dialectical Approach to Artifact Classification and Sorting*. Cambridge: Cambridge University Press.
- Aldenderfer, Mark. 1982. "Methods of Cluster Validation for Archaeology." *World Archaeology* 14 (1): 61–72.
- Alkım, U. Bahadır, Handan Alkım, and Önder Bilgi. 1988. *İkiztepe I, Birinci ve İkinci Dönem Kazıları*. Ankara: Turk Tarih Kurumu Basimevi.
- Aydingün, Şengül, and H. Ali Ekinci. 1999. "Burdur Müzesinde Korunan Çaykenar Tip İdollerin Öncüsü Pişmiş Toprak Bir İdol." *Arkeoloji ve Sanat* 90: 29–31.
- Bartel, Brad. 1981. "Cultural Associations and Mechanisms of Change in Anthropomorphic Figurines during the Neolithic in the Eastern Mediterranean Basin." *World Archaeology* 13 (1): 73–86.
- Biehl, Peter F. 1996. "Symbolic Communication Systems: Symbols on Anthropomorphic Figurines of the Neolithic and Chalcolithic from South-Eastern Europe." *JEurArch* 4: 153–176.
- Bilgi, Önder. 2001. *Metallurgists of the Central Black Sea Region: Protohistoric Age; A New Perspective on the Question of the Indo-Europeans' Original Homeland*. Istanbul: TASK.
- Brugger, Peter. 1999. "One Hundred Years of an Ambiguous Figure: Happy Birthday, Duck/Rabbit!" *Perceptual & Motor Skill* 89: 973–977.
- Brugger, Peter, and Susanne Brugger. 1993. "The Easter Bunny in October: Is It Disguised as a Duck?" *Perceptual & Motor Skills* 76: 577–578.
- Buccellati, Giorgio. 2007. "Non-linear Archaeology." *Backdirt: Annual Review of the Cotsen Institute of Archaeology*, 37–39.
- Carroll, Lewis, and John Tenniel. 1896. *Alice's Adventures in Wonderland*. New York: Hurst.
- Feldman, Marian H. 2014. *Communities of Style: Portable Luxury Arts, Identity, and Collective Memory in the Iron Age Levant*. Chicago: University of Chicago Press.
- Fliegende Blätter*. 1892. <<https://doi.org/10.11588/diglit.2137#0147>>.
- Ford, James A. 1954. "Spaulding's Review of Ford." *AmA* 56 (1): 109–112.
- Ford, James A., and Julian H. Steward. 1954. "On the Concept of Types." *AmA* 56 (1): 42–57.
- Gansell, Amy R., Jan-Willem van de Meent, Sakellarios Zairis, and Chris H. Wiggins. 2014. "Stylistic Clusters and the Syrian/South Syrian Tradition of First Millennium BCE Levantine Ivory Carving: A Machine Learning Approach." *Journal of Archaeological Science* 44: 194–205.
- Gnecco, Cristóbal, and Carl Langebaek, eds. 2014. *Against Typological Tyranny in Archaeology*. New York: Springer.
- Gordon, Allan D. 1999. *Classification*. Boca Raton, FL: Chapman and Hall.

- Guralnick, Eleanor. 1973. "Kouroi, Canon and Men: A Computer Study of Proportions." *Computer Studies in the Humanities and Verbal Behavior* 4: 77–80.
- Hansen, Svend. 2007. *Bilder vom Menschen der Steinzeit: Untersuchungen zur anthropomorphen Plastik der Jungsteinzeit und Kupferzeit in Südosteuropa*. Mainz: Philipp von Zabern.
- Hermón, Sorin, and Franco Niccolucci. 2002. "Estimating Subjectivity of Typologists and Typological Classification with Fuzzy Logic." *ACalc* 13: 217–232.
- Hodder, Ian. 2012. *Entangled: An Archaeology of the Relationships between Human and Things*. Malden, MA: Wiley-Blackwell.
- Korfmann, Manfred. 1983. *Demircihüyük: Die Ergebnisse der Ausgrabungen 1975–1978*. Vol. 1. *Architektur, Stratigraphie und Befunde*. Mainz: Philipp von Zabern.
- Korfmann, Manfred, and Bernd Kromer. 1993. "Demircihöyük, Beşik-Tepe, Troia- eine Zwischenbilanz zur Chronologie dreier Orte in Westanatolien." *StTr* 3: 135–171.
- Kökten, Kiliç, Nimet Özgüç, and Tahsin Özgüç. 1945. "1940 ve 1941 Yılında Tarih Kurumu Adına Yapılan Samsun Bölgesi Hakkında İlk Kısa Rapor." *Belleten* 9 (35): 361–400.
- Madsen, Torsten. 2007. "Multivariate Data Analysis with PCA, CA AND MS." *Introduction to CAPCA Programme*. <<http://www.archaeoinfo.dk/>>.
- Makowski, Maciej. 2005. "Anthropomorphic Figurines of Early Bronze Age Anatolia." *Archeologia* 56: 7–30.
- Martino, Shannon. 2012. "The Intersection of Culture and Agency as Seen Through the Shared Figurine Genre of the Prehistoric Southwest Black Sea." PhD diss., University of Philadelphia.
- Obladen-Kauder, Julia. 1996. "Die Klientfunde aus Ton, Knochen und Metall." In *Demircihüyük: Die Ergebnisse der Ausgrabungen 1975–8*. Vol. 4. *Die Klientfunde*, edited by Manfred Korfmann, 209–314. Mainz: Philipp von Zabern.
- Renfrew, Colin. 1969. "The Development and Chronology of the Early Cycladic Figurines." *AJA* 73: 1–32.
- Roy, Ellen, and David Reason. 1979. *Classifications in their Social Context*. London: Academic Press.
- Seeher, Jürgen. 1988. *Demircihüyük: Die Ergebnisse der Ausgrabungen 1975–1978* 3,2 3,2: *Die Kerami*; 2, C, *Die frühbronzezeitliche Keramik der jüngeren Phasen (ab Phase H)*. Mainz am Rhein: Philipp von Zabern.
- Seeher, Jürgen. 1992. "Die Nekropole von Demircihüyük-Sarıket." *IM*. 42: 5–19.
- Spaulding, Albert C. 1953. "Review: Measurements of Some Prehistoric Design Developments in the Southeastern States by James A. Ford." *AmA* 55 (4): 588–591.
- Spaulding, Albert C. 1954. "Spaulding's Review of Ford." *AmA* 56 (1): 109–114.
- Thissen, Laurens. 1993. "New Insights in Balkan-Anatolian Connections in the Late Chalcolithic: Old Evidence from the Turkish Black Sea Littoral." *AnatS* 43: 207–237.
- Tixier, Jacques. 1967. "Procédés d'Analyse et Questions de Terminologie dans l'Etude des Ensembles Industriels du Paléolithique Récent et de l'Épipaléolithique en Afrique



- du Nord-Ouest." In *Background to Evolution in Africa*, edited by Walter W. Bishop and J. Desmond Clark. Chicago: University of Chicago Press, 771–820.
- Ucko, Peter. 1968. *Anthropomorphic Figurines of Predynastic Egypt and Neolithic Crete, with Comparative Material from the Prehistoric Near East and Mainland Greece*. London: Andrew Szmidla.
- Weinberg, Saul S. 1951. "Neolithic Figurines and Aegean Interrelations." *AJA* 55 (2): 121–133.

## PART 3

### *Texts*

∴



# A Qualitative Approach Using Digital Analyses for the Study of Action in Narrative Texts: KTU 1.1-6 from the Scribe 'Ilimilku of Ugarit as a Case Study

*Vanessa Bigot Juloux*

## Introduction

The hermeneutics of action is key to understanding both an action and the interpersonal relationships narrated in literature.<sup>1</sup> Consider this short example as an illustration based on a summary of KTU 1.3:ii:5b–6a: “Anatu fights in the valley.”

We can interpret this action in the sense of inflicting either humiliation or physical injury on a group of people. In order to determine the specific meaning of such a phrase, we need to consider its wider context. An investigation into the interactions between characters in narrative texts and their agency can shed light on the author’s intentions.<sup>2</sup> In this chapter, I outline a digital method that facilitates this approach in the study of Ugaritic literature.<sup>3</sup> Rath-

- 
- 1 *In memoriam* my mother, Madeleine Bigot-Bovia. I am very grateful to Nicolas Wyatt (my co-adviser, emeritus of the University of Edinburgh) for his kind advice, to Glenn Roe (Australian National University) for his English-editing suggestions, to Terhi Nurmikko-Fuller (Australian National University) for her useful suggestions in anticipation of a digital ontology of power relationships and for reviewing my English, and to Daniel Stockholm (École Pratique des Hautes Études, Paris Sciences et Lettres) for his valued assistance on R. I am also grateful to the internal committee and peer reviewers for their kind suggestions. Of course, the final content is my own responsibility.
  - 2 This approach is also useful for the study of historical corpora (such as annals and chronicles) and some epistolary texts. According to Englehardt (2013, 4): “Agency is an open concept that can be employed in different theoretical contexts for different interpretative goals.” In a context of social construction, one speaks of behavior of agency that necessarily includes three elements: the ability to act, the willingness to act, and the power to act.
  - 3 For other Ugaritic texts, see also in this volume, Prosser, 324–328. I will not dwell on Hebrew narrative traditions, especially since I am not a Hebrew scholar; I will rather refer to contemporary or previous cultures to Ugarit.

er than describe an interface for working with textual data, I will explain the encoding process and the rationale for this interdisciplinary methodology.

I focus on ‘Anatu in the Ba’lu and the ‘Anatu Cycle (KTU 1.1–6),<sup>4</sup> a narrative story, traditionally viewed as a “myth” by scholars, about the fight between two clans over the throne (likely of the kingdom of Ugarit).<sup>5</sup> Written in alphabetic cuneiform on six double-sided clay tablets, the text is attributed to the scribe ’Ilimilku the Šubbanite of Ugarit (modern name: Ras Šamra), a Bronze Age kingdom situated on the northwestern coast of what is now Syria.<sup>6</sup> The date of the writing remains debated, but I follow Dennis Pardee’s hypothesis that it was composed during the last quarter of the thirteenth century BCE.<sup>7</sup> However, our understanding of this literary composition is complicated for two main reasons: first, it follows the archeological gap of the fifteenth century, which renders the chronology difficult to establish due to a lack of evidence;<sup>8</sup> and second, at least 50 percent of the text is missing, and only a single copy survives.<sup>9</sup> This could explain why Ugaritic narrative texts assigned to ’Ilimilku have been examined with great interest by Ugaritologists.<sup>10</sup> Nevertheless, what existing studies have in common is a lack of analysis of the relationships be-

4 This myth is well known as the Ba’lu Cycle, but, based on the preliminary results, I propose to rename it (as does Pitard [1999, 53]). The KTU reference follows *Die Keilalphabetischen Texte aus Ugarit (KTU)* (Dietrich, Loretz, and Sanmartín 2013–). Concordance: KTU 1.1 = RS 3.361; KTU 1.2 = RS 3.367 + 3.346; KTU 1.3 = RS 2.[014] + 3.363 + 3.364; KTU 1.4 = RS 2.[008] + 3.341 + 3.347; KTU 1.5 = RS 2.[022] + 3.[565]; KTU 1.6 = RS 2.[009] + 5.155.

5 See Fensham (1979, 273), who also postulates that they represent two clans in a struggle for supremacy between two groups from northern Syria.

6 For other narrative compositions by ’Ilimilku, see: KTU 1.14–16, KTU 1.17–19. For the most recent research in favor of ’Ilimilku’s authorship, see Wyatt (2015). See also Mazzini (2004, 68), who notes that it “is generally considered to have been written at the court of Niq-maddu II, so that ultimately he too might have been a witness of the political event mentioned.” The Kingdom of Ugarit had about two hundred villages, with an estimated population of thirty-five thousand inhabitants, including the capital (Liverani 2011, 326). The kingdom was destroyed at the end of the Late Bronze Age, c. 1200 BCE. For additional information about Ugarit (Ras Šamra), see in this volume, Prosser, 315–316.

7 Pardee 2012, 11–12, and following Mazzini (2004) and Wyatt (2015). However, I may enrich the discussion in my forthcoming doctoral dissertation.

8 Vidal 2006, 172–173.

9 Wyatt 2002, 36. The tablet was found in 1929 by Claude Schaeffer and Georges Chenet in the library of the chief priest.

10 Among other relevant research on KTU 1.1–6, see Smith and Pitard (2008), Wyatt (2002); on the kingdom of Ugarit, see Freu (2006), Yon (2006).

tween the characters and of the connection of the narrative to anthropological phenomena.<sup>11</sup>

Investigating an action introduces a new perspective for interpreting the role of each character, and it can hint at the motivations of the author, be they social, political, or otherwise. To reveal how this hermeneutics of action works, we need to look at empirical testimonies (such as annals and chronicles). It is pertinent to compare contexts, such as war contexts, especially since ‘Anatu is traditionally known as a bloodthirsty goddess.<sup>12</sup> My concern is the role assigned to ‘Anatu and her actions, rather than her divine figure.<sup>13</sup> The semantic field of a verb is one piece of evidence for identifying a role within the context of similar events.

For my initial analyses for my Master’s thesis, I counted all of the verbs used for each character in KTU 1.1–6.<sup>14</sup> Preliminary results showed that in this episode, Ba’lu is in a secondary role almost a quarter of the time (24 percent of the

11 With the notable exceptions of Page (1998), Karkajian (1999), Schloen (2001), and Murphy (2010). Natan-Yulzary’s (2009) study of ‘Anatu and the Aquatu’s Legend should be noted; this study has the merit of highlighting the characters’ roles and their relationships. However, one can only regret the anachronistic comparison with the book of Job.

12 Scholars have fixated on ‘Anatu’s violence and sexuality (e.g., Kapelrud 1969; Gray 1979). For an overview of most hypotheses, see Walls (1992, 161–175). This significant violence defines her as a cruel goddess, full of bloody violence that would be linked to either a fertility or seasons rituals (Hentrich 2001). Even though she is primarily known as a war goddess, ‘Anatu is often associated to the so-called fertility cult. However, I strongly disagree with this interpretation, as do others (notably, Day [1991, 142] and Wilson [2013, 179]: “assumptions that cannot be justified, assumptions biased by Western notions of womanhood that are more androcentric than Ugaritic culture may have been.” As Wyatt (email to author, April 2017, quoted with his kind agreement), says: “Fertility cult’ is basically derogatory and hysterical language to characterize supposedly ‘orgiastic’ and ‘licentious’ cults, such as anything Canaanite, and Ugaritic by definition ([as proposed by scholars including] John Gray, Johannes de Moor, Leila Leah Bronner, Ulf Oldenburg *et al.*).” For an example, see Gray (1965, 45). Wyatt is currently working on a project dealing with ritual: “Ritual is essentially the repetition of a stylized form of activity, which, in conforming to an archetype, is believed to “re-member” it, and perpetuate it in the continuing life of the community.” For some scholars, ‘Anatu acts with violence toward others without specific reason; de facto, her violence would be associated with her intrinsic personality. Jeffery Lloyd quite well reminded us that it is difficult to judge ‘Anatu’s violence in KTU 1.3.ii: “Since we are unaware of the nature of the material that preceded and followed this episode this judgment might seem harsh.” He added that “we can legitimately compare this narrative to historical practice” (Lloyd 1994, 164, 166), pointing to ‘Anatu’s violence and her relation to history.

13 For further explanation see below, under “Category: Role (@xml:id=“role”).”

14 Juloux 2013.

verbs). Since in the same episode ‘Anatu is the most active over a third of the time (34 percent), we can conclude that, contrary to earlier studies, ‘Anatu is the main protagonist and Ba’lu plays a secondary role.

To my knowledge, there is, as of yet, no digital analysis of ’Ilimilku’s corpus that includes the analytical taxonomies that I am suggesting here.<sup>15</sup> The approach presented here is experimental. I will first introduce the general methodology, in which useful definitions for understanding the chosen approach will be noted. The following two parts focus on an explanation of the digital humanities methodologies used for the analytical taxonomies and text mining, both encoded within TEI-XML.<sup>16</sup> The final stage in the preparation for the creation of a hermeneutics of action is reserved for R (a free software environment that enables data manipulation).<sup>17</sup> Here I will only give a short example illustrating the importance of my analytical taxonomies to the preliminary findings from parsing the text using an R environment.<sup>18</sup>

### Prior Reflection

Let me introduce the thought process that has led to the analytical approach. A variety of approaches can be applied to the investigation of a narrative text: pragmatics, semantics, philosophy of action,<sup>19</sup> empiricism, intersubjective phenomena, and Popperian deduction. These allow us to engage, *in fine*, an analysis close to anthroposocial sciences as defined by Jean-Michel Berthelot,<sup>20</sup>

15 “The Ras Shamra Tablet Inventory” (RSTI) is a research project of the Oriental Institute of the University of Chicago co-directed by Miller C. Prosser and Dennis Pardee (<<http://ods.uchicago.edu/rsti/>> [accessed May 28, 2017]). RSTI uses the Online Cultural and Historical Research Environment (OCHRE) (<<https://ochre.uchicago.edu/>> [accessed May 28, 2017]). While a hermeneutics of action has not yet been attempted in OCHRE, the flexible data model makes this sort of research possible. I have already discussed the possibility of a future collaboration with Prosser. Regarding RSTI and OCHRE, see in this volume, Prosser, 317–322.

16 I will return to a discussion of Extensible Markup Language (XML) and the Text Encoding Initiative (TEI) in the subsection “Elementary Explanation” below.

17 As for data manipulation, see in this volume, Svård, Jauhiainen, Sahala, and Lindén, 226.

18 For a short explanation on R, see below under “Text Mining.”

19 The French philosopher Maurice Blondel (1861–1949) first introduced the philosophy of action. See *Internet Encyclopedia of Philosophy*, s.v. “Maurice Blondel,” <<http://www.iep.utm.edu/blondel/>> (accessed May 28, 2017).

20 Berthelot 2004, 17: “Parler des sciences anthroposociales, c’est souligner une unité et interroger une diversité ... La diversité est celle des disciplines.” My translation: “To talk about

and that is, to some extent, similar to the ethnomethodological approach.<sup>21</sup> This approach is no exception, since, as shown by Roberto P. Franzosi, “in recent decades, social scientists, from psychologists to sociologists and anthropologists, have paid increasing attention to the study of narrative.”<sup>22</sup>

### *Hermeneutics*

What is hermeneutics, other than a method for understanding a text?<sup>23</sup> There are abundant extant varieties.<sup>24</sup> We are indebted to Aristotle for the original conception of hermeneutics based on the logical approach, followed by the logician Petrus Ramus in the sixteenth century who used a substantially similar method while criticising the Greek philosopher.<sup>25</sup> In the seventeenth century, from the logical tradition, first Johann Conrad Dannhauer, then Johann Clauberg introduced the scope of hermeneutics,<sup>26</sup> although with two different views,<sup>27</sup> followed chronologically by Friedrich Schleiermacher, Wilhelm Dilthey, and Paul Ricoeur—just to mention some scholars who have influenced

---

the anthroposocial sciences is to emphasize a unity and to question diversity ... The diversity is that of disciplines.”

- 21 The ethnomethodological approach includes political, historical, geopolitical, cognitive, legal, philological, and iconographical fields of expertise.
- 22 Franzosi 2010, 600.
- 23 *Stanford Encyclopedia of Philosophy*, s.v. “hermeneutics,” <<https://plato.stanford.edu/entries/hermeneutics/>> (accessed May 26, 2017). Moreover, Jean-Claude Gens (2006) has clearly shown the connection between logic and hermeneutics.
- 24 Types of hermeneutics include cultural (deriving from the “writing culture” debate on the interpretative anthropology of Clifford Geertz), literary (based on philological concepts, mostly influenced by Friedrich Schlegel), and philosophical (inspired by Martin Heidegger and his reflection on “Dasein”).
- 25 The concepts of analysis developed by Aristotle and Ramus are opposed to those of Dannhauer (Gens 2006, 21). *Stanford Encyclopedia of Philosophy*, s.v. “Petrus Ramus,” <<https://plato.stanford.edu/entries/ramus/>> (accessed March 4, 2018).
- 26 The term *hermeneutica* had been used once in 1629 by the Ramist philosopher Alexander Richardson in a commentary on the dialectics of Ramus (Gens, 2006, 16). *Peri Hermeneias* (second book of *Organon*). See also Grondin (2006, 2) and Molinié (2007, 433–444). Grondin 2006, 1: “Il s’est ouvertement inspiré du traité d’Aristote intitulé *Peri hermeneias* (*De interpretatione*) et a soutenu que la nouvelle science de l’interprétation n’était effectivement rien d’autre qu’un complément à l’organon aristotélicien.” My translation: “He was openly inspired by Aristotle’s treaty entitled *Peri hermeneias* (*De interpretatione*), and he argued that the new science of interpretation was actually nothing more than an addition to Aristotle’s *Organon*.” See also Gens (2006).
- 27 (1) Aristotelician and (2) Cartesian views (Gens, 2006, 5).



my approach, first for the practice of observation, second for investigating the observed data.<sup>28</sup>

With the digital age, two new types of hermeneutics have arisen: digital hermeneutics, based on digital ontologies,<sup>29</sup> and computational hermeneutics for big data studies.<sup>30</sup>

Within the scope of literature, specific hermeneutics have arisen, of which various methods are borrowed from more general hermeneutic methods: among others, Biblical hermeneutics, Talmudic hermeneutics, Babylonian hermeneutics, and Mesopotamian hermeneutics. Although the genre of texts is different (Mesopotamian commentaries about divination versus narrative literature), I will briefly talk about Mesopotamian hermeneutics,<sup>31</sup> in particular Uri Gabbay's notable approach that focuses on, among other topics, the intention of Mesopotamian commentaries.<sup>32</sup> One may think *prima facie* that Gabbay's approach is close to the intentionality concept (defined below), since we both focus on the same questions (what, how, and why). But in fact, it is not: our methods and purposes are different. While Gabbay examines the intention of the commented text according to the commentator's interpretation,<sup>33</sup> I am

28 Ricœur 1969 and 1986; Laks and Neschke 2008.

29 Capurro 2010, 37–38: "It deals with processes related to the digital network at the social level, autonomous systems of interpretation, communication and interaction (robotics), as well as all kinds of hybrid biologic systems (bionics) and digital manipulation at the nano level ... It aims at translating and interpreting *logos* and *arithmos* within the human realm but it is not restricted to this sphere."

30 Mohr, Wagner-Pacifici, and Breiger 2015, 3–4: "The central idea of a computational hermeneutics is that all available text analysis tools can and should be drawn upon as needed in order to pursue a particular theory of reading ... Instead of focusing on the main communicative intentions of a text, we are now able to push toward the kind of close reading that has traditionally been conducted by hermeneutically oriented scholars who find not one simple uncontested communication, but multiple, contradictory and overlapping meanings." Big data: large and complex datasets that require computational methods for their analysis (Dutcher, 2014, <<https://datascience.berkeley.edu/what-is-big-data>> [accessed June 2, 2017]).

31 Selz 2013, 48: "The term 'Mesopotamian hermeneutics' is used in the following contribution in a very broad and modest sense. We will search for indications of epistemic self-reflexivity within the framework of early Mesopotamian scholarship. Mesopotamian scholarship is always empirically based—that means knowledge is founded on various sorts of observations." It is based on signs signification according to the diviner who "holds the 'hermeneutic keys' to the divinatory code" (Koch 2010, 44).

32 Gabbay 2016.

33 The intention here relies on the literal meaning rather than philosophical scope described below. "The concept of 'literal meaning' can relate to two different categories. It can refer

concerned with the author's intention regarding his motivated choices of characteristics and the action's characters, by investigating characters intentionality. Actually, intention and choices are related in that both attempt to answer the questions: why does the author assign such and such characteristics to a specific character rather than to another, and how does this help the anthropological interpretation? Whereas Gabbay uses hermeneutical procedures based on "the lexical tradition and the divinatory tradition,"<sup>34</sup> I consider empiricism (mainly politico-historical) and pragmatics (defined below), which are essential in order to prepare to a hermeneutics of action.<sup>35</sup>

### *Pragmatics, Semantics, and Intentionality*

A text includes sentences and utterances. A sentence has grammatical units; an utterance has contextual information and relevant elements for the interpretation. These elements have several meanings. Using pragmatics, in particular those of Herbert Paul Grice,<sup>36</sup> invites us to consider the implicature,<sup>37</sup> or what is suggested by an utterance—in other words, what is beyond the conventional linguistic meaning: what the author wants to tell the audience according to both the context of the composition and the utterance itself.<sup>38</sup> From my point of view, the relevance of pragmatics lies here.

To get back to my first example, "Anatu fights in the valley," one needs to consider the two interrelated criteria: the implicature according to the author's intention, and semantics. On the latter criterion, my first concern is the verb,

---

to a lexical understanding of a word or phrase regardless of its context, for example, when figurative language occurs; or it can refer to the obvious intention of a sentence or passage, usually agreeing with the basic lexical meaning of the words that comprise it, as opposed to a more expository meaning achieved through exegesis. The lexical meaning of a word usually fits its meaning in the context it is found in, and the formulation of a text usually reflects the intention of the text" (Gabbay 2014, 335).

34 Gabbay 2014, 335.

35 Within the scope of Mesopotamian hermeneutics, "action" has different meanings. For example, for extispicy, action's diviner has to be understood as "instrument" (Koch 2010, 54). For further discussion of "action," see the subsection "Actancial Event," below.

36 Grice 1975; *Stanford Encyclopedia of Philosophy*, s.v. "implicature, Gricean theory," <<https://plato.stanford.edu/entries/implicature/#GriThe>> (accessed April 25, 2017).

37 Implicature is a specific term for pragmatics (*Stanford Encyclopedia of Philosophy*, s.v. "implicature," <<https://plato.stanford.edu/entries/implicature>> [accessed April 25, 2017]).

38 For example, "It is twelve" may mean either "It is lunchtime" or "It is time to leave for my appointment."

which has several senses belonging to significantly distinct semantic fields that affect its interpretation.<sup>39</sup>

However, I will mostly observe the character's intentionality rather than the authors's intention. I will, of course, explain the reason.

In the scope of philosophy of action, "intentionality" has a significantly different meaning than "intention." One speaks of intentionality when the action is achieved, and one speaks of intention when the action is hoped to be achieved. For example, (A) "Lena wants to go to the concert." There are two possible outcomes: either she goes or not. If (B) "Lena went to the concert," the action was achieved, and one would talk of Lena's intentionality. But if (C) "Lena did not go to the concert," then one would talk of her intention, albeit unachieved in this case. To sum up, intentionality includes states A, B, and C,<sup>40</sup> while intention is only demonstrated by state C.<sup>41</sup> Thus, whether the action is voluntary or not, intention to act is not always achieved. Regarding the intentionality, one looks to know for what reason the action was performed, while for intention, one only wonders about the motivation of the foreseen action.

One might object: what about an action of motion, since some actions may be not voluntary? For example, slipping on the floor is an involuntary action, while dancing is voluntary. Honestly, no one has a voluntary intention to slip on the floor—with few exceptions (theater, circus). But, nonetheless, it happens: the action of slipping is achieved, as is the action of dancing. Both no longer fall within intentional actions. A reason for acting in a particular way can also justify the action itself;<sup>42</sup> this is the case with dancing.

However, both "intention" and "intentionality" have consequences which need to be investigated, since an action achieved or not was decided for one or several reasons. These consequences concern either the person who performs the act, or other only or both.

---

39 I will return to this topic below in the subsections titled "Actancial Event" and "Category: Verb (@xml:id="Action")."

40 Davidson 2008, 129: "Former une intention peut être une action, mais ce n'est pas une réalisation." My translation: "To form an intention may be an action, but it is not a realization."

41 Anscombe 2002, 45. For additional information about several states of reason, see Audi (2003, 77, 102–104). According to Ricœur (1977, 13): "Le transfert de la connaissance à l'action repose sur le parallélisme entre l'objet et l'événement, entre être vrai et rendre vrai." My translation: "The transfer of awareness to action is based on the parallel between the object and the event, between being true and making true."

42 See Davidson 2008, 17.

### *Empiricism and Intentionality*

Of course, I am not considering the act of writing words on a clay tablet. To do so, one needs to consider that first the scribe 'Ilimilku intended to narrate a story, and he did. The act of writing was indeed achieved; thus, all three states A, B, and C would support the analysis of the author's intentionality.<sup>43</sup> The main problem is that we are not able to analyze the consequences of state B, among other reasons, we have no evidence for its reception by an audience of readers and listeners (in the case of oral delivery) in antiquity. Furthermore, I prefer to avoid assumptions for state B, so I only will look to the author's intention by investigating intentionality of characters in the narrative story.

I will use an empirical approach to observe an action that needs to be understood as a change of conditions from state A to state B. Regarding an act of violence, it needs to be contextualized according to a place where a parallel can be drawn with a similar event and place. For example, somebody beats a group of people: in which texts do we find this event other than the one we are currently studying? This analogical event needs to be known by the author and/or associated with an intersubjective phenomenon.<sup>44</sup> Based on the assumptions of Nicolas Wyatt about 'Ilimilku's authorship of this text,<sup>45</sup> it is quite believable that this scribe has accumulated much useful knowledge about history testimonies, since literature competencies are the highest level of his training.<sup>46</sup>

Going back to the previous example, "somebody" can be a sovereign. Before a military campaign, he requested a god's approbation in some ways—an intersubjective phenomenon. Then the question arises as to why the author has chosen a god to beat a group of people in a similar event—and beyond this by

43 See also the auctorial intentionalism (Bühler 2015, 240).

44 A community network is connected by a subjective consciousness. Harari 2015, 144–145: "Est intersubjectif ce qui existe au sein du réseau de communication qui lie la conscience subjective de nombreux individus ... Nombre de moteurs les plus importants dans l'histoire sont intersubjectifs: loi, argent, dieux et nations." My translation: "It is intersubjectivity that exists within the communication network that links the subjective consciousness of many people ... Many of the most important driving forces in history are intersubjective: law, money, gods, and nations." Harari takes the example of human rights. They belong to the imagination of billions of people. If one person no longer believes in human rights, there will be no overall impact, but if, during a long time, a very large group of people stop believing in them, then they may be called into question.

45 See note 6, above.

46 The training of an Ugaritic scribe included five levels: (1) lists of syllabic signs, (2) thematic lists of word signs, (3) model documents, (4) thematic list of "knowledge," and (5) "poetry/literature" (Hawley 2008, 59).

choosing a verb from a similar semantic group to the one used to describe the sovereign's action. Then it is easier to suggest the intention of an author, *in fine*, by the signification of an utterance, as well as his or her will to make significant reference to the mental realities of the addressee according to socialized codification.<sup>47</sup> This is particularly true when taking the assumption of Donald Davidson:

Les croyances et les désirs nous disent quelles raisons un agent a d'agir seulement si ces attitudes se trouvent reliées de manière appropriée à l'action telle que l'agent lui-même la considère.<sup>48</sup>

Regarding the previous explanation, the author of a narrative text is the primary agent. So the question is simple: why has the author intentionally assigned a type of action to a specific person? First one needs to analyze each action performed by a substitute agent, a fictional character, since I believe that the intentionality of a character in a literary narrative is intrinsically dictated by the author's will—for example, motivated by the correlation with history.

### *Actancial Event*

It goes without saying that such an analysis cannot be performed by a text-oriented approach alone.<sup>49</sup> I believe that the weakness of such an approach is due mainly to the negligence of analytical criteria.<sup>50</sup> These are also required in order to determine the analytical variables that, *de facto*, help to avoid a biased reading.

I describe the notion of action as an “actancial event” because it necessarily jointly implies an actant (subject = main agent, mostly active) and an event

---

47 Following Le Ny (2001, 32): “Chaque fois qu'un verbe désignateur d'événement est utilisé dans une phrase, la signification de celle-ci comporte une référence à un exemplaire de l'ensemble général de ces réalités du monde qu'on appelle des événements, c'est-à-dire une exemplification du concept d'événement.” My translation: “Each time a verb that shows an event is used in a sentence, its signification includes a reference to a model of the general set of these realities that is an exemplification of the event concept.”

48 Davidson 2008, 120. My translation: “Beliefs and desires tell us what reasons an agent has to act only if these attitudes are linked in an appropriate manner to the action as the agent himself understands it.”

49 This is also known as a “literary approach”: “In summary, a text-oriented approach focuses on the text” (Sun 2008).

50 Juloux 2016a.

(verb).<sup>51</sup> The focus on an actantial event will be useful within the scope of the study of a narrative story since a story “refers to a skeletal description of the fundamental events in their natural logical and chronological order or sequence” and the roles of the actants.<sup>52</sup> Thanks to the philosophy of action, an observation of all of the components related to an actantial event can be used to appraise quantifiable variations linked to the change of a state of the agent through an event.

These variations, like the actantial event itself, belong to distinct groups, and are sorted systematically prior to the extraction. I have defined three analytical taxonomies to investigate an actantial event: primary data, objective variables, and subjective variables—which I will address in detail in the next section.

What are objectivity and subjectivity? Their common characteristics are their affiliation with the field of knowledge. I will not dwell on definitions here. However, I will rely on the explanation of Popper, who defined objective knowledge as demonstrable knowledge that follows a certain truth, while subjective knowledge is linked both to our own beliefs and opinions.<sup>53</sup> Following Popper’s assumption, objective variables are lexical items for which meaning can easily be identified according to common knowledge (such as context, biological sex, verb).<sup>54</sup> Subjective variables fall within our own interpretations (consider, for example, an emotion). Popper suggested that subjective knowledge may become objective by deduction. In this, the proposed taxonomies find their pertinence for the purpose of a hermeneutics of action that follows the principle of a Popperian deduction, as soon as empiricism is also taken into account.

---

51 Following Greimas’s theory (1987, xxxiv) on the role of the actant subject and its competency: the actant subject will be endowed successively with the modalities of competence, and in this case the “subject assumes those actantial roles which manifest the subject in terms of wanting, the subject in terms of knowing, and the subject in terms of being able to do.” The main agent’s antonym is the auxiliary agent, which comes after the verb.

52 See Franzosi 2010, 597.

53 See Popper (1998, 138–139), in particular: “La connaissance subjective est un genre de disposition dont l’organisme peut parfois prendre conscience sous la forme d’une croyance, d’une opinion ou d’un état d’esprit.” My translation: “Subjective knowledge is a kind of disposition from which the being may sometimes become aware either of an opinion or a state of mind, in the form of a belief.”

54 Later in this paper, I come back to each of these terms as well as to the concept of “subjective variables.”

To get back to an actantial event, one has to distinguish the action itself from the result of the action. The two are intrinsically linked. Both can be useful to measure the capacity for setting up authority and power. There can be several outcomes from a single action, each with its own consequences. The result of an action is not necessarily what was anticipated at the outset. It is not always straightforward to take into account the reason for an action that assumes the notion of belief, notably as defended by Davidson;<sup>55</sup> thus, following the intersubjective phenomenon, an agent performs an action if he believes he will have a positive result.

More importantly, this process enables us to set up analytical taxonomies and a structural framework for a preliminary step toward an interpretation. I begin by focusing on 'Anatu's actions, establishing the groundwork for a hermeneutics of her action. This preliminary step, which uses mixed methods, includes three parts: analytical taxonomies, text mining both with TEI, and parsing with R. In this paper, I will mostly focus on TEI.

## Analytical Taxonomies in TEI

### *Research Method*

Quantitative and qualitative approaches are analytical methods of text and data mining.<sup>56</sup> This broad range of techniques enables the extraction of available knowledge from a large dataset in order to find correlations between variables and unexpected items or models (similar arguments). Among a wider range of datasets, there is textual data.

Quantitative methods allow the researcher to count and measure a group of data;<sup>57</sup> they use mathematical formulas and statistics to express results in terms of numbers or sets of numbers. Graphics or templates are often used to communicate results. In contrast, qualitative methods can be employed independently or in conjunction with quantitative methods. Qualitative methods

55 Davidson (2008, 17, 20) talks about a primary action that is its cause: a pro-attitude (the agent's mental attitude, which implies desire) and the associated belief. However, I will not develop both types of action here. For further discussion on the topic, see *Stanford Encyclopedia of Philosophy*, s.v. "action," <<https://plato.stanford.edu/entries/action/>> (accessed May 27, 2017).

56 In this volume, a) for qualitative method applied to objects, see Martino and Martino, 120; b) for quantitative method applied to semantics, see Svård, Jauhiainen, Sahala, and Lindén, 238–240; c) for measuring intertextual relationships, see Monroe, 266–268.

57 In 1851, Augustus de Morgan first proposed to use a quantitative method of text analysis for the study of the authorship of the Pauline Epistles (Hughes et al. 2012).

enable the collection of data by observation through the researcher's participation in the investigation; in order to show the purpose of the analysis, the most relevant data is used for a deeper investigation, such as parsing with R.<sup>58</sup> This data is recorded and sorted according to its nature in order to match categorical data. Data can be classified into two types: numerical and categorical data. Numerical data is used quantitatively to measure the numerical values' outcome, and categorical data is used qualitatively to organize nominal values into categories. Categorical data is used for text- and data-mining processes.<sup>59</sup> Qualitative methods do not necessarily exclude quantitative methods, because, in some cases, calculations based on significant data types are needed. This combination of approaches, which I have chosen to use in my research, is known as the mixed method. In the context of this essay, I will focus mainly on text mining. The technologies I employed in this analysis are TEI (an XML encoding standard) and, to some extent, R. I used the TEI to record and sort some textual data of KTU 1.1–6, and I relied on R for the data extraction and for the counting of occurrences of the relevant data.

### Elementary Explanation

Before going further, I will provide an elementary explanation of TEI-XML for neophytes.

My approach relies mostly on markup tagging. Marking up online content can be compared to annotating a hard-copy text by hand (i.e., writing notes with a pen on the paper manuscript).<sup>60</sup> As James Coombs, Allen Renear, and Steven DeRose wrote, "The markup is not part of the [online] text or content of the expression, but tells us something about it."<sup>61</sup> A tag is rather an indication of the classification of what is described: after a tag "word," one either has a verb, adverb, noun, or adjective, under its inflectional form or not.<sup>62</sup>

Standardized in the late 1990s, Extensible Markup Language (XML) is a pre-defined markup language that follows a standard syntax, enabling data

58 Relevant data are different, depending on whether the analysis is philological, anthropological, geographical, etc.

59 I provide an explanation below in the "Text Mining" section.

60 For further explanation of "markup" syntax, see W3C (last revised 2008, <<https://www.w3.org/TR/xml/#syntax>> [accessed June 10, 2017]). See also in this volume, Eraslan 289–290, for the markup in EpiDoc.

61 Coombs, Renear, and DeRose 1987, 934.

62 In this volume, a) for additional explanation, see Eraslan, 302n72, and Nurmikko-Fuller, 339; b) for an example of tagging a person in Oracc, see Pagé-Perron, 203–204; c) for further examples used for semantic analyses, see Svärd, Jauhiainen, Sahala, and Lindén, 234–238.



exchange to be easily machine readable.<sup>63</sup> One speaks then of interoperability data—for example, XML would be a kind of artificial constructed language such as Esperanto.<sup>64</sup>

The Text Encoding Initiative (TEI) was created in 1987. Then, in 2000, the TEI Consortium (TEI-C) was established. The TEI-C is a group of international scholars who collaborate on the development of a dedicated encoding standard for text analysis.<sup>65</sup> Unfortunately, using TEI is a “paradox,” since it does not yet enable interoperability, but rather the interchange of cross-corpora text analysis.<sup>66</sup> However, using TEI semantics, such as elements and attributes, enables one to structure a document for highlighting relevant pieces of data and the relationships between them for text analysis—one of the goals of TEI was

---

63 XML also offers to withstand time and new technologies deployed on the internet or locally (on a computer)—in other words, the current methodology is sustainable even as technology evolves (w3c, last revised 2016, <<https://www.w3.org/XML/>> [accessed April 1, 2017]). For other languages that are easily machine-readable, see in this volume, Matskevich and Sharon, 46; Pagé-Perron, 200; Nurmikko-Fuller, 336, 339–340.

64 Interoperability: “a measure of the degree to which diverse systems, organizations, and/or individuals are able to work together to achieve a common goal” (Ide and Pustejovsky 2010, <<https://www.cs.vassar.edu/~ide/papers/ICGL10.pdf>> [accessed April 1, 2017]). On the interoperability of transcription, see Schmidt (2014, 8): “The ability to load a transcription into various programs without modification.” See also Unsworth (2011, <<http://jte.revues.org/215>> [accessed April 12, 2017]). For further explanation about interoperability, especially on epigraphy, see in this volume, Eraslan, 284–285, 292–295, 309–310. Esperanto: an artificial international language (*Encyclopaedia Britannica*, s.v. “Esperanto,” <<https://www.britannica.com/topic/Esperanto>> [accessed May 2, 2017]).

65 For the history of TEI, see Ide and Sperberg-McQueen (1995); on TEI and TEI-C, see Vanhoutte (2004), and about XML and TEI, consult Nellhaus (2001, 257–260). Unsworth 2011: “The ‘I’ in TEI sometimes stands for interchange, but it never stands for interoperability. Interchange is the activity of reciprocating or exchanging, especially with respect to information (according to Wordnet), or, if you prefer the Oxford English Dictionary, it is ‘the act of exchanging reciprocally; giving and receiving with reciprocity.’” See also Nellhaus (2001, 258). A great advantage of TEI is that it can be readily understandable, even by a non-expert. Its encoding includes (1) elements (500 to date), (2) attributes, and (3) values, according to a pre-defined standard syntax. Generally, the semantics of the elements and attributes follows those proposed by the TEI, unlike the values that are mostly defined by the project manager. Its structure is hierarchical: each element has descendants and can have ancestors, in the manner of a family tree (see <<http://www.tei-c.org/index.xml>> [accessed April 15, 2017]). See also in this volume, Eraslan, 289–290, who describes EpiDoc, a structured markup language for epigraphic documents in TEI.

66 TEI interoperability is very difficult because of the level type of the tag (type 1 often used for TEI and type 2 for XML). See Schmidt (2014, <<http://jte.revues.org/979>> [accessed April 12, 2017], 4–5).

“to provide a markup scheme that will permit scholars to encode linguistic analyses of any text in any language to any desired degree of detail.”<sup>67</sup> TEI markup can be compared to either (1) a syntactic unit or (2) a lexical unit. Both are comparable in some ways to elements, and their values (e.g., adjective, common noun, subject) to attributes. An element, which is a markup tag, is the first criterion to analyze data (in the form of text, image, sound, etc.).<sup>68</sup> Each piece of text data is a glyph,<sup>69</sup> a word, a group of words, a reference, or a concept. An attribute adds useful precision both for text analysis and for interpretation during the process of the exchange of data.<sup>70</sup> An attribute stands within the element tag.

67 Langendoen and Simons 1995, 191.

68 An element is conventionally marked up `<element>`. The markup data in between the opening tag “`<`” and the closing tag “`>`” indicates the type of information analyzed: for example, `<persName>‘Anatu’</persName>`. One can easily understand that the element refers to a personal name, in this case ‘Anatu. One can also notice that the first opening tag, which is followed by “Anatu,” is then followed by a slash “/” to indicate the end of information related to that element `<persName>`, ‘Anatu. However, it is also possible to include information within `<element>`; it would be written `<persName [attribute] />`. One can notice that the slash precedes the closing tag. Generally, the semantic of an `<element>` is very close in its abbreviation to its English vocabulary. For example, `<interp>` stands for “interpretation,” `<w>` for “word,” `<l>` for “line,” and `<text>` for “text.” For a further explanation of “element,” see <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/SG.html#SG13> (accessed June 10, 2017). By convention, I will write an element between open and closing tags, and when I refer to an attribute, the name will be preceded “@.”

69 Glyph: a graphical representation, such as a character or an accent. In Ugaritic cuneiform script, it is one or several carved lines or impressed wedges expressing a letter, syllable, logogram, or separation mark (a dot) between two words. For the current text analysis, all glyphs, including the separation mark under the element `<g>`: `<g> . </g>`, are encoded. See also in this volume, Eraslan, 291.

70 An attribute is conventionally preceded by “@” within a tutorial, to distinguish it from an element. An attribute is always followed by an equals sign “=” with its value between quotation marks: `attribute name="value"`. Consider, for example, the attribute `@xml:id`, which always has a unique value: `xml:id="ktu1.1"`. I have defined the value of this attribute in KTU 1.1, a very well-known text of the Ba’lu and ‘Anatu Cycle. To refer to this unique attribute’s value within the TEI file, one uses the analytical pointer `@ana`. The value associated with `@ana` is preceded by a hashtag “#.” That is the main difference with the majority of other attributes. Therefore, when referring to `xml:id="ktu1.1"` within the text, one would write: `ana="#ktu1.1"`. Coming back to the syntactic and lexical units, consider the following example: “Anatu fights in the valley”: `<l><name type="character">‘Anatu’</name><w pos="verb"> fights</w><w pos="prep">in</w><w pos="noun">valley</w></l>` Syntactic

### Categorical Data: Taxonomies

Coming back to markup tags, which are already predefined in TEI P5,<sup>71</sup> their choice results from each project's focus. Regarding author intention and character action, I have used markup tags in order to work with analytical taxonomies and a structural framework (see the section "Analytical Taxonomies in TEI" below). Taxonomies allow one to classify, index, and search types of data by groups of categories and subcategories, which are also called categorical data. Analytical taxonomies are based on criteria for extracting information that is not always noticeable with more standard text-oriented approaches. I have defined three intrinsic analytical taxonomies as key components for a hermeneutics of action:<sup>72</sup> (1) primary data, (2) objective variables, and (3) subjective variables.

### *Prior to the Analysis: Transcription and Pre-processing*

The transcription of the text needs to be completed prior to the analysis in order to contextualize an actantial event,<sup>73</sup> and, de facto, to make the identification of the value of each category.<sup>74</sup> This is called the pre-processing step in order to tokenize the text. The tokenization process consists of splitting the group of words into individual words. Each word is sorted into a category, the equivalent of a TEI attribute value.<sup>75</sup>

---

unit: <l> for line; lexical unit, <w> for word. Example of <w> attribute and values: @pos="verb", @pos="noun". For further reading about "attribute," see <<http://www.tei-c.org/release/doc/tei-p5-doc/en/html/SG.html#SG16>> (accessed April 27, 2017). On "exchange of data," see "interchange," note 65.

71 For TEI P5 guidelines, see <<http://www.tei-c.org/Guidelines/P5/>> (accessed June 10, 2017).

72 For specific guidelines for a preliminary hermeneutics of action on the internet, see Bigot Juloux (2017, <<https://vbigot-juloux.github.io/hermeneutics-of-action/UserManual/out/webhelp/index.html#process.html>> [accessed April 2, 2017]).

73 Since Ugaritic is a cuneiform language, one needs to transcribe the cuneiform signs into Latin glyphs. For example, this Ugaritic verb: Unicode: 𐎗𐎟𐎧 is transcribed as "mḥṣ."

74 In various aspects, this analytical method is close to the "PropBank" annotation of English verbs (see English Propbank Annotation Guidelines, <<https://github.com/propbank/propbank-documentation/blob/master/annotation-guidelines/Propbank-Annotation-Guidelines.pdf>> [accessed April 15, 2017], 3). As shown by Langacker and Vandeloise (1991, <[http://www.persee.fr/doc/comm\\_0588-8018\\_1991\\_num\\_53\\_1\\_1804](http://www.persee.fr/doc/comm_0588-8018_1991_num_53_1_1804)> [accessed April 12, 2017], 103), a category is defined by a set of criteria.

75 An example is given below under "Semantic Categories (@xml:id="verb.category")."

*Taxonomy: Its Framework*

For my project, each taxonomy has different criteria related to the philosophy of action. This approach is influenced by the action-oriented solution.<sup>76</sup> This involves determining whether an action is willed and the “personal” aim of a character. In addition to investigating an action, I am considering whether the actant achieves an action according to his will or not. This question is relevant both for sphere of influence (to which I will return [category: sphere]) as well as for the agency.

Taxonomies (1) and (2) jointly contribute to an actantial framework of intelligibility. This framework enables us to study a character’s action in a way that tends to be closer to impartiality. The exception is the verbal semantic classification that will be developed below. Taxonomy (3), although subjective, proposes to provide measuring elements for evaluating the action according to objective variables.

In order to extract the required information according to categorical data, these three taxonomies are first introduced in the TEI code within the elements `<taxonomy>/<category>` related to the structural classification before the encoded transcription, and identified by `@xml:id`, respectively with:<sup>77</sup>

```
xml:id="primaryData",          xml:id="objectiveVar",          xml:
id="subjectiveVar".
```

The value enumeration follows a common schema for each taxonomy:

```
<taxonomy xml:id="taxonomy's value (1), (2) or (3)">
  <category xml:id="category's value"/>
</taxonomy>
```

The `@xml:id` within the taxonomic elements is necessary for counting the verbal occurrences and for the action’s interpretation for each character. This data will be used to parse relevant attributes’ values in R.

<sup>76</sup> Massin 2014, 91.

<sup>77</sup> `<category>` is a child of `<taxonomy>`, which can also have one or several children and grandchildren—in other words, the descendants of `<taxonomy>`. This notion of hierarchical structure is one of the most important for parsing with R (see an example below in the “Text Mining” section).

### *Primary Data and their Semantic Values*

Primary data is at the core of the analysis.<sup>78</sup> It represents the actantial event that was described above under “Actantial Event.”<sup>79</sup>

Category: Verb (@xml:id=“Action”)

In a previous paper,<sup>80</sup> I explained the role of a verb in a sentence. According to Louis Tesnière,<sup>81</sup> a verb is the node of a sentence—or of a group of words, for this study. According to Maurice Grevisse,<sup>82</sup> from a semantic point of view, a verb expresses an action made or undergone—in other words, a change of state from A to B. It acts as bridge between the subject (the agent = character) and the complement.<sup>83</sup> Therefore, within the framework of a hermeneutics of action, a verb is undoubtedly to be considered in two primary manners 1) in infinitive inflectional forms and 2) for the analysis of Semitic verbs, in terms of gender (which is essential to analysis of the gender role).<sup>84</sup> The latter can be defined thanks to the agency.

78 For all taxonomies, I have chosen to write the first letter of each value's word in lowercase, except for some values that are borrowed from an existing ontology in anticipation of a new ontology on power relationships. These primary data are located after the following elements /teiCorpus/TEI/teiHeader/encodingDesc/classDecl/taxonomy—thus these elements are the common ancestors of each category. For additional information on ontology, see in this volume, Nurmikko-Fuller, 347–348. Primary-data syntax is fully described within the guidelines, see Bigot Juloux (2017, <<https://vbigot-juloux.github.io/hermeneutics-of-action/UserManual/out/webhelp/index.html#PrimData.html>> [accessed April 15, 2017]).

79 See above, pages 160–162.

80 “Herméneutique de l'action pour l'étude des relations entre les entités animées et leur agency au Proche-Orient ancien: Hypatia et alii” (forthcoming in *Proceedings of the 61st RAI, Geneva and Bern, 22–26 June 2015*).

81 Tesnière 1959, 6.

82 Grevisse 1986, 1159.

83 The complement is made of one or several words subordinated to the verb, which bring additional information for the understanding of the sentence or a group of words.

84 In particular, for Semitic languages, a verb has three inflectional forms in order to indicate the gender, number, tense, voice, stem, or mood. The three inflection possibilities are the affix, suffix, and prefix. Although I will not go into details, among many cases, in this paper I will show four examples for the verb *mḥṣ*. (1) Its inflectional form “t-mḥṣ” provides the following information: the prefix /t-/ first, the verb is imperfective (“imperfective” refers to the actuation of an action; in other words, it expresses a present state), the person, gender, number, and tense is either third feminine singular indicative or second masculine singular indicative (I do not consider tenses apart from the indicative here). (2) Its inflectional form “t-m-tḥṣ” has a prefix /t-/; the infix /-t-/ (after the first radical) means a reflexive function (derived stem I called Gt). (3) Its inflectional form “t-m-t-ḥṣ-n,” the

### Category: Animated Entity (@xml:id="Being")

In order to avoid biased interpretations of actions, especially when studying an action, I have decided not to differentiate between divine and human words; this is why I opted for the expression “animated entity” (AE) to describe a character in the text. An AE is an agent (a “being”) associated with a verb.<sup>85</sup>

Each AE is named, one by one, with @xml:id="Character" within <text> /<stage>/<listPerson>/<person>:

```
<persName xml:id="character">'Anatu</persName>
```

The element <stage> was seen as the best element, as it captures the dramatic quality of the text. The element <text> is used following the sections of the taxonomies.<sup>86</sup> Besides, <stage> is also used within the text-mining section.

### Objective Variables

Objective variables have been set up according to pragmatics and semantics.<sup>87</sup>

An action belongs to a semantic category. It is performed by an agent, male or female, in a given context and acts with a peculiar role within a sphere. Each of the following categories is included within:

```
<taxonomy xml:id="objectiveVar" ana="#taxonomies">
[Category]
</taxonomy>
```

For some cases, we cannot clearly identify one of the objective variables (for example, unknown context), the glyphs are unreadable or lost, or there are two common values. It is the same for subjective variables.

---

prefix /t-/ with the suffix /-n/ indicates either the third male/female, second masculine/feminine, all plural. (4) Without a prefix, “mḥṣ-t,” its suffix /-t/ indicates that the verb is in the perfective (simple past) form: third female, second masculine/feminine, first, all singular, or third feminine dual. Regarding the imperfective/present-tense in Ugaritic narrative verse, I follow Greenstein (2006, 102), who has suggested that rather than the present-future/imperfective default sense of YQTL, “the action taking place as current, present-tense, dramatic.”

85 “Being” was taken from OntoMedia ontology (<<http://www.contextus.net/ontology/onto-media/ex/ext/common/being>> [accessed April 5, 2017]).

86 The transcription is also located within <text>. Objective variables are located within /teiCorpus/TEI/teiHeader/encodingDesc/classDecl/taxonomy.

87 Objective variables are located within /teiCorpus/TEI/teiHeader/encodingDesc/classDecl/taxonomy.

### Semantic Categories (@xml:id="verb.category")

The defining of semantic categories is an essential stage in qualitative methodologies, as it enables the usable quantification of verbal occurrences. A verb (@xml:id="Action") is associated with a semantic category. This association includes three distinct stages: (1) defining verbal categories, (2) identifying verbal subcategories, and (3) assigning a verb to one of the subcategories. Unfortunately, since a semantic database for ancient Near Eastern languages does not exist yet,<sup>88</sup> I have used four online databases of English semantic relations: BabelNet, FrameNet, VerbNet, and WordNet.<sup>89</sup> It goes without saying that the known analyses of verbs—in particular, synchronic semantic variations of Bronze Age verbs—needs to be considered, especially during the development of a hermeneutics of action. For the first stage, I selected fourteen verbal categories from WordNet, in order to follow its model of cognitive synonyms, which is significant given that verbs have several cognitive meanings:<sup>90</sup> for example, verbs may relate to realms of the body, change, cognition, communication, competition, consumption, contact, creation, emotion, motion, perception, possession, social interaction, and stative conditions.

For the second and third stages, I have mostly used the cognitive linguistics approach for the correlation of semantic contexts (in some ways such as a prototype),<sup>91</sup> from the four databases that make it possible to check whether the choice of classification is coherent—even though attention should be paid to the diachronic components, due notably to cultural variations. Of course, these two stages of classification can only be done during the transcription process, since verbal category assignment depends on the group of words of an actantial event.

A practical example for the category of emotion relating to humiliation is:

```
<category n="8" xml:id="verb.emotion" ana="#verb.
category #action">
  <!-- according to WordNet [37]: "verbs of feeling" -->
  <catDesc>taxonomy: emotion or psych verbs</catDesc>
```

88 I am currently developing an online, open-source semantic dictionary for Ugaritic.

89 BabelNet: <<http://babelnet.org/>>, FrameNet: <<https://framenet.icsi.berkeley.edu/fndru pal/>>, VerbNet: <<http://verbs.colorado.edu/>>, and WordNet: <<https://wordnet.princeton.edu/>> (all accessed April 15, 2017). For WordNet, see also Miller (1995).

90 The fourteen verbal categories are included in the WordNet lexicographer file as numbers 29 to 42 (<<https://wordnet.princeton.edu/documentation/lexnames5wn>> [accessed, April 25, 2017]). On cognitive linguistics, see also in this volume, Svärd, Jauhainen, Sahala, and Lindén, 230, who proposed another approach applied to Akkadian semantics.

91 Langacker and Vandeloise 1991, 103.

```

<category n="1" xml:id="humiliation" ana="#verb.emotion">
  <catDesc>subcategory of emotion's verb as a concept of:
humiliation
  <term ana="#mḥṣ02" type="baseForm">
    <ptr n="1" target="http://babelnet.org/
synset?word=bn:00086117v" source="BabelNet" />
    <ptr n="2" target="http://wordnet-
rdf.princeton.edu/id/01804206-v" source="WordNet" />
    <ptr n="3"
target="http://verbs.colorado.edu/propbank/framesets-eng-
lish-aliases/humiliate.html" source="VerbNet" />
    <ptr n="4"
target="http://verbs.colorado.edu/html_groupings/wound-
n.html" source="VerbNet" />
    <ptr n="5"
target="https://framenet2.icsi.berkeley.edu/fnReports/data/
frameIndex.xml?frame=Stimulate_emotion" source="FrameNet" />
  </term>
</catDesc>
<category ana="#verb.emotion #humiliation
" xml:lang="uga">
  <gloss n="1" xml:id="mḥṣ02" cert="high"/>
</category>
</category>92

```

For the following transcription example, I will take the verb *mḥṣ*, which means “to destroy” or “to fight.” Our understanding of “destroy” depends on the verbal semantic categories to which it belongs—more specifically, the categories of

92 Additional explanation for the elements and attributes within the elements <category> and <catDesc>: @n="8" indicates the eighth category of the semantic verbal categories. @n="1" refers to the first subcategory of this eighth category. Within the element <ptr>, each attribute @target points to (the destination of) the lexical and/or semantic website that is used to refer to a semantic category. The attribute @source indicates the semantic database's name to which @target points. The value “uga” of the @xml:lang specifies that Ugaritic is the reference language. The element <gloss> means that the verb *mḥṣ* (02, second meaning) has shared affinities (in the sense of cognitive meanings) with the @xml:id="humiliation" subcategory. I indicate that its level of certitude, according to my evaluation, is high (@cert="high").



competition and humiliation.<sup>93</sup> In the current example, the occurrence of “mhš” is related to the humiliation concept (verb.emotion), to which I will return:<sup>94</sup>

```
<l n="7" xml:id="ktul-3_ii_17" ana="#ktul-3_ii_17_int">
  <w pos="verb" ana="#mhš02 #yQTL #verb.emotion #humiliation
#ANT" xml:id="ktul-3_ii_17_tmhš" lemmaRef="../uga/verb.
xml#mhš">tmhš</w>
  <g>.</g>
  <w pos="noun" lemmaRef="uga/noun.xml#l'im">l'im</w>
  <g>.</g>
  <w pos="noun" lemmaRef="uga/noun.xml#hp">hp<gap
unit="chars"/></w>
  <lb/><w pos="noun" lemmaRef="uga/noun.xml#ym">y<damage
agent="unknown"><supplied resp="KTU">m</supplied></damage>
</w><gap unit="chars"/>
</l>95
```

93 “To humiliate” is defined as “to reduce (someone) to a lower position in one’s own eyes or others’ eyes” (*Merriam-Webster*, s.v., “humiliate,” <<https://www.merriam-webster.com/dictionary/humiliate>> [accessed July 29, 2017]).

94 Located within /teiCorpus/text/body/div1/div2/div3/div4/lg/1—the number of <div4> may vary since I have not completed my study of all the previous tablets.

95 The attribute <lemmaRef> and its value refer to an open-source Ugaritic semantic dictionary that am I currently developing. It will be specific to the corpus of the scribe ʾIlmilku. Each occurrence of a verb that refers to one or several interpretations has a short grammatical analysis—once again, *mhš* as a case studied (for additional information see Budin, Majewski, and Mörth, 2012, <<http://journals.openedition.org/jtei/522>> [accessed April 1, 2017]):

```
<entryFree n="6" xml:id="mhš">
  <form type="verb">
    <orth>mhš</orth>
  </form>
  <form type="inflected">
    [other forms of mhš]
    <w n="2" lemma="tmhš" xml:id="tmhš" ana="#mhš">
  <m type="base">
  <m type="pref" ana="#aff-pref.t">t</m>
  <m type="baseform">mhš</m>
</m>
```

---

```

    <gramGrp type="baseform" ana="#mḥṣ">
    <gramGrp>
    [other forms of mḥṣ]
    <gramGrp ana="#tmḥṣ">
        <iType ana="#stem.D" value="D" type="semantic-
variations"/>
        <mood ana="#mood.ind" value="ind"/>
        <tns ana="#tns.perf" value="perf"/>
        <subc ana="#prop.trans" value="trans"/>
        [just an example]
        <gramGrp n="1.1" ana="#actor-affixes"/>
            <per ana="#pers.s3" value="3"/>
            <gen ana="#gen.fem" value="f"/>
            <number ana="#num.sg" value="sg"/>
        </gramGrp>
        <cit ana="#tmḥṣ">
            <quote n="1" xml:id="tmḥṣ01">
                <ref target="corpus_ilimilku.xml#ktul1-
3_ii_17_tmḥṣ">ktul.3:ii:17</ref>
            </quote>
        </cit>
    </gramGrp>
</form>
[...]
```

<sense n="1" ana="#mḥṣ #tmḥṣ01" xml:id="mḥṣ01 xml:lang="en">to  
 fight</sense>
 <sense n="2" ana="#mḥṣ #tmḥṣ01" xml:id="mḥṣ01 xml:lang="en">to  
 destroy</sense>
 [other senses of mḥṣ]
 <re n="2" ana="#tmḥṣ #mḥṣ02" type="inflected">
 <sense>she destroys</sense>
 <span type="interp">hermeneutics
 <ref
 target=" ../computation/corpus\_ilimilku.xml#ktul1-
3\_ii\_17\_int">(1)
 </ref>
 </span>
 <span ana="corpus\_ilimilku.xml#humiliation"
 type="category">taxo., subcat. Of emotion's v. as a  
 concept of <ref
 target="corpus\_ilimilku.xml#mḥṣ02">humiliation
 </ref>

I will not go into details, since this is not the main topic of this paper, although the following text mining also relies on transcription and pre-processing. Rather, I will focus on the second line, starting with the element `<w>` with the `@ana="#mhş02"`. In this example (`@xml:id="ktu1-3_ii_17"`),<sup>96</sup> the hypothesis about the concept of humiliation was suggested by following the contextualized examples from the four databases, compared to the context from KTU 1.3:ii:17 as similar as possible.

Further to the previous example:

line 5b–6a: *w hln 'nt tmhş b 'mq* (then, 'Anatu fights in the valley)

line 7 (previous example): *tmhş lim hp ym* (she destroys the clan [coming] from the sea-shore).

In both lines, the verb *mhş* is used to describe 'Anatu's performance. But there are minor variations of meanings that I have seen following the same process: *mhş*, line 5b–6a, belongs to competitive subcategory.

Although I agree with Ronald Langacker, who suggests that a semantic category does not lie in an objective reality,<sup>97</sup> I think the affiliation of a specific verb to a semantic category can be readily demonstrated, at least, based on a substantial majority of objective components, related to other objective categories.

Category: Context (`@xml:id="context"`)

I am not considering the context of writing but in what context the action is or was performed. Concept has an important scope for helping both in the assignment of verb semantics and in interpretation. By saying this, each action needs to be contextualized according to the pragmatics evidenced either by an explicit word or by the utterance itself. For example:

---

```

    </span>
  </re>
</entryFree>

```

96 My translation: "She destroys the clan [coming] from the seashore." In my opinion, the psychological dimension prevails over the physical context. This assumption is further developed in my forthcoming work on the hermeneutics of action.

97 Langacker and Vandeloise 1991, 107.

line 23: *mid tmthšn w t'n* (they vigorously fight, and she looks).

The verb *mḥš* is under its inflectional form “t-m-t-ḥš-n” (prefix /t-/, infix /-t/, suffix /-n/).<sup>98</sup> Previous lines tell us that a fight has been organized between two warriors in the courtyard of a palace. By deduction, ‘Anatu (she) is looking (*t-n*) to a single combat.

Its non-exhaustive list of values is proposed according to different contexts in the studied text—in others words, contexts may be different according to the contexts found in a text: assembly, battle, burial, care, feast, hunt, landscaping project, lawsuit, maintenance, ritual, single combat, trade, visit, wedding.

Category: Result (@xml:id=“result”)

All action leads to a result distinct for the action itself.<sup>99</sup> Generally, a result is distinctly identified within the group of words to which the verb belongs or in the following lines. According to the action-oriented solution, the actantial event precedes a result. Of course, this list is non-exhaustive: death of opponent, injury of opponent, trauma, defeat of the opposition, deportation, reach household, regaining strength, take over the throne.

Following the example of lines 5b–6a (‘Anatu fights in the valley), the result was clearly given in lines 7 and 8, where both verbs belong to the humiliation subcategory, which means the defeat of the opposition.

Category: Sphere (@xml:id=“sphere”)

By considering a sphere, my interest focuses first on interactions within a political context—since the main story is about a fight between two clans—and to what extent these interactions can interfere within public and private relationships in regards of power and authority. As explained by Remigiusz Rosicki:

In the political context, it is attempted to connect sphere with a description of political phenomena, power, violence, force, the sphere of freedom, etc.<sup>100</sup>

The agent behavior is different according to whether the event takes place inside or outside a sphere that can be both a political and cultural center of

<sup>98</sup> See note 84.

<sup>99</sup> Massin 2014, 97.

<sup>100</sup> Rosicki 2012, 11.

power (such as a palace or temple). The palace can be owned either by the agent or a close member of the clan to which he or she belongs. In this manner, I am considering a spatial delimitation, where the agent performs an action to suggest his or her sphere of influence within a political context, more precisely with regard to deontic powers.<sup>101</sup> Since I am using a spatial criterion, the tag <location> is appropriate. Finding the sphere is easy, either thanks to the inflectional ending of a noun or the geographical indication:

line 3b-4a: *klt tgrt bht 'nt* ('Anatu closes the doors of her temple).

In the first example, *bht* means "her temple" (*bh*),<sup>102</sup> and the inflectional ending /-t/ is a personal pronominal suffix. Here 'Anatu is performing the action inside her sphere.

line 6b: *tḥtšb bn qrytm* (Once again, she strongly fights the sons of Ugarit).<sup>103</sup>

'Anatu is performing the action outside her sphere.

Category: Role (@xml:id="role")

In what manner is the AE performing the action? It either active or passive.

Coming back to deontic powers, looking at the role of an AE is relevant to investigating whether she has the right to be active within a spatial delimitation, according to a context. The same question is also useful for the character

101 "Deontic powers," as described by John Searle (1995, 133–134), are political powers that structure relationships between individuals who evolve in an institutionalized society, thus relying on conventional power. These individuals are called "agent" (Y, X) and belong to two major categories of deontic properties (or deontic status-functions): (1) "right" (positive power) where agents (Y) are vested with powers, (2) "duties" (negative power) where agents (X) have an obligation to act according to the powers of agents (Y). As stated by Searle, conventional power does not preclude physical power. For further consideration, see Juloux (2016, 134).

102 If we accept 'Anatu's ritual performance as preparation for a battle (KTU 1.3.ii:2–3a), translating *bht* as "her temple" seems appropriate, especially if we consider the gold cup RS 5.031 representing a hunting action found in the temple of Ba'lu (Yon 2006, 165; Vidal 2007, 710). In other words, this is a symbolic image of an armed hero leading a battle.

103 *qrt* means either the toponym *Qaritu* or "the City," with implicit understanding that the city is "Ugarit." The suffix /-y/ indicates the gentilic (*qrt-y* is also attested; see Huehnergard [1987, 239]); thus, with the plural suffix /-m/, one can translate by the inhabitants of the city of 'Ugarit, translation that I am more inclined to consider. For additional information regarding gentilic and *qrt-y-m*, see Soldt (2005, 39). Also see Smith (2012, 112–117).

following the verb (complement). Does she have the obligation to endure the action of the AE actant? This is particularly relevant in such cases as in the context of an assembly where political issues are discussed in KTU 1.2.i:21–47.

Category: Biological Sex (@xml:id="Genetic-sex")

As previously explained, I have chosen not to differentiate the origin of the character, either divine or human. By saying this, I apply the sexual distinction in a broader sense ("male" or "female").

Biological sex is innate; gender role is acquired through experience. The latter can be suggested by using the agency's approach during the hermeneutics of action process. Obviously, my ultimate concern is not the sexual distinction but rather the gender-related role, which relied on social constructions. Investigating the character agency then becomes relevant in order to suggest a social role gendered "male" or "female." This investigation is based on the number of occurrences and not based on gender-marked stereotypes or conventional codifications that rather fall under Judeo-Christian heritage in Western societies—a point that is especially relevant for power-relationship studies.<sup>104</sup> These conventional codifications created expectations for masculine and feminine behavior, as we are reminded by Jean Lipman-Blumen.<sup>105</sup> In fact, it has affected previous interpretations of a character's action, and it is not so easy to change

104 "Professional and popular historians in these fields continued to carry forward pre-modern Judeo-Christian and Classical historiographic traditions for generations" (Richardson 2011, 12). For a simple example of gender-marked stereotypes, consider the main actors in a television commercial: house-cleaning is (still) mostly cast as a feminine task, while driving a truck is more often cast as a masculine task. Returning to conventional codification in Ugaritic literature, Murphy (2010, 532) cautions us: "Anat's role in the Ba'al Cycle should not be assumed to have something to do with fertility simply because she is female." On secular institutions and their roles in gender systems within power relationships, see Lipman-Blumen (1984, 13–17). Regarding conventional codifications, my assumption is that they can be attributed to the "cultural lag" described by Lipman-Blumen (1984, 53–64).

105 Lipman-Blumen 1984, 2: "Gender roles ... are socially created expectations. Exaggerating both real and imagined aspects of biological sex, each society sorts certain polarized behaviors and attitudes into two sets it then labels 'male' and 'female.' Gender roles are social constructions; they contain self-concepts, psychological traits, as well as family, occupational, and political roles assigned dichotomously to members of each sex. For example, the traditional female gender role includes expectations for females to be passive, nurturant, and dependent. The standard male gender role incorporates alternative expectations—behavior that is aggressive, competitive, and independent. Women as mothers, nurses, and teachers, men as doctors, generals, and legislators are part of this pattern." See also Lipman-Blumen (1984, 21–54).

these interpretations. One way to help, however, is to present data—that is, to count the number of verbal occurrences and then to check the biological sex of the character who is performing the actions (either with name or prefix/suffix of the inflectional form of the verb). This also helps to analyze the agency's behavior of this character.<sup>106</sup>

### *Subjective Variables and its Semantic Values*

Category: Consequence (@xml:id="consequence")

The consideration of consequence is based on the actantial event and the result according to three possibilities:<sup>107</sup> 1) effect on the AE, 2) effect on the AE and other(s), or 3) effect on only the other(s). This analysis is relevant for the political scope (power and authority) of the AE, as well as his or her involvement in his or her performance to act. This evaluation is suggested by stage indications as well as by previous and future anecdotal evidence of the result of an event. The death of a warrior caused by the AE can have a consequence toward the AE, the clan of the AE, and the clan of the opponent; it is particularly true for the result of a single combat.

Category: Emotion and its Strength (@xml:id="Emotion  
Category", @xml:id="hasEmotionIntensity")

An emotion may fall within a range of subjectivity while being conceptualized, since we need to find what Robert Plutchik calls a "stimulus event."<sup>108</sup> This stimulus event in turn brings about an emotion. For example, a war's context brings out a trigger for anger that leads to a desire to kill. When observing this desire in you, your enemy wants to kill you to protect himself or herself. Observing that desire in them to kill you makes you want to protect yourself from death.

Emotion is the result of a mindful judgment when facing a situation,<sup>109</sup> and its intensity may be caused by a domino effect. The judgment of an AE is based

<sup>106</sup> Juloux 2016b, 131 n. 58.

<sup>107</sup> Located within /teiCorpus/TEI/teiHeader/encodingDesc/classDecl/taxonomy. For guidelines about subjective variables, see Bigot Juloux (2017, <<https://vbigo-juloux.github.io/hermeneutics-of-action/UserManual/out/webhelp/index.html#SubjVar.html>> [accessed April 15, 2017]).

<sup>108</sup> Plutchik 1980, 4–5, 11.

<sup>109</sup> Lyons 1985, 72, 85. Livet 2004, 136: "Emotions can thus be seen as linked to situations in which we are led to reassess our beliefs, expectations, and even our preferences." For Anna Wierzbicka, "emotions are founded on beliefs: in an emotion, one feels something similar to what one normally feels when one has such beliefs. In the revision-based analysis of emotions, beliefs correspond to the information that survives the revision process."

on her objective or subjective appraisal—even if irrational.<sup>110</sup> Of course, the emotion is often not explicitly stated; nonetheless, the six elements (indicated below as a thought of) relating to an emotion, as defined by Johnny Fontaine, Klaus Scherer, Etienne Roesch, and Phoebe Ellsworth, act as indices to define an emotion:

- (a) appraisals of events, (b) psychophysiological changes (bodily sensations), (c) motor expressions (face, voice, gestures), (d) action tendencies, (e) subjective experiences (feelings), and (f) emotion regulation.<sup>111</sup>

Their method enables us to take into account empirical considerations related to the human experience. The description of an emotion therefore relies on empirical observation, as suggested also by Plutchick: “inner emotional states usually are retrospective and depend on memory.”<sup>112</sup> Thus, an actantial event and its context, the verbal pattern of an action, the adverb, the emotional component (i.e., the gesture),<sup>113</sup> and the previous utterance are among the clues for suggesting a type of emotion. Of course, the consequence of an actantial event is a relevant criterion to ascertain the agent’s (AE) emotion. Among several emotion typologies, my first choice is that described above by Fontaine et al., who have developed a typology that includes languages other than those from Western civilizations.<sup>114</sup> I have selected twenty-four emotions from their typology:<sup>115</sup> anger, anxiety, being hurt, compassion, contempt, contentment, despair, disappointment, discouragement, disgust, fear, guilt, happiness, hate, interest, irritation, jealousy, joy, love, pride, sadness, shame, stress, surprise. To this list, I have added one more emotion: satisfaction.

Emotions enable us to propose character traits that will be suggested within the hermeneutics of action according to the number of occurrences of each emotion of an AE within a specific context.<sup>116</sup>

<sup>110</sup> Lyons 1985, 78–84, 100.

<sup>111</sup> Fontaine et al. 2007, 1050. See also James (2006, 60) and Plutchik (1980, 5): “There are physical symptoms, attitudes toward oneself, impulses to action, and physiological changes.”

<sup>112</sup> Plutchik 1980, 6.

<sup>113</sup> See the template of Fontaine et al. (2007, 1052–1054).

<sup>114</sup> Fontaine et al. 2007, 1056.

<sup>115</sup> These twenty-four emotions are known as the “FSRE categories,” as they were “used in the study by Fontaine, Scherer, Roesch and Ellsworth (Fontaine et al. 2007, 1055)” (W3C, 2014, <<https://www.w3.org/TR/emotion-voc/#fsre-categories>> [accessed April 25, 2017]).

<sup>116</sup> Following the proposition of Plutchik (1980, 21).



The strength of an emotion is a matter for consideration to temper (or not) the behavior. Its evaluation is made possible by the verbal pattern of the utterance to which the action belongs, in particular for the Ugaritic D verb (also called stem II).<sup>117</sup> The intensity is expressed according to five value levels: feeble, medium, normal, high, very high. The emotion intensity enables us to judge, even if expressed irrationally, what to do later on, following here the theory of William Lyons.<sup>118</sup> This judgment, which depends upon the author's intention, enables us to comprehend the result that follows the actantial event, as well as both the desire degree and the motivation of the AE.

Category: Degree of Desire (@xml:id="degreeDesire")

According to Andrew Ortony, Gerald Clore, and Allan Collins:

Affective reactions arise when a person construes the consequences of an event as being desirable or undesirable, so that judged desirability (including undesirability) is the most important, or the central, variable that affects the intensity of all these Event-based emotions ... The structure that falls below the *pleased/displeased* node divides first according to whether the person who experiences the emotions is reacting to the consequences of the focal event with respect only to himself, or also with respect to some other person.<sup>119</sup>

Here I focus on the notion of desirability, a reaction linked to the node "pleased/displeased." In other words, an action often gives rise to an affective reaction in response both to the actantial event and its consequences for others.<sup>120</sup> The affective reaction is closely linked to the emotion. Ortony, Clore, and Collins respectively define pleased/desirable and displeased/undesirable as "pleased about a desirable event" and "displeased about an undesirable event" that is eventually quite distant from the motivation of a voluntary intentionality. I am more interested in the AE's desire to act and that desire's quantifiable intensity.<sup>121</sup>

<sup>117</sup> Stem II or *Doppelungsstamm*, in German, is also called "intensive," since it shows the intensive aspect, both qualitatively and quantitatively (Lipiński 2001, 390). The group to which the verbs belong is very important in particular for the evaluation of the character's emotion and its strength.

<sup>118</sup> Lyons 1985, 72.

<sup>119</sup> Ortony, Clore, and Collins 1998, 20.

<sup>120</sup> Ortony, Clore, and Collins 1998, 92, table 5.2.

<sup>121</sup> Ortony, Clore, and Collins 1998, 87.

Category: Voluntary Intentionality and its Degree of Motivation

```
(@xml:id="voluntaryIntent", @xml:id=
  "motivation_vI")
```

Even though this category may be considered objective, it is not obvious from the Popperian point of view. It may be understood as the intentional nature of an action.

## Text Mining

Having defined analytical taxonomies (primary data, objective and subjective variables), one must next apply data processing in order to discover useful information.<sup>122</sup> The method used is “text mining” which, as defined by Marti Hearst, is

the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources ... Text mining is a variation on a field called data mining, that tries to find interesting patterns from large databases.<sup>123</sup>

The purpose of this text mining is to find patterns that match shared analytical variables in order to suggest a first interpretation of the transcription. As a preliminary step in the hermeneutics of the action,<sup>124</sup> it follows a structural unit that provides some clear answers to seven key issues: (1) what action, (2) what result, (3) what character, (4) what role, (5) what context, (6) what sphere, and (7) what behavior. Roberto P. Franzosi has demonstrated different disciplinary backgrounds, which are similar to mine in some ways:

a common understanding of narrative and social action in terms of agents (Who) and actions (What) in time (When) and space (Where), for some reasons (Why) and with certain outcomes and instruments.<sup>125</sup>

<sup>122</sup> “Data processing” is also called “data analysis.”

<sup>123</sup> Hearst 2003, <<http://people.ischool.berkeley.edu/~hearth/text-mining.html>> (accessed April 12, 2017). For additional information, see Kumar and Bathia (2013).

<sup>124</sup> Located within `teiCorpus/text/body/div1`. For guidelines about text mining, see Bigot Juloux (2017, <<https://vbigot-juloux.github.io/hermeneutics-of-action/UserManual/out/webhelp/index.html#TextMining.html>> [accessed April 15, 2017]).

<sup>125</sup> Franzosi 2010, 600.

Franzosi has designed a method that is based on story grammar and uses semantic triplets.<sup>126</sup>

Coming back to my approach, each element of the structural unit refers to a category of the analytical taxonomies:

```
<interpGrp xml:id="hermeneutics" type="structuralUnit"
  ana="#taxonomies">
  <interp n="1" xml:id="whatAction" ana="#primaryData #ob-
    jectiveVar #action #verb.category"/>
  <interp n="2" xml:id="whatResult" ana="#objectiveVar
    #result"/>
  <interp n="3" xml:id="whatCharacter" ana="#primaryData
    #objectiveVar #Character #Genetic-Sex"/>
  <interp n="4" xml:id="whatRole" ana="#objectiveVar
    #role"/>
  <interp n="5" xml:id="whatContext" ana="#objectiveVar
    #context"/>
  <interp n="6" xml:id="whatSphere" ana="#objectiveVar
    #sphere"/>
  <interp n="7" xml:id="whatBehavior" ana="#subjectiveVar
    #consequence #EmotionCategory #hasEmotionIntensity #de-
    greeDesire
    #voluntaryIntent"/>
</interpGrp>
```

The attributes @xml:id and @ana are of utmost importance, both for the parsing in R and the preliminary hermeneutics of the action within the elements <interp> </desc>.<sup>127</sup> To explain the text mining clearly, I think it is most efficient to give a specific example:

```
<interp xml:id="ktul-3_ii_17_int" ana="#ktul-3_ii_17">
```

126 Franzosi 2010, 602: "The relational properties of the grammar (e.g., with actors related to actions, actions to time and space and objects) make a story grammar a far more superior tool than content analysis, the traditional quantitative social science approach to texts." The semantic triplets used by Franzosi look like this: <semantic triplet> → {<participant>}, {<process>}; [{<participant>}]. For further explanation of the syntax and values within the angular brackets, see Franzosi (2010, 601).

127 R: an open-source software environment, mostly used for statistics, graphics, and data manipulation. One of the major benefits of R is its popularity, which guarantees regular developments by a broader community of R users. See <<https://www.r-project.org>> (accessed April 25, 2017).

```

<desc>
  <ref n="1" target="#whatAction #ktu1-3_ii_17_
tmḥṣ" ana="#verb.emotion #humiliation #mḥṣ02"/>
  <ref n="2" target="#whatResult"><stage
ana="#defeat_ofOpposition"/>
  <castList>
    <castItem n="1">
      <ref n="3" target="#whatCharacter">
        <persName type="character" ana="#ANT #Female"/>
      </ref>
      <ref n="4" target="#whatRole">
        <state ana="#active" cert="high"/>
      </ref>
    </castItem>
    <castItem n="2">
      <persName type="character" ana="#UNK
#Unknown_Sx" cert="low">
        Clan (coming) from the sea-shore
      </persName>
      <state ana="#passive" cert="medium"/>
    </castItem>
  </castList>
  <view>
    <ref n="5" target="#whatContext">
      <placeName ana="#battle"/>
    </ref>
    <ref n="6" target="#whatSphere">
      <location ana="#outside"/>
    </ref>
  </view>
  <stage>
    <ref n="7" target="#whatBehavior">
      <span ana="#toDestroy #free #five_dD"/>
      <span ana="#affectEntity_and_other"/>
    </ref>
  </stage>
</desc>
</interp>128

```

<sup>128</sup> <ref n="1">: Action belongs to verb emotion, subcategory humiliation: mḥṣ. <ref n="2">: defeat of opposition. <ref n="3">: 'Anatu. <ref n="4">: role: active. <ref

Each `<interp>` has a `@xml:id` and `@ana`, which refers to a group of words of the transcription within `<text>/<div>/<lg>/<l>`. I focus, in this example, on the line KTU 1.3:ii:17.<sup>129</sup>

Each taxonomic category is analyzed within the tag `<ref>` in which the pointer `@target` refers to an item for the structural unit, as this is the case for `<ref n="1">`; it gives semantic information for the verb of the referenced line.

The elements `<stage>`, `<castList>`, `<castItem>`, and `<view>` come from the vocabulary related to the dramatic aspect of the narrative text, which enables us to improve the contextualization of the actantial event.

The verb *mḥš* in the current example, under its inflectional form *tmḥš*, is an emotion verb—more precisely, one of humiliation as a conceptualized approach.<sup>130</sup> The result of this action is the defeat of the opponent. The feminine gender of the actant subject ‘Anatu is confirmed by the inflectional form of the verb. ‘Anatu is active and is facing the Western clan, which seems to suffer from the action, according to the author’s will. The actantial event takes place during a battle outside of ‘Anatu’s household. Her behavior characterizes a voluntary intentionality of destruction with a free will and a high level of rage toward the opponent. The consequence of this action affects ‘Anatu and the others, notably the clan of the West.

Although I will not go into computational details for the pre-processing step with R in order to refine methods for the quantification process, here a short example highlights the relevance of the structural unit for the last step of the preliminary of a hermeneutics of action:

```
listInterp=matrix(nrow=20,ncol=9)
colnames(listInterp)=c("Character", "TAXO", "subTAXO",
  "Role", "Context", "Sphere", "Behavior",
  "Consequence", "KTU")
for (i in 1:length(whatCharacter))
  listInterp[i,1]=word(xmlGetAttr(whatCharacter[[i]], "
  ana"), 1)
for (i in 1:length(whatActionSem))
```

---

`n="5">`: battle. `<ref n="6">`: outside her household. `<ref n="7">`: (a) to destroy of her free will, with rage (level five), (b) and consequence: an impact on ‘Anatu and others.

129 See my previous transcription, page 172.

130 The translation of *mḥš* is “to destroy,” which I understand first as the humiliation of an opponent, and, second, according to the previously introduced semantic analysis, a discussion of which follows.

```

listInterp[i,2]=word(whatActionSem[[i]], 1)
for (i in 1:length(whatActionSem))
  listInterp[i,3]=word(whatActionSem[[i]], 2)
for (i in 1:length(whatRole))
  listInterp[i,4]=xmlGetAttr(whatRole[[i]], "ana")
for (i in 1:length(whatContext))
  listInterp[i,5]=xmlGetAttr(whatContext[[i]], "ana")
for (i in 1:length(whatSphere))
  listInterp[i,6]=xmlGetAttr(whatSphere[[i]], "ana")
for (i in 1:length(whatBehavior))
  listInterp[i,7]=xmlGetAttr(whatBehavior[[i]], "ana")
for (i in 1:length(whatConseq))
  listInterp[i,8]=xmlGetAttr(whatConseq[[i]], "ana")
for (i in 1:length(whatAction))
  ListInterp[i,9]=(stri_replace_all_fixed(word
(xmlGet
  Attr(whatAction[[i]], "target"), -1), "_", ":"))131

```

The results of the pre-processing steps are summarized in templates (e.g., Appendix 5.1 for 'Anatu's actions), and ordered in the following way: data (@ana or @target) from each objective and subjective variable within the text mining are displayed by row (@xml:id of the transcription = a line of κTU) and column (colnames, a total of nine columns), according to each @xml:id of the structural unit and the line's text reference (κTU). Among other parsings, this template is used to count the occurrences of each variable. Afterward, one can proceed to some statistics for each character (AE) versus all characters (Being) and verbs (Action)—for example, one could quantify the verbs of emotion for each AE by context and sphere, as well as where the actant is active.

### Concluding Remarks

In this paper, I have attempted to show the usefulness of analytical taxonomies and text mining with TEI-XML as the first two preliminary steps for a new hermeneutics of action, in order to investigate the actions between AE using

<sup>131</sup> listInterp is the variable assigned to data within the template (matrix). I am currently developing a guide for parsing with R.

Popperian deduction. Of course, tagging words is time-consuming, but analytical markup makes data analysis easier with R, in order to manipulate the data (`@xml:id`, `@ana`) in many ways, to determine, for example, (1) how many times a character  $AE_x$  is performing an action in a specific context and sphere versus all AE and all actions?, or (2) what type of action (semantic category) is  $AE_x$  performing, in what context and sphere, with what result and consequence versus all characters AE? Afterward we can identify the distribution of action types for  $AE_x$  by increasing order versus all other AE.

In this project, a qualitative method was first used to sort relevant variables by category. Next, text mining was employed as an intermediate qualitative stage to enable parsing with R. This should not only make it possible to quantify occurrences related to the analytical taxonomies but also to suggest one or several authors' intention, thanks to his choice of occurrences according to the objective and subjective variables. The evidence of intention will also rely on the agency of a character, which will be displayed after the quantification, as well as an action between AE, especially according to a context and a sphere. So the first question will be: for  $n$  actancial events related to a peculiar subcategory of verbal semantics,<sup>132</sup> how many times did the actant achieve an action with free will, and in what context, sphere, and consequences?

A second concern is the question of whether the author's choices in the character of 'Anatu's characteristics could have been derived from a real-life experience, or whether they could be a reflection of gender roles in ancient Ugarit.<sup>133</sup>

These results and considerations, combined with the empirical observation of intersubjective phenomena, will enable us to suggest a hermeneutics of action for each AE, especially 'Anatu. But above all, this approach should be relevant for highlighting the weakness of the (un-)shared analytical variables that would have been preferable to support the consistency of previous interpretations—in particular, regarding the role of a female entity and her influence on others (including a community), notably from social and cultural anthropology's point of view.

To push the analysis further, and regarding the intersubjective phenomena and mental realities touched upon here in the introduction, it would be interesting to apply the same approach to historical texts (i.e., annals, chronicles) for counting semantic verbal occurrences for each actant (historical figure)

<sup>132</sup> Where  $n$  is a numerical value that is equal to the real actancial events' number.

<sup>133</sup> The answer to this question can provide a complementary perspective to Christine Neal Thomas's (2013) dissertation on gender roles at Ugarit.

according to a context and a sphere, in order to gain a better understanding of codification used by the elites, especially for political purposes—this falls within the analysis of a cognitive psychology of the elites, regarding their power and authority. By doing so, it would then be easier to speak of an objective interpretation of narrative texts according both to empirical context and to the codification used by the elites, who themselves authored the narratives.



Appendix 5.1. An Overview of ‘Anatu’s Actions

TAXO	subTAXO	Role	Context	Sphere	Behavior	Consequence	KTU
verb.contact	closing	active	battle	inside	toFight free unknown_E_NA_E unknown_dD	affectEntity_and_other	ktu1-3:iii:3b-4akl'at
verb.motion	meeting	active	unknown_C	outside	unknown_Vol free unknown_E_NA_E unknown...	affectEntity_and_other	ktu1-3:iii:4b-5atqry
verb.competition	contend	active	battle	outside	toDestroy free veryHight rage five_dD	affectEntity_and_other	ktu1-3:iii:5b-6atmthş
verb.competition	contend	active	battle	outside	toDestroy five_dD rage	affectEntity_and_other	ktu1-3:iii:6b:thtşb
verb.emotion	humiliation	active	battle	outside	ToDestroy free five_dD	affectEntity_and_other	ktu1-3:iii:7:tmhş
verb.emotion	humiliation	active	battle	outside	ToDestroy free rage five_dD	affectEntity_and_other	ktu1-3:iii:8:tşmt
verb.contact	attaching	active	battle	outside	ToDestroy free contentment five_dD	affectEntity_and_other	ktu1-3:iii:11b-12a:tkt
verb.contact	attaching	active	battle	outside	toDestroy free contentment five_dD	affectEntity_and_other	ktu1-3:iii:12b-13a:şnst
verb.contact	placing	active	ritual	outside	toPerform_aRitual free pride five_dD	affectEntity_and_other	ktu1-3:iii:13b-14a:tğll
verb.contact	removing	active	battle	outside	toThreaten free contempt five_dD	affectEntity_and_other	ktu1-3:iii:15b-16a:trş
verb.motion	self_motion	active	battle	outside	toProve free pride five_dD	affectEntity_and_other	ktu1-3:iii:17:tmğyn
verb.motion	arriving	active	battle	inside	toProve free pride five_dD	affectEntity_and_other	ktu1-3:iii:18:tştlql
verb.state	satisfaying	active	battle	inside	toComplain free unsatisfaction five_dD	affectEntity	ktu1-3:iii:19:şb't
verb.contact	placing	active	battle	inside	toKill free unsatisfaction five_dD	affectEntity_and_other	ktu1-3:iii:20b-21a:t'tr
verb.competition	contend	active	singleCombat	inside	toRevenge free happiness five_dD	affectEntity	ktu1-3:iii:24:thtşb
verb.state	fill_withEmotion	active	singleCombat	unknown_S	toKill free interest five_dD	affectEntity	ktu1-3:iii:25a:tdğdd

## References

- Anscombe, Gertrude E. M. 2002. *L'Intention*. Paris: Gallimard.
- Audi, Robert. 2003. "Acting for Reasons." In *The Philosophy of Action*, edited by Alfred R. Mele, 74–105. Oxford: Oxford University Press.
- Berthelot, Jean-Michel. 2004. *Les vertus de l'incertitude*. Paris: PUF.
- Bigot Juloux, Vanessa, and Alessandro di Ludovico. 2018. "Digital Practices vs. Digital Humanities: Reflections to Bridge the Gap in Order to Improve Research Methods and Collaboration." Paper presented at the CAA annual meeting, Tübingen, Germany.
- Budin, Gerhard, Stefan Majewski, and Karlheinz Mörrh. 2012. "Creating Lexical Resources in TEI P5." *JTEI* 3. <<http://journals.openedition.org/jtei/522>>.
- Bühler, Axel. 2015. "Intention, Intentionnalisme." In *L'interprétation: Un dictionnaire philosophique*, edited by Christian Berner and Denis Thouard, 235–247. Paris: Vrin.
- Capurro, Rafael. 2010. "Digital Hermeneutics: An Outline." *AI Society* 35 (1): 35–42.
- Coombs, James H., Allen H. Renear, and Steven J. DeRose. 1987. "Markup Systems and the Future of Scholarly Text Processing." *Communications of the ACM* 30: 933–947.
- Davidson, Donald. 2008. *Actions et événements*. Paris: PUF.
- Day, Peggy L. 1991. "Why Is Anat a Warrior and Hunter?" In *The Bible and the Politics of Exegesis: Essays in Honor of Norman K. Gottwald on His Sixty-Fifth Birthday*, edited by David Jobling, Peggy L. Day, and Gerald T. Sheppard, 141–146. Cleveland, OH: The Pilgrim Press.
- Dietrich, Manfred, Oswald Loretz, and Joaquín Sanmartín. 2013–. *Die Keilalphabetischen Texte aus Ugarit, Ras Ibn Hani und anderen Orten*. 3rd ed. AOAT 360 (1). Münster: Ugarit-Verlag.
- Dutcher, Jennifer. 2014. "What is Big Data?" Online Data Science Degree blog, Berkeley School of Information. Last modified September 3, 2014, <<https://datascience.berkeley.edu/what-is-big-data>>.
- Englehardt, Joshua. 2013. *Agency in Ancient writing*. Boulder: University Press of Colorado.
- Fensham, F. Charles. 1979. "Notes on Treaty Terminology in Ugaritic Epics." *UF* 11: 265–274.
- Fontaine, Johnny R. J., Klaus R. Scherer, Etienne B. Roesch, and Phoebe C. Ellsworth. 2007. "The World of Emotions Is Not Two-Dimensional." *Psychological Science* 18 (12): 1050–1057.
- Franzosi, Roberto P. 2010. "Sociology, Narrative, and the Quality Versus Quantity Debate (Goethe Versus Newton): Can Computer-Assisted Story Grammars Help Us Understand the Rise of Italian Fascism (1919–1922)?" *Theory and Society* 39 (6): 593–629.
- Freu, Jacques. 2006. *Histoire politique du royaume d'Ugarit*. Paris: L'Harmattan.
- Gabbay, Uri. 2014. "Actual Sense and Scriptural Intention: Literal Meaning and Its Terminology in Akkadian and Hebrew Commentaries." In *Encounters by the Rivers of Babylon*, edited by Uri Gabbay and Shai Secunda, 335–370. Tübingen: Mohr Siebeck.

- Gabbay, Uri. 2016. *Exegetical Terminology of Akkadian Commentaries*. Leiden: Brill.
- Gens, Jean-Claude. 2006. *La logique herméneutique du XVII<sup>e</sup> siècle*. Argenteuil: L'association Le Cercle Herméneutique.
- Gray, John. 1965. *The Legacy of Canaan*. Vetus Testamentum Supplements 5. Leiden: Brill.
- Gray, John. 1979. "The Blood Bath of the Goddess Anat in the Ras Shamra Texts." *UF* 11: 315–324.
- Greenstein, Edward L. 2006. "Forms and Functions of the Finite Verb in Ugaritic Narrative Verse." In *Bible Hebrew in its Northwest Semitic Setting*, edited by Steven E. Fassberg and Avi Hurvitz, 75–102. Jerusalem: The Hebrew University Magnes Press.
- Greimas, Algirdas J. 1987. *On Meaning: Selected Writings in Semiotic Theory*. Theory and History of Literature, vol. 38. Minneapolis: University of Minnesota Press.
- Grevisse, Maurice. 1986. *Le bon usage*. Paris: Ducolot.
- Grice, Herbert Paul. 1975. "Logic and Conversation." In *Syntax and Semantics 3, Speech Acts*, edited by Peter Cole and Jerry Morgan, 41–58. New York: Academic Press.
- Grondin, Jean. 2006. "La tâche de l'herméneutique dans la philosophie ancienne." In *Klêsis* 1: 1–18.
- Harari, Yuval N. 2015. *Sapiens: Une brève histoire de l'humanité*. Paris: Albin Michel.
- Hawley, Robert. 2008. "On the Alphabetic Scribal Curriculum at Ugarit." In *Proceedings of the 51st Rencontre Assyriologique Internationale, Held at the Oriental Institute of the University of Chicago, July 18–22, 2005*, edited by Robert D. Biggs, Jennie Myers, and Martha T. Roth, 57–67. Chicago: The Oriental Institute of the University of Chicago.
- Hearst, Marti. 2003. "What is Text Mining?" Last modified October 17, 2003, <<http://people.ischool.berkeley.edu/~hearst/text-mining.html>>.
- Hentrich, Thomas. 2001. "The Fertility Pair Ba'al and 'Anat in the Ugaritic Texts." In *Recherches canadiennes sur la Syrie*, edited by Michel Fortin, 115–122. Bulletin of the Canadian Society for Mesopotamian Studies 36. Montréal: Musée de la Civilisation.
- Huehnergard, John. 1987. *The Ugaritic Vocabulary in Syllabic Transcription*. Atlanta: Scholars.
- Hughes, James M., Nicholas J. Foti, David C. Krakauer, and Daniel N. Rockmore. 2012. "Quantitative Patterns of Stylistic Influence in the Evolution of Literature." *PNAS* 109 (20): 7682–7686.
- Ide, Nancy, and James Pustejovsky. 2010. "What Does Interoperability Mean, Anyway? Toward an Operational Definition of Interoperability for Language Technology." In *Proceedings of the Second International Conference on Global Interoperability for Language Resources, Hong Kong, 18–20 January 2010*, edited by Alex Chengyu Fang, Nancy Ide, and Jonathan Webster. Hong Kong: City University of Hong Kong. <<https://www.cs.vassar.edu/~ide/papers/ICGL10.pdf>>.
- Ide, Nancy, and C. M. Sperberg-McQueen. 1995. "The TEI: History, Goals, and Future." *CH* 29: 5–15.
- James, William. 2006. *La théorie de l'émotion*. Paris: L'Harmattan.

- Juloux, Vanessa. 2013. "Les liens de parenté entre les divinités dans le Cycle de Ba'lu: étude de genre et analyse ethno-historique." MA thesis, École Pratique des Hautes Études (EPHE).
- Juloux, Vanessa. 2016a. "Is the Violence of 'Anatu a Criterion of Sovereign Power? Using Combined Anthropological and Philosophical Approaches to the Study of Power and Agency in the Cycle of Ba'lu." Paper presented at the Society of Biblical Literature annual meeting, San Antonio, TX.
- Juloux, Vanessa. 2016b. "Prolégomènes de l'étude des relations de pouvoir entre les entités animées dans KTU 1.1–6." *RANT* 13: 123–164.
- Juloux, Vanessa. 2017. *Guidelines for a Hermeneutics of Action*. <<https://vbigot-juloux.github.io/hermeneutics-of-action/UserManual/out/webhelp/index.html>>.
- Juloux, Vanessa. Forthcoming. "Herméneutique de l'action pour l'étude des relations entre les entités animées et leur *agency* au Proche-Orient ancien: Hypatia et alii." In *Proceedings of the 61st RAI, Geneva and Bern, 22–26 June 2015*.
- Kapelrud, Arvid S. 1969. *The Violent Goddess. Anat in the Ras Shamra Texts*. Oslo: Universitetsforlaget.
- Karkajian, Lourik. 1999. "La maisonnée patrimoniale divine à Ougarit: Une analyse webérienne du dieu de la mort, Mot." PhD diss., Université de Montréal.
- Koch, Ulla S. 2010. "Three Strikes and You're Out! A View on Cognitive Theory and the First Millenium Extispicy Ritual." In *Divination and Interpretation of Signs in the Ancient World*, edited by Amar Annus, 43–59. Chicago: The Oriental Institute of the University of Chicago.
- Kumar, Lokesh, and Parul Kalra Bhatia. 2013. "Text Mining: Concepts, Process and Applications." *JGRSC* 4 (3): 36–39.
- Laks, André, and Ada Neschke, eds. 2008. *La naissance du paradigme herméneutique: De Kant et Schleiermacher à Dilthey*. Villeneuve d'Ascq: Presses Universitaires du Septentrion.
- Langacker, Ronald W., and Claude Vandeloise. 1991. "Noms et verbes." *Communications* 53, 103–153. <[http://www.persee.fr/doc/comm\\_0588-8018\\_1991\\_num\\_53\\_1\\_1804](http://www.persee.fr/doc/comm_0588-8018_1991_num_53_1_1804)>.
- Langendoen, Terence D., and Gary F. Simons. 1995. "A Rationale for the TEI Recommendations for Feature-Structure Markup." *CH* 29: 191–209.
- Le Ny, Jean-François. 2001. "La sémantique des verbes et la représentation des situations." *Syntaxe et sémantique* 2: 17–54.
- Lipiński, Edward. 2001. *Semitic Languages Outline of a Comparative Grammar*. Orientalia Lovaniensia Analecta 80. Louvain: Uitgeverij Peeters and Departement Oosterse Studies.
- Lipman-Blumen, Jean. 1984. *Gender Roles and Power*. Englewood Cliffs, NJ: Prentice-Hall.
- Liverani, Mario. 2011. *The Ancient Near East: History, Society and Economy*. London: Routledge.
- Livet, Pierre. 2004. "Emotion: Philosophy." In *Dictionary of Cognitive Science*, edited by Olivier Houdé, 134–137. Hove: Psychology.

- Lloyd, Jeffery B. 1994. "The Goddess Anat: An Examination of the Textual and Iconographic Evidence from the Second Millennium BCE." PhD diss., The University of Edinburgh.
- Lyons, William. 1985. *Emotion*. Cambridge: Cambridge University Press.
- Massin, Olivier. 2014. "Quand vouloir c'est faire." In *Analyses contemporaines: Recherches sur la philosophie et le langage* 30, edited by Rémi Clot-Goulard, 79–114. Grenoble: Université Pierre Mendès France.
- Mazzini, Giovanni. 2004. "Baal and Niqmaddu: A Suggestion to Ugaritic KTU 1.2 I, 36–38." *SEL* 21: 65–69.
- Miller, George A. 1995. "WordNet: A Lexical Database for English." *Communications of the ACM* 38 (11): 39–41.
- Mohr, John W., Robin Wagner-Pacifici, and Ronald L. Breiger. 2015. "Toward a Computational Hermeneutics." *Big Data & Society* 2: 1–7.
- Molinié, Georges. 2007. "Rhétorique et herméneutique." *Dix-septième siècle* 236 (3): 433–444.
- Murphy, Kelly J. 2010. "Myth, Reality, and the Goddess Anat." *UF* 41: 525–541.
- Natan-Yulzary, Shirly. 2009. "Divine Justice or Poetic Justice?" *UF* 41: 581–599.
- Nellhaus, Tobin. 2001. "XML, TEI, and Digital Libraries in the Humanities." *Libraries and the Academy* 1 (3): 257–277.
- Olmo Lete, Gregorio del. 1981. "Le mythe de la Vierge-Mère 'Anatu: une nouvelle interprétation de CTA/KTU 13." *UF* 13: 49–62.
- Ortony, Andrew, Gerald L. Clore, and Allan Collins. 1988. *The Cognitive Structure of Emotions*. Cambridge: Cambridge University Press.
- Page, Hugh R., Jr. 1998. "The Three Zone Theory and Ugaritic Conceptions of the Divine." *UF* 30: 615–631.
- Pardee, Dennis. 2012. *The Ugaritic Texts and the Origins of the West-Semitic Literary Composition*. Oxford: Oxford University Press.
- Pitard, Wayne. 1999. "The Written Source: The Alphabetic Ugaritic Tablets." In *Handbook of Ugaritic Studies*, HdO 39, edited by Wilfred G.E. Watson and Nicolas Wyatt, 46–57. Leiden: Brill.
- Plutchik, Robert. 1980. "A General Psychoevolutionary Theory of Emotion." In *Emotion Theory, Research, and Experience. Theories of Emotion*, vol. 1, edited by Robert Plutchik and Henry Kellerman, 3–33. Boston: Academic Press.
- Popper, Karl. 1998. *La connaissance objective*. Paris: Flammarion.
- Richardson, Seth F. C. 2011. "Mesopotamia and the 'New' Military History." In *Recent Directions in the Military History of the Ancient World. Publications of the Association of Ancient Historians*, edited by Lee L. Bryce and Jennifer T. Roberts, 11–51. Claremont, CA: Regina Books.
- Ricoeur, Paul. 1969. *Le conflit des interprétations: Essais d'herméneutique* 1. Paris: Le Seuil.
- Ricoeur, Paul. 1977. "Le discours de l'action." In *La sémantique de l'action*, edited by Dorian Tiffeneau, 1–137. Paris: CNRS.
- Ricoeur, Paul. 1986. *Du texte à l'action: Essais d'herméneutique* (2). Paris: Le Seuil.

- Rosicki, Remigiusz. 2012. "Public Sphere and Private Sphere—Masculinity and Femininity." In *Some Issues on Women in Political, Media and Socio-economic Space*, edited by Iwetta Andruszkiewicz and Alina Balczyńska-Kosman, 9–19. Poznań: Faculty of Political Science and Journalism, Adam Mickiewicz University.
- Schloen, John D. 2001. *The House of the Father as Fact and Symbol*. Winona Lake, IN: Eisenbrauns.
- Schmidt, Desmond. 2014. "Towards an Interoperable Digital Scholarly Edition." *Journal of the Text Encoding Initiative* 7. <<https://journals.openedition.org/jtei/979>>.
- Searle, John R. 1995. *The Construction of Social Reality*. New York: The Free Press.
- Selz, Gebhard J. 2013. "Texts, Textual Bilingualism, and the Evolution of Mesopotamian Hermeneutics." In *Between Text and Text—The Hermeneutics of Intertextuality in Ancient Cultures and Their Afterlife in Medieval and Modern Times*, edited by Michaela Bauks, Wayne Horowitz, and Armin Lange, 47–65. Göttingen: Vandenhoeck & Ruprecht.
- Smith, Mark S. 2012. "The Concept of the 'City' ('Town') in Ugarit." In *Die Stadt im Zwölfprophetenbuch*, edited by Aaron Scharf and Jutta Krispenz, 107–146. Berlin: De Gruyter.
- Smith, Mark S., and Wayne T. Pitard. 2008. *The Ugaritic Baal Cycle*, vol. 2. Leiden: Brill.
- Soldt, Wilfred H. van. 2005. *The Topography of the City-State of Ugarit*. Münster: Ugarit-Verlag.
- Sun, Chloé. 2008. *The Ethics of Violence is the Story of Aquat*. Piscataway, NJ: Gorgias Press.
- Tesnière, Louis. 1959. *Les éléments de syntaxe structurale*. Paris: Librairie C. Klincksieck.
- Thomas, Christine Neal. 2013. "Reconceiving the House of the Father: Royal Women at Ugarit." PhD diss., Harvard University.
- Unsworth, John. 2011. "Computational Work with Very Large Text Collections." *Journal of the Text Encoding Initiative* 1. <<http://jtei.revues.org/215>>.
- Vanhoutte, Edward. 2004. "An Introduction to the TEI and the TEI Consortium." *Literary and Linguistic Computing* 19 (1): 9–16.
- Vidal, Jordi. 2006. "The Origins of the Last Ugaritic Dynasty." *AoF* 33 (1): 168–175.
- Vidal, Jordi. 2007. "Ugarit at War (2): Military Equestrianism, Mercenaries, Fortifications and Single Combat." *UF* 38: 699–716.
- Walls, Neal. 1992. *The Goddess Anat in Ugaritic Myth*. SBL Dissertation Series 135. Atlanta: Scholars Press.
- Wilson, Eleanor A. 2013. *Women of Canaan. The Status of Women at Ugarit*. Whitewater, WI: Heartwell Productions.
- Wyatt, Nicolas. 2002. *Religious Texts from Ugarit*. London: Sheffield Academic Press.
- Wyatt, Nicolas. 2015. "The Evidence of the Colophons in the Assessment of Ilmilku's Scribal and Authorial Role." *UF* 46: 399–446.
- Yon, Marguerite. 2006. *The City of Ugarit at Tell Ras Shamra*. Winona Lake, IN: Eisenbrauns.

# Network Analysis for Reproducible Research on Large Administrative Cuneiform Corpora

*Émilie Pagé-Perron*

## Introduction

Although network analysis is becoming a more commonly utilized methodology in the study of Mesopotamian texts,<sup>1</sup> it is usually still employed on unique archives or small corpora. This is because there are no automated tools to reliably annotate large corpora, of which the administrative genre is the least annotated. Until we devise Natural Language Processing (NLP) tools with efficient machine learning components,<sup>2</sup> annotation requires expert knowledge and is very time-consuming. I wish to ease this burden by rethinking the ways in which we can simplify the process of extracting and analyzing relationships among entities.

Network analysis techniques can both expand Assyriological horizons in the study of Mesopotamian social history and enable research reproducibility. Network analysis permits one to approach the data in large amounts of texts through a new lens, by inquiring quantitatively with a focus on the relationships among entities of interest.<sup>3</sup> Such inquiries facilitate the detection of meaningful patterns that could not be easily perceived using traditional methods. Furthermore, when combined with practices of open access, open data, and method disclosure, network analysis research can be fully reproducible.

---

1 My deepest thanks to the editors, readers, and reviewers of earlier drafts of this chapter. A special mention is in order for Vanessa Bigot Juloux, Amy Rebecca Gansell, Terhi Nurmikko-Fuller, and Sarah Whitcher Kanza for their insightful comments and suggestions. Of course, any errors are my responsibility alone.

2 Regarding Natural Language Processing, see also in this volume, Prosser, 322, and Svärd, Jauhiainen, Sahala, and Lindén, 246. The Machine Translation and Automated Analysis of Cuneiform Languages project is working on this task (<<https://cdli-gh.github.io/mtaac/>> [accessed June 12, 2017]).

3 In textual studies, network analysis is a digital methodology that focuses on the relationships among entities in the written record and enables these relationships to be studied on a larger scale than is normally feasible with traditional philological methods.



Thus, a traditional philological approach can be supported by quantitative, reproducible, and fully verifiable arguments.

This chapter will show how network analysis can provide new insights into large amounts of data from cuneiform texts. I will demonstrate a method for the preparation and extraction of relevant data for network analysis, as well as present some techniques for graph visualization. A discussion will follow explaining how the methods themselves and the results they yield support research in Mesopotamian social history.

## Background

Legal and administrative cuneiform documents make up the largest subset of textual material from ancient Mesopotamia. These clay tablets contain essential details about the economic, social, religious, and political practices of the ancient societies that created them. Although individually these texts are short and contain little information, when grouped and analyzed as a corpus, they can offer insights into complex social topics. Despite comprising the most numerous type of surviving cuneiform documents, administrative texts are the least-annotated genre of Mesopotamian sources and are thus not prepared adequately for network analysis. Because they also remain untranslated, they have the additional drawback of being inaccessible to specialists in adjacent disciplines who are not trained to read cuneiform.<sup>4</sup> Through digital processing, however, these limitations may be mitigated.

Administrative texts mostly document transactions involving people, things, actions, and places. By using graphs and, more precisely, network theory,<sup>5</sup> scholars can examine the relationships among entities present in the text from new perspectives. The current practice in social network analysis based on cuneiform sources uses verbs to create directed links between named entities or concepts.<sup>6</sup> This requires an additional layer of analysis in which the researcher identifies not only verbs that are meaningful in the context of relationships but also who performs the action and who benefits from it. This type of analysis is manageable for smaller corpora—texts in the hundreds—but

---

4 Pagé-Perron et al. 2017.

5 Network theory: the study of systems and patterns found in network graphs. Social network analysis uses network theory to understand social relationships.

6 In this volume, see Bigot Juloux (161, 166–181), who proposes analytical taxonomies to analyze the relation between verbs and animated entities.



when a corpus has over a thousand texts, this manual operation becomes a serious investment of time.

When scholars extract information from cuneiform texts, they usually organize it in lists or tables to help with classification. In reality, when the focus of interest is, for example, the people, places, institutions, and goods present in the text, it is the relationships among these entities that enrich our understanding of them. The network graph presents a data structure that emphasizes the relationship among entities and can be analyzed and traveled (meaning queries can follow pathways from node to node) in a computationally less expensive (i.e., rapid and efficient) way compared with processing similar inquiries using text files or relational databases such as MySQL.<sup>7</sup>

The use of network analysis is well established in some disciplines, such as Biology and Sociology, but it is only slowly gaining popularity among Assyriologists. The first important work of prosopographic research using network analysis was Caroline Waerzeggers' study of Marduk-rēmanni's archive; she published a network analysis method overview the same year.<sup>8</sup> Allon Wagner and colleagues have prepared a method paper with interesting examples, modern and ancient, to demonstrate the process of network analysis.<sup>9</sup> Sara Brumfield's research on Akkadian imperial policy over controlled provinces benefits from Waerzeggers' innovative use of text mining techniques and network analysis but in novel ways to answer her research questions.<sup>10</sup> Further contributing to the Assyriological use of network analysis, David Bamman and colleagues published a statistical model that can infer missing elements in a network, and Eduardo Escobar has recently utilized semantic network analysis to trace the identity of ancient Mesopotamian ingredients.<sup>11</sup>

---

7 MySQL is a relational database software that employs Structured Query Language (SQL) to fetch data. Relational databases are a type of data format that uses unique identifiers (ID) to represent data from one table in another, so that when data is queried using SQL, it is possible to fetch related information from multiples tables at once.

8 Waerzeggers 2014a; 2014b.

9 Wagner 2013.

10 Brumfield 2013. The Old Akkadian period lasted approximately from 2340 to 2200 BCE. The "Old Akkadian period" describes a period in Mesopotamian history during which this region saw the development of an important political entity that, at its apogee, extended almost from the Mediterranean in the west, to the Persian Gulf in the south, and to the Zagros mountains in the east. The Sumerian and Akkadian languages were written using the cuneiform script. Prominent Mesopotamian rulers from the Old Akkadian period include Sargon and Naram-Sin. For another text-mining method, see in this volume, Bigot Juloux, 181–187.

11 Banman et al. 2013; Escobar 2017.

## Problem

The problem explored in this chapter concerns the following questions: How it is possible to manipulate information from larger cuneiform corpora with network analysis methods—especially in the case of unannotated texts? And how can this quantitative approach support the researcher in inquiries about Mesopotamian social history?

As previously established by Waerzeggers, network analysis has the advantage of providing an overview of the larger networks in which individuals of interest inscribed themselves.<sup>12</sup> Social network analysis's main task is to explore social relations, and the network graph data structure is the most appropriate data type to focus on relationships between entities. But there is an additional advantage to employing quantitative research methods: the possibility of generating reproducible research. As explained by Ben Marwick, a researcher can take dispositions, such as sharing data and code,<sup>13</sup> to increase the reproducibility of their work. In turn, papers published in this context are generally more widely read and cited.<sup>14</sup>

With these preceding questions in mind, this chapter provides an overview of the methods and tools used to extract and organize information from Mesopotamian cuneiform administrative texts. It then discusses the processes necessary to convert this data into network graph data, which enables the exploration of the relationships among entities. Transcription, standardization, tokenization, lemmatization, and information extraction are explained, followed by an introduction to exploratory visualization and graph manipulation algorithms. Figure 6.1 shows the flow of the tasks required to prepare graph data for network analysis.

Three major steps comprised in this method are pre-processing the source texts,<sup>15</sup> extracting pertinent information, and analyzing this extracted information. Using exploratory visualization and network analysis algorithms enables the detection of meaningful groups of individuals from the sources.<sup>16</sup>

---

<sup>12</sup> Waerzeggers 2014b.

<sup>13</sup> Sharing data and code can be achieved by offering a copy in a public repository and using a permissive license, such as the MIT license, for software and a creative commons attribution license for data. A practical aspect of public repositories is the possibility of collaboration for the maintenance and further development of the product.

<sup>14</sup> Marwick 2016.

<sup>15</sup> Pre-processing: the task of preparing the (textual) data before starting the annotation process or some computational analysis.

<sup>16</sup> In computer science, an algorithm is a set of instructions that a computer can execute to solve a (mathematical) problem. Network analysis algorithms: algorithms specifically geared to network analysis.

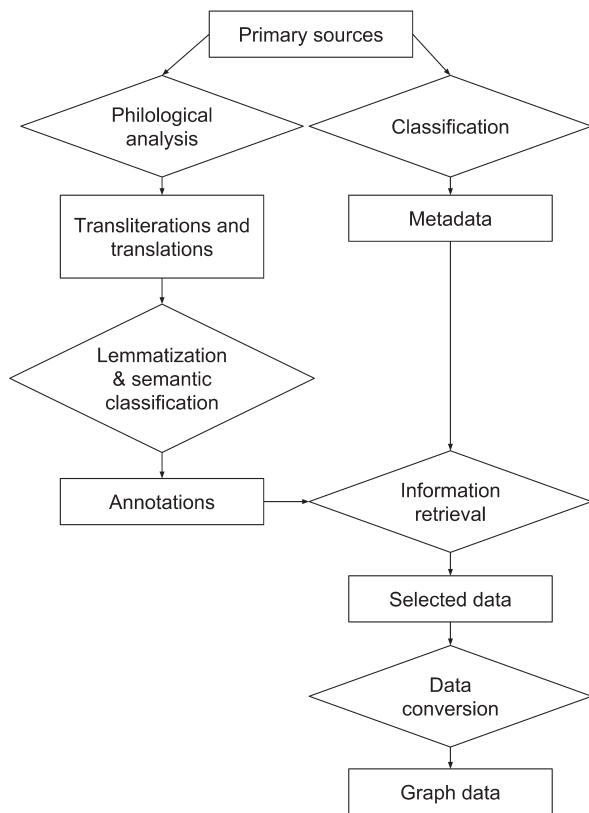


FIGURE 6.1  
*Information processing  
flowchart*

A corpus of approximately 2,700 texts from the city of Adab will serve as the main dataset, and the examples used will showcase the relationships among the people who appear in the texts, based on their co-occurrence in the sources.<sup>17</sup> The discussion section below follows up on how these steps can be made reproducible and how they can enhance traditional approaches.


### Sources Acquisition and Pre-processing

First, the text corpus must be digitized. Thankfully, there are plenty of already digitized corpora available online that can be reused for quantitative inquiries. The Cuneiform Digital Library Initiative (CDLI) and the Open Richly Annotated

<sup>17</sup> It is possible to use this method with other types of entities. For instance, Escobar (2017) uses network analysis to better understand ancient recipes.

Cuneiform Corpus (Oracc) both offer a wide range of prepared transliterations that can be used in the following steps.<sup>18</sup> Let us first consider CDLI. CDLI is the largest database of cuneiform artifacts; it seeks to collect information about all Mesopotamian inscribed objects. It stores this data as metadata and images named “fatcrosses,”<sup>19</sup> as well as transliterations, normalizations, and translations in various languages. The full search page of the CDLI website is complex but also powerful for the advanced user.<sup>20</sup> The Search Aid page explains the possibilities of each search field.<sup>21</sup> One can, for instance, find the exact way to indicate a specific time period or place name in order to restrict a search query. Catalog metadata and textual data, when available, can also be downloaded freely from the CDLI search results page.<sup>22</sup> For example, searching for “Ur-Ninsun,”<sup>23</sup> an inhabitant of the Mesopotamian city of Adab in the Old Akkadian period (2340–2200 BCE), will bring up all of the texts in which this person, and potentially his homonyms, occurs. Transliterations can thereafter be collected via a download link.

One of CDLI’s roles is to maintain a complete catalog and digital copy of all cuneiform texts. Contributions are updated periodically by CDLI staff based on new research, changes in standards, and direct contributions from schol-

- 
- 18 Transliteration: transcription of cuneiform inscriptions into a romanized rendering of the cuneiform sign readings. This includes marking word boundaries and some structural markers, such as line numbering, object surfaces, etc. See Table 6.1 of this paper for an example of transliteration. For example, if one finds  on a cuneiform tablet, the transliteration would read “lugal,” which means king. For CDLI, see <<https://cdli.ucla.edu>> (accessed May 19, 2017), and for Oracc, see <<http://oracc.org>> (accessed May 19, 2017). See also in this volume, Eraslan (285), who discusses special characters specific to c(anonical)-ATF, which is used to encode transcriptions of cuneiform signs in CDLI. For a practical example of Oracc data manipulation for semantics research, see in this volume Svård, Jauhiainen, Sahala, and Lindén, 227–229. See also in this volume, Bigot Juloux, 166, 166n73, 172.
- 19 Metadata is information about the text that is external to the textual information itself and can be used for classification purposes. It may describe provenience, period, genre, etc. See also in this volume, Matskevich and Sharon, 38.
- 20 CDLI: <<http://cdli.ucla.edu/search/>> (accessed May 19, 2017).
- 21 <<http://cdli.ucla.edu/?q=cdli-search-information>> (accessed May 19, 2017).
- 22 Most texts do not have an accompanying translation, but if the user searches with the expression “./.” in the translation field, the search engine will interpret the period as meaning any character and fetch entries that have at least one character in a translation line.
- 23 This search query must be entered in the transliteration field using the notation “ur-{d}nin-sun2,” based on the C-ATF encoding, which will be discussed below. For additional information about C-ATF, see in this volume, Eraslan, 285–286.

ars.<sup>24</sup> Additionally, CDLI has backups at other sites, presently in Oxford, Berlin, and Toronto. Because of these characteristics, CDLI is the ideal place to store cuneiform editions, as it will stand the test of time. When preparing transliterations in digital form for the first time, contributing the results to a well-known database such as CDLI will enable the preservation and accessibility of a researcher's work.

Oracc presents itself as a platform hosting sub-projects that are managed by their contributing teams.<sup>25</sup> CDLI and Oracc both use standardized encoding schemes that are based on the same original format: ATF. ATF was created by CDLI as a stable archiving format for the long-term storage of texts.<sup>26</sup> It has evolved, first to adapt to the usage of CDLI, and later it branched out into Oracc-ATF. Specific characteristics of Oracc-ATF include a wider array of characters permitted in the transliteration lines and the validation of data content managed at the project level. As a result of these characteristics, transcription standards vary. Whether one is working directly from the ancient tablet or from paper publications, the ATF format is an excellent choice for encoding one's work for long-term preservation. Preparing data in a machine-actionable format is essential for its reuse.<sup>27</sup> The Oracc help pages provide instructions with the most complete documentation about both ATF formats.<sup>28</sup> Other websites offer digital transcriptions of cuneiform texts, but some of those websites fail to meet the requirements for a strict and machine-actionable encoding format that is necessary for the success of the next steps. Note that checking licenses is essential, since some sites do not permit reuse of their data.

---

24 Properly prepared transliterations and translations in the C-ATF format made using the guidelines can be sent to [cdli@ucla.edu](mailto:cdli@ucla.edu). Quick pointers can be found here: <http://cdli.ucla.edu/?q=support-cdli> (accessed May 19, 2017).

25 To contribute to Oracc, see this web page: <http://oracc.museum.upenn.edu/doc/about/contributing/index.html> (accessed May 19, 2017).

26 Koslova and Damerow 2003.

27 Machine-actionable (also called "machine-readable") data is structured in a way that it can be processed by computer software. For other encodings that are easily machine-readable, see in this volume, Matskevich and Sharon, 46; Bigot Juloux, 163–164; Nurmikko-Fuller, 339–340, 353–354.

28 For the C-ATF Primer, see <http://oracc.museum.upenn.edu/doc/help/editinginatf/primer/index.html> (accessed May 19, 2017) for Oracc-ATF. There are also quite a few other help pages that can be useful when encoding cuneiform textual data to ATF. See, for example, <http://oracc.museum.upenn.edu/doc/help/> (accessed May 19, 2017). For an example of an encoded transcription of cuneiform in CDLI, see in this volume, Eraslan, Table 9.1, 286.

## Network Analysis

Network graphs are composed of nodes and links; nodes are data points representing entities.<sup>29</sup> Edges link nodes, representing the relationships between those nodes. This data takes the form of “triples” comprising two nodes and an edge that connects them (Fig. 6.2). In network analysis, nodes represent entities that can be of one or more types, such as people, institutions, or places. Edges between entities can be directed and weighted. Directed edges will not be discussed here, since we are only using the co-occurrence of entities in the texts to observe relationships; as such, these relationships are not directed.<sup>30</sup> The weight of an edge here represents the quantity of co-occurrences of two individuals in the corpus. It is important to note that because a relationship is assumed from the co-occurrence of individuals on the same tablet, the nature of this relationship is not similar to a network of Facebook users or employees of a company at their downtown building. For instance, long cuneiform ration lists enumerate people who might never have met but who were part of the same redistribution scheme. The relationship between them can be viewed as weak.<sup>31</sup> However, preparing a graph from this information makes it possible to observe the big picture of the social organization of the workforce, and it is especially helpful when working with a large number of sources.

## Information Type

Administrative texts are typically short and formulaic,<sup>32</sup> with simple sentences and few actions. This type of text is thus an ideal candidate for network graph analysis in which individuals who appear in the text are chosen as entities that will become nodes in the final graph, and their co-occurrence generates links that connect them. The corpus that will be used for the examples below com-

---

29 Network graph: the total of all data points, the nodes, and their relationships—i.e., the edges. Some nodes can be completely disconnected from others and form independent ensembles. Together, connected and disconnected nodes form the network graph.

30 The directionality of a link usually represents who acts upon or toward whom. In these cases, the relationship is more precise than a co-occurrence in a text. For example, if A delivers something to B, the edge can be named “delivers to” and has a direction, from A to B. Linked data is a directed network graph in which all edges are named and directional.

31 Brumfield (2013) has already noted this in her work.

32 Administrative texts of the Old Akkadian period are formulaic in the sense that they mostly consist of lists of people, things, or both, with verbs that have specific, technical meanings in particular contexts. There is little deviation from the usual formula.

prises around 2,700 administrative texts, all dated to the third millennium BCE and provenienced in Adab (a city in Southern Iraq). The corpus is mostly restricted to the Old Akkadian period (2340–2200 BCE).<sup>33</sup> The network it creates has over 800 individuals and 7,000 relationships.

The method proposed in this chapter can be applied to any corpus of administrative texts. A good strategy is to delimit a corpus based on region and temporality and to trim the unnecessary sources from the lot. By comparison, searching for a type of product or concept and harvesting the transliterations from this search query can be problematic; texts bearing the name of important individuals that appear in other contexts will be omitted too. These high-ranking individuals are often bridges, that is, nodes that connect otherwise unconnected groups of nodes together. They are usually very important in understanding social structure using network analysis.

Information about the individuals in a corpus is traditionally stored in the form of lists or as a relational database. These data structures do not, however, facilitate the manipulation of the relationships between the entities. To mitigate this rigidity, the data must be converted to graph data based on triples (Fig. 6.2).

Such data takes the form of triples,<sup>34</sup> where the triple always has two nodes and an edge (individual 1, individual 2, and the relationship between them). Their edge could be weighted, since those individuals can appear concomitantly in texts more than once. In the examples here, labels are attached to the nodes, so the names of the individuals represented by these nodes can be read while manipulating the network graph.<sup>35</sup> For instance, if the ancient Mesopotamian textile workers Geme-Enlil and Mama-ummi appear in the same text,<sup>36</sup> both of them would be nodes, and the edge that links them represents the text in question. If Geme-Enlil and Mama-ummi appear together in nineteen texts, then the edge linking them will be made visibly thicker and have a value of 21.<sup>37</sup> The individuals and their interpersonal relationships thus form a graph.

33 For a preliminary network graph of individuals appearing in third-millennium BCE Adab texts, Pagé-Perron 2017, (<<http://irkalla.net/adab/>> [accessed May 22, 2017]).

34 For another meaning, see in this volume, a) Matskevich and Sharon, 46; b) Nurmikko-Fuller, 345–347 (especially for online-publishing purposes).

35 Label: a form of attribute used to identify nodes and edges easily instead of using only their unique identifiers.

36 Such as in the texts MS 4049 and Lippmann Collection 189 (<[http://cdli.ucla.edu/search/search\\_results.php?SearchMode=Text&ObjectID=P472489,P253146](http://cdli.ucla.edu/search/search_results.php?SearchMode=Text&ObjectID=P472489,P253146)> (accessed May 20, 2017).

37 Searching for “geme2-{d}en-lil2, ma-ma-um-mi” in the transliteration field of the CDLI search page (<<http://cdli.ucla.edu/search/>> [accessed June 4, 2017]) yields the result that

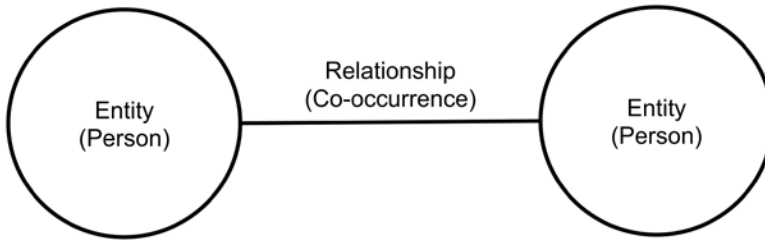


FIGURE 6.2 *A triple*

## Tokenization and Lemmatization

Tokenization and lemmatization are the next steps after the homogenization of the transliterations, meaning after the sign readings, hyphenation, and structural notation are made consistent throughout the corpus. “Tokenization” is the segmentation of a text into similar units called tokens. Tokens can be sentences or signs, but, in this case, the texts are divided at the word level. They are then assigned to lemmata, which are also called “dictionary entries,” in a process called “lemmatization.”<sup>38</sup> Annotations preserve this new information about the text. Various annotation methods employ XML and TEI, popular markup languages. Oracc uses inline glossing as a solution for adding informa-

---

these two individuals appear together 21 times. See such search results here: <[https://cdli.ucla.edu/search/search\\_results.php?SearchMode=Text&requestFrom=Search&TextSearch=geme2-%7Bd%7Den-lil2,ma-ma-um-mi](https://cdli.ucla.edu/search/search_results.php?SearchMode=Text&requestFrom=Search&TextSearch=geme2-%7Bd%7Den-lil2,ma-ma-um-mi)> (accessed June 4, 2017). Examples of such texts are Cornell University Studies in Assyriology and Sumerology (CUSAS) 20, 066: <<http://cdli.ucla.edu/P325058>> (accessed June 4, 2017) and CUSAS 20, 067 (<<http://cdli.ucla.edu/P323404>> [accessed June 4, 2017]). Following is this last text (the June 4th version in CDLI):

```

&P323404 = CUSAS 20, 067
#atf: lang sux
@tablet
@obverse
1. [1(asz@c)] [...] x
2. [1(asz@c)] ma#-ma-um-mi
3. 1(asz@c) geme2-{d}en#-lil2
4. 1(asz@c) GI4-NE-LI
5. 1(asz@c) da-[ni2]-a
6. 1(asz@c) nin-AD2-gal
@reverse
$ blank
    
```

38 Lemma (plural: lemmata): the headword used in a dictionary entry.



tion directly into the transliteration.<sup>39</sup> An alternative approach is to keep the annotations in spreadsheets or a database outside of the studied transliterations with formats such as CONLL.<sup>40</sup> Niek Veldhuis designed a scraper specialized in extracting the lemmata assigned to each token of a text annotated in Oracc.<sup>41</sup> When used on an already annotated text, this type of tool will make it easier to gather specific types of entities: for instance, people in a text are tagged with an indicator that the word is a personal name: “[PN]”. Disambiguation among people in a corpus with the same name is achieved by assigning a unique number to identify each one.

XML annotations, including TEI, are perhaps the most popular means of annotating texts across disciplines.<sup>42</sup> These types of annotations provide two major advantages: (1) ease of marking (with precision) the exact position of the annotated entity in the text and (2) flexibility of the markup,<sup>43</sup> making it possible to annotate multiple layers of information in one text. A good example of a successful annotated corpus available on the web using XML is the Electronic Text Corpus of Sumerian Literature (ETCSL).<sup>44</sup>

One of the possible ways to preserve complex relationships among elements of information is to use a relational database, in which unique ID represent recurring information in conjunction with intermediary tables that associate elements from different tables.<sup>45</sup> In a model to store texts, the tokens it contains, and a glossary, at least five tables are required: texts, tokens, and words, supplemented with relational tables that store the relationships among

39 Inline glossing results in annotations about a transliteration line are stored in the line beneath it.

40 Buchholz and Marsi 2006. CONLL is a column-based file format.

41 Scraper: a type of software that selectively retrieves information from a website when a machine-actionable version of the data is not available. See Veldhuis (2017, <<https://github.com/niekveldhuis/Digital-Assyriology/tree/master/Scrape-Oracc>> [accessed May 19, 2017]). Tokenization: the process by which occurrences of a word, or other units (such as signs or sentences), are separated as units from the original textual data. Working from transliterations, the major separator to take into account is of course the space character.

42 XML: a form of markup language, like HTML, that encodes annotations such as the codes hidden behind the text we see when we use a word processor, such as LibreOffice or Microsoft Word. See also in this volume, Bigot Juloux, 163–164.

43 Markup: a way to annotate information using tags. HTML is a markup language. For further explanation, see in this volume, Bigot Juloux, 163, and for markup to create taxonomies, Bigot Juloux, 166. For markup in EpiDoc, see in this volume, Eraslan, 289–290.

44 <<http://etcsl.orinst.ox.ac.uk/>> (accessed May 19, 2017). For further information, see in this volume, Nurmikko-Fuller, 351–352.

45 Identifier (ID): a unique number that does not have a specific meaning. Its quality resides in its uniqueness. See also in this volume, Matskevich and Sharon, 36.

tokens and texts and the relationships among words and tokens. This system provides the same advantages as annotations but without leaving traces in the transliterations. The alpha version of my database and interface are freely available on GitHub.<sup>46</sup>

At the moment, there is no multi-purpose tokenizer available for cuneiform transliterations, but judiciously chosen regular expressions can extract a text's vocabulary, which can then be stored in tables created to this effect (Table 6.1).<sup>47</sup> Only transliteration lines, which always start with a number followed by a period, are to be processed, and structure-marking elements are discarded. Special attention must be given to broken text. Square brackets and hashtags representing breaks are removed from the text as if there were no damage.<sup>48</sup>

TABLE 6.1 *From text to glossary*<sup>a</sup>

Text	Word tokens	Words/lemmata
&P329002 = CUSAS 13, 134	1(asz@c)	1(asz@c)
#atf: lang sux	sil4	sil4
@tablet	ur-{d}nin-sun2	ur-{d}nin-sun2
@obverse	muhaldim	muhaldim
1. 1(asz@c) sil4	1(asz@c)	udu
2. ur-{d}nin-sun2	udu	nam-ti-e2-mah-ta
3. muhaldim	nam-ti-e2-mah-ta	zi-ga
4. 1(asz@c) udu	udu	
5. nam-ti-e2-mah-ta#	zi-ga	
@reverse		
\$ blank space		
1. udu zi-ga		

a "Archival view of P329002," <<http://cdli.ucla.edu/P329002>> (accessed May 19, 2017).

46 Pagé-Perron 2016a, <[https://github.com/epageperron/cuneiform\\_mining](https://github.com/epageperron/cuneiform_mining)> (accessed May 19, 2017).

47 RegExr (<<http://www.regexr.com/>> [accessed May 19, 2017]) is an excellent online tool that features a find-and-replace function using regular expressions. For additional information on regular expressions, see in this volume, Eraslan (293), who uses them to work on damaged signs. The Adab corpus is stored in C-ATF format.

48 Note that the Assyriological field does not adhere to the Leiden conventions on representing the condition of the text, although most encoding schemes that are used display similar conventions.

In the above table, column two gathers word tokens found in the text, and the third column assembles only the unique tokens. Repeated tokens are assigned to the same lemma, ensuring the final list is duplicate-free. This information is then inserted into the database; new tokens create new entries in the token table, and re-occurring tokens are counted. This count is added to the join table between the tokens and texts. These tokens can be semi-automatically associated with lemmata. When this process has been applied to all of the texts in the corpus, the result is a word list—that is, a list of occurrences of these words as tokens and a list of associations between texts and tokens. The lemmata themselves should be associated with a part of speech and, in the case of personal names, with the specific named entity type “personal name.”

### Data Conversion

After the text has been processed and its vocabulary extracted and stored, the precise data to be represented in the network graph has to be filtered out. This is done by querying the MySQL database tables prepared in the preceding section and saving the results in comma separated value (CSV) text files.<sup>49</sup> The management tool phpMyAdmin can help prepare the queries by showing the results to the user before exporting the data to files.<sup>50</sup> Two files are required: one that lists the entities to study and one that lists the relationships among those entities based on the texts in which they appear (Table 6.2). In the case at hand, only personal names are extracted. The list of individuals can be used as is, and each person will be represented as a node in the graph.

Next, the join data must be converted to network graph triples, which generates an edge between all of the pairs of individuals occurring in the same text. This operation must be carried out for each text, with care taken to remove duplicates, or else to have a duplicate of each combination, so as not skew the edge weights.<sup>51</sup> This new list of pairs is the edge list for the network graph. Creating the combinations can be done using online tools, but working with over 50 nodes is made easier by using a combination algorithm to process all

49 CSV: a format that organizes data as a table, with new lines forming rows and commas delimiting columns. See also in this volume, Monroe, 274.

50 <<https://www.phpmyadmin.net/>> (accessed May 19, 2017).

51 If connections between “a” and “b” (i.e., “a, b”) and between “b” and “a” (“b, a”) are both present in an edge list, two edges will be created between these entities, or a weight of 2 will be attributed to a single edge between them.

TABLE 6.2 *Excerpt from a MySQL table storing information on the occurrence of tokens in the corpus*

Identification (ID) number	Tablet ID number	Token ID number	Occurrences <sup>a</sup>
1892083	652	3	0
1892085	652	4	0
1892086	652	20	0
1892101	652	21	0
1892092	652	312	0
1892095	652	868	0

a The occurrences field is specifically used to note cases in which a token appears more than once in a text. As such, when the field is set to “1,” it means that there is one extra occurrence present. Because in this example the token IDs belong to individuals, the value is set to “0.” It is rare that the same individual appears twice in the same text.

the relationships (Table 6.3).<sup>52</sup> At this stage, the data is now expressed as a

52 The Text Mechanic Combination Generator Tool is a good example of such a tool. (<<http://textmechanic.com/text-tools/combination-permutation-tools/combination-generator/>> [accessed May 19, 2017]). A short PHP script adapted to converting a join table to an edge list, thus generating the required combinations, can be found here: <<https://gist.github.com/epageperron/9f631e7898975653b7dc808586763787>> (accessed May 19, 2017). The code can be read like this:

```
<?php
$join_data = array(); // create an array to hold the join data
$fp = fopen('words_tablets.tab','r'); //read the join data file
exported from MySQL
// while reading the data line by line as tab separated values,
add data as a row in the array precedingly prepared
while (($line = fgetcsv($fp, 0, "\t")) !== FALSE) if ($line)
$join_data[] = $line;
fclose($fp); // close the file after reading
$fp = fopen('edges.csv', 'w'); // create the edges file
fwrite($fp, 'Source, Target, Type, id, Label'.PHP_EOL); //
create the headers for the data
$i=0; // set a counter
$pnos = array(); // set a new array
foreach ($join_data as $jd) { // for each row of the join data
array
```

network graph, and it is ready to be fed into a network graph visualization software.

TABLE 6.3 *Excerpt from an edge list*

Source <sup>a</sup>	Target	Type	Id	Weight
697	1000	Undirected	1	2.0
914	1290	Undirected	2	1.0
1577	1572	Undirected	3	1.0
1065	1062	Undirected	4	1.0
1426	1424	Undirected	5	1.0

a “Source” and “target” are word IDs for personal names. Since the graph is undirected, which ID comes first does not have any bearing on the results.

```
$pnos[] = $jd[0]; // add the tablet id to the pnos array
}
$uniquePnos = array_unique($pnos); // pick only unique tablets
foreach ($uniquePnos as $upno){ // going through tablet ids one
by one
$to_combine=''; //instantiate a variable
foreach ($join_data as $jd){ // for each line of the join_data
array
if ($upno==$jd[0]) { // if the unique tablet id appears in the
first column of that line
$to_combine[]=$jd[1]; // add the content of column 2 in the
array "to combine"
}}
while ($item = array_shift($to_combine)) { // while combining
all individuals appearing on each text
foreach ($to_combine as $val) {
fwrite($fp, "'. $item.'", "'. $val.'", "undirected", "'. $i++
.', "'. $jd[2]."''.PHP_EOL); // create a row to
represent an edge, the relationships between each
individual
}}}
fclose($fp); // close the file
?>
```

Among the software available on the market, Gephi, with its intuitive interface, is a good entry-level tool for producing a first network graph. Importing data into Gephi is straightforward, and it is easy to style nodes and edges depending on the user's needs. Gephi can also convert graph data to numerous formats; this comes in handy when one is using more than one tool to manipulate the data. Gephi also integrates the `sigma.js` plugin,<sup>53</sup> which can prepare a website on the fly for displaying an interactive network graph. Finally, Gephi includes a useful feature for isolating ego-networks that provide information about individuals' relationships.<sup>54</sup> Cytoscape is a more powerful software than Gephi.<sup>55</sup> It is used not only by digital humanists but also by biologists and other scientists. Cytoscape offers an extended number of features, especially for statistical data analysis. Complemented by modules such as NetworkX and `igraph`, the programming language Python can further facilitate exploration of the data in a network graph.<sup>56</sup>

To illustrate the possibilities of exploratory visualization and algorithmic analysis, the next section delineates processes for revealing patterns in the data. The question being asked is: can groups of individuals described in the secondary literature be identified in the dataset using exploratory visualization and algorithmic analysis?

### Identifying Individuals of Interest and Group Cores

Anybody visualizing graph data practices exploratory visualization, but it is often overlooked in discussions of research methods. Exploratory visualization exploits our natural ability to recognize patterns visually. Investigating these

---

53 `Sigma.js` is available as a Gephi module (<http://sigma.js.org/> [accessed May 19, 2017]; <https://gephi.org/plugins/#/plugin/sigmaexporter> [accessed May 19, 2017]).

54 Ego-network: a network graph of one individual, all of its neighbors, and optionally all of the neighbors of these neighbors. In this context, a neighbor refers to a node that is directly connected with the entity being investigated.

55 Both Gephi and Cytoscape are free, open source, and easy to install.

56 A graphic user interface development environment such as Spyder (<https://pythonhosted.org/spyder/index.html> [accessed May 19, 2017]) can be used. For more information about the programming language Python, visit the official website: <https://www.python.org/> (accessed May 19, 2017). NetworkX documentation can be found here: <http://networkx.readthedocs.io/en/networkx-1.11/> (accessed May 19, 2017), and `igraph` can be found here: <http://igraph.org/python/#docs> (accessed May 19, 2017). Note that `igraph` is also available for the programming language R: <http://igraph.org/r/#docs> (accessed May 19, 2017). A module, add-on, or plug-in are all ways to name a software part that can be added to a stand-alone computer program or to a website to increase functionality.

patterns can lead to clues about the organization of the network and help to refine research questions and stimulate new hypotheses. When using a corpus in which the texts, such as sub-archives discussing only one type of activity, are homogenous, network graph visualization is an efficient means for giving an overview of the relationships without requiring extensive manipulation of the graph. For example, a network graph of the fishermen associated with the temple of Bau of Girsu in the Early Dynastic period (c. 2500–2340 BCE) explicitly shows the work arrangements of these individuals. An interactive version of this graph is available online.<sup>57</sup> The graph shows that four major groups of fishermen and their foremen regularly performed jobs together, but the composition of the teams and the foreman in charge could vary slightly. We thus can see four tight groups but with some crossing relationships. The supervisors appear in a centralized position, acting as hubs between the different groups of fishermen.<sup>58</sup>

The more extensive network graph of individuals attested in third-millennium BCE Adab texts comprises over 600 nodes and 5,000 edges (Fig. 6.3).<sup>59</sup> When we look at a full network graph of the Adab corpus, two groups are most visible: textile workers from the Mama-ummi archive and individuals mentioned in the daily bread-and-beer temple rations of the E-tur temple and other institutions.<sup>60</sup> The individuals in these groups share

57 For an interactive version of the graph by Pagé-Perron (2015), see <<http://irkalla.net/fishermen>> (accessed May 19, 2017).

58 Hub: a node with an important number of connections compared to the average number of connections of nodes in a said network graph.

59 To navigate the Adab network graph, load the sigma.js interactive visualization by visiting <<http://irkalla.net/adab>> (accessed May 22, 2017) (Pagé-Perron 2017). Search for an individual using the search field on the left, then click on the individual of your choice in the list that appears below. For instance, search for “geme2-” and click on geme2-{d}en-lil2. On the right pane will appear all individuals connected to her. Her ego-network will be shown in the main visualization space. Clicking on a node will bring up that person’s ego-network. Manipulation of a graph is facilitated by stand-alone software, such as Gephi and Cytoscape. As such, when the Adab graph is set in its final stage, a link to download the graph data will be available at the same URL. In the meantime, inquiries should be sent by e-mail to Émilie Pagé-Perron—[epageperron@gmail.com](mailto:epageperron@gmail.com).

60 The Adab corpus dates to the third millennium BCE and comprises all texts from Adab. The texts include both those from official excavations and those that are lacking provenance but have been attributed to Adab based on factors such as the shape of the tablet, prosopography, and the institutions and rulers mentioned. The large majority of the texts are administrative in nature. The Mama-ummi archive is a group of texts recording transactions related to textile workers. See Maiocchi (2016) for an overview.



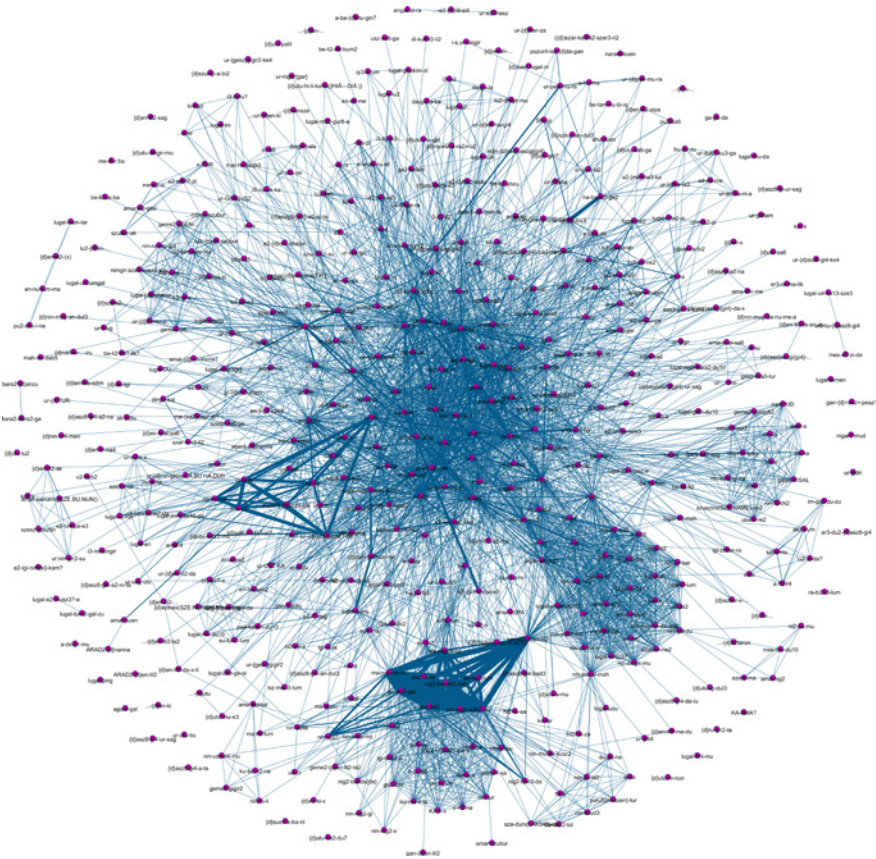


FIGURE 6.3 *The Adab network graph*

edges that have a high weight (i.e., they are thicker) because they co-occur frequently.<sup>61</sup>

In a network with few edge overlaps, most groups can be visually identified. In more complex graphs, extensive overlapping of edges makes it difficult to isolate specific triples. Thus, other means are necessary to indentify groups in the graph.

A possible way to highlight groups in a network graph is to remove edges that have a lower weight. Graphs become simpler and thus clearer when relationships that do not re-occur often are omitted. Nodes that become isolated

61 Weighted relationships can be highlighted in different ways. In Gephi, one can easily make the edges thicker based on the numeric value of their weight. In Cytoscape, a practical way to display the frequency or weight of an edge is either by applying a color range or by changing the thickness of the edge, with the variation depending on the weight value.



can be discarded in the operation. If most individuals in one text do not occur again in others, their contribution to understanding the makeup of groups is minimal. Removing the edges that connect these infrequent individuals reduces the number of edges in the full graph. This highlights the remaining relationships and helps to identify people who frequently appear together, and it also clarifies the role of individuals who act as bridges between separate clusters.

In the Adab corpus, bridges are high-ranking individuals, or they represent homonyms (different individuals who have the same name and who should be represented with two or more separate nodes). A graph-partitioning technique based on the identification of these individuals consists of removing entities that act as a sole or faster point of access between groups in a graph, thus separating these groups from each other. It may be evident from the visualization which individuals serve as bridges, but using a quantitative method to remove them is more efficient, since the threshold used will be numerical and not solely visual. In Cytoscape, it is possible to remove such nodes from a graph after applying the analysis setting and selecting the nodes with the largest edge betweenness,<sup>62</sup> hiding them from the display view. Bridges consist of nodes that usually have a high edge betweenness. Removing bridges fractures the graph and reveals communities by isolating groups of nodes that were connected. Coupled with removing low-weight edges as described above, this method clarifies the picture and highlights important ensembles of individuals. It is particularly efficient for revealing the stable core of archives where individuals co-occur frequently. In the Adab corpus, this equates to specialized workers who less often have hierarchical responsibilities. Individuals with more responsibilities are more likely to be bridges, as they will appear in multiple groups where they manage workers and perform other activities, such as bringing goods to the central authority or receiving rations to redistribute. When the bridges of a graph are removed, the formed groups are tight communities, but they are also missing some of their leaders (the bridges that were removed). It warrants searching for such individuals and their roles to get a full portrait of affairs. With the Adab dataset, groups highlighted with this method equate exactly with the core of the archives that have been discussed in the secondary literature.<sup>63</sup> Examples are: the Mama-ummi archive, the bread-and-

---

62 Edge betweenness: in network analysis, a mathematical centrality measure of relationships between network data points.

63 See: Zhi 1988; Pomponio et al. 2006; Maiocchi 2009; Maiocchi and Visicato 2012; Visicato and Westenholz 2012; Bartash 2013; Molina 2014; Maiocchi 2015; Pomponio and Visicato 2015.

beer lists to temples (E-tur, E-dam), the middle Sargonic herders, and individuals occurring in Early Dynastic ration lists.<sup>64</sup>

### Quantitatively Identifying Meaningful Groups of Individuals

A component is a connected network graph. Any two or more unconnected sections of a single graph are thus different components. These separated groups can be manually identified, but using an algorithm to perform calculations guarantees an exact count. Components can be calculated in Cytoscape or using the “connected\_components” function from the NetworkX Python module.<sup>65</sup> This function will traverse the graph and note the unconnected parts of the network, generating lists of entities present in each component. In the Adab corpus, most nodes are connected. This is because the choice was made to conflate entities bearing the same name into one node. These individuals should be subsequently divided into the adequate number of nodes when disambiguated.<sup>66</sup> Since the corpus is diverse and covers about 500 years of history, this high connectedness tells about the important level of homonymy in the corpus. Hence, the identification of groups with the verification of components should be complemented with other methods. In a graph in which the nodes have been fully disambiguated,<sup>67</sup> one would expect to find

64 Examples would be the texts CUSAS 11, 356 and CUSAS 11, 212. (<[http://cdli.ucla.edu/search/search\\_results.php?SearchMode=Text&ObjectID=P412171,P323060](http://cdli.ucla.edu/search/search_results.php?SearchMode=Text&ObjectID=P412171,P323060)> [accessed May 20, 2017]).

65 The appropriate code looks as follows:

```
import networkx as nx #import the NetworkX module to
use as nx
G = nx.read_gml('/path/file.gml') #read the graph data from a
gml format file and put into "G"
print(list(nx.connected_components(G))) #print on
screen a list of the found connected components
The Graph Modeling Language (.gml) file can be obtained by exporting the graph from
Gephi.
```

66 An alternative method takes the opposite approach in creating one component for each group of people in one transaction and, while performing disambiguation, merging the multiple nodes representing one individual. For a description of this method, see Broux (2017). Anderson (Bamman, Anderson, and Smith 2013) uses a similar method with the Old Assyrian corpus.

67 Although disambiguation can be facilitated using a statistical approach, in the case of individuals with very few occurrences, there can be few indicators to help with the process.

well-defined, medium-sized components, as opposed to what one finds in the central area of the graph above (Fig. 6.3).<sup>68</sup>

The clique is an important concept in network graph analysis, especially when working with administrative records for which the source data generates undirected links among all of the people appearing in the same text. Each tablet therefore forms a clique, that is, a fully connected subgraph where all nodes are interrelated. Cliques are not exclusive: they can overlap and are formed by any group of nodes that have a link between each of them. For example, in a group where a, b, and c are all related, there are also three cliques of two nodes: (a, b), (a, c), and (b, c).

Maximal cliques are the largest cliques identifiable in a network. If a text mentions “a” and “b,” and another text mentions “a,” “b,” and “c,” then “a, b, c” forms a maximal clique, overlapping with the “a-b” clique. As such, maximal cliques equate to the texts with the largest number of individuals who co-occur in other sources.

To find these groups in a graph, one can use the Python NetworkX function “find\_cliques”.<sup>69</sup> Calculating maximal cliques will generate several groups in a number smaller than the total number of texts but larger than the number of components of the graph. These resulting groups are formed by individuals that occur together at least once, but often more frequently. There are 1,430 maximal cliques in the Adab administrative corpus, about half of the total number of texts in the corpus. Here is an excerpted list of nodes of maximal cliques generated using a Python algorithm for the Adab corpus:

```
[ 'lu2-lil-la', 'ad-da', 'ur-gu', 'za3-mu', 'ur-
{d}{sze3}szer7-da', 'ur-ga2'],
[ 'lu2-lil-la', 'ad-da', 'ur-gu', 'za3-mu', 'ur-
{d}{sze3}szer7-da', 'ur-nu'],
[ 'lu2-lil-la', 'ad-da', 'ur-gu', 'za3-mu', 'di-
{d}Utu', 'ur-e2-igi-nim', 'ur-nu'],
[ 'lu2-lil-la', 'ad-da', 'ur-gu', 'za3-mu',
```

68 Future research in network analysis applied to cuneiform corpora could examine how advanced clustering methods can further investigations of groups of people and provide semi-automated and automated methods for disambiguation.

69 

```
import networkx as nxG =
nx.read_gml('/path/file.gml')
print(len(list(nx.find_cliques(G)))) #print on screen a count of
the listing of found maximal cliques
print(list(nx.find_cliques(G)))
This will give a count of maximal cliques.
```

```
\lugal-ezem', \lugal-KA', \ur-ga2'],  
[ \lu2-lil-la', \ad-da', \ur-gu', \za3-mu',  
\lugal-ezem', \lugal-KA', \ur-e2-igi-nim', \ur-  
nu'],  
[ \lu2-lil-la', \ad-da', \ur-gu', \za3-mu',  
\lugal-ezem', \lugal-KA', \ur-...'],  
[ \lu2-lil-la', \szu-na', \ur-{d}{sze3}szer7-da',  
\ur-ga2']70
```

As shown in this excerpt of the maximal cliques from the Adab corpus, one can visually detect the variety of arrangements of individuals that appear intermittently. Meaningful groups can be formed by associating some of the maximal cliques with each other.<sup>71</sup>

TABLE 6.4 Sample of entity frequency in maximal cliques

People	Appearance frequency	Times part of a max. cliques
Mesag	24	23
Mama-ummi	47	27
Ur-Ninsun	17	29

There is some degree of variation in the frequency of appearance of an individual in the whole corpus vis-à-vis its appearance in maximal cliques, as exemplified in Table 6.4. An individual appearing more often in maximal cliques than in the whole corpus is one who connects with diverse individuals. This can be explained by two phenomena: either the individual requires disambiguation, or he or she is a higher-ranking person who performs activities in different contexts. For example, “Ur-Ninsun” appears in 17 texts; a group of eight of these texts discusses expenditures of sheep dispensed to Ur-Ninsun the cook. In two other texts, he is attested as merchant and as gardener, co-occurring with different people. Based on this information, one might decide to

<sup>70</sup> Each list of people forming a maximal clique is enclosed in square brackets. Some of the texts that include maximal cliques are as follows: Adab 0800 + 1011; CUSAS 11, 050; CUSAS 11, 084; CUSAS 11, 129; CUSAS 13, 008; CUSAS 19, 118; CUSAS 19, 179; TCBI 1, 207. (<[http://cdli.ucla.edu/search/search\\_results.php?SearchMode=Text&ObjectID=P217531,P326098,P325845,P324963,P323569,P323191,P328938,P382459](http://cdli.ucla.edu/search/search_results.php?SearchMode=Text&ObjectID=P217531,P326098,P325845,P324963,P323569,P323191,P328938,P382459)> [accessed June 4, 2017]).

<sup>71</sup> See below for an elaboration on this topic in the *k*-plex section.

disambiguate “Ur-Ninsun” into two distinct individuals. A high frequency of occurrence in maximal cliques could also result from a high-ranking individual having influence in different spheres of activity.

A solution compensating for this high-frequency effect in maximal cliques and further clarifying meaningful groups is to relax the maximal clique rules to include more individuals. This relaxation should allow individuals to join the maximal clique members even though they are not fully connected with all other individuals. This procedure forms larger groups and reduces the quantity of groups. Consider these two maximal cliques from the Adab corpus:

```
[lu2-lil-la, ad-da, ur-gu, za3-mu,
ur{d}{sze}szer7-da, ur-ga2] [lu2-lil-la, ad-da,
ur-gu, za3-mu, ur{d}{sze}szer7-da, ur-nu]
```

“Ur-ga” and “Ur-nu” each appear in one list—but among the exact same co-occurring individuals. In a network graph, the individuals in these two maximal cliques are linked to one another, with the noted exceptions of Ur-ga and Ur-nu. Merging these maximal cliques creates what is called a 1-plex, that is, a maximal clique relaxed to accept individuals who are connected to all other individuals minus one connexion.

A  $k$ -plex is a relaxed maximal clique in which  $k$  represents the number of connections that can be missing between nodes when the nodes would still form a group, the  $k$ -plex.<sup>72</sup> In a  $k$ -plex, most nodes are interconnected, but with a tolerance of  $k$ -number of absent edges. In other words, an entity must be tied to all but  $k$  other entities in the group. A 2-plex would be a relaxed maximal clique in which some entities can be connected to all other nodes except two. Calculating  $k$ -plexes is a complex mathematical problem for which there is no Python implementation at present. Since the number of groups has already been reduced drastically,  $k$ -plexes can be formed manually. In the case of the Adab corpus, varying the relaxation variable yields results set between the full graph and the maximal cliques, and thus between one and 1,430 groups.

## Discussion

How can using network analysis help us better understand the social organization and dynamics of ancient Mesopotamians? Although Assyriologists are generally not acquainted with quantitative analysis methods, these techniques

<sup>72</sup> For more information on this mathematical problem, see Balasundaram et al. (2011).

are making a remarkable entry into the field. A compelling example is Paul Delnero's work concerning verbal prefixes whose frequency, used in conjunction with syntactic analysis, supports his assessments of their meaning and usage.<sup>73</sup> This is the first reason why using quantitative analyses is helpful for Assyriological studies: preparing data in a machine-actionable way enables researchers to explore a corpus using a large array of alternative and complementary approaches, from simple statistical models to machine learning algorithms. The results of such methods, including network analysis, yield exact results that can be counted, compared, and used as strong evidence to build an argument. For instance, the cohesion of a group can be said to be stronger in a 1-plex than in a 2-plex, and an individual with a high edge betweenness may either be reputed to be a bridge or require disambiguation.

Looking at larger sets of data, quantitative approaches also offer new means to systematically store information in a manner that retains factors that are not already known to be meaningful. When this data has been collected, network analysis offers an efficient approach to observing entities or individuals and the groups they form in their larger context.<sup>74</sup> One should note that network visualization is also a powerful communication tool. Information remains mostly inaccessible when hidden in tables, but a network graph has an important visual impact that can be used for teaching or to facilitate information dissemination.

Network analysis and visualization are particularly useful for comparative research questions, such as how the hierarchical structures of two different work groups compare and what factors seem to influence the differences. Research on economic, historic, and social questions is enhanced by exploring the relationship among different entities present in the texts but also by connecting entities outside of their restricted groups, between archives. Mama-ummi the textile-worker supervisor is a good example of this: she appears frequently in conjunction with textile workers in the context of the production of textiles and reception of wool. However, since she is a high-ranking manager, she also appears in texts discussing grain allocations.<sup>75</sup> By preparing an ego-network of Mama-ummi, including her direct connections and the connections of her connections, it is possible to situate her within a larger network and investigate the different circles in which she was involved.

---

73 Delnero 2009; 2012.

74 Waerzeggers 2014b.

75 For example: Lippmann Coll 211 (<<http://cdli.ucla.edu/P472511>> [accessed June 14 2017]) and Lippmann Coll 209 (<<http://cdli.ucla.edu/P472509>> [accessed June 14 2017]).

Sometimes, disambiguation by disconnecting homonyms is necessary, and it has the noted advantage of detecting individuals who span archives and periods—a task that is much more difficult using other methods. For instance, when the occurrence of the same personal name spans periods or archives, it is necessary to determine whether the personal name represents a single individual. For example, “ama-kesz3”<sup>76</sup> and “a-tu” appear together only once in the Adab corpus, but they also appear multiple times with other individuals who interconnect in texts classified into both the Early Dynastic IIB and Old Akkadian periods. The analysis of the larger context of their networks, however, reveals that they probably are the same individuals.<sup>77</sup>

Digital, quantitative techniques, as demonstrated in the examples from the Adab corpus case study presented here, can help to detect unseen but meaningful patterns, especially where there are numerous sources to examine. They can also provide quantitative data and a visually compelling display to support or supplement arguments based on traditional analyses.<sup>78</sup>

Increasing the reproducibility of the research is possible by contributing transliterations to a digital library,<sup>79</sup> sharing the code used to produce some results, and/or explaining the steps taken using particular algorithms, coding languages, and software to analyze information. Using or offering open data, along with opting to use open-source software,<sup>80</sup> further supports reproducibility. Moreover, rather than only sharing research results, opening access to the whole methodology employed, as I have done here, increases the possi-

76 Note that the normalized form of this name is “Ama-Keš.”

77 CUSAS 11, 052; CUSAS 11, 238; CUSAS 11, 285 (<[https://cdli.ucla.edu/search/search\\_results.php?SearchMode=Text&ObjectID=P322838,P322862,P328956](https://cdli.ucla.edu/search/search_results.php?SearchMode=Text&ObjectID=P322838,P322862,P328956)> [accessed March 21, 2017]). Note that the distinction between the Early Dynastic IIB and Old Akkadian periods is contentious for some texts from Adab, and some authors assign some texts to one or the other period, although the texts are from the same archive. This is due to an overlapping period where Meskigalla was in power. It is thus possible that the same individual appears across periods, but it is also possible that the period assigned to a tablet is simply not refined enough, causing it to appear as though one individual spanned multiple periods.

78 Quantitative data: countable and measurable specimens on which mathematical operations, such as statistical analysis, can be performed.

79 In cyber-research contexts, “library” has two distinct usages: 1) a “digital library” is an online collection of digital objects that can be surrogates of actual objects or parts of objects, or born digital objects (i.e., Cuneiform Digital Library Initiative) 2) a “software library” is a bundle of code that has a specific focus of application and that can be reused by many while developing software (i.e., Python library).

80 For those pursuing digital projects such as this, it is advisable, if possible, to use free and open-source software, such as Gephi, Cytoscape, MySQL (or MariaDB), PHP, and Python, in order to democratize access to digital-research methods and results.

bility for others to validate and reproduce the research steps for their own purposes.<sup>81</sup> Reproducible research saves labor, allowing more analyses to be conducted, since new investigations do not need to start from the ground up.<sup>82</sup> Information processed during research should be shared as Open Data; this data can then be reused by specialists and non-specialists alike. In his recent article, Marwick explains that enhancing the reproducibility of research also increases the value of a contribution,<sup>83</sup> since reproducibility facilitates replicability—a desirable attribute of studies we call “scientific.” Furthermore, it helps to preserve one’s research, and, as such, the research will better endure the test of time.

### Final Words

From raw data to graph data, manipulating information extracted from cuneiform corpora, algorithmic analysis, and exploratory visualization can be instrumental in discovering new patterns in datasets, including those that have already been studied for a long time. By using network analysis to identify meaningful groups of people, it is possible to perform tasks, such as disambiguation, comparisons of group structure, investigations of meta-interactions between groups, and the exploration of ego-networks. Many compelling hypotheses and interpretations can result from this type of research. For example, of particular interest to this project are questions related to work organization and structure and to the mobility of individuals within this structure, as is shown to be pertinent to the Girsu Early Dynastic fishermen archive and the Adab Mama-Ummi textile workers archive.<sup>84</sup> In both cases, statistical and network analysis support the hypothesis that individuals in positions of power do not simply rest atop a power pyramid: other factors influence their appearance in the texts, and they do not always receive the largest quantity of rations or bring back or produce the most goods. They may also alternate with other central individuals in their respective work groups.

The case study of the Adab corpus presents not only results but also a foundation for further research. A next step in this methodology would be to refine

---

81 See Marwick’s 2017 article, which lays out a comprehensive model for reproducible science that can be consulted for further insight into this topic.

82 The best example of an ancient world project of this sort would be the Perseus Digital Library and its extensions (<<http://www.perseus.tufts.edu/hopper/>> [accessed May 19, 2017]).

83 Marwick (2017) lays out a comprehensive model for reproducible science that can be consulted for further insight into this topic.

84 Pagé-Perron 2016b; Maiocchi 2016.



the investigation results by giving additional attributes to the nodes based on the metadata collected from the texts,<sup>85</sup> such as sub-period, date,<sup>86</sup> and known archive or find spot. Also, the tools discussed in this paper could be explored further. For instance, the Python modules NetworkX and igraph offer many other algorithms that can help resolve other research questions. More complex programming models could be explored, such as MapReduce programming,<sup>87</sup> which can handle both larger datasets and more insense computations to process graph data. This would be a good avenue to investigate in order to implement an algorithm computing  $k$ -plexes. Using information based on the verbs present in the texts, furthermore, it would be possible to build a directed network graph and use other, more refined, tools to investigative algorithms geared to the manipulation of directed edges. Because of the way the network is built, it is easy to identify individuals in positions of power and to trace the changes in their influence over time to some degree. Another interesting path to pursue would be to introduce other types of entities, such as institutions, into the graph in order to enrich the network and the interpretations we could draw through its analysis.

By using the techniques discussed in this chapter, we also come closer to the possibility of reproducing one's research. The Assyriological reader is invited to consider preparing machine-actionable text editions that could extend the use of the published information for digital research and quantitative inquiries, such as in the exciting domain of network analysis. Future research in network analysis applied to cuneiform corpora should examine how advanced clustering methods can further this investigation of groups of people and provide semi-automated and automated methods for disambiguation.<sup>88</sup> Clustering methods have proven useful in social network analysis for many decades. A classic example is Wayne Zachary's 1977 analysis of a karate club: he was able to predict where the network would break when the club split apart.<sup>89</sup> Although we do not have evidence of any ancient Mesopotamian karate clubs, dynamic

---

85 Attributes take the form of extra columns of information in the nodes and edges lists from which data can be filtered, such as by time period or transaction verb.

86 The date should be encoded in a manner that makes it easy to use to form a timeline for all the texts that have such a date; for example, kings can be numbered in order of reign, year placed before the month, month noted as a numeral value, etc.

87 A MapReduce model handles running parallel and distributed computing tasks on a cluster of machines.

88 Clustering methods utilize a statistical approach to group together nodes that resemble each other. Those methods tell us about similarities in the dataset, but the researcher has to interpret the meaning of the groups. Pearce's 2017 Berkeley Prosopography Service might offer solutions along these lines.

89 Zachary 1977.

social networks permeated ancient society, as they do today. Network analysis, therefore, can and should be adapted to the study of Mesopotamian entities, while an array of other digital methodologies can also be employed to tame and interpret larger cuneiform corpora, the scale and complexity of which can obscure meaningful information when treated solely by traditional, qualitative techniques.<sup>90</sup> Especially by ensuring the reproducibility of research by using Open Data and open source software, cuneiform scholars will be able to build upon each other's work and also make it accessible to queries and projects pursued by researchers in adjacent fields who do not read cuneiform. Together we can push the limits of research in our field, while also stimulating new dialogues by opening knowledge of the field to other disciplines.

## References

- Balasundaram, Balabhaskar, Sergiy Butenko, and Illya V. Hicks. 2011. "Clique relaxations in social network analysis: The maximum k-plex problem." *Operations Research* 59 (1): 133–142. <<http://pubsonline.informs.org/doi/abs/10.1287/opre.1100.0851>>.
- Bamman, David, Adam Anderson, and Noah A. Smith. 2013. "Inferring Social Rank in an Old Assyrian Trade Network." *arXiv.org* cs.CY:1–6. <<https://arxiv.org/abs/1303.2873>>.
- Bartash, Vitali. 2013. *Miscellaneous Early Dynastic and Sargonic Texts in the Cornell University Collections*. CUSAS 23. Bethesda, MD: CDL Press.
- Broux, Yanne. 2017. "Identifying individuals through network visualizations (Perfecting prosopographies)." *Historical Dataninjas*. <<http://historicaldataninjas.com/identifying-individuals-network-visualizations>>.
- Brumfield, Sara. 2013. "Imperial Methods: Using Text Mining and Social Network Analysis to Detect Regional Strategies in the Akkadian Empire." PhD diss. University of California, Los Angeles.
- Buchholz, Sabine, and Erwin Marsi. 2006. "CONLL-X Shared Task on Multilingual Dependency Parsing." In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, 149–164. CONLL-X '06. New York: The Association for Computational Linguistics.
- Delnero, Paul A. 2010. "The Sumerian Verbal Prefixes im-ma- and im-mi-." In *Language in the Ancient Near East - Proceedings of the 53e Rencontre Assyriologique Internationale, Moscow-Saint Petersburg, 23-28 July 2007*, edited by Leonid Kogan, Natalja Koslova, Sergej Loesov, and Sergej Tishchenko, 535–561. Babel und Bibel 4 (2). Winona Lake, IN: Eisenbrauns.

---

90 For instructional information on network analysis and other cyber-procedures, see Weingart 2011, Padilla and Locke 2014, and Posner and Lincoln 2016.

- Delnero, Paul A. 2012. "The Sumerian Verbal Prefixes mu-ni- and mi-ni-." In *Altorientalische Studien zu Ehren von Pascal Attinger*, edited by Catherine Mittermayer and Sabine Ecklin, 139–164. OBO 256. Fribourg: Academic Press.
- Escobar, Eduardo. 2017. "Cuneiform Technical Recipes as Semantic Network." Paper presented at the AOS annual meeting, Los Angeles.
- Koslova, Natalia and Peter Damerow. 2003. "From Cuneiform Archives to Digital Libraries: The Hermitage Museum Joins the Cuneiform Digital Library Initiative." *Proceedings of the 5th Russian Conference on Digital Libraries RCDL 2003*. <<http://etana.org/node/6359>>.
- Maiocchi, Massimo. 2009. *Classical Sargonic Tablets Chiefly from Adab in the Cornell University Collections*. CUSAS 13. Bethesda, MD: CDL Press.
- Maiocchi, Massimo. 2016. "Women and Production in Sargonic Adab." In *The Role of Women in Work and Society in the Ancient Near East*, edited by Brigitte Lion and Cécile Michel, 90–111. Studies in Ancient Near Eastern Records 13. Berlin: De Gruyter.
- Maiocchi, Massimo, and Giuseppe Visicato. 2012. *Classical Sargonic Tablets Chiefly from Adab in the Cornell University Collections*. CUSAS 19. Bethesda, MD: CDL Press.
- Marwick, Ben. 2017. "Computational Reproducibility in Archaeological Research: Basic Principles and a Case Study of Their Implementation." *JAMT* 24 (2): 424–450.
- Molina, Manuel. 2014. *Sargonic Cuneiform Tablets in the Real Academia de la Historia: The Carl L. Lippmann Collection*. Catálogo del Gabinete de Antigüedades 1. Antigüedades Epigrafía 1. Madrid: Real Academia de la Historia, Ministerio de Cultura de la República de Iraq.
- Pagé-Perron, Émilie. 2012. "L'industrie de la pêche à Lagaš: Analyse d'un corpus de tablettes cunéiformes provenant des archives du temple de Baba." MA thesis, University of Geneva.
- Pagé-Perron, Émilie. 2015. *Network Graph of the Early Dynastic Lagash Fishermen* <<http://irkalla.net/fishermen>>.
- Pagé-Perron, Émilie. 2016a. "Cuneiform\_mining." Last modified April 1st, 2018. <[https://github.com/epageperron/cuneiform\\_mining](https://github.com/epageperron/cuneiform_mining)>.
- Pagé-Perron, Émilie. 2016b. "The Fishermen of Early Dynastic Lagash." Paper presented at the AOS annual meeting, Boston.
- Pagé-Perron, Émilie. 2017. *Preliminary Adab Network Graph*. <<http://irkalla.net/adab/>>.
- Pagé-Perron, Émilie, Maria Sukhareva, Ilya Khait, and Christian Chiarcos. 2017. "Machine translation and automated analysis of Sumerian." *Association for Computational Linguistics Anthology*. <<http://aclweb.org/anthology/W/W17/W17-2202.pdf>>.
- Pearce, Laurie. 2017. "They all have the same name! Using Berkeley Prosopography Services to Tame the Hellenistic Uruk Onomasticon." Paper presented at the AOS annual meeting, Los Angeles.
- Pomponio, Francesco, and Giuseppe Visicato. 2015. *Middle Sargonic Tablets Chiefly from Adab in the Cornell University Collections*. CUSAS 20. Bethesda, MD: CDL Press.

- Pomponio, Francesco, Giuseppe Visicato, Aage Westenholz, Odoardo Bulgarelli, and Marten Stol. 2006. *Le tavolette cuneiformi di Adab delle collezioni della Banca d'Italia; Tavolette cuneiformi di varia provenienza*. Rome: Banca d'Italia.
- Veldhuis, Niek. 2017. "Scrape Oracc repository." <<https://github.com/niekveldhuis/Digital-Assyriology/tree/master/Scrape-Oracc>>.
- Visicato, Giuseppe, and Aage Westenholz. 2010. *Early Dynastic and Early Sargonic Tablets from Adab in the Cornell University Collections*. CUSAS 11. Bethesda, MD: CDL Press.
- Waerzeggers, Caroline. 2014a. *Marduk-Rēmanni: Local Networks and Imperial Politics in Achaemenid Babylonia*. Leuven: Peeters.
- Waerzeggers, Caroline. 2014b. "Social Network Analysis of Cuneiform Archives: A New Approach." In *Documentary Sources in Ancient Near Eastern and Greco-Roman Economic History: Methodology and Practice*, edited by Heather D. Baker and Michael Jursa, 207–233. Oxford: Oxbow.
- Wagner, Allon, Yuval Levavi, Siram Kedar, Kathleen Abraham, Yoram Cohen, and Ran Zadok. 2013. "Quantitative Social Network Analysis (SNA) and the Study of Cuneiform Archives: A Test-case based on the Murašû Archive." *Akkadica* 134: 117–134.
- Zachary, Wayne W. 1977. "An Information Flow Model for Conflict and Fission in Small Groups." *JAR* 33 (4): 452–473.
- Zhi, Yang. 1989. *Sargonic Inscriptions from Adab*. Changchun: The Institute for the History of Ancient Civilizations.

### ***Tutorials and Learning Material***

- Padilla, Thomas, and Brandon Locke. 2014. *Introduction to Network Analysis*. <<http://thomaspadilla.org/na2014/>>.
- Posner, Miriam, and Matthew Lincoln. 2016. *Creating Network Graphs with Cytoscape*. <<http://doi.org/10.5281/zenodo.56245>>.
- Weingart, Scott B. 2011. "Demystifying Networks, Parts I & II." *Journal of Digital Humanities*. <<http://journalofdigitalhumanities.org/1-1/demystifying-networks-by-scott-weingart/>>.

## Semantic Domains in Akkadian Texts

*Saana Svärd, Heidi Jauhiainen, Aleksi Sahala, and Krister Lindén*

### Introduction<sup>1</sup>

The Mesopotamian civilizations have left behind rich textual material in the Akkadian language, written in cuneiform script on artifacts excavated from that area.<sup>2</sup> A variety of cuneiform scripts were used to write different languages in the ancient Near East,<sup>3</sup> but we are here interested in what has often been called the Mesopotamian cuneiform tradition, which means texts mostly written in Sumerian (which has no known language relatives) and Akkadian (a Semitic language). Michael Streck estimated in 2010 that the number of published texts in Sumerian was approximately 102,300 (3.076 million words) and in Akkadian was approximately 144,000 (9.9 million words).<sup>4</sup> Of course, the published texts are only a segment of the total number of objects with cuneiform text in the museums of the world, and there are doubtless still more texts that have not been excavated that are still unknown to us. Due to the challenges of the cuneiform script, as well as the small number of Assyriologists, the texts are being published at a slow rate, but the number of published texts is continuously growing. According to some estimates, only 10 percent of the excavated texts on cuneiform tablets have been published. Since the late 1990s,

---

1 We thank Dr. Sebastian Fink for his feedback and Dr. Albion Butters for helping to improve the English of this essay. We also gratefully recognize the financial support of the Academy of Finland for the writing of this essay.

2 Mesopotamia refers to a region situated within the Tigris–Euphrates river system, in modern days roughly corresponding to most of Iraq and the eastern parts of Syria. A number of cultures and political units flourished in the area. The most prominent political entities that used Akkadian are usually referred to as Assyria and Babylonian.

3 The cuneiform script was originally developed by Sumerians c. 3000 BCE. In the following centuries, the script was adapted for Akkadian (a Semitic language in use from about 2500 to 500 BCE). It was also used to write several other ancient Near East languages, such as Hittite and Elamite. Although passing centuries and new languages caused the form of the signs to change, the basic shapes—triangular wedges left by a reed stylus—stayed the same.

4 Streck 2010; see esp. the summary on pages 53–54. Here “published text” refers to a cuneiform text that has been made available as a drawing, transliteration, or translation in a credible scholarly publication.

Assyriology has made significant advances in making the texts available in electronic corpora, although it is still not self-evident that all published texts should be made electronically available. The aim of this contribution is to examine two possible language-technology methodologies for analyzing semantic fields in Akkadian.<sup>5</sup> Our research group used the existing electronic Open Richly Annotated Cuneiform Corpus (Oracc).<sup>6</sup> Oracc is one of the largest corpora of Sumerian and Akkadian texts, consisting of over seventeen thousand texts (almost two million words).<sup>7</sup> Roughly half of it has been annotated. Annotation and metadata are crucially important,<sup>8</sup> since not even a well-trained Assyriologist finds transliterated Akkadian as easy to read as a newspaper. The script uses both syllabic and logographic signs, and each sign can be read in many different ways.<sup>9</sup> Akkadian also uses complex inflection; that is, words are modified in order to express various grammatical categories. Furthermore, the precise genre, time period, and expected vocabulary and orthography need to be taken into account. In addition to more traditional Assyriological methods, this project explores two language-technological methods: Pointwise Mutual Information (PMI) and Word2vec.

In the second section of this chapter, we discuss the background for the research and our source material. The third section outlines the theoretical framework underlying our work and how that framework connects with language technology. The fourth section presents hypothetical semantic fields for our test lexemes “horse,” “to speak,” and “power,” based on Assyriological research. These semantic fields will act as a hypothetical baseline against which the results of our chosen language-technological methods may be compared. The fifth section describes the PMI method and reports the results of our test lexemes. The sixth section describes the Word2vec method and reports on the

---

5 Language Technology is a multidisciplinary field that studies and develops methods for processing human language with the help of computers.

6 The research group's work consists of two projects: “Semantic Domains in Akkadian texts” (principal investigator: Krister Lindén) and “Deep Learning and Semantic Domains in Akkadian Texts” (principal investigator: Saana Svärd). Oracc: <<http://oracc.museum.upenn.edu/>> (accessed June 30, 2017). We would like to thank Professor Niek Veldhuis (University of California, Berkeley) for his assistance with Oracc. For further information, see in this volume, Pagé-Perron, 198–200.

7 Our numbers are from September 2017.

8 In corpus linguistics, an annotation is a comment that specifies the various linguistic features of a word. Metadata provides additional information about the text in question. In Assyriology, this information includes date and provenance.

9 A syllabic sign represents a syllable, and a logographic sign represents a word.

results achieved with it so far. The concluding section discusses the results and outlines some paths forward.

### Theoretical and Methodological Background and Source Material

In this section, we will first highlight the general aims of our research group, setting out the larger framework for our whole endeavor. We conclude the section with a more detailed discussion of our source material and the specific aim of this project: to examine two language-technological methodologies for analyzing semantic fields in Akkadian.

The theoretical basis of our research lies in the well-tested and well-founded approach that research needs to consider the meaning of concepts from more than just an outsider's analytical perspective, but also from inside the group that is being studied. This principle of *emic* research underlies, for example, many anthropological studies of human culture.<sup>10</sup> In general, concepts that are foreign to the culture that is being researched do not necessarily have much explanatory power. However, applying an *emic* approach is challenging in the case of the distant Mesopotamian civilizations, which are known through their texts, mostly written in Akkadian and Sumerian, as well as through archaeological finds from the region. We aim to answer this challenge through language itself.

Semantic research in Assyriology has very much consisted of qualitative research on individual concepts—that is, research done on a small amount of data without mathematical methods. Our research group, however, aims to do quantitative research—that is, to use statistical and computational techniques employed in the natural sciences on large amounts of Mesopotamian textual data. We seek to use state-of-the-art technology to handle the data in order to gain new semantic insights into ancient texts and cultures. Without the use of automated methods, this kind of research is extremely slow and often possible for only the few senior Assyriologists who have read through and absorbed thousands of Akkadian texts. Offering a quantitative perspective will broaden the possibilities of semantic and linguistic research on these text corpora by linguists and historians who have only a basic knowledge of the language. Overall, we aim to generate semantic domains for Akkadian lexemes using methods from language technology.

Word sense induction (also called word sense discrimination or word sense discovery) is the task of determining what meaning a word may have in

---

<sup>10</sup> See, for example, Eriksen (2010, 39–40).



different contexts. In some ways, the manual way of doing this has been one of the core concerns of Assyriology ever since scholars started to decipher the first cuneiform texts in the 1850s. However, no determined effort has been made to fully utilize the existing electronic corpora and language-technological approaches. Therefore, dependency parsers or other sophisticated automated tools are not available to create pre-analyzed contexts for words. Starting from scratch, our project plans to apply a number of different approaches to the corpus.

In general, language-technology-related research has rarely been pursued in Assyriology. For the Sumerian language, there are some studies, perhaps because Sumerian material has been available in useful formats longer than Akkadian. Stephen Tyndall used Naive Bayes and Maximum Entropy classifiers to reunite different fragments of the same text in Hittite.<sup>11</sup> In Akkadian, a social network analysis approach has been used as an analytical tool, but not automatically.<sup>12</sup> One previous, older work relating to Akkadian comes from Helsinki, where Laura Kataja and Kimmo Koskeniemi experimented on using computational morphology for Akkadian texts.<sup>13</sup> Recently, Terhi Nurmikko-Fuller has been analyzing narrative structures in Sumerian literature from the point of view of the Semantic Web.<sup>14</sup>

Our data comes from the electronic text corpus Oracc. The corpus is regularly updated, but our data for this paper is a snapshot of the corpus from October 2016.<sup>15</sup> Akkadian uses inflection; that is, words are modified in order to express various grammatical categories: for example, *šarrum*, “king,” for the singular nominative form of the noun and *šarram* for the singular accusative form. This is why we decided to use only the lemmata—that is, the dictionary or citation forms—of various words when looking for semantic domains. Not all of our texts have been annotated with these dictionary forms, and most texts also contain unannotated words; therefore, we created a file containing all the texts where at least 10 percent of the words have been annotated with a dictionary form such as *šarru*, “king”; or *kakku*, “weapon.” In general, in the

11 Tyndall 2012. Hittite was an Indo-European language of Anatolia (Asia Minor) written with cuneiform signs.

12 Waerzeggers, forthcoming.

13 Kataja and Koskeniemi 1988.

14 Nurmikko-Fuller 2016. The Semantic Web provides a common framework that allows data to be shared and reused across applications, enterprises, and community boundaries. For more information, see “w3c Semantic Web Activity,” updated December 11, 2013, <<https://www.w3.org/2001/sw/>> (accessed June 21, 2017).

15 The data had 13,662 texts and almost 1.7 million words. We would like to thank Niek Veldhuis and Steve Tinney for providing us with this data.



Oracc corpus, the metadata added to the texts has been done during different projects over a number of years. This is why some texts have more metadata than others. Some texts include information on provenance, the period when they were written, and genre. Individual words may have such tags as transcription, dictionary form, translation, part of speech, and language. For example, a well-annotated sample word in our material, {GIŠ}TUKUL.MEŠ-ia, includes the following: the base form in Akkadian (*kakku*), part of speech (noun), transcription (*kakkīya*), translation lemma (“stick”), translation sense (weapon), and language (Akkadian). Additional metadata for the whole text can include its provenance (the city of Uruk) and period (Neo-Assyrian).<sup>16</sup>

In our processing of Oracc, such words as numerals, personal names, and divine names, as well as place names, were grouped together under the lemmata “numeral,” “person,” “divinity,” and “place,” respectively. Broken-off words and missing dictionary forms were replaced with the underline character (“\_”), while partly broken words were left as they are in Oracc. Furthermore, the cuneiform script does not mark the end of a sentence, and this is not indicated in any way in our texts. Hence, each cuneiform document was considered as one line of text. When experimenting with the file produced this way, we noticed that various prepositions and adverbs were too prominently present in the results of the analysis. Hence, we modified the file by changing all words that had not been annotated as nouns, verbs, or adjectives into an underline character (“\_”). All in all, this latter file contains 962,868 words from 8,392 different texts. We used the file for analysis with both methods of language technology chosen for this paper.

Regarding the genre of our texts, most of our material are either letters, historical texts, legal texts, or scholarly texts. Counted together, these genres form an overwhelming majority (more than seven thousand texts). Chronologically speaking, about two-thirds of our material (more than six thousand texts) is from the late Neo-Assyrian period or later (written roughly after 800 BCE). Finally, almost half of the texts come from Nineveh, and a significant portion (more than a thousand) are from Uruk. As expected, a large number of texts (approximately 1,500) have uncertain origins. These factors naturally influence our results.

In the context of language technology, distributional semantic models keep track of the appearances of words according to their proximity to each other in

<sup>16</sup> For further examples, see the Korp interface at <[https://korp.csc.fi/?mode=other\\_languages#!lang=en&stats\\_reduce=word&cqp=\[\]&corpus=oracc\\_dcclt&search=word|lugal](https://korp.csc.fi/?mode=other_languages#!lang=en&stats_reduce=word&cqp=[]&corpus=oracc_dcclt&search=word|lugal)> (accessed July 1, 2017).

order to measure their similarity.<sup>17</sup> These models can be computed by counting the frequency of all the words in relation to their neighboring words or, more recently, by predicting the context where words appear. The Pointwise Mutual Information (PMI)<sup>18</sup> method is a count model, whereas Word2vec uses so-called Artificial Neural Networks (ANNS) to predict word co-occurrence.<sup>19</sup> In the current essay, we take the meanings of Akkadian words, as defined in the *Chicago Assyrian Dictionary* (CAD), as a departure point.<sup>20</sup> The results gleaned from analyzing the CAD will then act as a comparison point to the results gleaned by PMI and Word2vec from the file generated from the Oracc data. The following sections will explore in more detail how we analyze semantic fields in Akkadian with the help of PMI and Word2vec.

### Emic Approach and Linguistic Departure Points

As the aim of this section is to discuss our theoretical and methodological departure points, we start by indulging in some history of the field of Assyriology. The eminent Assyriologist Benno Landsberger emphasized the conceptual autonomy of Mesopotamian culture as early as 1926, arguing that Mesopotamian concepts need to be understood on their own terms. He suggested that if one is approaching “the alien mind from a fixed system of conceptual referents ... I could always only find again in my object what I already had within my own perspective.”<sup>21</sup> In some ways, we see this as a kind of precursor to the emic principle (see “Theoretical and Methodological Background and Source Material,” above), which suggests that a culture needs to be understood on its own terms.

17 The field of distributional semantics encompasses the study of semantic similarities between linguistic items by using mathematical methods to see how they are spread throughout a large dataset. See, for example, Baroni, Dinu, and Kruszewski (2014). See also early work by John R. Firth.

18 Church and Hanks 1989.

19 Mikolov et al. 2013a; Mikolov, Yih, and Zweig 2013. Neural networks: a computational model inspired by the human brain where interconnected nodes (neurons) work in parallel to find out how to solve a problem by themselves when given an example. See in this volume, applied a) to landscape archaeology, Ramazzotti, 63–65; b) to objects, in particular cylinder seals, Ludovico, 92–94.

20 The CAD (see in bibliography Roth 1956–2011) is the best and most extensive dictionary of Akkadian, although some parts of it are already fairly old, as the dictionary project took place from 1921 until 2011. The volumes are freely downloadable at <<https://oi.uchicago.edu/research/publications/assyrian-dictionary-oriental-institute-university-chicago-cad>> (accessed June 21, 2017), but they are not fully searchable.

21 This English translation appears in Landsberger (1976, 60), but the original idea was first published in 1926.

Anthropologists have engaged the same problem.<sup>22</sup> At the root of the dilemma is that Western models of thinking are far from universal. At the same time, while concepts that are foreign to the culture that is being researched can be used in research (the so-called *etic* approach), if the evidence is shoe-horned into a modern mold, there is a danger of anachronistic explanations. As a simple example, taking marriage as a topic of study, defining marriage as a legal contract between men and women as a cornerstone of one's research is not very fruitful if one is researching a cultural system where there are more than two genders or where marriage does not have legal implications in the same sense as in the modern Western world.

A similar dilemma of *emic* vs. *etic* is well known in linguistics. In the twentieth century, formal linguists argued that individual languages were mere variations of the underlying universal cognitive language model. On the other hand, linguistic relativists opposed this and proved that different languages, as they are spoken by native speakers, do have real differences on the conceptual level.<sup>23</sup> Linguistic relativity has a long and debated history, but here we concentrate on its connection to cognitive linguistics.<sup>24</sup> A classic in that regard is George Lakoff's work, which argues against an "objectivist" view of categories reflecting reality, but suggests that categories are rooted in the body ("experimental realism") and have a lot to do with the physical and social environment.<sup>25</sup> We subscribe to this view, which has been further explored in a field that is parallel to Assyriology: namely, the study of the Hebrew Bible.

In biblical studies, cognitive linguistics has been approached, for example, by James Barr and Ellen van Wolde.<sup>26</sup> We found the work of Reinier de Blois to be especially useful; it argues strongly for the application of cognitive linguistics. According to Blois, "Where traditional linguistic theory claims that words have meanings, the cognitive linguist would say that meanings have words."<sup>27</sup>

Every word is a member of a larger group of words and shares certain aspects of meaning with them. From a paradigmatic perspective, people categorize concepts in paradigmatic cognitive categories, also referred to as lexical semantic domains. For example, in English, the concept "orange" belongs to the lexical semantic domain "fruit," together with pears and apples.

22 For one example, see Arens and Karp (1989).

23 Levinson 2003, 14–16; Fleisch 2007, 41–42, 46. See also Dirven and Verspoor (2004) and Geeraerts (2010).

24 See also in this volume, Bigot Juloux, 170–174, who uses cognitive linguistics for verbal categories.

25 Lakoff 1987.

26 Barr 1961; Wolde 2009.

27 Blois 2008b, 266.

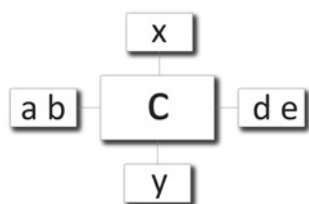


FIGURE 7.1

*Syntagmatic (abCde) and paradigmatic (XCY) relationships between words*

Simultaneously, concepts always associate with other concepts syntagmatically, forming contextual semantic domains. In other words, the lexeme “orange” can function “in different settings ... each of which evokes a different image in our minds.” The HORTICULTURE-domain for “orange” evokes images of “tree,” “ripe,” “picking,” etc., whereas the COMMERCE-domain for “orange” implies “booth,” “seller,” “money,” etc.<sup>28</sup>

Thus, syntagmatic semantic domains are groups of words that “occur together in a prototypical scenario.”<sup>29</sup> We see the methods of language technology as the key to tracing such semantically relevant properties in large corpora and to providing Assyriologists with tools to reflect on the semantic domains of the words. The traditional Assyriological approach to building semantic fields is thus connected with the language-technological approaches by means of our focus on syntagmatic and paradigmatic relationships between words. In Figure 7.1 below, these relationships are arranged on two axes, *abCde* and *xcy*. To give a concrete example of these relationships, Figure 7.2 shows the example relating to “orange” when it is situated in this diagram.

The following three methodological sections examine syntagmatic and paradigmatic relationships for three specific Akkadian lexemes: *sīsû*, “horse”; *qabû*, “to speak”; and *danānu*, “to be strong, powerful.” These lexemes were chosen because they are well attested in Akkadian and sufficiently distinct from each other. Another topic altogether is the connection of individual lexemes to more complex cultural concepts, but this question is not within the scope of the current paper. As we are, at the moment, only testing which language-technology methods might be useful for analyzing our data, in the first of the following three sections (“Traditional Examples of Semantic Fields”) we use the traditional Assyriological method of examining words “by hand” with the literature of the field. In practice, we built a suggestion for both syntagmatic and paradigmatic fields with the help of the *CAD*. The two language-technological

<sup>28</sup> Blois 2008b, 274.

<sup>29</sup> Blois 2008a, 206.

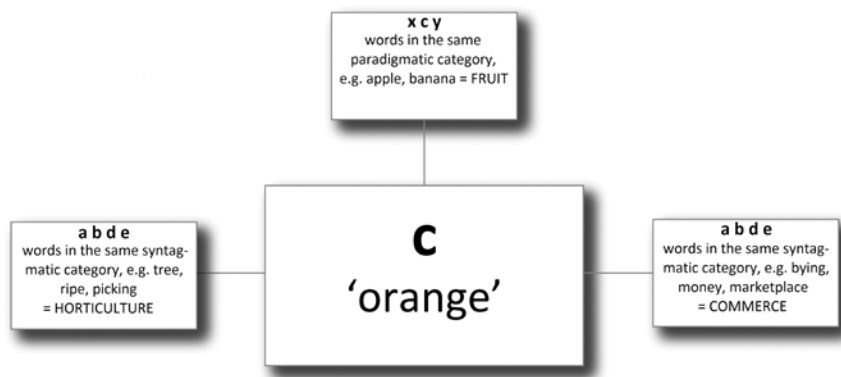


FIGURE 7.2 Syntagmatic and paradigmatic relationships, using “orange” as an example

methods which we chose to test here are PMI and Word2vec. PMI is a statistical method that finds the words that appear close to a target word,<sup>30</sup> whereas Word2vec has been shown to capture the semantic relationships of words.<sup>31</sup> These particular methods were chosen because PMI can be used to analyze the syntagmatic axis *abCde*, and Word2vec may be able to yield results regarding the paradigmatic *xcY* axis of our test words. The results of these analyses will be compared to the results of the Assyriological inquiry in “Pointwise Mutual Information” and “Word2vec.” In other words, the results of the traditional method described in the following section will act as a kind of hypothesis in relation to which the results of PMI and Word2vec can be compared.

### Traditional Examples of Semantic Fields

In the three subsections of this section, we outline both syntagmatic and paradigmatic relationships of three Akkadian lexemes: *sisû*, “horse”; *qabû*, “to speak”; and *danānu*, “to be strong, powerful.” Analysis of these relationships is based on existing Assyriological research, namely the *CAD*.<sup>32</sup> In practice, we researched the entry in question (*sisû*, *qabû*, and *danānu*) in the *CAD* and formulated a hypothetical syntagmatic and paradigmatic field based on the infor-

<sup>30</sup> Church and Hanks 1989.

<sup>31</sup> Mikolov et al. 2013a; Mikolov, Yih, and Zweig 2013.

<sup>32</sup> Although we use the *CAD* as our source material, we follow here the writing conventions of *A Concise Dictionary of Akkadian (CDA)* (Black, George, and Postgate 2000), as that is the guideline volume used in our text corpus, the Oracc.

mation and examples there. The *CAD* is not fully searchable, so this part of the research needed to be done by hand. As these proposals are based on the *CAD*, they represent a potentially etic view of the lexemes. This is why the suggestions of this section will be evaluated in light of PMI and Word2vec results in later sections. Tables 7.1–3 highlight syntagmatic relationships. A discussion of the paradigmatic relationships follows each table.

In our close reading of the *CAD*, the aim regarding syntagmatic relationships was to identify the probable text scenarios where the lexeme of interest could appear, the kind of “prototypical” scenarios described above (see Fig. 7.1 and Fig. 7.2). Each such domain has been labeled with a letter and a heading in the table. For example, in Table 7.1 is found category C: *Horse as an animal in royal service and war*, which includes lexemes that (based on the *CAD*) tend to occur with “horse” in that particular scenario. This approach has meant that we have excluded words that have semantically a great number of possible contexts (e.g., prepositions) or that appear very rarely with the word of interest. Details of such excluded words are given below in the subsections. Because the analysis is qualitative and the method involves close reading, there is no simple list available for which words were discarded. The endeavor was based on *CAD* evidence, which of course had to be balanced with the fact that the text material available for the *CAD* project is not identical with our corpus material (described above in “Theoretical and Methodological Background and Source Material”).

We compared the tables that were thus created against a list detailing how many times the words appeared in the text corpus used for this study. Words appearing fewer than ten times were eliminated from the tables. Additionally, one should bear in mind that most compound words (e.g., *rabi sisī*, lit. “master of horses”) are not identified as individual lexemes in our corpus, which is why they did not show up in our analysis.<sup>33</sup> The lexemes in tables are presented (within each category) in the order of how common they are in our corpus. The number of occurrences is indicated in parentheses after the Akkadian word. Thus, this listing at the same time indicates a rough probability of the individual lexemes appearing in the PMI analyses.

33 Note that approximately 380 compound words are marked with a single lemma in the corpus.

### *sisû*, horse

Based on the work of Blois,<sup>34</sup> it was suggested earlier that in a small sample corpus<sup>35</sup> the syntagmatic semantic domain of “horse” (Akkadian *sisû*) could have at least two categories: 1) the WARFARE-domain associates “horse” with “garrison,” “chariot,” “cavalry,” and “attack,” whereas 2) the ECONOMIC-domain gives associated words such as “silver,” “gold,” “to buy,” and “gift.” Hence, one can suggest that the writers considered the horse mainly as an economic and military resource and not—as in Finland—an important animal in the AGRICULTURE-domain.<sup>36</sup>

This initial analysis was compared to the *CAD*,<sup>37</sup> which suggests the following categories: **a** in general; **b** for transport; **c** for riding; **d** draft horses; **e** military contexts; **f** booty and tribute; **g** as gifts; **h** care, pasturing, and training; **i** provenance; **j** price; **k** colors and markings; **l** qualification by breed, age, and sex; **m** trappings; and **n** personnel.<sup>38</sup> These categories, which highlight some syntagmatic relationships of “horse,” have helped us to compile Table 7.1 below. In order to build syntagmatic categories for *sisû* (686), “horse,” as outlined in the beginning of “Traditional Examples of Semantic Fields,” we disregarded some words when creating Table 7.1. For example, *nagû*, “to bray, bellow, bawl,” can in some contexts be translated as “to neigh.” Instead, we concentrated on those words that occur in *CAD* evidence consistently with “horse” in many different *CAD* categories. The *CAD* lists approximately fifty lexemes for *CAD* categories **k**, **l**, **m** and **n**, but it is impossible to evaluate how commonly these were connected with “horse.” We would expect that individual qualifying terms would not show up in quantitative analysis because horses are rarely qualified with any one of these terms. Moreover, these categories were excluded from Table 7.1 because of our focus on finding prototypical scenarios

34 Blois 2008b.

35 SAA 18 (Reynolds 2003), consisting of 204 texts and dealing mainly with the Babylonian correspondence of King Esarhaddon, served as an example corpus: “horse” appeared in SAA 18 on pages 7, 15, 56, 66, 80, 112, 125, 175, and 197.

36 Svärd 2012, 58–61; Svärd 2015, 15–19.

37 *CAD*, vol. S, 328–334.

38 The lexeme can also refer to the constellation Horse or to a sea creature (literally “horse of the sea”), but these meanings are rare and not relevant to this discussion.

TABLE 7.1 *Three prototypical contexts for sisû, “horse” (syntagmatic relationships)*

Akkadian lexeme (base form according to the CAD)	English translation (according to the CDA)
<b>A. Valuable commodity</b>	
<i>kaspu</i> (1,686)	silver
<i>maddattu</i> (209)	payment, obligation, tribute
<i>šulmānu</i> (22)	greeting gift
<b>B. Equipment connected with using a horse and taking care of it</b>	
<i>kurummatu</i> (317)	food allocation, ration
<i>nīru</i> (166)	yoke
<i>šūšānu</i> (56)	horse trainer, groom
<b>C. Horse as an animal in royal service and war</b>	
<i>šābu</i> (750)	people, troops
<i>emūqu</i> (323)	strength, force, military force
<i>narkabtu</i> (234)	chariot
<i>karāšu</i> (93)	military camp
<i>rakābum</i> (91)	to ride
<i>šamādu</i> (31)	to tie up, yoke, hitch up

The CAD is not ideal for identifying paradigmatic relationships because its categorizations are mostly based on syntagmatic relationships. However, at least the following comparable animals and designations of different types of horses may be expected to belong to its paradigmatic category: *imēru* (612), “donkey”; *alpum* (539), “bull, ox”; *kūdānu* (114), “mule”; and *atānu* (31), “she-ass, mare.” These lexemes are also likely to show up in an automatic analysis of syntagmatic relationships because they often appear *both* close to each other *and* in similar contexts.

### *qabû*, “to speak”

The verb *qabû* (2,353), “to speak,” is a common verb with a wide semantic range. In Table 7.2, we seek to identify prototypical scenarios where the category can be connected with a fairly specific set of lexemes (based on CAD, vol. Q 22–42). There are certain lexemes that seem to often appear together with *qabû*, but which cannot be connected to any specific semantic domain based



on the *CAD* information. Such words include *awātu* (556), “word, matter, order”; *muḥḥu* (3,493), “skull, top, concerning (something)”; *magāru* (150), “to consent, agree”; *pānu* (3,233), “front, face” (in the expression *ina/ana pāni*); *aššum* (6), “because of, concerning”; *umma* (3,303), “saying” (a particle introducing direct speech); *kīam* (102), “so, thus”; and *kīma* (1,339), “like, when, thus.” Given the broad semantic nature of the word *qabû*, we expect that its syntagmatic categories will be difficult to identify. Nonetheless, Table 7.2 presents some possibilities, based on our analysis of the *CAD*.

TABLE 7.2 *Three prototypical contexts for qabû, “to speak” (syntagmatic relationships)*

Akkadian lexeme (base form according to the <i>CDA</i> )	English translation (the first meaning according to the <i>CDA</i> )
<b>A. To state officially (e.g., in court)</b>	
<i>maḥru</i> (1,101)	front, before
<i>dīnu</i> (873)	legal decision, judgment, verdict
<i>mimma</i> (531)	anything, something, everything
<i>dayyānu</i> (82)	judge
<b>B. To order</b>	
<i>divine names</i>	any name of a deity
<i>šarru</i> (14,389)	king
<i>šulmu</i> (1,874)	completeness, well-being (in letters)
<i>pû</i> (870)	mouth
<i>tūbu</i> (287)	goodness, happiness (in letters)
<b>C. To promise (financially)</b>	
<i>kaspu</i> (1,686)	silver
<i>šiqḷu</i> (974)	shekel
<i>še’u</i> (181)	barley, grain

The verb *qabû* was chosen as an object of study nonetheless, because of its potentially interesting paradigmatic relationships. There is reason to suspect that its meaning is not sufficiently conveyed in English expressions. It seems to have specific connotations with authoritative speech, whereas another verb, *dabābu* (1,163), “to speak,” has connotations with informal speech.<sup>39</sup> Therefore,

39 This is already briefly stated in Svärd (2015, 169). According to the *CAD*, vol. D, 4–14, *dabābu* can mean “to recite” (referring to cultic actions, typically close to divine names or prayer

the paradigmatic relationships of this word are of great interest. Does *dabābu* appear in similar contexts as *qabû*? Or are there other lexemes denoting speech that are closer to *qabû* than *dabābu*? A possible candidate for this may be *zakāru* (209), “to speak.”

### *danānu*, “to be strong, powerful”

Svård’s previous research has explored the concept of “power” in the Neo-Assyrian Empire.<sup>40</sup> There are many lexemes, usually translated with the English “power” or the German “Macht,” but here we are concentrating on one of the most obvious ones, *danānu* (221), “to be strong, powerful” (*CAD* vol. D, 83–86), which is connected with the idea of physical strength and force.<sup>41</sup> Table 7.3 sums up our hypothesis based on the *CAD*.

TABLE 7.3 Five prototypical contexts for *danānu*, “to be strong” (syntagmatic relationships)

Akkadian lexeme (base form according to the <i>CDA</i> )	English translation (at least the first meaning according to the <i>CDA</i> )
<b>A. The crown and military</b>	
<i>šarru</i> (14,389)	king
<i>mātu</i> (4,039)	land, country
<i>nakru</i> (711)	strange, foreign, enemy
<i>emūqu</i> (323)	strength, force, army
<i>kabātu</i> (29)	to be heavy, to be important
<b>B. To guard</b>	
<i>maššartu</i> (491)	observation, guard, watch
<b>C. To make valid, binding</b>	
<i>šibu</i> (4,232)	old, elder, witness
<i>tuppu</i> (537)	tablet, document

terminology) or “to plead in court” (typically close to words related to the derivatives of the verbal root *djn*, such as *diānu* in the infinitive “to judge”; *dajjānu*, “the judge”; *dīnum*, “court case”). It also frequently occurs with the preposition *itti* (with), because *dabābu* often refers to reciprocal action.

40 Svård 2012, 2015.

41 Svård 2012, 61–65.

TABLE 7.3 Five prototypical contexts for *danānu*, “to be strong” (cont.)

Akkadian lexeme (base form according to the CDA)	English translation (at least the first meaning according to the CDA)
<b>D. To strengthen man-made structures</b> (canals, foundations, dikes, temple of baked bricks, structure, reservoir)	
<i>bītu</i> (5,927)	house
<i>nāru</i> (460)	river, canal
<i>dūru</i> (313)	city wall, rampart
<i>igāru</i> (113)	wall
<b>E. To speak/write severely/strongly</b>	
<i>šapāru</i> (2,187)	to send, send a message
<i>awātu</i> (556)	word, matter, order
<i>šiptu</i> (31)	judgment, verdict

As for paradigmatic relationships, the root of *danānu* (*dnn*) has many derivatives, which might show up in similar contexts as *danānu*. These include *dan-nu* (825), “strong, mighty”; *dannūtu* (50), “strength, power”; and *dunnu* (16), “power, strength.” Other possibilities for paradigmatic companions are derivatives of the root *gšr*: *gašāru* (3), “to be strong, powerful”; and *gašru* (60), “very strong, powerful.” Finally, it seems possible that *emūqu* (323), “strength, force,” could appear in similar contexts as *danānu*.

Pointwise Mutual Information (PMI)

PMI is a statistical method used to find collocations and associations between words. It measures the reduction of uncertainty about the occurrence of word<sub>1</sub> (*w*<sub>1</sub>) when word<sub>2</sub> (*w*<sub>2</sub>) is known to be present within a given distance. Thus, we may be able to generate groups of words that are syntagmatically related to each other (see *abCde* axis above in Figures 7.1 and 7.2).

To calculate the PMI scores, we used the Collocations module provided by the Natural Language Toolkit (NLTK), which is the most popular platform for building Natural Language Processing (NLP) programs for Python.<sup>42</sup> The NLTK

42 Natural Language Toolkit (NLTK): <<http://www.nltk.org/>> (accessed July 1, 2017). See also Bird, Klein, and Loper (2009). For the Python programming language, see <<https://www.python.org/>> (accessed July 1, 2017).

is open-source software, and it includes several powerful and easy-to-use tools, from tokenizers to parsers and visualization.

The method splits the input text into smaller segments of fixed length, called a (collocational) window. The size of the window defines the maximum distance between the elements of the examined bigram consisting of  $w_1$  and  $w_2$ . Then the frequency of the bigram and the individual frequencies of  $w_1$  and  $w_2$  are calculated and converted into probabilities. The PMI score is a logarithmic ratio of the bigram's actual probability to the expected chance that the words of the bigram would co-occur independently.<sup>43</sup> Mathematically:

$$PMI(w_1; w_2) = \log \frac{p(w_1, w_2)}{p(w_1)p(w_2)}$$

FIGURE 7.3  
Formula for PMI score

where  $p(w_1, w_2)$  stands for the joint probability of the words and  $p(w_1)$  and  $p(w_2)$  for their individual probabilities. For example, let us assume that we have a corpus of 1,000,000 words and we want to measure the PMI score for a bigram consisting of words  $w_1$  and  $w_2$ . Let us also assume that our bigram occurs in the corpus 100 times within a given window, and the separate frequencies for  $w_1$  and  $w_2$  are 400 and 160, respectively. By using the formula described above (without yet using the logarithm), we get:

$$\frac{\frac{100}{1000000}}{\frac{400}{1000000} \times \frac{160}{1000000}} \approx 1562.5$$

FIGURE 7.4  
Example of the formula for the PMI score

If the individual frequencies of  $w_1$  and  $w_2$  were higher, the expected probability of their independent co-occurrence would increase, and the formula would give us a lower result. If the result drops below 1.0, it indicates that there is a higher probability that the words occur together coincidentally rather than as a true bigram.

By using a logarithm,<sup>44</sup> we can normalize the outcome in a way that values greater than 1.0 are converted into positive PMI scores and values less than 1.0 into negative PMI scores. With the values given above, we would have a PMI score of 7.35. If the individual frequencies were 4,000 and 1,600, the PMI score would decrease to 2.75 ( $\log 15.63$ ), thus indicating a weaker association.

43 Church and Hanks 1989.

44 By common definition, a logarithm (abbreviated "log") is a quantity representing the power to which a fixed number (the base) must be raised to produce a given number.

As the PMI has a well-known tendency to give high scores for bigrams with low frequency,<sup>45</sup> several different ways have been proposed to normalize and improve the reliability of the PMI scores. One option is simply to use a high frequency threshold (i.e., to filter out bigrams that occur rarely), but due to the small size of our corpus, this would have caused a loss of potentially interesting data. In order to get better results with low frequency thresholds, we chose to normalize our scores by using Normalized PMI (NPMI).<sup>46</sup> The NPMI is an extended version of the (PMI) formula, which removes some of the low-frequency bias and gives the score fixed upper and lower boundaries.

$$NPMI(w_1; w_2) = \frac{PMI(w_1; w_2)}{-\log p(w_1, w_2)}$$

FIGURE 7.5  
*Formula for NPMI*

To avoid subjectively choosing one window size and frequency threshold, we used a range of different window sizes from 5 to 25 in increments of 5 with five different frequency thresholds each: 10, 15, 20, 25 and 30. This produced 25 score tables for each word. We chose to pick our final results from the best 20 collocations from each score table. The results were merged by calculating the collocate frequencies ( $freq_c$ ) from each score table (i.e., in how many tables they were present at least once) and by using a natural logarithm of this frequency as a weight for their average NPMI scores:

$$SCORE = NPMI_{avg} \times \log freq_c$$

FIGURE 7.6  
*Formula for the final score*

Here the logarithm was used for two purposes: to avoid giving high-frequency collocates too-dominant weights and to filter out all collocates that occurred only once in our score tables (as  $\ln 1 = 0$ ). For example, the following simplified score tables would have given words *a* and *c* the  $freq_c$  of 2 (as they occur in Part I and Part II of Table 7.4) and a logarithmic weight of  $\ln 2 \approx .3$ ; thus, they would have been merged as follows:<sup>47</sup>

45 Manning and Schütze 1999; Pantel and Lin 2002.

46 Bouma 2009.

47 Naturally, with only two score tables, filtering out single collocates would not have been justified.

TABLE 7.4 *Merging the score tables*

Part I Collocate	NPMI	Part II Collocate	→ NPMI	Merged Collocate	NPMI <sub>AVG</sub>	Score
<i>a</i>	.3	<i>d</i>	.3	<i>a</i>	.25	.075
<i>b</i>	.2	<i>a</i>	.2	<i>c</i>	.15	.045
<i>c</i>	.1	<i>c</i>	.2	<i>b, d</i>	–	0

The scoring did not only filter out unique collocates, but it also gave a little bias for the high-frequency collocates and bigrams,<sup>48</sup> of which the latter, as mentioned earlier, are generally considered more reliable. It also allowed some frequent but not so highly scored collocates to appear in our final results. Very similar results could have been achieved by using slightly higher frequency thresholds, but that would have caused more data loss, especially with infrequent words of interest. By using the scoring method described here, the threshold is more adaptive to the words of interest with varying frequencies.

The tables below summarize our final results, which are sorted by the average NPMI scores. The weighted NPMI scores are not shown in the tables, as they were only used for merging the results and filtering out potentially irrelevant collocations. Column “Freq<sub>AVG</sub>” indicates the average bigram frequency.

TABLE 7.5 *NPMI results for the top fifteen suggestions for sisû, “horse”*

	Word	Translation	Freq <sub>AVG</sub>	NPMI <sub>AVG</sub>
1.	<i>Kusaya</i>	Kushite	91	.418
2.	<i>pēthallu</i>	riding horse	82	.363
3.	<i>parû</i>	mule	72	.353
4.	<i>Mesaya</i>	Mesaeian	16	.349
5.	<i>kūdanu</i>	mule	109	.337
6.	<i>udru</i>	Bactrian camel	20	.329
7.	<i>sisû</i>	horse	219	.293
8.	<i>ṣēnu</i>	flock	66	.293
9.	<i>gammalu</i>	camel	34	.288

48 The bigram and collocate frequencies are related as averages, and thus giving a penalty for the infrequent collocations in score tables increases the average bigram frequency in the merged table.

TABLE 7.5 NPMI results for the top fifteen suggestions for *sisû*, horse (cont.)

	Word	Translation	Freq <sub>AVG</sub>	NPMI <sub>AVG</sub>
10.	<i>nību</i>	naming	34	.254
11.	<i>narkabtu</i>	chariot	57	.248
12.	<i>nīru</i>	yoke	43	.246
13.	<i>gimru</i>	totality	129	.229
14.	<i>alpu</i>	ox	92	.215
15.	<i>pešû</i>	white	36	.208

At first glance, the PMI results of Table 7.5 for syntagmatic fields of “horse” do not match up well with the syntagmatic fields proposed by traditional *CAD* analysis presented in Table 7.1. Table 7.1 presents three hypothetical fields: (A) valuable commodity, (B) equipment relating to horses, and (C) horses in war and royal service. The only two lexemes from Table 7.1 that appear in Table 7.5 are *narkabtu* (chariot), connecting “horse” to field C, and *nīru* (yoke), connecting “horse” to field B.

We should note, however, that some Table 7.1 words received high NPMI scores but do not show up in Table 7.5 due to their low collocational frequency (see Table 7.4 above). Words that appear in the *CAD* analysis of the section titled “*sisû* (horse)” and had a positive NPMI score (but which do not show up in Table 7.5) include: *šūšānu(tu)*, “position of horse-trainer” (.317); *rakābu*, “to ride” (.206); *karāšu*, “military camp” (.186); *emūqu*, “strength, soldier” (.13); and *kaspu*, “silver” (.003).

In general, the results seem to consist of different draft animals and animals used in expeditions and caravans. Therefore, Table 7.5 seems better at presenting the paradigmatic relationships than syntagmatic relationships. Two of the words in Table 7.5 (“mule” and “ox”) appear as hypothetical paradigmatic relationships for “horse” in the section titled “*sisû* (horse).” The results of PMI also match well the results of Word2vec analysis (see the section “Word2vec”), which specifically aimed to explore the paradigmatic relationships of “horse.” This was partially expected in the section titled “*sisû*, ‘horse,’” however, as the lexemes of the paradigmatic group (such as “mule” and “donkey”) appear both in similar contexts and close to each other. At least to some degree, this is because many of the occurrences of “horse” and the lexemes of its paradigmatic group appear in lists enumerating, for example, taxes or tribute from foreign lands. This explanation is also supported by the appearance of *gimru* (total) in the PMI results, as this term is often used to sum up contents in lists. It is also

supported by the fact that the PMI results exclusively show nouns—a feature that is also typical for lists.

An unexpected result of the PMI analysis is the presence of designations of horses: “Kushite,” “Mesaeen,” and “white.” “Kushite” is even the term most often connected to “horse.” This is surprising, because there are very many qualifying adjectives for horses. The assumption in the *CAD* analysis of the section “*sisû*, ‘horse,’” was that because of the large number of these terms, none of them would show up in PMI analysis. The dominance of “Kushite” might indicate that it is rarely connected to any other word apart from “horse,” whereas “white,” for example, of course has many other uses.

TABLE 7.6 NPMI results for the top fifteen suggestions for *qabû*, “to speak”

	Word	Translation	Freq <sub>AVG</sub>	NPMI <sub>AVG</sub>
1.	<i>niġak</i>	deed	26	.249
2.	<i>naqbîtu</i>	utterance	15	.247
3.	<i>teslîtu</i>	appeal	42	.204
4.	<i>bibbulu</i>	flood	15	.173
5.	<i>ibru</i>	friend	17	.159
6.	<i>qabû</i>	to say	388	.138
7.	<i>apālu</i>	to pay	32	.136
8.	<i>awātu</i>	word	97	.128
9.	<i>zakāru</i>	to speak	43	.124
10.	<i>šālu</i>	to ask	69	.121
11.	<i>bîtu&amp;criđûtu<sup>a</sup></i>	“House of Succession”	23	.113
12.	<i>epištu</i>	deed	30	.112
13.	<i>magāru</i>	to consent	29	.111
14.	<i>arazu</i>	supplication	26	.111
15.	<i>puḥru</i>	assembly	26	.109

a The symbols “&” indicate a compound word.

In the case of *qabû* (and verbs in general), using only the average NPMI scores would have produced very obscure results,<sup>49</sup> and even with the weighted

49 For example, the top ten results contain such words as *baḥru*, “boiling hot one” (.252); *anam*, “what” (.205); *tartugallu*, “hen,” from KUR DAR.LUGAL.MEŠ<sup>MUŠEN</sup>, “land of roosters” (.177); and *sittûtu*, “those remaining” (.174).



scores the results are open to some interpretation. Somewhat better results can be achieved by using very narrow collocational windows:

TABLE 7.7 *NPMI results for the top fifteen suggestions for qabû, “to speak,” when using a window size of 5*

	Word	Translation	Freq <sub>AVG</sub>	NPMI <sub>AVG</sub>
1.	<i>teslîtu</i>	appeal	27	.275
2.	<i>zakāru</i>	to speak	27	.192
3.	<i>awātu</i>	word	38	.143
4.	<i>šālu</i>	to ask	26	.138
5.	<i>šiptu</i>	incantation	20	.117
6.	<i>pû</i>	mouth	47	.116
7.	<i>mahāru</i>	face	20	.106
8.	<i>ṭēmu</i>	(fore)thought	30	.104
9.	<i>alāku</i>	to go	136	.098
10.	<i>šarru</i>	king	577	.095
11.	<i>šemû</i>	to hear	28	.094
12.	<i>qabû</i>	to say	73	.093
13.	<i>epēšu</i>	to do	75	.082
14.	<i>bēlu</i>	lord	467	.081
15.	<i>balātu</i>	life	22	.078

We then compared Tables 7.6 and 7.7 to the hypotheses formulated with the help of the *CAD* in the section titled “*qabû*, ‘to speak.’” As expected, in this section the syntagmatic relationships of *qabû* were difficult to establish with PMI, due to its wide semantic range. The results for words that appear close to *qabû* in both tables include *awātu* (word, matter) and, in Table 7.6, *magāru* (to consent). These were expected to occur with *qabû*, but they cannot be connected to any clear semantic domain or prototypical scenario.

The word *qabû*, “to speak,” could have hypothetically appeared in three prototypical scenarios: (A) to state officially (e.g., in court), (B) to order, and (C) to promise (financially). The only one of these that has a connection to PMI results is B, with the words *pû* and *šarru* appearing in Table 7.7. We should note, however, that when looking at simple average NPMI scores, the following additional words from Table 7.2 appeared in our results: *dayyānu*, “judge” (.064); *šulmu*, “well-being” (.035); and *maḥru*, “front” (.012).

In these results, we also note a verb associated with speaking: *šemû*, “to hear.” We can also see the typical participants in Mesopotamian texts involving speaking, lords (*bēlu*) and kings (*šarru*), who typically order someone to do (*epēšu*) something, such as going (*alāku*) somewhere, and who are usually appealed to (*teslītu*) by their subjects. Finally, as with “horse” above, the paradigmatic relationships seem to be represented here, or at least the appearance of *zakāru* (to speak) suggests it.

TABLE 7.8 NPMI results for the top fifteen suggestions for *danānu*, “to be strong”

	Word	Translation	Freq <sub>AVG</sub>	NPMI <sub>AVG</sub>
1.	<i>šupku</i>	foundation	13	.387
2.	<i>birtūtu</i>	function of fort	11	.382
3.	<i>enēšu</i>	to be(come) weak	14	.303
4.	<i>takālu</i>	to trust	15	.277
5.	<i>pīlu</i>	limestone	22	.273
6.	<i>ewû</i>	to become	18	.271
7.	<i>epištu</i>	deed	16	.251
8.	<i>temmēnu</i>	foundation	16	.251
9.	<i>lītu</i>	victory	12	.241
10.	<i>mušarû</i>	(royal) inscription	14	.22
11.	<i>šaṭāru</i>	to write	28	.215
12.	<i>bēlūtu</i>	rule	23	.188
13.	<i>dūru</i>	city wall	16	.188
14.	<i>šēru</i>	back	24	.177
15.	<i>nišu</i>	people	29	.135

Finally, we compared Table 7.8 with Table 7.3 (in the section titled “*danānu*, ‘to be strong, powerful’”). The word *danānu*, “to be strong,” was expected to appear in five syntagmatic domains: (A) the crown and military, (B) to guard, (C) to make valid, (D) to strengthen man-made structures, and (E) to speak or write strongly.

Interestingly, the results of Table 7.8 seem to be connected with three of the suggested syntagmatic fields of Table 7.3: namely, A, D and E. Only one of the lexemes in Table 7.3 appears explicitly—namely, “city wall”—but three other terms in Table 7.8 fit the bill for syntagmatic field D: namely, the two terms translated as “foundation” and “limestone.” Regarding the proposed field E, although we do not find *tuppu*, “tablet”; or *šapāru*, “to send,” in our results, some

words with similar meanings such as *musarû*, “(royal) inscription”; and *saṭāru*, “to write,” are present instead in Table 7.8. Finally, regarding A, there are several terms in Table 7.8 that are connected with crown and military: “function as fort,” “victory,” “(royal) inscription,” and “to rule.”

As far as *danānu* is concerned, the following expected words of syntagmatic relationships (see “*danānu* [to be strong, powerful]”) had positive NPMI values: *dūru*, “city wall” (.188); *emūqu*, “soldier, strength” (.168); *maššartu*, “to guard, watch” (.139); *nakru*, “enemy” (.109); *mātu*, “land” (.097); and *šarru*, “king” (.009). The low NPMI score of *šarru* can be explained by the word’s very high individual frequency and by the fact that Oracc does not lemmatize *dannu* as *danānu*, which prevents its most obvious bigram *šarru dannu*, “a mighty king,” from showing up in our results. In general, several collocates seem to come from different texts describing the strengthening of structures or conquering new lands. The antonym *enēšu* can be explained by formulas such as: “I strengthened it, so that it would not become weak.” For paradigmatic relationships, as outlined based on the CAD in “*danānu*, ‘to be strong, powerful,’” there is little evidence in the PMI results.

### Word2vec

Whereas PMI might suggest syntagmatic semantic fields, Word2vec should be highlighting paradigmatic relationships of the words of interest (see Figs. 7.1 and 7.2 above). Word2vec is one of the methods in the field of Natural Language Processing (NLP) that can map vocabulary units to vectors of real numbers.<sup>50</sup> Word2vec models plot a unique vector in space for each lexeme in the corpus. The models align these vectors in such a way that words appearing in similar linguistic contexts cluster near each other, thus giving us information on possible paradigmatic groups of words. The text to be analyzed is converted via a hidden layer to a multidimensional output vector that gives the probability distribution of the words.<sup>51</sup> Word2vec uses so-called ANNs to predict word co-occurrence.

Word2vec was developed by Tomas Mikolov and his team in 2013,<sup>52</sup> but the underlying idea that “a word is characterized by the company it keeps” was popularized as early as the 1950s by John R. Firth,<sup>53</sup> who in turn was influenced

50 Mikolov et al. 2013a; Mikolov, Yih, and Zweig 2013.

51 Mikolov, Yih, and Zweig 2013.

52 Mikolov et al. 2013a; Mikolov, Yih, and Zweig 2013.

53 Firth 1957.

by early anthropological studies. The ideas presented by Firth definitely have the essence of other branches of the humanities emphasizing the importance of context in any endeavor of researching culture and language.<sup>54</sup> Word2vec is an open-source toolkit for producing word vectors and querying semantic relationships between words.<sup>55</sup>

There are two alternative models one can choose from when using Word2vec: the Continuous Skip-gram model and the Continuous Bag-of-Words (CBOW) model.<sup>56</sup> When training the data, the Continuous Skip-gram model predicts the words that may appear near the target word,<sup>57</sup> whereas the CBOW model uses the various words that can appear in the same context as the target word for the prediction. A detailed explanation of these methods is beyond the scope of this paper.

The last stage of converting the hidden layer into the output vector is computationally expensive and time-consuming. Hence, there are several ways of approximating this step. In Word2vec, one can choose between two such methods for creating the output vector: hierarchical softmax and negative sampling. In hierarchical softmax, the values for the words are stored in such a way that not all of the words have to be processed when calculating the probability of a word. Negative sampling is a procedure in which only a small part of the model's weights is updated. According to Mikolov,<sup>58</sup> these methods not only speed up the training, but they also improve the accuracy of the prediction. When looking for words appearing together with other words, one is not interested in the most frequent words, such as "the" in English. According to Mikolov,<sup>59</sup> a method called subsampling efficiently reduces the effect these very frequent words have on the training. In Word2vec, it is also possible to

---

54 Firth 1957. See also the discussion on emic approaches above in "Emic Approach and Linguistic Departure Points."

55 Mikolov et al. 2013a. The original link is not working anymore, but many copies of the toolkit can still be found on Github. The kit we used was downloaded on November 14, 2016 from <<https://github.com/dav/word2vec>> (accessed July 7, 2017).

56 Mikolov et al. 2013a. Continuous Bag-of-Words: one of two alternative models one can choose from when training data in Word2vec. While training, the CBOW model uses the various words that can appear in the same context as the target words for the prediction of semantic relationships. For further explanation of the bag-of-words model, see in this volume, Monroe, 270–272.

57 Continuous Skip-gram model: one of two alternative models one can choose from when training data in Word2vec. While training, the Continuous Skip-gram model predicts the words that may appear near the target word.

58 Mikolov et al. 2013a.

59 Mikolov et al. 2013a.

choose the window size of the context in which the words are considered. For example, a window size of 5 would consider five words before and five after a word. However, in practice, Word2vec picks a random number  $R$  within the range chosen (1 to 5 in the previous example), and it considers  $R$  words on each side of the target word.<sup>60</sup>

Depending on the values chosen for the different parameters used with Word2vec, the resulting multidimensional vectors differ from each other. Furthermore, because of randomness and subsampling, the results also change from one computing session to another. Thus, the results presented here are just an approximation. We experimented with different parameters and noticed that the dimension of the vector has a smaller effect on the results than the threshold used for subsampling. According to Mikolov,<sup>61</sup> a typical threshold for this would be .00001. That would mean that words with a relative frequency greater than this might be disregarded. As our texts have only 12,067 different word types, using a small threshold did not give good results. In the tests conducted by Mikolov and his team,<sup>62</sup> negative sampling worked better than hierarchical softmax. Therefore, we used negative sampling. Although values of 5–20 are recommended by Mikolov and his team for negative sampling in small training sets and 2–5 in large ones, we found that using too many samples for training did not work well with our data. We used the following parameters in our analysis: vector dimensionality 200, negative sampling with 4 samples, subsampling with threshold .01, and window size 7. We ran the script provided in the Word2vec toolkit with these parameters 500 times, using the CBOW model architecture and our own text file. After each run, the word vector produced by Word2vec was discarded, as we wanted the tool to create a new one with different random samples. From each run, we queried for the 50 words that were semantically closest to each of the test words (*sisû*, *qabû*, and *danānu*). After the runs, we scored the words received from Word2vec, so that the closest match to a test word was scored 1 and the furthest 50. For every run a word that was counted at least once not appearing in the list, a penalty of 51 points was scored. The average score was then counted for each word. In this way, we ended up with a list of words that were closest to the target word in all 500 runs. Ten of the closest words for each of the target words *sisû*, *qabû*, and *danānu* can be seen in Tables 7.9–11. The numbers after the words indicate the average placement that a particular word appeared in during the 500 runs exe-

60 Mikolov et al. 2013b; Levy and Goldberg 2014.

61 Mikolov et al. 2013a.

62 Mikolov et al. 2013a.

cuted, thus reflecting how high in the list of words returned by Word2vec that word was on average.

TABLE 7.9 *Word2vec results for the top ten suggestions for sisû, “horse”*

Word	Score	Translation
<i>kūdānu</i>	1.3	mule
<i>šumbu</i>	1.7	wheel
<i>parû</i>	3.3	mule
<i>gammalu</i>	5.6	camel
<i>udru</i>	5.8	Bactrian camel
<i>narkabtu</i>	7.1	chariot
<i>šēnu</i>	7.9	flock
<i>namrāšu</i>	8.6	hardship
<i>Kusaya</i>	9.3	Kushite
<i>mānu</i>	9.7	counting

As expected, the Word2vec analysis for “horse” found the words that appear in similar contexts as “horse”—that is to say, potential paradigmatic relationships (see “Emic Approach and Linguistic Departure Points”). Lexeme to lexeme, the results are not an exact match, and only 1 out of 4 suggestions (*kudānu*) made its way onto the Word2vec list. However, most of the other words appearing in the Word2vec list clearly belong to the same conceptual domain. In addition to *kudānu*, “mule,” which was suggested in traditional analysis, we find two words for “camel” and a different word for “mule,” for example. From the words suggested by the traditional method, we find *alpu*, “ox,” in position 11, with an average score of 9.7, and *imēru*, “donkey,” in position 12, with a score of 15.9. The word *atānu* can be found in shared 179th position from a total of 372 words. The average scores (1.3–9.7) of the ten closest words to *sisû*, “horse,” reveal that these words appeared very high on the list on almost every run.

A little surprisingly, we find a clear indication here of the semantic domain of war as well, with “chariot” and “wheel” being mentioned. These match the proposed syntagmatic group for “horse” (see “*sisû*, ‘horse’”), but we would have expected to find them surrounding “horse” in prototypical scenarios in the PMI analysis (see “Pointwise Mutual Information”), not as words that appear in a similar environment. Such results could suggest that separating syntagmatic and paradigmatic relationships statistically might be more difficult than we anticipated.

TABLE 7.10 *Word2vec results for the top ten suggestions for qabû, “speak”*

Word	Score	Translation
<i>magāru</i>	2.7	to consent
<i>awātu</i>	2.9	word
<i>wadû</i>	2.9	to know
<i>hasāsu</i>	5.2	to be(come) conscious
<i>kamsu</i>	9.3	gathered
<i>šālu</i>	12.3	to ask
<i>šipirtu</i>	14.9	message
<i>mītu</i>	19.9	dead
<i>hibiltu</i>	21.2	wrongdoing
<i>abāku</i>	21.7	to overturn

The results of Word2vec for *qabû* were generally quite similar to the PMI results. Interestingly, the Word2vec analysis did not show any other words that are traditionally translated as “to speak.” This might support the idea that each of the verbs related to speaking (e.g., *dabābu*, *zakāru*, and perhaps *awû*) has a semantic domain of its own, as hypothesized in “*sisû*, ‘horse’” above.

The average scores (2.7–21.7) in the lists of words given by Word2vec for *qabû* are quite a lot higher than for *sisû*, which might indicate that there is a larger variety of words appearing close to the word *qabû* than to *sisû*. Outside the list of the top ten words found, the word *dabābu*, “to speak,” is on the list in the shared 91st place, with the average score of 49.1. The word *awû*, “to speak,” was present on only two runs. The word *zakāru*, however, is not on the list of closest words to *qabû*; this is interesting, as it does show up in Tables 7.6 and 7.7 of the PMI results that presumably suggest syntagmatic categories.

The paradigmatic relationships suggested by Word2vec for *danānu* bear some similarity to the proposals based on the *CAD* (see “*danānu*, ‘to be strong, powerful’”). Regarding words suggested by the analysis of the *CAD*, one finds the word *gašru* in the first position in Table 7.11. Words with the root *dnn* are present in the form of *dandannu*, “all-powerful.”<sup>63</sup> The words *gašāru* and *emūqu* were not found in the analysis. When we compare this to the syntagmatic proposals of PMI, we find that the lexemes are fairly dissimilar, as there is only one word that appears directly in both: *lītu*, “victory.”

63 The word *dunnu* appeared in only two runs.

TABLE 7.11 *Word2vec results for the top ten suggestions for danānu (dnn), “to be strong, powerful”*

Word	Score	Translation
<i>gašru</i>	5.2	strong
<i>šeriktu</i>	13.3	present
<i>ilūtu</i>	14.9	divinity
<i>lītu</i>	16.6	victory
<i>qurdu</i>	17.6	warriorhood
<i>liptu</i>	18.4	undertaking
<i>dandannu</i>	19.3	all-powerful
<i>rā’imu</i>	19.5	loving
<i>agû</i>	21.4	crown
<i>rimītu</i>	21.4	residence

The average scores (5.2–21.4) for the words returned for *danānu*, “to be strong,” are even higher than the ones for *qabû*. It is difficult at this point in the research to say whether this is due to the fact that *danānu* was a less frequently used word than the other two. It appears in the test corpus only 221 times, whereas *sisû* appears 686 times and *qabû* appears 2,353 times. This could also be the reason why the results of Word2vec analysis for *danānu* were more varied than for *sisû* or *qabû*.

### Discussions and Future Prospects

To sum up the results, our aim was to examine the usefulness of the PMI and Word2vec methods for constructing syntagmatic and paradigmatic semantic fields for Akkadian by using *sisû*, “horse”; *qabû*, “to speak”; and *danānu*, “to be strong, powerful,” as test cases. After a discussion of theory, methodology, and our sources (“Theoretical and Methodological Background and Source Material” and “Emic Approach and Linguistic Departure Points”), we suggested both paradigmatic and syntagmatic semantic domains of Akkadian based on the current semantic knowledge of Assyriology (“Traditional Examples of Semantic Fields”). These suggestions were compared with the results of PMI analysis for syntagmatic relationships (“Pointwise Mutual Information”) and Word2vec for potential paradigmatic relationships (“Word2vec”). These first results of our research group offer some promise that quantitative data on the



connections between individual lexemes can indeed help Assyriologists understand the ancient contexts better. At the very least, the detailed suggestions for semantic domains generated with the help of PMI and Word2vec enable Assyriologists to look in the right direction. Any detailed Assyriological work on any specific semantic field will of course need to be backed up with solid philological work on the actual word occurrences in primary sources. We see the importance of the current contribution not so much in the light that it can shed on our three sample lexemes: *sisû*, “horse”; *qabû*, “to speak”; and *danānu*, “to be strong, powerful.” Rather, we see its main value in showing that methods like PMI and Word2vec can be profitably used to help define the semantics of individual words. For language-technological work, the interconnections between results gleaned with PMI and results from Word2vec suggest that automatically differentiating between paradigmatic and syntagmatic semantic fields will need more research.

As for the future plans of our research group, we intend to proceed with a diachronic approach and also analyze our results by text genre. Such comparisons may be able to illuminate possible changes in semantic domains. The written history of Mesopotamia in the Akkadian language covers roughly the years 2300–300 BCE. The dialects of Akkadian during these centuries have been studied, of course, but not from the perspectives suggested here: semantic domains and long-term continuities and changes in these domains. Thus, an interesting and important research project would be to do similar analyses as here, but only for parts of our corpus. As we are dealing with several different dialects of Akkadian and several different text genres, it would be interesting to see what kinds of differences there are between them and how such differences might be explained from the perspective of the history of the Akkadian language. For example, the possibility of identifying language contacts and influences from other languages is intriguing.

For archaeology, understanding the paradigmatic relationships better would help interpretations of the material records. From a methodological perspective, understanding the methods of word sense induction (paradigmatic and syntagmatic concepts) and how they can be applied could be a useful tool when developing semiotic models for interpreting archaeological records.

Finally, further results might also offer valuable input for a comparison of languages: do the contextual semantic frames of Akkadian match those constructed for modern languages or for biblical Hebrew? Furthermore, they might shed light on the questions: how much of the contextual semantic frames stay the same from language to language and to what degree are languages really unique cognitive constructions?

On a practical level, we might need to take better into account the specific challenges presented by the corpus. Many of our texts are damaged or only fragments of the original texts. Ideally, we would find a way to model the missing pieces of information in such a way that the methods designed for complete texts or sentences will still be able to provide meaningful results. However, at least for the results presented here, our rather simple solution of excluding damaged parts from the analysis seems to have worked, at least to some extent.

Another challenge is that many of the culturally significant texts, such as the Epic of Gilgameš or the royal inscriptions, were found in several partially preserved copies. The presence of duplicate or near-duplicate texts could be considered in a more nuanced way when using statistical methods. Members of our team have encountered the same problem when harvesting Finno-Ugric texts from the internet.<sup>64</sup> For this paper, we have simply treated the composite master text as “the text,” without taking into account manuscript variants.

The amount of information created by looking for semantic fields with the methods of computational linguistics is immense. That is why one of the long-term goals of our research group is to create a web interface for browsing the different points of view created by varying the time, type, quality, etc., of the texts being analyzed. As filters, we may be able to use some of the metadata that accompanies the texts in the corpus (on the corpus metadata, see the discussion in “Theoretical and Methodological Background and Source Material”). Such an interface could perhaps also be used for other, relatively small text corpora, especially those in extinct and endangered languages.

In terms of visualizing the information on semantic domains, the semantic relationships between lexemes can also be conceptualized as networks, with nodes and edges. Thus, the changing relationships between semantic fields throughout the timeframes could be demonstrated via Gephi or some other software designed to visualize and analyze large networks.<sup>65</sup>

These are some of the team’s first results. As we have years of work ahead of us, we look forward to continuing to develop these approaches further. The results will be open to the research community, and all the software tools built in the project will be published as open-source. In the future, the research will benefit from additional sources, as texts are continuously being added to Oracc.

---

64 Jauhiainen, Jauhiainen, and Lindén 2015.

65 Gephi: <<https://gephi.org/about/>> (accessed July 1, 2017). See also Pagé-Perron, 209.

## References

- Arens, William, and Ivan Karp. 1989. "Introduction." In *Creativity of Power: Cosmology and Action in African Societies*, edited by William Arens and Ivan Karp, xi–xxix. Washington, DC: Smithsonian Institution.
- Baroni, Marco, Georgiana Dinu, and Germán Kruszewski. 2014. "Don't Count, Predict! A Systematic Comparison of Context-Counting vs. Context-Predicting Semantic Vectors." In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 22–27 June 2014, Baltimore, Maryland, USA*, vol. 1, 238–247. Stroudsburg, PA: Association for Computational Linguistics.
- Barr, James. 1961. *The Semantics of Biblical Language*. London: Oxford University Press.
- Bird, Steven, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. Sebastopol, CA: O'Reilly Media Inc.
- Black, Jeremy, Andrew George, and John Nicholas Postgate. 2000. *A Concise Dictionary of Akkadian*, 2nd corrected ed. SANTAG 5. Wiesbaden: Harrassowitz.
- Blois, Reinier de. 2008a. "New Tools and Methodologies for Biblical Lexicography." In *Foundations for Syriac Lexicography III: Colloquia of the International Syriac Language Project*, edited by Janet Dyk and Wido van Peursen, 203–216. Perspectives on Syriac Linguistics 4. Piscataway, NJ: Gorgias Press.
- Blois, Reinier de. 2008b. "Semantic Domains for Biblical Greek: Louw and Nida's Framework Evaluated from a Cognitive Perspective." In *Foundations for Syriac Lexicography III: Colloquia of the International Syriac Language Project*, edited by Janet Dyk and Wido van Peursen, 265–278. Perspectives on Syriac Linguistics 4. Piscataway, NJ: Gorgias Press.
- Bouma, Gerlof. 2009. "Normalized (Pointwise) Mutual Information in Collocation Extraction." In *Von der Form zur Bedeutung: Texte automatisch verarbeiten / From Form to Meaning: Processing Texts Automatically, Proceedings of the Biennial GSCl Conference*, edited by Christian Chiacos, Richard Eckart de Castilho, and Manfred Stede, 31–40. Tübingen: Narr.
- Church, Kenneth Ward, and Patrick Hanks. 1989. "Word Association Norms, Mutual Information, and Lexicography." In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics, 26–29 June 1989, Vancouver, British Columbia, Canada*, 76–83. Stroudsburg, PA: Association for Computational Linguistics.
- Dirven, René, and Marjolijn Verspoor, eds. 2004. *Cognitive Exploration of Language and Linguistics*. Amsterdam: J. Benjamins.
- Eriksen, Thomas H. 2010. *Small Places, Large Issues: An Introduction to Social and Cultural Anthropology*, 3rd ed. Anthropology, Culture and Society. London: Pluto Press.
- Firth, John R. 1957. *Studies in Linguistic Analysis*. Oxford: Blackwell.
- Fleisch, Axel. 2007. "How Cognitive Semantics Relate to Comparative Linguistics: A Case Study from Nguni." In *Viva Africa 2007: Proceedings of the 2nd International Conference*

- on *African Studies*, edited by Tomáš Machalík and Jan Záhorský, 39–53. Pilsen: University of West Bohemia.
- Geeraerts, Dirk. 2010. *Theories of Lexical Semantics*. Oxford: Oxford University Press.
- Jauhiainen, Heidi, Tommi Jauhiainen, and Krister Lindén. 2015. “The Finno-Ugric Languages and The Internet Project.” In *First International Workshop on Computational Linguistics for Uralic Languages*, edited by Tommi A. Pirinen, Francis M. Tyers, and Trond Trosterud, 87–98. Septentrio Conference Series 2. Tromsø, Norway: Septentrio Academic Publishing.
- Kataja, Laura and Kimmo Koskeniemi. 1988. “Finite-state Description of Semitic Morphology: A Case Study of Ancient Accadian.” In *COLING Bdapest: Proceedings of the 12th International Conference on Computational Linguistics, Budapest 22-27.08.1988*, edited by Dénes Vargha, 313–315. Budapest: John von Neumann Society for Computing Sciences.
- Lakoff, George. 1987. *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*. Chicago: University of Chicago Press.
- Landsberger, Benno. 1976. *The Conceptual Autonomy of the Babylonian World*. Translated by Benjamin Foster, Thorkild Jacobsen, and Heinrich von Siebenthal. Monographs on the Ancient Near East 1 (4). Malibu, CA: Undena. Originally published in 1926 as *Die Eigenbegrifflichkeit der babylonischen Welt: Ein Vortrag*. Sonderausdruck aus *Islamica* 2 (3), 355–372. Leipzig: Asia Major.
- Levinson, Stephen C. 2003. *Space in Language and Cognition: Explorations in Cognitive Diversity*. Language, Culture, and Cognition 5. Cambridge: Cambridge University Press.
- Levy, Omer and Yoav Goldberg. 2014. “Dependency-Based Word Embeddings.” In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 22–27 June 2014, Baltimore, Maryland, USA*, vol. 1, 302–308. Stroudsburg, PA: Association for Computational Linguistics.
- Manning Christopher D., and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.
- Mikolov, Thomas, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. “Distributed Representations of Words and Phrases and their Compositionality.” In *Advances in Neural Information Processing Systems 26 (NIPS 2013): 27th Annual Conference on Neural Information Processing Systems 2013; December 5 - 10, Lake Tahoe, Nevada*, edited by Christopher J.C. Burges, Léon Bottou, Max Welling, Zoubin Ghahramani, and Kilian Q. Weinberger. Red Hook, NY: Curran. <<https://arxiv.org/abs/1310.4546>>.
- Mikolov, Thomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. “Efficient Estimation of Word Representations in Vector Space.” Last modified September 7, 2013. <<https://arxiv.org/abs/1301.3781>>.

- Mikolov, Thomas, Wen-tau Yih, and Geoffrey Zweig. 2013. "Linguistic Regularities in Continuous Space Word Representations." In *Proceedings of NAACL-HLT 2013, Atlanta, Georgia, USA, 9–14 June 2013*, 746–751. Stroudsburg, PA: Association for Computational Linguistics.
- Nurmikko-Fuller, Terhi. 2016. "Publishing Sumerian Literature on the Semantic Web." Paper presented at the ASOR annual meeting, San Antonio, TX.
- Pantel, Patrick, and Dekang Lin. 2002. "Discovering Word Senses from Text." In *Proceedings of the SIGKDD Conference on Knowledge Discovery and Data Mining, New York, NY, USA—August 24–27, 2014*, edited by Sofus Macskassy and Claudia Perlich, 613–619. New York: Association for Computing Machinery.
- Reynolds, Frances. 2003. *The Babylonian Correspondence of Esarhaddon and Letters to Assurbanipal and Sin-šarru-iškun from Northern and Central Babylonia*. SAA 18. Helsinki: Helsinki University Press.
- Roth, Martha T., ed. 1956–2011. *The Assyrian Dictionary of the Oriental Institute of the University of Chicago (CAD)*, 26 volumes. Chicago: The Oriental Institute of the University of Chicago.
- Streck, Michael P. 2010. *Großes Fach Altorientalistik: Der Umfang des keilschriftlichen Textkorpus*. MDOG 142: 35–58.
- Svärd, Saana. 2012. "Power and Women in the Neo-Assyrian Palaces." PhD diss., University of Helsinki.
- Svärd, Saana. 2015. *Women and Power in Neo-Assyrian Palaces*. SAAS 23. Helsinki: Neo-Assyrian Text Corpus Project.
- Tyndall, Stephen. 2012. "Towards Automatically Assembling Hittite-Language Cuneiform Tablet Fragments into Larger Texts." In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, July 8–14, 2012, Jeju Island, Korea*, vol. 2, 243–247. Stroudsburg, PA: Association for Computational Linguistics.
- w3c. "w3c Semantic Web Activity." Last modified December 11, 2013. <<https://www.w3.org/2001/sw/>>.
- Waerzeggers, Caroline. Forthcoming. "Social Network Analysis of Cuneiform Archives: A New Approach." In *Proceedings of the Second START Conference in Vienna*, edited by Heather D. Baker and Michael Jursa.
- Wolde, Ellen van. 2009. *Reframing Biblical Studies: When Language and Text Meet Culture, Cognition, and Context*. Winona Lake, IN: Eisenbrauns.

# Using Quantitative Methods for Measuring Inter-Textual Relations in Cuneiform

*M. Willis Monroe*

## Introduction

One of the many benefits of quantitative methods in the digital humanities is their ability to test intuitions and assumptions about a body of textual material quickly and efficiently. This is often most useful when the corpus is extremely large, and regular methods of reading or analysis are not possible. While quantitative methods can provide a level of certainty to results or conclusions, their use in exploratory hermeneutics is equally valuable.<sup>1</sup> This chapter intends to explore the viability of these methods in order to investigate a small corpus of Babylonian astrological material concerned with the association between ingredients used in medical treatment and the signs of the zodiac. The methods employed are relatively simple and can be performed using free open-source software, yet they allow for an iterative process of knowledge creation and investigation that produces multiple results at different stages of the research.<sup>2</sup> In the spirit of this volume, a significant amount of explanation will be included to break down the techniques and methodology behind the processing and analysis of the texts.

The underlying goal of this analysis is to understand the scholarly practices behind the composition and editing of the some unique examples of Late Babylonian astrology. The period under Achaemenid and Hellenistic rule in Babylonia (late sixth to second centuries BCE) saw a flourishing of scholarly activity recorded on cuneiform tablets in southern Iraq. Astronomy and astrology were, in particular, realms experimentation and rapid development. The texts under consideration in this chapter comprise one of many new types of astrology, each borrowing and building on previous forms of knowledge. The

---

1 “There is a tendency in debates to reduce the potential of computation to a methodology of quantification. The nature of digital humanities is hybrid, however, and there is not an a priori discontinuity with the hermeneutic traditions” (Zundert 2015, 340).

2 For iterative processes (developed in C++) and for projects on object typologies, especially to set up qualitative-data processing, see in this volume, Martino and Martino, 120.

material represents a collection of tables of astrological significance; the area under investigation is a row of medical ingredients linked with pairs of zodiacal signs. The text is called the “Micro-zodiac,” so named for its use of a smaller set of zodiacal signs to subdivide the regular 12 signs of the traditional zodiac.<sup>3</sup> The traditional Babylonian zodiac is the antecedent to the western zodiacal signs, as illustrated in the following table:

TABLE 8.1    *Traditional Babylonian zodiac*

Babylonian sign	Translation	Modern equivalent
ḪUN	Hired man	Aries
GU <sub>4</sub> .AN.NA	Bull of heaven	Taurus
MAŠ.MAŠ	Twins	Gemini
ALLA	Crab	Cancer
A	Lion	Leo
ABSIN	Furrow	Virgo
RIN <sub>2</sub>	Scales	Libra
GIR <sub>2</sub> .TAB	Scorpion	Scorpio
PA	Pabilsag	Sagittarius
MAŠ <sub>2</sub>	Goat-fish	Capricorn
GU	Great one	Aquarius
ZIB.ME	Tails	Pisces

The methods employed here are often used with much larger and more complete corpora that represent bodies of material beyond the ability of a single scholar to read closely. In those cases, the methods allow for a process of analysis not otherwise feasible. Widely used in the field of Early Modern Literature, computer-aided textual analysis can answer questions about themes or the usage of terms across a corpus containing millions of words. The corpus with which this paper is concerned differs in that it is small enough to read closely, but it is lacking in coherence. The individual words within the corpus are grouped in clusters on the text and lack any linguistic syntax. The damaged and incomplete nature of the corpus makes traditional forms of reading and analysis difficult. So a quantitative method allows for more concrete conclu-

3 The Micro-zodiac series was the subject of my dissertation research (Monroe 2016).



sions or insight into the organization of the text obscured by its format.<sup>4</sup> Through computer-aided textual analysis, we can visualize what would otherwise be mere hunches gleaned from a close reading of the texts.

Modern techniques in textual analysis are often used to perform a so-called “close reading” of a textual corpus, a method of analyzing the constituent parts and makeup of a text to reach a deeper understanding.<sup>5</sup> By leveraging the analytical power of digital methods, the researcher can survey an entire corpus of material for a particular term or area of interest, or, in other cases, allow the material to speak for itself.<sup>6</sup> This chapter will attempt to show another use for these methods. The corpus used here is smaller than most of those used in other forms of textual analysis, but the disjointed nature of its content is immediately analyzable through similar methods. The traditional methods of textual analysis often remove words or terms from their local context and analyze them as a whole, looking for patterns among groups of words represented in documents or manuscripts. In the case of the Micro-zodiac, the first step is already completed; the lack of any syntactic grouping allows for the bunches of terms within each cell of the tables to represent a “document” in our analysis.

## Theory

Much of the theory behind the methods employed in this chapter in particular and in the field of digital humanities in general were first concretely laid out by Franco Moretti in his book *Graph, Maps, Trees*, in which he advocated for a form of “distant reading.”<sup>7</sup> The “distant” in “distant reading” need not refer to a lack of physical proximity between the reader and text; it refers, rather, to methods of analysis that involve mapping and the distance between objects of the text (or texts). The reader or scholar is still very close to the text, in some cases using minute elements of the text as the unit of their analysis. The overall

---

4 The corpus of the Micro-zodiac texts consists of 15 known tablets of varying degrees of preservation. The most complete have more than 20 cells (the unit of analysis for this paper) of ingredients on them. The dataset contains 476 individual ingredients pulled from the preserved cells on the known Micro-zodiac tablets.

5 Jänicke et al. 2016, 2.

6 These techniques are increasing rapidly in frequency of use within the digital humanities. A recent survey of journal articles charted a large increase in papers from 2005 to 2014 that used techniques classified as “distant reading” (Jänicke et al. 2015, fig. 3).

7 Moretti 2007.



aim of the methodology employed here is to produce a graph used to visualize the connections between cells. A graph is made of nodes and edges, both of which are determined by the nature of the evidence. In many cases nodes represent texts within a corpus or documents within a genre. Edges represent connections between nodes, often formed by shared terms, words, or other meaningful units of analysis. The choice of what represents a node or an edge can change as the analysis proceeds; in the method that follows nodes and edges are processed one way and then inverted for another view of the same data.

The methods of distant reading are most commonly divided into two categories, “supervised” and “unsupervised.” The distinction between the two is at times fluid, and there are many overlapping terms that encompass one or both,<sup>8</sup> but at its simplest level, the difference between “supervised” and “unsupervised” methods of text analysis involves whether or not the researcher has chosen to “classify” the material. The processing of classifying textual data involves identifying important features, for example, length, word counts, and format, that can be represented as numbers and ascribing these features to certain classes of documents. Analysis with a text that has already been partially classified is considered to be supervised, whereas an analysis with an unclassified dataset is unsupervised. For instance, a study might involve attempting to associate similar authors with an unattributed new trove of anonymous novels in order to discover the unnamed author. A supervised method of analysis would involve classifying sets of novels (by their features) for which the authors are known, comparing the new material against the known corpus, and then assigning it to one of the existing groups.<sup>9</sup> This is a supervised method because the researcher is supervising, commonly called “training,” the algorithm by providing it with known data prior to performing analysis on the unknown material. The unsupervised method would involve not classifying any of the existing corpus, but, rather, letting the algorithm decide which records (i.e., novels) out of the entire corpus belonged together. Here, defining features is also important, but it is crucial to note that the documents are never grouped into classes based on these features; that task is left to the algorithm with the hope

---

8 “Today, *data mining* often implies unsupervised learning, whereas *machine learning* is more commonly applied to supervised learning processes. But this boundary can be drawn in several different ways: sometimes *data mining* names the practice that corresponds to *machine learning*’s theory” (Jockers and Underwood 2015, 292).

9 Supervised method: more generally, a method in which the researcher has chosen to “classify” part of the material before the analysis.

that the known works of certain authors would be found in similar groups. The benefit of the unsupervised method here, however, is that it breaks what are otherwise researcher-oriented biases and presents an “objective” view of the material, free from prior classification. Perhaps the unsupervised method detected that a group of novels thought to belong to one author most closely match the unknown material in our example. The conclusion is up to the researcher here, but the unsupervised method can offer insight into a dataset without any preconceived notions of structure or definition.<sup>10</sup> Unsupervised methods are commonly used to explore a dataset, after which a supervised method can be used to hone in on a particular area of interest.<sup>11</sup>

### Text Background

The Micro-zodiac material comes from Hellenistic Babylonia, where it was written by scribes educated and working in the Late Babylonian scholarly communities of southern Mesopotamia. In this period local rule in Mesopotamia had been eclipsed, and the cities of Babylonia were part of the wider Seleucid Empire. The scribes were attached to the temples through cultic roles and family ties. These scholars wrote extensively on many forms of traditional Mesopotamian scholarly knowledge, from divination to ritual and from mythology to astrology. They have taken their place in our modern history of science as skilled mathematical astronomers; their methods for calculating the planetary ephemerides and solar and lunar eclipses were renowned throughout the ancient world. Terms and constants computed in Babylonia show up in other traditions, and the ethnonym “Chaldean,” referring to Babylonian scholars, became synonymous with ancient astronomers/astrologers of the Middle East.<sup>12</sup>

---

<sup>10</sup> An important caveat here is that the output generated by unsupervised methods is only as good and as accurate as the input data. An unsupervised method could very well suggest theories and connections between points in the data that are not borne out by the evidence because particularly crucial differentiating variables were left out of the analysis. See in this volume, two other examples of unsupervised methods: Artificial Neural Network (ANN) algorithms applied to landscape archaeology by Ramazzotti (63–65), and the multivariate analysis of figurines by Martino and Martino (118). As for a supervised example (on cylinder seal imagery), see also in this volume, Ludovico (92–95).

<sup>11</sup> Jockers and Underwood 2015, 294–295.

<sup>12</sup> Useful introductions to Babylonian astronomy and astrology can be found in Rochberg (2004) and Steele (2008).

Along with performing mathematical calculations, these same scribes wrote a large corpus of astrology that exhibits a similar complexity and development of thought. These texts take the configuration of planets and stars and compute new meaning and significance based on their positions and movement. Much of their work borrowed from a long tradition of astrology going back roughly two millennia to the earliest forms of Mesopotamian divination. A large astrological compendium of omens, called *Enūma Anu Enlil*, formed the basis of much of their education. During this period, however, these ideas were modified and further developed, producing new types of texts that experimented with novel methods of association to produce new meaning. The Micro-zodiac series in particular shows how threads of previous forms of knowledge were compiled and woven together using the recently invented zodiac as their organizing principle.

The Micro-zodiac tablets come from the cities of Babylon and Uruk and date roughly to the third and second centuries BCE.<sup>13</sup> Each tablet preserves one or more (most have two) tables of astrological knowledge. It is important to note that these are formal tables demarcated by incised lines separating the columns and rows from each other. Often the tables have a header containing information about celestial omens related to the lunar eclipse. The examples from Uruk are known for their fabulous imagery of the zodiacal signs located in a horizontal band above the table itself. The material is organized around the zodiac—hence the previously mentioned imager—and each table is associated with one of the 12 signs of the zodiac. The twist in the Micro-zodiac is the subdivision of each sign into twelve more signs, each of which governs only  $2.5^\circ$  of the entire sky or year.<sup>14</sup> Each of the subdivisions represents one of the twelve columns within each table. In a sense the entire series is a large spreadsheet, with each table representing one-twelfth of the overall content. Tablets from the series generally have two tables, and therefore one-sixth of the entire text. The two tables are always in sequence, starting with an odd-numbered sign on the obverse of the tablet and the subsequent even-numbered sign on the reverse. Returning to the organization, each column of the entire text can

---

13 The known exemplars of the series are currently housed in the Iraq Museum in Baghdad, The British Museum in London, the Louvre in Paris, and the Vorderasiatisches Museum in Berlin. Only one exemplar is derived from controlled excavation (by the Germans in Uruk). The rest are from uncertain provenance collected during the late nineteenth and early twentieth century by the aforementioned museums. It is reasonably certain that the known exemplars all come originally from either Babylon or Uruk.

14  $1^\circ$  of the Sun's motion through the zodiac is roughly equivalent to 1 day of the ideal year. The sun travels  $360^\circ$  around the earth in a year, and there are 360 days in an ideal Babylonian calendrical year.

be identified by the zodiacal sign to which its table is assigned, hereafter the “Major sign,” and the zodiacal sign to which the individual column has been assigned, hereafter the “minor sign.” In simpler terms, each column shares a Major sign with the other 11 columns in its table and has a minor sign that is unique within its table. This system of Major-minor sign pairs creates a unique pattern of 144 couples, each of which defines the location of a column within the entire text.

The Micro-zodiac table itself contains four major rows of content: medical ingredients, celestial divination, a cultic calendar, and daily advice. Each row has many parallels in existing textual traditions; the final row, concerning daily advice, for example, borrows much of its content from the well-known hemerological texts, including the Babylonian Almanac. For this paper, however, only the first row, the medical ingredients, will factor into the analysis.<sup>15</sup> The rows interact with the organization of the columns in different ways. The middle two rows, divination and cult, repeat their content every twelve columns. This means that, in a sense, their contents only depend on the latter of the two zodiacal signs, the minor sign, which indicates the column. This pattern restricts the total amount of content in these two rows to only 12 possibilities.

The first and final row, the medical material and daily advice, respectively, do not show any discernable pattern. Under the same exact column on another text, the two cells will contain the same material, i.e., material found under one pair of signs will be the same as material found under the same pair on another tablet. However, the material does not repeat on each table under the same minor sign, as it does with the middle two rows. The fact that material is found on a different table under the same unique pairing of major and minor signs suggests that the inclusion of material in this row does follow a predictable pattern within the context of the Micro-zodiac.

Closely related to the Micro-zodiac texts are another genre of astrological material called Calendar Texts; these tablets contain much the same information organized under the same system as the Micro-zodiac but are structured in a slightly different way. They factor into the analysis below because the content of the medical-ingredients section follows the same rules and parallels the Micro-zodiac material.

The medical ingredient cells contain four to six individual items. The first is always a location, either a city or a temple. The next three ingredients are always found in the following order: wood, plant, stone. However, while some texts stick to a rigid set of four total ingredients, others might have double the number of ingredients in one or more of these three categories. Often there are

---

15 A good overview of Babylonian medicine can be found in Geller 2010.

two ingredients listed for the final category, stone. Below is an excerpt from the middle of the Leo side of tablet VAT 7847 (text 7 in the corpus).<sup>16</sup> Here, because of the location of Leo in the zodiac (the fifth sign), the boundary between Pisces and Aries (the last and first signs, respectively) is found in the physical middle of the tablet itself. The two cells list cultic buildings first, the Urinnu and the Ekur in Nippur. Next they list a sequence of ingredients following the paradigm of place, wood, plant, and stone

TABLE 8.2 *List of ingredients*

Leo-Pisces	Leo-Aries
E <sub>2</sub> U <sub>4</sub> .RI <sub>2</sub> .IN	NIBRU <sup>ki</sup> E <sub>2</sub> .KUR
giš <sup>š</sup> KUR.RA	giš <sup>š</sup> EŠ <sub>22</sub>
u <sub>2</sub> si-ḫa	u <sub>2</sub> A.ZAL.LA <sub>2</sub>
na <sub>4</sub> BAL	na <sub>4</sub> AN.ZAḪ

Most of the ingredients found within these rows are well-known from the existing medical tradition in Mesopotamia.<sup>17</sup> The woods, plants, and stones were used to create amulets and phylacteries employed in the treatment of illnesses. It is most likely that the first ingredient in each column, listing religiously significant places, refers to the dust from that location used in treatment.<sup>18</sup> Prior to the Late Babylonian period, these ingredients could be found in the therapeutic corpus used in the construction of various remedies for ailments. In fact, the categorical scheme is well known from medical texts contemporary

<sup>16</sup> VAT 7847 was first edited by Ernst F. Weidner in his 1967 edition of the first identified Micro-zodiac texts.

<sup>17</sup> The *bit urinnu* (E<sub>2</sub> U<sub>4</sub>.RI<sub>2</sub>.IN) was a generic term for a temple treasury (*CAD*, vol. U, 227, *urinnu* B). giš<sup>š</sup>KUR.RA is found in the lexical lists and must refer to a “mountain-tree,” as resin of the *siḫu*-tree was used in recipes (*CAD*, vol. S, 242, *siḫu* d). The E<sub>2</sub>.KUR in Nippur is the chief temple of the god Enlil, the *šiqdu*-tree (EŠ<sub>22</sub>) is probably an almond (*CAD*, vol. Š/III, 94, *šiqdu*), the *azallû*-plant is well known from the medical corpus (*CAD*, vol. A/II, 524, *azallû*), and finally *anzahḫu*-stone is a glass-like material also used in medical contexts (*CAD*, vol. A/II, 151, *anzahḫu*). Very few of the ingredients are translated in this chapter, partly because their identification is not certain, but also because their actual correspondence with any botanical name is not important for the method employed here.

<sup>18</sup> Note the parallel with the text edited by Nils Heeßel (2005), in which a similar group of four medical ingredients is included, with the last one in each case being the “dust” (SAḪAR) of a location.

with the Micro-zodiac.<sup>19</sup> Interestingly other texts using this scheme order the categories in slightly different ways: the texts that Heeßel has published use the order: stone, wood, plant, place. This scheme seems to be well known; in fact, one text in particular directly references these groupings of ingredients as a medical device, referring to the generic terms for the ingredients “stone, plant, tree” in connection with the zodiacal schemes.<sup>20</sup> It should be noted that the inclusion of multiple related zodiacal schemes in this tablet suggests that the associations present in this text between medical ingredients and signs of the zodiac constituted a paradigmatic structure that existed outside of the Micro-zodiac series.

The guiding question behind this study was whether there was an underlying system of organization centered around these medical ingredients and their association with certain zodiacal signs. Perhaps certain signs were linked in ways that were not immediately obvious; for instance, a medical ingredient might be linked to a certain minor sign and only appear when that sign was present within the column's major-minor sign pair. If the other rows displayed some form of organizational structure, albeit quite simple, perhaps the medical-ingredients row could also contain an underlying logic—one not immediately visible. A good example of this practice is an astrological text written by the scribe Iqīša, which links signs of the zodiac with medical ingredients made from the animals associated with the zodiacal signs.<sup>21</sup> In particular, the initial assumption of this method is that the organizational structure will be based on one of the two zodiacal signs that identify the location of each column, similar to two of the other rows. Astrology of this period is full of patterns and repetitive sequences, many of which found their way into Hellenistic geometric descriptions of astrological significance.<sup>22</sup> Some of these patterns date to earlier periods; for example, the linking of diseases and signs of the zodiac was shown by Mark Geller to be a development based on calendrical medicine from the earlier Neo-Assyrian period.<sup>23</sup>

The theory of borrowed associations and existing paradigms can easily be tested with other shorter contemporary texts. When the ingredients and

<sup>19</sup> Heeßel 2008, 9.

<sup>20</sup> This text, LBAT 1593, refers to both the Micro-zodiac and Calendar text schemes as “the animals of 13 and 4,37” (Reiner 2000, 424).

<sup>21</sup> Steele 2011, 338–339.

<sup>22</sup> Rochberg-Halton 1988.

<sup>23</sup> “The profound change taking place is that a traditional hemerology-based system of favourable and unfavourable days of the month for various rituals (STT 300) has been replaced by zodiac-based system which assumes astral influences over the same spells and rituals (BRM 4 20 and 19)” (Geller 2014, 57).

schemes within the Micro-zodiac are compared to other known texts such as the aforementioned material edited by Heeßel, there is a lack of parallel content; this suggests a familiarity with material but not the adoption of an organizational scheme. For instance, many of the medical ingredients are similar, and the methods of association, i.e., the idea of linking ingredients to zodiacal signs, are clearly related, but the connections themselves are not shared between corpora. Another astrological text links signs of the zodiac with cities, many of which are found within the Micro-zodiac, but under different signs.<sup>24</sup> Making this type of comparison is relatively easy to do, as the external material to the Micro-zodiac is more concise and simpler in format. The question remains whether or not there are other forms of organization underlying the material in the Micro-zodiac that are not detectable by comparison with other existing textual exemplars. If there were, we would assume these connections would not be bound to the Micro-zodiac scheme, but rather would present themselves through repetitive usage of ingredients in different sign pairings.

The methodology outlined below allows the text to be analyzed in a way that should bring to light any schemes of organization that are not immediately obvious. By analyzing the network of ingredients as a whole, we can highlight areas of commonality between cells. Crucially this would suggest the existence of a scheme without needing to have a prior example of the text returning to the difference between supervised and unsupervised methods, this method of generating groupings and structures *ex novo* without needing *comparanda* is one of the benefits of the unsupervised technique. It is important to note, however, that it is not without its pitfalls. The methods employed below are entirely dependent on the quality and form of the data entered. The unsupervised methods in particular may produce results that do not reflect a true understanding of the material, without the types of regulatory metrics that derive from a more thorough understanding of research questions, as is the case with supervised methods.

## Methodology

Since one of the aims of this volume is to explain some of the methods of the digital humanities in understandable terms, it is worth running through the process of going from text to analysis in detail.<sup>25</sup> The following section will run

---

<sup>24</sup> Steele 2015.

<sup>25</sup> The following methods were all completed with a digital “notebook” using the IPython platform. This allows the researcher to work iteratively through processing their data,



through the steps in turn, explaining the methods and terminology behind each step. Parallel to the explanation, I will include a small subset of the data to serve as an illustration of the steps. We will start with the text objects themselves and then progress through philological work, digital coding, and finally analysis and visualization. This process spans both work in museum collections as well as digital processing with the Python programming language.

It is important to state at the outset that none of this work would be possible without very traditional philological study.<sup>26</sup> In order to begin the work, the data had to be collected. The texts were located in museum collections, visited, photographed, studied, translated, and finally edited. The output of this work, of course, has many uses, and it forms the bedrock of any future study of the material, this investigation being only one part.

After the texts have been edited, the relevant sections have to be selected and converted into a format that allows for easy manipulation by data analysis tools. In this case I transcribed the data into a spreadsheet. It is important to choose a format and method of coding that is both flexible and standardized from the outset. This allows for any type of textual data to be represented in a form that is translatable and readable by later stages of your analysis.<sup>27</sup> Each row represented one ingredient, and the columns were laid out as follows (broken into two lines here):

There are more columns than necessary here, but it is best to strive for completeness rather than to have to go back through the dataset to re-enter a crucial bit of information later on. The redundancy, in particular between the signs and numbers, also makes things like labels for visualization easier later. A few notes on the columns: “Index” serves as an overall index of every row; “text” represents an internal count of the texts in the corpus; “text\_type” records the type of text, allowing for future expansion; “face” refers to the physical side of the tablet; “major\_sign,” “major\_number,” “minor\_sign,” “minor\_number,” and “zodiacal\_location” locate the ingredient on the text; “raw” represents the ingredient as it appears in the transliteration (the Latin-character equivalent of the cuneiform script); “name\_only” removes the determinative; “type” is

---

leaving detailed explanations, visualizations, and comments (Pérez and Granger 2007, <<https://www.computer.org/csdl/mags/cs/2007/03/index.html>> [accessed July 2, 2017]). For the notebook, see Monroe (2017, <<https://zenodo.org/record/827359>> [accessed July 1, 2017]).

26 One important facet to the development of techniques in digital humanities is their ability not to replace but to “enhance” the traditional work of text scholars (Jänicke et al. 2015).

27 A good description of digital encoding of textual data can be found in Sinclair and Rockwell’s chapter, “Text Analysis and Visualization” (Sinclair and Rockwell 2015, 279).



TABLE 8.3 *Example row from the input data*

Index	Text	text_type	face	major_sign	major_ number	minor_ sign	minor_ number
1	1	Micro-zodiac	Obv	Aries	1	Cancer	4
zodiacal_ location	Raw	name_only	type	translation	damage	unclear	Notes
1-4	bar <sub>2</sub> - sip <sub>2</sub> {ki}	bar <sub>2</sub> -sip <sub>2</sub>	place	Borsippa			

essentially a translation of the determinative; “translation” is an English translation when relevant; “damage” and “unclear” are just flags for when the raw transliterated string is damaged or an unclear reading—these can be used later on to restrict the dataset to certain confidence levels—and notes is just for internal comments about the data.

Once the data is in a digital form, it can then be imported into any number of tools; in this case the programming language Python was used, as well as the Pandas and NumPy libraries for data processing.<sup>28</sup> The Pandas library<sup>29</sup> functions around DataFrames, in essence a form of digital table. They come with their own built-in methods and tools for managing and viewing the data. Pandas allows for logical operations to be applied quickly and efficiently to the entire dataset. First, any cells with no relevant data in the “name\_only” field were removed; these included cells where there was no preserved information beyond a determinative, or where there were unclear readings of ingredients. Next a few custom columns were constructed out of previous columns: a unique location identifier that combined the text number and the two zodiacal locations, and a composite column including the name of the ingredient and its type. At this point, because I included the “text\_type” field I was able to easily combine or separate out various groups of texts into new DataFrames for later analysis. For instance, I could analyze the Micro-zodiac material by itself or include the Calendar Texts as well.

28 McKinney 2010, <<http://conference.scipy.org/proceedings/scipy2010/>> (accessed July 2, 2017); Walt, Colbert, and Varoquaux 2011.

29 A “library” is a collection of commonly used programming functions that are distributed as a package for use by a wide range of users.

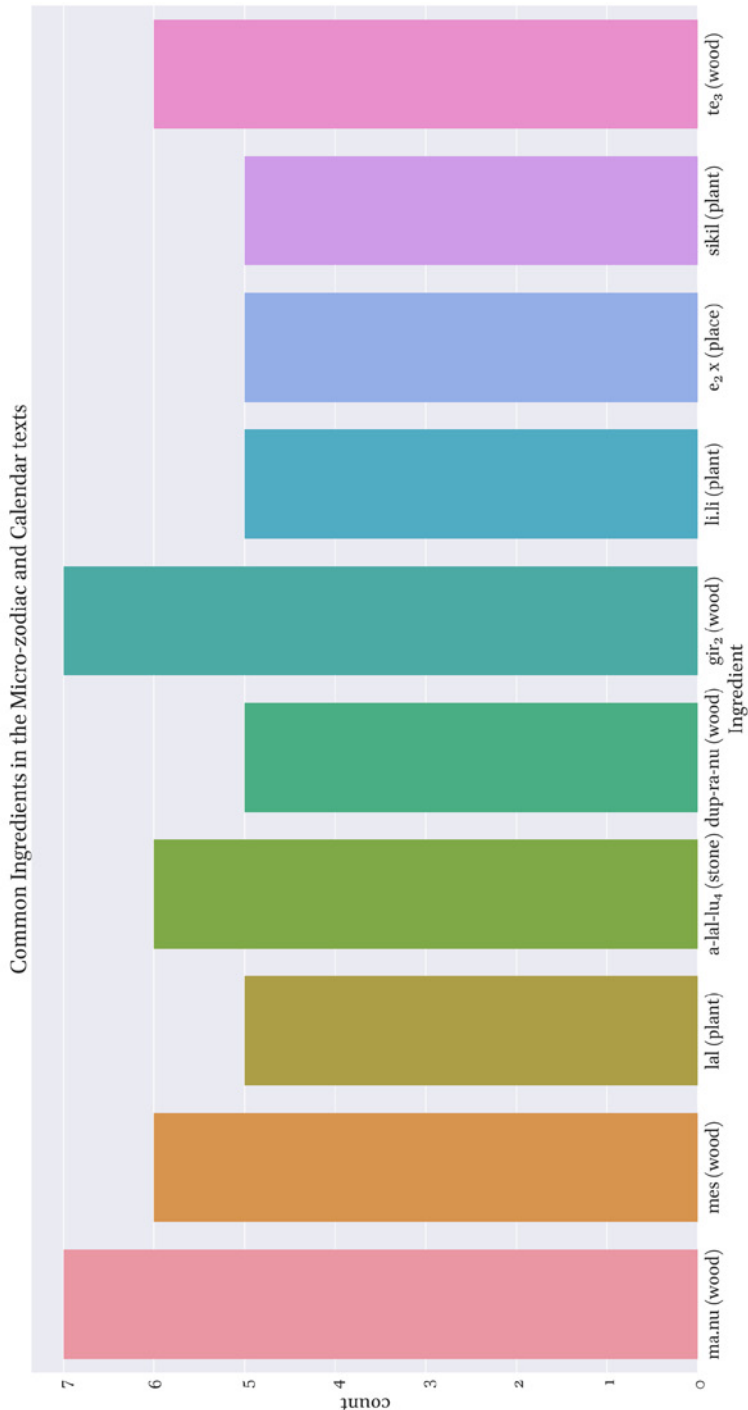


FIGURE 8.1 Counts of ingredients in the Micro-zodiac

As mentioned above, the process of producing and analyzing data is an iterative process that generates results while moving toward an end research goal. At this point the coded and imported data can already be processed and visualized in interesting ways. The power of using scripting languages and their associated libraries is that it is easy to explore the data in order to produce a general sense of its shape and make-up. Using some simple logic, I was able to produce a graph of the counts of various ingredients within the entire corpus using the Matplotlib and Seaborn libraries.<sup>30</sup> In this case I selected ingredients that appeared more than four times in the entire corpus and graphed them against each other.

This type of counting and visualization is feasible by hand, but the power inherent in this technique is that while one prepares the data for further, more complicated analysis, these intermediate steps can help one explore the data and refine the research question while developing further methods. This initial count gave me a sense of the most common ingredients in the corpus. It is interesting to note the low total numbers for all of these ingredients, yet the numbers are high enough to suggest that there could be shared content between different cells. If there were no ingredients with higher counts—if all ingredients had a count of one, for example—then it would be clear that there would be no value in graphing the connections between cells. The inverse is also true: if the counts were too high, it would suggest that the total pool of ingredients was too small and too interconnected to make a meaningful differentiation between groups of interconnected cells. Also notice that the third ingredient from the right is  $e_2 \times$ , meaning an unknown temple or cultic building. This is an obvious example of where damage or unclear readings can introduce errors into the processing. All of the instances of this ingredient are considered the same and if not accounted for may lead to false similarities between cells later on in the analysis.

After the data had been cleaned and new columns added, it was run through a number of functions to perform the analysis. The first step was to form bags of words for each cell in each text. The bag-of-words model is a well-known method in textual analysis in which all the individual lexemes from a text are flattened and placed in a large group.<sup>31</sup> Syntax and order are lost and do not figure into the final analysis, but this model offers a level of simplicity that functions very well for comparing the similarity of texts within a corpus. The

30 Hunter 2007; Waskom et al. 2016, <<https://zenodo.org/record/54844>> (accessed July 14, 2017).

31 See also in this volume, Svärd, Jauhiainen, Sahala, and Lindén (247–248), who applied the Continuous Bag-of-Words model to semantic analyses.

table below is an example of the bag-of-words model for three cells from the dataset. Each cell is given its own row, and the ingredients are listed in one line. While this model looks relatively simple for this dataset, most applications of the bag-of-words model involve listing the entire contents of, for instance, a novel, in one line.

TABLE 8.4 *Bag-of-words model for three cells*

Cell	Items
#3/4-11 <sup>a</sup>	e <sub>2</sub> nin.gir <sub>2</sub> .su, dup-ra-nu, ak-tam, mar-ḫal-lu <sub>4</sub>
#6/4-11	dup-ra-nu, ak-tam, mar-ḫal-lu <sub>4</sub>
#7/5-8	sig <sub>4</sub> unug, dup-ra-nu, ši-im-ra-nu, nir <sub>2</sub>

a The convention used throughout this chapter for naming cells is as follows: the character “#” followed by the number of the text within the corpus, then a forward slash, and finally a pair of numbers representing the Major and minor signs of the zodiac, respectively. “#3/4-11” stands for the third text in the corpus and the cell assigned to Cancer-Aquarius.

With a bag of words (in this case ingredients) for each cell, the entire group could then be run through a function to produce a Document Term Matrix (DTM). The matrix is a large “vector space” in which the cells reside. Each unique term within the corpus becomes a dimension on which the documents are plotted; for instance, if your corpus consists of documents that only use three words, the resulting vector space would be recognizable as a three-dimensional plot.<sup>32</sup> In practical terms the DTM is represented by a large table in which each row is one of the input documents, in this case a cell from one of our texts, and each column represents a unique term that shows up somewhere in the entire corpus.<sup>33</sup> The function, CountVectorizer from the Scikit-learn module,<sup>34</sup> runs through each bag and tallies a total list of all terms, then

32

See also in this volume, Eraslan (284–285, 305), on vector space and interoperability.

33

See also in this volume, two other examples of structured matrices used for identifying relationships: Auto-Contractive Map (Auto-CM) applied to landscape archaeology by Buscema in Ramazzoti (68 fig. 2.1a, 73n36, 74), and textual correspondence analysis for the investigation of cylinder seals by Ludovico (94–97).

34

CountVectorizer was chosen over another popular method in textual analysis called Term Frequency–Inverse Document Frequency (TF-IDF). The latter is quite popular because it evaluates the importance of terms within one text in inverse relationship to their frequency in the entire corpus, essentially identifying the terms that are most important for identifying a document. This method is based on frequency counts for terms within

marks, for each bag, which terms appear in that bag.<sup>35</sup> The end result is a matrix of all cells and all terms, with the count for each term in each cell tallied in the table. The table below is a DTM for the example given above. As you can see, the columns consist of a list of every term that exists in the example corpus. Each row is one of the cells in our example corpus. The values of the cells in this table are the number of occurrences of a term in a cell, in this case either zero or one.

TABLE 8.5 Document Term Matrix for example dataset

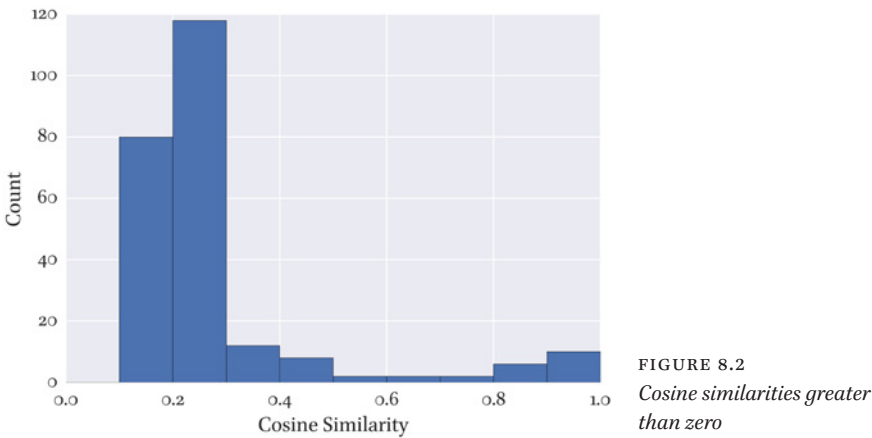
	ak-tam	dup-ra-nu	e <sub>2</sub> nin.gir <sub>2</sub> .su	mar-ḥal-lu <sub>4</sub>	nir <sub>2</sub>	sig <sub>4</sub> unug	ši-im-ra-nu
#3/4-11	1	1	1	1	0	0	0
#6/4-11	1	1	0	1	0	0	0
#7/5-8	0	1	0	0	1	1	1

Immediately some patterns begin to emerge; it is clear that the term *duprānu* is present in all cells, and the last three terms are only present in the final cell. This example is relatively simple and compact, but on a real corpus this step produces a huge dataset; the resultant table must include a cell for every word in the corpus and every document. On a small, example dataset such as the one above, the result is readable and understandable, but a real DTM is of little use without further processing.

With the matrix of terms prepared for the entire corpus, a function can be applied to the data to find the similarity between cells based on their shared terms. In this case the metric used was one called “cosine similarity.” This is a geometric method for determining similarity between rows in the DTM by assigning each row to a vector in multi-dimensional space and calculating the cosine of the angle between each vector. This method is well suited to the sparse binary nature of this dataset. Because the values from the DTM are either one or zero, representing the presence or absence of a term, respectively, the cosine of the angle is best suited for an analysis of similarity. If, on the other hand, we were concerned with the number of terms within each cell and

documents and within the corpus as a whole. It works well when there are many terms shared among all documents, as it weeds out common terms and focuses on the important rare terms. Because all terms in the present corpus are important, and counts greater than one are rare, a simple count of terms was used.

35 Pedregosa et al. 2011, <<http://jmlr.org/papers/v12/>> (accessed July 5, 2017).



took the frequency of terms to be an integral part of the difference or similarity between cells, then another method (for instance, Euclidean distance) would be better suited. With the cosine similarity computed across the entire corpus, a matrix can be constructed that shows each cell and its similarity with every other cell within a range of zero to one (with one being identical). Like the graph above of counts of ingredients, this step can be graphed as well, giving us a quick window into the data.

From the graph above (Fig. 8.2), you can see that only a few of the cells, on the far right, are exactly identical to each other. The majority of cells that have any similarity at all (this graph does not include cells with zero similarity) are relatively dissimilar. Applied across the entire corpus, this method produces an adjacency matrix, essentially a table in which the rows and columns are cells from the corpus, and at the intersection of each row and cell, a number is given for how similar the two cells are. As expected there is a diagonal line of 1's running down through the middle of the table where each cell is exactly similar with itself. When the similarity is computed across the entire example dataset, the following table is the result:

TABLE 8.6    *Adjacency matrix for example dataset*

	#3/4-11	#6/4-11	#7/5-8
#3/4-11	1	.866025	.25
#6/4-11	.866025	1	.288675
#7/5-8	.25	.288675	1

This table shows us that each cell is similar to itself, which is expected. The more interesting results are at the intersections of different cells, where we find differing counts. From this example dataset, it is clear that the cells located under Cancer-Pisces in texts 3 and 6 are very closely related, sharing three terms between them, resulting in a cosine of .866025. Meanwhile, the cell under Leo-Scorpio from text 7 shares only one term with the previous two cells, which, if you return to our DTM table above, was the term *duprānu*. This results in two different cosine values because texts 3 and 6 differ in length. Text 7 is more similar to text 6 than is text 3, because text 3 has one more term than does text 6, and text 7 shares only one term with both. As with the production of a DTM, this example dataset is easy to understand; on a real dataset the result would be too large to comprehend. This is where the visualization of data is necessary for understanding the results. The adjacency matrix produced above can then be exported to a Comma Separated Value (csv) file and imported into a graphing tool such as Gephi.<sup>36</sup> The procedure for creating data that is digestible by graphing programs is somewhat opaque and is left as an exercise to the reader. Gephi in particular will silently reject data in many cases, for instance if labels contain blank spaces.

Graphing of data can also be accomplished by utilizing many of the tools used above to process the data. In particular the SciPy library offers a number of metrics for performing cluster analysis of the adjacency matrices produced above.<sup>37</sup> Clustering methods are complicated and are an area of academic research in their own right, but various metrics can be suggested for best performance. In particular, “cophenetic correlation coefficient” is a metric that can evaluate the validity of a particular clustering technique together with a measure of distance or dissimilarity. Because we have already calculated the cosine similarity for each cell, we can simply subtract it from one to get the equivalent cosine distance, which can be used as the input for a clustering technique. When cosine distance is used to determine the closeness of data points, in this case cells from the Micro-zodiac, the best-performing clustering method was Unweighted Pair Group Method with Arithmetic Mean (UPGMA).<sup>38</sup> While

36 Bastian et al. 2009, <<https://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154>> (accessed July 2, 2017). “Graphing programs” are tools that allow one to visually interpret numerical data and export the results as images. On csv file format, see in this volume, Martino and Martino, 137. For further information about Gephi, see in this volume, Pagé-Perron, 209.

37 <<http://www.scipy.org/>> (accessed July 14, 2017). For an example of performing cluster analysis to reveal objects typologies, see in this volume, Martino and Martino, 118–124, 133–134.

38 Saraçlı, Doğan, and Doğan 2013, <<https://doi.org/10.1186/1029-242X-2013-203>> (accessed July 2, 2017), 203, table 2.

there are many methods of clustering, the one chosen here is relatively simple in its procedure. It starts by looking for the two closest cells within the adjacency matrix and merging those into a cluster to which it assigns the average of their distance. It then looks for the next two closest cells, or, if the next cell is closest to the newly created cluster, it merges those and assigns it a new average distance. This process is completed for the entire adjacency matrix, merging close items whether they are cells or newly created clusters. When UPGMA clustering is applied to the adjacency matrix of cells from the Micro-zodiac, a dendrogram can be created which sorts the cells into hierarchical bunches based on their similarity.

This illustration shows the bottom part of the dendrogram, the cells which were easily clustered by the algorithm. The location of the cells is listed along the left side of the graph, and vertical connecting lines link cells or groups of cells into clusters. The vertical lines' location along the bottom axis represents how similar the contained elements are. The left edge of the graph shows elements that are identical, and the right edge represents highly different elements. The graph clearly shows that cells with the same major and minor sign pairs are highly similar. The dotted line at .75 represents a subjective marker for where the clustering no longer seemed to produce relevant results. Each of these dendrograms encompasses the entire dataset regardless of the particular metrics of interest for the research. The .75 cutoff seemed to mark a point at which the clustering of cells or existing clusters was no longer based on shared ingredients but rather on average distances of existing clusters. The results point to a very strong association of cells with other cells sharing the same major and minor sign pairs regardless of the particular text. The bottom group of Gemini-Capricorn (3-10) come from three different tablets but all are identical in their medical ingredients. What is perhaps more telling is what clusters are not present. There are three clusters that have the major sign Cancer (4), each with a different minor sign, and they are all highly distinct. The inverse pattern is also true; there are clusters assigned to the minor signs Capricorn (10) and Aquarius (11), each with different major signs, and they are all highly distinct. If the association of ingredients was dependent on only the major or minor sign, we would expect the cells to cluster in a pattern in which either the major or minor sign were shared.

The same process can be applied to the inverse of the entire dataset. Instead of basing our analysis on the cells and the ingredients they contain, we can flip the paradigm around and look for ingredients that share similar cells. Here a similar picture plays out; the above graph is a detail from a dendrogram created using the exact same procedure outlined above (one of the great advantages is the rapid re-usability of methods on similar data). Naturally these clusters are slightly less organized because the dataset is considerably larger, and



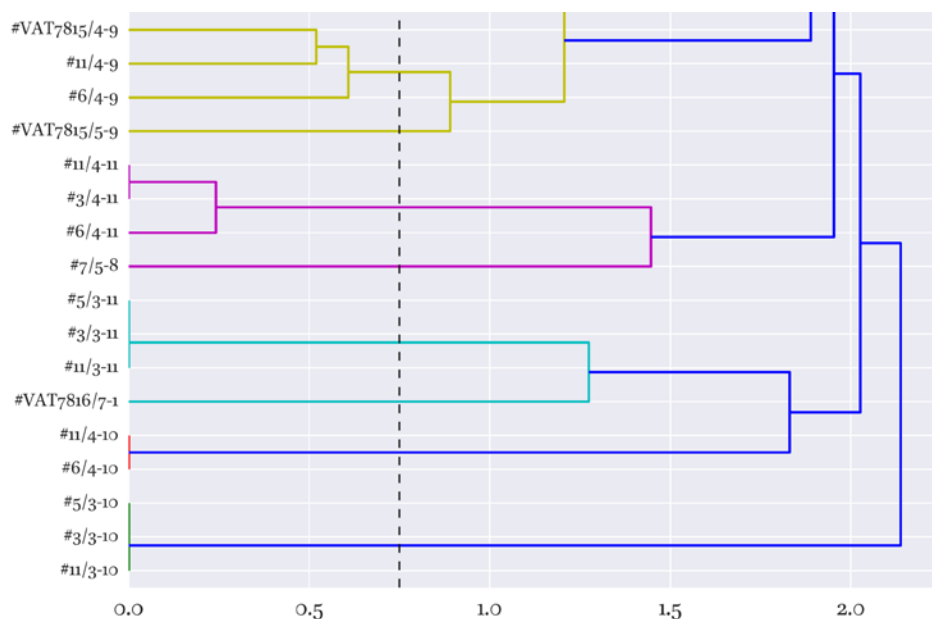
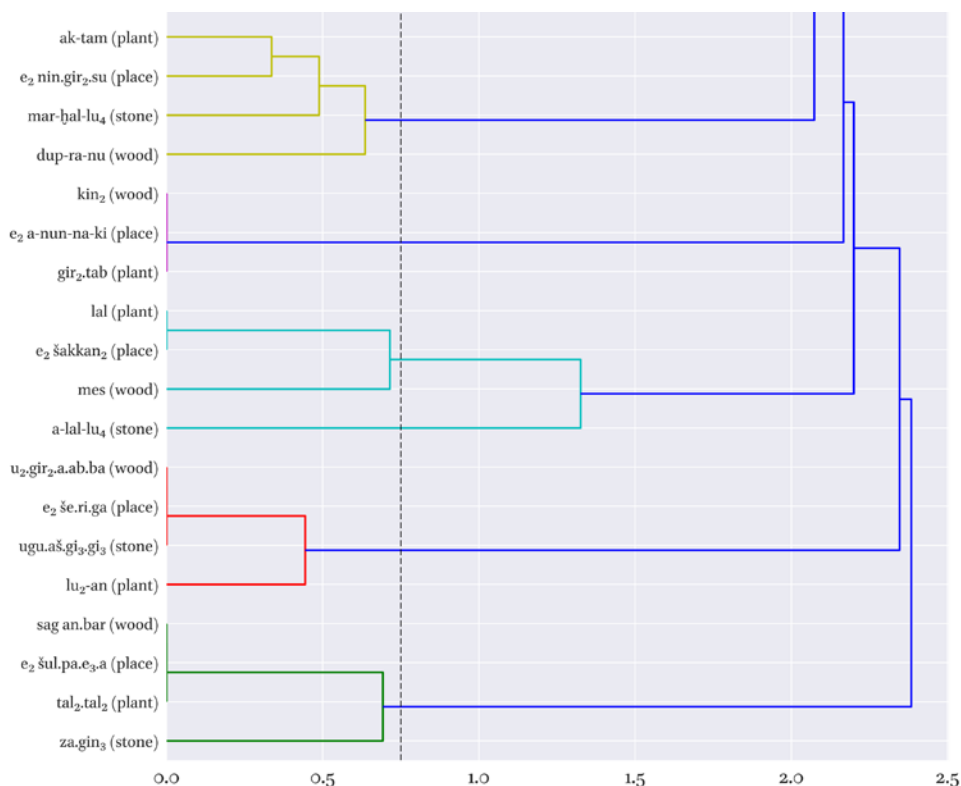


FIGURE 8.3 *Detail from dendrogram of Micro-zodiac cells*

connections are less common. Notice, however, that the ingredients cluster roughly into patterns that mirror the inclusion of ingredients in the individual cells: place, wood, plant, and stone. Interestingly stone is often slightly removed from the cluster of the first three ingredients. The likely cause of this pattern is the aforementioned common practice of including two stone ingredients in the cells, meaning that stone ingredients are more common throughout the entire dataset and thus slightly more distant from the other tightly connected ingredients.

## Conclusions

The methods employed above took what was originally a damaged and fragmentary corpus of Late Babylonian astrological data and, through various quantitative processes, extracted meaningful information about its organization. While generally these methods are used in the analysis of much larger and more coherent datasets, the results here show that even text that is difficult to read in conventional ways still offers an opportunity for analysis with digital tools. The tools used here integrate well into the traditional philological

FIGURE 8.4 *Detail from dendrogram of Micro-zodiac ingredients*

work of editing a text and offer another vantage point from which to gain insight about the patterns or structures perhaps not immediately obvious within a given corpus. The disjointed nature of the evidence is particularly well suited to quantitative methods that treat the fragmentary text as a series of data points and allow for the application of powerful algorithms. The patterns discoverable here are widely applicable to other corpora. For instance, one could use this same method to analyze the adjacency of the corpus of Neo-Assyrian letters to the king; texts clustered around each other might be a function of a single author or a particular topic. Similarly, a dendrogram could be constructed for the phrasing of royal inscriptions, illustrating the terms and forms shared among inscriptions by a single king over the course of his reign, or between or among kings

As for the data derived from the Micro-zodiac texts, after the data for both the cells and the ingredients themselves were processed, the resulting graphs suggest that the medical ingredient cells of the Micro-zodiac are related to

each other only by the internal organization scheme inherent in the Micro-zodiac itself. If the cells are clustered according to shared ingredients, the most common clusters are oriented around cells sharing the exact same major-minor sign pairings. No other system of linkage or connection between ingredients appears in the clustered cells, and, even when the data is inverted to look for patterns between ingredients, no other organizational structure emerges. The results of the analysis show that the connections between medical ingredients in the Micro-zodiac text do not preserve an alternate system of association as might have been borrowed from another textual tradition. As mentioned in the introduction, there were a number of texts associating medical ingredients and zodiacal signs written during this period, and it is interesting that the Micro-zodiac does not seem to borrow from these.

The lack of a connection with contemporary texts or earlier forms of astrological knowledge suggest that the associations found within the ingredient row of the Micro-zodiac are a novel invention unique to this series. It is important to note that there are systems of association and connections that existed outside of the ingredient row, attesting to the Micro-zodiac's borrowing of previous text and content. However, the lack of any explicit pattern in the ingredient row is important because it further illustrates the complexity of astrology during the late period in Babylonia. As more of the late astrological texts are studied, it is clear that the science of astrology was developing new genres of knowledge both by borrowing from previous traditions and inventing new forms of meaning.<sup>39</sup> The Micro-zodiac texts were trying something distinctly new, but they were written in a context in which many other texts were also attempting new systems of organization or association.

## References

- Bastian, Mathieu, Sebastien Heymann, and Mathieu Jacomy. 2009. "Gephi: An Open Source Software for Exploring and Manipulating Networks." *Proceedings of the Third International Conference on Weblogs and Social Media*, edited by Eytan Adar, Matthew Hurst, Tim Finin, Natalie Glance, Nicolas Nicolov, and Belle Tseng, 361–362. Menlo Park, CA: AAAI. <<https://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154>>.
- Geller, Mark. 2010. *Ancient Babylonian Medicine*. Malden, MA: Wiley-Blackwell.

---

39 Koch 2015, 197–199.

- Geller, Mark. 2014. *Melothesia in Babylonia: Medicine, Magic, and Astrology in the Ancient Near East*. Science, Technology, and Medicine in Ancient Cultures 2. Boston: De Gruyter.
- Heeßel, Nils. 2005. "Stein, Pflanze und Holz. Ein neuer Text zur 'medizinischen Astrologie.'" *Orientalia* 74: 1–22.
- Heeßel, Nils. 2008. "Astrological Medicine in Babylonia." In *Astro-Medicine: Astrology and Medicine, East and West*, edited by Anna Akasoy, Charles Burnett, and Ronit Yoeli-Tlalim, 1–16. Micrologus' Library 25. Florence: SISMEL - Edizioni del Galluzzo.
- Hunter, John D. 2007. "Matplotlib: A 2D Graphics Environment." *CISE* 9: 90–95.
- Jänicke, Stefan, Greta Franzini, Muhammad F. Cheema, and Gerik Scheuermann. 2015. "On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges." In *Proceedings of the Eurographics Conference on Visualization (EuroVis) (2015) STAR – State of The Art Report*, edited by Rita Borgo, Fabio Ganovelli, and Ivan Viola, 83–103. Aire-la-Ville, Switzerland: Eurographics Association.
- Jänicke, Stefan, Greta Franzini, Muhammad F. Cheema, and Gerik Scheuermann. 2016. "Visual Text Analysis in Digital Humanities: Visual Text Analysis in Digital Humanities." *Computer Graphics Forum* 36 (6): 1–25.
- Jockers, Matthew L., and Ted Underwood. 2015. "Text-Mining the Humanities." In *A New Companion to Digital Humanities*, edited by Susan Schreibman, Raymond G. Siemens, and John Unsworth, 291–306. Chichester: John Wiley & Sons.
- Koch, Ulla. 2015. *Mesopotamian Divination Texts: Conversing with the Gods Sources from the First Millennium BCE*. Münster: Ugarit-Verlag.
- McKinney, Wes. 2010. "Data Structures for Statistical Computing in Python." In *Proceedings of the 9th Python in Science Conference, June 28–July 3 2010, Austin, TX (SciPy 2010)*, edited by Stéfan van der Walt, and Jarrod Millman, 51–56. <<http://conference.scipy.org/proceedings/scipy2010/>>.
- Monroe, M. Willis. 2016. "Advice from the Stars: The Micro-Zodiac in Seleucid Babylonia." PhD diss. Brown University.
- Monroe, M. Willis. 2017. "Willismonroe/Monroe2017DH: Initial Upload." Last modified July 14, 2017. *Zenodo*. <<https://zenodo.org/record/827359>>.
- Moretti, Franco. 2007. *Graphs, Maps, Trees. Abstract Models for Literary History*. London-New York: Verso.
- Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. 2011. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research* 12: 2825–2830. <<http://jmlr.org/papers/v12/>>.
- Pérez, Fernando, and Brian E. Granger. 2007. "IPython: A System for Interactive Scientific Computing." *CISE* 9 (3): 21–29. <<https://www.computer.org/csdl/mags/cs/2007/03/index.html>>.

- Reiner, Erica. 2000. "Early Zodiologia and Related Matters." In *Wisdom, Gods and Literature: Studies in Assyriology in Honour of W.G. Lambert*, edited by Andrew R. George, and Irving L. Finkel, 421–427. Winona Lake, IN: Eisenbrauns.
- Rochberg-Halton, Francesca. 1988. "Elements of the Babylonian Contribution to Hellenistic Astrology." *JAOS* 108 (1): 51–62.
- Rochberg, Francesca. 2004. *The Heavenly Writing: Divination, Horoscopy, and Astronomy in Mesopotamian Culture*. Cambridge: Cambridge University Press.
- Saraçlı, Sinan, Nurhan Doğan, and İsmet Doğan. 2013. "Comparison of Hierarchical Cluster Analysis Methods by Cophenetic Correlation." *Journal of Inequalities and Applications* 2013: 203. <<https://doi.org/10.1186/1029-242X-2013-203>>.
- Sinclair, Stéfan, and Geoffrey Rockwell. 2015. "Text Analysis and Visualization: Making Meaning Count." In *A New Companion to Digital Humanities*, edited by Susan Schreibman, Raymond G. Siemens, and John Unsworth, 274–290. Chichester: John Wiley & Sons.
- Steele, John M. 2008. *A Brief Introduction to Astronomy in the Middle East*. London: SAQI.
- Steele, John M. "Astronomy and Culture in Late Babylonian Uruk." 2011. In "*Oxford IX*" *International Symposium on Archaeoastronomy Proceedings IAU Symposium* 278, edited by Clive L.N. Ruggles, 331–341. Cambridge: Cambridge University Press.
- Steele, John M. 2015. "A Late Babylonian Compendium of Calendrical and Stellar Astrology." *JCS* 67: 187–215.
- Walt, Stéfan van der, S. Chris Colbert, and Gaël Varoquaux. 2011. "The NumPy Array: A Structure for Efficient Numerical Computation." *CISE* 13 (2): 22–30.
- Waskom, Michael, Olga Botvinnik, Drew O'Kane, Paul Hobson, David C. Gemperline, Yaroslav Halchenko, Saulius Lukauskas, 2016. "Seaborn: v0.7.1 (June 2016)." *Zenodo*. Last modified June 6, 2016. <<https://zenodo.org/record/54844>>.
- Weidner, Ernst F. 1967. *Gestirn-Darstellungen auf babylonischen Tontafeln*. Vienna: Böhlau in Kommission.
- Zundert, Joris J. van. 2015 "Screwmenetics and Hermenumericals." In *A New Companion to Digital Humanities*, edited by Susan Schreibman, Raymond G. Siemens, and John Unsworth, 331–347. Chichester: John Wiley & Sons.

**PART 4**

*Online Publishing, Digital Archiving, and  
Preservation*





# On the Problem of the Epigraphic Interoperability of Digitized Texts of the Mediterranean and Near Eastern Regions from the First Millennium BCE

*Doğu Kaan Eraslan*

## Introduction<sup>1</sup>

Recently there have been impressive efforts to digitalize ancient texts of the first millennium BCE. Their digitalization has not only facilitated access to the information contained in these documents, but it has also given us a chance to use computational methods to analyze them.<sup>2</sup> The question of how projects that digitize these texts interact with each other, however, has not been a domain of interest for most researchers.<sup>3</sup> Yet for those who are trying to understand the history of international relations in the ancient Mediterranean

- 
- 1 The present communication has benefited tremendously from the suggestions and discussions that took place at the Caf'E.PHE (École Pratique des Hautes Études, Paris) during January and February of 2017 (Eraslan 2017a). I would like to express my gratitude to all of the Caf'E.PHE attendees, noting especially the very insightful questions of Daniel Stockholm and also Émilie Pagé-Perron's suggestion to add the section on practical concerns. My remarks on the use of OCR were also incorporated following discussion with Émilie Pagé-Perron. Furthermore, I was fortunate to hear Terhi Nurmikko-Fuller's 2017 paper at Caf'E.PHE on Linked Data technologies. This motivated my section on CIDOC CRM. Marine Béranger, one of the few people who dared to encode hundreds of Akkadian letters in TEI, laid the foundations for the observations that led to this paper. In addition, I would like to express my gratitude to Vanessa Bigot Juloux for her unwavering encouragement and faith in me as a colleague. I am also very grateful to Amy Rebecca Gansell for her English-editing suggestions. Her particular attention to technical details with regard to terminology and general phrasing increased considerably the overall comprehensibility of the paper. Of course, the content is my own responsibility. Lastly, I would like to thank my father, Eyüp Eraslan, for his constant support.
  - 2 For example, a library such as Classical Language Toolkit (CLTK) (<<http://cltk.org/>> [accessed May 6, 2017]) would not have been possible without the digitalization of these texts. See Johnson (2014, <<https://doi.org/10.5281/zenodo.60021>> [accessed May 6, 2017]).
  - 3 This was probably due to the isolated nature of the disciplines, but it has become increasingly clear that this practice of isolation cannot endure. For a notable exception to this phenomenon, see Gippert (1999).



region during this period, it is necessary to work with a multilingual corpus of inscriptions. Multilingual researchers are impeded by the incompatibility of the various digitalizations.

This paper demonstrates one of the incompatibilities that arises from the current state of the encoding schemes.<sup>4</sup> After a brief survey of the encoding schemes used by the major corpora online, I review the primary trends and the differences among them and highlight the problem of how to facilitate their interoperability.<sup>5</sup> In the interest of moving toward a solution, I propose a shift in the way we look at the encoding schemes. Until the present, we have *read* the ancient texts and deemed what we *read* as worthy of conservation; however, current technology makes it possible to analyze what we *see*, rather than what we have read. Although linguistic differences may make the encoding of what we have read incompatible, the unit we employ to encode what we see promises a solution of compatibility, since all 2D shapes used by the scripts of the first millennium BCE can be expressed in the vector spaces of linear algebra.<sup>6</sup>

---

4 Throughout this paper, I use the terms “encodings,” “schemes,” and “encoding schemes” interchangeably. In the most general sense, an encoding scheme is a set of rules that standardizes the representation of an object in a digital environment.

5 I use the terms “epigraphic interoperability” and “interoperability” to refer to the same phenomenon. Epigraphic interoperability refers to the ability to transform the encoding scheme of an encoded epigraphic phenomenon into another encoding scheme without losing data.

6 Although the scripts under analysis are all attested in 3D environments, we can at least work with their 2D models in a vector space without data loss. This is necessary because not all texts have been digitized in 3D. With the continual advances in 3D scanning technologies, however, an upcoming task for digital epigraphists will be the development of standard encoding schemes for 3D shapes. Vector space: in linear algebra, an additive group that is associated with a field of real numbers and permits vector addition and scalar multiplication. For more detailed information, including formulas, see Weisstein (n.d., <<http://mathworld.wolfram.com/VectorSpace.html>> [accessed February 26, 2018]). For additional information, see in this volume, Monroe (271), who used vector space for processing a Document Term Matrix (DTM) for a bag-of-words method. Vector: Mathematically speaking, a vector is anything that represents something that has a size and a direction. Size is the distance between the origin and end points, and direction is represented by the angle that a line makes with an axis. For a more technical definition, see Lang (1986, 9–12), who uses the term “located vector” for the concept we have defined above, although this appellation does not represent a general consensus. Vector addition simply entails adding two or more vectors together. Scalar multiplication is a multiplication operation between a vector and a number that belongs to the same set of numbers contained in the vector.

Toward this end I present some tools to enable digital epigraphers to surmount the problem of the current lack of multilingual interoperability.<sup>7</sup> The solution offered here consists of the projection of signs from a physical medium to a vector space where they are mapped to their semantic units,<sup>8</sup> facilitating the querying of the shape in the vector space.<sup>9</sup> For encoding technology, my choice of EpiDoc, SVGs, and Unicode (which are discussed later in this paper) is motivated by purely practical concerns, but I discuss alternatives as well. Finally, this paper provides some practical insights on how different components of the model can interact and describes what would be necessary to implement the suggested solution for an ongoing project.

### Encodings of the First Millennium BCE: A Brief Survey

#### *c(anonical)-ATF*

C(anonical)-ATF is the backbone of the texts displayed by the Cuneiform Digital Library Initiative (CDLI), which is currently the major online database for cuneiform texts.<sup>10</sup> As the name implies, CDLI conserves epigraphical and semantic data of cuneiform texts, rather than for a specific language.<sup>11</sup> It uses American Standard Code for Information Interchange (ASCII) characters.<sup>12</sup> Special characters such as “\_” or “\$” are reserved for indicating sign functions and information about the state of conservation of the script and its physi-

- 
- 7 These tools are part of a possible solution; certainly, there may be additional approaches with different tools.
  - 8 Sign corresponds to an elementary semantic unit comprising alphabetical and non-alphabetical languages. The concept of elementary semantic unit is defined more precisely later on. Physical medium refers to the object on which we observe the signs. It can be papyrus, a vase, a stone surface, etc.
  - 9 Querying: extracting information from a database based on the user's input.
  - 10 For details regarding the specific procedures used in the encoding process, see Tinney (2017, <<http://oracc.museum.upenn.edu/doc/help/editinginatf/cdliatf/>> [accessed January 26, 2017]).
  - 11 CDLI was initially only for Sumero-Akkadian cuneiform (Anderson et al. 2000, <<http://pages.jh.edu/~dighamm/ice/iceireport.html>> [accessed April 4, 2017]). Regarding the “c” in “C-ATF,” it can stand for either “canonical” or “CDLI,” since the latter uses ATF encoding. For further information, see in this volume, Pagé-Perron, 198–200, 200n28. For the representation of cuneiform characters in unicode, see Everson, Feuerheim, and Tinney (2004, <<http://std.dkuug.dk/jtc1/sc2/wg2/docs/n2786.pdf>> [accessed April 7, 2017]).
  - 12 ASCII characters are limited to those that are used in the writing of English words without accents. The characters include but are not limited to: “3, b, R, a, m, k, [, ,”.



FIGURE 9.1  
*C-ATF: Cropped detail of  
Tablet PF404 (CDLI)*



FIGURE 9.2 *Bavant-XML Standard: Cropped detail of DNa (Schmidt 1970, pl. 32)*

cal medium.<sup>13</sup> The minimal semantic unit of the encoding corresponds to the language’s minimal semantic unit.<sup>14</sup> In most cases, the minimal semantic unit is a cuneiform sign.

TABLE 9.1 *Elamite text in C-ATF*

<b>CDLI no: P383090<sup>a</sup> = PF 404<sup>b</sup></b>	<b>Hallock transliteration:</b> <b>20 ZÍD.DA.lg kur-min<sup>c</sup></b>
	<b>C-ATF:</b> 2 (u) _zi3-da-mesz_ kur-min2 <sup>d</sup>

a <[http://cdli.ucla.edu/search/archival\\_view.php?ObjectID=P383090](http://cdli.ucla.edu/search/archival_view.php?ObjectID=P383090)> (accessed January 25, 2017).

b Hallock 1969, 163.

c Author’s translation: “20 flour supplied by.”

d The differences, such as zi3=ZÍD, are related to different readings/interpretations of the signs and fall outside the scope of this paper. One should note, however, the typological differences between representations of the sign readings.

13 Special characters are exemplified by “\*, \_ , =” as opposed to alphanumeric characters such as “a, b, 1, 2,” etc.

14 I do not use the term “semantic unit” in its strictly linguistic sense. By minimal semantic unit, I mean the smallest distinctly perceived element. This corresponds to the smallest delimited element in the case of an encoding scheme, since the perception is done by the machine.

*Bavant XML-Elamite Standard*

XML-Elamite standard, developed by Marc Bavant,<sup>15</sup> is tailored for conserving Elamite texts along with their transliterations and translations.<sup>16</sup> It is based on Unicode characters, and it follows widely used conventions in Elamite scholarship, making it human readable.<sup>17</sup>

TABLE 9.2 *Elamite text in Bavant-XML***DNa 1§1<sup>a</sup>**


---

```
<inscription name="DNa"><metadata><king dates="522-486"
name="Darius I" />
<source></source>
<url>http://www.um.es/ipoa/cuneiforme/elamita/archi
vosreales/dario1/dario_i_dna.htm</url>
</metadata><notes/><section id='DNa:§1'>
<T1>(1) d.na-ap ir-šá-ir-ra</T1>
<T2>(1) Ahuramazda es el gran dios</T2>
<T3>d.na-ap ir-šá-ir-ra</T3>
<T4>nap iršara</T4></section>
</inscription>
```

---

a "Great god (Auramazda)" (Schmidt 1970, pl. 32).

*M(anuel) d(e) C(odage)*

Manuel de Codage (mdc) is the standard developed for processing ancient Egyptian hieroglyphs.<sup>18</sup> mdc, specifically its last version, mdc-88, has seen a lot of variants, and it uses ASCII characters. The most frequently used special characters are reserved for expressing hieroglyph positions, such as "\*" for

15 The 2014 version of Bavant's Elamite-XML corpus is accessible through the University of Amsterdam's Digital Academic Repository: <<http://hdl.handle.net/11245/1.348271>> (accessed February 13, 2017).

16 Elamite: an ancient language used in Elam, a civilization existing from the third into the mid-first millennium BCE based in the southwest region of the Iranian plateau, just north-east of the Persian Gulf (in the modern provinces of Ilam and Khuzestan).

17 Human readable: a text that is destined to be read by humans, as opposed to machine readable, which means it is intended to be processed by machines.

18 Buurman et al. 1988. Note that, in theory mdc can be extended to process Hieratic, an ancient Egyptian cursive script, as well. See also the critical overview by Nederhof (2013).



TABLE 9.4 *Imperial Aramaic text in CAL Code*


---

[כסף גבר] יא זי עביר על בית מלכא

Hebrew Transliteration: [כסף גבר] יא זי עביר על בית מלכא

CAL Code: [k]sP g? [br]y) zy? [ (byd (l byt mlk)

---

a “silver of men that is done towards the house of king” (inscription on TAD C3.7, *The Comprehensive Aramaic Lexicon*, <[http://cal.huc.edu/get\\_a\\_chapter.php?file=23350](http://cal.huc.edu/get_a_chapter.php?file=23350)> [accessed November 4, 2016]).

### *EpiDoc*

Strictly speaking, EpiDoc is an XML-scheme for adding markup to ancient texts.<sup>23</sup> Thus, it is applicable to all ancient languages, but it is predominantly used for Graeco-Roman corpora. It employs the Text Encoding Initiative (TEI)<sup>24</sup> scheme as its backbone, and it accepts anything encoded with UTF-8.<sup>25</sup> Since EpiDoc is fundamentally a markup language that tries to be applicable to a broad range of languages,<sup>26</sup> it is fairly liberal when it comes to using values for attributes. A significant issue with EpiDoc, however, is that any text encoded with it is dependent on the project’s financial capacity to maintain itself long-term. Increased costs are due to the high overhead costs for the server and the need to retain a project staff person who knows how different versions of the schema map to each other in a particular project.<sup>27</sup> One potential way to at

23 Elliot et al. 2017, <<http://www.stoa.org/epidoc/gl/latest/>> (accessed June 1, 2017).

24 Text Encoding Initiative (TEI): a consortium that maintains the scheme for evaluating XML-encoded documents. For the history of the consortium, see TEI Consortium (2018, <<http://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf>> [accessed May 15, 2017], xxv). For further explanation, see in this volume, Bigot Juloux, 164, 164n65.

25 “UTF-8” stands for Unicode Transformation Format, with “8” referring to the use of 8-bit (a numerical value that equals 0 or 1) sequences to represent a character. It is a method for encoding Unicode characters. There are also UTF-16 and UTF-32, which have, respectively, 16- and 32-bit character sequences.

26 Markup language: “a set of markup conventions used together for encoding texts. A markup language must specify how markup is to be distinguished from text, what markup is allowed, what markup is required, and what the markup means” (TEI Consortium 2018, <<http://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf>> [accessed May 15, 2017], xxv). For further explanation in this volume, see Bigot Juloux (163, 165–166, 186) regarding creating analytical taxonomies.

27 The cost is due to a specific situation with XML. During the encoding process, XML requires one to end a markup section with the tag that is used at the beginning of the section, making one repeat information in an inefficient way, since one would be using extra characters for information that does not necessarily need those characters. The gain is that this

least get around the server issue would be to make a detailed database of encoding projects, so that already encoded material can be scraped from the internet,<sup>28</sup> which would eliminate the need to redo the whole process. The database should also provide an application programming interface (or API, a set of rules and tools that defines how computers can interact)<sup>29</sup> for accessing it and scripts to add the markup to the source-text material based on the scraped material. The use of scripts in the automatic conversion of one database to another, however, could create what are known as cross-walk problems.<sup>30</sup>

TABLE 9.5    *Ancient Greek text in EpiDoc-XML*

Ἀπόλλω- νει θεῶ <sup>a</sup>	<pre>&lt;div type="edition"&gt;&lt;head&gt;edition&lt;/head&gt; &lt;ab&gt;&lt;lb xml:id="line-1" n="1"/&gt; &lt;persName type="divine"&gt; &lt;name nymRef="Ἀπόλλων"&gt;Ἀπόλλω- &lt;lb xml:id="line-2" n="2"/&gt; νει&lt;/name&gt; &lt;w lemma="θεός"&gt;θεῶ&lt;/w&gt; &lt;/ab&gt; &lt;/div&gt;</pre>
Normal	EpiDoc

a The inscription reads “To the god Apollo;” see Thonemann et al. (2012, <<http://mama.csad.ox.ac.uk/monuments/MAMA-XI-314.html>> [accessed January 26, 2017]).

Major Trends and Differences

There are two major trends in all of these encoding schemes:

---

makes XML very human readable; the cost is the extra transactions between the client and the server.

28    Scraping: a process of selectively retrieving online data using special software. For further discussion of “scraping,” see in this volume, Pagé-Perron 204, 204n41.

29    See in this volume, Prosser, 318, who describes OCHRE API.

30    Cross-walk problems occur when you want to move the content of a database to another database that is arranged in a different manner. One knows that the content is essentially the same, but different representations of it make it difficult to transfer content from one to another.

1. Minimal distinctly perceived units are mapped to intermediary elements.<sup>31</sup>
2. Glyphs are treated as characters/letters.<sup>32</sup>

The key difference among schemes is the nature of the intermediary elements. CAL Code, C-ATF, Bavant-XML, and EpiDoc use semantic elements in a manner that is similar to a transliteration, and Mdc uses arbitrary codes. Two problems result:

1. Apart from Mdc, the encoded texts are conceived so that the majority of the readers of the texts are humans.
2. Some schemes attribute glyphic properties to their elements, and some do not. Therefore, the resulting encoded document for multilingual corpora has the problem of a lack of interoperability.<sup>33</sup>

The first problem concerns the design of the encoding schemes, and it probably cannot be remedied quickly. In sum, a document that is digitally conserved as it is observed by humans suffers from the imposition of a hierarchy by the human perception of the document. The semantic hierarchy imposed by the observer creates substantial differences in the retention rate of the data held by the original document. This means that since we wanted to conserve what the text said, we conceived systems that conserve in detail how that which is said is expressed but give less attention to the actual state of the document in

---

31 Minimal distinctly perceived units: units that could correspond to letters for alphabetical languages or signs for logo-syllabic languages. Intermediary elements: the vehicle between what we see in the physical medium and what we understand. This corresponds, for example, to the alpha-numeric characters that are employed to represent the glyph in the physical medium.

32 Glyphs: the most basic parts of the visual representation of an elementary semantic unit. In the case of an Egyptian hieroglyph, for example, a glyph could be a sitting man or a bird. For cuneiform, a glyph would be a horizontal, vertical, or diagonal line comprising a cuneiform sign, and for alphabetic languages, such as Greek and Aramaic, a glyph corresponds to a visual representation of a letter, such as an alpha.

33 For more technically oriented reader, the problem can be summarized as a class inheritance problem. If we take the elementary semantic unit as the basic class of architecture, which would be the encoding scheme, the above encoding schemes instantiate this first basic class with different attributes in their constructor, making it difficult do transformations later on. The logical solution would be to add the lacking attributes to the constructor, but this still would not enable us to transform the already encoded material, since what we want to express with the attributes is defined with methods later than at the current state of these encoding schemes.



question.<sup>34</sup> Imposing the semantic hierarchy of the human observer therefore makes encoded documents less computable because it reduces our capacity to compute non-semantic data by presenting it, if at all, in a less precise and less organized manner. Thus, arranging encoding schemes on the premise that the majority of readers would be humans is a design flaw that cannot be addressed easily. The present paper deals with a subset of the problems that are generated by this flaw.

### The Problem of Achieving Interoperability

The problem with achieving interoperability comes from the fact that the elementary unit of the above encoding schemes, corresponding to a minimal semantic unit in the relative language, does not have the same visual decomposability throughout the above-mentioned languages, making the units unreliable in representing the state of conservation of a multilingual document. This is the result of the human-oriented approach, which relies on characters rather than glyphs as the basis of encoding.

Consider, for example, the quadrilingual vase of Darius I.<sup>35</sup> It bears four inscriptions, which are written with Elamite cuneiform, Akkadian cuneiform, Egyptian hieroglyphs, and Old Persian cuneiform. A fracture has damaged three of the four inscriptions. Each inscription uses more or less the same formula for venerating the king. In this case, to represent the texts on the vase, one would need to resort to at least three different encoding schemes.<sup>36</sup> What if there were a thousand vases like this, and a researcher was trying to observe the extent to which the king's name was preserved across all of the inscriptions on all of the vases?<sup>37</sup> Such a scenario would require three steps:

1. extracting the king's name from the documents
2. assessing whether the king's name is damaged or not

---

34 For instance, a dot above an Aramaic word in CAL does not mean that there is a damage at that point of the sign; instead it means that the sign is damaged, so the reading can be doubtful.

35 Darius I ruled the Achaemenid Empire from 521 to 486 BCE. For the publication of the vase, see Stolper and Goodnick (2002).

36 In this case, the encoding schemes would be C-ATF, Unicode (for Old Persian), and mdc-88.

37 We are assuming here that the texts address the king with the same name, but in different languages and thus with slightly different phonetic variations.

3. finding a way to express in a common unit the extent to which the king's name is preserved

The first step would involve a parser for the encoding schemes.<sup>38</sup> The second step would involve a unit test for the parsed elements, and the third step would require the normalization of the parsed elements.<sup>39</sup> However, several obstacles hinder these steps. First, as of February 2017, there is no parser that can deal with all of the encoding schemes mentioned above. This can be remedied with some effort, because there are parsers for some of the encoding schemes.<sup>40</sup> Nonetheless, there can be no unit test for it, although we can try to work with the damaged signs in the king's name by parsing the encoding schemes with regular expressions.<sup>41</sup> In addition, the normalization effort required by the third step implies the following two assumptions:

1. The same input value of the normalization process always represents the same phenomenon in real life.
2. There is a homogeneity in the types of the input value of the normalization process.

The first assumption holds that each encoded minimal distinct unit regularly refers to its correspondent (that is, there is a one-to-one correspondence

---

38 Parser: a procedure or a set of procedures, for determining syntactical structures of a given. The "given" can be a series of instructions for executing a computer program or a sentence attested in a natural language (Chapman 1987, 2).

39 Normalization: the process of expressing different phenomena with a common unit. For example, one cannot add apples to oranges, but one can add fruits to fruits. Thus, in a context in which we are interested only in adding fruits, since apples and oranges are fruits, we can add apples to oranges. The passage from apples to fruits is a case of the application of a normalization procedure used throughout the paper.

40 For example, a parser for MdC is written by Serge Rosmorduc in his Jsesh software: <<http://jseshdoc.qenherkhopeshef.org>> (accessed March 13, 2017). The current author is also working on a feature extractor for MdC to be incorporated into PySesh (Eraslan 2017b, <<https://github.com/D-K-E/PySesh>> [accessed March 11, 2017]). Another parser for C-ATF, now in the alpha version, is being prepared by the current author (Eraslan 2017c, <<https://github.com/D-K-E/c-atf-feature-extractor>> [accessed May 30, 2017]).

41 Regular expressions: a group of characters corresponding to a multitude of characters; thus, the characters of a regular expression are meta-characters. For example, in the Python implementation of regular expressions, the characters `\w`, `\d` correspond to an alphabetical character and a digit, respectively. There are various implementations of regular expressions throughout the programming languages. See in this volume, Pagé-Perron (205), who uses regular expressions to extract the text's vocabulary.

between the encoded distinct unit and its correspondent).<sup>42</sup> The second assumption holds that all the encoded distinct units refer to their correspondents in the same manner. The first assumption is false.<sup>43</sup> The second one is also not true, but it requires a little more attention.

42 The terms minimal distinctly perceived unit and minimal distinct unit are interchangeable.

43 The proof of this statement lies in the comparison of the encoded documents to their real correspondents; it is as if we are trying to prove that the ocean is made up of water. The following comparisons can also help the computer scientist understand the extent to which the specialist in ancient history considers the character of their work to be the same. From there, the computer scientist can get an idea about how much the encodings represent the actual documents. To observe the difference between the encoded document and its correspondent, the reader can refer to the following documents as examples: For MdC, compare the photograph of Caire O.25338 with the digital version distributed with Jsesh. Specifically, the first two signs of “V30” and “nb,” in the eighth column from the right in the plate, and the seventh line from the top in the digital version. In the photograph the first one is a closed shape, whereas the second one is clearly not a closed shape. However, it is without a doubt that both signs represented the same semantic value; thus, they were encoded with the same character, “V30.” See Daressy (1901, pl. 60, no. 25338). For C-ATF, compare the photograph of P480793 to its transliteration. At the second line of the first column, the last two “a”s are encoded in the same manner. In the photograph, one sees that the last “a” is actually slightly damaged on the superior side, but this does not change the reading of sign because one can count all of the elements of the sign. Nevertheless, it is clear that the “a”s do not represent the same thing here. For P480793, see <[http://cdli.ucla.edu/search/archival\\_view.php?ObjectID=P480793](http://cdli.ucla.edu/search/archival_view.php?ObjectID=P480793)> (accessed January 28, 2017). For CAL Code, compare the hand drawing of the TAD C3.7 with its encoding: <[http://cali.cn.huc.edu/get\\_a\\_chapter.php?file=23350](http://cali.cn.huc.edu/get_a_chapter.php?file=23350)> (accessed November 4, 2016). See the “n” of the “ryqnn” in Column E., Recto Line 6, and the “n” of the “mndt”: the first one clearly has a longer superior part, but it is not visible in the encoding. Actually, the last “n” of the “ryqnn” is encoded as “ryqnnN,” with the “N” corresponding to word final form of the “n” character. One can see that this does not go well with the hand drawing given in the publication, because, in that position, the hand drawing presents a normal form of the “n” character, as opposed to the word final form. For the hand drawing, see Yardeni (1994, fig. 1). For EpiDoc, compare the photography of the TM 25656 with its encoding. Compare the omegas of the Apollon, the omega of the toi in the first line, and the omega at the end of the fourteenth line. In terms of the encoding and interpretation, they are all omegas. However, the photograph shows that the “t” of the toi touches the omega, and that there is clearly a space in the middle of the sign at the final omega on the fourteenth line. For the encoding, see Duke Databank of Documentary Papyri (<<http://papyri.info/ddbdp/bgu;249>> [accessed March 24, 2017]). For the photo, see the online catalogue of the Ägyptisches Museum und Papyrusslung, Staatliche Museen zu Berlin (<<http://berlpap.smb.museum/02131/>> [accessed March 24, 2017]). Since the Bavant-XML standard directly follows the transliteration conventions in Elamite, there is no point in discussing

The second assumption claims that encoding schemes admit the same attributes for their elementary unit. On the contrary, there is not homogeneity among the types of the input value in the normalization process. For example, consider, hypothetically, that one wants to represent the damage on a certain “na” sign in an Elamite inscription.<sup>44</sup> The bottom part of the vertical wedge is damaged. This information is impossible to express in C-ATF, EpiDoc,<sup>45</sup> Bavant-XML, and CAL Code. MdC does, however, provide an indirect way to express this information with its “\shading” attribute,<sup>46</sup> through which one can pinpoint which part of a sign is damaged. In its current stage of development, the “\shading” divides the sign to four sectors, but, with a little extension of its syntax, we can use it to pinpoint the damage. Therefore, we can say that there is a substantial difference between how the “na” sign is conceived throughout the encoding schemes. For C-ATF, EpiDoc, Bavant-XML, and CAL Code, the sign would be considered as a monolithic block, but for MdC, a sign can be considered in quadrants. Hence, there would not, and cannot, be homogeneity in the initial input types.

The empirical refutation of the first assumption, combined with the logical refutation of the second one,<sup>47</sup> shows us that the premises required by the normalization process are indeed impossible, impeding the normalization process itself. Thus, the third step required by an epigraphically reasonable query of an ancient multilingual document is not possible. This proves effectively that the current status quo maintained by the providers of these encoding schemes, if it rests unchallenged, does not, cannot, and will not support a viable way to interact with multilingual documents on an epigraphic plane.<sup>48</sup>

---

whether the encoding scheme reproduces the inscription or not. In essence, we would be discussing whether the transliteration conventions of the Elamite studies represent their object well or not, and this is a topic beyond the scope of this paper.

44 An example of the “na” sign can be found in the picture referenced in note 43, above, for the Bavant-XML standard. It is the second sign from the left.

45 For EpiDoc, it depends on the namespace of the project, and, in any case, there is no native support for it.

46 Shading attribute: an attribute that corresponds to a feature of the MdC encoding scheme: for example, `A1\shading13` which would add diagonal lines to the first half of the sign, indicating damage on that particular area.

47 In the  $p \Rightarrow q$  propositional logic, our refutation follows the format of *reductio ad absurdum*; for a small example, see Copi, Cohen, and McMahon (2016, 420–422).

48 As of February 2017, with major exceptions being C-ATF and EpiDoc, we lack means for even recording documents as multilingual. Because multilingual documents are typically considered outliers for the projects that have provided these encoding schemes, the absence of support for these documents should not be surprising. As a result, with the major exception of C-ATF, with its language switch, and EpiDoc, the encoding schemes

### (A Possible Part of a) Solution

The only real challenge for ensuring epigraphic interoperability is to find a solution to the normalization problem discussed above. There are two possible solutions:

1. Extend the syntax of the encoding schemes so that they treat signs as a set of sectors,<sup>49</sup> as they are treated by mdc.
2. Encode visual phenomena rather than semantic units, then map the encoding to the semantic unit.

Both of these solutions would work for achieving homogeneity at the input level of the normalization process. However, due to the practical implications of the extension of the syntaxes, the second option—the encoding of the visual phenomena—is a more viable option in the long run.

What would the extension of the syntaxes of the encoding schemes imply? It would imply more precision regarding the state of conservation of the sign. How could one attribute a value to the extended syntax? Since such a value has not been generated before, one would have to go back to a photograph or facsimile to record the value in question. This would entail a lot of repetitive manual work, but it is possible to write a program so that the computer can do the heavy lifting.

Optical Character Recognition (OCR)<sup>50</sup> has become a viable option to use in epigraphy largely due to the popularization of the deep-learning algorithms.<sup>51</sup> The passage, however, between the raster image and the vector graphic, which can be stored in encodings, is actually quite tricky. The difference between the raster image and a vector graphic is that a raster image is made up of a matrix of numerical data for the computer called pixels, whereas the vector image

---

devised to conserve the documents of the first-millennium BCE Mediterranean area empires do not offer the necessary means to represent the multilingualism of the cultures.

49 Extending the syntax: understood here as adding additional features to the encoding scheme.

50 OCR: a subfield of computer vision that permits machines to extract text characters from a pixel-based image. The overall structure of the form recognition is concisely described by Cheriet et al. (2007, 6–7). See in this volume, Prosser, 322–323. For other practical applications of computer vision, see Dawson-Howe (2014), and, more recently Peters (2017).

51 Deep Learning: a subfield of machine learning. It is characterized by the breakdown of complex representations into a hierarchy of simple components during the learning process. For a concise introduction into its theory and practice, see Goodfellow, Bengio, and Courville (2016, 1–8).

comprises a group of lines passing through points defined in a coordinate system, that is, a set of equations instantiated by a set of points. Roughly speaking, the vectorization process of a raster image and the identification of it as a character address two related goals. The latter is relevant to our discussion because the languages of some of our texts are written without dividers between words. One cannot understand where one sign ends and another begins without recognizing both signs at first.<sup>52</sup> Thus, in order to isolate the signs for the purpose of recording a value to the extended syntax, some sort of OCR-based method is necessary.

Since an OCR-based approach is inevitable, even in the simple case of extending the syntaxes, there is no reason to content ourselves only with recording values to an extended encoding syntax. This would neglect a lot of the OCR data, most notably the local visual features that are associated with the provenance of the script.<sup>53</sup> The OCR process would actually generate a features' vector that could be compared against the abstract vector of the sign.<sup>54</sup> If the comparison gives a positive result, then the image from which the features' vector is created is assigned (or mapped)<sup>55</sup> to the sign in

---

52 The scope of the current paper does not permit further discussion of the application of OCR methods to ancient languages treated through the encoding schemes, but OCR methods remain vital for the study of previously encoded material.

53 Examples of local features include the handwriting of individual scribes or small, regional, or historical variants in the appearance of certain signs.

54 Since documents written in ancient languages are not reproducible products like fonts, we dismiss the matrix-based approach used in early OCR technologies. Features' vector is a convenient term for the mathematical construct that contains the patterns of the image of the sign. Technically, depending on the architecture one is using, one can generate anything between a tensor and a vector as the output of the process. An abstract vector in this case refers to practically any representation of a sign that can be found in a sign list or a dictionary. These are called "abstract" because no sign in its native form would correspond to it 100%. The comparison would also implicitly require that the image of the sign in the sign list be treated by the same architecture that treats other images of the sign in order to generate the abstract vector. For a typical authoritative sign list, see Borger (2004).

55 In the context of a dictionary-data structure, to map is to assign a value to a key. Dictionary: a type of abstract data structure (also known as an associative array or symbol table) used in programming languages, implemented under the name `dictionary` in Python. A dictionary stores information as key value pairs. Use of dictionaries, at least in the context of Python, implies that we would have a non-negligible performance gain for operations in which we want to process a value by using its associated key. For a more detailed explanation, see the Python tutorial: <<https://docs.python.org/3.6/tutorial/datastructures.html>> (accessed March 24, 2017).

question.<sup>56</sup> Thus, the comparison of the features' vector derived from the input image can be used against the abstract vector of the sign in order to map the input image to the sign.

In the case of previously encoded documents, one could take the generated features' vector that would be mapped to the encoded sign and store it along with the encoded documents. This summarizes the second option (to encode visual phenomena rather than semantic units, then map the encoding to the semantic unit). Its viability boils down to the following:

1. Can a features' vector accurately represent the sign in the photograph?
2. Can a features' vector be converted to an encoding scheme of the type mentioned here?
3. How does a previously encoded document's scheme interact with the features' vector?
4. What tools can one use to accomplish this task?

The answer to the first question is that it depends on the expected accuracy. Of course, 100% accuracy would not be possible, but somewhere around 90–95% may be expected.<sup>57</sup> Although not all scholars agree that methods that provide less than 100% accuracy are acceptable, we should also consider the sheer quantity of the inscriptions. Manual intervention, which may seem like it would guarantee 100% accuracy, can also introduce errors. However, we strive for perfection in our manual and digital efforts, and certainly increasing the precision of the OCR results is an important objective.<sup>58</sup>

The answer to the second question is EpiDoc. The problem is that the encoding schemes we have discussed up to this point were not designed to support machine-to-machine interactions; rather, except for MDC, they were mostly concerned with human-to-machine interactions. This resulted in the

---

56 A note to the technical reader: I am simply describing very broadly a supervised learning process in which the ground truth is produced from the entries of sign lists.

57 Holley 2009, <<http://www.dlib.org/dlib/marchog/holley/o3holley.html>> (accessed April 4, 2017). This might be confusing to some. OCR technology, like all deep-learning- and machine-learning-based technology, reasons based on probabilities. Consequently, there is always a risk of having a false positive as the result of the process even if that risk is very low in terms of probability. Since the image from which the features' vector is created might not be mapped to the right sign, there is, again, a very small risk that the features' vector might not represent the sign contained in the image.

58 The most viable option for increasing precision may be to allow public users to edit and save OCR-results, see Holley (2009, <<http://www.dlib.org/dlib/marchog/holley/o3holley.html>> [accessed April 4, 2017]).

use of encoding schemes that are not compatible with the standard data serialization formats, such as JSON,<sup>59</sup> XDR,<sup>60</sup> and XML, used in computer science. This, in turn, has hindered our computational capacity with regard to encoded documents. It is also true that at the time these encoding schemes were produced, the data serialization formats mentioned above did not exist,<sup>61</sup> with the exception of XDR. EpiDoc is the only encoding scheme that uses a data serialization format as its basis.<sup>62</sup> This means that if a features' vector can be serialized into XML format, it can be integrated into EpiDoc with relative ease. Furthermore, it can be serialized.<sup>63</sup>

The answer to the third question (How does a previously encoded document's scheme interact with the features' vector?) is that previously encoded documents can provide the semantic content that the features' vector would be assigned. The real question is whether or not the features' vector and the semantic content would be stored in the same document. For the reasons stated above, storing both in the same document is only possible for documents encoded with EpiDoc. A viable approach that would not change the structure of the encoding scheme itself, but would serve as a markup, would be to use

---

59 JSON stands for JavaScript Object Notation. It is a serialization format, meaning that it allows us to save the state of our data. JSON is well standardized and well supported as an interchange format between different platforms.

60 XDR stands for External Data Representation. Like JSON, it is a standardized serialization format.

61 Data serialization: the process that enables us to save the data we are working on so that it can be reconstructed just as we had saved it in different platforms.

62 My experience working with around 30,000 texts encoded in EpiDoc has revealed a number of shortcomings of the format, mostly due to the changes in the scheme. Nevertheless, EpiDoc is the only encoding scheme that is based on a data serialization format, so we must work with it.

63 As I have stated, the features' vector is a mathematical construct, and representing mathematical constructs like matrices and tensors inside XML has a result that is quite verbose. JSON would be a much better choice here. The structure of the content of the serialized data, however, is not a trivial subject. I discuss this further in the "Serialization and Epigraphic Interoperability" section, below. Note that a features' vector cannot be directly included in an EpiDoc document due to the restrictions of the schema. It can be included if one modifies its schema.



linked data technologies based on ontologies,<sup>64</sup> such as CIDOC CRM.<sup>65</sup> CIDOC CRM, for example, presents classes, such as: `class E24 Physical Man-Made Thing`, to which would belong the medium of the inscription; `class E36 Visual Item`, to which would belong the picture comprising the general view of the medium; `class E37 Mark`, to which would belong a features' vector of a particular sign; `class E38 Image`, to which would belong the vector graphic of the sign; and finally `class E34 Inscription`, to which would belong the sign "transcription/reading."<sup>66</sup> This leads to our next question: are there any tools we can use to encode visual phenomena rather than the semantic unit? Here I discuss Unicode and encoding schemes, EpiDoc, and Scalable Vector Graphics (SVG).

### *Unicode and Encoding Schemes*

Unicode is a system to map characters to bytes through code points. Characters are treated as semantic units, and code points are arbitrary constructions. The mapping requires two different reversible procedures that make the system rather complex.<sup>67</sup> Most modern platforms use a Unicode-based encoding

- 
- 64 For "Linked Data" (LD) as "a set of techniques for the publication of data on the Web using standard formats and interfaces," see Wood et al. (2014, 5). For further information in this volume, see Nurmikko-Fuller, 344, and see Matskevich and Sharon, 46–47, for practical examples in Archaeology. Ontology: a term borrowed from Philosophy referring to Conceptual Reference Models (CRM), which are models that explain what things are in the digital world and how an object relates itself to the digital world. For additional information, see in this volume, Nurmikko-Fuller, 343, 347–349.
- 65 CIDOC is the International Committee for Documentation of the International Council of Museums. The CIDOC Conceptual Reference Model (CRM) "provides definitions and a formal structure for describing the implicit and explicit concepts and relationships used in cultural heritage documentation" (<<http://www.cidoc-crm.org/>> [accessed March 30, 2017]). See also Le Boeuf et al. (2015, <<http://cidoc-crm.org/Version/version-6.2>> [accessed March 30, 2017]). An interesting epigraphic application of CIDOC CRM ontology is found in Felicetti et al. (2016, <<https://pdfs.semanticscholar.org/5ac4/f10e5f824bb35ede6af40ee36489408017ca.pdf>> [accessed March 30, 2017]). For further discussions and evaluations related to the use of CIDOC CRM, see in this volume, Matskevich and Sharon, 45, on material heritage and field archaeology recording; Nurmikko-Fuller, 353–355, on the web publication of Sumerian texts.
- 66 Since the empirical observations that led to this paper were based on the provided encoding schemes and sign drawings in SVG format, not on the technologies that were recently developed to link them, the question of properties remains to be explored in a future project.
- 67 For more technical information, see Whistler, Davis, and Freytag (2008, <<https://www.unicode.org/reports/tr17/>> [accessed January 20, 2017]).

scheme since it was designed to support every character used in every language.<sup>68</sup>

Both Unicode and the encoding schemes covered above can be used as stable class identifiers for the glyphs. With encoding schemes, we can use the already encoded semantic unit as an identifier. I, however, do not recommend this option because sometimes the same signs can have different readings, and they could acquire other readings in the future. Nonetheless, since Unicode code points do not actually cover all of the signs encoded by these languages,<sup>69</sup> encoded readings can be used for class identifiers while waiting for more robust Unicode coverage.

### *EpiDoc*

EpiDoc is, as stated above, the only encoding scheme that is based on a data-serialization format. The abstract nature of the markup language lends itself well to the multilingual document. With a little parsing, we can pursue syntactic influences among contemporary languages and extract phonetic data from different languages for a relative lemma,<sup>70</sup> such as a king's name. EpiDoc generally has all the advantages and disadvantages of using an XML-based encoding scheme.

The disadvantages of using EpiDoc include:

1. EpiDoc's hierarchy of the elements is sometimes too strict to accurately express what is happening. This is especially troublesome when one is working on texts that are copies of another text in different language,

68 This includes languages such as UTF-8 and UTF-16.

69 This is most notable for ancient Egyptian, for example.

70 Syntactic influence: the observation of the order of linguistic elements in a sentence of one language in a sentence of another, linguistically unrelated, language. This is exemplified by the formula "x-š-a-y-θ-i-y : x-š-a-y-θ-i-y-a-n-a-m," "king of kings," in Old Persian royal inscriptions. Normally in Old Persian, the genitive precedes its object \*("p-a-r-s-h-y-a : p-u-ç" [Xph line 12], son [p-u-ç] of Persian [p-a-r-s-h-y-a])\* , but in Akkadian the genitive follows its object, as in "šarrū māti," meaning "kings of the land." Hence, with the genitive "x-š-a-y-θ-i-y-a-n-a-m" following its object "x-š-a-y-θ-i-y," as in the Akkadian example, the occurrence of the genitive "māti" after its object "šarrū" represents syntactic influence. Old Persian is an Indo-Iranian language descended from Old Iranian. It was used mostly, but not exclusively, in royal inscriptions in the Achaemenid Empire (550–330 BCE). Lemma: the dictionary form of an attested word.

such as those on the quadrilingual vase or Behistun inscription of Darius.<sup>71</sup>

2. To ensure a reliable result, EpiDoc requires a lot of planning on the namespace and tag usage.<sup>72</sup> Without sufficient planning on this level, especially regarding the values given to the attributes of the elements, encoding would present an inconsistent representation of the document.<sup>73</sup>
3. As a scheme for recording epigraphic phenomena, EpiDoc is rather limited in terms of the number of attributes that can be used for recording linguistic phenomena. For example, consider that we want to encode the word “λυθήσομαι” in first person singular, future tense, indicative passive.<sup>74</sup> To store the information, we would use the <w> element in EpiDoc.<sup>75</sup> TEI, the main scheme in which EpiDoc is based, supports more attributes

71 One case that truly exemplifies this dilemma is that of the Behistun inscription (522–486 BCE). The traditional division of the inscription follows the Old Persian text, but the Elamite version of the same inscription does not coincide exactly with the divisions of the Old Persian version. See King and Thompson (1907); Shahbazi (2012, <<http://www.iranicaonline.org/articles/darius-iii>> [accessed April 2, 2017]).

72 Namespace refers to the appellation of the tags with relation to the information they convey. There could be several options in determining which tag should be used for a personal name, such as John Doe Smith. We could, for example, decide to use “name,” “personalName,” or both. Which appellation better frames the interpretation of the name depends on the context, and, thus, on the project, although it should be kept in mind that the more interpretation one provides, the more time one must invest in constructing a coherent namespace. Among other examples: 1) When one uses an element and attribute in TEI (see in this volume, Bigot Juloux, 165), one must refer to TEI namespace URI. 2) There are URIs that are namespace names in XML (see also in this volume, Nurmikko-Fuller, 338–339, who describes them). The appellation of tags filling the role of markups, as we have seen in the case of XML documents, must be decided in the process of developing the namespace. Deciding what a tag should be called, however, does not necessarily result in its coherent usage. Taking the example of the name John Doe Smith, we would not know whether the name belongs to a real person or a fictional character. If we were to distinguish this aspect of the names in a document, the distinction would not come from the domain of the namespace, since it is possible for the same name to belong to both real and fictional persons. Rather, the differentiation would need to be made through the manner in which we apply the tags. Therefore, it is essential that tag usage is a planned part of a project. For a further definition of tags, see in this volume, Bigot Juloux, 163.

73 I address this issue in a forthcoming article on encoding Elamite inscriptions.

74 “Λυθήσομαι” is Greek for “I’ll be set free” (author’s translation).

75 <w> is the markup for a word in TEI. Elements: equivalent to markups in this context, they correspond to “w” in <w ana=’singular’>apple</w>. For further information, see in this volume, Bigot Juloux 165n68.

than EpiDoc,<sup>76</sup> but, even with its support, at most we would have something like: `<w type='verb' subtype='not-deponent' function='indicative' ana='future' n='1' lemma='λύω'>λυθήσομαι</w>`. We can also add the value “singular” to any of the attributes used above, but that would be encoding a semantically different feature along with another feature, which would create problems later if someone wanted to analyze the attestations of those features separately.

4. There is no Graphical User Interface (GUI).<sup>77</sup> This can be remedied on project basis, but the lack of a GUI can be a disadvantage to colleagues with less technical backgrounds.

There are several advantages to using EpiDoc:

1. EpiDoc's text-based format is easily exchangeable with peers and is platform independent.<sup>78</sup>
2. EpiDoc's XML-based format is easily exchangeable with outside parties, such as computer scientists, who may provide help with certain issues. Since XML is a mature technology, most computer scientists and engineers have experience with it. Thus, help for technical problems is readily available.
3. EpiDoc is human readable to some extent. This is important for sharing one's work with less technically trained colleagues, who would be able to understand what is happening in the document with a little effort.
4. EpiDoc has style sheets that facilitate the creation of the digital editions of the conserved texts.<sup>79</sup>
5. With EpiDoc, it is easy to reuse any already encoded material. This is tricky to notice, however, because most of the time the encodings are full of project-based decisions. However, once a document is encoded with a

76 Attributes: precisions indicating an aspect of an element, they correspond to “ana” in `<w ana='singular'>apple</w>`. For additional explanation, see in this volume, Bigot Juloux, 165n70.

77 Graphical User Interface (GUI) refers to the medium in which users interact with the functions of the computer program through visual signs and symbols as opposed to written commands, as in the case of command line interface. Programs with GUI include, but are not limited to, Microsoft Word, Open Office Writer, Internet Explorer, and Mozilla Firefox.

78 Platform independent means that it does not depend on the user's operating system, such as Windows, Linux, or macOS.

79 As far as the conservation of the document in a digital environment and the computations that can follow, electronic editions are of secondary importance, although they can support a variety of important digital humanities research projects.

scheme, it becomes a source that can be harvested for future projects that could be encoded with the same scheme. This advantage is mostly due to the maturity of the XML technology. Since there are lots of parsers available for XML, an EpiDoc-encoded document can be parsed and harvested with most of the major programming languages.

6. Although the lack of a GUI can present a disadvantage to some users, it can also be an advantage to technically trained users. By not having a GUI, EpiDoc gives the knowledgeable user more control. The absence of a GUI forces this user to become familiar with the XML structure and the limits of the scheme, which ultimately motivates better decisions regarding the project's goals.

### *Scalable Vector Graphics (svg)*

SVG is an XML-based language for describing 2D graphics. I suggest that SVG presents the most viable option for storing local glyph representations.<sup>80</sup> SVGs are a joint product of the web development and design communities, making them reliable for expressing any type of 2D graphics that require cross-platform exchange. There are also different types of cross-platform open-source software for visualizing SVGs.<sup>81</sup>

The XML nature of SVGs implies a relatively simple integration with EpiDoc. For the purpose of encoding visual phenomena, rather than semantic units, SVGs provide two very important elements for making any reasonable normalization attempt concerning the above-mentioned encoding schemes possible:

1. the abstract vector graphic to which OCR algorithms can be compared to the processed raster image<sup>82</sup>
2. a means of visualization for the generated features' vectors of the OCR procedures<sup>83</sup>

80 Dahlstörn et al. 2011, <<https://www.w3.org/TR/SVG11/>> (accessed 1 April 2017), 20.

81 Inkscape is a popular example of this software (<<https://inkscape.org/en/>> [accessed April 1, 2017]). See Bah (2011).

82 Most of these are available as fonts, so projects should not have difficulties finding materials to train their OCR algorithms. I thank Daniel Stockholm for bringing this to my attention through our discussion at Caf'E.PHE (January 27, 2017).

83 This is partially true, because feature extraction algorithms differ in an OCR process. Extracted features are especially sensitive to preprocessing of the image, so SVG cannot incorporate or visualize every feature that can be extracted with every available feature extraction algorithm. Nevertheless, if one could convert the features of a simple tracing algorithm, such as the Potrace engine, to SVG, it would be major step toward the normalization of the input required for reasonable epigraphic queries throughout the corpora.

### *Serialization and Epigraphic Interoperability*

Up to now I have tried to avoid technical terms and issues with regard to my subject, which may have resulted in oversimplified statements on technical procedures. Now I would like to address the readers who have the necessary background to follow the discussion behind these technologies. A reader who has the sufficient technical background would notice that once vector space representation is adopted, a key problem is that of the difficulty of data serialization, since serialization ensures interoperability across systems.

The choice of format, i.e., the technology, e.g., JSON, XML, etc., is actually trivial, because transformation between major serialization formats is already extensively supported by programming languages. But the interior structure of the serialized data, that is, the representation of the shape contained within the vector space inside the serialization, is what matters the most. In more abstract terms, we should be asking ourselves what kind of graph the shape generates and how we can process it with computers.<sup>84</sup> Since we are dealing with 2D structures here, the simple solution would be to represent the shape as a list of points. As the technical reader might have noticed by now, however, the general mistake of the presently analyzed encoding schemes was their lack of mathematical foundation, that is, they did not aspire to represent their data in the form of mathematical constructs that would have given them long-term flexibility and technological independence. This lack of mathematical foundation is something the new generation of research should avoid. The solution has to be extensible, especially to 3D shapes. In my humble opinion, the ideal solution would be along the lines of the following:

Let  $G$  be the graph generated by the shape  $\theta$ ,  
 let  $V = \{v_1, v_2, \dots, v_n\}$  be the vertex set of  $G$ ,  
 let  $E = \{e_1, e_2, \dots, e_j\}$  be the edge set of  $G$ ,  
 let  $v_k, v_i \in V$ , and let  $U = \{u_1, u_2, \dots, u_x\}$  where  $\forall u \in U, u(v_k) = v_i$ ,  
 let  $I = \{v_k, u_j(v_k) \mid v_k \in V, u_j \in U\}$ ,  
 then  $I \subseteq E$ .

Simply put, if  $G$  is the graph generated by the shape, the relationship between the nodes of the graph is modeled as a set of functions that maps one node to

---

For the Potrace engine, see Selinger (2003, <<http://potrace.sourceforge.net>> [accessed April 1, 2017]).

84 Graph as in graph theory:  $G = [\text{Vertices}, \text{Edges}]$ .

another node.<sup>85</sup> The application of these functions to a node simply outputs another node. As the reader might have noticed by now, I am simply proposing to model each edge between two nodes as a function. In this way, adding new information with regard to the relationship between any two points on the graph would simply mean adding new functions to their relative set of functions. Notice that this way of representing shapes gives great flexibility.<sup>86</sup> We can even represent the relationships within entire corpora of inscriptions if we consider them as components of the same graph. In terms of data structures used in programming languages, we simply have an array *S*, which contains an array *V*, an array *E*, and an associative array *T*.<sup>87</sup> The keys of *T* are pairs that contain a node as the first term and the function value of first term as the second term. These two terms, which are essentially two nodes, are mapped to an array of functions.<sup>88</sup> Notice that we do not propose to store the result of the functions; we propose to store the functions themselves.<sup>89</sup>

There are already some properties that we can infer if we model a shape in this way. For example, we would know that every vertex is related to every other vertex with a distance function, and that a damage is simply an edge cut creating different components within the shape graph, of which some are visible, and some are not, visibility being a simple coefficient between 0 and 1. Increasing the dimensionality of the vector space is also possible without changing the structure of the serialization.<sup>90</sup>

---

85 I use the terms “node” and “vertex” interchangeably.

86 As usual, the flexibility comes at the price of difficulty in standardization. A project-based standardization can be obtained by implementing type definitions for functions, that is, defining which function maps which type to which type. Even some global standardization is possible by defining custom algebraic data types, as in Haskell programming language.

87 Notice that this is not the most efficient way to represent graphs in a computing environment. One efficient way would be to use an adjacency matrix for representing *E* and *V* together. An adjacency matrix is a matrix whose rows and columns are labeled with the nodes, and whose cell value represents whether an edge is present between the node labeled in the row and the node labeled in the column.

88 The signature of the data structure *S* in Python would be: *S* = [set(), set(), {(): []}], where each element of *S* corresponds to *V*, *E*, and *T*, respectively.

89 This should give flexibility in implementing their computations, depending on practical circumstances.

90 An increase or decrease in the dimensionality of the vector space can have a non-trivial impact on the related function set of the node group. Adapting a given function to higher dimensionality, however, provides much more backward compatibility(!) than any encoding scheme we have seen so far.

Though this is the ideal solution, it would be unrealistic to expect that it would be adopted within the community of researchers in the humanities or ancient history apart from those who are particularly interested in mathematics, graph theory, and the like. Unfortunately for a potential solution, it would be difficult to convince colleagues that it is a solution that is worth spreading. Thus it would be more practical to use a technology that is easier to work with but also is capable of being transformed into the above-mentioned structure.

Here SVG technology, in my opinion, provides the middle ground, not as a serialization technology, but as the first step to transforming the physical shape into a graph. We can add other qualities of the shape with respect to its physical state by other means. Plus, there are already a lot of fonts dealing with ancient languages, including those that are supported by presently analyzed encoding schemes. Most of these fonts, if not all of them, use SVG to represent signs. Even in the worst-case scenario, in which reusing an SVG drawing from a font for modeling a physical sign being studied by an epigrapher is not possible, SVG encoded signs present a good resource as a training dataset if the dataset is well augmented.<sup>91</sup>

Ultimately, what we want is a representation that contains everything we know about the shape, and one that is open to the adding of supplementary information once discoveries are made. Computationally, this representation would provide a lot of labeled data, which in turn can be used with supervised learning algorithms, opening up all the benefits of using artificial-intelligence technologies. Unfortunately, the current scope of this paper does not permit me to detail the application of these technologies to numerous aspects of epigraphy and ancient history in general.<sup>92</sup>

### Some Practical Notes

Generally speaking, the tools, or rather the existing elements, that can help us to deal with the problem of normalization have a very simple architecture. Epi-Doc should hold the SVG graph objects, which are based on the results of

---

91 Augmenting a dataset in this context means modifying the dataset in order to increase the invariance of the algorithms working on a classification task. This technique is mostly used where Convolutional Neural Networks (CNNs) are used as a classifier. CNNs are neural networks that have a special layer called a convolutional layer, which consists of convolutional filters. They provide the state-of-the-art performance in image-classification tasks. See also in this volume, Svärd, Jauhiainen, Sahala, and Lindén, 229n19.

92 For a recent application of AI to Akkadian which incorporated the ATF scheme, see Homburg and Chiarcos (2016).



feature extraction algorithms.<sup>93</sup> First, each should have an identifier indicating its class, which can be either a Unicode code point or an encoded sign in the encoded document. Each should also have an identifier indicating the source features' vector. The namespace of the identifier of the features' vector should include: a stable identifier for the document; an arbitrary unique integer for instantiating this features' vector so that if, in the future, we have better algorithms that can extract better features, we can cast this one aside and add the updated result with another arbitrary unique integer; and, lastly, the position of the sign over an arbitrary unique decimal that represents the total number of signs in the document in order to instantiate the current sign count for marking changes in the future.<sup>94</sup> If the use of EpiDoc is impossible as a scheme, or if it is not really desired, then SVGs should be mapped to their relative objects at first again in a separate file. Then, the signs in the encoded document would be mapped to the SVGs with LD technologies as mentioned above.

Currently, no project uses the implementation described above, but there is an ongoing effort in that direction.<sup>95</sup> In terms of staff, any project that wants to deploy the implementation described above would need someone experienced in OCR technologies, equipped with working knowledge of LD technologies, and possessed of computational experience with the encoding schemes mentioned above.<sup>96</sup> This staff is of course a supplement to the actual encoding

---

93 Feature extraction algorithm: the process through which one extracts the factors of variance from a digital representation of an object. For example, it could be an algorithm that detects the corners and sides in a picture of a triangle. Factors of variance (also called factors of variation) refers to elements that are believed to be constitutive of the observed state of that which is observed. This could include things such as elements of form (e.g., lines, angles), degrees of arc or convexity, color, and light intensity. Such features make the object distinguishable in a representation.

94 This last abstraction might seem unnecessary, but since the counted signs can be changed if a new interpretation of the document arises, and since change in the counted signs does not necessarily imply a change in the total count of the signs but may imply a change in the sign position, it would be wiser to instantiate the total sign count. Total sign count should be based on the already encoded material. It should include damaged signs but not gaps presented as sign numbers. For example, a gap may be as wide as three signs.

95 The main source for the observation of the problem of achieving epigraphic interoperability that is described above, the quadrilingual vase of Darius I, is being processed with the tools described above. Unfortunately, time and financial constraints have not yet permitted this research to be fully realized. However, at least the encoding work, with the kind help of Marine Béranger, continues.

96 Even if all of these qualities can be found in an individual, it is certain that that individual is an exception rather than the rule because OCR and LD technologies are themselves the

project, which could recruit as many encoders as needed, as well as expert epigraphers, philologists, and historians.

### Conclusion

This paper began by presenting schemes used for encoding the texts from the first millennium BCE. The list of schemes is not exhaustive, and their descriptions are only summaries. This discussion was intended to give a brief presentation, after which it could be seen that these encoding schemes have some interesting similarities. All of them use an intermediary element for mapping signs to the encoded document. The major difference is the nature of the intermediary element. It might be semantic or non-semantic. Another revealed issue is the problem resulting from the design decisions of the encoding schemes. Apart from the MDC, most of the encoding schemes were designed to work on a human-to-machine interaction level, meaning that the majority of the users of the encoded documents were thought to be humans, so the role of the machine was to provide and conserve the document for a human reader trained in the original language of the text. This was a design flaw, with many consequences, one of them being the problem of achieving epigraphic interoperability.

The problem of achieving epigraphic interoperability as discussed here is due to the different attributes of the elementary unit of the encoding schemes. MDC and EpiDoc consider that a sign is made up of different parts, for which it is more or less possible to describe the state of physical preservation, while the others consider a sign as a monolithic block that is described either as damaged or intact. The difference in the attributes creates an incompatibility in the normalization process required by the queries on multilingual documents. The low-level nature of the problem prevents any working solutions from within the encoding schemes.

In the end, I described part of a possible solution to this problem. I have demonstrated that if normalization is to be attained on reasonable grounds, some form of OCR would be necessary. The requirements of this process and

---

domains of specialized professional research. At the planning stage of a project, a team should not cut the budget in this area, because even if the team can find a talented individual who can do all of those things, it would be very difficult to maintain the code she or he has generated if said individual departs the project. A viable compromise would be to recruit two or three engineers to the institutional faculty and share them among the projects.

the data generated by it invite us to use a vector graph associated with a semantic unit. In light of this tool set, I recommended including some combination of EpiDoc, SVGs, Unicode, relative encoding schemes, and LD technologies. How they are integrated can be evaluated on a case-by-case basis. Some other practical concerns about the necessary staff for the integration models proposed here were also addressed. The effort to create an exemplary implementation of the EpiDoc model based on the quadrilingual vase of Darius I is ongoing.

## References

- Anderson, Lloyd, Karljuergen Feuerheim, John Jenkins, Rick McGowan, and Dean Snyder. 2000. *Initiative for Cuneiform Encoding (ICE): Report of the First Conference*. Initiative for Cuneiform Encoding, updated 3 November 2000. <<http://pages.jh.edu/~dighamm/ice/iceireport.html>>.
- Bah, Tavmjong. 2011. *Inkscape: Guide to a Vector Drawing Program*. 4th ed. Boston: Pearson Education, Inc.
- Bavant, Marc. 2014. *SVG Cuneiform Tool (v4.3)*. <<http://kursoj.pagesperso-orange.fr/cunei/>>.
- Borger, Rykle. 2004. *Mesopotamisches Zeichenlexikon*. AOAT 305. Münster: Ugarit-Verlag.
- Buurman, Jan, Nicolas Grimal, Michael Hainsworth, Jochem Hallof, and Dirk van der Plas. 1988. *Manuel de codage des textes hiéroglyphiques en vue de leur saisie sur ordinateur*. Paris: Institut de France.
- Chapman, Nigel. 1987. *LR Parsing: Theory and Practice*. Cambridge: Cambridge University Press.
- Cheriet, Mohamed, Nawwaf Kharmah, Liu Cheng-Lin, and Suen Ching. 2007. *Character Recognition Systems: A Guide for Students and Practitioners*. Hoboken, NJ: John Wiley & Sons, Inc.
- Copi, Irving M., Carl Cohen, and Kenneth McMahon. 2016. *Introduction to Logic*. 14th ed. London-New York: Routledge.
- Dahlstör, Erik, Patrick Dengler, Anthony Grasso, Chris Lilley, Cameron McCormack, Doug Schepers, Jonathan Watt, Jon Ferraiolo, Fujisawa Jun, and Dean Jackson. 2011. *Scalable Vector Graphics (SVG) 1.1*. 2nd ed. W3C. <<https://www.w3.org/TR/SVG11/>>.
- Daressy, Georges. 1901. *Catalogue Générale des Antiquités Égyptiennes du Musée du Caire, 25001–25338: Ostraca*. Cairo: Le Caire Impr. de l'Institut français d'archéologie orientale.
- Dawson-Howe, Kenneth. 2014. *A Practical Introduction to Computer Vision with Open cv*. West Sussex, UK: John Wiley & Sons, Ltd.

- Elliot, Tom, Gabriel Bodard, Simona Stoyanova, Charlotte Tupman, Scott Vanderbilt, and Elli Mylonas. 2017. *EpiDoc Guidelines: Ancient Documents in TEI XML*. Last modified December 12, 2017. <<http://www.stoa.org/epidoc/gl/latest/>>.
- Eraslan, Doğu Kaan. 2017a. "On the Use of Existing Resources for Ensuring Epigraphic Interoperability of Ancient Texts: What Do We Do? How Do We Do It? How Can We Make It Better?" Paper presented at the Café.PHE, École Pratique des Hautes Études, Paris.
- Eraslan, Doğu Kaan. 2017b. *PySesh: A Python NLP Complement to Jsesh*. Last modified March 22, 2017. <<https://github.com/D-K-E/PySesh>>.
- Eraslan, Doğu Kaan. 2017c. *C-ATF Feature Extractor*. Last modified May 29, 2017. <<https://github.com/D-K-E/c-atf-feature-extractor>>.
- Everson, Michael, Karljuergen Feuerheim, and Steve Tinney. 2004. *Final Proposal to Encode the Cuneiform Script in the SMP of the UCS*. <<http://std.dkuug.dk/jtc1/sc2/wg2/docs/n2786.pdf>>.
- Felicetti, Achille, Francesca Murano, Paola Ronzino, and Franco Niccolucci. 2016. "CIDOC CRM and Epigraphy: A Hermeneutic Challenge." In *Extending, Mapping and Focusing the CRM. Proceedings of the Workshop EMF-CRM 2015, Poznań, Poland, September 17, 2015*, edited by Paola Ronzion, 55–68. Prato, Italy: PIN, Servizi Didattici e Scientifici per l'Università di Firenze. <<http://ceur-ws.org/Vol-1656/paper5.pdf>>.
- Gippert, Jost. 1999. "Language-Specific Encoding in Multilingual Corpora: Requirements and Solutions." In *Multilinguale Corpora: Codierung, Strukturierung, Analyse. 11. Jahrestagung der Gesellschaft für Linguistische Datenverarbeitung*, edited by Jost Gippert, 371–384. Prague: Enigma Corporation, Inc.
- Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. Cambridge, MA: MIT Press.
- Hallock, Richard. 1969. *Persepolis Fortification Tablets*. OIP 92. Chicago: The Oriental Institute of the University of Chicago.
- Holley, Rose. 2009. "How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs." *D-Lib Magazine* 15 (3/4). <<http://www.dlib.org/dlib/marchog/holley/03holley.html>>.
- Homburg, Timo, and Christian Chiarcos. 2016. "Akkadian Word Segmentation." In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Paris: European Language Resources Association.
- Jansen-Winkel, Karl. 2014. *Inschriften Der Spätzeit IV: Die 26. Dynastie*. Vol. 1. *Psametik I.–Psametik III*. Wiesbaden: Harrassowitz.
- Johnson, Kyle P., Patrick J. Burns, Tyler Kirby, Luke Hollis, Sourav Singh, Chaitanya Sai Alaparthi, Bhupendra Singh Chauhan, et al. 2016. *CLTK: v0.1.41*. Last modified August 11, 2016. <<https://doi.org/10.5281/zenodo.60021>>.
- Kaufman, Stephen A. 1987. "Coding Principles." In *The Comprehensive Aramaic Lexicon: Text Entry and Format Manual*, 5–19. Baltimore: The Comprehensive Aramaic Lexicon. <<http://cali.cn.huc.edu/pdfs/CalManualIntrol.pdf>>.

- King, Leonard William, and Reginald Campbell Thompson. 1907. *The Sculptures and Inscriptions of Darius the Great on the Rock of Behistûn in Persia: A New Collation of the Persian, Susian and Babylonian Texts*. London: The British Museum.
- Lang, Serge. 1986. *Introduction to Linear Algebra*. Undergraduate Texts in Mathematics. 2nd ed. New York: Springer.
- Le Boeuf, Patrick, Martin Doerr, Christian-Emil Ore, and Stephen Stead, eds. 2015. *Definition of the CIDOC Conceptual Reference Model*, 6.2. <[http://cidoc-crm.org/sites/default/files/cidoc\\_crm\\_version\\_6.2.pdf](http://cidoc-crm.org/sites/default/files/cidoc_crm_version_6.2.pdf)>.
- Nederhof, Mark-Jan. 2013. "The Manuel de Codage Encoding of Hieroglyphs Impedes Development of Corpora." In *Texts, Languages, Information Technology in Egyptology: Selected Papers from the Meeting of the Computer Working Group of the International Association of Egyptologists (Informatique & Égyptologie)*, Liège, 6–8 July 2010, edited by Jean Winand and Stéphane Polis, 103–110. *Ægyptiaca Leodiensia* 9. Liège: Presses universitaires de Liège.
- Nurmikko-Fuller, Terhi. 2017. "From the Philology of Ancient Sumer to the Semantic Web." Paper presented at the Caf'E.PHE, École Pratique des Hautes Études, Paris.
- Peters, James F. 2017. *Foundations of Computer Vision: Computational Geometry, Visual Image Structures and Object Shape Detection*. Intelligent Systems Reference Library 124. Cham, Switzerland: Springer.
- Schmidt, Erich. 1970. *Persepolis III: Royal Tombs and Other Monuments*. OIP 70. Chicago: The Oriental Institute of the University of Chicago.
- Selinger, Peter. 2003. "Potrace." <<http://potrace.sourceforge.net>>.
- Shahbazi, Alireza Shapur. 2012. "Darius III. Darius I the Great." *Encyclopaedia Iranica*. Last modified May 9, 2012. <<http://www.iranicaonline.org/articles/darius-iii>>.
- Stolper, Matthew, and Joan Goodnick Westenholz. 2002. "A Stone Jar with Inscriptions of Darius I in Four Languages." *Arta* 5: 1–13.
- TEI Consortium. 2018. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*, revision f4d8439. The TEI Consortium. <<http://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf>>.
- Thonemann, Peter, Charles Crowther, Edouard Chiricat, Maggy Sasanow, eds. 2012. "MAMA 314 (Perta): Votive Bomos Dedicated by Papias to Apollo." *Monuments Asiae Minoris Antiqua XI*. University of Oxford and the Center for the Study of Ancient Documents. <<http://mama.csad.ox.ac.uk/monuments/MAMA-XI-314.html>>.
- Tinney, Steve. 2017. "CDLI ATF Primer." *Oracc: The Open Richly Annotated Cuneiform Corpus*. Oracc. <<http://oracc.museum.upenn.edu/doc/help/editinginatf/cdliatf/>>.
- Weisstein, Eric W. n.d. "Vector Space." *MathWorld*—A Wolfram Web Resource. <<http://mathworld.wolfram.com/VectorSpace.html>>.
- Whistler, Ken, Mark Davis, and Asmus Freytag. 2008. *Unicode Character Encoding Model*. Unicode Technical Report 17. <<http://www.unicode.org/reports/tr17/>>.

- Wood, David, Marsha Zaidman, Luke Ruth, and Michael Hausenblas. 2014. *Linked Data: Structured Data on the Web*. Shelter Island, NY: Manning Publications.
- Yardeni, Ada. 1994. "Maritime Trade and Royal Accountancy in an Erased Customs Account from 475 BCE on the Aḥiqar Scroll from Elephantine." *BASOR* 293: 67–78.

# Digital Philology in the Ras Shamra Tablet Inventory Project: Text Curation through Computational Intelligence

*Miller C. Prosser*

The Ras Shamra Tablet Inventory (RSTI) is a research project co-directed by Miller C. Prosser and Dennis Pardee.<sup>1</sup> A primary goal of the project is to create reliable digital editions of the texts in the Ras Šamra-Ugarit corpus within a research-database environment.<sup>2</sup> More than just a data store of texts translated from their ancient languages, RSTI serves as a tool for addressing research questions. To this end, the project also seeks to integrate archaeological data from the excavations at Ras Šamra, including published archaeological plans, grid and square systems, and any other information freely available. Using the Online Cultural and Historical Research Environment (OCHRE), we are currently adding and curating data with the help of various workflow wizards.<sup>3</sup> To add new data to RSTI, we begin with a standard text transliteration saved as a Microsoft Word document or in another common document format. We load this document into OCHRE, which uses intelligent functions to atomize the linear transcription into individual signs or letters.<sup>4</sup> As part of this process, the application validates these signs and letters to make sure there are no typo-

- 
- 1 Miller C. Prosser is a Research Database Specialist of the OCHRE Data Service of the Oriental Institute of the University of Chicago. Dennis Pardee is the Henry Crown Professor of Hebrew Studies at the Oriental Institute. See <<https://ods.uchicago.edu/rsti/>>.
  - 2 Most people are familiar with the idea of a database. A research-database environment expands on the core structure of the database with tools and other features to help users work with their data.
  - 3 A workflow wizard is an interactive database tool that guides the user through a series of common actions. For further information on OCHRE, see <<https://ochre.uchicago.edu/>> (accessed May 1, 2017).
  - 4 Atomization refers to the process of dividing data into many individual database items. A text is atomized into many database items, each of which represents a single grapheme, either a letter or a logosyllabic sign.

graphical errors.<sup>5</sup> Once the text is added to the database, analytical wizards guide the user through the tasks of finding words in project dictionaries, adding grammatical properties to the words, and identifying people and places in the texts. The importation and curation steps both employ processes developed specifically for the task of knowledge representation of philological data.

In this chapter, we explain the OCHRE data model and many of the tools developed for textual analysis. At points the discussion turns to the technical and may sound overly complex. However, it is important to remember that most of these complexities are hidden from the user. A typical user need not understand the logic behind a complex query. They need only know that they are able to search their entire corpus for texts that attest a specific phrase, and that the query will take into account all possible grammatical variants of the words in the phrase. This approach has evolved over the course of more than a decade to address some of the most complex writing systems from the ancient world.<sup>6</sup> In a sense, the underlying data in OCHRE is complex, but no more complex than the written language, no more complex than the original text, and only as complex as necessary to address the research problem. From the users' perspective, the texts look very familiar, and the complexities live mostly under the surface, hidden behind familiar-looking views of the text.

### A Brief Introduction to Ras Šamra

The archaeological site of Ras Šamra is located on the eastern Mediterranean, near Lattakia, Syria. The ancient name for this site—and for the surrounding kingdom—was Ugarit. Ugarit was well situated, with access to trade routes, both land and sea, and to arable lands. The site was occupied almost continuously from the Neolithic period (c. 7000 BCE) through its eventual destruction in the twelfth century BCE during a regional period of instability. Later, Greek, Persian, and Roman garrisons occupied the site.<sup>7</sup> The historical significance of

---

5 OCHRE checks the text against a list of all known letters and signs in our ancient languages. If it finds a letter or sign that it does not recognize, it then warns the user that there may be an error.

6 The examples given in this chapter are taken from the languages and writing systems of the ancient Near East. However, the database and the various tools are appropriate for all languages and writing systems. Other projects using OCHRE are working with texts in modern languages such as English and German. Many of the complexities inherent in ancient languages such as Akkadian do not apply to modern languages, such as English. OCHRE is prepared to handle easier examples as much as it is ready to handle complicated examples.

7 Yon 2006, 15–18, 24.



the site was rediscovered nearly 100 years ago when a local herder happened upon underground tombs at the nearby harbor-side port of Minet al-Beida. The French Archaeological Mission was alerted and began investigations the following year. Excavations have continued from 1929 to the present, interrupted only by World War II and, more recently, by the Syrian civil war.<sup>8</sup>

In the centuries prior to its final conflagration, Ugarit was a cosmopolitan culture, exhibiting artistic and stylistic influences from Egypt, Mesopotamia, the Aegean, and Anatolia.<sup>9</sup> To date, Ugarit has yielded approximately 4,500 texts in various languages and writing systems.<sup>10</sup> The texts attest various genres, including texts conveying the economic concerns of the palace and ruling class, personal letters, accounts of ritual practice, and the famous mythological texts. This last genre has drawn a great deal of attention, as it provides an early example of a literary and religious tradition that is attested in a later form in the literature of the Hebrew Bible.

The scribes of Ugarit used a newly invented alphabetic writing system consisting of 30 letters. These same scribes were also trained in the Sumero-Akkadian logossyllabic writing system, which includes hundreds of characters in the local dialect.<sup>11</sup> Like the logossyllabic Mesopotamian writing system, Ugaritic letters are formed by impressing a stylus into a clay tablet, creating a series of wedges.<sup>12</sup> This type of writing is known as cuneiform, from the Latin *cuneus*, “wedge.”<sup>13</sup>

---

8 Yon and Arnaud 2001, 7–8.

9 Matoian 2008, 17–71; Akkermans and Schwartz 2003, 335–341.

10 Bordreuil and Pardee 2009, 8.

11 A logossyllabic writing system uses characters—or “signs”—that can represent words or syllables. For the Akkadian logossyllabic writing system, we call these word-signs logograms and the syllable-signs phonograms. Thus this type of writing is called logossyllabic, a combination of logograms and syllables. A third category of signs called determinatives plays a special role, sometimes marking grammatical information such as plurality, and sometimes marking the semantic category of a word. For example, the determinative <sup>d</sup> in the word <sup>d</sup>Aššur indicates that the word is a divine name. For the purposes of this discussion, it is only important to understand that scholars typically use font styles and formatting to differentiate the transcription of these three categories.

12 When a letter is written on a hard surface such as a stone object, the perceived outline of the wedge is scratched into the surface.

13 It is widely accepted that cuneiform writing was invented to express the Sumerian language (Woods, Emberling, and Teeter, 2010, 33). The writing system was later adopted by other cultures to express their own languages.

## The Ras Shamra Tablet Inventory (RSTI)

At the Oriental Institute of the University of Chicago, Professor Pardee's work on Ugarit has now spanned five decades, over which time he has produced a substantial corpus of published and unpublished text editions. One of the first goals of RSTI was to create a framework within which to capture, preserve, and share this work. To be clear, RSTI presents a standard text edition with translation, epigraphic commentary, philological explanations, and interpretation. We present a recomposed view of each text that resembles a traditional print publication. However, we are also using various analytical wizards to add properties in order to create text editions that can be queried, summarized, and mined as data.

Among Pardee's published works, one of the first volumes transformed and ingested into RSTI was *La Trouvaille Épigraphique de l'Ougarit*, a volume written jointly with Pierre Bordreuil.<sup>14</sup> The goal of this volume is to provide a systematic accounting of all inscribed objects discovered at Ras Šamra and of Ugaritic texts discovered at other sites. The printed volume consists of a year-by-year, object-by-object presentation of archaeological inventory numbers, find spots, measurements, and other descriptive details. Once imported into RSTI, this data provided the primary spatial outline of the site of Ugarit, from the broadest excavation area down to each specific find spot.

Objects, texts, and images are the three main categories of data in RSTI. The database has entries for 5,700 objects, mostly tablets, but also vessels, seals, axe heads, and other items. To date we have added 950 text transliterations. The project currently includes over 32,000 tablet photos and drawings. Because one of our primary goals is to create a corpus of reliable text editions, we are taking care to add transliterations that meet project standards. This means that we are beginning with text editions created from first-person inspection of tablets in the National Museums of Damascus and Aleppo and in the Louvre.<sup>15</sup> It is our vision that RSTI will become a digital publication platform for all of these text editions. In the end, a text edition will include information about the tablet: its find spot, dimensions, and other observations about its physical characteristics. The edition will also include a text transliteration, a translation where

<sup>14</sup> Bordreuil and Pardee 1989.

<sup>15</sup> I was fortunate to gain access to study the Ras Šamra materials in the National Museums of Damascus, Aleppo, and in the Louvre thanks to generous research grants from the University of Chicago and to permission granted by the joint Syrian and French Mission at Ras Shamra (officially titled in French, "Mission archéologique syro-française de Ras Shamra – Ougarit"). Over the course of four separate visits, I was able to study and photograph hundreds of texts.

useful, specific epigraphic commentary, commentary on structure and interpretation, and bibliographic references.

### **An Overview of the OCHRE Ecosystem**

The OCHRE database runs on a server professionally supported by the Digital Library Development Center at the Regenstein Library on the University of Chicago campus. All core data is stored in this database.<sup>16</sup> Users access data through the OCHRE Java application client. This client interface is a Java Web Start application that launches on any computer with an internet connection.<sup>17</sup> Because OCHRE is an online-database environment, project members from anywhere in the world have access to data live and in real time. If a user in Europe edits an item, users in North America immediately see these edits. The database, through the mediation of the OCHRE application, communicates with various external web servers and external databases. Project resources such as digital images, PDFs, and other supporting files are stored on these external servers and accessed for viewing and manipulation in the OCHRE client.

OUCHRE data can be sent to external programs through an API.<sup>18</sup> For example, core OCHRE data can be sent to an R Server for statistical analysis and visualization, then returned through the database to the OCHRE client for viewing.<sup>19</sup>

---

16 RSTI is an OCHRE project. There are various other projects working on philological, archaeological, and other types of research in OCHRE. Projects vary greatly in size, from a few researchers to a network of international institutions. Each project has a discrete set of data that is unavailable to other projects by default. Project data is made accessible through credentials that are specific to users and projects. All data is stored securely and backed up on University of Chicago servers.

17 Java Web Start: a technology that allows one to launch a computer program without going through an installation process. For more information, see <<https://ochre.uchicago.edu/page/java-user-interface>> (accessed May 1, 2017). Java was chosen as the development language because it offers powerful features and allows deployment on various operating systems.

18 Application programming interface (API): a set of rules and tools that defines how computers can interact. In this context, the OCHRE API defines what database items are available to extract and send to other programs.

19 R: an open source programming language used for statistical data analysis. See <<https://www.r-project.org>> (accessed May 1, 2017). An implementation of the R software environment can be installed on a server and accessed remotely by OCHRE, saving the user from having to install and maintain R on their local computer.

## An Overview of the OCHRE Data Model

One of the simple principles underlying the OCHRE data model is that each discrete meaningful unit of observation is a separate database item.<sup>20</sup> Each database item is stored as an XML file.<sup>21</sup> Items may represent things that vary greatly in scale and type. An item could be an entire archaeological site or a single seed discovered in the course of excavation. An item can be anything from a place, to a person, to an image, to a text, to just about anything observable or conceptual. As mentioned above, the process of dividing data into these discrete units of observation is called atomization. In the end, these units function much like atoms, coming together to form larger and more complex entities.

Millions of individual XML files are related to one another through a variety of organizational methods, with the primary among these being the hierarchy. For archaeological contexts, the hierarchy expresses the relationship between broad spatial areas and more specific areas contained therein. To take an example from RSTI, the database item that represents the site of Ras Šamra stands in the hierarchy above, and contains, the area of the site called the Royal Palace.<sup>22</sup> The Royal Palace contains various rooms beneath it in the hierarchy. In these various rooms, we find many of the items that represent the inscribed tablets. In this way, the hierarchy organizes locations and objects into discrete areas into which all items can be contextualized spatially. As we shall see below, hierarchies play a central role in organizing textual data. In addition to organization through hierarchical relationships, database items can be linked in a wide range of ad-hoc and cross-cutting ways.<sup>23</sup> This approach, called the semistructured item-based approach, is in contrast to the class-based or relational data model in which similar items are stored in tables of columns and rows, then joined with other tables based on a common column called a key. In

---

20 All databases have an underlying data model, which is simply an abstraction that defines how data is connected and processed in the database. A data model is like a framework with rules. Applied to Archaeology, see in this volume, Matskevich and Sharon, 48.

21 XML stands for Extensible Markup Language and is a flexible data format. For further information, see in this volume, Bigot Juloux, 163–164.

22 For an accessible discussion of the site of Ugarit, see Yon (2006), specifically pages 36 and following for a discussion of the Royal Palace.

23 In the world of the digital humanities, the term “linked” can have a specific connotation, meaning the mode of modelling data for sharing across the web with other datasets with which one’s data was previously not connected. Within the OCHRE system, the term refers to database items that “point” to each other, thereby creating a link between them. The database is a network of millions of files pointing at each other, i.e., linked to each other.

a semistructured item-based data model, each individual unit has a universally unique key and is free from the inherent restrictions of a table.<sup>24</sup>

### The OCHRE Ontology and Data Types

We have just hinted at some of the ontological categories of data in the OCHRE data model. OCHRE employs what, in the world of information science, is called an upper ontology.<sup>25</sup> It is a highly generic, non-specifying schema of data categories applicable for use across many research domains. The categories of data types, such as Location, Person, Text, and Resource, are very broad. Each category of data is slightly different from every other, both in conceptual definition and in practical implementation. These data types are presented to the user as different hierarchies. What follows is a brief description of the data types that play a central role in RSTI.

The Locations & Objects data type is used for items that exist in space. Typically these are places and physical objects. These can be real places, observable in our current world or in excavated ruins, such as cities, neighbourhoods, streets, buildings, and rooms. These items can be movable objects that exist in space, such as an ancient coin, a modern book, or a clay vessel. Locations & Objects can be defined by spatial-coordinate data to indicate precise location in space.

The Persons & Organizations category is fairly self-explanatory. The primary unit of data in this category is either a person—living, deceased, real, or fictional—or an organization—anything from a publisher to any other conceptual group of persons. A person from this category of data is associated with the occurrence of their name in the text. In this way, a project can build relationships between names in texts. Properties can be added to each person to identify their familial connections, vocational roles, or any other piece of information that may help define them. In RSTI we are interested in identifying relationships of power among people of differing social strata. Below we explain some of the database tools that we have developed to aid in adding this type of information to the database.

<sup>24</sup> Thuraisingham 2002, 155–171.

<sup>25</sup> Schloen and Schloen 2014, <<http://www.digitalhumanities.org/dhq/vol/8/4/000196/000196.html>> (accessed May 1, 2017). For additional discussion, see in this volume, a) Bigot Juloux, in particular, 165–181; b) especially as applied to online publishing, Nurmikko-Fuller, 343, 348–350, 353–360; c) briefly, as applied to ontology in data-sharing in the archaeological field, Matskevich and Sharon, 47.

The Writing Systems category is foundational to all philology projects in the OCHRE database. Therefore, it is important to define some terms and explain the structure of this data type. A writing system is defined by a series of script units. A script unit is more than just a letter or a sign in the writing system. Each script unit is defined by various readings and allographs.<sup>26</sup> The English writing system would be fairly simple. There are only 26 script units, one for each letter of the alphabet. Languages from the ancient world are slightly more complex. In OCHRE we have created a writing system that represents the logosyllabic cuneiform writing system used in the ancient Near East. All research projects in OCHRE have access to OCHRE's standardized logosyllabic writing system. Like the list of letters that represents the English alphabet, this list represents all the known signs from ancient languages such as Sumerian and Akkadian. The list is standardized in the sense that it represents an attempt to create an architecture into which every sign from every logosyllabic writing system in the Sumero-Akkadian tradition can be recorded. In other words, the sign list is not divided into separate sign lists based on language or dialect.<sup>27</sup> On a practical level, the sign list is used for importing and performing automated processing of textual data at the data ingestion stage.

The Texts category will receive more attention in the following sections, but a few remarks will place this category into context alongside the other data types. As mentioned above, tablets and other objects are recorded in the Locations & Objects hierarchy. The object is not the same as the text inscribed on the object. The text is the series of signs used to communicate information. Inscribed objects from the Locations & Objects category are linked to the texts recorded on those objects.

The Dictionary is another category of data in OCHRE, and it has its own organizational structure. Dictionary data is highly integrated with textual data and plays another central role in RSTI. Any given dictionary is populated with lemma items. A lemma can be simple, with a basic definition and description, or it can be complex, with nested sub-entries of meanings. A lemma is further

---

26 A reading is a value represented by a sign. An allograph is a variant form of the letter, exemplified in most modern alphabetic systems by uppercase and lowercase letters.

27 The OCHRE sign list is based in part on Rykle Borger's *Mesopotamisches Zeichenlexikon* (2004) and supplemented as needed. In the next iteration of the sign list, we will add properties to the various signs and readings to indicate in which languages and dialects they are attested. This will allow the user to extract a list of signs that represent a dialect-specific sign list. We have published a version of this sign list online at <<https://ods.uchicago.edu/signary/>> (accessed May 1, 2017).

defined by phonemic forms, which in turn are defined by attested forms.<sup>28</sup> To explain a bit further, the Akkadian writing system allowed for variation in the spelling of a given word, even among words with the same grammatical properties. These various spellings are recorded in the database as attested forms. An attested form is a form of the lemma as it occurs in a text, represented by a sequence of signs. A phonemic form represents a grammatical interpretation of an attested form. In many cases, a given phonemic form is represented by many attested forms. To view this system as a hierarchy, an attested form is the lowest level of the hierarchical organization of items: lemma > phonemic form > attested form.

The Resources category organizes various supporting files, including images, PDFs, audio files, videos, GIS files, or webpages. Typically a resource will include an address where the file can be found, either on the web or on a project server. Most of these files can be viewed directly in OCHRE even though they are loaded from external servers.

### Modelling Texts in a Database

To make philological data useful both to the researcher and to the interested public, it is essential to model the data in such a way that signs and words can be queried, referenced, and recombined easily and accurately. For English and some other modern languages, one can employ various computational methods that fall under the broad umbrella of Natural Language Processing (NLP) to analyze large blocks of text without first storing the text as tokenized units such as words.<sup>29</sup> Even if a text is not yet digitized, Optical Character Recognition (OCR) processing can usually produce a sufficiently accurate text. This process typically fails to identify a small percentage of letters in a printed document. However, this level of inaccuracy does not usually interfere with the analysis. Presently, and for the foreseeable future, OCR is not available for clay

---

28 Again, this complexity is less critical for modern languages; however, the highly variable and defective writing systems of the ancient world require these divisions.

29 NLP is a broad term that refers to methods for teaching computers to understand human language. At the risk of over-simplifying, the idea is that after a computer has processed a large number of words in a text, it should be able to extract meaning from a text it has never seen. Most work in this field has focused on teaching computers to understand English. The benefits of NLP have not yet reached the study of ancient languages. In this volume, see Svård, Jauhiainen, Sahala, and Lindén (238–246), who focus on Pointwise Mutual Information (PMI), related to NLP, for Akkadian semantic research.



tablets.<sup>30</sup> In the meantime, we are left to determine the most effective method for handling ancient texts.

We propose a basic principle of best practice: textual data should be stored in the database in units that do not require further atomization before meaningful analysis can be performed. In other words, textual data should be stored as very small atomic units, either as words or as letters. How would one create a dictionary of attested forms when data is stored as a continuous text or as lines? One would be required either to duplicate words in another table that represents lemmata or to transform the existing data into discrete words. Both these extra steps are error-prone and cumbersome. In contrast, data that is highly atomized and granular, that is organized and described, can be accessed and recombined quite easily for multiple types of display and analysis. In essence, this is the power of a semistructured item-based data model.

As is the case with the other categories of data presented above, textual data in OCHRE is atomized into minimal meaningful units. When we think about texts from the ancient Near East, I think it is very clear that the minimal meaningful part is the grapheme (i.e., the syllabic sign or the alphabetic letter). There are many reasons to atomize texts into individual graphemes. First, many projects in our field are working on establishing reliable text editions, either as first editions of newly discovered texts or as re-editions of texts that deserve further attention. In this process, the project will require the ability to make comments and record metadata about every sign. Is the sign partially broken? Is it a scribal correction? Is it written above the line? From the perspective of someone outside the field, this may seem like an extreme degree of atomization. But from the perspective of a philologist, this is absolutely necessary. Therefore, each letter is its own database item. When the user wishes to view a text, OCHRE recomposes a view of the text based on hundreds or thousands of database items.

We have seen above how the hierarchy is a primary organizational structure for data in the OCHRE database. Textual data is also organized into hierarchies.

---

30 OCR: a process by which a machine scans a typed or handwritten document and converts the text into digital characters. For decades now, various attempts have been made to produce a system that can analyze digital images and identify cuneiform signs (Dirksen and von Bally 1997). No doubt advances will continue to be made toward this end. However, the variation evident across scripts and languages will make it difficult to achieve the level of accuracy currently available from OCR of printed English texts. If it becomes possible to capture even half of a cuneiform text accurately, this would save some effort on the part of the scholar. In the end, however, the scholar will still need to verify and correct the text edition because no level of inaccuracy is acceptable. For additional information, in this volume, see Eraslan, 296–297.



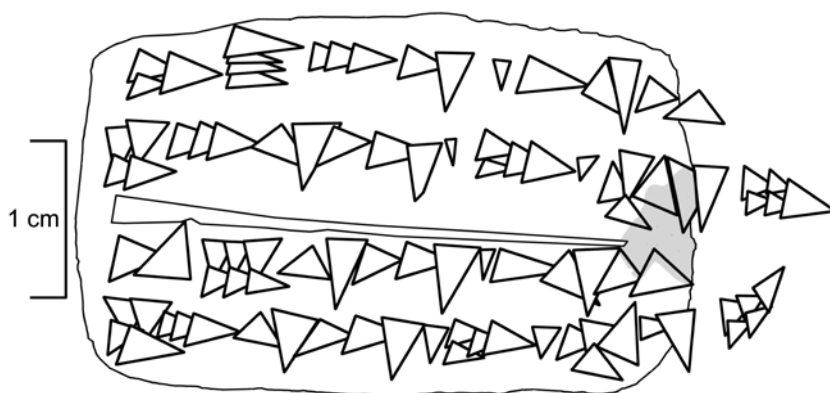


FIGURE 10.1 *Drawing of RS 3.320*

OCHRE maintains a conceptual distinction between the signs as they are visible on the object and the signs as they are interpreted as words by the scholar. The signs as they are visible on the object are organized into an epigraphic hierarchy, sign by sign. These same signs—as they are interpreted as words by the scholar—are organized into a discourse hierarchy. Any given word in the discourse hierarchy includes a list of links to the signs from the epigraphic hierarchy that compose that word. In this way, every word is associated with its constituent signs. This relationship between the grammatical form of the word and the attested spelling is leveraged to produce a dynamic dictionary.

Any typical text of 20 or 30 lines is composed of hundreds of database items. Taking a fairly simple example, the Ugaritic text RS 3.320 is a short text (Fig. 10.1).

It is composed of 36 signs and ten words. The text records two categories of cultic personnel (i.e., some type of priests or workers associated with a temple), each group with nine men and a donkey.<sup>31</sup> Following is the transliteration of the text I produced while studying the tablet in the National Museum of Aleppo.

<sup>31</sup> The text is probably a census taken by the royal palace, but one may ask if the real-world events behind the text are best described as a conscription or a work-assignment. The information that we would need to reach this type of conclusion was omitted by the scribe because it was either obvious to all parties involved or, possibly, irrelevant to the situation.

Transliteration	Vocalization	Translation
(01) khnm . tš <sup>1c</sup>	(01) kāhinūma tiš <sup>c</sup> u	Priests, nine
(02) bnšm . w . ḥm <sup>1r</sup>	(02) bunušūma wa ḥimāru	men and a donkey;
-----	-----	
(03) qdšm . tš <sup>c</sup>	(03) q-d-šūma tiš <sup>c</sup> u	cultic personnel, nine
(04) bnšm . w . ḥmr	(04) bunušūma wa ḥimāru	men and a donkey.

The first word of the text is *khnm*, the masculine plural nominative form of the noun, which may be vocalized as “*kāhinūma*,” “priests.”<sup>32</sup> The four letters k-h-n-m are organized in the epigraphic hierarchy. My interpretation of the word as “*kāhinūma*” is organized in the discourse hierarchy, but it is also linked to the letters in the epigraphic hierarchy. To be clear, the letters that make up the word in the discourse hierarchy are the same database items found in the epigraphic hierarchy.

### Recomposed Texts

Even though its textual data is atomized into individual letters, OCHRE produces recomposed views that look like text editions that are familiar to scholars. The image in Figure 10.2 is a screen capture from RSTI. On the left, we see the first part of the epigraphic hierarchy, expanded to show the individual letters. These letters are used to recompose the transliteration view of the text. The discourse hierarchy is used to create the recomposed view of the vocalized text (Fig. 10.2).

Textual data can be transformed very easily into other data formats, such as tables, PDFs, or word-processing documents. By creating a PDF directly from OCHRE, the user has an easy mechanism for publishing texts in a traditional paginated format. OCHRE can also publish textual data in a format appropriate for publication on the web. In this model, OCHRE data is published to an internal publications database in the OCHRE database environment, which in turn is made available to be accessed by a standard website. One final note on this

32 The Ugaritic alphabetic writing system indicates vowels only partially and only indirectly by use of three aleph signs: ā, i, and ū. Otherwise, no vowels were written (Bordreuil and Pardee 2009, §3.2-3.3) The vocalized form of the word is meant to convey the grammatical interpretation of the word.

RS 3.320 (KTU 4.29, UT 63)	RS 3.320: Transliteration	RS 3.320: Discourse view
Epigraphic hierarchy ▼ € Obverse ▼ € 01 € k € h € n € m	<b>Obverse</b> (01) khnm . t <sup>f</sup> š <sup>lc</sup> (02) bnšm . w . ḥm <sup>l</sup> r ----- (03) qdšm . tš <sup>c</sup> (04) bnšm . w . ḥmr	<b>Obverse</b> (01) kāhinūmatiš <sup>c</sup> u (02) bunušūmawa ḥimāru ----- (03) q-d-šūmatiš <sup>c</sup> u (04) bunušūmawa ḥimāru

FIGURE 10.2 *Text, recomposed view in RSTI*

topic: OCHRE can also publish data in any digital standard, such as TEI.<sup>33</sup> Because OCHRE data is highly atomized, it is simply a matter of defining the desired format of the recomposed document.

### Importing Texts

Textual data can be imported into OCHRE either from legacy formats or by typing directly into the database. Because OCHRE has been developed to handle all writing systems, the import process is highly flexible, customizable, and powerful.

POCHRE does not impose a transliteration system upon projects. Each project is free to customize OCHRE to understand the significance of lowercase, uppercase, italic, and superscript transliteration. For example, RSTI uses the following transliteration style for texts in which a logosyllabic script is used to record the Akkadian language:

Lowercase italic = phonogram, Akkadian language

Uppercase regular = logogram, Akkadian language

Lowercase regular, superscript = determinative, Akkadian language

For example, the following lines are a transcription of RS 17.238:11-12:

(11) *šum-ma* DUMU<sup>meš</sup> KUR *ú-ga-ri-it*

(12) *ša* KUR-ti *ša-ni-ti*

33 The “Text Encoding Initiative (TEI) is a consortium which collectively develops and maintains a standard for the representation of texts in digital form. Its chief deliverable is a set of Guidelines which specify encoding methods for machine-readable texts, chiefly in the humanities, social sciences and linguistics” (<<http://www.tei-c.org>> [accessed May 1, 2017]). For additional information, see in this volume, Bigot Juloux, 164–165.

During the import stage, OCHRE examines a transliterated sign, uses the specification to identify the type of sign based on formatting, then finds this specific reading in the appropriate writing system. Once imported, any given sign in a text is linked to a specific reading of a sign in a writing system. So, in the excerpt above from RS 17.238, the import process examines the first transliterated sign, “*šum*,” and looks it up in the Sumero-Akkadian writing system. Because the sign is transliterated in lowercase italic, and because I instructed OCHRE that this format is used for phonograms, the query considers only phonograms in the writing system. It finds the value “*šum*” as a phonogram in the writing system under the sign “TAG.” The sign “DUMU” is recognized as a logogram (a word sign), which in this case stands for an Akkadian word that means “son.” The “*meš*” sign after “DUMU” is recognized as a determinative, which in this case marks the noun as plural, “sons.”

For the sake of illustration, we will follow part of the text-import process for one RSTI text. The following text (RS 15.076) is a bilingual text. The recto, written in alphabetic Ugaritic, lists proper names, each with a number noun. For example, line one says, “*sākinu*, thirty.” The verso, written in logosyllabic Akkadian, lists quantities of garments.

#### Recto

- (01) *skn . tltm*
- (02) *iytlm . tltm*
- (03) *hymł . tltm*
- (04) *głkz . tltm*
- (05) *mlkn'm . šrm*
- (06) *mr'm . šrm*
- (07) *mlbù . šrm*
- (08) *mtđł . šrm*
- (09) *y'drd . šrm*
- (10) *gmrd . šrm*
- (11) *šdqšłm . šrm*
- (12) *yknıl . hmš*
- (13) *ılmlk . hmš*
- (14) *prt . šr*
- (15) *ubn . šr*

#### Verso

- (16) 3DIŠ TÚG<sup>meš</sup> GAL
- (17) 1AŠ TÚG<sup>meš</sup> TUR<sup>meš</sup>
- (18) 2DIŠ TÚG<sup>meš</sup> MÍ<sup>meš</sup>

(19) 𐤀𐤁𐤕 𐤕𐤕𐤂 𐤂𐤕.𐤀<sup>meš</sup>

(20) 𐤀𐤕 𐤕𐤕𐤂<sup>meš</sup> *ku-ub-šu*

There is no special encoding necessary to communicate with the import process. One simply supplies the import wizard with a block of text such as the one above, which looks like a standard text transliteration. In fact, this block of text was simply copied from a Microsoft Word document and pasted into RSTI. OCHRE has been trained and instructed on how to understand this document. The import process has been told that lowercase non-italicized text is alphabetic Ugaritic, that lowercase italicized text represents Akkadian phonograms, that uppercase text represents logograms, and that superscripted text represents determinatives. So, it finds “s” at the beginning of line one and understands it as a letter in the Ugaritic writing system. The import performs a query and finds “s” in the alphabetic Ugaritic writing system as a valid letter. By virtue of this link to the writing system, where we specify various properties, including Unicode encoding, OCHRE now knows that this letter can be represented as the Ugaritic cuneiform letter identified by Unicode point 𐤓.<sup>34</sup> At this point, the handling for the first epigraphic unit is complete. The import created an epigraphic unit called “s,” linked it to the script unit “s” in the Ugaritic writing system, and placed the letter in the hierarchy that represents the Recto of the text in line 1. This logical loop continues through the end of the text, handling each section, line, word, and sign.

In addition to identifying the structure of the text and the value of each sign, the import process also creates words from the signs. It understands that a space indicates a word boundary. As such, the import creates a word, “*skn*,” from the first three epigraphic units. Again, the individual epigraphic units exist in an epigraphic branch of the text’s hierarchical structure. These same epigraphic units are linked to a single word in a separate discourse hierarchy in the text’s hierarchical structure. In the end, every text is a network of signs and words.

34 We need our computers to communicate effectively with each other (and with us). On a very basic level, this means that computers must use an agreed-upon system of representing characters. The Unicode Consortium was created to promote this standard. One of the most significant contributions of this group is the creation of a standard that defines which underlying computer code is used to represent (nearly) every character in every writing system, even Ugaritic. The number of each character can be referred to as a Unicode point. For further information on Unicode, see in this volume, Eraslan 300–301.

## Text Analysis Workflow Wizards

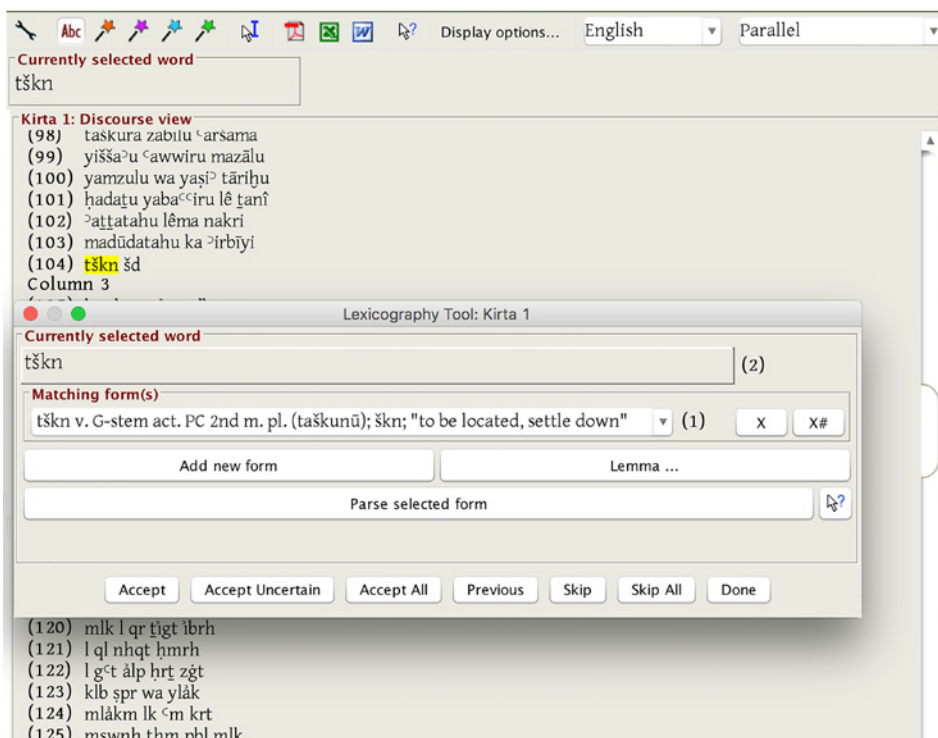
With RSTI, we aim to use our text editions to address various research questions. For example, we hope to identify socio-economic networks among the persons named in the texts. It is our working hypothesis that relations among persons of varying levels of power may be described using the terminology of patron-client relations.<sup>35</sup> To test this hypothesis, we are using OCHRE to help us identify the social positions of individuals named in texts. While this could prove to be a time-intensive activity, we have attempted to reduce the potential time needed by developing some database tools to help move the process along more smoothly. These tools, which we have been calling wizards, aid us in a variety of tasks that help us address this research question. There are currently three wizards for philological analysis: (1) the lexicography wizard; (2) the prosopography analysis wizard; and (3) the gazetteer wizard. Following is a brief overview of each of these workflow wizards.

Lexicography is the practice of compiling glossaries and dictionaries, a task we wish to perform in RSTI. Therefore, we developed a lexicography workflow wizard to link words in the text to lemmata in the dictionary, add parsing properties, and even add new words to the dictionary. To use the lexicography wizard, the user views a text in OCHRE, then launches the wizard (Fig. 10.3).

The wizard checks the words in the discourse hierarchy and finds the first word that is not yet linked to the dictionary. It searches the dictionary for attested forms that match the attested form of the word in the text. The wizard returns a list of all matches, allowing the user to select the correct word. If no match is found, the user may choose to add a new form to the dictionary while still in the workflow wizard. Once the new form is added to the dictionary, it is also linked to the current word in the text. At this point in the wizard, the user has the option to add grammatical parsing properties to the phonemic form of the lemma in the dictionary. These properties can include the typical range of nominal declension and verbal parsing properties such as person, number,

---

35 Patronage is one of the connective tissues of social integration joining individuals within and across the core institutional structures, such as kinship groups, vocational groups, and political authority. The ties between patron and client frequently augment pre-existing modes of social or economic protection such as kinship, vocational, or village ties. The relationship between a patron and his client(s) is based on a restricted access to the available productive resources (Eisenstadt and Roniger 1984, 171). This restriction does not represent an attempt to exclude a significant sector of the population from productive resources. On the contrary, the principle of restrictive access allows a wide range of individuals to gain conditional access to the available resources, primarily through the intermediary mode of exchange called the patron-client relationship.

FIGURE 10.3 *The Lexicography Wizard*

gender, and case. Over time, the import process and wizard are likely to find more matches automatically. Essentially, the user is training the database to become smarter.

The lexicographic work on the texts yields a dictionary populated with live links to the attested forms in the texts (Fig. 10.4).

To review this network of data briefly, a series of epigraphic units forms a discourse unit (i.e., a word). A word is linked as an attested form to the dictionary. Attested forms are organized under grammatical forms. The grammatical forms in the dictionary have properties that indicate the parsing of the word. The grammatical forms are organized under the lemma heading. At the level of the lemma, a project can define various meanings, assign properties, and indicate cross-references or other details. This graph network of data is one of the primary characteristics of an item-based semistructured data model. New connections are made on the fly, not based on predefined table structures.

The second workflow wizard aids the user in adding prosopography properties to personal names in the text, specifically the types of properties that

**bnš** noun; man, individual, person, worker.  
 bnš (5x) (bunušu) n. m. sg. abs. nom. QvTvl.  
 bnš (2x) (bunušu) n. m. sg. construct nom. QvTvl.  
 bnš (14x) (bunuši) n. m. sg. construct gen. QvTvl.  
 bnš (1x) (bunuša) n. m. sg. abs. acc. QvTvl.  
 bnšm (bunušāma).  
 bnš (bunušā).  
 bnšm (7x) (bunušūma) n. m. pl. abs. nom. QvTvl.

References for bnš (maximum 50 references shown).

Text	Location	Context
RS 11.858	Verso, standard orientati...	šadûbini gaṭrāni bīdi <b>bunuši</b> ʾaḡlikuzi
RS 9.453	Recto/line 03	tinā šaʾuratāmalē <b>bunuši</b>
RS 9.453	Recto/line 04	ʾarbaʿu šaʾurātulē <b>bunuši</b>
RS 9.453	Recto/line 06	ṭalātū šaʾurātulē <b>bunuši</b>
RS 9.453	Recto/line 07	ṭittu šaʾurātulē <b>bunuši</b>
RS 9.453	Recto/line 08	tinā šaʾuratāmalē <b>bunuši</b>
RS 9.453	Recto/line 09	ṭalātūmašaʾuratulē <b>bunuši</b>
RS 9.453	Recto/line 11	ṭittu šaʾurātulē <b>bunuši</b>
RS 9.453	Verso/line 23'	ṭalātūmasappulē <b>bunuši</b> tuppanuri
RS 9.453	Verso/line 24'	ʾarbaʿu sappūmadū lē <b>bunuši</b> PRWSDY
RS 9.453	Verso/line 25'	ṭittu sappūmalē <b>bunuši</b> kulanimuwa
RS 9.453	Verso/line 28'	lē <b>bunuši</b> tuppanuri dū yiʾaḥid- lē GYNM
RS 15.098	Recto/line 07	kīya m— wa pr tištaʾilū ʿimmānu <b>bunuši</b> — lā yblt ḥābiṭīma ʾapa kaspahumulā yblt
RS 15.111	Recto/line 06	lē yōmi hannadī ʿammiṭtamrubinu niqmēpaʿ malku ʾuḡārit yatana bēta ʾananīḡari bini —ytṇ <b>bunuši</b> malki dābi raʾši

FIGURE 10.4 Dictionary lemma in RSTI

would help identify a person's social, vocational, and familial connections (Fig. 10.5).

Like the lexicography wizard, the prosopography wizard iterates over the words in the text and stops at the first word it recognizes as a personal name. The wizard returns a list of matching persons, allowing the user to select the correct person. This wizard uses the properties added during the lexicography process to determine if a word is a personal name. When the wizard finds a personal name, it performs a query of all known names in the Persons & Organizations category. Because names can be spelled many ways, and in different writing systems—many names are attested in alphabetic Ugaritic and logosyllabic Akkadian—the query takes into consideration all these spellings when matching against names in the Persons & Organizations category.



**RIH 78/12: Discourse view**

Recto

(01) lê malkati ʔummiya  
 (02) ʔadattiya lê paʕnêki  
 (03) qālātu ʔida laʔikti  
 (04) ʕakkuya npl̥t



Prosopography Tool: RIH 78/12

**Currently selected word**

ʕky (1)

**Matching person(s)**

(0)

Add new person  

Describe selected person

Accept; Describe current word

Apply predefinition: Prosopography analysis

Accept Accept Uncertain Accept All Previous Skip Skip All

FIGURE 10.5 *The Prosopography Wizard*

Once enough names are identified and associated with persons, and enough relationships are defined, this data serves as the basis for social network analysis. Who is related to whom in real or fictive kinship relationships?<sup>36</sup> Who provides goods, services, or information? And what sort of hierarchical or other relationships of power can be detected in these connections?

The Ras Šamra texts frequently record the names of towns, villages, and estates that are located inside the kingdom of Ugarit or even farther afield. Using the final textual analysis workflow wizard, the gazetteer wizard, the user can associate proper nouns in the texts with places in the Locations & Objects category, essentially building a dynamic geographical index of all known places. In the economic texts from Ras Šamra, we are interested in identifying the various villages and estates throughout the kingdom that are obliged to send taxes of various sorts to the central administration. In general terms, scholars

36 The practice of adoption in the ancient world includes relationships that would be more properly described as being based on economic terms than as being based on terms of familial guardianship. These business arrangements are sometimes referred to as fictive kinship relationships (Bordreuil 1981).

have been able to place the villages into regions throughout the kingdom.<sup>37</sup> This has been achieved primarily through a detailed analysis of the co-occurrence of place names in royal administrative lists. It is presumed that cities that are frequently listed together are likely to have been located near each other. This conclusion could be tested in RSTI and further expanded to include other considerations that leverage the vast network of data.

With the textual data well modelled and described by properties, the user can begin to employ powerful queries to investigate the data. I will highlight two of the more interesting query types: the co-occurrence and sequence queries. These two queries are variations on a theme. The co-occurrence query allows the user to find any texts that attest a list of words, no matter the order in which they appear in the text. One may even query categories of words based on grammatical or other properties. The sequence query performs a similar search but returns only texts in which the words occur in a given order. This type of query may seem trivial on its face, but consider the complexities of ancient writing systems. How would we search for texts that contain the words for beer, distribution, and workers when each of these words may be spelled in five or six different ways? If we were dealing with textual data stored as lines in a text, this query would be very difficult. However, because textual data in OCHRE is organized into lemmata, with phonemic and attested forms contained therein, the query can look for texts that attest the lemma, regardless of the attested spelling. The query simply follows the links from the lemma in the dictionary to the discourse units in the texts.

## Conclusion

Textual data is complex. As scholars who study this complexity, we need digital tools that accurately model this complexity. If we are to employ a database in our work, it must meet the challenges of the textual data. RSTI uses OCHRE for this very reason. The item-based semistructured data model suits textual data more appropriately than any other current data model. The reality is that epigraphers and philologists make observations at every level, from the entire text down to the individual sign. For this reason, it is our view that textual data should be atomized into individual database items that represent the signs of a text.

In addition to meeting these basic criteria, OCHRE also provides powerful analytical tools to guide the scholar through common analytical activities.

---

37 Soldt 2005.

From generating dictionary entries based on texts to documenting social-network relationships, the scholar can work with their data in a single unified platform. Data can be reformatted and exported for use in other programs, but the primary task of describing the data can take place entirely in OCHRE. The centralized database obviates the need to maintain a separate image database, GIS database, text database, and object inventory. The OCHRE Java client makes it possible for users without programming or advanced computational skills to work with their data in a natural and familiar way. Even though the underlying data is stored in XML, the user need never interact directly with these files.

While OCHRE was developed for the archaeology and complex languages of the ancient world, the underlying data model and tools are just as applicable for modern languages or archaeological sites in other contexts. Logosyllabic Akkadian is complex, so RSTI requires all the complexities offered by OCHRE. But the available complexities of the system may be used only when needed. With a highly customizable taxonomy and flexible data model, OCHRE is ready to accommodate a wide variety of projects.

## References

- Akkermans, Peter M. M. G., and Glenn M. Schwartz. 2003. *The Archaeology of Syria: From Complex Hunter-Gatherers to Early Urban Societies (c. 16,000-300 BC)*. Cambridge: Cambridge University Press.
- Bordreuil, Pierre. 1981. "Production-pouvoir-parenté dans le royaume d'Ougarit (14ème-13ème s. av. J.C. environs)." In *Production, pouvoir et parenté dans le monde méditerranéen de Sumer à nos jours: Actes du colloque / organisé par l'E.R.A. 357, CNRS/EHES, Paris, décembre 1976*, edited by Claude-H. Breteau, 117-131. Paris: Geuthner.
- Bordreuil, Pierre, and Dennis Pardee. 1989. *La Trouvaille Épigraphique de l'Ougarit 1: Concordance*. Ras Shamra-Ougarit 5, no. 1. Paris: Éditions Recherche sur les Civilisations.
- Bordreuil, Pierre, and Dennis Pardee. 2009. *Manual of Ugaritic*. Linguistic Studies in Ancient West Semitic 3. Winona Lake, IN: Eisenbrauns.
- Borger, Rykle. 2004. *Mesopotamisches Zeichenlexikon*. AOAT 305. Münster: Ugarit-Verlag.
- Dirksen, Dieter, and Gert von Bally, eds. 1997. *Optical Technologies in the Humanities: Selected Contributions to the International Conference on New Technologies in the Humanities and Fourth International Conference on Optics within Life Sciences OWLS IV Münster, Germany, 9-13 July 1996*. Series of the International Society on Optics within Life Sciences 4. Berlin: Springer.

- Eisenstadt, Shmuel N., and Luis Roniger. 1984. *Patrons, Clients, and Friends: Interpersonal Relations and the Structure of Trust in Society*. Themes in the Social Sciences. Cambridge: Cambridge University Press.
- Matoïan, Valérie. 2008. *Le Mobilier Du Palais Royal d'Ougarit*. Ras Shamra-Ougarit 17. Lyon: Maison de l'Orient et de la Méditerranée.
- Schloen, David, and Sandra Schloen. 2014. "Beyond Gutenberg: Transcending the Document Paradigm in Digital Humanities." *DHQ* 8 (4). <<http://www.digitalhumanities.org/dhq/vol/8/4/000196/000196.html>>.
- Soldt, Wilfred H. van. 2005. *The Topography of the City-State of Ugarit*. AOAT 324. Münster: Ugarit-Verlag.
- Thuraisingham, Bhavani. 2002. *XML Databases and the Semantic Web*. Boca Raton, FL: CRC Press.
- Woods, Christopher, Geoff Emberling, and Emily Teeter, eds. 2010. *Visible Language: Inventions of Writing in the Ancient Middle East and Beyond*. Oriental Institute Museum Publications 32. Chicago: The Oriental Institute of the University of Chicago.
- Yon, Marguerite. 2006. *The Royal City of Ugarit on the Tell of Ras Shamra*. Winona Lake, IN: Eisenbrauns.
- Yon, Marguerite, and Daniel Arnaud. 2001. *Études Ougaritiques*. Ras Shamra-Ougarit 14. Paris: Éditions Recherche sur les Civilisations.

# Publishing Sumerian Literature on the Semantic Web

*Terhi Nurmikko-Fuller*

## Introduction

In the three decades since the invention of the World Wide Web (henceforth, “the web”), archaeologists, museums, and experts in the study of ancient languages have published their knowledge regarding the archaeological, curatorial, and philological material of the ancient world online. For the most part, it has been in the form of text and images intended for humans, both experts and interested laypersons, to read. Technological developments, however, now enable us to publish data online in machine-readable formats, where it is as accessible to software (a program or application which runs on a computer, as opposed to the hardware, which is the physical machine itself) as it is to a human being. It is in those formats that a family of technologies around the concept of the Semantic Web (sw) comes into play. These tools can be used to make information accessible in ways that allows machines to “understand” the meaning of the content they encounter.

The pragmatic implications of publishing datasets as Linked Data (LD) involve the social and academic considerations of trust, authority, data ownership, and copyright. The technical practicalities of data storage, server management, and user-interface generation represent another set of issues. From the perspective of capturing and representing expert knowledge, the development of the appropriate structural frameworks (known as “ontologies”)—through both the evaluation of existing models and the creation of new ones—plays a crucial role. In this chapter, three existing ontologies—the CIDOC CRM,<sup>1</sup> FRBROO,<sup>2</sup> and OntoMedia (OM)<sup>3</sup>—are evaluated for their suitability for adequately representing and capturing data made publicly available by an existing web resource, the Electronic Text Corpus of Sumerian Literature

1 Crofts 2008, <<http://www.cidoc-crm.org/get-last-official-release>> (accessed May 9, 2017).

2 Bekiari et al. 2015, <[https://www.ifla.org/files/assets/cataloguing/frbr/frbroo\\_v2.2.pdf](https://www.ifla.org/files/assets/cataloguing/frbr/frbroo_v2.2.pdf)> (accessed May 9, 2017).

3 Jewell et al. 2005, <<https://eprints.soton.ac.uk/261024/>> (accessed May 7, 2017).

(ETCSL).<sup>4</sup> The applied perspective is that of critical evaluation in the context of the inter- and multidisciplinary digital humanities, with equal consideration given to the opportunities and challenges of the data and the technical implementations and solutions.

### On Data, Information, and Knowledge

In order to understand the motivations behind the adoption of sw technologies and their unique advantages, we need to differentiate between three key concepts: data, information, and knowledge. The first, sometimes described in its un-analyzed state as being “raw,” consists of values and both quantitative and qualitative variables. We can easily recognize a spreadsheet as data: there is a column of numbers and, adjacent to it, a column with corresponding unit descriptions. The ancient Near East provides examples of clay tablets, which closely resemble modern spreadsheets, created almost five millennia before the invention of Microsoft Excel. We can use these to illustrate the differences between data, information, and knowledge: data is a list that stipulates that there are things that are sheep, and that there are, for example, five of them. The more data there is, the more information we can gain from it: if the list also contains goats, we can see that the owner has a mixed herd of animals of the subfamily of *Caprinae*. Knowledge is often generated as a result of much available, diverse information. The knowledge generated here would be that such mixed herds were and continue to be a popular approach to keeping livestock, due to their similar needs and non-competitive grazing habits (low and high, respectively). There are some clear caveats, too: the same data can be interpreted in many different ways; information is affected by bias and the limitations of the data; knowledge is often tacit and thus difficult to precisely pinpoint or define. Without further data about the animals (their sale price, for example) and additional information about the owner, our knowledge of the wider context remains incomplete. Was this a poor man who lived in a society in which sheep were more expensive than goats, and who could afford more animals if he had a mixed herd? Or was this a well-off individual wanting to display his wealth in a society that valued goats?

---

4 <<http://etcsl.orinst.ox.ac.uk/>> (accessed May 10, 2017).

## From Lovelace to Berners-Lee

Traditionally, computers—both the electronic machines we know today and their mechanical predecessors—have been well suited to creating and carrying out calculations on data. Several technological revolutions with far-reaching social ramifications have occurred since Ada Lovelace, the first programmer, wrote an algorithm for Charles Babbage's Analytical Engine in 1842. Not the least of these are the realization and continuing growth of the internet (a global system of networks of interconnected computers) since the 1960s and the invention of the web (a hypermedia system of pages, sites, and other types of resources, accessed and connected through the use of hyperlinks) by Sir Tim Berners-Lee at the end of the 1980s (and its exponential growth and worldwide adoption since the early 1990s). Although its current manifestation is one of interconnected pages, the three technologies that lie at the heart of the web—HTML, HTTP, and URIS, all of which are discussed in greater detail below—are also at the core of LD. This information publication paradigm is a practical solution for bringing forth a sw, where connections occur at the level of data entities, not sites or pages.

## Acronyms of Web Architecture

In the course of our daily browsing, we encounter a series of platforms, applications, projects, resources, and content. Behind the scenes, a number of technologies are at play.

The web currently consists of over 4 billion pages, all connected internally (within the page and a site) and externally (to other pages on other sites) through hyperlinks. Often visualized as blue font and underlined text, hyperlinks allow users to move between websites. The technology enabling this interconnectedness is the HyperText Transfer Protocol (HTTP); the acronym is familiar to us from the address bar displayed at the top of our browsers, starting with “http://”. These clusters of letters and characters form unique identifiers for each and every page on the web. They are known as Universal Resource Identifiers (URIS), and although other types of identifiers are possible, the ones we see when browsing are HTTP URIS. So common are the “http://” and “www.” prefixes that, when we discuss these resources, both are taken to be givens and thus are completely omitted; we refer to “Google,” for example, rather than “http://www.google.com.” On a conceptual level, these URIS do not differ from any other unique identifier, such as an ISBN for a book, a national identity

number for a person, or an object identifier in a museum, but they are unique on the scale of the entire web.

For the most part, HTTP URIs point to specific pages; in this case, they are largely URLs, or Universal Resource Locators. The majority of the content of these pages is text, encoded using the HyperText Markup Language (HTML) to display in desired sizes, colors, and fonts. HTML does not capture any meaning, and, as such, is not strictly speaking fundamental to sw technologies—unlike HTTP URIs, which are intrinsically necessary. On the *Semantic Web*, rather than pointing to web pages, HTTP URIs serve as identifiers for specific data entities and information published on a page, a site, or in a resource. When utilized consistently and documented appropriately, URIs can and are used by large numbers of data publishers (people and institutions who make data available online) to show that completely disparate and unconnected pages contain information about the same entity or resource. As they refer to the same “thing,” which can occur on many different pages, these URIs are identifiers for resources, not for locations. They are thus URIs and not URLs. They are fundamental in the move from a “Web of Documents” to a “Web of Data.”<sup>5</sup>

HTML is used for formatting and the visual appearance of the content of a web page, but it does not capture meaning. Enter the Extensible Markup Language (XML) and the tags, which appear in the code as angled brackets that allow us to encode the content as machine-readable. Agreed-upon standards such as the eponymous one maintained by the Text Encoding Initiative (TEI-XML) ensure that the tags used in the encoding of texts are systematically adhered to, and that different projects and sites can be shown to share relevant information. Consider two or more texts from independent projects that have been tagged to contain instances of people. If, for one, the utilized tag is `<person>`, but, for the other, it is `<human>`, an XML-based application would not necessarily consider them as containing information about the same kind of entity, unless such a relationship is explicitly defined in a relevant schema.

Machine comprehension can be achieved through RDF-XML, a syntax that can be used to express Resource Description Framework (RDF) in an XML document. RDF is an abstract data model with roots in metadata modelling, and it is used to represent information, knowledge, and data entities in a graph structure. This approach captures the relationships between entities by representing these entities and the connections between them using HTTP URIs. A more in-depth description of this technology follows later in this chapter, but it is RDF that enables the sw by allowing us to represent not just data, but, also, information and knowledge, in machine-understandable, as well as

---

5 “Semantic Web,” w3c, <http://www.w3.org/standards/semanticweb/> (accessed May 9, 2017).



machine-readable, ways. The use of RDF and HTTP URIs to publish information online is the fundamental approach at the heart of LD.<sup>6</sup>

RDF can be represented in different but equally valid serialization formats. Of these, the most commonly occurring are Turtle (.TTL), JSON-LD, and RDF/XML, each of which has its own proponents based on the needs, skills, and personal preferences of the people using them.

### Linked Ancient World Data Online

The success of online projects as diverse as Nomisma.org, which has a focus on numismatics, a term applied to the study of coins and currency, and the Pleiades.stoa.org gazetteer, an index or dictionary of historic places and spaces, illustrates the suitability of ancient-world data as the focal point for SW technologies.

LD projects such as these utilize W3C-endorsed technologies such as RDF to represent and capture information and SPARQL (SPARQL Protocol and RDF Query Language)<sup>7</sup> to access, retrieve, and manage it. The rich material from the ancient Near East is no different and can contribute to the development of LD through the promotion of shared vocabularies or schemas and exchange protocols.<sup>8</sup> Assyriological data can be used to evaluate and test the robustness and flexibility of existing models and structural frameworks (ontologies). As a discipline, Assyriology would benefit from improved interconnectedness of information, from enriching external sources, improved discoverability, and, ultimately, the inference of new, implicit knowledge from explicitly declared facts.

### Data

A multitude of domains, ranging from archaeology to curatorial and philological specializations regarding different aspects of the cultures of the ancient Near East, are brought together under the umbrella of Assyriological scholarship. Conservatively, the scope of the discipline is limited to the presence of objects carrying texts written in the distinctive cuneiform script. Even then,

6 For an example of the concrete use of RDF, see in this volume, Matskevich and Sharon, 46–47.

7 W3C is an acronym referring to the World Wide Web Consortium, the international standards organization for the World Wide Web, founded and chaired by Sir Tim Berners-Lee.

8 Protocol: a set of rules for the exchange or transmission of data between machines.

exemplars spanning the three millennia BCE can easily be referenced, provenanced to sites separated by time and space, and containing inscriptions in a number of unrelated ancient languages. The Semitic Akkadian, with two distinct dialects, Babylonian and Assyrian, and the linguistic isolate Sumerian, appear as a general rule; there are also the Indo-European Hittite, Hurrian, and Elamite, as well as other smaller language groups. Major professional entities, such as the International Association of Assyriologists,<sup>9</sup> are similarly polyglot, accepting submissions in English, French, and German. At the digital periphery of the discipline, programming languages and diverse technologies are added to the mix. Geographically, the Mesopotamian plateau constitutes modern-day Iraq as well as much of the surrounding area. Specialties by modern scholars and major research threads within the community are frequently focused on specific temporal, spatial, and socio-cultural subsets or niches within this wider context.

The biographies of Assyriological objects are complex. Historically, collaborative excavations have led to the division of material culture from a single archaeological site into various collections. Charles Leonard Woolley's excavations at the Royal Cemetery of Ur,<sup>10</sup> for example, saw material allocated and distributed between funding institutions, and the material is currently shared between the University of Pennsylvania Museum of Archaeology and Anthropology in Philadelphia, the British Museum in London, the National Museum of Iraq, and a number of other sites. Limited access to data that is physically distant or stored in closed-off information siloes or entirely off-line collections management systems complicates the types of philological scholarship that can be based on the analysis of ancient texts. The online publication of object metadata and high-resolution images accompanied by the translations and transliterations of the texts can help bridge collections and contribute to solutions to overcome the challenges of access and distance.<sup>11</sup> sw technologies have the potential to link heterogeneous datasets and to open up Assyriology as a discipline to enriching external data streams. Much work remains to be done until the full potential of the "Web of Data" can be harnessed and used to answer new and diverse scholarly questions, but the suitability of existing technologies for supporting this research is clear.

---

9 <https://iaassyriology.com/> (accessed May 10, 2017).

10 Woolley 1934.

11 Transliterations are conversions of the cuneiform script into the Latin alphabet without the translation of the ancient language to a modern one.

## On Composite Texts

This chapter examines the suitability of existing technologies for the representation of information and narrative content of ancient pieces of narrative fiction, published as composite texts by the ETCSL. Consisting of parts of the text from different surviving physical tablets, these manifestations of the inscriptions represent the result of “an attempt, based on certain theoretical presuppositions, to reconstruct a hypothetical original version of a composition, when none of the original sources of the text have survived.”<sup>12</sup> Composites may represent the content of a single instance of a tablet of any degree of preservation or completeness, or of any number of witness tablets, numbering in some cases in the hundreds.<sup>13</sup> The process of creating these digitally published amalgamations was described by Jeremy Black and Gábor Zólyomi as working with ‘jigsaw pieces, which may not overlap at all.’<sup>14</sup> Each of the 400 or so inscriptions is a new entity and the product of a multi-stage process of interpretation. Where corrections or changes to the text have happened in the hard-coded HTML alone, the translation or transliteration in question can be described as being “born-digital”: this text, in this specific form, has no analogue or tangible version in existence. For a humanities scholar, the distinction is important, as it affects the critical framework in which the text is to be understood. From the perspective of a digital humanist, or a working ontologist,<sup>15</sup> it is essential to identify and separate the composite text at a technical level from other manifestations. At the same time, it is crucial to capture and represent corresponding overlaps, original sources, and other inherent relationships between the witness tables—and their digital amalgam—in a way that makes those differences explicit to a specialist piece of software known as a “reasoner.”

## Neutrality of Technology

Three decades prior to the publication of this book, Melvin Kranzberg published his article outlining the relationships between the social and the technological.<sup>16</sup> The first of Kranzberg’s “laws,” where a law is a collection of truisms, describes technology as “neither good, nor bad; nor is it neutral.” Socio-politi-

---

<sup>12</sup> Delnero 2012.

<sup>13</sup> Delnero 2012, 1; Robson 2013.

<sup>14</sup> Black and Zólyomi 2007.

<sup>15</sup> Allemang and Hendler 2011.

<sup>16</sup> Kranzberg 1986, 544–560.

cal, economic, cultural, and intellectual aspects—all things human—play a role within the design, development, implementation, evaluation, alteration, and use of technology. For LD, and in particular in the evaluation and creation of ontologies, which by their very function capture human understanding of fact and human interpretation of the truth, the personal and idiosyncratic biases of those designing and using these structures must be acknowledged. It is similarly so with the heterogeneous and incomplete datasets we have regarding the ancient world, where gaps can only be bridged by the informed deductions of modern scholars. Investigative agendas utilizing sw technologies should at all times maintain a hyper-awareness of the limitations imposed by the data, appreciate the likelihood of differences between modern understanding and ancient thought and perspectives, and endeavor to explicitly capture tacit information. Extensive documentation of each aspect and stage of the project workflow as well as the datasets, including version control, should sit at the core of each research process in order to ensure accountability, transparency, and reproducibility.

### **On the Importance of Words**

Disambiguation and clarity of communication are essential to the interdisciplinary research of the digital humanities. Even the simplest of terminologies can be misinterpreted. Consider the term “tablet”: it has clear and unambiguous meaning to an Assyriologist, who envisions a clay object, probably one that fits comfortably in the hand and is inscribed with the distinctive wedge-like signs of the cuneiform script. The term has an equally clear meaning to those from the broader field of historical written media, but as a different type of writing implement, perhaps one based on wax (e.g., the Vindolanda tablets). The same word is just as unambiguously applied by all of us when referring to iPads or other similar mobile technologies. Context, then, is key. Other such examples can be readily identified in the data of the GLAM (Galleries, Libraries, Archives, and Museums) sector and the wider digital humanities: “record,” which could be understood as object metadata, digital metadata, or an object carrying a piece of audio, and “ontology” which could be understood either within the remit of the analysis of existence (philosophy), or within a structural framework for capturing information about entities and their relationships (computer science and digital information technologies, the sw). Other terms, such as “subject,” “predicate,” and “object,” have clear meanings to linguists, but those meanings have little in common with the semantics an ontologist would assign to them. Even the term “semantic” is understood by

linguists as relating to the human process of assigning meaning, but, for those working in Knowledge Representation (KR), the term refers specifically to machine-readability and beyond, into the domain of machine-understanding, accomplished not necessarily through the complexities of Artificial Intelligence,<sup>17</sup> but, rather, through the assignment of HTTP URIs.

The SW and LD as terms are frequently used erroneously as if they were interchangeable nouns referring to the same thing, much like the internet and the web. In fact, the SW refers to a thing, while LD is a process, namely one of data publication online in accordance with determined and agreed-upon criteria.<sup>18</sup> Within this terminology there are clear, unambiguously separated terms which ought to be used precisely, not synonymously: Linked Data, Linked Open Data, and Open Data: the first refers to information that has been captured and published using the W3C standards of RDF and HTTP URIs.<sup>19</sup> The term does not, however, signify that the data in question has been made openly available. RDF contained on a computer that is not connected to the internet is still Linked Data; RDF published online but only accessible to certain individuals is similarly not Open Data (OD). OD is data that has been made publicly available without access restrictions resulting from copyrights or paywalls. It often manifests as raw, or largely unanalyzed, datasets, made downloadable in formats such as CSV files. Linked Open Data (LOD) combines both these types: it is often based on OD that has been converted to RDF, and it is similarly available for consumption and reuse without restrictions or limitations. Since the term is more inclusive, this chapter uses Linked Data (LD) throughout.

A clear and distinct line needs to be drawn between semantic technologies and SW technologies: the former are understood to constitute a heterogeneous collection of tools and algorithms that can be used to bring structure to information. These include Natural Language Processing (NLP), data mining, Artificial Intelligence (AI),<sup>20</sup> and a number of other approaches and methodologies. SW technologies refers to a family of technologies and standards provided by the W3C for the purposes of describing and connecting information online. A by-no-means-exhaustive list of these technologies includes RDF, RDFS, OWL (RDF Schema and the Web Ontology Language; for ontologies as will be dis-

---

17 For further information, see in this volume, Ramazzotti, 62.

18 "Linked Data," W3C, <<https://www.w3.org/DesignIssues/LinkedData.html>> (accessed May 7, 2017).

19 For general but concrete examples related to the use of Linked Data, see in this volume, Matskevich and Sharon, 47.

20 For further information, see in this volume, Ramazzotti, 62n9.

cussed in detail momentarily),<sup>21</sup> and SPARQL. It is on these w3C standards that this chapter will focus.

## The Triple

The grand promise of LD has been one of harnessing all the intellectual, factual, and raw data anywhere online. Information amalgamated from all corners of the web, across disciplinary barriers, data siloes, and automated inference, would, at the stroke of a key, bring to light the types of implicit connections that would take human scholars generations to uncover. As a flexible data structure, RDF is both powerful and capable of capturing nuanced and domain-specific relationships between data entities. Information stored in this way can also be restructured, manipulated, and edited with comparative ease without the extensive restructuring that a relational database would require. RDF as a technology imposes few limitations on the type of information that can be captured and represented, whether at the level of entities (things, people, places, concepts, notions) or at the level of the relationships between them. All RDF data structures are based around a data cluster of three components, known as a “triple.”<sup>22</sup>

A triple consists of three parts: a *SUBJECT*, a *PREDICATE*, and an *OBJECT*. These are used to define entity types and directional relationships between them, both at the general (or schema) and specific (instance) level.

Let us consider the instance level first, since the schema level is captured using ontologies, which are described in greater detail in the section below. At an instance level, we would consider a specific cuneiform tablet. It will have an author, perhaps an ancient scribe named in the colophon.<sup>23</sup> A natural language utterance capturing that information might take the form of the sentence “Tablets are written by scribes.” Since not all scribes wrote all tablets, differentiation at the level of specific instances is necessary, and it is achieved through the use of personal names for scribes (e.g., Lukalla) and museum

21 “OWL 2 Web Ontology Language Document Overview (Second Edition),” w3C, <<http://www.w3.org/TR/owl2-overview/>> (accessed May 7, 2017).

22 See in this volume, a) applied to graph database, Matskevich and Sharon, 46–47, and b) for a different meaning, Pagé-Perron, 202–203, who uses triples for network graph.

23 A colophon is a brief statement on a text, such as a cuneiform tablet, that captures the equivalent of what could now be considered bibliographical metadata and can contain information such as the name of the authoring individual.

numbers for the tablets they wrote (e.g., Soo2932).<sup>24</sup> The same utterance now becomes more specific: “the specific tablet to which we now refer by its modern identifier Soo2932 was written by the named scribe Lukalla.”

Both granularities of information can be captured using RDF triples. In the case of the specific (instance-level) relationship between Lukalla and Soo2932, the `subject` could be the tablet, the `predicate` would be the notion of creation, and the `object` could be the scribe. This would result in a triple of

```
tabletSoo2932(SUBJECT)  was written by(PREDICATE)  Lukalla(OBJECT).
```

If we define the relationship between the tablet and the scribe as “wrote,” we can capture this relationship from the other direction, resulting in an equally true and accurate but slightly different triple:

```
Lukalla(SUBJECT)  wrote(PREDICATE)  tabletSoo2932(OBJECT).
```

In this example, the personal name and the museum number are expressed as strings of characters (letters and numbers), rather than as semantically meaningful identifiers. When this information is captured as RDF, each of the three parts is assigned a HTTP URI:<sup>25</sup>

```
<http://example.org/abcd> <http://example.
org/12345> <http://example.org/xyz>
```

HTTP URIs can be simplified in .TTL, making them easier for humans to read. This is accomplished using a prefix or a disclosed shorthand: if we define “http://example.org/” as `prefix:example`, the same triple can be expressed more succinctly (without any loss of meaning) as:

24 Tablet Soo2932 is also known as P249185 in CDLI (<[http://cdli.ucla.edu/search/search\\_results.php?SearchMode=Text&ObjectID=P249089](http://cdli.ucla.edu/search/search_results.php?SearchMode=Text&ObjectID=P249089)> [accessed May 10, 2017]).

25 This example may have been clearer if HTTP URIs with embedded semantics had been used, e.g., Lukalla had been identified as <http://example.org/scribe/Lukalla> and the tablet as <http://example.org/tablet/Soo2932>. Although these are easy for humans to read and understand, the current best practice in the LD and SW communities state that it is better to use URIs that do not have human-readable meanings. In this example, <http://example.org/abcde> refers to the scribe, <http://example.org/12345> to the relationship between scribe and tablet, and <http://example.org/xyz> to the tablet.

example:abcde    example:12345    example:xyzz

It is here that a knowledge graph begins to emerge: the object in one triple, when represented using an HTTP URI, can be the subject in another. This enables us to produce an interconnected graph of nodes (the entities) and arcs (the relationships). Unlike hierarchical data structures such as XML, RDF graphs can extend to direction at any point. The end of the sprawling network only comes at points in which the object of the triple is a string (or indeed anything but a HTTP URI), as no new triples can be added.

Navigating and querying the resulting knowledge graph necessitates understanding the entity and relationship types that have been used in capturing this information both on the instance and schema levels. We use information structures expressed in .TTL to do this.

## Ontologies

The sw is not a new concept. KR, one of its cornerstones, has been discussed in the context of the Artificial Intelligence community since at least the early 1980s.<sup>26</sup> Berners-Lee's original vision for the web was one of a hypertext system not only linking human-readable documents but also including machine-readability from the onset.<sup>27</sup> Yorick Wilks and Christopher Brewster described the sw as a "more powerful, more functional and more capable version of [the] document and language-centric Web."<sup>28</sup> In recent years, the sw has increasingly manifested through the adoption of the LD publication paradigm, with the more complex processes of KR relying on complex information structures (ontologies).

Ontologies are formalized structures that form parts of expert systems. They constitute the underlying data infrastructure for projects, tools, and datasets published as LD. Described as "crucial" to the realization of the sw,<sup>29</sup> they are frequently defined through Thomas R. Gruber's description as "an explicit specification of a conceptualization."<sup>30</sup> Ontologies are used to define the types of entities that occur in the dataset (classes) and the possible relationships

---

26    Levesque 1984.

27    Berners-Lee 2000.

28    Wilks and Brewster 2006.

29    Wilks and Brewster, 2006.

30    Gruber 1993, 199.



between those entities (properties). The classes of an ontology define the type of entity a `subject` or `object` is, and the properties describe `predicates`, including the directionality of the relationship it captures. The class from which the property runs is the `domain`; the class to which it runs is the `range`.<sup>31</sup> While many of the relationships between entities are explicitly declared, one of the main advantages of sw technologies, and the publication of data in machine-readable formats, is automated reasoning (the finding of implicit connections between explicitly declared facts by software agents).

There are two main schools of thought regarding ontology design:<sup>32</sup> the Newtonian reductionist model and the Leibnizian model, which concerned with the capture of the nuanced complexities of experience. The former allows for a greater degree of control and administrative clarity, while the fuzziness of the latter can make it easier to apply. Newtonian models depend on a degree of omniscience over a domain; Leibnizian models assert that true and absolute objectivity is in all pragmatic terms unachievable. Classification of ontologies is frequently limited to two broad types: domain-specific or upper-level. The former focus on a given, defined, and specific area of information. The latter are more generic and have applicability across a large number of disciplines and areas of expertise.

One of the fundamental cornerstones of the successful application of LD is the ability to share information and knowledge across disparate datasets. This is accomplished through several methods: adherence to w3C recommendations (RDF, HTTP URIs) and the publication of data according to the Five Star standard are examples of the adoption and use of agreed-upon standards to facilitate this information exchange.<sup>33</sup> The reuse of existing ontologies to point to shared entity types is another significant approach. While it is possible to define a new ontological structure for each new project or resource, reusing existing ones is a simple, relatively straightforward method that is highly likely to effectively and efficiently facilitate links between datasets. Reuse is dependent on awareness of an existing model, and the latter's adequate, often necessarily extensive, documentation enables the data publisher to be sure that they are using the ontology appropriately. Evaluations of existing ontologies are

31 For examples of concrete applications of ontologies in the philological field, see in this volume, Bigot Juloux, in particular 165–181; Prosser, 320. For ontology in data sharing, see also in this volume, Matskevich and Sharon, 47.

32 Brewster and O'Hara 2004.

33 "Linked Data Glossary," w3C, <<https://www.w3.org/TR/ld-glossary/>> (accessed May 7, 2017).

heavily dependent on the availability and completeness of relevant documentation and thus can be time-consuming. Yet they are a significant part of the interlinking process for LD.

Assertion of equivalence between ontological class and dataset entity may seem relatively straightforward when considering things that have a degree of tangibility or generic familiarity, such as person, place, or event. A closer examination highlights a number of complicating uncertainties, such as our frequent inability to point to very specific start- or endpoints for events (those moments when something begins, or when an event draws to a conclusion), or the challenges created by spatio-temporal changes to the geographical areas of given locations, changed as territory is won and lost. Disambiguation between individuals may also be complicated by naming conventions and inherited professions, resulting in a number of persons from the same location and time's sharing both name and societal role. Incomplete data exacerbates the problem. Consideration and evaluation of existing ontologies ought to form a significant part of any research project examining and using LD.

### Ontologies for Sumerian Data

The Assyriological community has yet to extensively embrace SW technologies. Although other existing models will be shown to be relevant and appropriate for the representation of our data, at the time of this printing, only two ontologies specific to Sumerological data are known (with a third included in a PhD dissertation).<sup>34</sup>

Wojciech Jaworski describes an application-oriented system focused on economic tables of the Ur III period.<sup>35</sup> It includes an ontology representing a selected branch of economic activities, with translations of texts into a meaning representation language through a semantic grammar. This approach allows for the representation of ambiguities caused by the limitations of the document simplicity, any incompleteness of current understanding of the Sumerian language, and omissions in the texts caused by damage to the original primary sources.

---

<sup>34</sup> Nurmikko-Fuller 2015.

<sup>35</sup> Jaworski 2008.

Sumerian was used by Sergio Alivernini as an example for the representation of natural language grammars using ontologies.<sup>36</sup> Described as “an experiment,”<sup>37</sup> it consists of two parts: the T-BOX (terminological) and A-BOX (assertions), where the former is Sumerian grammar, and the latter is the content of ancient inscriptions. Of these, the latter utilizes foundation-brick inscriptions from the reign of Ur-Nammu (c. 2000 BCE), but the source material for the grammatical features is not declared.

Although focused on data that is socio-culturally, linguistically, and temporally complementary, these models are insufficient or unsuitable for the representation of the content of the literary inscriptions as published on the ETCSL. Before we evaluate the suitability of ontologies created from the perspective of other domains, the niche of Sumerian literature deserves definition and description.

### Literary Sumerology

Literary Sumerology (the study of literature written in Sumerian) is one of the most recent areas of expertise within the sphere of Assyriology, having emerged only in the 1950s.<sup>38</sup> The term “Sumerian” is here to be understood as an exclusively linguistic label, referring to the language of the compositions themselves, rather than being indicative of the ethnic or socio-cultural identity of the authoring scribe or the patron commissioning the piece. Data is not limited to sources with a proven provenance in Sumerian periods for two reasons: First, literary pieces unearthed from third-millennium BCE contexts at geographically diverse sites (Ebla and Mari in the north and Girsu, Abu Salabiḥ, Nippur, and Adab in the south) suggest these compositions were conceived of even earlier;<sup>39</sup> second, while primary sources for Sumerian literature originate from a range of temporal, geographical, and cultural settings,<sup>40</sup> many are dated to the Old Babylonian (OB) period (c. 20th-16th centuries BCE). Sumerian is absent from administrative and legal documents, as well as letters, from around 1730 BCE onwards,<sup>41</sup> supporting the theory that at this time it no longer sur-

36 Alivernini, D'Agostino, and Romano 2006, <[http://www.epistemica.com/docs/Ur\\_Namma.pdf](http://www.epistemica.com/docs/Ur_Namma.pdf)> (accessed May 12, 2017).

37 Alivernini, D'Agostino, and Romano, 2006, 3.

38 Robson 2013.

39 Taylor 2013.

40 Goodnick-Westenholz 2013.

41 Black and Zólyomi 2007.

vived as a spoken language in any of the urban centers of the Mesopotamian plateau. Its use in the vernacular registers of personal correspondences had ceased some two centuries earlier.<sup>42</sup> Scribes who most likely spoke Akkadian as their mother tongue carried on the tradition of Sumerian literature via copies of earlier compositions (e.g., Instructions of Šuruppak) as well as new compositions and those created through extensive modification of known material such as *Lugalbanda*.<sup>43</sup> Separating instances of deliberate variation (a reflection of a particular scribal tradition, for example) from those of scribal error can be difficult,<sup>44</sup> but it is possible that the practice of creating Sumerian literary compositions outspan, in both directions, the third millennium BCE.

### The Electronic Text Corpus of Sumerian Literature (ETCSL)

The ETCSL is an online resource that provides access to the transliterations (in both Unicode and ASCII) and translations of some 400 literary compositions, all dated to the late third or early second millennium BCE. The name encapsulates the form and function of this bi-part, web-based resource, which Jarle Ebeling describes as a “diachronic, transliterated, annotated corpus of Sumerian literature, and a Sumerian-English parallel corpus.”<sup>45</sup> Internally, it has a seven-part thematic structure that mirrors the original non-digital and largely unpublished corpus created by Miguel Civil.<sup>46</sup> For the Sumerian, there are transliterations rendered using the Latin alphabet and modern philological conventions (with no accompanying cuneiform, either Unicode or otherwise) on a line-by-line basis. For the English translation, there are paragraphs of prose.

The ETCSL provides external links to other related and complementary projects. These include other online corpora, resources for museological and palaeographical data, and a dictionary. These projects include the Diachronic Corpus of Sumerian Literature,<sup>47</sup> Database of Neo-Sumerian Texts,<sup>48</sup> the Cuneiform Digital Palaeography Project,<sup>49</sup> the Cuneiform Digital Library Initia-

42 Black and Zólyomi 2007.

43 Taylor 2013.

44 Black and Zólyomi 2007.

45 Ebeling 2007.

46 Ebeling and Cunningham 2007.

47 <<https://digital.humanities.ox.ac.uk/project/diachronic-corpus-sumerian-literature>> (accessed May 7, 2017).

48 <<http://bdtns.filol.csic.es/>> (accessed May 7, 2017).

49 <<http://www.etana.org/node/6969>> (accessed May 7, 2017).

tive (CDLI),<sup>50</sup> and the Pennsylvania Sumerian Dictionary (ePSD).<sup>51</sup> Many of the links run in one direction: the Diachronic Corpus links to the transliterations published on the ETCSL, but cannot be accessed via it. The ePSD links externally, providing a list of texts and a short passage in which a specific lexical item occurs, but the hyperlink is not bidirectional. The ETCSL record is a pop-up window and can be compared with the original search results from the ePSD.

Although not actively developed since 2006, the ETCSL is a cited resource and is itself the focus of research.<sup>52</sup> It was an innovative project in the realm of study of Sumerian literature in that it offered free and unrestricted online access to the data, with the aim of interlinking with external resources. The live site was created using HTML and JavaScript (a programming language for HTML and the web) only, with the lemmatisation hard-coded into the latter, but the content is exclusively output as customized TEI-XML P4, available from the Oxford Text Archive.<sup>53</sup>

### Research Questions for the ETCSL

Ontological representation and the publication of data as RDF allows for a new, more flexible and complex type of research question. These enabled avenues of investigation can span beyond the expected searches facilitated by most search fields on online projects and resources. Traditional investigative agendas are frequently limited to the defining characteristics of geographical location, time, and an area of expertise. Capturing the narrative thread of these literary compositions will enable a new type of question: Which kings have claimed a divine or semi-divine character as their brother? What are the demographics and dynamics of giving and receiving advice? Are men more likely to advise or be advised by women? Is advice predominantly given by seniors to juniors? What are the shared characteristics of persons associated with a birth legend? Adherence to LD will not accomplish data collection tasks that would be beyond a human scholar intellectually, but LD can speed them up, make

50 <<http://cdli.ucla.edu/>> (accessed May 10, 2017). For further information, see in this volume, Pagé-Perron, 198–200, and Eraslan 285.

51 <<http://psd.museum.upenn.edu/epsd1/index.html>> (accessed May 9, 2017).

52 Ebeling and Cunningham 2007; Delnero 2012; Crawford 2013.

53 <<https://ota.ox.ac.uk/>>; <<http://etcsl.orinst.ox.ac.uk/>> (both accessed May 7, 2017). Robson 2013; Nurmikko-Fuller 2014, <<http://dlib.nyu.edu/awdl/isaw/isaw-papers/7/nurmikko-fuller/>> (accessed May 9, 2017).

them more precise, and bring to light complementary compositions from the surrounding areas (such as the oral traditions of the Bedouin). The addition of data features published by the ETCSL for compositions could add layers of complexity: Are there objects types that are frequently associated with tablets containing a specified literary motif? What types of material culture are curated by professionals responsible for tablets that carry a specific narrative feature? Do tablets carrying specific storylines tend to be displayed in a particular interpretative context within a museum setting? On what other areas do scholars who have published on a given literary composition, or one containing a particular feature, tend to focus? These research agendas are possible not only through the richness of the transcriptions contained in ETCSL data, which open up the analysis to specific narrative motifs, but also through museological information regarding specific witness tablets and through bibliographical metadata regarding publications of translations.

### **Evaluating the Suitability of Existing OWL Ontologies to Represent ETCSL Data**

The three existing OWL ontologies with structures complementary to different aspects of ETCSL data that are evaluated here are the CIDOC CRM, FRBROO, and OM.

#### ***CIDOC CRM***

CIDOC CRM is a domain-specific, event-based ontology designed for the representation of cultural heritage data.<sup>54</sup> It has been an official International Organization for Standardization (ISO)<sup>55</sup> standard from 2006 and lists some 30 examples of extensions and compatible models that have been launched since. The purpose of the CIDOC CRM is to establish a common framework for the sharing of data between GLAM institutions, and to function as an example of best practices in the cultural heritage domain.

A number of other ontologies have already been incorporated into the CIDOC CRM. These include Dublin Core, SKOS, and FOAF, with additional available extensions and use cases further enabling the representation of archaeological

---

54 For a description of CIDOC CRM and its use in material heritage and field archaeology recording, see in this volume, Matskevich and Sharon, 45, 48–49.

55 International Organization for Standardization (ISO): the world's largest developer and publisher of international standards. It is an independent, voluntary organization, with representatives from each of the 162 member countries.

data and processes,<sup>56</sup> the provenance of digital artifacts, and bibliographies. All of these provide opportunities for linking to other data streams.

CIDOC CRM is a large and complex ontology consisting of 90 distinct entities and 149 property declarations.<sup>57</sup> Extensively documented, it makes the identification of suitable classes and properties relatively easy. The focus of the structure is to map cultural heritage information or details regarding the biography of a given object, but neither is directly available through ETCSL. The CIDOC CRM alone is insufficient for the representation of the project data, but it is crucial in its role linking FRBROO and OM, as well as in its enabling possible future data exchange with other digital heritage projects.

### Mapping ETCSL Data onto the CIDOC CRM

One future potential is clear: the CIDOC CRM allows for the differentiation between the text content (E33 Linguistic Object) and the physical item that carries a composition (E84 Information Carrier). They are linked through E73 Information Object (superclass of E33).<sup>58</sup> This is particularly useful when representing a composite text, an intangible construct, and any instance of E33, manifesting as a transliteration on the ETCSL site (itself a digital instance of E73). The transliteration and translation can each be mapped to have a language (Sumerian and English, respectively). Similarly, each witness tablet is an instance of E73, and each text carried on each physical tablet is a separate instance of E33. The use of E90 Symbolic Object (the superclass of E73)<sup>59</sup> makes it possible to represent the notion of the composite as a separate entity. The P106 is composed of—property maps the composite (an instance of E73) as an amalgamation of several other texts (also all instances of E73). This nesting of super- and subclasses is significant because subclasses inherit the properties of the superclasses.

The carrier in this case is a cuneiform tablet purposely made for the task of storing data in the form of written text (E84 Information Carrier is a nested subclass of E24 Physical Man-Made Thing, connected to E90 via the P128 carried by (is carried by)—property). The CIDOC CRM enables the mapping of the physical features of the physical items (E57 Material, E58

56 For information on Dublin Core, see <<http://dublincore.org/>> (accessed May 12, 2017). “SKOS Simple Knowledge Organization System,” w3c, <<https://www.w3.org/2004/02/skos/>> (accessed May 7, 2017). For further explanation, see in this volume, Matskevich and Sharon, 46. On FOAF, see “FOAF Vocabulary Specification 0.99” (*Xmles.com*, <<http://xmles.com/foaf/spec/>> [accessed May 7, 2017]).

57 Crofts et al 2008.

58 Crofts et al. 2008, 29.

59 Instances of E90 are clusters of characters (including writing) that have a recognizable structure.

Measurement Unit), and although presently that data is not directly available through the ETCSL, it could be found by identifying the corresponding record from the housing institution's online collections (Musée du Louvre, the British Museum).<sup>60</sup> This is an ideal example of the type of information-symbiosis that could be used to enrich the ETCSL and museum records, providing access to the transliteration and translation of the text content and helping to place objects in a wider literary context. With the addition of high-resolution photography from projects such as CDLI, the collections and the knowledge within the discipline could be brought together efficiently and accessed from a single entry point by anyone with access to the web.

### **FRBROO**

FRBROO is a formal ontology for the representation of bibliographic information. It has been designed to merge with the CIDOC CRM and to facilitate the integration of museum and library data. It is based on the Functional Requirements for Bibliographic Records (FRBR), which was originally designed as an entity-relationship model. FRBR was developed independently of CIDOC CRM, but, coincidentally, at approximately the same time (1991–1997). It was designed by a group appointed by the IFLA (International Federation of Library Associations and Institutions) and approved by the IFLA Cataloguing Section in 1997. In 2003, an international working group began to examine the potential of merging FRBR with the CIDOC CRM; the first draft of FRBROO was completed in 2006, with the official publication of version 1.0 at the end of 2009. It was approved and issued by January 2010.<sup>61</sup> Unlike the CIDOC CRM, FRBR models products, not processes. It is smaller and less complex, with 52 classes and 64 properties, compared to 90 classes and 149 properties in the CIDOC CRM. The process of merging the two existing structures necessitated changes in both, resulting in changes to the CIDOC CRM ISO standard.<sup>62</sup> The combination of FRBR and the CIDOC CRM brought about FRBROO—a new, object-oriented mapping. The process was one of extensive merging: FRBROO is referred to in 60 of the CIDOC CRM classes and in 55 of the properties.

The international working group for FRBROO cited the complementary and interlinked nature of bibliographic and cultural heritage data as the incentive for the project: “Libraries and museums are memory institutions—both strive to preserve cultural heritage objects, and information about such objects, and they often share the same users ... the boundary between them is often blurred

60 This assumes that these heritage institutions' required object catalogues and collections management data had been published in adherence to the Five Star LD criteria.

61 Bekiari et al. 2015.

62 Bekiari et al. 2015, 11.



... the cultural heritage objects preserved in both types of institutions were created in the same cultural context or period, sometimes by the same agents ... it seems therefore appropriate to build a common conceptualisation of the information gathered by the two types of organization.”<sup>63</sup> An example of the use of CIDOC CRM and FRBROO for the representation of data related to ancient material is the British Museum’s project for the digitization of Malcolm Moshier’s work on the ancient Egyptian Book of the Dead.<sup>64</sup>

#### Mapping ETCSL Data onto FRBROO

There are two types of bibliographies embedded into the ETCSL. The first is the metadata of the transliteration, including the title and the revision history. The second is the text-only page listing the items cited in the transliterations, including the author name, article title, year, name of the publication type, and name of the publisher. These six categories of information can be mapped onto the classes of FRBROO and connected to the aforementioned class of E73 Information Object: FRBROO class F2 Expression is a subclass of E73. F2 has two subclasses, which allow for the mapping of the composite text: F22 Self-contained Expression is a class for the composites as they appear on ETCSL (as a cohesive text), while each separate segment is an instance of F23 Expression Fragment, since they are sections of the composite and sections of the original witness tablets. The two classes are connected as subclasses of E2, and their relationships to E2 additionally include R5 has Component (is component of) and R14 incorporates (is incorporated in) for E22, and R15b is fragment of (has Fragment) for E23. The relationship of text to the object is via P128 carries (is carried by). It connects F4 Manifestation Singleton to F2. Modern scholars named in the bibliography can be mapped as instances of F10 Person (an equivalent of E21), so these classes provide an example of one of the many points where the two ontologies merge. Another example relevant here is F4 (a subclass of E24). FRBROO allows for the representation of both physical and electronic publishing.

In terms of content, FRBROO includes F38 Character, ideally suited for the capture of protagonists, antagonists, and other characters in Sumerian literature. The scope notes specify this class as being for “fictional and iconographic individuals ... appearing in works in a way relevant as subjects. Characters may be purely fictitious or based on real persons.”<sup>65</sup> Applicable to a wide spectrum

63 Bekiari et al. 2015, 11.

64 Oldman and Norton 2014, <<https://www.youtube.com/watch?v=cK54YIY-xZs>> (accessed May 12, 2017).

65 Bekiari et al. 2015.

of protagonists, it is equally valid for the unnamed characters (the slave-girl, the young scribe, the ox-driver) as it is for Gilgameš, the eponymous main character of the oldest known piece of epic literature,<sup>66</sup> a semi-divine being, and a powerful ruler.

For Gilgameš, there are three disjoint categories: the man, who may have been a genuine, historical person, and is treated in the context of the literature as such, and ought to be mapped as an instance of  $F10 \equiv E21$ ; the later-deified, <sup>d</sup>Gilgameš,<sup>67</sup> a myth, a fabricated construct serving an educational, cultic or socio-political purpose; and the character in a literary composition ( $F38$ ). This character is separate from the divine manifestation, as he exists solely in the fabula of the story and could, from the perspective of those who do not believe in the existence of a divine entity, be considered wholly fictitious. The deified <sup>d</sup>Gilgameš is disjoint from the man, as the latter has to transform into the former (they cannot coexist). Neither CIDOC CRM nor FRBROO has a class that truly captures this notion, although FRBROO does allow for the capture of the third manifestation of Gilgameš: the literary representation. It is possible to map the relationship between the protagonist and his historical counterpart by  $R57$  is based on (with  $E39$  Actor as range).

### *OntoMedia (OM)*

OM, like the CIDOC CRM, is event-based. It focuses on the representation of the narrative in multi-media and has been designed as linkable to the CIDOC CRM.<sup>68</sup> The aim of this ontology is to enable the human-like, vague questions we might use in conversation when trying to identify a given story.<sup>69</sup> OM, unlike CIDOC CRM or FRBROO, is not a widely utilized ontology, nor has it been awarded ISO (or equivalent) status; the first version was the product of a collaborative doctoral thesis, and, since then, the constituent .OWL files have been made (somewhat inconsistently) accessible online. The widespread adoption of OM is likely to be hindered by the lack of extensive, clear, up-to-date, and systematic documentation.

OM represents the narrative content of heterogeneous media. It is based largely on two interlinked topics: the literary genres of fantasy and science fiction and the fan-fiction associated with them, both manifesting as a number of genre-specific classes (Void-Travel, Pegasii, Unicorn and Faerie).

66 George 1999; 2010.

67 This addition of the lowercase “d” in superscript before the personal name denotes that the entity in question is a deity.

68 Jewell et al. 2005.

69 Lawrence 2008.

Some classes are indicative not only of genre, but, also, the works of a specific author (*Hobbit*).

Design decisions complicate the structure and result in the repetition of equivalent classes across sub-ontologies. Reliance on nested subclasses and the duplication of classes is illustrated by the *Trait*: it is simultaneously a parallel class and a superclass for over 50 other classes. Another example is *Person*≡*Person*≡*Person* in the *Being* sub-ontology, where one is *foaf:Person*,<sup>70</sup> and the other two are nested classes of OM repeated within separate sub-ontologies.

Protagonists are mapped as instances of *Character*, although this class does not allow for any differentiation between people, deities, sentient supernatural beings, demigods, or anthropomorphized creatures. Incorporation of the FOAF ontology enables the mapping of interpersonal relationships and physical characteristics through a series of additional classes. Regarding relationships, it is possible to differentiate between four distinct subclasses of *Family Bonds* (*Adopted*, *Blood*, *Foster*, and *Step*), as well as to map alliances (including a separate subclass *Friendship*). There are classes for *Deal*, *Enmity*, *Pledge*, and *Possession*. The *Agent* class has subclasses for *Group*, *Organisation*, and *Person*.

#### Mapping ETCSL Data onto OM

OM's *Character Trait* class contains many characteristics of individuals, including *Gender*, *Name*, *Stages Of Life*, *Stages Of Being*, and *State Of Consciousness*, as well as *Knowledge* and *Motivation*. Various subclasses of the *Detail* class map human physical features (*Eye Colour*, *Body Type*, *Hair Colour*), and enable the mapping of the descriptions of protagonists: "his angry brow...his bison eyes...his lapis lazuli beard...his elegant fingers..." (lines 70–75 in *Gilgameš and Aga*).<sup>71</sup> The *Attire* class can also be used to represent different parts of the arc of the story, such as the undressing of the eponymous protagonist in *Inanna's descent to the Netherworld*.<sup>72</sup>

Another tangible data entity within the content of the literary corpora is geographical location. The OM class *Space* has three distinct subclasses. Two of these are structurally simple, consisting of a small number of nested sub-

<sup>70</sup> That is to say, instances of this class are people, as defined by the FOAF ontology.

<sup>71</sup> "Gilgameš and Aga," *Electronic Text Corpus of Sumerian Literature*, <<http://etcsl.orinst.ox.ac.uk/section1/tr1811.htm>> (accessed May 10, 2017).

<sup>72</sup> "Inanna's Descent to the Netherworld," *Electronic Text Corpus of Sumerian Literature*, <<http://etcsl.orinst.ox.ac.uk/section1/tr141.htm>> (accessed May 10, 2017).

classes. AKT Abstract Space (relevant nested subclasses include Surface Space and Biological Surface Space) and AKT Enclosed Space (including Vessel, Portal, and Container) allow for the representation of, for example, the coracle that held the infant Sargon as he was sent down the river according to his eponymous *Birth Legend*, or the much larger version built by *Atrahasis*.<sup>73</sup> The extensive AKT Open Space contains almost 40 distinct classes and subclasses, most of which (with the exception, perhaps, of World, Galaxy, and Universe) are directly useful for the representation of place in these ancient literary compositions.

Although extensive, the Space sub-ontology does not include those architectural entities and features familiar to us from Sumerian narratives: the temple, the palace, the city wall, the canal, the dyke, and the reed bed. The closest class is Building, but since temples were often individually named, it does not seem infeasible to suggest that they were seen as a separate type of entity, disjoint from other, more mundane buildings, and that representing them as instances of this class would reflect a perspective radically differ from the perspectives of the ancient peoples.

The Core Expression class includes aspects with universal applicability. These include Abstract Item and Physical Item, as well as the classes of Timeline, Occurrence (with the subclass Event), Introduction, Gain, Loss, Transformation (with subclass Travel), Social, and Action. All encountered events in the Sumerian literary canon can be mapped using these classes, although the lack of granularity would result in many false positives when identifying stories that share some broad narrative motif. Query-based, automated differentiation between different narrative structures would necessarily be quite limited.

Although OM Profession class includes Military, challenges remain as to whether the differences between ancient (as described and recorded in the texts), modern (as per our own cultural perspectives), and hypothetical future (as expressed in the narratives of science fiction) warfare are sufficiently similar to justify the use of this class. The Bestiary and Zoology sub-ontologies allow for the specification of many animals, but they are not based on Linnaean taxonomy and rarely match those identified in the ETCSL.

Thirteen classes allow for the mapping of human and animal actions, movements of celestial bodies, and natural phenomena: Action, Being, Celestial, Environment, Eventprop, Gain, Group, Introduction, Loss, Social, Space, Transformation, and Travel. Action contains the parallel subclass system discussed in the context of the Trait class: here, Celestial

73 Finkel 2014.

and *Environmental* are subclasses of *Action*, as well as parallel to it. When parallel, they contain further subclasses. *Environmental Event* subclasses enable the mapping of natural phenomena that appear in the context of the Sumerian narrative, such as rising waters (Enki and Ninhursaga)<sup>74</sup> or growing of crops (Enki and the World Order).<sup>75</sup> *Celestial* is more complex due to the Mesopotamian perspective on the cosmos. At the most reductionist of levels, the Mesopotamian perspective could be described as viewing gods and goddesses as anthropomorphized celestial objects. Utu rises and sets; the sun is the god.

Several of the subclasses of the *Action* class can be used directly to represent the elements of a Sumerian story. The Victory of Utu-hegal has instances of *Battle* and *War*; the beating received by Birḫar-tura in Gilgameš and Aga populates *Corporal Punishment*. Other classes of *Events* can be shown to be equally relevant, and others as unnecessary (*Space Travel*).

For all the complexity of OM, its classes fall short of representing the potential richness of ETCSL data when they attempt to capture verbal exchanges. Various sorts of oration are a frequently utilized rhetoric device and contain additional explanations and information. Extensions to *Social*, including the subclasses of *Monologue*, *Dialogue*, *Polylogue*, *Advice*, *Warning*, *Lamentation*, *Curse*, *Prayer*, and *Rejoicing* would be necessary to adequately capture the richness of the rhetoric in these compositions.

## Conclusion

sw technologies facilitate the sharing of knowledge regarding identified entities, a concept familiar to Assyriologists, who already use non-LD online resources such as CDLI to access data collated from a large number of different databases into one cohesive unit and to query them via a single point of entry. LD differs from these existing models in that it is not limited to the content and structure of the selected databases. Relevant information, could potentially be identified anywhere online and incorporated into the processes of automated inference. Data published as RDF can be queried starting from any desired or arbitrary point in the knowledge graph without the need to adhere to a strict hierarchical structure. Both RDF and ontologies are as inherently flexible as they are extensible: it is possible to merge ontologies and to add RDF triples as

74 "Enki and Ninhursaga," *Electronic Text Corpus of Sumerian Literature*, <[#">http://etcsl.orinst.ox.ac.uk/cgi-bin/etcsl.cgi?text=t.1.1.1#](http://etcsl.orinst.ox.ac.uk/cgi-bin/etcsl.cgi?text=t.1.1.1)> (accessed May 10, 2017).

75 "Enki and the World Order," *Electronic Text Corpus of Sumerian Literature*, <<http://etcsl.orinst.ox.ac.uk/section1/tr13.htm>> (accessed May 10, 2017).

knowledge increases and new discoveries are made. Crucially, the addition of semantics via the ontology facilitates a degree of understanding of the content of a text by software agents as well as human readers. Data stored as RDF can be queried using another W3C standard, SPARQL, effectively creating searches that extend beyond keywords and Boolean operators.

The rich, heterogeneous, and at times incomplete information contained within the ETCSL and the wider corpora of Sumerian literature offers many challenges to these technologies. Three ontologies were evaluated in the context of this data: the CIDOC CRM, FRBROO, and OM. The first captures cultural heritage data, the second represents bibliographic data, and the third maps narrative content. Each of the three was assessed in turn, and they were all found to be in many ways adequate for the representation of this information but also to contain numerous superfluous classes. They are all predominantly designed to capture data-types that do not exist in the ETCSL.

The need for universality and the benefits of reuse are compelling arguments in favor of the use of these ontologies in further Assyriological investigations and of the large-scale publication of Sumerological data as RDF. The ontologies also facilitate linking between the considered data and myriad other contexts, both historical and contemporary.

Three conclusions are clear. First, the ontological representation of Sumerian literary narratives remains an interesting, rich, and engaging topic, with great potential for future development and investigation. Second, existing research has done little more than scratch the surface, and future scholarship is likely to bring to light further parallels and patterns that help us better understand the literature of ancient Sumer. Third, SW technologies are in a position to help support existing investigative paradigms across an array of topics, enabling new types of scholarship which can enrich—and be enriched by—the philological analysis of ancient Sumerian literary compositions.

## References

- Alivernini, Sergio, Franco D'Agostino, and Marco Romano. 2006. "Ur\_Namma: An OWL Ontology of a Sumerian Grammar." *Epistematica*. <[http://www.epistematica.com/docs/Ur\\_Namma.pdf](http://www.epistematica.com/docs/Ur_Namma.pdf)>.
- Allemang, Dean, and Jim Hendler. 2011. *Semantic Web for the Working Ontologist: Effective Modelling in RDFS and OWL*. Amsterdam: Elsevier.
- Bekiari, Chryssoula, Martin Doerr, Patrick Le Bœuf, and Pat Riva. 2015. "FRBR Object-orientated Definition and Mapping from FRBRER, FRAD and FRSAID (version 2.2)."

- International Working Group on FRBR and CIDOC CRM Harmonisation. <[https://www.ifla.org/files/assets/cataloguing/frbr/frbroo\\_v2.2.pdf](https://www.ifla.org/files/assets/cataloguing/frbr/frbroo_v2.2.pdf)>.
- Berners-Lee, Tim. 2000. *Weaving the Web: The Past, Present and Future of the World Wide Web by its Inventor*. London: Butler & Tanner.
- Black, Jeremy, and Gábor Zólyomi. 2007. 'Introduction to the Study of Sumerian.' In *Analysing Literary Sumerian Corpus-based Approaches*, edited by Jarle Ebeling and Graham Cunningham, 1–32. London: Equinox.
- Brewster, Christopher, and Kieron O'Hara. 2004. "Knowledge Representation with Ontologies: The Present and Future." *IEEE Intelligent Systems* 19 (1): 72–73.
- Crawford, Harriet. 2013. *The Sumerian World*. London: Routledge.
- Crofts, Nick, Martin Doerr, Tony Gill, Stephen Stead, and Matthew Stiff. 2008. *Definition of the CIDOC Conceptual Reference Model*. ICOM/CIDOC Documentation Standards Group. CIDOC CRM Special Interest Group 5. <<http://www.cidoc-crm.org/get-last-official-release>>.
- Delnero, Paul. 2012. *The Textual Criticism of Sumerian Literature*. JCSSS 3. Boston: ASOR.
- Ebeling, Jarle. 2007. "Corpora, Corpus Linguistics and the Electronic Text Corpus of Sumerian Literature." In *Analysing Literary Sumerian Corpus-based Approaches*, edited by Jarle Ebeling and Graham Cunningham, 33–50. London: Equinox.
- Ebeling, Jarle, and Graham Cunningham, eds. 2007. *Analysing Literary Sumerian Corpus-based Approaches*. London: Equinox.
- Finkel, Irving L. *The Ark Before Noah: Decoding the Story of the Flood*. London: Hodder & Stoughton, 2014.
- George, Andrew. 1999. *The Epic of Gilgamesh: The Babylonian Epic Poem and Other Texts in Akkadian and Sumerian*. London: Penguin.
- George, Andrew. 2010. *The Epic of Gilgamesh: The Babylonian Epic Poem and Other Texts in Akkadian and Sumerian*. London: The Folio Society.
- Goodnick Westenholz, Joan. 2013. "In the Service of the Gods: The Ministering Clergy." In *The Sumerian World*, edited by Harriet Crawford, 246–276. London: Routledge.
- Gruber, Thomas R. 1993. "Translation Approach to Portable Ontology Specifications." *Knowledge Acquisition* 5 (2): 199–220.
- Jaworski, Wojciech. 2008. "Contents Modelling of Neo-Sumerian Ur III Economic Text Corpus." In *Coling 2008: Proceedings of the 22nd International Conference on Computational Linguistics, 24 August 2008, Manchester, UK*, Vol. 1, 369–376. Stroudsburg, PA: Association for Computational Linguistics.
- Jewell, Michael O., Faith K. Lawrence, Mischa M. Tuffield, Adam Prugel-Bennett, David E. Millard, Mark S. Nixon, m.c. schraefel, and Nigel R. Shadbolt. 2005. "OntoMedia: An Ontology for the Representation of Heterogeneous Media." <<https://eprints.soton.ac.uk/261024/>>.
- Kranzberg, Melvin. 1986. "Technology and History: 'Kranzberg's Laws.'" *Technology and Culture* 27 (3): 544–560.

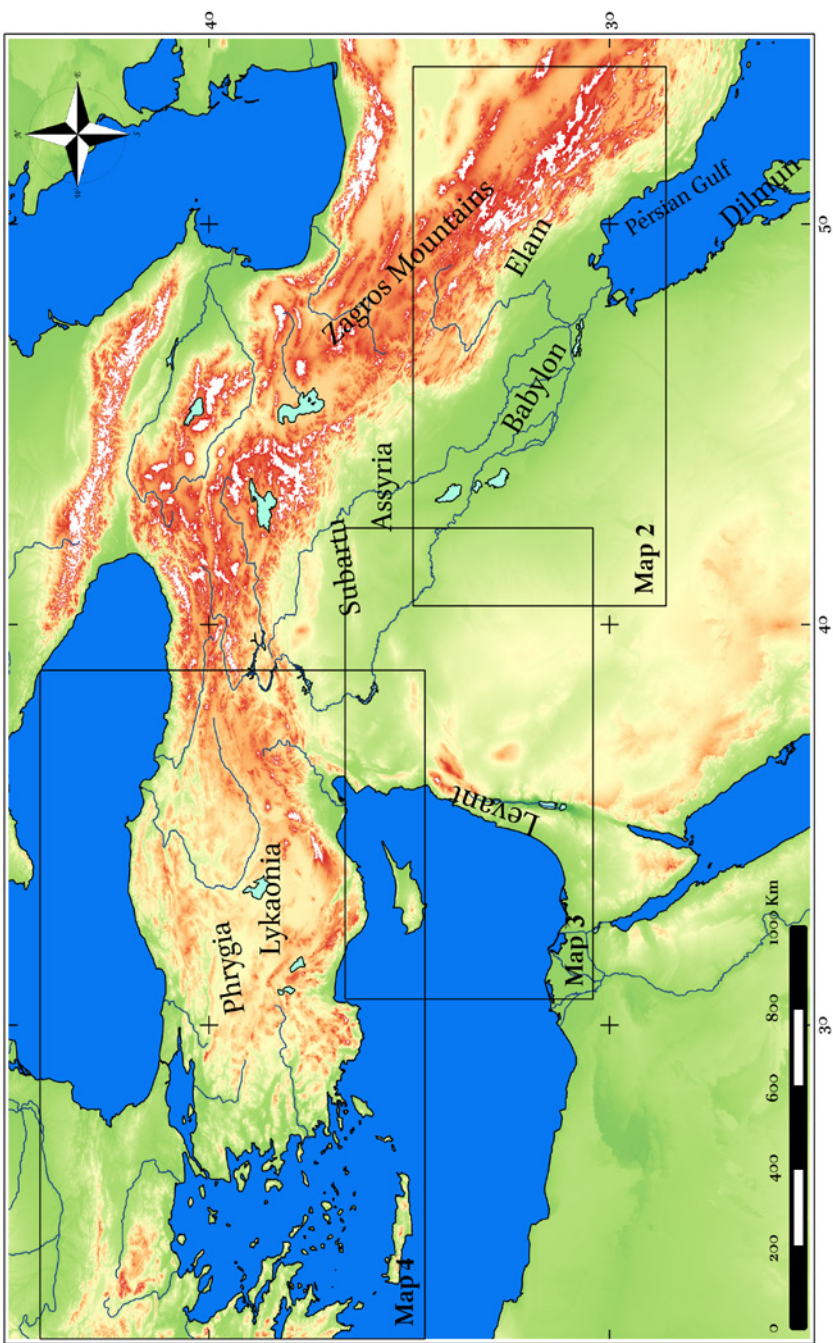


- Lawrence, Faith K. 2008. "The Web of Community Trust Amateur Fiction Online: A Case Study in Community Focused Design for the Semantic Web." PhD diss., University of Southampton.
- Levesque, Hector J. 1984. "Foundations of a Functional Approach to Knowledge Representation." *Artificial Intelligence* 23: 155–212.
- Nurmikko-Fuller, Terhi. 2014. "Assessing the Suitability of Existing OWL Ontologies for the Representation of Narrative Structures in Sumerian Literature." *Current Practice in Linked Open Data for the Ancient World, ISAW Papers*, 7, edited by Thomas Elliott, Sebastian Heath, and John Muccigrosso. <<http://dlib.nyu.edu/awdl/isaw/isaw-papers/7/nurmikko-fuller/>>.
- Nurmikko-Fuller, Terhi. 2015. "Telling Ancient Tales to Modern Machines: The Ontological Representation of Sumerian Literary Narratives." PhD diss., University of Southampton.
- Oldman, Dominic, and Barry Norton. 2014. "A New Approach to Digital Editions of Ancient Manuscripts using CIDOC CRM, FRBROO and RDFa." *The Digital Classicist*. <<https://www.youtube.com/watch?v=cK54YLY-xZs>>.
- Robson, Eleanor. 2013. "Lone Heroes or Collaborative Communities? On Sumerian Literature and its Practitioners." In *Ancient Egyptian Literature: Theory and Practice*, edited by Roland Enmarch, and Verena M. Lepper, 45–61. Oxford: Oxford University Press.
- Taylor, Jon. 2013. "Administrators and Scholars: The First Scribes." In *The Sumerian World*, edited by Harriet Crawford, 290–305. London: Routledge.
- Wilks, Yorick, and Christopher Brewster. 2006. "Natural Language Processing as a Foundation of the Semantic Web." *Foundation and Trends\* in Web Science* 1 (3–4): 199–327.
- Woolley, Charles Leonard. 1934. *Ur Excavations 2. The Royal Cemetery: A Report on the Predynastic and Sargonid Graves Excavated between 1926 and 1931*. London: The Joint Expedition of The British Museum and of the Museum of the University of Pennsylvania to Mesopotamia.

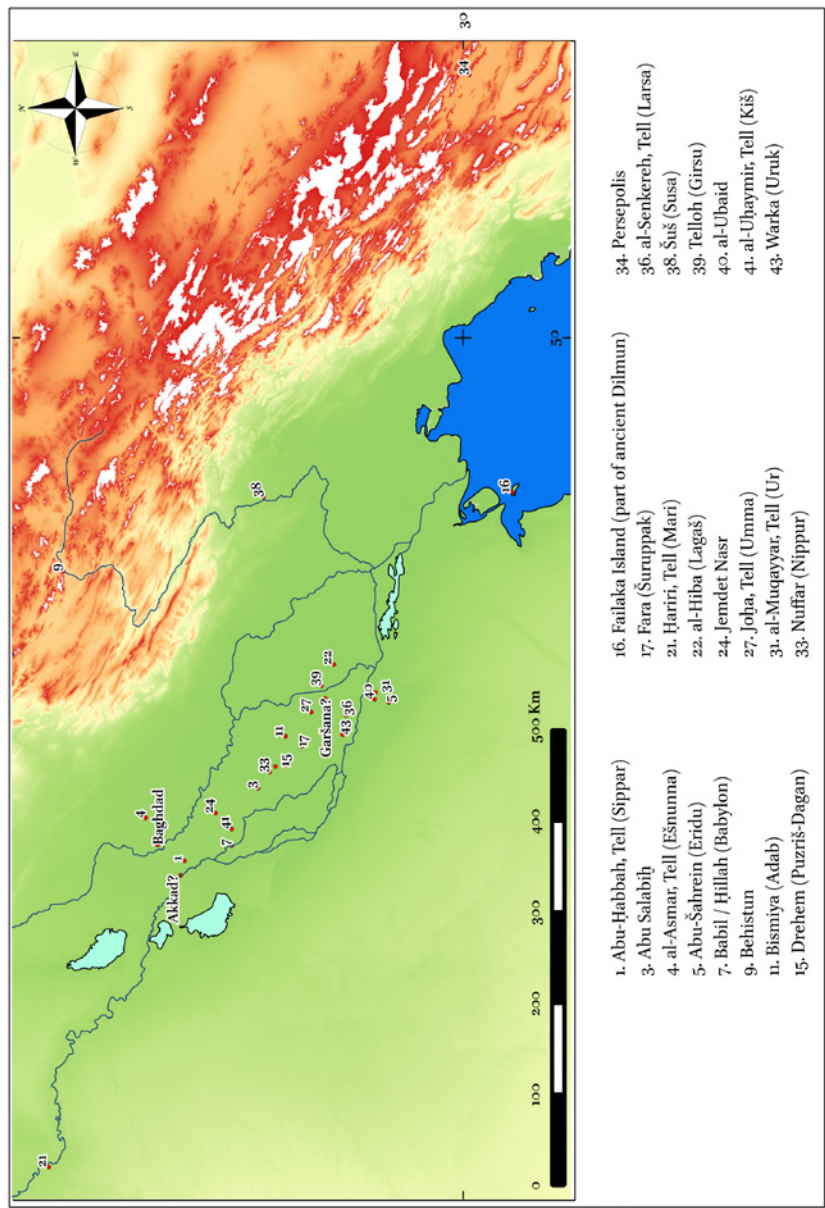


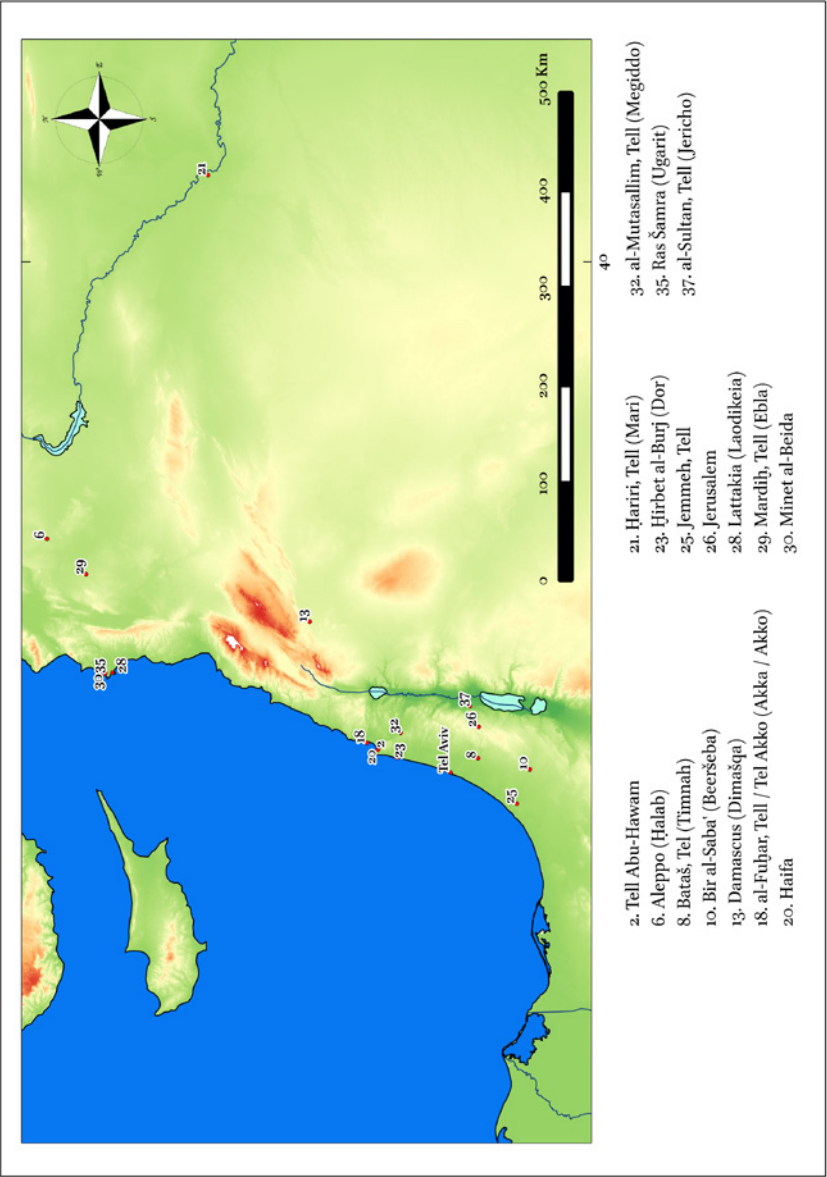


# Maps

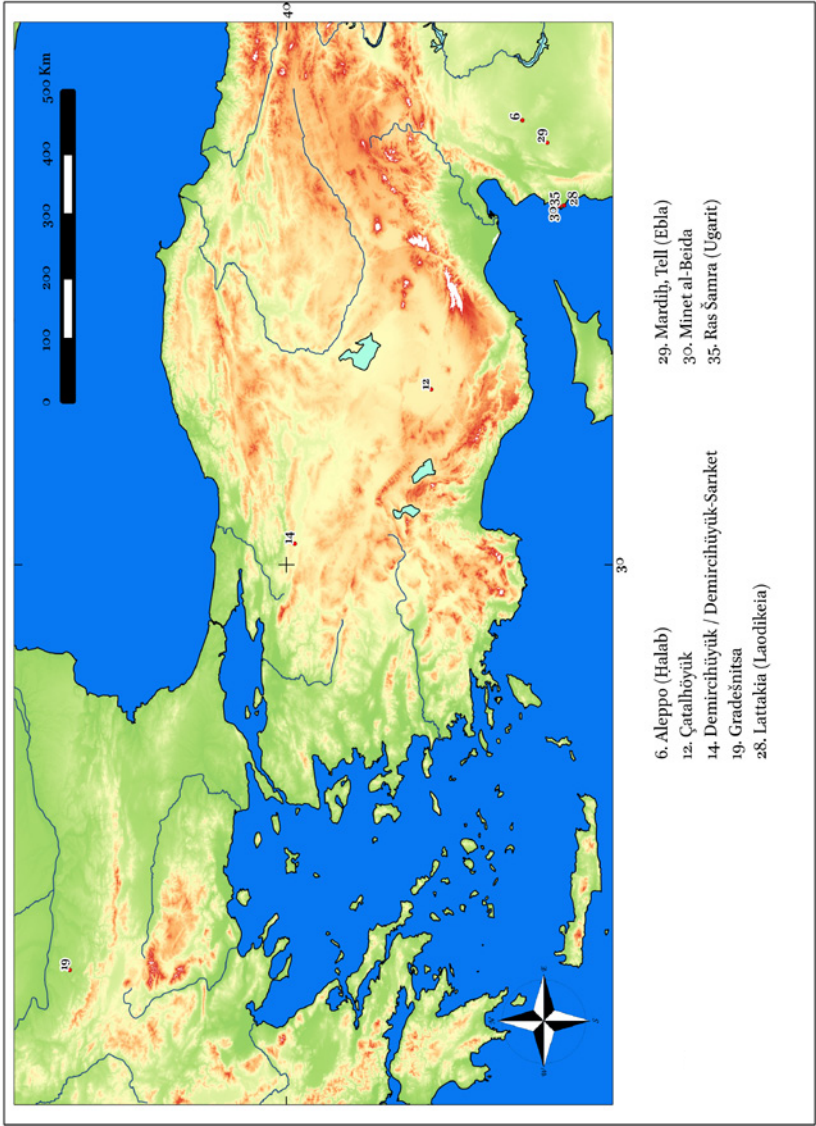


MAP 1 General map

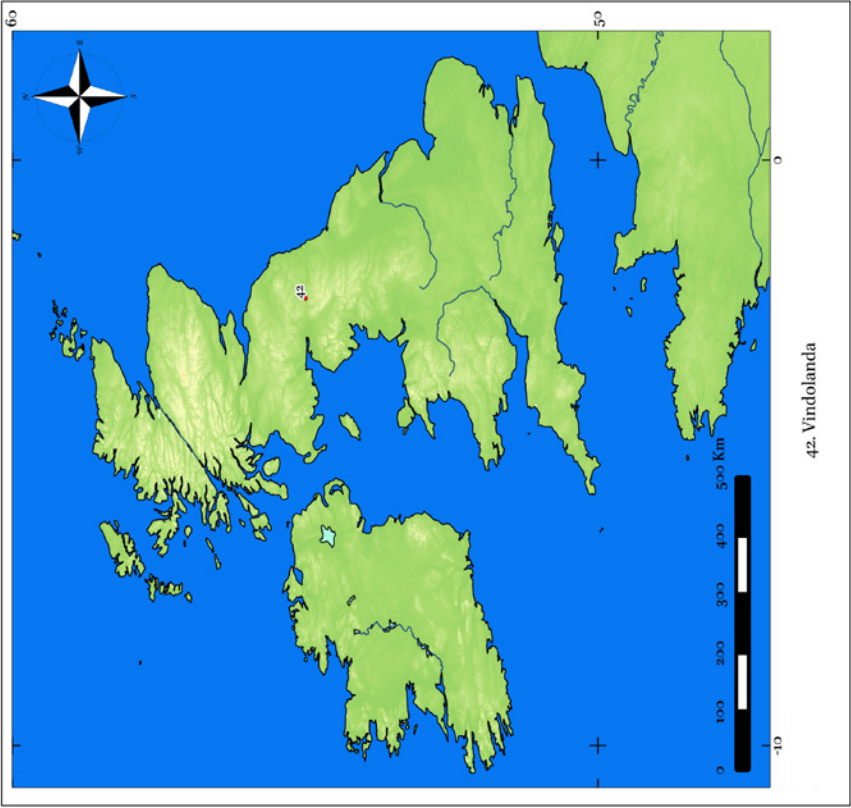




MAP 3 *Toponyms for the Levant and Syria*



MAP 4 *Toponyms for Anatolia*



MAP 5    *Location of Vindolanda (England)*



# Toponyms Related to Ancient Settlements or Regions\*

A     *List by modern toponyms (in brackets the ancient equivalent, if the correspondence is known)*

Toponym	Map(s)
1. Abu-Ḥabbah, Tell (Sippar)	1
2. Tell Abu-Hawam	2
3. Abu Salabiḥ	1
4. al-Asmar, Tell (Ešnunna)	1
5. Abu-Šahreïn (Eridu)	1
6. Aleppo (Ḥalab)	2, 3
7. Babil / Ḥillah (Babylon)	1
8. Bataš, Tel (Timnah)	2
9. Behistun	1
10. Bir al-Saba‘ (Beeršeba)	2
11. Bismiya (Adab)	1
12. Çatalhöyük	3
13. Damascus (Dimašqa)	2
14. Demircihüyük / Demircihüyük-Sarıket	3
15. Drehem (Puzriš-Dagan)	1
16. Failaka Island (part of ancient Dilmun)	1
17. Fara (Šuruppak)	1
18. al-Fuḥar, Tell / Tel Akko (Akka / Akko)	2
19. Gradešnitsa	3
20. Haifa	2
21. Ḥariri, Tell (Mari)	1, 2
22. al-Hiba (Lagaš)	1
23. Ḥirbet al-Burj (Dor)	2
24. Jemdet Nasr	1
25. Jemmeh, Tell	2
26. Jerusalem	2
27. Joḥa, Tell (Umma)	1
28. Lattakia (Laodikeia)	2, 3

\* For further information on site locations, see *Ancient Locations*, <<http://www.ancientlocations.net>>.

Toponym	Map(s)
29. Mardiḥ, Tell (Ebla)	2, 3
30. Minet al-Beida	2, 3
31. al-Muqayyar, Tell (Ur)	1
32. al-Mutasallim, Tell (Megiddo)	2
33. Nuffar (Nippur)	1
34. Persepolis	1
35. Ras Šamra (Ugarit)	2, 3
36. al-Senkereh, Tell (Larsa)	1
37. al-Sultan, Tell (Jericho)	2
38. Šuš (Susa)	1
39. Telloh (Girsu)	1
40. al-Ubaid	1
41. al-Uḫaymir, Tell (Kiš)	1
42. Vindolanda	4
43. Warka (Uruk)	1

B *List by ancient toponyms (in brackets the modern equivalent, if the correspondence is known)*

Toponym	Map(s)
11. Adab (Bismiya)	1
18. Akka / Akko (al-Fuḫar, Tell / Tel Akko)	2
Akkad (exact location unknown)	1
7. Babylon (Babil / Ḥillah)	1
10. Beeršeḇa (Bir al-Sabaʿ)	2
16. Dilmun (part of which is modern Failaka Island)	1
13. Dimašqa (Damascus)	2
23. Dor (Ḥirbet al-Burj)	2
29. Ebla (Mardiḥ, Tell)	2, 3
5. Eridu (Abu-Šahreïn)	1
4. Ešnunna (al-Asmar, Tell)	1
Garšana (exact location unknown)	1
39. Girsu (Telloh)	1
20. Haifa (Haifa)	2
6. Ḥalab (Aleppo)	2, 3



B *List by ancient toponyms (cont.)*

Toponym	Map(s)
37. Jericho (al-Sultan, Tell)	2
26. Jerusalem (Jerusalem)	2
41. Kiš (al-Uḥaymir, Tell)	1
22. Lagaš (al-Hiba)	1
28. Laodikeia (Lattakia)	2, 3
36. Larsa (al-Senkereh, Tell)	1
21. Mari (Ḥariri, Tell)	1, 2
32. Megiddo (al-Mutasallim, Tell)	2
33. Nippur (Nuffar)	1
34. Persepolis (Persepolis)	1
15. Puzriš-Dagan (Drehem)	1
1. Sippar (Abu-Ḥabbah, Tell)	1
38. Susa (Šuš)	1
17. Šuruppak (Fara)	1
8. Timnah (Bataš, Tel)	2
35. Ugarit (Ras Šamra)	2, 3
27. Umma (Joḥa, Tell)	1
31. Ur (al-Muqayyar, Tell)	1
43. Uruk (Warka)	1

# Glossaries

## CyberResearch\*

(including computer science, mathematical, and technological terms)

### Abstract Vector

Any representation of a  $\rightarrow$  sign that can be found in a sign list or a dictionary. These are called “abstract” because no sign in its native form would correspond to it 100%.

*Doğu Kaan Eraslan, 297n54*

$\rightarrow$  *features’ vector, Optical Character Recognition, sign*

**Adab Corpus**  $\rightarrow$  see General Glossary

### Adjacency Matrix

A matrix whose rows and columns are labeled with the  $\rightarrow$  nodes, and whose cell value represents whether an  $\rightarrow$  edge is present between the node labeled in the row and the node labeled in the column.

*Doğu Kaan Eraslan, 306n87*

It is essentially a table in which the rows and columns are cells from the corpus, and at the intersection of each row and cell, a number is given for how similar the two cells are.

*M. Willis Monroe, 273*

$\rightarrow$  *edge, node*

**Administrative Texts**  $\rightarrow$  see General Glossary

**AI**  $\rightarrow$  Artificial Intelligence

**Akkadian**  $\rightarrow$  see General Glossary

---

\* Although all words of an expression are capitalized in the list of terms below, when indicated after the “see also” arrow, words are capitalized when the expression has an acronym (for example, Conceptual Reference Model, CRM) but not capitalized when the expression does not have an acronym. There are some exceptions, such as Python, R, and Word2vec, that are applications or tools and thus always capitalized. The “definitions” provided in this glossary are neither exact nor comprehensive; rather, they are informative statements extracted from the chapters. In some cases, minor modifications (such as verb tense, or the addition of “it is”) have been made in order for the statement to make sense out of context. When multiple statements are presented for a term, they are listed according to the alphabetical order of their authors’ names. Cross-references point both to entries in this glossary as well as in the General Glossary.

**Algorithm**

A step-by-step procedure for solving a problem or accomplishing some end, especially by a computer.

*Shannon Martino and Matthew Martino, 118n29*

In Computer Science, an algorithm is a set of instructions a computer can execute to solve a (mathematical) problem.

*Émilie Pagé-Perron, 197n16*

→ *data mining*

**American Standard Code for Information Interchange (ASCII) characters**

Limited to characters that are used in the writing of English words without accents. They include but are not limited to: “3, b, R, a, m, k, [, ;”

*Doğu Kaan Eraslan, 285n12*

**Analyse Logiciste**

Analytical procedure developed in France focusing on outlining and highlighting the symbolic elements (and their relevant functions) that can describe the structure of an observed phenomenon. The logicist procedure is also a critical review of the methodologies and the epistemological perspectives adopted by different scholars.

*Alessandro di Ludovico, 87n4*

**ANN** → Artificial Neural Network

**Annotation** → see General Glossary

**Application Programming Interface (API)**

A set of rules and tools that defines how computers can interact. In this context, the → Online Cultural and Historical Research Environment API defines what database → items are available to extract and send to other programs.

*Miller C. Prosser, 318n18*

→ *item, Online Cultural and Historical Research Environment*

**Aramaic** → see General Glossary

**Arbitrary Unit**

A → spatial unit used as a primary spatial designator mostly on sites that lack any architectural remains or other clear spatial features that would allow for horizontal subdivision of an excavation area and defining vertical layers.

*Sveta Matskevich and Ilan Sharon, 39*

→ *spatial unit*

**Artificial Adaptive Systems (AAS)**

Biological computing methods, techniques, and → algorithms forming part of the vast world of → Natural Computation/Natural Computing, which is itself a subset of the → Artificial Sciences (AS).

*Marco Ramazzotti, 61n5*

→ *algorithm, Artificial Sciences, Natural Computing*

**Artificial Intelligence (AI)**

The development of models using symbol manipulation. The computation in the models is based on explicit representations that contain symbols organized in specific structures. The connectionist paradigm aims at massively parallel models that consist of a large number of simple and uniform processing elements interconnected with extensive links, the → Artificial Neural Networks .

*Marco Ramazzotti, 62n9*

→ *Artificial Neural Network*

**Artificial Intelligence Model**

Applied to → Mesopotamian Urban Revolution Landscape (MURL), Artificial Intelligence models are a type of artificial simulation of physical or cultural phenomena and features that have been previously encoded, so that from processing such models previously unknown relationships can be revealed through the features.

*Marco Ramazzotti, 63n12*

→ *Mesopotamian Urban Revolution Landscape*

**Artificial Neural Network (ANN)**

Artificial Neural Networks are a class of → algorithms for which the design is inspired by the observation of the mechanisms of the human brain. One of the most interesting features of ANNs is their ability to learn, which implies an attitude of flexibility, and to adapt to specific situations and datasets.

*Alessandro di Ludovico, 94n26*

They can be defined as Learning Systems; these are → algorithms for processing information that allow for the reconstruction, in a particularly effective way, of the approximate rules relating to a set of “explanatory” data concerning the considered problem (the input), with a set of data (the output) for which it is requested to make a correct forecast or reproduction in conditions of incomplete information.

*Marco Ramazzotti, 64n16*

→ *algorithm, Artificial Intelligence, neural networks, Self-Organizing Map, Word2vec*

**Artificial Sciences (AS)**

Sciences for which an understanding of natural and/or cultural processes is achieved by the recreation of those processes through automatic models.

*Marco Ramazzotti, 61n5*

→ *Artificial Adaptive Systems*

AS → Artificial Sciences

**ASCII Characters** → American Standard Code for Information Interchange characters

**ASCII Transliteration Format (ATF)**

ATF was created by the → Cuneiform Digital Library Initiative (CDLI) as a stable archiving format for the long-term storage of texts. It has evolved, first to adapt to the usage of CDLI, and later it branched out into → Oracc-ATF ... Whether one is working directly from the ancient → tablet or from paper publications, the ATF format is an excellent choice for encoding one's work for long-term preservation.

*Émilie Pagé-Perron, 200*

→ *American Standard Code for Information Interchange characters, Cuneiform Digital Library Initiative, Open Richly Annotated Cuneiform Corpus, tablet*

ASS → Artificial Adaptive System

ATF → ASCII Transliteration Format

**Atomization**

By atomization, we refer to the process of dividing data into many individual database → items. A text is atomized into many database items, each of which represents a single grapheme, either a letter or a → logosyllabic → sign.

*Miller C. Prosser, 314n4*

→ *item, logosyllabic, sign*

**Attribute (for digital practice)**

In the context of → Text Encoding Initiative (TEI), an attribute adds useful precision both for text analysis and for interpretation during the process of the exchange of data. An attribute stands within the element → tag.

*Vanessa Bigot Juloux, 165*

Attributes take the form of extra columns of information in the → nodes and → edges lists from which data can be filtered, such as by time period or transaction verb.

*Emile Pagé-Perron, 220n85*

→ *edge, node, tag, Text Encoding Initiative*

**Attribute (in Archaeology)** → see General Glossary

**Auto-CM** → Auto Contractive Map Neural Network

**Auto-Contractive Map (Auto-CM)**

An → Artificial Neural Network (ANN) that is characterized by a three-layer architecture consisting of 1) an input layer, where the signal is captured from the environment; 2) a hidden layer, where the signal is modulated inside the Auto-CM; and 3) an output layer, through which the Auto-CM feeds back upon the environment on the basis of the stimuli previously received and processed. The top layer is the input layer, and these are equivalent to the variables among which we seek to identify relationships.

*Giulia Massini, in Marco Ramazzotti, 73n36*

→ *Artificial Neural Network*

**Bag-of-Words Model**

A well-known method in textual analysis in which all of the individual lexemes from a text are flattened and placed in a large group. Syntax and order are lost and do not figure into the final analysis, but this model offers a level of simplicity that functions very well for comparing the similarity of texts within a corpus.

*M. Willis Monroe, 270*

→ *Continuous Bag-of-Words model*

**Big Data**

Large and complex datasets that require computational methods for their analysis.

*Vanessa Bigot Juloux, 156n30*

**Bridge**

→ Nodes that connect otherwise unconnected groups of nodes together ... Bridges consist of nodes that usually have a high → edge betweenness ... In the → Adab corpus, bridges are high-ranking individuals, or they represent homonyms (different individuals who have the same name and who should be represented with two or more separate nodes).

*Émilie Pagé-Perron, 202, 212*

→ *Adab corpus, edge betweenness, graph partitioning, node*

**C(anonical)-ATF (C-ATF)**

The backbone of the texts displayed by the → Cuneiform Digital Library Initiative (CDLI) ... It uses → American Standard Code for Information Interchange (ASCII) characters.

*Doğu Kaan Eraslan, 285*

→ *American Standard Code for Information Interchange characters, Cuneiform Digital Library Initiative*

**C-ATF** → **C(ANONICAL)-ATF**

### **CAL Code**

Adopted by the Comprehensive Aramaic Lexicon (CAL) project, it is used for storing → Aramaic texts from the ninth century BCE to the thirteenth century CE. CAL Code was developed in the 1980s. It has changed considerably with the arrival of the internet and → Unicode, and it remains effective in giving a better understanding of displayed text. CAL Code uses → ASCII characters and was designed so that Aramaic scholars could use it without intensive training.

*Doğu Kaan Eraslan, 288*

→ *American Standard Code for Information Interchange characters, Aramaic, Unicode*

**CBOW** → Continuous Bag-of-Words

**CDLI** → Cuneiform Digital Library Initiative

### **CIDOC Conceptual Reference Model (CIDOC CRM)**

A leading standard (→ ISO 21127/2006) for conceptual referencing of cultural heritage data. Developed in the last decade by one of the special-interest groups of ICOMOS for museums and object collections, it gained popularity in other domains of the humanities and cultural heritage. The CIDOC CRM was first adopted into the archaeological domain by English Heritage. After some adaptations and additions (the modified version was dubbed CRM-EH), they successfully mapped their field recording systems. Due to the standard's growing popularity, the initial group of researchers developed several extensions of the basic → CRM, aiming at incorporating additional information about the objects, such as bibliographic references, provenience, and other scientific analyses, as well as spatio-temporal models. The archaeological fieldwork extension, CRMarchaeo, accommodates stratigraphy-related concepts that were missing from CIDOC CRM.

*Sveta Matskevich and Ilan Sharon, 45*

The purpose of the CIDOC CRM is to establish a common framework for the sharing of data between GLAM institutions, and to function as an example of best practices in the cultural heritage domain.

A number of other → ontologies have already been incorporated into the CIDOC CRM ... with additional available extensions and use cases further enabling the representation of archaeological data and processes, the provenance of digital artifacts, and bibliographies. All of these provide opportunities for linking to other data streams. CIDOC CRM is a large and complex ontology consisting of 90 distinct entities and 149 property declarations.

*Terhi Nurmikko-Fuller, 353–354*

→ *Conceptual Reference Model, International Organization for Standardization, ontology*

**CIDOC CRM** → CIDOC Conceptual Reference Model**Clique**

An important concept in → network graph analysis, especially when working with → administrative records for which the source data generates undirected links among all of the people appearing in the same text. Each → tablet therefore forms a clique, that is, a fully connected subgraph where all → nodes are interrelated. Cliques are not exclusive: they can overlap and are formed by any group of nodes that have a link between each of them. For example, in a group where a, b, and c are all related, there are also three cliques of two nodes: (a, b), (a, c), and (b, c).

*Émilie Pagé-Perron, 214*

→ *administrative text, maximal clique, network graph analysis, node, tablet*

**Cluster Analysis**

A particular form of → multivariate analysis that divides a dataset into groups based on how near the data points are to each other and gives one the ability to analyze subgroups of the dataset using an → algorithm ... A type of analysis that divides a set of objects into groups (clusters) so that objects of one group are similar to each other, whereas objects within groups are dissimilar from the objects within other groups.

*Shannon Martino and Matthew Martino, 118, 118n28*

→ *algorithm, multivariate analysis*

**Clustering Method**

A statistical approach to group together → nodes that resemble each other. Those methods tell us about similarities in the dataset, but the researcher has to interpret the meaning of the groups ... Clustering methods have proven useful in → social network analysis for many decades.

*Émilie Pagé-Perron, 220, 220n88*

→ *cophenetic correlation coefficient, network analysis, node, social network analysis, Unweighted Pair Group Method with Arithmetic Mean*

**Comma Separated Value (csv)**

A format that most programs can interact with given the simplicity of its formatting and coding.

*Shannon Martino and Matthew Martino, 137n62*

A format that organizes data as a table, with new lines forming rows and commas delimiting columns.

*Émilie Pagé-Perron, 206n49*



### Computer-aided Textual Analysis

A research method that can answer questions about themes or the usage of terms across a corpus containing millions of words ... Through computer-aided textual analysis, we can visualize what would otherwise be mere hunches gleaned from a close reading of the texts.

*M. Willis Monroe, 258–259*

### Computer Science

Computer Science brings together disciplines including Mathematics, Engineering, the Natural Sciences, Psychology, and Linguistics.

*Marco Ramazzotti, 60n2*

### Computer Semiotics

Any empirical approach mainly interested in the ways that humans and machines may communicate with each other using written languages, texts, and/or codes.

*Marco Ramazzotti, 60n1*

### Computerized Data Modeling

The process of creating a digital → data model.

*Sveta Matskevich and Ilan Sharon, 55n108*

→ *data model*

### Conceptual Reference Model (CRM)

Models that explain what things are in the digital world and how an object relates itself to the digital world.

*Doğu Kaan Eraslan, 300n64*

A *lingua franca* of recording systems, and the main purpose of ontological → mapping is to ensure compatibility of the terminologies used by various data management systems, which is one of the necessary conditions for creating → Linked Open Data clouds.

*Sveta Matskevich and Ilan Sharon, 45*

→ *CIDOC Conceptual Reference Model, data cloud, Database Management System, Linked Open Data, mapping, ontology*

### Continuous Bag-of-Words (CBOW) Model

One of two alternative models one can choose from when training data in → Word2vec. While training, the CBOW model uses the various words that can appear in the same context as the target word for the prediction of semantic relationships.

*Saana Svärd, Heidi Jauhiainen, Aleksi Sahala, Krister Lindén, 247n56*

→ *bag-of-words model, Word2vec*

### Continuous Skip-gram Model

One of two alternative models one can choose from when training data in → Word2vec. While training, the Continuous Skip-gram model predicts the words that may appear near the target word.

*Saana Svärd, Heidi Jauhiainen, Aleksi Sahala, Krister Lindén, 247n57*

→ *Word2vec*

### Cophenetic Correlation Coefficient

A metric that can evaluate the validity of a particular → clustering technique together with a measure of distance or dissimilarity.

*M. Willis Monroe, 274*

→ *clustering method*

### Correspondence Analysis

A → quantitative method used to describe a corpus of qualitative data. It falls within the statistical realm of “factor methods” and can be described as the analysis of the dependent relationships existing among the different features of the observed data. When the data are collected in a table, for example, the rows can be referred to as the observed specimens (in this case the seals’ scenes), while the columns refer to the different features (here, the different possible iconographic elements that appear in the scenes) that can pertain to the specimens. Correspondence analysis permits the investigation of the associations between rows and columns of the table and enables associations to be displayed in a graphic form.

*Alessandro di Ludovico, 94–95n29*

→ *hierarchical classification, quantitative analysis/method*

### Cosine Similarity

A geometric method for determining similarity between rows in the → Document Term Matrix (DTM) by assigning each row to a → vector in multi-dimensional space and calculating the cosine of the angle between each vector.

*M. Willis Monroe, 272*

→ *Document Term Matrix, vector*

### CountVectorizer

The function, CountVectorizer from the Scikit-learn module, runs through each bag and tallies a total list of all terms, then marks, for each bag, which terms appear in that bag. The end result is a matrix of all cells and all terms, with the count for each term in each cell tallied in the table.

*M. Willis Monroe, 271–272*

**CRM** → Conceptual Reference Model

**Cuneiform** → see General Glossary

### **Cuneiform Digital Library Initiative (CDLI)**

Currently the major online database for → cuneiform texts. As the name implies, CDLI conserves epigraphical and semantic data of cuneiform texts, rather than for a specific language. It uses → American Standard Code for Information Interchange (ASCII) characters.

*Doğu Kaan Eraslan, 285*

The largest database of → cuneiform artifacts; it seeks to collect information about all → Mesopotamian inscribed objects. It stores this data as → metadata and images named “fatcrosses,” as well as → transliterations, normalizations, and translations in various languages ... One of CDLI’s roles is to maintain a complete catalog and digital copy of all cuneiform texts. Contributions are updated periodically by CDLI staff based on new research, changes in standards, and direct contributions from scholars.

*Émilie Pagé-Perron, 199*

→ *American Standard Code for Information Interchange characters, cuneiform, Mesopotamia, metadata, transliteration*

### **Cytoscape**

A more powerful software than → Gephi. It is used not only by digital humanists but also by biologists and other scientists. It offers an extended number of features, especially for the statistical data analysis ... Cytoscape is free and open-source.

*Émilie Pagé-Perron, 209, 218n80*

→ *Gephi, graphing programs, weighted*

**DANA** → Digital Archaeology and National Archive

### **Data Cloud**

A visualization diagram for a → triple store or multiple → graph datasets in the → Linked Open Data initiative.

*Sveta Matskevich and Ilan Sharon, 49n95*

→ *graph, Linked Open Data, triple store*

### **Data Mining**

Such a task can be accomplished through a large number of different types of → algorithms and strategies. Typical of data mining is the search for correlations among features, like variables, within a certain dataset, so that the results of an investigation

of that dataset can give clues to predict the figures related to those features in similar cases, or the possible development of some features in case some other features change in a certain way. Just as resources can be found and extracted from a mine through proper types of investigation and extraction techniques, useful information that was not evident at a first look can be extracted from a dataset through data-mining techniques.

*Alessandro di Ludovico, 91n21*

→ *algorithm, Système Portable pour l'Analyse des Données, text mining*

### **Data Model**

An abstract model that describes data structure and relations within a system, where the data is stored, and how it is processed.

*Sveta Matskevich and Ilan Sharon, 48n91*

A framework with rules. All databases have an underlying data model, which is simply an abstraction that defines how data is connected and processed in the database.

*Miller C. Prosser, 319n20*

→ *computerized data modeling*

### **Data Serialization**

The process that enables us to save the data we are working on so that it can be reconstructed just as we had saved it in different platforms.

*Doğu Kaan Eraslan, 299n61*

→ *External Data Representation, JavaScript Object Notation*

### **Database Management System (DBMS)**

A software that enables an administrator to create and manage databases and to monitor, modify, and analyze their interactions with users and other applications.

*Sveta Matskevich and Ilan Sharon, 47n84*

→ *Conceptual Reference Model, Identifier, meta-system*

**DBMS** → Database Management System

### **Deep Learning**

A subfield of machine learning. It is characterized by the breakdown of complex representations into a hierarchy of simple components during the learning process.

*Doğu Kaan Eraslan, 296n51*

### **Dictionary**

A type of abstract data structure (also known as an associative array or symbol table)

used in programming languages, implemented under the name `dictionary` in Python. A dictionary stores information as key value pairs. Use of dictionaries, at least in the context of Python, implies that we would have a non-negligible performance gain for operations in which we want to process a value by using its associated key.

*Doğu Kaan Eraslan, 297n55*

→ *map/mapping*

### **Digital Archaeology and National Archive (DANA)**

Software for archaeological field recording developed by the Israel Antiquities Authority.

*Sveta Matskevich and Ilan Sharon, 28n20*

### **Distributional Semantic Models**

In the context of → language technology, distributional semantic models keep track of the appearances of words according to their proximity to each other in order to measure their similarity. These models can be computed by counting the frequency of all the words in relation to their neighboring words or, more recently, by predicting the context where words appear.

*Saana Svärd, Heidi Jauhiainen, Aleksi Sahala, Krister Lindén, 228–229*

→ *language technology, Pointwise Mutual Information, Word2vec*

### **Distributional Semantics**

The study of semantic similarities between linguistic → items by using mathematical methods to see how they are spread throughout a large dataset.

*Saana Svärd, Heidi Jauhiainen, Aleksi Sahala, Krister Lindén, 229n17*

→ *item*

### **Document-oriented Database**

A data store, the main concept of which is the document. Documents can be organized hierarchically, grouped into collections by some criteria, or → tagged. Entities in a document store do not necessarily share a structure; differently structured documents can be stored in the same database.

*Sveta Matskevich and Ilan Sharon, 51n99*

→ *tag*

### **Document Term Matrix (DTM)**

A large → “vector space” in which the cells reside. Each unique term within the corpus becomes a dimension on which the documents are plotted; for instance, if your corpus consists of documents that only use three words, the resulting vector space would be

recognizable as a three-dimensional plot. In practical terms the DTM is represented by a large table in which each row is one of the input documents, in this case a cell from one of our texts, and each column represents a unique term that shows up in the entire corpus.

*M. Willis Monroe, 271*

→ *cosine similarity, vector space*

**DTM** → Document Term Matrix

### Edge

Connection between two vertices on a → graph.

*Giulia Massini, in Marco Ramazzotti, 73n2*

Connections between → nodes, often formed by shared terms, words, or other meaningful units of analysis.

*M. Willis Monroe, 260*

→ *graph, label, node*

### Edge Betweenness

In → network analysis, a mathematical centrality measure of relationships between network data points.

*Émilie Pagé-Perron, 212n62*

→ *bridge, edge, network analysis*

### Ego-network

A → network graph of one individual, and of its neighbors, and optionally all of the neighbors of these neighbors. In this context, a neighbor refers to a → node that is directly connected with the entity being investigated.

*Émilie Pagé-Perron, 209n54*

→ *Gephi, network graph, node*

**Elamite** → see General Glossary

### Electronic Text Corpus of Sumerian Literature (ETCSL)

An online resource that provides access to the → transliterations (in both → Unicode and → American Standard Code for Information Interchange, ASCII) and translations of some 400 literary compositions, all dated to the late third or early second millennium BCE. The name encapsulates the form and function of this bi-part → web-based resource ... For the → Sumerian, there are transliterations rendered using the Latin alphabet and modern philological conventions (with no accompanying cuneiform, either

Unicode or otherwise) on a line-by-line basis. For the English translation, there are paragraphs of prose. The ETCSL provides external links to other related and complementary projects. These include other online corpora, resources for museological and palaeographical data, and a dictionary ... Although not actively developed since 2006, the ETCSL is a cited resource and in itself the focus of research. It was an innovative project in the realm of study of Sumerian literature in that it offered free and unrestricted online access to the data, with the aim of interlinking with external resources.

*Terhi Nurmikko-Fuller, 351–352*

→ *American Standard Code for Information Interchange characters, Sumerian, transliteration, Unicode, web*

### Element

In the context of → Text Encoding Initiative (TEI), an element, which is a → markup → tag, is the first criterion to analyze data (in the form of text, image, sound, etc.). An element is conventionally marked up <element>. The markup data in between the opening tag "<" and the closing tag ">" indicates the type of information analyzed: for example, <persName>'Anatu'</persName>. One can easily understand that the element refers to a personal name, in this example 'Anatu ... Generally, the semantic of an <element> is very close in its abbreviation to its English vocabulary. For example, <interp> stands for "interpretation," <w> for "word," <l> for "line," and <text> for "text."

*Vanessa Bigot Juloux, 165n68*

→ *attribute, markup/markup tagging, tag, Text Encoding Initiative*

### Encoding Scheme

A set of rules that standardizes the representation of an object in a digital environment.

*Doğu Kaan Eraslan, 284n4*

→ *interoperability/epigraphic interoperability, minimal semantic unit*

### EpiDoc

Strictly speaking, EpiDoc is an → XML-scheme for adding → markup to ancient texts. Thus, it is applicable to all ancient languages, but it is predominantly used for Graeco-Roman corpora. It employs → Text Encoding Initiative (TEI) scheme as its backbone, and it accepts anything encoded with UTF-8. Since EpiDoc is fundamentally a markup language that tries to be applicable to a broad range of languages, it is fairly liberal when it comes to using values for attributes.

*Doğu Kaan Eraslan, 289*

→ *Extensible Markup Language, markup/markup tagging, Text Encoding Initiative, Unicode Transformation Format-8*

ETCSL → Electronic Text Corpus of Sumerian Literature

### Event

A piece of contents (visual, textual, numeric measurement, for example) recorded by a particular observer at a particular time and place that conveys information concerning one or more semantic entities.

*Sveta Matskevich and Ilan Sharon, 36*

→ *syntactic elements*

### Extensible Markup Language (XML)

Standardized in the late 1990s, Extensible Markup Language (XML) is a predefined → markup language that follows a standard syntax enabling data exchange to be easily → machine readable. One speaks then of → interoperability data—for example, XML would be a kind of artificial constructed language such as Esperanto, an artificial international language ... XML also offers to withstand time and new technologies deployed on the internet or locally (on a computer)—in other words, the current methodology is sustainable even as technology evolves.

*Vanessa Bigot Juloux, 163–164, 164n63*

A form of → markup language, like → HTML, which encodes → annotations, like the codes hidden behind the text we see when we use a word processing program, such as LibreOffice or Microsoft Word.

*Emilie Pagé-Perron, 204n42*

→ *annotation, data serialization, EpiDoc, HyperText Markup Language, interoperability/epigraphic interoperability, machine-actionable data/machine-readable data, markup/markup tagging, RDF-XML*

### Extensible Markup Language (XML) Annotation

→ XML → annotations, including → Text Encoding Initiative (TEI), are perhaps the most popular means of annotating texts across disciplines. These types of annotations provide two major advantages: (1) ease of marking (with precision) the exact position of the annotated entity in the text and (2) flexibility of the → markup, making it possible to annotate multiple layers of information in one text. A good example of a successful annotated corpus available on the → web using XML is the → Electronic Text Corpus of Sumerian Literature (ETCSL).

*Émilie Pagé-Perron, 204*

→ *annotation, Electronic Text Corpus of Sumerian Literature, Extensible Markup Language, markup/markup tagging, Text Encoding Initiative, web*



### External Data Representation (XDR)

Like → JavaScript Object Notation (JSON), it is a standardized serialization format.

*Doğu Kaan Eraslan, 299n60*

→ *data serialization, JavaScript Object Notation*

### Factors of Variance

Also called factors of variation, they refer to elements that are believed to be constitutive of the observed state of that which is observed. This could include things such as elements of form (e.g., lines, angles), degrees of arc or convexity, color, and light intensity. Such features make the object distinguishable in a representation.

*Doğu Kaan Eraslan, 308n93*

→ *feature extraction algorithm*

### Feature Extraction Algorithm

The process through which one extracts the → factors of variance from a digital representation of an object. For example, it could be an → algorithm that detects the corners and sides in a picture of a triangle.

*Doğu Kaan Eraslan, 308n93*

→ *algorithm, factors of variance*

### Features' Vector

It is a convenient term for the mathematical construct that contains the patterns of the image of the → sign.

*Doğu Kaan Eraslan, 297n54*

→ *sign*

### FRBR → Functional Requirements for Bibliographic Records

#### FRBROO

A formal → ontology for the representation of bibliographic information. It has been designed to merge with the → CIDOC Conceptual Reference Model (CIDOC CRM) and to facilitate the integration of museum and library data. It is based on the → Functional Requirements for Bibliographic Records (FRBR), which was originally designed as an entity-relationship model ... The first draft of FRBROO was completed in 2006, with the official publication of the version 1.0 at the end of 2009. It was approved and issued by January 2010 ... FRBROO is referred to in 60 of the CIDOC CRM's classes and in 55 properties.

*Terhi Nurmikko-Fuller, 355*

→ *CIDOC Conceptual Reference Model, Functional Requirements for Bibliographic Records, ontology*

### Functional Requirements for Bibliographic Records (FRBR)

FRBR was developed independently of → CIDOC CRM, but coincidentally at approximately the same time (1991–1997). It was designed by a group appointed by the International Federation of Library Associations and Institutions (IFLA), and approved by IFLA Cataloguing Section in 1997. In 2003, an international working group began to examine the potential of merging FRBR with the CIDOC CRM. Unlike the CIDOC CRM, FRBR models products, not processes. It is smaller and less complex, with 52 entities and 64 properties. The combination of FRBR and the CIDOC CRM brought about → FRBROO—a new, object-oriented → mapping.

*Terhi Nurmikko-Fuller, 355*

→ *CIDOC Conceptual Reference Model, FRBROO, map/mapping*

### Gaussian Distribution

A (bell curve) that has a central peak, is evenly distributed about the mean, and can be quantified with just the mean and the standard deviation (related to the width of the bell curve).

*Shannon Martino and Matthew Martino, 121*

### Gephi

Gephi is free, open source, and easy to install ... Among the software available on the market, Gephi, with its intuitive interface, is a good entry-level tool for producing a first → network graph. Importing data into Gephi is straightforward, and it is easy to style → nodes and → edges depending on the user's needs. Gephi can also convert → graph data to numerous formats; this comes in handy when one is using more than one tool to manipulate the data. Gephi also integrates the sigma.js plugin, which can prepare a website on-the-fly for displaying an interactive network graph. Finally, Gephi includes a useful feature for isolating → ego-networks that provide information about individuals' relationships. An alternative to Gephi is → Cytoscape.

*Émilie Pagé-Perron, 209, 209n55*

→ *Cytoscape, edge, ego-network, graph data, graphing programs, network graph, node*

**Glyph** → see General Glossary

### Graph

A mathematical abstraction that basically consists of a set of vertices and a set of → edges, where an edge represents the connection between two vertices on a graph.

*Giulia Massini, in Marco Ramazzotti, 73*

Used to visualize the connections between cells, a graph is made of → nodes and → edges, both of which are determined by the nature of the evidence.

*M. Willis Monroe, 260*

→ *edge, Gephi, graph data, node*

### Graph Data

It enables the exploration of the relationships among entities.

*Émilie Pagé-Perron, 197*

→ *Gephi, graph*

### Graph Model

A model that describes relations between objects as a collection of → nodes connected by → edges ... In a graph model, the data are structured so that each entity (subject), its attribute (object), and the relationships between them (predicate) create a → triple ... The main advantage of the graph model is its flexibility, which allows for querying large datasets by setting relations as query criteria.

*Sveta Matskevich and Ilan Sharon, 46, 46n80*

→ *edge, node, triple*

### Graph Partitioning

Technique ... that consists of removing entities that act as a sole or faster point of access between groups in a → graph, thus separating these groups from each other. It may be evident from the visualization which individuals serve as → bridges, but using a → quantitative method to remove them is more efficient, since the threshold used will be numerical and not solely visual.

*Émilie Pagé-Perron, 212*

→ *bridge, graph, quantitative analysis/method*

### Graph Visualization Software → Gephi

#### Graphical User Interface (GUI)

The medium in which users interact with the functions of the computer program through visual signs and symbols as opposed to written commands, as in the case of command line interface. Programs with GUI include, but are not limited to, Microsoft Word, Open Office Writer, Internet Explorer, and Mozilla Firefox.

*Doğu Kaan Eraslan, 303n77*

→ *Système Portable pour l'Analyse des Données*

#### Graphing Programs

Tools that allow one to visually interpret numerical data and export the results as images.

*M. Willis Monroe, 274n36*

→ *Cytoscape, Gephi*

GUI → Graphical User Interface

**Hierarchical Classification**

In the context of → correspondence analysis, it is based on similarity and proximity. It classifies groups and shows hierarchical dependences among them.

*Alessandro di Ludovico, 95n29*

→ *correspondence analysis*

**Hierarchical Clustering Algorithm**

Procedure of defining clusters (types) by considering how similar objects are to each other (in a quantitative fashion) and grouping “close” objects into the same type.

*Shannon Martino and Matthew Martino, 120*

→ *algorithm, cluster analysis, quantitative analysis/method*

**Hierarchical Softmax**

When using → Word2vec, a method for creating the output vector, the values for the words are stored in such a way that not all of the words have to be processed when calculating the probability of a word.

*Saana Svärd, Heidi Jauhiainen, Aleksi Sahala, Krister Lindén, 247*

→ *negative sampling, Word2vec*

**HTML** → HyperText Markup Language

**HTTP** → HyperText Transfer Protocol

**HTTP URIs** → HyperText Transfer Protocol Universal Resource Identifiers

**Hub**

A → node with an important number of connections compared to the average number of connections of nodes in a said → network graph.

*Émilie Pagé-Perron, 210n58*

→ *network graph, node*

**Human Readable**

A text that is destined to be read by humans, as opposed to → “machine readable,” which means it is intended to be processed by machines.

*Doğu Kaan Eraslan, 287n17*

→ *machine-actionable/machine-readable*

**Hyperlinks**

Often visualized as blue font and underlined text, hyperlinks allow users to move between → websites.

*Terhi Nurmikko-Fuller, 338*

→ *HyperText Transfer Protocol, web*

### HyperText Markup Language (HTML)

HTML is used for formatting and the visual appearance of content of a → web page, but it does not capture meaning.

*Terhi Nurmikko-Fuller, 339*

→ *HyperText Transfer Protocol, web*

### HyperText Transfer Protocol (HTTP)

The technology enabling interconnectedness over the → web through → hyperlinks. The acronym is familiar to us from the address bar displayed at the top of our browsers, starting with “http://”.

*Terhi Nurmikko-Fuller, 338*

→ *hyperlinks, HyperText Markup Language, HyperText Transfer Protocol Universal Resource Identifiers, web*

### HyperText Transfer Protocol Universal Resource Identifiers (HTTP URIs)

These clusters of letters and characters form unique identifiers for each and every page on the → web. They are known as → Universal Resource Identifiers (URIs), and although other types of identifiers are possible, the ones we see when browsing are HTTP URIs. So common are the “http://” and “www.” prefixes that, when we discuss these resources, both are taken to be givens and thus are completely omitted; we refer to “Google,” for example, rather than to “http://www.google.com” ... For the most part, HTTP URIs point to specific pages. The majority of the content of these pages is text, encoded using the → HyperText Markup Language (HTML) ... On the → Semantic Web, rather than pointing to web pages, HTTP URIs serve as identifiers for specific data entities and information published on a page, a site, or in a resource.

*Terhi Nurmikko-Fuller, 338–339*

→ *HyperText Markup Language, HyperText Transfer Protocol, Semantic Web, Universal Resource Identifiers, web*

**ID** → Identifier

### Identifier (ID)

In archaeological contexts, the identifier of an object is expressed in a label that consists of a combination of any alphanumeric characters attached to that object. Most data management systems (and most archaeological recording systems, be they manual or computerized) require that any entity has a unique ID.

*Sveta Matskevich and Ilan Sharon, 36–37n52*

A unique number that does not have a specific meaning. Its quality resides in its uniqueness.

*Émilie Pagé-Perron, 204n45*

### **International Organization for Standardization (ISO)**

It is a non-governmental body that develops and publishes international standards of products, services, and systems.

*Sveta Matskevich and Ilan Sharon, 45n72*

It is the world's largest developer and publisher of international standards. It is an independent, voluntary organization, with representatives from each of the 162 member countries.

*Terhi Nurmikko-Fuller, 353n55*

### **Interoperability/Epigraphic Interoperability**

The ability to transform the → encoding scheme of an encoded epigraphic phenomenon into another encoding scheme without losing data.

*Doğu Kaan Eraslan, 284n5*

→ *encoding scheme*

**ISO** → *International Organization for Standardization*

### **Item**

An item could be an entire archaeological site or a single seed discovered in the course of excavation. An item can be anything from a place, to a person, to an image, to a text, to just about anything observable or conceptual.

*Miller C. Prosser, 319*

### **Iterative Process**

The process of producing and analyzing data that generates results while moving toward an end research goal.

*M. Willis Monroe, 270*

### **Java Application Client**

This client interface is a → Java Web Start application that launches on any computer with an internet connection.

*Miller C. Prosser, 318*

→ *Java Web Start, Online Cultural and Historical Research Environment, web*

**Java Client** → Java application client

### **Java Web Start**

A technology that allows one to launch a computer program without going through an installation process.

*Miller C. Prosser, 318n17*

→ *Online Cultural and Historical Research Environment*

### **JavaScript Object Notation (JSON)**

A serialization format, meaning that it allows us to save the state of our data. JSON is well standardized and well supported as an interchange format between different platforms.

*Doğu Kaan Eraslan, 299n59*

→ *data serialization, External Data Representation*

**JSON** → *JavaScript Object Notation*

### **$k$ -plex**

A relaxed → maximal clique in which  $k$  represents the number of connections that can be missing between → nodes when the nodes would still form a group, the  $k$ -plex. In a  $k$ -plex, most nodes are interconnected, but with a tolerance of  $k$ -number of absent → edges. In other words, an entity must be tied to all but  $k$  other entities in the group.

*Émilie Pagé-Perron, 216*

→ *edge, maximal clique, node*

### **Label**

A form of → attribute used to identify → nodes and → edges easily instead of using only their unique → identifiers.

*Émilie Pagé-Perron, 202n35*

→ *attribute, edge, identifier, node*

### **Language Technology**

A multidisciplinary field that studies and develops methods for processing human language with the help of computers.

*Saana Svärd, Heidi Jauhiainen, Aleksi Sahala, Krister Lindén, 225n5*

→ *distributional semantic models*

**Lemma** → see General Glossary

**Lexicography** → see General Glossary

### Library

A collection of commonly used programming functions that are distributed as a package for use by a wide range of users.

*M. Willis Monroe, 268n29*

In cyber-research contexts, “library” has two distinct usages: 1) a “digital library” is an online collection of digital objects that can be surrogates of actual objects or parts of objects, or born digital objects (i.e., → Cuneiform Digital Library Initiative) 2) a “software library” is a bundle of code that has a specific focus of application and that can be reused by many while developing software (i.e., Python library).

*Emilie Pagé-Perron, 218n79*

→ *Cuneiform Digital Library Initiative*

### Linked

In the world of the digital humanities, the term “linked” can have a specific connotation, meaning the mode of modelling data for sharing across the → web with other datasets with which one’s data was previously not connected. Within the → Online Cultural and Historical Research Environment (OCHRE) system, the term refers to database → items that “point” to each other, thereby creating a link between them. The database is a network of millions of files pointing at each other, i.e., “linked” to each other.

*Miller C. Prosser, 319n23*

→ *item, Online Cultural and Historical Research Environment, web*

### Linked Data (LD)

Information that has been captured and published using the → w3c standards of → Resource Description Framework (RDF) and → HyperText Transfer Protocol (HTTP) URIs. The term does not, however, signify that the data in question has been made openly available.

*Terhi Nurmikko-Fuller, 344*

A directed → network graph in which all → edges are named and directional.

*Émilie Pagé-Perron, 201n30*

→ *edge, HyperText Transfer Protocol Universal Resource Identifiers, linked, network graph, ontology, Resource Description Framework, Simple Knowledge Organization System, World Wide Web Consortium*

### Linked Open ‘Data (LOD)

It combines both → Linked Data (LD) and → Open Data (OD): it is often based on OD that has been converted to → Resource Description Framework (RDF), and it is similarly available for consumption and reuse without restrictions or limitations.

*Terhi Nurmikko-Fuller, 344*

→ *linked, Linked Data, Open Data, Resource Description Framework*



**Locus** → see General Glossary

**LOD** → Linked Open Data

### **Logarithm**

By common definition, a logarithm (abbreviated “log”) is a quantity representing the power to which a fixed number (the base) must be raised to produce a given number.

*Saana Svärd, Heidi Jauhiainen, Aleksi Sahala, Krister Lindén, 239n44*

→ *Pointwise Mutual Information*

**Logosyllabic** → see General Glossary

### **Machine-actionable/Machine-readable**

A text intended to be processed by machines.

*Doğu Kaan Eraslan, 287n17*

### **Machine-actionable Data/Machine-readable Data**

Data structured in a way that it can be processed by computer software.

*Émilie Pagé-Perron, 200n27*

→ *scraper*

### **Manuel de Codage (mdc)**

The standard developed for processing ancient Egyptian hieroglyphs. mdc, specifically its last version, mdc-88, has seen a lot of variants, and it uses → American Standard Code for Information Interchange (ASCII) characters.

*Doğu Kaan Eraslan, 287*

→ *American Standard Code for Information Interchange characters*

### **Map/Mapping**

In the context of a → dictionary-data structure, it is to assign a value to a key.

*Doğu Kaan Eraslan, 297n55*

→ *dictionary*

### **Markup/Markup Tagging**

Marking up online content can be compared to annotating a hard-copy text by hand (i.e., writing notes with a pen on the paper manuscript)

*Vanessa Bigot Juloux, 163*

→ *annotation, element, Extensible Markup Language annotation, HyperText Markup Language, tag*

A way to annotate information using → tags. → HyperText Markup Language (HTML) is a markup language.

*Émilie Pagé-Perron, 204n43*

### **Maximal Cliques**

They are the largest → cliques identifiable in a network. If a text mentions “a” and “b,” and another text mentions “a,” “b,” and “c,” then “a, b, c” forms a maximal clique, overlapping with the “a-b” clique.

*Émilie Pagé-Perron, 214*

→ *clique, k-plex*

### **Maximally Regular Graph (MRG)**

In the context of the → Mesopotamian Urban Revolution Landscape (MURL) study, MRG represents an alternative graph-theory representation of the underlying power network with respect to the one obtained by simply reconstructing the geography of common board affiliation ... From a → Minimum Spanning Tree (MST), generated from any metric, the MRG reshapes the links among nodes in order to maximize the fundamental and the most regular structures implicated in any dataset ... Compared to the MST, therefore, the MRG adds all (and only) those extra features that are really useful in understanding the prototypes that are hidden in the database; in other words, it adds the optimal amount of complexity that is necessary to read the phenomenon.

*Marco Ramazzotti, 67n28*

→ *Mesopotamian Urban Revolution Landscape, Minimum Spanning Tree*

### **Maximally Regular Graph (MRG) Algorithm**

A new type of semantic → graph that uses a new index for detecting structural/topological complexity information in undirected graphs (H function) ... The MRG → algorithm generates, starting from the → Minimum Spanning Tree (MST), the graph presenting the highest number of regular microstructures that make use of the dataset's most important connections.

*Marco Ramazzotti, 67n28*

→ *algorithm, graph, Maximally Regular Graph, Minimum Spanning Tree*

**Mesopotamian Urban Revolution Landscape** → see General Glossary

### **Metadata**

Information about the text that is external to the textual information itself and can be used for classification purposes. It may describe provenience, period, genre, etc.

*Émilie Pagé-Perron, 199n19*

Metadata provides additional information about the text in question. In Assyriology, this information includes date and provenance.

*Saana Svärd, Heidi Jauhiainen, Aleksi Sahala, Krister Lindén, 225n8*

→ *Cuneiform Digital Library Initiative, Open Richly Annotated Cuneiform Corpus*

### **Meta-model**

A → model that describes a model. More specifically, it is a high-level abstraction that uses modeling language to describe a model.

*Sveta Matskevich and Ilan Sharon, 35n51*

→ *model*

### **Meta-system**

Data description system external to a data management system; a system that describes a system.

*Sveta Matskevich and Ilan Sharon, 38n55*

→ *Database Management System*

### **Metric Scaling**

It is the process of considering each → item as a point in a multidimensional space and calculating the “distance” between the points using a metric, just as you find the distance between points on a map.

*Shannon Martino and Matthew Martino, 120*

→ *item*

### **Minimal Semantic Unit**

The smallest distinctly perceived element. This corresponds to the smallest delimited element in the case of an → encoding scheme, since the perception is done by the machine.

*Doğu Kaan Eraslan, 286n14*

→ *encoding scheme*

### **Minimum Spanning Tree (MST)**

From a conceptual point of view, the MST represents the energy-minimization state of a structure. In fact, if we consider the atomic elements of a structure as the vertices of a → graph and the strength among them as the weight of each → edge linking a pair of vertices, the MST represents the minimum of energy needed so that all the elements of the structure preserve their mutual coherence. In a closed system, all the components tend to minimize the overall energy. So the MST, in specific situations, can represent the most probable state for the system.

*Marco Ramazzotti, 67n28*

→ *edge, graph, Maximally Regular Graph*

### **Minimum Spanning Tree (MST) Data Mining**

A method that permits branching structures that also reveal clustering in archaeological datasets; this method has been applied as a model for → spatial analysis in different ways.

*Marco Ramazzotti, 67n28*

→ *data mining, Minimum Spanning Tree, spatial analysis*

### **Mixed Method**

→ Quantitative and → qualitative methods combined.

*Vanessa Bigot Juloux, 163*

→ *qualitative method, quantitative analysis/method*

### **Model**

It is as an external and explicit representation of a system or a part of reality.

*Sveta Matskevich and Ilan Sharon, 35n51*

→ *meta-model*

**MRG Algorithm** → Maximally Regular Graph algorithm

**MST** → Minimum Spanning Tree

**MST Data Mining** → Minimum Spanning Tree data mining

### **Multivariate Analysis**

One method often utilized by archaeologists to create typologies, but it is not simply one tool that is easily employed in the same way by everyone regardless of the purpose. It is, rather, a general description for a type of analysis that takes many forms. One must always adapt the method to the material that one wishes to analyze. Using a multivariate analysis, archaeologists can determine the types that define a typology through the analysis of multiple variables and their interdependency.

*Shannon Martino and Matthew Martino, 118*

### **MySQL**

A → relational database software that employs the → Structured Query Language (SQL) to fetch data.

*Émilie Pagé-Perron, 196n7*

→ *querying, relational database, Structured Query Language*

### Namespace

In → EpiDoc, it refers to the appellation of the → tags with relation to the information they convey. There could be several options in determining which tag should be used for a personal name, such as John Doe Smith. We could, for example, decide to use “name,” “personalName,” or both. Which appellation better frames the interpretation of the name depends on the context, and, thus, on the project, although it should be kept in mind that the more interpretation one provides, the more time one must invest in constructing a coherent namespace.

*Doğu Kaan Eraslan, 302n72*

→ *EpiDoc, tag*

### Natural Computing (NC)

Nonlinear dynamics and pattern formation computing inspired by concepts, principles, and mechanisms underlying natural systems.

*Marco Ramazzotti, 60n2*

→ *Artificial Adaptive System, Natural Computing algorithms*

### Natural Computing (NC) Algorithms

Inspired by natural and biological phenomena, they include evolutionary → algorithms, → neural networks, molecular computing, and quantum computing.

*Marco Ramazzotti, 60n2*

→ *algorithm, Natural Computing, neural networks*

### Natural Language Processing (NLP)

A broad term that refers to methods for teaching computers to understand human language. At the risk of over-simplifying, the idea is that after a computer has processed a large number of words in a text, it should be able to extract meaning from a text it has never seen. Most work in this field has focused on teaching computers to understand English. The benefits of NLP have not yet reached the study of the languages of the ancient world.

*Miller C. Prosser, 322n29*

NC → Natural Computing

NC Algorithms → Natural Computing algorithms

### Negative Sampling

When using → Word2vec, a method for creating the output vector, which enables a procedure in which only a small part of the model's weights is updated.

*Saana Svärd, Heidi Jauhiainen, Aleksi Sahala, Krister Lindén, 247*

→ *hierarchical softmax, Word2vec*

### Network Analysis

It permits one to approach the data in large amounts of texts through a new lens, by inquiring quantitatively with a focus on the relationships among entities of interest. Such inquiries facilitate the detection of meaningful patterns that could not be easily perceived using traditional methods ... The use of network analysis is well established in some fields, such as biology and sociology, but it is only slowly gaining popularity among Assyriologists ... In textual studies, network analysis is a digital methodology that focuses on the relationships among entities in the written record and enables these relationships to be studied at a larger scale than is normally feasible with traditional philological methods.

*Émilie Pagé-Perron, 194, 194n3, 196*

→ *quantitative analysis/method*

### Network Analysis Algorithms

→ Algorithms specifically geared to → network analysis.

*Émilie Pagé-Perron, 197n16*

→ *algorithm, network analysis*

### Network Graph

The total of all data points, the → nodes, and their relationships—i.e., the → edges. Some nodes can be completely disconnected from others and form independent ensembles. Together, connected and disconnected nodes form the network graph ... Network graphs are composed of nodes and links; nodes are data points representing entities. Edges link nodes, representing the relationships between those nodes. This data takes the form of → “triples” comprising two nodes and an edge that connects them.

*Émilie Pagé-Perron, 201, 201n29*

→ *edge, graph, node, triple*

### Network Graph Analysis

→ An administrative text ... is an ideal candidate for → network graph analysis in which individuals who appear in the text are chosen as entities that will become nodes in the final → graph, and their co-occurrence generates links that connect them together.

*Émilie Pagé-Perron, 201*

→ *administrative text, graph, network graph*

### Network Graph Triples

In a → network graph, data takes the form of → “triples” comprising two → nodes and an → edge that connects them.

*Émilie Pagé-Perron, 201*

→ *edge, network graph, node, triple*

### Network Graph Visualization

When using a corpus in which the texts, such as sub-archives discussing only one type of activity, are homogenous, network graph visualization is an efficient means for giving an overview of the relationships without requiring extensive manipulation of the → graph.

*Émilie Pagé-Perron, 210*

→ *graph, network graph*

### Network Theory

The study of systems and patterns found in → network graphs. → Social network analysis uses network theory to understand social relationships.

*Émilie Pagé-Perron, 195n5*

→ *network graph, social network analysis*

### Neural Networks

A computational model inspired by the human brain where interconnected nodes (neurons) work in parallel to find out how to solve a problem by themselves when given an example.

*Saana Svärd, Heidi Jauhiainen, Aleksi Sahala, Krister Lindén, 229n19*

→ *Artificial Neural Network, Natural Computing algorithms*

**NLP** → Natural Language Processing

### Node

In many cases, nodes represent texts within a corpus or documents within a genre.

*M. Willis Monroe, 260*

In → network analysis, nodes represent entities that can be of one or more types, such as people, institutions, or places.

*Émilie Pagé-Perron, 201*

In the context of → Mesopotamian Urban Revolution Landscape (MURL), nodes are the spatial relations between points (sites) ... Each distance between a pair of nodes becomes the → weighted → edge between this pair of nodes.

*Marco Ramazzotti, 66, 73*

→ *edge, edge betweenness, ego-network, Mesopotamian Urban Revolution Landscape, network analysis, Self-Organizing Map, weighted*

**Normalization**

The process of expressing different phenomena with a common unit. For example, you cannot add apples to oranges, but you can add fruits to fruits. Thus, in a context in which we are interested only in adding fruits, since apples and oranges are fruits, we can add apples to oranges. The passage from apples to fruits is a case of the application of a normalization procedure.

*Doğu Kaan Eraslan, 293n39*

**Normalized Pointwise Mutual Information (NPMI)**

An extended version of the  $\rightarrow$  Pointwise Mutual Information (PMI) formula, which removes some of the low-frequency bias and gives the score fixed upper and lower boundaries.

*Saana Svärd, Heidi Jauhiainen, Aleksi Sahala, Krister Lindén, 240*  
 $\rightarrow$  *logarithm, Pointwise Mutual Information*

**NoSQL**

$\rightarrow$  Database models that utilize a non-tabular or not only tabular data structure. Examples include a document store (such as  $\rightarrow$  Extensible Markup Language),  $\rightarrow$  triple store ( $\rightarrow$  Resource Description Framework), graph databases, and object-oriented databases.

*Sveta Matskevich and Ilan Sharon, 51n98*  
 $\rightarrow$  *database model, Extensible Markup Language, Resource Description Framework, triple, triple store*

**NPMI**  $\rightarrow$  Normalized Pointwise Mutual Information

**OCHRE**  $\rightarrow$  Online Cultural and Historical Research Environment

**OCR**  $\rightarrow$  Optical Character Recognition

**OD**  $\rightarrow$  Open Data

**OM**  $\rightarrow$  OntoMmedia

**Online Cultural and Historical Research Environment (OCHRE)**

The OCHRE database runs on a server professionally supported by the Digital Library Development Center at the Regenstein Library on the University of Chicago campus. All core data is stored in this database. Users access data through the OCHRE  $\rightarrow$  Java application client. This client interface is a  $\rightarrow$  Java Web Start application that launches on any computer with an internet connection. Because OCHRE is an online-database



environment, project members from anywhere in the world have access to data live and in real time ... One of the simple principles underlying the OCHRE → data model is that each discrete meaningful unit of observation is a separate database → item. Each database item is stored as an → Extensible Markup Language (XML) file ... Millions of individual XML files are related one to another through a variety of organizational methods, with the primary among these being the hierarchy.

*Miller C. Prosser, 318–319*

→ *data model, Extensible Markup Language, item, Java application client, Java Web Start, linked*

### **Ontology**

Formalized structures that form part of expert systems. They constitute the underlying data infrastructure for projects, tools, and datasets published as → Linked Data ... Ontologies are used to define the types of entities that occur in the dataset (classes) and the possible relationships between those entities (properties) define the type of entity a subject or object is, and the properties describe predicates, including the directionality of the relationship it captures. The class from which the property runs is the domain; the class to which it runs is the range.

*Terhi Nurmikko-Fuller, 347–348*

→ *Linked Data, upper ontology*

### **OntoMedia (OM)**

Like the → CIDOC Conceptual Reference Model (CIDOC CRM), OM is event-based. It focuses on the representation of the narrative in multi-media and has been designed as linkable to the CIDOC CRM. The aim of this → ontology is to enable the human-like, vague questions we might use in conversation when trying to identify a given story ... OM represents the narrative content of heterogeneous media. It is based largely on two interlinked topics: the literary genres of fantasy and science fiction and the fan-fiction associated with them, both manifesting as a number of genre-specific classes.

*Terhi Nurmikko-Fuller, 357*

→ *CIDOC Conceptual Reference Model, ontology*

### **Open Data (OD)**

Data that has been made publicly available without access restrictions resulting from copyrights or paywalls. It often manifests as raw, or largely unanalyzed datasets, made downloadable in formats such as → Comma Separated Value (csv) files.

*Terhi Nurmikko-Fuller, 344*

→ *Comma Separated Value, Linked Data, Linked Open Data*

### Open Richly Annotated Cuneiform Corpus (Oracc)

Oracc presents itself as a platform hosting sub-projects that are managed by their contributing teams. Oracc uses standardized encoding schemes that are based on the same original format: → ASCII Transliteration Format (ATF).

*Émilie Pagé-Perron, 200*

One of the largest corpora of → Sumerian and → Akkadian texts, consisting of over seventeen thousand texts (almost two million words). Roughly half of it has been annotated ... In general, in the Oracc corpus, the → metadata added to the texts has been done during different projects over a number of years. This is why some texts have more metadata than others. Some texts include information on provenance, the period when they were written, and genre. Individual words may have such → tags as transcription, dictionary form, translation, part of speech, and language.

*Saana Svärd, Heidi Jauhiainen, Aleksi Sahala, Krister Lindén, 225, 228*  
→ Akkadian, annotation, ASCII Transliteration Format, metadata, Sumerian, tag

### Optical Character Recognition (OCR)

A subfield of computer vision that permits machines to extract text characters from a pixel-based image ... OCR has become a viable option to use in epigraphy largely due to the popularization of → deep-learning → algorithms. The passage, however, between the → raster image and → vector graphic, which can be stored in encodings, is actually quite tricky ... The OCR process would actually generate a → features' vector that could be compared against the → abstract vector of the → sign.

*Doğu Kaan Eraslan, 296, 296n50, 297*

A process by which a machine scans a typed or handwritten document and converts the text into digital characters.

*Miller C. Prosser, 323n30*

→ abstract vector, algorithm, deep learning, features' vector, raster image, sign, vector, vector graphic

**Oracc** → Open Richly Annotated Cuneiform Corpus

### Oracc-ATF

Specific characteristics of Oracc-ATF include a wider array of characters permitted in the → transliterations lines and the validation of data content managed at the project level. As a result of these characteristics, transcription standards vary.

*Émilie Pagé-Perron, 200*

→ ASCII Transliteration Format, Open Richly Annotated Cuneiform Corpus, transliteration

**Physical Medium** → see General Glossary

**PMI** → Pointwise Mutual Information

**Pointwise Mutual Information (PMI)**

A statistical method used to find collocations and associations between words. It measures the reduction of uncertainty about the occurrence of *word<sub>1</sub>* when *word<sub>2</sub>* is known to be present within a given distance. Thus, we may be able to generate groups of words that are syntagmatically related to each other ... PMI is a count model whereas → Word2vec uses so-called → Artificial Neural Networks (ANNs) to predict word co-occurrence.

*Saana Svärd, Heidi Jauhiainen, Aleksi Sahala, Krister Lindén, 229, 238*  
→ *Artificial Neural Network, Word2vec*

**Pre-processing**

The task of preparing the (textual) data before starting the → annotation process or some computational analysis.

*Émilie Pagé-Perron, 197m15*  
→ *annotation*

**Protocol**

A set of rules for the exchange or transmission of data between machines.

*Terhi Nurmikko-Fuller, 340n8*

**Qualitative Method**

Qualitative methods enable the collection of data by observation through the researcher's participation in the investigation; in order to reflect the purpose of the analysis, the most relevant data is used for a deeper investigation, such as parsing with → R. This data is recorded and sorted according to its nature in order to match categorical data. Data can be classified into two types: numerical and categorical data. Numerical data is used → quantitatively to measure the numerical values' outcome, and categorical data is used qualitatively to organize nominal values into categories. Categorical data is used for → text and → data mining processes. Qualitative methods do not exclude the quantitative methods because, in some cases, calculations based on significant data types are needed.

*Vanessa Bigot Juloux, 162–163*  
→ *data mining, mixed method, quantitative analysis/method, R, text mining*

### Quantitative Analysis/Method

Quantitative methods allow the researcher to count and measure a group of data; they use mathematical formulas and statistics to express results in terms of numbers or sets of numbers. Graphics or templates are often used to communicate results.

*Vanessa Bigot Juloux, 162*

Although Assyriologists are generally not acquainted with quantitative analysis methods, these techniques are making a remarkable entry into the field ... using quantitative analyses is helpful for Assyriological studies: preparing data in a → machine-actionable way enables researchers to explore a corpus using a large array of alternative and complementary approaches, from simple statistical models to machine learning → algorithms. The results of such methods, including → network analysis, yield exact results that can be counted, compared, and used as strong evidence to build an argument.

*Émilie Pagé-Perron, 216–217*

→ *algorithm, correspondence analysis, graph partitioning, machine-actionable/machine-readable, mixed method, network analysis, qualitative method*

### Quantitative Data

Countable and measurable specimens on which mathematical operations, such as statistical analysis, can be performed.

*Émilie Pagé-Perron, 218n78*

### Querying

The processes of extracting information from a database based on the user's input.

*Doğu Kaan Eraslan, 285n9*

→ *MySQL, Structured Query Language*

### R

An open-source software environment, mostly used for statistics, graphics, and data manipulation. One of the major benefits of R is its popularity, which guarantees regular developments by a broader community of R users.

*Vanessa Bigot Juloux, 182n127*

→ *qualitative method*

**Ras Šamra** → see General Glossary

### Ras Shamra Tablet Inventory (RSTI)

RSTI is an → OCHRE research project co-directed by Miller C. Prosser and Dennis Pardee. A primary goal of the project is to create reliable digital editions of the texts in the

Ras Šamra-Ugarit corpus within a → research database environment. More than just a data store of texts translated from their ancient languages, RSTI serves as a tool for addressing research questions. To this end, the project also seeks to integrate archaeological data from the excavations at → Ras Šamra, including published archaeological plans, grid and square systems, and any other information freely available ... RSTI presents a standard text edition with translation, epigraphic commentary, philological explanations, and interpretation ... Objects, texts, and images are the three main categories of data in RSTI. The database has entries for 5,700 objects, mostly → tablets, but also vessels, seals, axe heads, and other → items. To date we have added 950 text → transliterations. The project currently includes over 32,000 tablet photos and drawings.

*Miller C. Prosser, 314, 317*

→ *item, Online Cultural and Historical Research Environment, Ras Šamra, research database environment, tablet, transliteration*

### **Raster Image**

A matrix of numerical data for the computer called pixels.

*Doğu Kaan Eraslan, 296*

→ *Optical Character Recognition*

**RDF** → Resource Description Framework

### **RDF-XML**

A syntax that can be used to express → Resource Description Framework (RDF) in an → Extensible Markup Language (XML) document.

*Terhi Nurmikko-Fuller, 339*

→ *Extensible Markup Language, Resource Description Framework*

### **Regular Expressions**

A group of characters corresponding to a multitude of characters; thus, the characters of a regular expression are meta-characters. For example, in the Python implementation of regular expressions, the characters \w, \d correspond to an alphabetical character and a digit, respectively. There are various implementations of regular expressions throughout the programming languages.

*Doğu Kaan Eraslan, 293n41*

### **Relational Data Model**

Similar → items are stored in tables of columns and rows, then joined with other tables based on a common column called a key.

*Miller C. Prosser, 319*

→ *item*

### Relational Database

Computer database in which data are organized in tables, where each table contains all the instances of one entity. Each → “tuple” in a table represents one instance of the entity and must have a unique → identifier as one of its attributes; the columns (fields) of the table represent all other attributes of the entity, including keys of related entities. The relationships between the tables are defined as links between keys.

*Sveta Matskevich and Ilan Sharon, 47n84*

Relational databases are a type of data format that uses unique → identifiers (ID) to represent data from one table in another, so that when the user → queries the data using a query language called → Structured Query Language (SQL), it is possible to fetch related information from multiple tables at once.

*Émilie Pagé-Perron, 196n7*

→ *identifier, MySQL, querying, Structured Query Language, tuple*

### Research Database Environment

It expands on the core structure of the database with tools and other features to help users work with their data.

*Miller C. Prosser, 314n2*

→ *Ras Shamra Tablet Inventory*

### Resource Description Framework (RDF)

An abstract → data model with roots in → metadata modelling, it is used to represent information, knowledge, and data entities in a → graph structure. This approach captures the relationships between entities by representing these entities and the connections between them using → HyperText Transfer Protocol (HTTP) URIs ... As a flexible data structure, RDF is both powerful and capable of capturing nuanced and domain-specific relationships between data entities. Information stored in this way can also be restructured, manipulated, and edited with comparative ease without the extensive restructuring that a → relational database would require. RDF as a technology imposes few limitations on the type of information that can be captured and represented, whether at the level of entities (things, people, places, concepts, notions) or at the level of the relationships between them. All RDF data structures are based around a data cluster of three components, known as a → “triple.”

*Terhi Nurmikko-Fuller, 339, 345*

→ *data model, graph, HyperText Transfer Protocol Universal Resource Identifiers, meta-data, relational database, Simple Knowledge Organization System, triple*

**RSTI** → *Ras Shamra Tablet Inventory*

### Scalable Vector Graphics (SVG)

→ Extensible Markup Language (XML)-based language for describing 2D graphics. SVG presents the most viable option for storing local glyph representations. SVGs are a joint product of the web development and design communities, making them reliable for expressing any type of 2D graphics that require cross-platform exchange. There are also different types of cross-platform open-source software for visualizing SVGs. The XML nature of SVGs implies a relatively simple integration with → EpiDoc.

*Doğu Kaan Eraslan, 304*

→ *EpiDoc, Extensible Markup Language, vector*

### Scalar Multiplication

A multiplication operation between a → vector and a number that belongs to the same set of numbers contained in the vector.

*Doğu Kaan Eraslan, 284n6*

→ *vector*

### Scraper

A type of software that selectively retrieves information from a website when a → machine-actionable version of the data is not available ... Nick Veldhuis designed a scraper specialized in extracting the → lemmata assigned to each → token of a text annotated in → Open Richly Annotated Cuneiform Corpus (Oracc). When used on an already annotated text, this type of tool will make it easier to gather specific types of entities: for instance, people in a text are → tagged with an indicator that the word is a personal name: “ [ PN ] ”.

*Émilie Pagé-Perron, 204, 204n41*

→ *annotation, lemma, machine-actionable/machine-readable, Open Richly Annotated Cuneiform Corpus, tag, token*

### Self-Organizing Map (SOM)

Self-Organizing Maps are a type of → Artificial Neural Network (ANN); they have had great success and a number of concrete applications in many different fields. The logic underlying this type of ANN is the progressive distribution of data in an abstract map to form groups (potentially classes or sub-classes), which can show, based on their measurable distances, the degree of similarity that has been recognized among them.

*Alessandro di Ludovico, 94n26*

A type of learning → algorithm used to draw inferences from datasets and to order high-dimensional statistical data so that neighboring → nodes on the map represent similar inputs. Often the SOM is applied to numerical data in application areas such as pattern recognition, signal processing, and → multivariate statistical analysis. In other

words, the SOM is an → unsupervised type of network that offers a classification of the input vectors, creating a prototype of the classes and a projection of the prototypes on a map having two dimensions (but  $n$ -dimensional maps are also possible) that is able to record the relative proximity (or neighborhood) between the classes.

*Marco Ramazzotti, 65n20*

→ *algorithm, Artificial Neural Network, multivariate analysis, node, unsupervised method*

### Semantics of a Recording System

In archaeological context, this is what is being recorded, irrespective of syntax. Only two types of archaeological entities are common across the whole spectrum of existing recording methods, systems, and schools. The two mandatory pieces of information that must be recorded by an excavator in every system are a find designator and the spatial unit from which that find originated. Many systems have more than one unit of either type.

*Sveta Matskevich and Ilan Sharon, 38*

→ *arbitrary unit*

### Semantic Technologies

A heterogeneous collection of tools and → algorithms that can be used to bring structure to information. These include → Natural Language Processing, → data mining, → Artificial Intelligence, and a number of other approaches and methodologies.

*Terhi Nurmikko-Fuller, 344*

→ *algorithm, Artificial Intelligence, data mining, Natural Language Processing*

### Semantic Web (sw) Technologies

A family of technologies and standards provided by the → w3c for the purposes of describing and connecting information online. A by-no-mean-exhaustive list of these technologies includes → Resource Description Framework (RDF), RDF Schema (RDFS), Web Ontology Language (OWL), and SPARQL.

*Terhi Nurmikko-Fuller, 344*

→ *Resource Description Framework, Simple Knowledge Organization System, web, World Wide Web Consortium*

### Shading Attribute

It corresponds to a feature of the → Manuel de Codage (Mdc) encoding scheme: for example, A1\shading13, which would add diagonal lines to the first half of the → sign, indicating damage on that particular area.

*Doğu Kaan Eraslan, 295n46*

→ *Manuel de Codage, sign*



**Sign** → see General Glossary

### **Simple Knowledge Organization System (skos)**

A model for expressing vocabularies, thesauri, and taxonomies in a → machine-readable format, namely → Resource Description Framework (RDF). SKOS is a → World Wide Web Consortium (w3c) standard employed for over a decade for linking data in the → Semantic Web. Controlled vocabularies and taxonomies are essential for efficient data retrieval in archaeological practice in general, and field recording in particular. Presented in SKOS format, they can be integrated into the → Linked Data cloud to serve all participant datasets.

*Sveta Matskevich and Ilan Sharon, 46*

→ *data cloud, Linked Data, machine-actionable/machine-readable, Resource Description Framework, Semantic Web, web, World Wide Web Consortium*

**skos** → Simple Knowledge Organization System

### **skosifying**

A process of → mapping user-defined terminology (thesaurus or vocabulary) to → Simple Knowledge Organization System (SKOS).

*Sveta Matskevich and Ilan Sharon, 48n91*

→ *map/mapping, Simple Knowledge Organization System*

### **Social Network Analysis**

Its main task is to explore social relations, and the → network graph data structure is the most appropriate data type to focus on relationships between entities ... The current practice in social network analysis based on cuneiform sources uses verbs to create directed links between entities or concepts.

*Émilie Pagé-Perron, 195, 197*

→ *administrative text, network graph*

**som** → Self-Organizing Map

**SPAD** → Système Portable pour l'Analyse des Données

**Spatial Unit** → see General Glossary

### **Special Characters**

Exemplified by “\*, \_ , =,” as opposed to alphanumeric characters such as “a, b, 1, 2,” etc. ... Special characters do not solely express a particular semantic or syntactical field of the language. Some, for example, express phonetic values or diacritics ... Special

characters such as “\_” or “\$” are reserved for indicating → sign functions and information about the state of conservation of the script and its → physical medium.

*Doğu Kaan Eraslan, 285, 286n13, 288*

→ *CAL Code, physical medium, sign*

**SQL** → *Structured Query Language*

**Structured Query Language (SQL)**

It is used to → query → relational databases to fetch data.

*Emilie Pagé-Perron, 196n7*

→ *MySQL, querying, relational database*

**Sumerian** → see General Glossary

**Supervised Method**

Method in which the researcher has chosen to “classify” part of the material before the analysis. The researcher is supervising, commonly called *training*, the → algorithm by providing it with known data prior to performing analysis on the unknown material.

*M. Willis Monroe, 260, 260ng*

→ *algorithm, unsupervised method*

**SVG** → Scalable Vector Graphics

**Syntactic Elements**

The components of the record itself, irrespective of what is being recorded. The most basic of these is the individual recording → event.

*Sveta Matskevich and Ilan Sharon, 36*

→ *event*

**Système Portable pour l'Analyse des Données (SPAD)**

A software for statistical analysis that has a → Graphical User Interface and includes a large series of different packages and functions. With the tools included in SPAD, one can perform different types of analyses and also manage data and whole datasets, with the aim, for example, of investigating their structure or preparing them to be processed. It is a proprietary software offering user interfaces in French. SPAD was first conceived as academic software, but after 1987 it became a commercial product developed by CISIA/DECISIA.

*Alessandro di Ludovico, 97n35*

→ *data mining, Graphical User Interface*

**Tablet** → see General Glossary

### Tag

When using → Extensible Markup Language (XML) or → Text Encoding Initiative (TEI), it is rather an indication of the classification of what is described: after a tag “word,” one either has a verb, adverb, noun, or adjective, under its inflectional form or not.

*Vanessa Bigot Juloux, 163*

In → EpiDoc, the appellation of tags filling the role of → markups, as in the case of → Extensible Markup Language (XML) documents, must be decided in the process of developing the → namespace. Deciding what a tag should be called, however, does not necessarily result in its coherent usage. Taking the example of the name John Doe Smith, we would not know whether the name belongs to a real person or a fictional character. If we were to distinguish this aspect of the names in a document, the distinction would not come from the domain of the namespace, since it is possible for the same name to belong to both real and fictional persons. Rather, the differentiation would need to be made through the manner in which we apply the tags. Therefore, it is essential that tag usage is a planned part of a project.

*Doğu Kaan Eraslan, 302n72*

→ *attribute, EpiDoc, Extensible Markup Language, markup/markup tagging, namespace, Text Encoding Initiative*

**TEI** → Text Encoding Initiative

### Term Frequency–Inverse Document Frequency (TF-IDF)

A quite popular method that evaluates the importance of terms within one text inverse to their frequency in the entire corpus, essentially identifying the terms that are most important for identifying a document. This method is based on frequency counts for terms within documents and within the corpus as a whole. It works well when there are many terms shared among all documents, as it weeds out common terms and focusing on the important rare terms.

*M. Willis Monroe, 271–272n34*

### Text Data

A glyph, a word, a group of words, a reference, or a concept.

*Vanessa Bigot Juloux, 165*

### Text Encoding Initiative (TEI)

TEI was created in 1987. Then, in 2000, the TEI Consortium (TEI-C) was established. The TEI-C is a group of international scholars who collaborate on the development of a

dedicated encoding standard for text analysis. Unfortunately, using TEI is a “paradox,” since it does not yet enable → interoperability, but rather the interchange of cross-corpora text analysis.

*Vanessa Bigot Juloux, 164*

→ *attribute, element, EpiDoc, interoperability/epigraphic interoperability*

### **Text Mining**

Its purpose is to find patterns that match shared analytical variables in order to suggest a first interpretation of the transcription.

*Vanessa Bigot Juloux, 181*

→ *qualitative method*

**TF-IDF** → Term Frequency–Inverse Document Frequency

### **Token**

A sentence or → sign, but, in this case, the texts are divided at the word level. Tokens are then assigned to → lemmata, which are also called “dictionary entries,” in a process called “lemmatization.”

*Émilie Pagé-Perron, 203*

→ *lemma, sign*

### **Tokenization**

The tokenization process consists of splitting the group of words into individual words. Each word is sorted into a category, the equivalent of a → Text Encoding Initiative (TEI) → attribute value.

*Vanessa Bigot Juloux, 166, 203*

It is the segmentation of a text into similar units called → tokens ... in a process by which occurrences of a word, or other units (such as → signs or sentences), are separated as units from the original textual data. Working from → transliterations, the major separator to take into account is of course the space character.

*Émilie Pagé-Perron, 204n41*

→ *attribute, sign, Text Encoding Initiative, token, transliteration*

### **Topological Weighted Centroid (twc)**

A topological approach with a set of ordered mathematical quantities that transforms the discrete dataset into different scalar fields: Alpha, Beta, Gamma, Theta, and Iota.

*Paolo Massimo Buscema, in Marco Ramazzotti, 74*

### Topological Weighted Centroid (twc) Mathematical Approach

It explores the natural and anthropic landscape with respect to certain quantities (entropy) spatializing the optimal solutions in the centers of the masses.

*Marco Ramazzotti, 70n30*

→ *Topological Weighted Centroid*

**Transliteration** → see General Glossary

### Triple

*It has two distinct meanings:*

1. Triples consist of three parts: a SUBJECT, a PREDICATE, and an OBJECT. These are used to define entity types and directional relationships between them, both at the general (or schema) and specific (instance) level.

*Terhi Nurmikko-Fuller, 345*

2. It always has two → nodes and an → edge (individual 1, individual 2, and the relationship between them).

*Émilie Pagé-Pérroon, 202*

→ *edge, network graph triples, node, triple store*

### Triple Store

A database for the storage and retrieval of → triples.

*Sveta Matskevich and Ilan Sharon, 46n82*

→ *triple*

### Tuple

An ordered list or row in a table.

*Sveta Matskevich and Ilan Sharon, 47n84*

**twc** → Topological Weighted Centroid

**twc Mathematical Approach** → Topological Weighted Centroid Mathematical Approach

### Unicode

A system to map characters to bytes through code points. Characters are treated as semantic units, and code points are arbitrary constructions. The → mapping requires two different reversible procedures that make the system rather complex. Most modern platforms use a Unicode-based encoding scheme, since it was designed to support every character used in every language.

*Doğu Kaan Eraslan, 300–301*

→ *map/mapping, Unicode Consortium, Unicode Transformation Format-8*

### Unicode Consortium

Computers must use an agreed-upon system of representing characters. The Unicode Consortium was created to promote this standard. One of the most significant contributions of this group is the creation of a standard that defines which underlying computer code is used to represent (nearly) every character in every writing system, even → Ugaritic. The number of each character can be referred to as a → Unicode point.

*Miller C. Prosser, 328n34*

→ *Ugaritic, Unicode*

### Unicode Transformation Format (UTF)-8

“8” refers to the use of 8-bit (a numerical value that equals 0 or 1) sequences to represent a character. It is a method for encoding → Unicode characters. There is also UTF-16 and UTF-32, which have, respectively 16- and 32-bit character sequences.

*Doğu Kaan Eraslan, 289n25*

→ *EpiDoc, Unicode*

### Universal Resource Identifiers (URIs)

These clusters of letters and characters form unique → identifiers for each and every page on the → web ... On a conceptual level, these URIs do not differ from any other unique identifier, such as an ISBN for a book, a national identity number for a person, or an object identifier in a museum, but they are unique on the scale of the entire web ... When utilized consistently and documented appropriately, URIs can and are used by large numbers of data publishers (people and institutions who make data available online) to show that completely disparate and unconnected pages contain information about the same entity or resource.

*Terhi Nurmikko-Fuller, 338–339*

→ *HyperText Transfer Protocol, identifier, HyperText Transfer Protocol Universal Resource Identifiers, web*

### Unsupervised Method

Classification method that involves not classifying any of the existing corpus, but, rather, letting the → algorithm decide which records out of the entire corpus belonged together ... The benefit of the unsupervised method here, however, is that it breaks what are otherwise researcher-oriented biases and presents an “objective” view of the material, free from prior classification. Perhaps the unsupervised method detected that a group of novels thought to belong to one author most closely match the unknown material in our example. The conclusion is up to the researcher here, but the unsupervised method can offer insight into a dataset without any preconceived notions of structure and definition. Unsupervised methods are commonly used to explore a dataset, after which a → supervised method can be used to hone in on a particular area of interest.

*M. Willis Monroe, 260–261*

→ *algorithm, Self-Organizing Map, supervised method*

### Unweighted Pair Group Method with Arithmetic Mean (UPGMA)

A → clustering method that is relatively simple in its procedure. It starts by looking for the two closest cells within the adjacency matrix and merging those into a cluster to which it assigns the average of their distance. It then looks for the next two closest cells, or, if the next cell is closest to the newly created cluster, it merges those and assigns it a new average distance. This process is completed for the entire adjacency matrix, merging close items whether they are cells or newly created clusters.

*M. Willis Monroe, 275*

→ *clustering method*

UPGMA → Unweighted Pair Group Method with Arithmetic Mean

### Upper Ontology

A highly generic, non-specifying schema of data categories applicable for use across many research domains. The categories of data types, such as Location, Person, Text, and Resource, are very broad. Each category of data is slightly different from every other, both in conceptual definition and in practical implementation. These data types are presented to the user as different hierarchies.

*Miller C. Prosser, 320*

→ *ontology*

URIs → Universal Resource Identifiers

### Vector

Mathematically speaking, for example in the context of → Scalable Vector Graphics (svg) and → Optical Character Recognition (OCR), a vector is anything that represents something that has a size and a direction. Size is the distance between the origin and end points, and direction is represented by the angle that a line makes with an axis.

*Doğu Kaan Eraslan, 284n6*

→ *Optical Character Recognition, Scalable Vector Graphics, scalar multiplication, vector addition, vector graphic, vector space*

### Vector Addition

It simply entails adding two or more → vectors together.

*Doğu Kaan Eraslan, 284n6*

→ *vector*

### Vector Graphic

A group of lines passing through points defined in a coordinate system.

*Doğu Kaan Eraslan, 297*

→ *Optical Character Recognition, vector*

### Vector Space

In linear algebra, an additive group that is associated with a field of real numbers and permits → vector addition and → scalar multiplication.

*Doğu Kaan Eraslan, 284n6*

→ *scalar multiplication, vector, vector addition*

### W3C → World Wide Web Consortium

### Web

A hypermedia system of pages, sites, and other types of resources, accessed and connected through the use of → hyperlinks ... Although its current manifestation is one of interconnected pages, the three technologies that lie at the heart of the web—→ HyperText Markup Language (HTML), → HyperText Transfer Protocol (HTTP), and → Universal Resource Identifiers (URIs)—are also at the core of → Linked Data (LD) ... The web currently consists of over 4 billion pages, all connected internally (within the page and a site) and externally (to other pages on other sites) through hyperlinks.

*Terhi Nurmikko-Fuller, 338*

→ *hyperlinks, HyperText Markup Language, HyperText Transfer Protocol, Linked Data, Universal Resource Identifiers, World Wide Web Consortium*

### Weight

The weight of an → edge represents here the quantity of co-occurrences of two individuals in the corpus.

*Émilie Pagé-Perron, 201*

→ *edge, weighted*

### Weighted

Weighted relationships can be highlighted in different ways. In → Gephi, one can easily make the edges thicker based on the numeric value of their → weight. In → Cytoscape, a practical way to display the frequency or weight of an → edge is either by applying a color range or by changing the thickness of the edge, with the variation depending on the weight value.

*Émilie Pagé-Perron, 211n61*

→ *Cytoscape, edge, Gephi, weight*

### Wizard

Wizards aid us in a variety of tasks that help us address a research question. There are currently three wizards for philological analysis: (1) the → lexicography wizard; (2) the prosopography analysis wizard; and (3) the gazetteer wizard.

*Miller C. Prosser, 329*

→ *lexicography, workflow wizard*



### Word Sense Induction

Also called word sense discrimination or word sense discovery, it is the task of determining what meaning a word may have in different contexts.

*Saana Svärd, Heidi Jauhiainen, Aleksi Sahala, Krister Lindén, 226–227*

### Word2vec

Developed by Tomas Mikolov and his team in 2013 ... Word2vec is an open-source toolkit for producing word vectors and → querying semantic relationships between words ... It uses so-called → Artificial Neural Networks to predict word co-occurrence ... There are two alternative models one can choose from when using Word2vec: the → Continuous Skip-gram model and the → Continuous Bag-of-Words (CBOW) model.

*Saana Svärd, Heidi Jauhiainen, Aleksi Sahala, Krister Lindén, 229, 246–247, 247n57 → Artificial Neural Network, Continuous Bag-of-Words model, Continuous Skip-gram model, Natural Language Processing, Pointwise Mutual Information, querying*

### Workflow Wizard

An interactive database tool that guides the user through a series of common actions.

*Miller C. Prosser, 314n3*

→ *wizard*

### World Wide Web Consortium (w3c)

w3c is an acronym referring to the World Wide Web Consortium, the international standards organization for the World Wide Web.

*Terhi Nurmikko-Fuller, 340n7*

→ *web*

### Writing System

In → Online Cultural and Historical Research Environment (OCHRE) we have created a writing system that represents the → logosyllabic → cuneiform writing system used in the ancient Near East ... Like the list of letters that represents the English alphabet, this list represents all the known → signs from ancient languages such as → Sumerian and → Akkadian.

*Miller C. Prosser, 321*

→ *Akkadian, cuneiform, logosyllabic, Online Cultural and Historical Research Environment, sign, Sumerian*

**XDR** → External Data Representation

**XML** → Extensible Markup Language

**XML Annotation** → Extensible Markup Language annotation

### **XML-Elamite Standard**

Developed by Marc Bavant, it is tailored for conserving → Elamite texts along with their → transliterations and translations. It is based on → Unicode characters, and it follows widely used conventions in Elamite scholarship, making it → human readable.

*Doğu Kaan Eraslan, 287*

→ *Elamite, human readable, transliteration, Unicode*

### **General\***

(including terms associated with archaeology, history, geography, literature, and philology)

### **Adab Corpus**

It dates to the third millennium BCE and comprises all texts from Adab. The texts include both those from official excavations and those that are lacking provenance but have been attributed to Adab based on factors such as the shape of the → tablet, prosopography, and the institutions and rulers mentioned. The large majority of the texts are → administrative in nature.

*Émilie Pagé-Perron, 210n60*

→ *administrative text, tablet*

### **Administrative Texts**

They mostly document transactions involving people, things, actions, and places ... Administrative texts are typically short and → formulaic, with simple sentences and few actions ... Despite comprising the most numerous type of surviving → cuneiform documents, administrative texts are the least-annotated genre of → Mesopotamian sources ... Because they also remain untranslated, they have the additional drawback of being inaccessible to specialists in adjacent disciplines who are not trained to read cuneiform.

*Émilie Pagé-Perron, 195*

→ *Adab corpus, annotation, cuneiform, formulaic, Mesopotamia*

---

\* Although all words of an expression are capitalized in the list of terms below, when indicated after the “see also” arrow, words are capitalized when the expression has an acronym (for example, Electronic Distance Measurement, EDM) but not capitalized when the expression does not have an acronym. The “definitions” provided in this glossary are neither exact nor comprehensive; rather, they are informative statements extracted from the chapters. In some cases, minor modifications (such as verb tense, or the addition of “it is”) have been made in order for the statement to make sense out of context. When multiple statements are presented for a term, they are listed according to the alphabetical order of their authors’ names. Cross-references point both to entries in this glossary as well as in the CyberResearch Glossary.

## Akkadian

The Akkadian → writing system allowed for variation in the spelling of a given word, even among words with the same grammatical properties.

*Miller C. Prosser, 322*

A Semitic language in use from about 2500 until 500 BCE ... The script uses both → syllabic and logographic signs, and each → sign can be read in many different ways. Akkadian also uses complex inflection; that is, words are modified in order to express various grammatical categories ... For example, *šarrum*, “king,” for the singular nominative form of the noun and *šarram* for the singular accusative form.

*Saana Svärd, Heidi Jauhiainen, Aleksi Sahala, Krister Lindén, 224n3, 227*

→ *logosyllabic, sign, syllabic sign, writing system*

## Allograph

A variant form of the letter, exemplified in most modern alphabetic systems by uppercase and lowercase letters.

*Miller C. Prosser, 321n26*

→ *script unit*

## Annotation

In corpus linguistics, an annotation is a comment that specifies the various linguistic features of a word.

*Saana Svärd, Heidi Jauhiainen, Aleksi Sahala, Krister Lindén, 225n8*

## Aramaic

An ancient Northwest Semitic language attested throughout eastern Turkey, northwest Iran, Iraq, Syria, the Levant, and parts of Arabia at the beginning of the 1st millennium BCE and onwards. Its descendants, such as Mandaic, are still spoken today.

*Doğu Kaan Eraslan, 288n20*

**Arbitrary Unit** → see CyberResearch Glossary

## Archaeological Site

An accumulation of archaeological remains spatially related to each other.

*Sveta Matskevich and Ilan Sharon, 48*

## Area

A subdivision of the → archaeological site (usually several contiguous grid squares, encompassing at least one, possibly several, architectural features).

*Sveta Matskevich and Ilan Sharon, 48*

→ *archaeological site*

**Attested Form**

A form of the → lemma as it occurs in a text, represented by a sequence of → signs.

*Miller C. Prosser, 322*

→ *lemma, sign*

**Attribute (for digital practice)** → see CyberResearch Glossary

**Attribute (in Archaeology)**

It is a minimal characteristic of an artifact such that it cannot be further subdivided.

*Sveta Matskevich and Ilan Sharon, 33n37*

**Ba'lu and 'Anatu Cycle (KTU 1.1–6)**

A narrative story, traditionally viewed as a “myth” by scholars, about the fight between two clans over the throne. Written in alphabetic → cuneiform on six double-sided clay → tablets, the text is attributed to the scribe 'Ilimilku the Šubbanite of → Ugarit.

*Vanessa Bigot Juloux, 152*

→ *cuneiform, tablet, Ugarit*

**Basic Volume Unit (BVU)**

In archaeological context, the primary excavation unit of a recording system, defined arbitrarily (grid square), stratigraphically (→ locus, context), or behaviorally (feature).

*Sveta Matskevich and Ilan Sharon, 49n93*

→ *locus*

**BVU** → Basic Volume Unit

**Colophon**

A brief statement on a text, such as a → cuneiform → tablet, that captures the equivalent of what could now be considered bibliographical → metadata and can contain information such as the name of the authoring individual.

*Terhi Nurmikko-Fuller, 345n23*

→ *cuneiform, metadata, tablet*

**Composite Texts**

Composites may represent the content of a single instance of a → tablet of any degree of preservation or completeness, or of any number of witness tablets, numbering in some cases in the hundreds.

*Terhi Nurmikko-Fuller, 342*

→ *tablet*

### Cuneiform

It is widely accepted that cuneiform writing was invented to express the → Sumerian language. The → writing system was later adopted by other cultures to express their own languages.

*Miller C. Prosser, 316n13*

The cuneiform script was originally developed by Sumerians c. 3000 BCE. In the following centuries, the script was adapted for → Akkadian. It was also used to write several other ancient Near East languages, such as → Hittite and → Elamite. Although passing centuries and new languages caused the form of the → signs to change, the basic shapes—triangular wedges left by a reed stylus—stayed the same.

*Saana Svärd, Heidi Jauhiainen, Aleksi Sahala, Krister Lindén, 224n3*  
→ *Akkadian, Elamite, Hittite, sign, Sumerian, tablet, Ugaritic, writing system*

### Daily Top-plan

Daily graphical record of an excavation area or part of it. The background plan shows all architectural features extant in the area on that particular day. The foreground information reflects how the excavation proceeded during the day. It comprises several types of data, such as outlines of excavation units, elevations, and the registration numbers and coordinates of finds.

*Sveta Matskevich and Ilan Sharon, 26n10*

### Demircihüyük

Site located in northwest Turkey near modern-day Eskişehir. It has remains from the Neolithic to the second millennium BCE.

*Shannon Martino and Matthew Martino, 124*

### Dor (Tel Dor)

A major port town on the Mediterranean, between present-day Tel Aviv and Haifa in Israel. It was active from the Middle Bronze Age to the Roman era.

*Sveta Matskevich and Ilan Sharon, 25*

### Ecofact

A non-artifactual organic (botanical or biological) or inorganic (result of geological processes) object recovered from an archaeological context.

*Sveta Matskevich and Ilan Sharon, 42n68*  
→ *archaeological site*

**EDM** → Electronic Distance Measurement

**Elam**

Ancient civilization existing from the third into the mid-first millennium BCE based in the southwest region of the Iranian plateau, just northeast of the Persian Gulf (in the modern provinces of Ilam and Khuzestan).

*Doğu Kaan Eraslan, 287n16*

**Elamite**

An ancient language used in → Elam.

*Doğu Kaan Eraslan, 287n16*

→ *cuneiform, Elam*

**Electronic Distance Measurement (EDM)**

An instrument that uses reflected light (infrared, laser, microwave) to calculate distances to remote objects; the models produced in the 1970s through the 1990s were mounted on top and externally connected to a theodolite. Newer models, in which a digital theodolite and an EDM are integrated within the same instrument, are known as “Total Stations.”

*Sveta Matskevich and Ilan Sharon, 27n15*

**Formulaic**

Administrative texts of the → Old Akkadian period are formulaic in the sense that they mostly consist of lists of people, things, or both, with verbs that have specific, technical meanings in particular contexts. There is little deviation from the usual formula.

*Émilie Pagé-Perron, 201n32*

→ *Old Akkadian period*

**Glyph**

A graphical representation, such as a character or an accent. In → Ugaritic → cuneiform script, it is one or several carved lines or impressed wedges expressing a letter, syllable, logogram, or a separation mark (a dot) between two words.

*Vanessa Bigot Juloux, 165n69*

The most basic part of the visual representation of an elementary semantic unit. In the case of an Egyptian hieroglyph, for example, a glyph could be a sitting man or a bird. For → cuneiform, a glyph would be a horizontal, vertical, or diagonal line comprising a cuneiform → sign, and for alphabetic languages, such as Greek and Aramaic, a glyph corresponds to a visual representation of a letter, such as an alpha.

*Doğu Kaan Eraslan, 291n32*

→ *cuneiform, sign, Ugaritic*

**Grapheme**

→ Syllabic → sign or the alphabetic letter.

*Miller C. Prosser, 323*

→ *sign, syllabic sign*

**Hittite**

An Indo-European language of Anatolia (Asia Minor) written with → cuneiform → signs.

*Saana Svärd, Heidi Jauhiainen, Aleksi Sahala, Krister Lindén, 227n11*

→ *cuneiform, sign*

**Intermediary Elements**

Vehicle between what we see in the → physical medium and what we understand. This corresponds, for example, to the alpha-numeric characters that are employed to represent the → glyph in the physical medium.

*Doğu Kaan Eraslan, 291n31*

→ *glyph, physical medium*

**KTU 1.1–6** → Ba'lu and 'Anatu Cycle

**Layer**

In Archaeology, a unit of sediments in a stratified → archaeological site, created as a result of one of the site-formation cycles.

*Sveta Matskevich and Ilan Sharon, 42n67*

→ *archaeological site, stratum, tell*

**Lemma** (pl. lemmata)

A lemma is the headword used in a dictionary entry.

*Émilie Pagé-Perron, 203n38*

In → OCHRE, a lemma is further defined by → phonemic forms, which in turn are defined by → attested forms.

*Miller C. Prosser, 321–322*

→ *attested form, Online Cultural and Historical Research Environment, phonemic form*

**Lexicography**

The practice of compiling glossaries and dictionaries.

*Miller C. Prosser, 329*

**Locus** (pl. loci)

The basic → spatial unit in most recording systems in Near Eastern Archaeology. The exact denotation of this term (i.e., the archaeological entity that is modeled by the abstract unit) can vary among different systems ... The de facto standard definition of the “locus” in Levantine Archaeology today on sites with architecture is a contiguous segment of the site volume interpreted as being the result of a single depositional event.

*Sveta Matskevich and Ilan Sharon, 26n9, 40*

→ *spatial unit*

**Logosyllabic**

A → writing system that uses characters—or → “signs”—that can represent words or syllables. For the → Akkadian logosyllabic writing system, we call these word-signs logograms and the syllable-signs phonograms. Thus this type of writing is called logosyllabic, a combination of logograms and syllables.

*Miller C. Prosser, 316n11*

→ *Akkadian, sign, syllabic sign, writing system*

**Mesopotamia**

A region situated within the Tigris–Euphrates river system, in modern days roughly corresponding to most of Iraq and the eastern parts of Syria. A number of ancient cultures and political units flourished in the area.

*Saana Svärd, Heidi Jauhiainen, Aleksi Sahala, Krister Lindén, 224n2*

→ *Mesopotamian civilizations*

**Mesopotamian Civilizations**

Mesopotamian civilizations are known through their texts, mostly written in → Akkadian and → Sumerian, as well as through archaeological finds from the region.

*Saana Svärd, Heidi Jauhiainen, Aleksi Sahala, Krister Lindén, 226*

→ *Akkadian, Mesopotamia, Sumerian*

**Mesopotamian Urban Revolution Landscape (MURL)**

Historical period and economic concept of the Urban Revolution, as well as the area of the world's first urbanism—central-southern Iraq.

*Marco Ramazzotti, 62n10*

→ *Mesopotamia, Mesopotamian civilizations*

**Metadata** → see CyberResearch Glossary



### Micro-zodiac Text

The Micro-zodiac material comes from Hellenistic Babylonia, where it was written by scribes educated and working in the Late Babylonian scholarly communities of southern → Mesopotamia ... The corpus of the Micro-zodiac texts consists of 15 known → tablets of varying degrees of preservation. The most complete have more than 20 cells (the unit of analysis for this paper) of ingredients on them.

*M. Willis Monroe, 259n4, 261*

→ *Mesopotamia, Mesopotamian civilizations, tablet*

**MURL** → Mesopotamian Urban Revolution Landscape

### Neolithic Package

A collection of technologies and beliefs that was once thought to be wholeheartedly and without exception adopted along with agriculture.

*Shannon Martino and Matthew Martino, 115*

### Old Akkadian Period

Approximately from 2340 BCE to 2200 BCE: a period in → Mesopotamian history during which this region saw the development of an important political entity that, at its apogee, extended almost from the Mediterranean in the west, to the Persian Gulf in the south, and to the Zagros mountains in the east.

*Émilie Pagé-Perron, 196n10*

→ *Mesopotamia, Mesopotamian civilizations*

### Old Persian

An Indo-Iranian language descended from Old Iranian. It was used mostly, but not exclusively, in royal inscriptions in the Achaemenid Empire (550-330 BCE).

*Doğu Kaan Eraslan, 301n70*

**Online Cultural and Historical Research Environment (OCHRE)** → see CyberResearch Glossary

### Phonemic Form

It represents a grammatical interpretation of an → attested form. In many cases, a given phonemic form is represented by many attested forms.

*Miller C. Prosser, 322*

→ *attested form*

**Physical Medium**

An object on which we observe the → signs. It can be papyrus, a vase, a stone surface, etc.

*Doğu Kaan Eraslan, 285n8*

→ *sign*

**Ras Šamra**

The → archaeological site of Ras Šamra is located on the eastern Mediterranean, near Lattakia, Syria. The ancient name for this site—and for the surrounding kingdom—was → Ugarit. Ugarit was well situated, with access to trade routes, both land and sea, and to arable lands. The site was occupied almost continuously from the Neolithic period (c. 7000 BCE) through its eventual destruction in the twelfth century BCE during a regional period of instability. Later, Greek, Persian, and Roman garrisons occupied the site.

*Miller C. Prosser, 315*

→ *archaeological site, Ugarit*

**Raw Data of an Excavation**

A collection of interpretations fixed in pre-printed forms, diaries, and tables.

*Sveta Matskevich and Ilan Sharon, 44*

**Reading**

A value represented by a → sign [as part of a → script unit].

*Miller C. Prosser, 321n26*

→ *script unit, sign*

**Script Unit**

It is more than just a letter or a → sign in the → writing system. Each script unit is defined by various → readings and → allographs.

*Miller C. Prosser, 321*

→ *allograph, reading, sign, writing system*

**Sign**

Elementary semantic unit comprising alphabetical and non-alphabetical languages.

*Doğu Kaan Eraslan, 285n8*

**Spatial Unit**

A locational unit of excavation defined in absolute coordinates, relatively to the site grid, or to architectural features and other units in its vicinity.

*Sveta Matskevich and Ilan Sharon, 26n9*

→ *arbitrary unit*

### Static Archaeological Record

One of the keywords of the middle-range theory in Archaeology; it refers to all findings and the fact that, although uncovered almost simultaneously and in close physical proximity, they represent some diachronic reality.

*Sveta Matskevich and Ilan Sharon, 31n26*

### Stratum

A single construction–destruction cycle within a stratified sequence of deposits that form a multi-layered site (e.g., → tell).

*Sveta Matskevich and Ilan Sharon, 40n61*

→ *archaeological site, layer, tell*

### Sumerian

In the context of literary Sumerology, “Sumerian” is to be understood as an exclusively linguistic label, referring to the language of the compositions themselves, rather than being indicative of the ethnic or socio-cultural identity of the authoring scribe or the patron commissioning the piece. Data is not limited to sources with a proven provenance in Sumerian periods for two reasons: First, literary pieces unearthed from third-millennium BCE contexts at geographically diverse sites (Ebla and Mari in the north and Girsu, Abu Salabiḥ, Nippur, and Adab in the south) suggest these compositions were conceived of even earlier; second, while primary sources for Sumerian literature originate from a range of temporal, geographical, and cultural settings, many are dated to the Old Babylonian (OB) period (c. 20th–16th centuries BCE). Sumerian is absent from → administrative and legal documents, as well as letters, from around 1730 BCE onwards, supporting the theory that at this time it no longer survived as a spoken language in any of the urban centers of the → Mesopotamian plateau. Its use in the vernacular registers of personal correspondences had ceased some two centuries earlier. Scribes who most likely spoke → Akkadian as their mother tongue carried on the tradition of Sumerian literature via copies of earlier compositions (e.g., Instructions of Šuruppak) as well as new compositions and those created through extensive modification of known material such as *Lugalbanda*.

*Terhi Nurmikko-Fuller, 350–351*

→ *administrative texts, Akkadian, cuneiform, Mesopotamia, Mesopotamian civilizations*

### Syllabic Sign

It represents a syllable, whereas a logographic → sign represents a word.

*Saana Svärd, Heidi Jauhiainen, Aleksi Sahala, Krister Lindén, 225n9*

→ *cuneiform, logosyllabic, sign*

### Syntactic Influence

It refers to the observation of the order of linguistic elements in a sentence of one language in a sentence of another, linguistically unrelated, language. This is exemplified by the formula “x-š-a-y-θ-i-y : x-š-a-y-θ-i-y-a-n-a-m,” “king of kings” in → Old Persian royal inscriptions. Normally in Old Persian, the genitive precedes its object \*(“p-a-r-s-h-y-a : p-u-ç” [Xph line 12], son [p-u-ç] of Persian [p-a-r-s-h-y-a])\* , but in → Akkadian the genitive follows its object, as in “šarrū māti,” meaning “kings of the land.” Hence, with the genitive “x-š-a-y-θ-i-y-a-n-a-m” following its object “x-š-a-y-θ-i-y,” as in the Akkadian example, the occurrence of the genitive “māti” after its object “šarrū” represents syntactic influence.

*Doğu Kaan Eraslan, 301n70*

→ *Akkadian, Old Persian*

### Tablet

A clay object, probably one that fits comfortably in the hand and is inscribed with the distinctive wedge-like → signs of the → cuneiform script.

*Terhi Nurmikko-Fuller, 343*

→ *cuneiform, sign*

### Tell

An artificial mound formed as a result of successive cycles of construction, occupation, and destruction, sometimes separated by periods of abandonment.

*Sveta Matskevich and Ilan Sharon, 25n3*

→ *archaeological site, layer, stratum*

### Transliteration

Conversions of the → cuneiform script into the Latin alphabet without the translation of the ancient language to a modern one.

*Terhi Nurmikko-Fuller, 341n11*

Transcription of cuneiform inscriptions into a romanized rendering of the → cuneiform → sign readings. This includes marking word boundaries and some structural markers, such as line numbering, object surfaces, etc.

*Émilie Pagé-Perron, 199n18*

→ *cuneiform, sign*

### Ubaid Clay Sickles

The Ubaid clay sickles (fifth to fourth millennium BCE) are usually associated with the first intensive agricultural activities in the alluvial plain.

*Marco Ramazzotti, 67n29*

→ *Mesopotamia, Mesopotamian civilizations*

### Ugarit

The Kingdom of Ugarit had about two hundred villages, with an estimated population of thirty-five thousand inhabitants, including the capital. The kingdom was destroyed at the end of the Late Bronze Age, c. 1200 BCE.

*Vanessa Bigot Juloux, 152n6*

In the centuries prior to its final conflagration, Ugarit was a cosmopolitan culture, exhibiting artistic and stylistic influences from Egypt, → Mesopotamia, the Aegean, and Anatolia. To date, Ugarit has yielded approximately 4,500 texts in various languages and writing systems.

*Miller C. Prosser, 316*

→ *Ba'lu and 'Anatu Cycle, Mesopotamia, Ras Šamra, Ugaritic*

### Ugaritic

The scribes of → Ugarit used a newly invented alphabetic writing system consisting of 30 letters ... Like the → logosyllabic Mesopotamian writing system, Ugaritic letters are formed by impressing a stylus into a clay → tablet, creating a series of wedges. This type of writing is known as → cuneiform, from the Latin *cuneus*, “wedge” ... The Ugaritic alphabetic → writing system indicates vowels only partially and only indirectly by use of three aleph → signs ʾ, i, and ū. Otherwise, no vowels were written. The vocalized form of the word is meant to convey the grammatical interpretation of the word.

*Miller C. Prosser, 316, 325n32*

→ *cuneiform, logosyllabic, sign, tablet, Ugarit, writing system*

### Uruk Clay Cones

Uruk clay cones (fourth millennium BCE) are linked to the ideographical and symbolical representations of the most important religious buildings of the first southern → Mesopotamian cities.

*Marco Ramazzotti, 67n29*

→ *Mesopotamia, Mesopotamian civilizations*

**Writing System** → see *CyberResearch Glossary*

# Index of Authors and Researchers (including authors in bibliographies, scholars, and historical figures)

- Abraham, Kathleen 223  
 Adams, Ernest W. 112, 114, 117, 146  
 Adams, Robert McCormick 63n14, 65n19, 65n21, 75–76  
 Adams, William Y. 112, 114, 117, 146  
 Aharoni, Yohanan 26n9, 41n62, 55  
 Akkermans, Peter M.M.G. 316n9, 334  
 Alaparthi, Chaitanya Sai 311  
 Albert, Rosa-Maria 59  
 Aldenderfer, Mark 124, 146  
 Alivernini, Sergio 350, 361  
 Alkim, Handan 114n18, 146  
 Alkim, U. Bahadır 114n18, 146  
 Allemang, Dean 342n15, 361  
 Alvarado, Rafael C. 9n36, 19  
 Anati, Emmanuel 36–37  
 Andersen, Peter B. 60n1, 76  
 Anderson, Adam 15n57, 213n66, 221  
 Anderson, James A. 60n2, 76  
 Anderson, Lloyd 285n11, 310  
 Angiolini, Andrea 6n17, 19  
 Anscombe, Gertrude E. M. 158n41, 189  
 Arens, William 230n22, 254  
 Aristotle 155  
 Arnaud, Daniel 316n8, 335  
 Artzi, Michal 26  
 Asadi-Zeydabadi, Masoud 77  
 Asher-Greve, Julia M. 89, 106  
 Audi, Robert 158n41, 189  
 Aydingün, Şengül 125n50, 146  
  
 Babbage, Charles 338  
 Babcock, Sidney H. 79  
 Bah, Tavmjong 304n81, 310  
 Bahn, Paul 28n17, 31n28, 33n37, 58  
 Bahrani, Zainab 89n16, 106  
 Balasundaram, Balabhaskar 216n72, 221  
 Balensi, Jacqueline 37, 55  
 Bally, Gert von 323n30, 334  
 Bamman, David 196, 213n66, 221  
 Barceló, Juan A. 62n7, 76  
 Barker, Philip 34, 53, 55  
 Baroni, Marco 229n17, 254  
 Barr, James 230, 254  
 Bartash, Vitali 212n63, 221  
 Bartel, Brad 115n20, 118–119, 134, 146  
 Barthélemy, Marc 61n4, 76  
 Bastian, Mathieu 274n36, 278  
 Bavant, Marc 287, 310  
 Beckerman, Martin 60n2, 76  
 Bekiari, Chryssoula 336n2, 355n61–62, 356n63, 356n65, 361  
 Benco, Nancy L. 67n29, 76  
 Bender, Barbara 34n43, 55  
 Bengio, Yoshua 296n51, 311  
 Bentley, R. Alexander 61n4, 76  
 Béranger, Marine 13n50, 283n1, 308n95  
 Berg, John 59  
 Bernard, Lou 7n23, 19  
 Berners-Lee, Tim 338, 340n7, 347, 362  
 Berra, Aurélien 7–8, 19,  
 Berry, David M. 8, 19  
 Berthelot, Jean-Michel 154, 189  
 Bhatia, Parul Kalra 191  
 Biehl, Peter F. 119, 146  
 Bigot Juloux, Vanessa 1, 2n4, 3n6, 5n17, 6n19, 7n20, 8n27, 13, 15, 16n60, 19, 151, 166n72, 168n78, 178n107, 181n124, 189, 194n1, 283n1  
 Bilgi, Önder 114n18, 146  
 Binding, Ceri 47n89, 56, 58  
 Binford, Lewis 30–31, 33, 56  
 Bintliff, John 62n7, 76  
 Bird, Steven 238n, 254  
 Bittel, Kurt 125  
 Black, Jeremy 232n32, 254, 342, 350n41, 351n42, 351n44, 362  
 Blin, Guillaume 46n82, 56  
 Blois, Reinier de 230, 231n28–29, 234, 254  
 Blondel, Mathieu 279  
 Blondel, Maurice 154n19  
 Bodard, Gabriel 311  
 Boon, Paul 90, 106

- Bordreuil, Pierre 316n10, 317, 325n, 332n, 334  
 Borger, Rykle 297n54, 310, 321n27, 334  
 Borgman, Christine 8, 19  
 Botvinnik, Olga 280  
 Bouma, Gerlof 240n46, 254  
 Brabazon, Anthony 60n2, 77  
 Bradbury, Jennie 55n109, 57  
 Brandes, Mark A. 67n29, 77  
 Breda, Marco 70n30, 77  
 Breiger, Ronald L. 156n30, 192  
 Brent, James 67n28, 78  
 Brewster, Christopher 347, 348n32, 362–363  
 Brinkman, John A. 65n19, 77  
 Bronner, Leila Leah 153n12  
 Brouwer Burg, Marieka 55n108, 56  
 Brugger, Peter 113n12–13, 146  
 Brugger, Susanne 113n13, 146  
 Brughmans, Tom 61n4, 77–78  
 Brumfield, Sara 196, 201n31, 221  
 Buccellati, Giorgio 114n16, 146  
 Buchholz, Sabine 204n40, 221  
 Budin, Gerhard 172n95, 189  
 Bühler, Axel 159n43, 189  
 Bulgarelli, Odoardo 222  
 Burdick, Anne 4n10, 7n22, 10n39, 19  
 Burns, Patrick J. 311  
 Busa, Roberto A. 7n22, 19  
 Buscema, Paolo M. 12, 60, 61n5, 64n16,  
 66n26, 67n27–28, 68, 70n30, 71n32, 73n,  
 74n, 75, 77–78, 415  
 Butenko, Sergiy 221  
 Butters, Albion 224n1  
 Buurman, Jan 287n18, 310  
  
 Camiz, Sergio 12n42, 13n48, 85n\*, 88n11–12,  
 92n23, 94n27, 94n29, 95n30–32, 106n38,  
 107–109  
 Campbell Thompson, Reginald 312  
 Camps, Jean-Baptiste 15n58  
 Capurro, Rafael 156n29, 189  
 Carayol, Valérie 6n18  
 Carroll, Lewis (Charles Lutwidge Dodgson)  
 111, 146  
 Cassin, Elena 89n14, 107  
 Catzola, Luigi 70n30, 77  
 Chapman, Nigel 293n38, 310  
 Chapman, Rupert 42n65, 56  
 Chauhan, Bhupendra Singh 311  
 Cheema, Muhammad F. 279  
  
 Chen, Kai 255  
 Chenet, Georges 152n9  
 Cheng-Lin, Liu 310  
 Cheriet, Mohamed 296n50, 310  
 Chircos, Christian 222, 254, 307n92, 311  
 Childe, V. Gordon 62n10, 78  
 Ching, Suen 310  
 Chiricat, Edouard 312  
 Church, Kenneth Ward 229n18, 232n30,  
 239n43, 254  
 Cioè, Francesca 12n43–44  
 Ciotti, Fabio 5n16, 19  
 Ciraci, Fabio 5n15, 19  
 Civil, Miguel 351  
 Clarke, David 28, 56, 62n11, 77–78  
 Clauber, Johann 155  
 Clivaz, Claire 5n14, 6n18, 7n21, 19  
 Clore, Gerald L. 180, 192  
 Cohen, Carl 295n47, 310  
 Cohen, Yoram 223  
 Colbert, S. Chris 268n28, 280  
 Collar, Anna 61n4, 78  
 Collins, Allan 180, 192  
 Collon, Dominique 89n14–15, 97n34, 107  
 Conway, Paul 11n41, 20  
 Coombs, James H. 163, 189  
 Copi, Irving M. 295n47, 310  
 Corrado, Greg 255  
 Courville, Aaron 296n51, 311  
 Coward, Fiona 78  
 Crawford, Harriet 352n52, 362  
 Cripps, Paul 45n75, 56  
 Crofts, Nick 336n1, 354n57–58, 362  
 Crowther, Charles 312  
 Cunningham, Graham 351n46, 352n52, 362  
 Curé, Olivier 46n82, 56  
 Čapek, Karel 85  
  
 D'Agostino, Franco 350n36–37, 361  
 Dahlstör, Erik 304n80, 310  
 Damerow, Peter 200n26, 222  
 Danielová, Mariana 55n110, 56  
 Dannhauer, Johann Conrad 155  
 Darbellay, Frédéric 8, 16n60, 20  
 Daressy, Georges 294n43, 310  
 Dauphin, Claudine 26, 56  
 Davidson, Donald 158n40, 158n42, 160, 162,  
 189  
 Davis, Mark 300n67, 312

- Dawson-Howe, Kenneth 296n50, 310  
 Day, Peggy L. 153n12, 189  
 Dean, Jeffrey 255  
 Delnero, Paul A. 217, 221, 342n12–13, 352n52, 362  
 Dengler, Patrick 310  
 DeRose, Steven J. 163, 189  
 Di Donato, Francesca 19  
 Dietrich, Manfred 152n4, 189  
 Digard, François 87n9, 107  
 Dilthey, Wilhelm 155  
 Dinu, Georgiana 229n17, 254  
 Dirksen, Dieter 323n, 334  
 Dirven, René 230n23, 254  
 Dixon, Helen 14n54  
 Doerr, Martin 45n73, 56, 312, 361–362  
 Doğan, Ismet 274n38, 280  
 Doğan, Nurhan 274n38, 280  
 Drucker, Johanna 19  
 Dunford, Robert 109n55, 57  
 Dunham, Sally 79  
 Dutcher, Jennifer 156n30, 189
- Earl, Graeme 57  
 Ebeling, Jarle 351, 352n52, 362  
 Eisenstadt, Shmuel N. 329n, 335  
 Ekinci, H. Ali 146  
 Elliot, Tom 289n23, 311  
 Ellsworth, Phoebe C. 179, 189  
 Emberling, Geoff 316n13, 335  
 Englehardt, Joshua 151n2, 189  
 Eraslan, Doğu Kaan 5n12, 10n38, 14, 283, 293n40, 311  
 Eriksen, Thomas H. 228n, 254  
 Escobar, Eduardo 196, 198n, 222  
 Evans, Leighton 4n10, 20  
 Everson, Michael 285n11, 311
- Fawcett, Clare 49n96, 57  
 Feldman, Marian H. 117, 118n27, 146  
 Felicetti, Achille 300n65, 311  
 Fellows, Dave 56  
 Fensham, F. Charles 152n5, 189  
 Ferilli, Guido 77  
 Ferraiolo, Jon 310  
 Feuerheim, Karljuergen 285n11, 310–311  
 Figge, Udo L. 60n1, 78  
 Fink, Sebastian 224n1  
 Finkbeiner, Üwe 65n18, 78
- Finkel, Irving L. 359n, 362  
 Finkelstein, Israel 52n103, 57  
 Firth, John R. 229n17, 246–247, 254  
 Fischer, Claudia 97n34, 107  
 Flanders, Julia 9n37, 18n62, 20  
 Fleisch, Axel 230n23, 254  
 Fontaine, Johnny R. J. 179, 189  
 Ford, James A. 115–117, 146  
 Foti, Nicholas J. 190  
 Fraassen, Bas C. van 32n32, 57  
 Franzini, Greta 279  
 Franzosi, Roberto P. 155, 161n52, 181–182, 189  
 Freu, Jacques 152n10, 189  
 Freytag, Asmus 300n67, 312
- Gabbay, Uri 156–157, 189–190  
 Gansell, Amy R. 1, 115n20, 136n, 146, 194n1, 283n1  
 Gardin, Jean-Claude 87–88, 90, 107  
 Gardiner, Eileen 8, 20  
 Garstang, John 25, 33, 38, 43–44, 57  
 Geeraerts, Dirk 230n23, 255  
 Geertz, Clifford 155n24  
 Geller, Mark 263n, 265, 278–279  
 Gemperline, David C. 280  
 Gens, Jean-Claude 155n23, 155n25–27, 190  
 George, Andrew 232n32, 254, 357n66, 362  
 Gibbins, Nicholas 57  
 Gibson, McGuire 65n19, 78  
 Gibson, Shimon 26, 56  
 Gilboa, Ayelet 27n13, 59  
 Gill, Tony 262  
 Gippert, Jost 283n3, 311  
 Gnecco, Cristóbal 112n3, 146  
 Goldberg, Yoav 248n60, 255  
 Goodfellow, Ian 296n51, 311  
 Goodnick Westenholz, Joan 292n35, 312, 350n40, 362  
 Goody, Jack 86n, 107  
 Goodyear, Albert C. 31, 58  
 Gordon, Allan D. 11n1, 146  
 Gramfort, Alexandre 279  
 Granger, Brian E. 267n25, 279  
 Grasso, Anthony 310  
 Gray, John 153n12, 190  
 Greenhalgh, Anne 56  
 Greenstein, Edward L. 169n84, 190  
 Greimas, Algirdas J. 161n51, 190  
 Grevisse, Maurice 168, 190



- Grice, Herbert Paul 157, 190  
 Grimal, Nicolas 310  
 Grisel, Olivier 279  
 Grondin, Jean 155n26, 190  
 Grossi, Enzo 70n30, 77–78  
 Gruber, Thomas R. 347, 362  
 Guralnick, Eleanor 115n20, 147  
 Guz-Zilberstein, Bracha 59
- Habu, Junko 49n76, 57  
 Hage, Per 67n28, 78  
 Hainsworth, Michael 310  
 Halchenko, Yaroslav 280  
 Hallock, Richard 286, 311  
 Hallof, Jochem 310  
 Hamilton, Sue 34n43, 55  
 Hanen, Marsha P. 32n32, 57  
 Hanks, Patrick 229n18, 232n30, 239n43, 254  
 Hansen, Donald P. 79  
 Hansen, Svend 115, 147  
 Harari, Yuval N. 159n44, 190  
 Harary, Frank 67n28, 78  
 Hausenblas, Michael 313  
 Haussperger, Martha 89n14–15, 107  
 Hawley, Robert 159n46, 190  
 Hearst, Marti 181, 190  
 Heeßel, Nils 264n18, 265, 266, 279  
 Heidegger, Martin 155n24  
 Heimpel, Wolfgang 96n, 107  
 Heizer, Robert F. 34, 39n59, 57  
 Hempel, Carl 30n22, 57  
 Hendler, Jim 342n15, 361  
 Hentrich, Thomas 153n12, 190  
 Heppler, Jason 8  
 Hermon, Sorin 112, 115n20, 115n22, 122, 123n41, 147  
 Herzog, Zeev 55  
 Heymann, Sebastian 278  
 Hicks, Illya V. 221  
 Hinrich, Schütze 255  
 Hirschberg, Israel 33  
 Hobson, Paul 280  
 Hockey, Susan 7n23, 20  
 Hodder, Ian 32n31, 33n36, 34n38, 34n43–44, 35, 35n46, 53n105, 57, 117n26, 147  
 Hole, Frank 34, 39n59, 57  
 Holl, Austin F. C. 28n18, 58  
 Holley, Rose 298n57–58, 311  
 Hollis, Luke 311
- Homburg, Timo 307n92, 311  
 Huehnergard, John 176n103, 190  
 Hughes, James M. 162n57, 190  
 Hunter, John D. 270n30, 279
- Ide, Nancy 164n64–65, 190  
 Iliffe, John H. 44, 57  
 Isaksen, Leif 45n77, 55n111, 57
- Jackson, Dean 310  
 Jacomy, Mathieu 278  
 James, William 179n111, 190  
 Jänicke, Stefan 259n5–6, 267n26, 279  
 Jansen-Winkel, Karl 288, 311  
 Jauhiainen, Heidi 13, 224, 253n64, 255  
 Jauhiainen, Tommi 253n64, 255  
 Jaworski, Wojciech 349, 362  
 Jefferson, Tom 70n30, 77  
 Jeffrey, Stewart 58  
 Jenkins, John 310  
 Jerrold, Cooper 78  
 Jewell, Michael O. 336n3, 357n68, 362  
 Jockers, Matthew L. 260n8, 261n11, 279  
 Johnson, Gregory A. 67n29, 82  
 Johnson, Kyle P. 283n2, 311  
 Joukowsky, Martha 34, 57  
 Juloux, Vanessa 153n14, 160n50, 176n101, 178n106, 191  
 Jun, Fujisawa 310
- Kanza, Sarah W. 194n1  
 Kapelrud, Arvid S. 153n12, 191  
 Kaplan, Abraham 32, 33n33, 57  
 Karkajian, Lourik 153n11, 191  
 Karp, Ivan 230n22, 254  
 Kataja, Laura 227, 255  
 Kaufman, Stephen A. 288n21, 311  
 Keay, Simon 57  
 Kedar, Siram 223  
 Kelley, Jane H. 32n32, 57  
 Kendall, David G. 62n8, 78  
 Kenyon, Kathleen M. 39, 57  
 Khait, Illya 222  
 Kharma, Nawwaf 310  
 Khwārizmī, Abū Ja'far Muḥammad ibn Mūsā al- 3n8  
 King, Leonard William 302n71, 312  
 Kingsley, Sean 26  
 Kirby, Tyler 311

- Kirschenbaum, Matthew 7n23, 8, 8n24, 20  
Klein, Ewan 238n, 254  
Klengel-Brandt, Evelyn 89n14, 107  
Knappett, Carl 78  
Koch, Ulla S. 156n31, 157n35, 191, 278n, 279  
Kochavi, Moshe 55  
Kohler, Timothy 61n4, 78  
Kohonen, Teuvo 65n20, 78, 94n26, 107  
Kökten, Kiliç 114n18, 147  
Korfmann, Manfred 125, 147  
Koskenniemi, Kimmo 227, 255  
Koslova, Natalia 200n26, 222  
Krakauer, David C. 190  
Kranzberg, Melvin 342, 362  
Kromer, Bernd 125n46, 147  
Kruskal, Joseph B. 67n28, 78  
Kruszewski, Germán 229n17, 254  
Kuhn, Thomas 18n64, 20  
Kumar, Lokesh 181n23, 191  
Kumke, Holger 55n110, 56  
  
Lakoff, George 230, 255  
Laks, André 156n28, 191  
Lamon, Robert S. 40n61, 57  
Landsberger, Benno 229, 255  
Langacker, Ronald W. 166n74, 170n91, 174, 191  
Langebaek, Carl 112n3, 146  
Langendoen, Terence D. 165n67, 191  
Laron, Gabi 44  
Lawrence, Dan 55n109, 57  
Lawrence, Faith K. 357n69, 362–363  
Le Boeuf, Patrick 300n65, 312, 361  
Le Ny, Jean-François 160n47, 191  
Lebart, Ludovic 94n29, 97n35, 108  
Leeuw, Sander van der 61n4, 78  
Leibowitz, Joseph 25–26, 57  
Levavi, Yuval 223  
Levesque, Hector J. 347n26, 363  
Levinson, Stephen C. 230n23, 255  
Levy, Omer 248n60, 255  
Levy, Thomas E. 28n18, 58  
Lilley, Chris 310  
Lin, Dekang 240n45, 256  
Lincoln, Matthew 223  
Lindén, Krister 13, 224, 225n6, 253n64, 255  
Lipiński, Edward 180n17, 191  
Lipman-Blumen, Jean 177, 191  
Liu, Alan 4n9, 20  
Liverani, Mario 63n11, 79, 152n6, 191  
  
Livet, Pierre 178n109, 191  
Lloyd, Jeffery B. 153n12, 192  
Locke, Brandon 223  
Lodwick, Weldon A. 77–78  
Loper, Edward 238n, 254  
Loretz, Ozwald 152n4, 189  
Losee, John 29n, 32n32, 35n45, 58  
Loud, Gordon 41, 52n102, 58  
Lovelace, Ada 338  
Lovis, William A. 56  
Ludovico, Alessandro di 1, 2n4, 3n6, 5n17, 12, 19, 62n7, 79, 85, 88n11–12, 90n19, 91n20, 92n23–24, 94n27–29, 95n30–32, 100n, 105n, 106n, 108–109, 189  
Lukauskas, Saulius 280  
Lunenfeld, Peter 19  
Lyons, William 178n109, 179n110, 180, 192  
  
Madsen, Torsten 120, 147  
Maedche, Alexander 47n87, 58  
Maiocchi, Massimo 2n2, 210n60, 212n63, 219n84, 222  
Majewski, Stefan 172n95, 189  
Makowski, Maciej 126n53, 147  
Manning, Christopher D. 240n45, 255  
Marsi, Erwin 204n40, 221  
Martin, S. Rebecca 43–44, 51–52, 58–59  
Martinez, Kirk 57  
Martino, Matthew 3n6, 12–13, 111, 120  
Martino, Shannon 3n6, 12–13, 111, 114n17, 119, 124–125, 147  
Marwick, Ben 197, 219, 219n81, 222  
Marzahn, Joachim 89n14, 107  
Massin, Olivier 167n76, 175n99, 192  
Massini, Giulia 12, 60, 65n20, 68, 72n, 73, 79  
Mathé, Anthony 6  
Matoian, Valérie 316n9, 335  
Matskevich, Sveta 12, 25, 32n29, 37, 39n58, 40, 58  
Matsunaga, John M. 49n96, 57  
May, Keith 45n77, 47n89, 56, 58  
Mazar, Amihai 27n11, 58  
Mazzini, Giovanni 152n6–7, 192  
McClelland, James L. 60n2, 79  
McCormack, Cameron 310  
McGarrahy, Seán 60n2, 77  
McGowan, Rick 310  
McKinney, Wes 268n28, 279  
McMahon, Kenneth 295n47, 310

- Meent, Jan-Willem van de 146  
 Melero, Francisco J. 106  
 Meredith-Lobay, Megan 20  
 Merton, Robert K. 30–31, 58  
 Mézard, Marc 64n17, 79  
 Michel, Vincent 279  
 Mikolov, Thomas 229n19, 232n31, 246–248,  
     255–256  
 Millard, David E. 362  
 Miller, George A. 170n89, 192  
 Miller, John H. 60n2, 79  
 Mills, Barbara J. 78  
 Minsky, Marvin 60, 79  
 Mlekuž, Dimitrij 70n31, 79  
 Mohr, John W. 156n30, 192  
 Molina, Manuel 212n63, 222  
 Molinié, Georges 155n26, 192  
 Monroe, M. Willis 13–14, 257, 258n, 267n25,  
     279  
 Moor, Johannes de 153  
 Mora, Thierry 64n17, 79  
 Morello, Nathan 2n2, 13n51–52  
 Moretti, Franco 259, 279  
 Morgan, Augustus de 162n57  
 Mörth, Karlheinz 172n95, 189  
 Moscati, Paola 87n5, 109  
 Mosher, Malcolm 356  
 Moshkovitz, Shmuel 55  
 Murano, Francesca 311  
 Murphy, Kelly J. 153n11, 177n104, 192  
 Musto, Ronald G. 8, 20  
 Mylonas, Elli 311  
  
 Nagar-Hilman, Orna 59  
 Nam, Roger S. 89n14, 109  
 Natan-Yulzary, Shirly 153n11, 192  
 Nederhof, Mark-Jan 287n18, 312  
 Nellhaus, Tobin 164n65, 192  
 Neschke, Ada 156n28, 191  
 Niccolucci, Franco 112, 115n20, 115n22, 122,  
     123n41, 147, 311  
 Nissen, Hans J. 63n14, 65n21, 76, 79  
 Norton, Barry 356n64, 363  
 Nurmikko-Fuller, Terhi 2n2, 14, 14n56, 15,  
     151n1, 194n1, 227, 256, 283n1, 312, 336,  
     349n34, 352n53, 363  
 Nyhan, Julianne 20  
  
 O'Hara, Kieron 348n32, 362  
  
 O'Kane, Drew 280  
 O'Neill, Michael 77  
 Obladen-Kauder, Julia 124–125, 126n54–56,  
     129, 147  
 Oldenburg, Ulf 153n12  
 Oldman, Dominic 356n64, 363  
 Olmo Lete, Gregorio del 192  
 Ore, Christian-Emil 312  
 Organisciak, Peter 20  
 Ortony, Andrew 180, 192  
 Özgüç, Nimet 114n18, 147  
 Özgüç, Tahsin 114n18, 147  
  
 Padilla, Thomas 223  
 Page, Hugh R., Jr. 153n11, 192  
 Pagé-Perron, Émilie 2n2, 3n6, 13, 15, 194,  
     195n4, 202n33, 205n46, 210n57, 210n59,  
     219n84, 222, 283n1  
 Pages, Scott E. 60n2, 79  
 Pannapacker, William 18, 20  
 Pantel, Patrick 240n45, 256  
 Pardee, Dennis 152, 154n15, 192, 314, 316n10,  
     317, 325n, 334  
 Pearce, Laurie 220n88, 222  
 Pedregosa, Fabian 272n, 279  
 Peeters, Hans 55n108, 56  
 Pérez, Fernando 267n25, 279  
 Peters, James F. 296n50, 312  
 Peters, Stefan 55n110, 56  
 Petrie, William M. Flinders 53, 58  
 Petritoli, Riccardo 77  
 Phillips, Philip 30n25, 59  
 Pieri, Giovanni 77, 92n23, 94n27–28, 95n30,  
     109  
 Piez, Wendell 9n37, 18n62, 20  
 Pitard, Wayne T. 152n4, 152n10, 192–193  
 Pitt-Rivers, Augustus H. L.-F. 53, 58  
 Plas, Dirk van der 310  
 Plato 2n4  
 Plutchik, Robert 178, 179n111–112, 179n116, 192  
 Pomponio, Francesco 212n63, 222  
 Popper, Karl 161, 192  
 Porada, Edith 64n15, 79, 89n14, 109  
 Posner, Miriam 223  
 Postgate, John Nicholas 232n32, 254  
 Pottorf, Andrew 12n45  
 Presner, Todd 7n22, 9n36, 19, 20  
 Prosser, Miller C. 3n6, 14–15, 154n15, 314  
 Prugel-Bennett, Adam 362

- Pustejovsky, James 164n64, 190
- Raab, L. Mark 31, 58
- Raban, Avner 26, 28, 48, 58–59
- Rainey, Anson F. 55
- Ramazzotti, Marco 12, 60, 61n3–4, 61n6,  
62n7, 62n10–11, 63n12–13, 64n15, 64n17,  
65n18, 65n21, 66n22–23, 66n26, 71n34,  
79–81, 92n23, 94n27, 109
- Ramus, Petrus 155
- Ranaweera, Kamal 20
- Raveh, Kurt 26
- Reason, David 111n2, 147
- Redman, Charles L. 62n11, 81
- Reeler, Claire 62n7, 81
- Rees, Sian 4m10, 20
- Reiner, Erica 265n20, 280
- Rendu Loisel, Anne-Caroline 13n49
- Renear, Allen H. 163, 189
- Renfrew, Colin 28n17, 31n28, 33n37, 58,  
126n53, 147
- Reynolds, Frances 234n35, 256
- Richardson, Alexander 155n26
- Richardson, Seth F. C. 177n104, 192
- Ricœur, Paul 155, 156n28, 158n41, 192
- Riva, Pat 361
- Robinson, David 56
- Robson, Eleanor 342n13, 350n38, 352n53,  
363
- Rochberg, Francesca 261n12, 265n22, 280
- Rochberg-Halton, Francesca 265n22, 280
- Rockmore, Daniel N. 190
- Rockwell, Geoffrey 9n33, 20, 267n27, 280
- Roe, Glenn 151
- Roesch, Etienne B. 179, 189
- Romano, Marco 350n36–37, 361
- Roniger, Luis 329n, 335
- Ronzino, Paola 311
- Ropars, Fabian 61n9, 20
- Rosati, Luca 19
- Rosenfeld, Edward 60n2, 76
- Rosenthal-Heginbottom, Renate 59
- Rosicki, Remigiusz 175, 193
- Roskams, Steve 39n59, 59
- Rosmorduc, Serge 288, 293n40
- Ross, Jennifer C. 88n13, 109
- Rossi, Federica 19
- Roth, Martha T. 229n20, 256
- Rothman, Mitchell S. 66n24, 81
- Rova, Elena 94n29, 107, 113n10
- Roy, Ellen 111n2, 147
- Ruecker, Stan 20
- Rumelhart, David E. 60n2, 79
- Ruth, Luke 313
- Sacco, Pierluigi 67n28, 77–78
- Sahala, Aleksis 13, 224
- Salah, Almila Akdağ 5n12, 21
- Salem, André 94n29, 97n35, 108
- Salvatori, Enrica 5, 19, 21
- Sanmartín, Joaquín 152n4, 189
- Saraçlı, Sinan 274n38, 280
- Sasanow, Maggy 312
- Schaeffer, Claude F.-A. 152n9
- Schaller, Kurt 45n73, 56
- Schepers, Doug 310
- Scherer, Klaus R. 179, 189
- Scheuermann, Gerik 279
- Schlegel, Friedrich 155n24
- Schleiermacher, Friedrich 155
- Schloen, John D. 153n11, 193, 320n25, 335
- Schloen, Sandra 320n25, 335
- Schmidt, Desmond 164n64, 164n66, 193
- Schmidt, Erich 286–287, 312
- Schnapp, Jeffrey 19
- schraefel, m.c. 362
- Schreibman, Susan 8n24, 19, 21
- Schwartz, Glenn M. 316n9, 334
- Searle, John R. 176n101, 193
- Seeher, Jürgen 125n50, 128, 130–133, 147
- Selinger, Peter 305n83, 312
- Selz, Gebhard J. 156n31, 193
- Shadbolt, Nigel R. 362
- Shahack-Gross, Ruth 42n66, 59
- Shahbazi, Alireza Shapur 302n71, 312
- Sharon, Ilan 12, 25, 27n12–13, 39n58, 40n60,  
58–59
- Shennan, Stephen J. 61n4, 76
- Shipton, Geoffrey M. 40n61, 57
- Shuchat, Alan H. 62n8, 81
- Siemens, Ray 21
- Simons, Gary F. 165n67, 191
- Sinclair, Stéfan 267n27, 280
- Singh, Sourav 311
- Slaatte, Howard A. 33n34, 59
- Smith, Abby 111n41, 21
- Smith, Mark S. 152n10, 176n103, 193
- Smith, Michael E. 62n11, 81

- Smith, Noah A. 213n66, 221  
 Smolensky, Paul 62n9, 81  
 Snyder, Dean 310  
 Soldt, Wilfred H. van 176n103, 193, 333n, 335  
 Spaulding, Albert C. 115–117, 147  
 Sperberg-McQueen, C. M. 164n65, 190  
 Stead, Stephen 312, 362  
 Steele, John M. 261n12, 265n21, 266n24, 280  
 Stein, Gil 67n29, 81  
 Steinkeller, Piotr 96n, 110  
 Stern, Ephraim 26n8, 27, 43, 51, 59  
 Stern, Willem B. 89, 106  
 Stewart, Andrew 43n, 44, 58–59  
 Stiff, Matthew 362  
 Stockholm, Daniel 151n1, 283n1, 304n82  
 Stol, Marten 222  
 Stolper, Matthew 292n35, 312  
 Stoyanova, Simona 311  
 Streck, Michael P. 224, 256  
 Sukhareva, Maria 222  
 Sun, Chloé 160n49, 193  
 Sutskever, Ilya 255  
 Svård, Saana 13, 224, 225n6, 234n36, 236n, 237, 256  
 Svensson, Patrik 8, 21
- Taylor, Jon 350n39, 351n43, 363  
 Teeter, Emily 316n13, 335  
 Tenniel, John 146  
 Terras, Melissa 9n37, 18n62, 20  
 Tesnière, Louis 168, 193  
 Theodoridou, Maria 45n73, 56  
 Thirion, Bertrand 279  
 Thissen, Laurens 114n18, 147  
 Thomas Aquinas (St.) 7n22  
 Thomas, Christine Neal 186n133, 193  
 Thonemann, Peter 290, 312  
 Thuraisingham, Bhavani 320n24, 335  
 Tilley, Christopher 34n43, 55  
 Tinney, Steve 227n15, 285n10–11, 311–312  
 Tixier, Jacques 112n6, 147  
 Tudhope, Doug 56, 58  
 Tüfekçi, Şevket 10n38, 21  
 Tuffield, Mischa M. 362  
 Tupman, Charlotte 311  
 Tyndall, Stephen 227, 256
- Ucko, Peter 118, 148  
 Underwood, Ted 260n8, 261n11, 279
- Unsworth, John 7, 21, 164n64–65, 193
- Van Beek, Gus W. 41n63, 59  
 Vandeloise, Claude 166n74, 170n91, 174n97, 191  
 Vanderbilt, Scott 311  
 Vanhoutte, Edward 164n65, 193  
 Varoquaux, Gaël 268n28, 279–280  
 Veldhuis, Nick 204, 223, 225n6, 227n15  
 Verspoor, Marjolijn 230n23, 254  
 Vidal, Jordi 152n8, 176n102, 193  
 Viiri, Sampo 5n13, 21  
 Vincent, Matthew 13n46  
 Visicato, Giuseppe 212n63, 222–223  
 Vitali, Stefano 19  
 Vries-Melein, Martine de 89, 90n18, 106
- Wachsmann, Shelley 26  
 Waerzeggers, Caroline 196–197, 217n74, 223, 227n12, 256  
 Wagner, Allon 196, 223  
 Wagner-Pacifici, Robin 156n30, 192  
 Walls, Neal 153n12, 193  
 Walt, Stéfan van der 268n28, 280  
 Waskom, Michael 270n30, 280  
 Watt, Jonathan 310  
 Weidner, Ernst F. 264n16, 280  
 Weinberg, Saul S. 118, 148  
 Weiner, Steve 59  
 Weingart, Scott B. 223  
 Weiss, Harvey 65n19, 81  
 Weisstein, Eric W. 284n6, 312  
 Westenholz, Aage 212n63, 222–223  
 Whistler, Ken 300n67, 312  
 Wiener, Norbert 3n4, 21  
 Wierzbicka, Anna 178n109  
 Wiggins, Chris H. 146  
 Wilks, Yorick 347, 363  
 Willey, Gordon R. 30n25, 59  
 Wilson, Eleanor A. 153n12, 193  
 Winter, Irene J. 89n14, 110  
 Wolde, Ellen van 230, 256  
 Wood, David 300n64, 313  
 Woods, Christopher 316n13, 335  
 Woolley, Charles Leonard 341, 363  
 Wright, Henry T. 63n14, 67n29, 81–82  
 Wyatt, Nicolas 151n1, 152n6–7, 152n9–10, 153n12, 159, 193

- Yadin, Yigael 41  
Yardeni, Ada 294n43, 313  
Yasur-Landau, Assaf 28  
Yih, Wen-tau 229n19, 232n31, 246n50–52, 256  
Yon, Marguerite 152n10, 176n102, 193, 315n7, 316n8, 319n22, 335  
Young, Timothy C. 65n19, 82  
  
Zachary, Wayne W. 220, 223  
Zadok, Ran 223  
  
Zaidman, Marsha 313  
Zairis, Sakellarios 146  
Zarzecki-Peleg, Anabel 41n64, 59  
Zhi, Yang 212n63, 223  
Zólyomi, Gábor 342, 350n41, 351n42, 351n44, 362  
Zorn, Jeff R. 27n13–14, 48n92, 59  
Zubrow, Ezra B. W. 62n7, 82  
Zundert, Joris J. van 257n1, 280  
Zweig, Geoffrey 229n19, 232n31, 246n50–52, 256

# Index of CyberResearch\* (including computer science, digital practice, mathematical, and technological terms)

- 2D 284, 304–305  
3D 37, 284n6, 305
- A-Temporal Diffusion Model (ATDM) 66n27, 71
- AAS (Artificial Adaptive System) 61, 63, 66
- Abstract vector 297, 297n54, 298, 304  
*See also* features' vector; Optical Character Recognition  
*See also in General Index* sign
- Adab network graph 210n59, 211
- Adjacency matrix 273–275, 306n87  
*See also* edge; node
- AI (Artificial Intelligence) 62, 63n12, 307, 344, 347
- Algorithm 60n2, 61n5, 63n12, 64n16–17, 65n20, 66–67, 70, 91–92, 94–95, 98, 106, 118, 120–122, 133–134, 197, 206, 209, 213–214, 217–220, 260, 261n10, 275, 277, 296, 304, 307–308, 338, 344  
*See also* Artificial Neural Network
- American Standard Code for Information Interchange (ASCII) characters 285, 287–288
- Analyse logiciste* 87
- Analytical taxonomies 154, 161–162, 166, 181–182, 185–186
- ANN (Artificial Neural Network) 62n9, 64, 66, 94–95, 229, 246, 261
- Application Programming Interface (API) 290, 318  
*See also* Online Cultural and Historical Research Environment
- Arbitrary unit 39  
*See also in General Index* spatial unit
- Arc 74, 308n93, 347
- Architecture (in CyberResearch) 73n36, 248, 291n33, 297n54, 307, 321, 338
- Artificial Adaptive Systems (AAS) 61, 63, 66  
*See also* algorithm; Artificial Sciences; Natural Computing
- Artificial Intelligence (AI) 62, 63n12, 307, 344, 347  
*See also* Artificial Neural Network
- Artificial intelligence model 63n12  
*See also* Mesopotamian Urban Revolution Landscape
- Artificial Neural Network (ANN) 62n9, 64, 66, 94–95, 229, 246, 261  
*See also* algorithm; Artificial Intelligence; neural networks; Self-Organizing Map; Word2vec
- Artificial Sciences (AS) 61n5  
*See also* Artificial Adaptive Systems
- AS (Artificial Sciences) 61n5
- ASCII characters (American Standard Code for Information Interchange characters) 285, 287–288
- ASCII Transliteration Format (ATF) 200, 285n11  
*See also* American Standard Code for Information Interchange characters; Cuneiform Digital Library Initiative; Open Richly Annotated Cuneiform Corpus  
*See also in General Index* tablet

---

\* The words following the first word of a term are capitalized when the expression has an acronym but not capitalized when the expression has no acronym. The related entries after “see also” follow the same rule for acronyms but are not capitalized when the expression does not have an acronym. There are some exceptions, such as Python, R, and Word2vec, that are applications or tools, and thus always capitalized. The index entries followed by “see also” refer to related term(s)/expression(s) in glossaries.

- ATDM (A-Temporal Diffusion Model) 66n27, 71
- ATF (ASCII Transliteration Format) 200, 285n11
- Atomization 314n4, 319, 323  
*See also in General Index* logossyllabic; sign
- Attribute (substantive) 34, 43–44, 46, 47n84, 50–51, 118–121, 123–124, 127, 134, 136–137, 164–166, 171n92, 172n95, 182, 202n35, 220, 289, 295, 302–303, 309  
*See also* edge; node; tag; Text Encoding Initiative
- Attribute-analysis approach 118–119
- Augmenting a dataset 307n91
- Auto-CM (Auto-Contractive Map) 66n26, 68–74
- Auto-Contractive Map (Auto-CM) 66n26, 68–74  
*See also* Artificial Neural Network
- Automatic conversion 290
- BabelNet 170
- Bag-of-words model 270–271  
*See also* Continuous Bag-of-Words model
- Bavant-XML 286–287, 291, 294n43, 295
- Big data 156
- Bigram 239–241, 246
- Binary correspondence analysis 99
- Bridge 202, 212, 217  
*See also* edge betweenness; graph partitioning; node  
*See also in General Index* Adab corpus
- C++ language 120, 257n2
- c(anonical)-ATF (C-ATF) 199n18, 199n23, 200n24, 205n47, 285–286, 291, 292n36, 294n43, 295  
*See also* American Standard Code for Information Interchange characters; Cuneiform Digital Library Initiative
- C-ATF (c(anonical)-ATF) 199n18, 199n23, 200n24, 205n47, 285–286, 291, 292n36, 294n43, 295
- CAL (Comprehensive Aramaic Lexicon) 288, 292n34
- CAL Code 288–289, 291, 294n43, 295  
*See also* American Standard Code for Information Interchange characters  
*See also in General Index* Aramaic
- CBOW model (Continuous Bag-of-Words model) 247–248
- CDLI (Cuneiform Digital Library Initiative) 88, 198–200, 202n37, 203, 218n79, 285–286, 346n24, 352, 355, 360
- CIDOC Conceptual Reference Model (CIDOC CRM) 45, 48, 300, 336, 353–357, 361  
*See also* Conceptual Reference Model; International Organization for Standardization; ontology
- CIDOC CRM (CIDOC Conceptual Reference Model) 45, 48, 300, 336, 353–357, 361
- Class 42, 45n71, 64n17, 65n20, 94, 119, 122–124, 127, 291n33, 300, 308, 347–348, 354–361
- Class identifiers 301
- Classical Language Toolkit (CLTK) 283n2
- Classification of ontologies 348
- Clique 214–216  
*See also* maximal clique; network graph analysis; node  
*See also in General Index* tablet
- Close reading 156n30, 233, 259
- CLTK (Classical Language Toolkit) 283n2
- Cluster 119–120, 123–124, 133–134, 212, 220n87, 246, 258, 275–276, 278, 338, 345
- Cluster analysis 118, 274  
*See also* algorithm; multivariate analysis
- Clustering method 214n68, 220, 274  
*See also* cophenetic correlation coefficient; network analysis; node; social network analysis; Unweighted Pair-Group Average
- CNNs (Convolutional Neural Networks) 307n91
- Comma Separated Value (CSV) 137, 206, 274, 344
- Comprehensive Aramaic Lexicon (CAL) 288, 292n34
- Computer-aided textual analysis 258–259
- Computer Science 60, 197n16, 299, 343
- Computer semiotics 60
- Computerized data modeling 55  
*See also* data model
- Conceptual Reference Model (CRM) 45, 300  
*See also* CIDOC Conceptual Reference Model; Data Management System; Linked Open Data; ontology
- Continuous Bag-of-Words (CBOW) model 247–248  
*See also* bag-of-words model; Word2vec



- Continuous Skip-gram model 247  
*See also* Word2vec
- Convolutional layer 307n91
- Convolutional Neural Networks (CNNs) 307n91
- Cophenetic correlation coefficient 274  
*See also* clustering method
- Correspondence analysis 94–95, 99, 105–106  
*See also* hierarchical classification; quantitative analysis/method
- Cosine similarity 272–274  
*See also* Document Term Matrix
- CountVectorizer 271
- CRM (Conceptual Reference Model) 45, 300
- Cross-platform 304
- csv (Comma Separated Value) 137, 206, 274, 344
- Cultural heritage data 45, 353, 355, 361
- Cuneiform Digital Library Initiative (CDLI) 88, 198–200, 202n37, 203, 218n79, 285–286, 346n24, 352, 355, 360  
*See also* American Standard Code for Information Interchange characters; metadata  
*See also in General Index* cuneiform; transliteration
- Cuneiform Digital Palaeography Project 351
- Cytoscape 209, 210n59, 211n61, 212–213, 218n80  
*See also* Gephi; graphing programs; weighted
- DANA (Digital Archaeology and National Archive) 28
- Data-management system 36n52, 38n55, 45
- Data cloud 49, 55  
*See also* graph; Linked Open Data; triple store
- Data mining 91n21, 162, 181, 260n8, 344  
*See also* Système Portable pour l'Analyse des Données; text mining
- Data model 48–49, 55, 154n15, 315, 319–320, 323, 330, 333–334, 339  
*See also* computerized data modeling
- Data processing 181, 257n2, 268
- Data serialization 299, 305  
*See also* External Data Representation; JavaScript Object Notation
- Data structure 47n83, 48n91 51n98, 196–197, 202, 297n55, 306, 345, 347
- Database 27–29, 32, 46n82, 47, 49, 51, 67n28, 114, 119, 123, 136–137, 170, 171n92, 174, 181, 199–200, 204–206, 285, 290, 314n3–4, 315, 317–326, 329–330, 333–334, 345, 360
- Database Management System (DBMS) 47n84  
*See also* Conceptual Reference Model; identifier; meta-system
- Database of Neo-Sumerian Texts 351
- Dataset 25, 27, 46–48, 49n95, 50, 55, 65n20, 66n27, 67n28, 71, 74, 90, 91n21, 92, 94n26, 95–100, 106, 117–118, 121–122, 124, 128, 133–135, 137, 156n30, 162, 198, 209, 212, 219–220, 229n17, 259n4, 260–261, 267–268, 271–276, 307, 319n23, 336, 341, 343–344, 347–349
- DBMS (Database Management System) 47n84
- Deep learning 66, 225n6, 296, 298n57
- Diachronic Corpus of Sumerian Literature 351
- Dictionary 203, 297n55, 321, 324  
*See also* map
- Digital Archaeology and National Archive (DANA) 28
- Digital Library Development Center (University of Chicago) 318
- Digital theodolite 27n15
- Distant reading 259–260
- Distributional semantic models 228  
*See also* language technology; Pointwise Mutual Information; Word2vec
- Distributional semantics 229n17
- Document-oriented database 51  
*See also* tag
- Document Term Matrix (DTM) 271–272, 274, 284n6  
*See also* cosine similarity; vector space
- DTM (Document Term Matrix) 271–272, 274, 284n6
- Dublin Core 353, 354n56
- Edge 38, 46n80, 67n28, 73, 104, 201–202, 206, 207n52, 208–212, 216–217, 220, 253, 260, 275, 305–306  
*See also* graph; label; node

- Edge betweenness 212, 217  
*See also* bridge; edge; network analysis
- Ego-network 209, 210n59, 217, 219  
*See also* Gephi; network graph; node
- Electronic Text Corpus of Sumerian Literature (ETCSL) 204, 336, 342, 350–356, 358–60, 360n74–75, 361  
*See also* American Standard Code for Information Interchange characters; Unicode  
*See also in General Index* Sumerian; transliteration
- Element 36, 62n9, 67n28, 100, 104, 164–165, 167, 168n78, 169, 171n92, 174, 182, 184, 204, 239, 259, 293, 294n43, 301–302, 303n76, 306n88  
*See also* attribute; markup; tag; Text Encoding Initiative
- Elementary semantic unit 285n8, 291n32
- Encode/encoding 60–63, 64n17, 72, 87–88, 91–93, 96–97, 100, 152, 154, 164–165, 167, 199n18, 199n23, 200, 204n42, 220n86, 267n27, 284–286, 289–294, 297–304, 307–309, 326n33, 328, 339
- Encoding process 152, 285n10, 289n27
- Encoding scheme 200, 205n48, 284n5, 286n14, 290–293, 295–296, 297n52, 298–301, 304–305, 306n90, 307–310  
*See also* interoperability; minimal semantic unit
- Entity 30, 37n52, 38–39, 42–44, 46–47, 49–50, 55, 71, 204, 206, 209n54, 215–216, 339, 345, 347–349, 354, 358–359
- Entity-relationship model 355
- EpiDoc 285, 289–291, 294n43, 295, 298–304, 308–310  
*See also* Extensible Markup Language; Text Encoding Initiative; Unicode Transformation Format (UTF)-8
- Epigraphic interoperability 284n5, 296, 299n63, 305, 308n95, 309  
*See also* interoperability
- epSD (Pennsylvania Sumerian Dictionary) 352
- ETCSL (Electronic Text Corpus of Sumerian Literature) 204, 336, 342, 350–356, 358–60, 360n74–75, 361
- Euclidean metric 121
- Event 36–37, 39–40, 42, 44, 50, 66, 74  
*See also* syntactic elements
- Extensible Markup Language (XML) 51n98, 163–164, 203–204, 289n27, 299, 302n72, 303–305, 319, 334, 339–340, 347  
*See also* data serialization; EpiDoc; interoperability; machine-actionable data/machine-readable data; markup; RDF/XML
- Extensible Markup Language (XML) annotation 204  
*See also* Electronic Text Corpus of Sumerian Literature; markup; Text Encoding Initiative; web
- Extension 45, 122, 219n82, 295–296, 353, 360
- External Data Representation (XDR) 299  
*See also* data serialization; JavaScript Object Notation
- Extract/Extracting 91n21, 163, 166–167, 181, 194–197, 204–206, 219, 276, 285n9, 292, 293n41, 296n50, 301, 304n83, 318n18, 321n27, 322n29
- Extraction/extractor 91n21, 161–163, 197, 293n40, 304n83, 308
- Factor(s) 95n29, 100, 115n21, 127, 210n60, 217, 228, 263
- Factor analysis 115, 119
- Factors of variance/factors of variation 308n93  
*See also* feature extraction algorithm
- Feature(s) 36, 37n54, 54, 65n20–21, 67n28, 91n21, 94n26, 94n28, 209, 293n40, 295n46, 296n49, 297, 308, 314n2, 318n17, 350, 353
- Feature extraction algorithm 304n83, 308  
*See also* algorithm; factors of variance
- Features' vector 45n71, 297–300, 304, 308  
*See also in General Index* sign
- FileMaker 114, 124
- FOAF ontology 358
- Format 46–47, 98, 106, 137, 196n7, 200, 204, 205n47, 206n49, 209, 267, 274n36, 295n47, 299, 300n64, 300n66, 301, 303, 305, 314, 319n21, 336, 344, 348
- FrameNet 170
- FRBR (Functional Requirements for Bibliographic Records) 355

- FRBROO 336, 353–357, 361  
*See also* CIDOC Conceptual Reference Model; Functional Requirements for Bibliographic Records; ontology
- Frequency analysis 134
- Frequency counts 271n34
- Function 60–62, 67n28, 87n4, 97n35, 205n47, 213–214, 268, 270–272, 303n77, 305–306, 314, 351
- Functional Requirements for Bibliographic Records (FRBR) 355  
*See also* CIDOC Conceptual Reference Model; FRBROO
- Gaussian distribution 121, 133
- Gephi 209, 210n59, 211n61, 213n65, 218n80, 253, 274  
*See also* Cytoscape; edge; ego-network; graph data; graphing programs; network graph; node
- Graph 46n80, 51, 67n28, 99, 211, 206, 213n65  
*See also* edge; Gephi; graph data; node
- Graph-algorithms 66
- Graph analysis 66–67, 201, 214
- Graph data 46–47, 49, 197, 202, 209, 210n59, 219–220  
*See also* Gephi; graph
- Graph database 46–47, 49–51, 55
- Graph model 46  
*See also* edge; node; triple
- Graph partitioning 212  
*See also* bridge; graph; quantitative analysis/method
- Graph theory 46n80, 66n24, 73, 305n84, 307
- Graph visualization software 208  
*See also* Cytoscape; Gephi
- Graphic 94, 162, 182n127, 304
- Graphical User Interface (GUI) 97n35, 303–304  
*See also* data mining; Système Portable pour l'Analyse des Données
- Graphing programs 274  
*See also* Cytoscape; Gephi
- GUI (Graphical User Interface) 97n35, 303–304
- Hierarchical classification 94–95  
*See also* correspondence analysis
- Hierarchical clustering algorithm 120–121, 133
- Hierarchical softmax 247–248  
*See also* negative sampling; Word2vec
- HTML (HyperText Markup Language) 204n43, 338–339, 342, 352
- HTTP (HyperText Transfer Protocol) 338–340
- HTTP URIs (HyperText Transfer Protocol Universal Resource Identifiers) 344, 346–348
- Hub 210  
*See also* network graph; node
- Human-oriented 292
- Human readable 287, 290n27, 303, 346n25, 347
- Hyperlinks 338  
*See also* HyperText Transfer Protocol; web
- HyperText Markup Language (HTML) 204n43, 338–339, 342, 352  
*See also* HyperText Transfer Protocol
- HyperText Transfer Protocol (HTTP) 338–340  
*See also* HyperText Markup Language; HyperText Transfer Protocol Universal Resource Identifiers; web
- HyperText Transfer Protocol Universal Resource Identifiers (HTTP URIs) 344, 346–348  
*See also* HyperText Markup Language; HyperText Transfer Protocol; Semantic Web; Universal Resource Identifiers; web
- ID (Identifier) 36–37, 43–44, 47n84, 50, 55, 123, 196n7, 202n35, 204, 207–208, 268, 301, 308, 338–339  
*See also* Database Management System
- Identifier (ID) 36–37, 43–44, 47n84, 50, 55, 123, 196n7, 202n35, 204, 207–208, 268, 301, 308, 338–339
- IFLA (International Federation of Library Associations and Institutions) 355
- Inheritance 291n33
- Input 64n16, 65n20–21, 66n25, 72, 73n36, 252, 261n10, 268, 271, 274, 285n9, 293, 295–296, 298, 304n83
- Instance 47n84, 270, 345, 347, 354, 356–360

- Interchangeable 344
- International Federation of Library Associations and Institutions (IFLA) 355
- International Organization for Standardization (ISO) 45, 353, 355, 357
- Interoperability/epigraphic interoperability 29, 164, 284–285, 291–292, 296, 299n63, 305, 308n95, 309
  - See also* encoding scheme
- ISO (International Organization for Standardization) 45, 353, 355, 357
- Item 42n69, 120, 123, 184, 229n17, 271, 275, 314n4, 318–325, 333, 352, 354, 356
- Iterative process 120, 257, 270
- Java 318n17
- Java application client/Java client 318, 334
  - See also* Java Web Start; Online Cultural and Historical Research Environment
- Java Web Start 318
  - See also* Online Cultural and Historical Research Environment
- JavaScript 352
- JavaScript Object Notation (JSON) 299, 305, 340
  - See also* data serialization; External Data Representation
- JSON (JavaScript Object Notation) 299, 305, 340
- JSON-LD 340
- k*-plex 216, 220
  - See also* edge; maximal clique; node
- Key 297n55, 319–320
- Knowledge Representation (KR) 315, 344, 347
- KR (Knowledge Representation) 315, 344, 347
- Label 36n52, 202, 267, 274
  - See also* attribute; edge; identifier; node
- Language technology 225–228, 231
  - See also* distributional semantic models
- Language-technology-related research 227
- Layer 73, 195, 204, 246–247, 307n91
- LD (Linked Data) 47, 55, 201n39, 300, 308, 310, 336, 338, 340, 343–345, 346n25, 347–349, 352, 355n60, 360
- Learning process 296n51, 298n56
- Lemmatization 197, 203
- Library 218n79, 268, 270, 274, 283n2
  - See also* Cuneiform Digital Library Initiative
- Linked 46–48, 74, 216, 319, 321, 328, 330, 354
  - See also* Online Cultural and Historical Research Environment; web
- Linked Data (LD) 47, 55, 201n39, 300, 308, 310, 336, 338, 340, 343–345, 346n25, 347–349, 352, 355n60, 360
  - See also* edge; HyperText Transfer Protocol Universal Resource Identifiers; network graph; ontology; Simple Knowledge Organization System
- Linked Data cloud 46
- Linked Open Data (LOD) 45, 49n95, 344
  - See also* Linked Data; Open Data; Resource Description Framework
- Linked Open Data cloud 45
- LOD (Linked Open Data) 45, 49n95, 344
- log (logarithm) 239
- Logarithm 239
  - See also* Pointwise Mutual Information
- Machine-actionable/machine-readable 46, 164, 200, 204n41, 217, 220, 326n, 336, 339–340
- Machine-actionable data/machine-readable data 200n27
  - See also* scraper
- Machine learning 72–73, 194, 217, 260n8, 296n50, 298n57
- Machine-readability 344
- Machine-understanding 344
- Manuel de Codage (MDC) 287–288, 291, 292n36, 293n40, 294n43, 295–296, 298, 309
  - See also* American Standard Code for Information Interchange characters
- Map/mapping/to map (in the context of a dictionary-data structure) 246, 289, 296, 297n55, 298, 300, 354, 357–358
  - See also* dictionary

- Map/mapping (in other contexts) 30, 45,  
47–49, 65n20–21, 71–72, 94n26, 94n28,  
120, 309, 354–360
- MapReduce model 220n87
- Markup/markup tag/markup tagging 163,  
165–166, 186, 204, 289–290, 299, 302n73  
*See also* element; Extensible Markup  
Language annotation; HyperText  
Markup Language; tag
- Markup language 163, 203, 204n42–43, 259,  
289
- Markup scheme 165
- Matrix/matrices 62–64, 66, 72–74, 271–275,  
296, 299, 306
- Maximal cliques 214–216  
*See also* clique; *k*-plex
- Maximally Regular Graph (MRG) 67, 69, 74  
*See also* Minimum Spanning Tree  
*See also in General Index* Mesopotamian  
Urban Revolution Landscape
- Maximally Regular Graph (MRG) algorithm  
67  
*See also* graph; Maximally Regular Graph;  
Minimum Spanning Tree
- Meta-model 35  
*See also* model
- Meta-system 37, 38n55  
*See also* Data Management System
- Metadata 36, 38, 199, 220, 225, 228, 253, 323,  
339, 341, 343, 345n23, 353, 356  
*See also* Cuneiform Digital Library  
Initiative; Open Richly Annotated  
Cuneiform Corpus
- Method 120
- Metric scaling 120
- Minimal semantic unit 286, 292  
*See also* encoding scheme
- Minimum Spanning Tree (MST) 67, 69–70,  
73–75  
*See also* edge; graph; Maximally Regular  
Graph
- Minimum Spanning Tree (MST) data mining  
67n28  
*See also* data mining; Minimum Spanning  
Tree; spatial analysis
- Mixed method 163  
*See also* qualitative method; quantitative  
analysis/method
- Model (substantive) 30, 32, 35, 45–46,  
48n91, 49, 51, 61, 154n15, 196, 204,  
220n87, 229, 247, 270–271, 300n64, 310,  
315, 319–320, 330, 333–334, 339, 348, 355  
*See also* meta-model
- MRG algorithm (Maximally Regular Graph  
algorithm) 67
- MST (Minimum Spanning Tree) 67, 69–70,  
73–75
- MST data mining (Minimum Spanning Tree  
data mining) 67n28
- Multivariate-attribute-analysis approach 118
- Multivariate analysis 65n20, 118
- MySQL 196, 206–207, 218n80  
*See also* querying; relational database;  
Structured Query Language
- Namespace 295n45, 302, 308  
*See also* EpiDoc; tag
- Natural Computing (NC) 60, 61n5, 61n7,  
63n12  
*See also* Artificial Adaptive System;  
Natural Computing algorithms
- Natural Computing (NC) algorithms 60n2  
*See also* algorithm; Natural Computing;  
neural networks
- Natural Language Processing (NLP) 194, 238,  
246, 322, 344
- Natural Language Toolkit (NLTK) 238
- NC (Natural Computing) 60, 61n5, 61n7,  
63n12
- NC algorithms (Natural Computing  
algorithms) 60n2
- Negative sampling 247–248  
*See also* hierarchical softmax; Word2vec
- Network analysis 61n4, 194–197, 201–202,  
212n62, 214, 216–217, 219–221, 227  
*See also* quantitative analysis/method
- Network analysis algorithms 197  
*See also* algorithm; network analysis
- Network graph 196–197, 201, 202, 206,  
208–211, 213–214, 216–217, 220  
*See also* edge; graph; node; triple
- Network graph analysis 201, 214  
*See also* graph; network graph  
*See also in General Index* administrative  
texts
- Network graph triples 206  
*See also* edge; network graph; node; triple

- Network graph visualization 208, 210  
*See also* graph; network graph
- Network theory 195  
*See also* network graph; social network analysis
- Neural networks 60n2, 62n9, 229n19, 307n91  
*See also* Artificial Neural Network; Natural Computing algorithms
- NLP (Natural Language Processing) 194, 238, 246, 322, 344
- NLTK (Natural Language Toolkit) 238
- Node 46n80, 61, 65n20, 66, 67n28, 73, 168, 180, 196, 201–202, 206, 209–214, 216, 220, 229n19, 253, 260, 305–306, 347  
*See also* edge; edge betweenness; ego-network; network analysis; Self-Organizing Map; weighted  
*See also in General Index* Mesopotamian Urban Revolution Landscape
- Normalization 199, 293, 295–296, 304, 307, 309
- Normalization process 293, 295–296, 309
- Normalized Pointwise Mutual Information (NPMI) 240–246  
*See also* logarithm; Pointwise Mutual Information
- NosQL 51  
*See also* database model; Extensible Markup Language; Resource Description Framework; triple; triple store
- NPMI (Normalized Pointwise Mutual Information) 240–246
- Object 45–46, 60, 115, 118n28, 120, 133–134, 137, 218n79, 259, 307–308, 345–348, 354, 356
- Object-oriented mapping 355
- Objective variables 161, 166–167, 169
- OCHRE (Online Cultural and Historical Research Environment) 314–315, 318–329, 333–334
- OCR (Optical Character Recognition) 296–298, 304, 308–309, 322, 323n
- OD (Open Data) 194, 218–219, 221, 344
- OM (OntoMedia) 169n85, 336, 353–354, 357–361
- Online Cultural and Historical Research Environment (OCHRE) 314–315, 318–329, 333–334  
*See also* data model; Extensible Markup Language; item; Java application client; Java Web Start; linked
- Ontological class 349
- Ontology 47, 169n85, 300n64–65, 320, 348–349, 353–355, 357–358, 361  
*See also* Linked Data; upper ontology
- OntoMedia (OM) 169n85, 336, 353–354, 357–361  
*See also* CIDOC Conceptual Reference Model; ontology
- Open Data (OD) 194, 218–219, 221, 344  
*See also* Comma Separated Value; Linked Data; Linked Open Data
- Open Richly Annotated Cuneiform Corpus (Oracc) 198–200, 203–204, 225, 227–229, 232n32, 246, 253  
*See also* annotation; ASCII Transliteration Format; metadata; tag  
*See also in General Index* Akkadian; Sumerian
- Optical Character Recognition (OCR) 296–298, 304, 308–309, 322, 323n  
*See also* abstract vector; algorithm; deep learning; features' vector; raster image; vector graphic  
*See also in General Index* sign
- Oracc (Open Richly Annotated Cuneiform Corpus) 198–200, 203–204, 225, 227–229, 232n32, 246, 253
- Oracc-ATF 200  
*See also* ASCII Transliteration Format; Open Richly Annotated Cuneiform Corpus  
*See also in General Index* transliteration
- Output vector 246–247
- OWL (Web Ontology Language) 344, 345n21, 353, 357
- Oxford Text Archive 352
- Parser 227, 239, 293, 304
- Parsing 154, 163, 182, 186, 293, 301, 329–330
- Pennsylvania Sumerian Dictionary (ePSD) 352

- Perseus Digital Library 219n82
- Pixel 296
- Platform 48, 200, 238, 303, 317, 334
- PMI (Pointwise Mutual Information) 225, 229, 232–233, 238–240, 242–244, 246, 249–252
- Pointwise Mutual Information (PMI) 225, 229, 232–233, 238–240, 242–244, 246, 249–252  
*See also* Artificial Neural Network; Word2vec
- Potrace engine 304n83
- Pre-processing 166, 174, 184–185, 197–198
- Predicate 46, 345–346, 348
- Prosopography 210n60, 329–332
- Protocol 29–30, 340
- Python 209, 213–214, 216, 218n79, 220, 238, 267–268, 293n41, 297n55, 306n88
- Qualitative method (and related terms) 62, 162–163, 170, 186, 221, 226, 233  
*See also* data mining; mixed method; quantitative analysis/method; R; text mining
- Qualitative data 94n29, 120–121
- Quantitative analysis/method 94n29, 118, 162–163, 212, 216–217, 234, 257–258, 277  
*See also* correspondence analysis; graph partitioning; machine-actionable data/machine-readable data; mixed method; network analysis; qualitative method
- Quantitative data 121, 218, 251
- Query/querying 46, 49, 196n7, 199, 202, 206, 247, 285, 295, 315, 327–328, 331, 347  
*See also* MySQL; Structured Query Language
- R 154, 162–163, 167, 182, 184, 186, 318  
*See also* qualitative method
- Ras Shamra Tablet Inventory (RSTI) 154n15, 314, 317, 318n16, 319–321, 325–329, 331, 333–334  
*See also* Online Cultural and Historical Research Environment; research database environment  
*See also in General Index* Ras Šamra; tablet; transliteration
- Raster image 296–297
- See also* Optical Character Recognition
- Recording 27–29, 33, 35–36, 38, 44, 46, 50, 55, 297, 302
- Recording system 26–27, 29, 33, 36, 38, 44–45, 48–49
- RDF (Resource Description Framework) 46–47, 51n98, 339–340, 344–348, 352, 360–361
- RDF Schema (RDFS) 344
- RDF-XML 339  
*See also* Extensible Markup Language; Resource Description Framework
- RDFS (RDF Schema) 344
- Regular expressions 205, 293
- Relational data model 319  
*See also* item
- Relational database 47, 202, 204, 345  
*See also* identifier; MySQL; querying; Structured Query Language; tuple
- Research database environment 314  
*See also* Ras Shamra Tablet Inventory
- Resource Description Framework (RDF) 46–47, 51n98, 339–340, 344–348, 352, 360–361  
*See also* data model; graph; HyperText Transfer Protocol Universal Resource Identifiers; relational database; Simple Knowledge Organization System; triple
- RSTI (Ras Shamra Tablet Inventory) 154n15, 314, 317, 318n16, 319–321, 325–329, 331, 333–334
- Scalable Vector Graphics (svg) 300, 304, 307  
*See also* EpiDoc; Extensible Markup Language
- Scalar multiplication 284n6  
*See also* vector
- Scoring method 241
- Scraper 204  
*See also* machine-actionable data/machine-readable data; Open Richly Annotated Cuneiform Corpus; tag; token  
*See also in General Index* lemma
- Scraping 290n28
- Self-Organizing Map (SOM) 65n20–21, 94  
*See also* Artificial Neural Network; multivariate analysis; node; unsupervised method



- Semantics 343, 346n25, 361
- Semantic technologies 344
  - See also* algorithm; Artificial Intelligence; data mining; Natural Language Processing
- Semantic unit 285, 286n14, 296, 298, 300–301, 304, 310
- Semantic Web (sw) 46, 227, 336, 338–339, 343–344, 347
- Semantic Web (sw) technologies 337, 339–341, 343–344, 348–349, 360–361
  - See also* Resource Description Framework; Simple Knowledge Organization System
- Semantic website 171n92
- Semantics of a recording system 38
  - See also* arbitrary unit
- Serialization format 299
- Shading attribute 295
  - See also* Manuel de Codage
  - See also in General Index* sign
- Simple Knowledge Organization System (skos) 46, 48n91, 353
  - See also* data cloud; Linked Data; machine-actionable data/machine-readable data; Resource Description Framework; Semantic Web
- skos (Simple Knowledge Organization System) 46, 48n91, 353
- skosifying 48
  - See also* Simple Knowledge Organization System
- Social network analysis 195, 220, 227, 332
  - See also* network graph
  - See also in General Index* administrative texts
- SOM (Self-Organizing Map) 65n20–21, 94
- SPAD (Système Portable pour l'Analyse des Données) 97–98
- SPARQL 340, 345, 361
- Spatial semantic 69–70
- Special characters 285, 286n13, 287–288
  - See also* CAL Code
  - See also in General Index* sign
- SQL (Structured Query Language) 123n42, 196n7
- Standardization/standardized 197, 200, 267, 299n59–60, 306n86, 321
- STAR project 47
- Statistical method 115, 117–118, 232, 238, 253
- STELLAR project 47
- String 268, 347
- Structured Query Language (SQL) 123n42, 196n7
  - See also* MySQL; querying; relational database
- Subject 46, 345–348
- Subjective variables 161, 169, 178, 181, 185–186
- Supervised method 260–261, 266
  - See also* algorithm; unsupervised method
- svg (Scalable Vector Graphics) 300, 304, 307
- sw (Semantic Web) 46, 227, 336, 338–339, 343–344, 347
- sw technologies (Semantic Web technologies) 337, 339–341, 343–344, 348–349, 360–361
- Syntactic elements 36
  - See also* event
- Système Portable pour l'Analyse des Données (SPAD) 97–98
  - See also* data mining; Graphical User Interface
- Tag 42, 44, 163, 164n66, 165, 176, 184, 289n27, 302, 339
  - See also* attribute; EpiDoc; Extensible Markup Language; markup; namespace; Text Encoding Initiative
- TEI (Text Encoding Initiative)/TEI-XML 154, 162–167, 185, 203–204, 289, 302, 326, 352
- TEI-C (Text Encoding Initiative Consortium) 164, 326n
- Term Frequency–Inverse Document Frequency (TF-IDF) 271n34
- Text data 165
- Text Encoding Initiative (TEI) 154, 162–167, 185, 203–204, 289, 302, 326, 352
  - See also* attribute; element; EpiDoc; interoperability
- Text Encoding Initiative Consortium (TEI-C) 164, 326n
- Text Mechanic Combination Generator Tool 207n52
- Text mining 154, 162–163, 181–182, 185–186, 196
  - See also* qualitative method



- TF-IDF (Term Frequency–Inverse Document Frequency) 271n34
- Token 203–207  
*See also* dictionary  
*See also in General Index* lemma; sign
- Tokenization/tokenization process 166, 197, 203, 204n41  
*See also* attribute; Text Encoding Initiative  
*See also in General Index* sign; transliteration
- Toolkit 238, 247–248, 283n2
- Topological Weighted Centroid (TWC) 70–72, 74–75
- Topological Weighted Centroid (TWC) mathematical approach 70n30  
*See also* Topological Weighted Centroid
- Tree-graph 66–67, 69
- Triple 46, 49n95, 201–203, 211, 345–347, 360  
*See also* edge; network graph triples; node  
*See also in General Index* tablet
- Triple store 46, 49n95, 51n98  
*See also* triple
- Tuple 47n84
- TURTLE 340
- TWC (Topological Weighted Centroid) 70–72, 74–75
- TWC mathematical approach (Topological Weighted Centroid mathematical approach) 70n30
- Unicode 285, 287–288, 300–301, 308, 310, 328, 351  
*See also* mapping; Unicode Consortium; Unicode Transformation Format (UTF)-8
- Unicode Consortium 328n34  
*See also* Unicode
- Unicode Transformation Format (UTF)-8 289n25, 301n68  
*See also* EpiDoc
- Universal Resource Identifiers (URIS) 302n72, 338–340, 344, 346, 348  
*See also* HyperText Transfer Protocol; HyperText Transfer Protocol Universal Resource Identifiers; identifier; web
- Unsupervised method 260–261, 266  
*See also* algorithm; Self-Organizing Map; supervised method
- Unweighted Pair Group Method with Arithmetic Mean (UPGMA) 274–275  
*See also* clustering method
- UPGMA (Unweighted Pair Group Method with Arithmetic Mean) 274–275
- Upper Ontology 320  
*See also* ontology
- URIS (Universal Resource Identifiers) 302n72, 338–340, 344, 346, 348
- Variable(s) 61, 63–64, 66n26, 67n27, 69–72, 73n, 91n21, 99, 118, 120, 160, 162, 181, 185–186, 261n10, 337
- Vector 65n20, 246–248, 272, 284n6, 297n54  
*See also* Optical Character Recognition; Scalable Vector Graphics; scalar multiplication
- Vector addition 284n6  
*See also* vector
- Vector graph 310
- Vector graphic 296, 300, 304  
*See also* Optical Character Recognition; vector
- Vector image 296
- Vector space 271, 284n6, 285, 305–306  
*See also* scalar multiplication; vector; vector addition
- Vectorization process 297
- VerbNet 170
- Vertex 73, 306
- Visualization diagram 49n95
- w3c (World Wide Web Consortium) 46, 340, 344–345, 348, 361
- Web 253, 318, 319n23, 338–339, 347  
*See also* hyperlinks; HyperText Markup Language; HyperText Transfer Protocol; Linked Data; Universal Resource Identifiers; World Wide Web Consortium
- Web of data 339, 341
- Web of documents 339
- Web Ontology Language (OWL) 344, 345n21, 353, 357
- Weight 43, 67n28, 73, 201, 206, 208, 211, 240, 247  
*See also* edge; weighted

- Weighted 71, 73, 201–202, 211n61, 241, 243  
*See also* Cytoscape; edge; Gephi; weight
- Wizard 329–332  
*See also* workflow wizard  
*See also in General Index* lexicography
- Word sense induction 226, 252
- Word2vec 225, 229, 232–233, 242, 246–252  
*See also* Artificial Neural Network;  
 Continuous Bag-of-Words model;  
 Natural Language Processing;  
 Pointwise Mutual Information;  
 querying; vector
- WordNet 164n65, 170
- Workflow wizard 314, 329, 332  
*See also* wizard
- World Wide Web Consortium (w3c) 46,  
 340, 344–345, 348, 361  
*See also* web
- Writing system 321, 327, 328n34  
*See also* Online Cultural and Historical  
 Research Environment  
*See also in General Index* Akkadian;  
 logossyllabic; Sumerian
- XDR (External Data Representation) 299
- XML (Extensible Markup Language) 51n98,  
 163–164, 203–204, 289n27, 299, 302n72,  
 303–305, 319, 334, 339–340, 347
- XML annotation (Extensible Markup  
 Language annotation) 204
- XML-Elamite standard 287  
*See also* human readable  
*See also in General Index* Elamite;  
 transliteration
- XML-scheme 289

# General Index\* (including terms associated with Archaeology, History, Geography, Literature, Philology, and their methods)

- ʾIlimilku (the Šubbanite scribe of Ugarit)  
151–152, 154, 159, 172n95
- ʾAnatu 13, 151–154, 157, 162, 165n68, 165n70,  
169, 174–176, 183n, 184–186, 188
- a-tu 218
- Abu Salabiḥ 350, 370
- Abu-Šahreīn (ancient name Eridu) 370–371
- Achaemenid Empire 292n35, 301n70
- Actancial event 160, 186
- Action(s) 13, 151, 153–154, 157–162, 167–170,  
174–182, 183n, 184–186, 188
- Action-oriented solution 167, 175
- Adab (modern name Bismiya) 96–97,  
198–199, 202, 210–212, 214, 215n70, 216,  
218n77, 350, 370–371
- Adab 0800 + 1011 215n70
- Adab corpus 205n47, 210, 212–216, 218–219
- Adab Mama-Ummi textile workers archive  
219
- Administrative texts 195, 197, 201–202
- Agency 112, 151, 167–168, 177, 186
- Akka/Akko (modern name Tell al-Fuḥar/Tel  
Akko) 370–371
- Akkad 64n15, 371
- Akkadian 13–14, 90, 92, 196, 224–229,  
231–238, 251–252, 283n1, 292, 301n70,  
307n92, 315n6, 316n11, 321–322, 326–328,  
331, 334, 341, 351
- Akkadian period 90n19
- al-Hiba (ancient name Lagaš) 370, 372
- al-Ubaid 371
- Aleppo (modern name Ḥalab) 317, 324,  
370–371
- Allograph 321n26
- ama-keš (ama-kesz3) 218
- Annotation 166n74, 194, 203–205, 225
- Apollo 290
- Aramaic 288–289, 291n32, 292n34
- Archaeological site 42n67, 48, 63–64, 65n21,  
66, 71, 73, 85, 315, 319, 334, 341
- Area (in the context of archaeology) 26n10,  
34, 39, 40n61, 41, 43, 48–49, 51, 69, 96n33
- Artifact 12, 17, 31, 33–34, 37–38, 42–44, 52,  
85–90, 92, 96, 111, 113, 116–119, 136, 199, 224,  
354
- Assyrian 341
- Astrology 257, 261–262, 265, 278
- Attested form 322–323, 329–330, 333
- Attribute (in Archaeology) 33–34, 43, 50
- Auramazda 287
- Baʿlu 13, 152–154, 165n70, 176n102
- Baʿlu and ʾAnatu Cycle (KTU 1.1–6) 13,  
152–153, 163, 165n70
- Babil/Ḥillah (ancient name Babylon)  
370–371
- Babylon (modern name Babil/Ḥillah)  
96–97, 262, 370–371

\* Note that the index follows the system for alphabetical order used in Gregorio del Olmo Lete and Joaquín Sanmartín, 2015, *A Dictionary of the Ugaritic Language in the Alphabetic Tradition*, Leiden: Brill. Thus, the Ugaritic signs *aleph*, ʾ, (here transliterated and vocalized /ʾi/) and *ayin*, ʿ, (here transliterated and vocalized /ʾa/ for “ʾAnatu”) stand at the top of the index list when one of these diacritic signs is the first letter of a noun—for further explanation, see in this volume, the page (xiii) on “Phonology.” For a proper noun, we traditionally add a capital letter after a diacritic sign, i.e., “ʾIlimilku.” Commonly, as well as for practical reasons, a consonant with a diacritic sign follows the consonant without diacritic sign in the index, i.e., “Hittite, Ḥalab.”

- Babylonian 62n11, 156, 234n35, 257, 341, 261, 262n14, 263
- Babylonian zodiac 258
- Batch 42–43
- Beeršeba (modern name Bir al-Sabaʿ) 370–371
- Behistun inscription 302
- Bir al-Sabaʿ (ancient name Beeršeba) 370–371
- Bismiya (ancient name Adab) 370–371
- BM 103232 97n34
- BM 13032A 97n34
- BM 13080A 97n34
- BM 130707 97n34
- BM 115418 97n34
- British Museum (London) 262n13, 341, 355
- Calendar Texts 263, 268–269
- Cemetery of Demircihüyük-Sarıket 125
- Clay tablet/cuneiform tablet 199n18, 316, 345, 354
- Cognitive linguistics 170, 230
- Colophon 345
- Composite texts 342, 354, 356
- Cuneiform/cuneiform script 13–15, 152, 165n69, 166n73, 194–197, 199–201, 205, 214n68, 219–221, 224, 227–228, 257, 267, 285–286, 291n32, 292, 316, 321, 323n, 328, 340, 341n11, 343, 345, 351, 354
- CUSAS 11, 050 215n70
- CUSAS 11, 052 218n77
- CUSAS 11, 084 215n70
- CUSAS 11, 129 215n70
- CUSAS 11, 212 213n64
- CUSAS 11, 238 218n77
- CUSAS 11, 285 218n77
- CUSAS 11, 356 213n64
- CUSAS 13, 008 215n70
- CUSAS 13, 134 205
- CUSAS 19, 118 215n70
- CUSAS 19, 179 215n70
- CUSAS 20, 066 203n37
- CUSAS 20, 067 203n37
- Cylinder seals 85, 87–92
- Çatalhöyük 35, 370
- Daily top-plan 26
- Damascus (ancient name Dimašqa) 317, 370–371
- Darius I 14, 292, 308n95, 310
- Demircihüyük/Demircihüyük-Sarıket 124–129, 133, 370
- Demircihüyük figurines 125
- Deontic powers 176
- Desire 160n48, 162n55, 178, 180
- Determinative 267–268, 316n11, 326–328
- Dilmun (part of modern Failaka Island) 366, 370–371
- Dimašqa (modern name Damascus) 370–371
- Dor (Tel Dor, modern name Ḥirbet al-Burj) 25, 27, 29, 33, 43, 48, 51–52, 370–371
- Dor expedition 25, 27
- Drehem (ancient name Puzriš-Dagan) 96–97, 370, 372
- E-dam temple 213
- E-tur temple 210, 213
- Early Bronze Age 12, 114–115, 124–125, 128
- Early Dynastic period 65n21, 210
- Early Ur III 100
- Ebla (modern name Tell Mardih) 66, 350, 371
- Ebla Royal Mausoleums 66
- Ecofact 42n68, 43
- EDM (Electronic Distance Measurement) 27
- Elam 287n16, 365
- Elamite 14, 224n3, 286–287, 292, 294–295n43, 295, 302n71, 341
- Electronic Distance Measurement (EDM) 27
- Emic approach 226, 229, 247n54, 249, 251
- Empiricism 154, 157, 159, 161
- Emotion 161, 170, 178–180, 183n, 184–185
- Entanglement (in Archaeology) 117
- Eridu (modern name Abu-Šahreïn) 370–371
- Ešnunna (modern name Tell al-Asmar) 96–97, 370–371
- Failaka Island (part of ancient Dilmun) 370–371
- Fara (ancient name Šuruppak) 370, 372
- Formulaic 201
- Garšana 96–97, 371
- Geme-Enlil 202
- geme2 202n37, 210n59
- Gender role 168, 177, 186

- Girsu (modern name Telloh) 210, 350, 371  
 Girsu Early Dynastic fishermen archive 219  
 GLAM (Galleries, Libraries, Archives, and Museums) 343  
 Glyph 165, 166n73, 169, 291–292, 301, 304  
 Glyptic 12, 87, 89, 91, 105  
 Gradešnitsa 370  
 Gradešnitsa figurines 119  
 Grapheme 314n4, 323
- Haifa 25–26, 370–371  
 Hellenistic Babylonia 261  
 Hermeneutics 155–156, 257  
 Hermeneutics of action 13, 151, 153–154, 157, 161–162, 166, 168, 170, 174n96, 177, 179, 181–182, 184, 186  
 Hittite 224n3, 227, 341  
 Ḫalab (ancient name Aleppo) 370–371  
 Ḫillah/Babil (ancient name Babylon) 370–371  
 Ḫirbet al-Burj (ancient name Dor) 370–371
- ICOMOS (International Council on Monuments and Sites) 45  
 Implicature 157  
 Inclusion 30, 128, 263, 265, 276  
 Indo-Iranian language 301n70  
 Inflection 168n84, 225, 227  
 Intention 151, 156–160, 166, 180, 186  
 Intentionality 156–160, 180–181, 184  
 Intersubjective phenomenon 159, 162  
 Intermediary element 291, 309  
 International Council on Monuments and Sites (ICOMOS) 45  
 Iqīša (a scribe) 265  
 Iraq Museum (Baghdad) 262n13
- Jansen-Winkel 57103 288  
 Jemdet Nasr 12, 64n15, 65n18, 72, 370  
 Jemdet Nasr period 65n18, 65n21  
 Jericho (modern name Tell al-Sultan) 39, 371–372  
 Jerusalem 25, 27, 370, 372
- Kiš (modern name Tell al-Uḫaymir) 64n15, 371–372  
 KTU 1.1–6 (Baʿlu and ʿAnatu Cycle) 13, 151–152, 163
- La Sapienza University of Rome 63  
 Lagaš (modern name al-Hiba) 96–97, 103–105, 370, 372  
 Laodikeia (modern name Lattakia) 370, 372  
 Larsa (modern name Tell al-Senkereh) 96–97, 371–372  
 Late Babylonian 257, 261, 264, 276  
 Late-Babylonian astrology 257  
 Late Chalcolithic 114  
 Late Uruk period 65n21  
 Lattakia (ancient name Laodikeia) 315, 370, 372  
 Layer (in Archaeology) 38, 42, 126  
 LBAT 1593 265n20  
 Lemma (pl. lemmata) 203–206, 227–228, 233n, 301, 321–323, 329–331, 333  
 Lexicography 329–331  
 Linguistic departure points 229, 251  
 Lippmann Coll 189 202n36  
 Lippmann Coll 211 217n75  
 Literary Sumerology 350  
 Locus (pl. loci) 26, 28, 33, 36, 38–42, 48–49, 51–52  
 Logographic sign 225  
 Logograms 165n69, 316n11 326–327  
 Logosyllabic 314n4, 316, 321, 326–327, 331, 334  
 Louvre (Paris) 262n13, 317, 355  
 Lukalla (a scribe) 345–346
- Mama-ummi 202, 215, 217, 219  
 Mama-ummi archive 210, 212  
 Mari (modern name Tell Ḫariri) 350, 370, 372  
 Medical ingredient 258, 263–266, 275, 277–278  
 Megiddo (modern name Tell al-Mutasallim) 41, 52, 371–372  
 Mesag 215  
 Mesopotamia 12–13, 85, 195, 224n2, 252, 261, 264, 316  
 Mesopotamian Civilizations 224, 226  
 Mesopotamian Urban Revolution 12, 60–63, 65n18, 71  
 Mesopotamian Urban Revolution Landscape (MURL) 62–64, 66, 69–73  
 Micro-zodiac text 259n4, 263, 264n16, 269, 277–278

- Minet al-Beida 316, 371  
 MS 4049 202n36  
 MURL (Mesopotamian Urban Revolution Landscape) 62–64, 66, 69–73
- National Museum (Aleppo) 317, 324  
 National Museum (Iraq) 341  
 National Museum (Damascus) 317  
 Neo-Assyrian 228, 237, 265, 277  
 Neo-Sumerian 90n19  
 Neolithic anthropomorphic figurines 118  
 Neolithic figurines 118  
 Neolithic package 115  
 Nineveh 228  
 Nippur (modern name Nuffar) 64n15, 65n21, 96–97, 264, 350, 371–372  
 Nuffar (ancient name Nippur) 371–372
- Objectivity 12, 38n55, 112, 115, 117, 161, 348  
 Old Akkadian period 196n10, 199, 201n32, 202, 218  
 Old Assyrian 213n66  
 Old Persian 14, 292, 301n70, 302n71  
 Oriental Institute of the University of Chicago 314n1, 317
- Paradigmatic relationships 231–233, 235–238, 242, 245–246, 249–252  
 Patron-client relationship 329n  
 Persepolis 371–372  
 Persian period 50–51  
 PF 404 286  
 Philosophy of action 154, 158, 161, 167  
 Phonemic form 322, 329  
 Phonograms 316n11, 326–328  
 Physical medium 285, 291n31  
 Pragmatics 154, 157, 169, 174  
 Puzriš-Dagan (modern name Drehem) 370, 372
- Ras Šamra (ancient name Ugarit) 14, 152, 314–315, 317, 319, 332, 371–372  
 Raw data of an excavation 44  
 Reading 156n30, 160, 257–258, 300, 321n26, 327  
 Regenstein Library (Chicago) 318  
 Royal Cemetery of Ur 341  
 Royal inscriptions 245–246, 253, 277, 301n70
- Royal Palace (archaeological site, Ugarit) 319, 324n  
 RS 3.320 324  
 RS 5.031 176n102  
 RS 15.076 327  
 RS 17.238 326–327
- SAA 18 234n35  
 Script unit 321, 328  
 Semantics (in the context of Linguistics and Philology) 154, 157, 169, 174, 186, 229n17, 252, 343  
 Semeion Research Center 60n\*, 63  
 Sign (in the context of semantic unit, including its digital analyze) 199n18, 203, 204n41, 224n3, 225, 258, 262–263, 265–266, 285–286, 291n31–32, 294n43, 295–298, 300–301, 307–309, 314, 315n5, 316n11, 321–324, 327–328, 333, 343  
 Sippar (modern name Tell Abu-Ḥabbah) 96–97, 370, 372  
 Slip 119, 128, 141  
 Spatial unit 26, 38–39, 43, 49  
 Static archaeological record 31  
 Stratum 40–41, 52  
 Subartu 96–97  
 Subjectivity 12, 112, 115, 117, 119, 123, 137, 161, 178  
 Sumer 64, 361  
 Sumerian (language) 13–14, 17, 196n10, 224–227, 316n13, 321, 336, 341, 349–351, 354, 359  
 Sumerian narrative/Sumerian literature 227, 350–352, 356, 359–361  
 Sumerian periods 350  
 Sumero-Akkadian writing system 327  
 Susa (ancient name Šuš) 96–97, 100, 102, 104–105, 371–372  
 Syllabic sign 159n46, 225, 323  
 Syntactic influence 301  
 Syntagmatic relationships 233–237, 242, 244, 246, 251  
 Syrian and French Mission at Ras Shamra 317n15  
 Šuruppak (modern name Fara) 351, 370, 372  
 Šuš (ancient name Susa) 371–372
- TAD C3.7 289

- Tablet (in the context of the Assyriology) 343  
 TCBI 1, 207 215n70  
 Tel Bataš (ancient name Timnah) 370, 372  
 Tel Dor (Dor, modern name Hîrbet al-Burj)  
 12, 25, 33  
 Tell Abu-Ḥabbah (ancient name Sippar)  
 370, 372  
 Tell Abu-Hawam 36–37, 370  
 Tell al-Asmar (ancient name Ešnunna)  
 370–371  
 Tell al-Fuḥar/Tel Akko (ancient name Akka/  
 Akko) 26, 370–371  
 Tell al-Mutasallim (ancient name Megiddo)  
 371–372  
 Tell al-Senkereh (ancient name Larsa)  
 371–372  
 Tell al-Sultan (ancient name Jericho)  
 371–372  
 Tell al-Uḥaymir (ancient name Kiš) 371–372  
 Tell Ḥariri (ancient name Mari) 370, 372  
 Tell Jemmeh 41, 370  
 Tell Joḥa (ancient name Umma) 370, 372  
 Tell Mardīḥ (ancient name Ebla) 66, 371  
 Tell Muqayyar (ancient name Ur) 70,  
 371–372  
 Telloh (ancient name Girsu) 96–97, 100–101,  
 104–105, 371  
 Temple (at Dor) 25, 38–39, 51  
 Temple of Bau of Girsu 210  
 Terracotta 51  
 Theodolite 27  
 Timnah (modern name Tel Bataš) 370, 372  
 Transliteration 199–200, 202–205, 218,  
 224n4, 267, 286–289, 291, 294n43, 314, 317,  
 324–326, 328, 341–342, 351–352, 354–356  
 Ubaid 12, 64, 66–67, 71–72  
 Ubaid clay sickles 67  
 Ugarit (modern name Ras Šamra) 14,  
 151–152, 176, 186, 314–317, 332, 371–372  
 Ugaritic 13, 153n12, 165n69, 166n73, 170n88,  
 171n92, 172n95, 180, 316–317, 324, 325n,  
 327–328  
 Ugaritic literature/Ugaritic narrative 17,  
 151–152, 169n84, 177n104  
 Umma (modern name Tell Joḥa) 96–97,  
 102–105, 370, 372  
 University of Pennsylvania Museum of  
 Archaeology and Anthropology  
 (Philadelphia) 341  
 Ur (modern name Tell al-Muqayyar) 69–71,  
 96–97, 100, 104–105, 341, 371–372  
 Ur III period 12, 90, 92, 100, 105, 349  
 Ur-ga 216  
 Ur-Nammu 350  
 Ur-Ninsun 199, 215–216  
 Ur-nu 216  
 Urban Revolution 62, 64n15, 64n17, 65n18,  
 71–72  
 Uruk (modern name Warka) 12, 64–66,  
 70–72, 96–97, 228, 262, 371–372  
 Uruk clay cones 67  
 Vase of Darius 292, 308n95, 310  
 VAT 7815 276  
 VAT 7816 276  
 VAT 7847 264n16  
 Verbal semantics 167, 171  
 Vindolanda 343, 369, 371  
 Vorderasiatisches Museum (Berlin) 262n13  
 Warka (ancient name Uruk) 65, 70, 371–372  
 Word sense induction 226, 252  
 Zodiacal sign 258, 262–263, 265–266, 278