# Graph Data-Models and Semantic Web Technologies in Scholarly Digital Editing

Schriften des
Instituts für Dokumentologie und Editorik



Band 15

Schriften des Instituts für Dokumentologie und Editorik — Band 15

# Graph Data-Models and Semantic Web Technologies in Scholarly Digital Editing

edited by

Elena Spadini, Francesca Tomasi, Georg Vogeler

2021

BoD, Norderstedt

**Bibliografische Information der Deutschen Nationalbibliothek:**
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über **http://dnb.d-nb.de/** abrufbar.

**Digitale Parallelfassung der gedruckten Publikation zur Archivierung im Kölner Universitäts-Publikations-Server (KUPS). Stand 5. Dezember 2021.**



Herstellung und Verlag: Books on Demand GmbH, Norderstedt
ISBN: 978-3-7543-4369-2
Einbandgestaltung: Stefan Dumont nach Vorarbeiten von Johanna Puhl und Katharina Weber
Satz: LuaTEX, Bernhard Assmann

# Contents

## Projects and Editions



## Appendices

# Infrastructures and Technologies

# The Power of OCHRE's Highly Atomic Graph Database Model for the Creation and Curation of Digital Text Editions

Miller C. Prosser, Sandra R. Schloen

**Abstract**

The Online Cultural and Historical Research Environment (OCHRE) is a research database platform that provides a suite of tools to aid in the curation of digital text editions. The power and flexibility of the OCHRE system is predicated on the underlying data model, which is constructed as a graph database. OCHRE is based on a semi-structured, item-based data model where data are atomized into granular items and organized through hierarchical arrangement and cross-cutting links. In this paper, we describe OCHRE's graph data model and demonstrate how this approach revolutionizes digital philology.

## 1 Introduction

The Online Cultural and Historical Research Environment (OCHRE) is a research database platform that provides a suite of tools to aid in the curation of digital text editions.[1] But that was not the original mandate of the program. OCHRE was created as a tool for archaeological data management.[2] Because no two archaeologists can be compelled to agree on a common system for excavation or on a single controlled vocabulary for describing their data, OCHRE was created to be customizable. Further, because an archaeologist typically needs to describe a wide variety of data, from a single botanical sample to an entire watershed region, OCHRE was designed to manage data at any level of abstraction or observation – indeed, at multiple levels of abstraction, or with multiple observations. Finally, because archaeology creates data of many types – such as images, geo-spatial (GIS), and daily journal entries to

[1] We use the term curation to refer to the activities and processes performed to capture, create, edit, publish and archive data—essentially the entire data life-cycle.

[2] Sandra R. Schloen and J. David Schloen invented OCHRE as a data management system for David Schloen's archaeological research. As a trained software developer, Sandra Schloen implemented their plan as a database platform. Over decades of use and decades of technological advancement, OCHRE has evolved through many phases. For more information on OCHRE in general, see Schloen and Schloen (2012).

*Graph Data-Models and Semantic Web Technologies in Scholarly Digital Editing.* Ed. by Elena Spadini, Francesca Tomasi and Georg Vogeler. Schriften des Instituts für Dokumentologie und Editorik 15. Norderstedt: Books on Demand, 2021. 55–71.

name only a few – OCHRE was built to integrate any and all types of project data.[3] The database, the underlying data model, and the user interface evolved to become highly flexible, generic, and extensible. For these reasons, it became clear that the system could be applied to a variety of other research domains. Based as it is at the Oriental Institute (OI) of the University of Chicago, the next logical application of the OCHRE system was the field of philology, one of the other core research areas at the OI. The same affordances granted the archaeologist are available to the philologist. Researchers are free to use a set of descriptive terms that are recognized in their specific area of study or to create their own knowledge representation vocabulary. Observations can be made by multiple authors or editors at any level of detail, from a single grapheme to an entire corpus. Data of all types are integrated in a common platform, allowing the presentation of text images, bibliography, and commentary along with textual data.

The power and flexibility of the OCHRE system is predicated on the underlying data model, which is constructed as a graph database. OCHRE is based on a semi-structured, item-based data model, where data are atomized into granular items and organized through hierarchical arrangement and cross-cutting links. In this paper, we describe OCHRE's graph data model and demonstrate how this approach revolutionizes digital philology.

## 2 The OCHRE Data Model

As mentioned already, data in OCHRE is highly atomized. By atomized, we mean that data is broken down into minimal meaningful parts, each of which is stored as a separate XML document.[4] Data is not stored in a tabular format – the data model used by a typical relational database. Data is not stored in fully composed and marked-up documents – the data model often used as the default approach for textual research. Instead, in OCHRE, items are arranged in a semi-structured, hierarchical model, an arrangement that is supplemented by cross-cutting links between items. The result is an integrated graph of data nodes. These nodes are categorized according to a generic upper ontology that specifies the classes and relationships between the nodes. The high-level data categories in OCHRE are the following: Agents, Bibliography, Concepts, Dictionary units, Periods, Resources, Spatial units, Texts, and Writing systems. Agents are people, real or fictional, ancient or modern, including even

[3] On issues related to using digital tools for archaeological data management, see Prosser (2020), "Digging for Data."

[4] OCHRE is implemented using the Tamino XML database, the first enterprise-level native-XML database (developed by Software AG, Germany). While data in OCHRE is stored as documents, the database is more accurately described as a semi-structured graph database rather than as a document-oriented database because it is characterized by relationships between highly atomized database items.

project team members. Resources are images, PDFs, audio files, or any other external file. Spatial units are any items, real or otherwise, that can be contextualized in space, meaning that they can be organized according to their location. A Spatial unit may have a latitude-longitude coordinate; it may be a single point or an entire region. Along with space, time is also data, which is recorded in OCHRE's Periods category. Even the controlled vocabulary of the project is stored as data. A hierarchy of variables and values that define the descriptive properties of all the other items forms a project Taxonomy. OCHRE manages a master taxonomy, from which projects may borrow to create their own local taxonomy. However, project personnel can customize their taxonomy to include unique variables and values needed to describe their research. Any OCHRE item can be described with properties as allowed by the project taxonomy. In addition, any OCHRE item can be linked to any other OCHRE item(s) using a variety of mechanisms. As we go on to define the details of the Text category of items below, keep in mind that, at its core, all data in OCHRE is a network of data organized within these high-level categories, or node classes, described by properties, and related by links.

### 2.1 Textual Data

Contrasting with what can be called the document model, wherein a series of string characters are stored in a sequence that corresponds to the layout of the text on the page, OCHRE's data model breaks down textual data into items that correspond to either words or graphemes.[5] Each item is uniquely identified and stored as a separate XML document. These items can be combined and organized to produce any variety of derived formats appropriate for viewing, analysis, publication, or even good old-fashioned printing. Each item can be addressed individually by the researcher: identified, commented upon, or reused in a variety of overlapping hierarchical contexts.

To make this point from another perspective, when guidelines from the Text Encoding Initiative (TEI consortium 2019) are used to record and describe the structure of a text, the result is a richly marked up single XML document for the entire text.[6]

---

[5] We use the linguistic term grapheme to refer to any minimal and meaningful unit of writing, more colloquially referred to as a letter, sign, accent, or punctuation. On the difference between a document model and a database model, see Schloen and Schloen (2014).

[6] We bring TEI into the discussion here because it is so commonly implemented in textual studies and because we want to emphasize that our use of XML is not the same as TEI-XML. In the simplest implementation of TEI markup, an entire text is stored as a single XML file. There is good reason for doing this and we have no criticism of this approach. It has utility in certain contexts. Textual data from OCHRE can be exported, then styled with an XSLT stylesheet to create well-formed TEI-XML. We allow the researcher to make this transformation to TEI to avoid imposing any single implementation of TEI on a researcher.

By contrast, in the OCHRE data model, we create a separate XML document for every word or grapheme or *minimal meaningful part* of the text. So, instead of one XML document per text, we work with hundreds or thousands of XML documents for each text.[7] The structure of the text is represented, in part, by hierarchical arrangement of these items. The hierarchy, itself, is a separate XML document that organizes its content. That is, the hierarchy has links to those items that it contains, and the items in turn link back to the (potentially many different) hierarchies in which they are contained. This approach allows any item to participate in multiple hierarchies. OCHRE represents this complex network of data in a natural and intuitive user interface. To the end user, a text view looks very much like a sequence of string characters.[8]

To illustrate the highly atomic and granular data model implemented in OCHRE, let us turn to the manner in which textual data is represented. First, we separate the idea of the object and the text. The object is what the paleographer refers to as the writing support, i.e., the surface or object on which the text is present. The object, a Spatial unit in OCHRE's ontology, has its own set of metadata properties and descriptions. Objects may have coordinates that represent where they were discovered – important for archaeology projects; or a designation to indicate where they are stored – useful for managing an inventory of objects in a museum context.

As distinct from the object, the Text item in OCHRE is composed of two major classes of items.[9] Specifically, a text consists of collections of epigraphic units and discourse units. The class of items called epigraphic units are organized hierarchically as recursive elements to describe the epigraphic layout of the text on the writing support, whether it be a folio, page, or ancient clay tablet. From broadest to most narrow, the recursive hierarchy of epigraphic units may be, as one example, Recto, Line 01, Latin uppercase D.

In addition to the epigraphic hierarchy, a text is defined using a discourse hierarchy, which represents a scholarly interpretation of the text. Any text may consist of multiple discourse hierarchies: for example, one for a word-by-word interpretation, one for a poetic analysis, or one for a syntactic analysis. A text may have multiple

---

7 We admit that it may well be possible to implement a highly atomized and granular approach to texts using TEI. It may be possible to utilize the pointing mechanism of TEI to define a text as being composed of thousands of external files. However, in this scenario, one wonders what value is gained by encoding the XML files as TEI. The highly atomized approach we use in OCHRE is better implemented as a graph database. On the use of pointers in TEI see TEI Consortium 2019, ch. 16. https://www.tei-c.org/release/doc/tei-p5-doc/en/html/SA.html#SAXP, last accessed April 2, 2019.

8 OCHRE has a sophisticated import tool built in to atomize a text document of the ordinary kind into a network of highly granular but related parts. The user need only set a few options and click the *Import* button.

9 Here we use the word "class" to refer to a component of an ontology, a category of similar concepts.

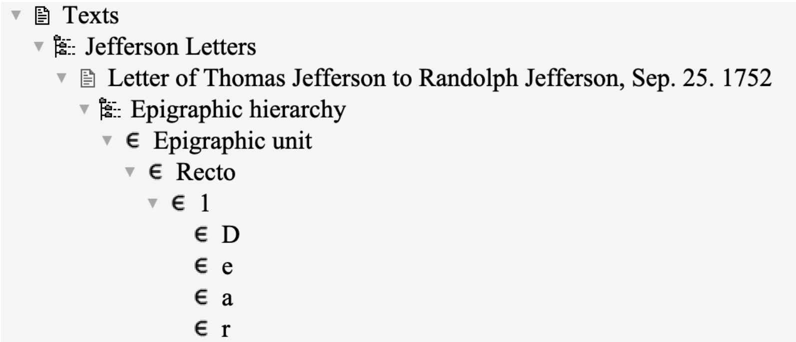


Figure 1. Partial view of the epigraphic structure of a Jefferson Letter.

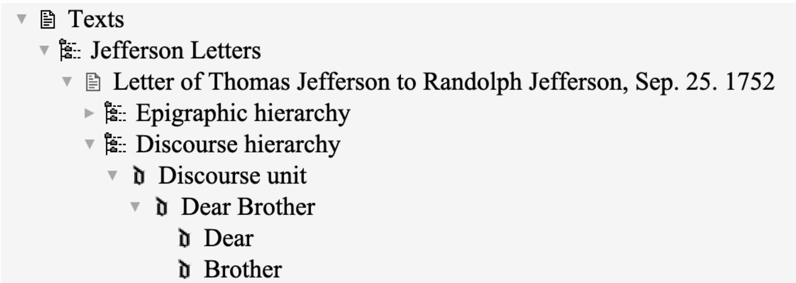


Figure 2. Partial view of the discourse hierarchy of the Jefferson Letter.

discourse hierarchies that represent interpretations of various scholars.[10] A discourse hierarchy consists of discourse units such as paragraph, sentence, stich, phrase, clause, or word. The smallest discourse unit is usually the word. Any given discourse unit can be reused in multiple discourse hierarchies. The same XML file that represents the word, for example, may point to a discourse hierarchy representing poetic analysis and a discourse hierarchy representing syntactic analysis. Reuse of data in this fashion solves the problem of multiple overlapping hierarchies in textual data. Further, the database item that represents the word is linked to the epigraphic units that represent the graphemes that are its constituent parts. Stated inversely, a series of epigraphic units is linked to a discourse unit.

To use the semantics of graph theory, a text is an atomized network of nodes and edges. An epigraphic unit – whether it be a grapheme or one of the larger epigraphic

[10] Especially in the world of ancient texts, there are often disagreements about the interpretations of graphemes and words.

sections – is available for reuse and sharing among multiple texts, because it is simply a node in a network. These multiple contexts may be editions prepared by different editors, or even separate texts copied by ancient scribes. We demonstrate below a practical application of this approach. An epigraphic unit records these network connections as a list of links that represents the edges that point to other nodes to which it is related. Each epigraphic unit is a node that may have an unlimited number of edges, i.e., the pointers that record where in the text a given node exists. In other words, there is no proscribed definitional boundary to the set of nodes that can link to each other. The node that represents a given word may link to any number of other larger discourse units such as couplets, lines, sentences, or paragraphs. Further, that same word may link to competing scholarly interpretations of couplets, lines, sentences, or paragraphs. This highly atomized graph approach to textual data provides an elegant solution to the problem of representing the same data in overlapping or competing hierarchies.[11]

Although the underlying OCHRE structure is modelled as a graph, OCHRE uses hierarchical structures to organize the nodes (items) and edges (links), rather than using node-edge style visualizations commonly associated with network analysis and graph databases. (See the two figures above for sample hierarchies.) It is worth noting that there is no need to choose between a hierarchy and a graph. A hierarchy *is* a graph, and OCHRE exploits the advantages of graphs in general, and hierarchies more specifically.[12] The hierarchical arrangement of graphemes within words, within phrases/clauses, within sentences/lines, within pages, and so on, is an intuitive construct, and closely parallels how scholars naturally work with, and think about, their textual data (and lexical, taxonomic, archaeological, and many other types of research data). OCHRE also recombines, on demand, the highly atomic items into composite views that are displayed to the scholar in familiar formats.

[11] The so-called *problem* of multiple, overlapping hierarchies when representing texts is well attested in the literature, usually in the context of applying markup (e.g., TEI) to documents. For a discussion of this issue see Schloen and Schloen (2014). But multiple overlapping hierarchies are unproblematic in the context of representing texts using an item-based approach, where each hierarchy is treated as a discrete item and even, potentially, representative of a work of scholarship. That is, each epigraphic or discourse hierarchy is simply a node of a graph that links to many other nodes (epigraphic or discourse units) in conformance to carefully applied rules.

[12] According to Robinson et al (2015, 109) "A graph database's structured yet schema-free data model" makes them "ideally applied to the modeling, storing, and querying of hierarchies…". Examples are given of how to represent "cross domain models" (*ibid.*, 41ff), which is analogous to how OCHRE manages multiple overlapping hierarchies without difficulty using a graph approach.

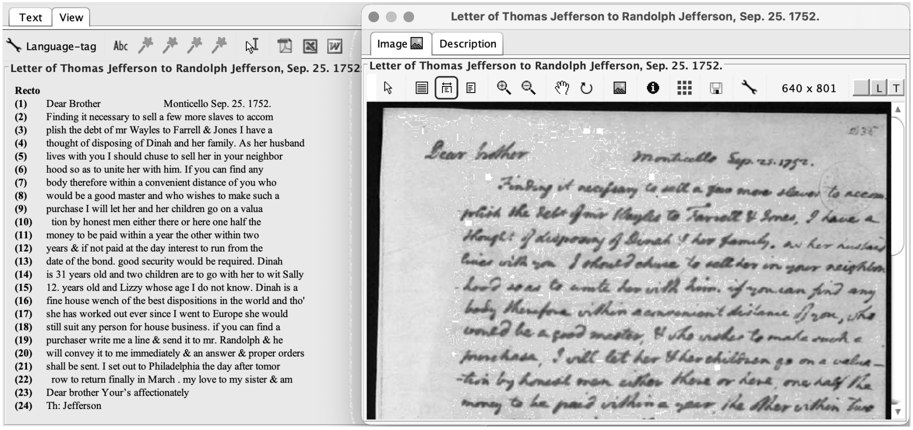


Figure 3. Recomposed document view of the Jefferson Letter example with associated image. (http://www.loc.gov/exhibits/jefferson/images/vc109a.jpg).

# 3 Critical Editions, Collations, and other Implementations

In the following section, we present two models for working with textual data in OCHRE: the text corpus model and the critical edition model. In the text corpus model, the researcher creates diplomatic editions of texts in a corpus. Typically, the researcher is establishing a new or updated edition of a text. In the critical edition model, the researcher may be creating text editions, but is also leveraging the graph data model to align various copies of a given text to compare manuscripts and trace transmission variations.

## 3.1 The Text Corpus Model

First, we illustrate a standard text corpus project and outline some of the most common tasks that take advantage of OCHRE's item-based approach. The *Ras Shamra Tablet Inventory* (RSTI) is an OCHRE research project, based at the Oriental Institute of the University of Chicago. RSTI is directed by the author (Prosser) and Dennis Pardee, professor of Northwest Semitic Philology in the Department of Near Eastern Languages and Civilizations at the University of Chicago. In this project, we organize our research on the culture of Late Bronze Age Ugarit. Ancient Ugarit (modern Ras Shamra) was a city and minor kingdom of the same name in what is now modern-day Syria. The site of Ugarit was occupied almost continuously for nearly six millennia,

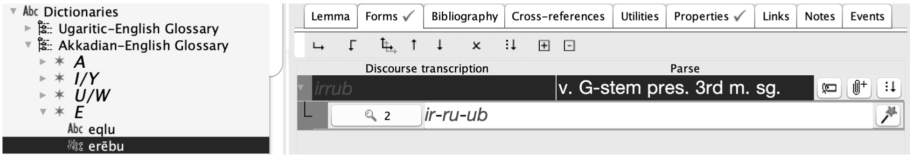


Figure 4. Dictionary unit for the Akkadian verb *erēbu*, "to enter," with the grammatical form *irrub*, and attested form *ir-ru-ub*.

from the Neolithic Period through to the beginning of the $12^{th}$ century BCE.[13] Since its discovery in 1929, archaeologists have uncovered thousands of tablets and other inscribed objects. To date, we have assembled a catalog of over five thousand objects in the project.[14] We are in the process of creating digital text editions of the texts recorded on these objects. Also, we have integrated the largest body of digital images of the Ras Shamra tablets.[15] The project serves two purposes: (1) it serves as a central repository for our research, and (2) it provides a mechanism through which we publish our data online. From a practical perspective, in RSTI we make a declarative statement – an interpretive observation – about every grapheme in every text. On this very specific level, we identify the reading of each cuneiform sign, record metadata about the level of preservation, and even write sometimes lengthy prose descriptions to document our epigraphic observations.

The glossary is where we record information about every lexeme attested in our corpus. Every word in a text – a unit of discourse and, conceptually, a node of the graph that represents the text – is associated with a Dictionary unit in a project glossary. A Dictionary unit consists of various grammatical forms, each of which may consist of various attested forms or spellings.

In RSTI, we have two glossaries, one for words in the Akkadian language and one for words in the Ugaritic language, each represented as a hierarchy of lexical entries. The Akkadian texts from Ras Shamra are written in a logosyllabic cuneiform writing system. The Ugaritic texts are, for the most part, written in an alphabetic cuneiform writing system. We transcribe both languages using the Latin script. The writing systems are also OCHRE items, capturing the attributes of different languages and

[13] See Yon et al (1995). Among the many reasons that Ugarit draws the attention of researchers is that the inhabitants of the site produced a fascinating corpus of textual material, from grand mythological tales, to personal letters, down to simple administrative records. See the contributions in Chapter 3 of Watson and Wyatt (1999).

[14] See the project website http://ochre.lib.uchicago.edu/RSTI/teo.html (last accessed September 22, 2020) and blog https://voices.uchicago.edu/rsti/ (last accessed March 17, 2019). For a description of how we have integrated GIS data with the object and text data, see Prosser (2018).

[15] Many of these photos were produced by the epigraphic team of the joint Syro-French Mission de Ras Shamra.

providing a catalog of all valid graphemes and values against which to match and validate textual content. In either system, any given grammatical form of a word may be written in a variety of orthographic forms. Further, there is a high degree of homography in both writing systems. Any given spelling may represent one of many words, or even one of many grammatical forms of those words. The structure of the glossary allows the researcher to disambiguate various word forms.

A word in a text is linked to an attested form in the glossary. Because this attested form is defined as a hierarchical child of a grammatical form, the word in the text inherits the grammatical properties assigned to the grammatical form in the dictionary.[16] In other words, OCHRE's data model represents the attested and grammatical forms as nodes in the graph that captures the relationships among these lexical entries.

Many of these texts are personal correspondence and other types of administrative documents from the royal palace: letters from the king and queen of Ugarit, letters to foreign dignitaries, letters of international intrigue, lists of land distribution to and from named individuals. These texts attest a wide variety of personal names. We are working to identify a prosopography of the persons named in these texts.[17] The graph data model has proven to be a powerful and flexible approach to this research goal. Specifically, the problem is to identify discrete individuals mentioned in texts, what we refer to in OCHRE as *agents*. In the Agent category, we identify an ancient person.[18] Here we can disambiguate various persons who share the same personal name.[19] In the semantics of graph networks, this person is a node in the graph. The many attestations of the person's name in texts are nodes of a discourse unit type. These are linked to the items (nodes) representing the persons being identified as agents. The network extends to include the glossary. Each attestation in a text is linked to a node in the glossary that represents the grammatical form of the name. In the end, we create a network of three nodal hubs: the agent, the textual attestations, and the grammatical entry. This arrangement helps us disambiguate names that are shared by different persons.

As mentioned above, RSTI includes a large collection of digital images of clay tablets. These images are all accessible through OCHRE, where they are presented alongside views of the text. To further integrate the textual and image data, we use a process called hotspotting to link graphemes to regions of the image where

---

[16] For a longer discussion of RSTI and its use of OCHRE to perform digital philology, see Prosser (2018), *Digital Philology*.

[17] In this context, we use the term prosopography to refer to the assembled set of familial, occupational, and other information that identifies a person and their relationships to other persons.

[18] See below a discussion of the ontological categories of data in OCHRE.

[19] In our texts, there are many persons identified by a single name only. In this period, there are no surnames, per se. Sometimes persons are listed with a patronymic affiliation, i.e., PN1 son of PN2. However, it is very common for persons to be identified by a single name only.

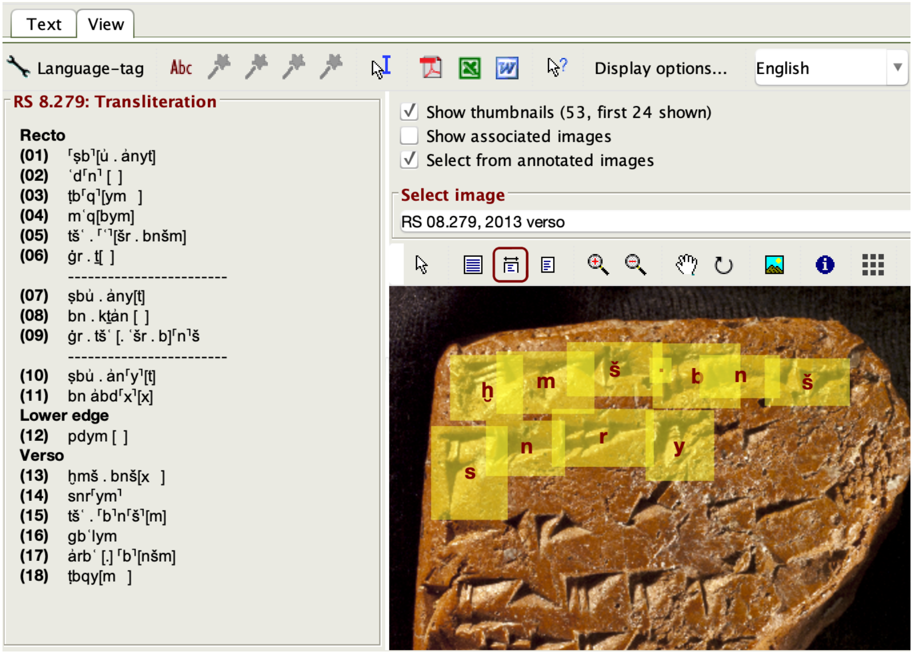


Figure 5. Image showing hotspot links between transcription and image (a click on ḫ in the image lights up the letter in line 13).

these graphemes are visible. In other words, each image can be marked up with polygons that are linked to epigraphic units in the text. When viewed together in OCHRE, the transcribed text can be synchronized with the hotspot polygons so that a click on a polygon in the image highlights the associated epigraphic unit in the text transcription. This tool has pedagogical utility, but it also useful for clarifying one's interpretation of a damaged sign.

## 3.2 Text Critical Model

The RSTI text corpus project in OCHRE is representative of numerous other projects that follow this same model: objects, texts, persons, images, analysis, and publication. A different group of text-based projects having different research goals and different source materials falls in a category that we think of as the text critical model. These projects, such as the *Critical Editions for Digital Analysis and Research* project (CEDAR), use the OCHRE platform to perform text criticism and to produce critical editions.[20] From the CEDAR website:

> [t]he goal of the project is to develop, test, and document new methods of digitally representing, displaying, and analyzing manuscripts, textual variants, and diverse editorial readings and translations, enabling views of these data that are not possible using traditional printed editions, with explicit representation of all the intra- and intertextual relationships a scholar may wish to note.[21]

These new methods benefit from reuse and sharing of items made possible by OCHRE's item-based approach, creating somewhat different kinds of graphs as compared to the text corpus projects, but using the same strategies of organizing (hierarchies) and linking of nodes. CEDAR demonstrates that the granular and generic structure of OCHRE suits text critical studies over a wide variety of text corpora. To date, CEDAR includes: (1) the Sumerian editions of the Gilgamesh epic; (2) Hebrew, Greek, Latin, Syriac, and Coptic manuscripts of selected chapters of Genesis, Proverbs, and Daniel; and (3) Shakespeare's Hamlet and Taming of the Shrew. Plans are already underway to expand CEDAR to include additional Biblical books, selections of Sanskrit and Middle Bengali literature, as well as editions of the Egyptian Book of the Dead. Note that text critical projects are an extended use case of the text corpora model, benefiting from all the features described above.

Using OCHRE's item-based approach, the CEDAR project compares textual variants across manuscripts on a letter-by-letter basis. For Biblical scholars interested in text criticism, it is standard to compare a wide variety of manuscripts to investigate the transmission history of a given text. For example, one may wish to compare Medieval Hebrew manuscripts like the Leningrad Codex, Dead Sea Scrolls that attest Biblical passages in Hebrew, Greek manuscripts of the Septuagint, and later Latin and Coptic

---

[20] The CEDAR project brings together University of Chicago faculty member from various departments: Simeon Chavel (Associate Professor of Hebrew Bible, The Divinity School), Whitney Cox (Associate Professor and Chair, Department of South Asian Languages & Civilizations), Thibaut d'Hubert (Associate Professor, Department of South Asian Languages & Civilizations), Ellen MacKay (Associate Professor, Department of English Language & Literature), David Schloen (Professor of Near Eastern Archaeology, Department of Near Eastern Languages and Civilizations), Jeffrey Stackert (Associate Professor of Hebrew Bible, The Divinity School), and Christopher Woods (John A. Wilson Professor of Sumerian, Department of Near Eastern Languages and Civilizations). The project has benefitted greatly from project personnel: Sarah Yardney (PhD, Divinity), Joseph Cross (PhD Candidate, NELC), Doren Snoek (PhD Candidate, NELC), Andrew Wilent (PhD Candidate, NELC), Ashleigh Cassemere-Stanfield (PhD student, English Language & Literature), Arianna Gass (PhD student, English Language & Literature), Sarah-Gray Lesley (PhD student, English Language & Literature), and Colton Siegmund (PhD student, Near Eastern Languages and Civilizations).

[21] See https://cedar.uchicago.edu/ (Last accessed March 19, 2019).

manuscripts. For these scholars, it is critically important that every letter – and even every accent, vowel mark, or punctuation – is a discrete unit for study.

Let's take the Hebrew text of Genesis chapter 1 as an illustrative example. The CEDAR team wishes to compare roughly a dozen Hebrew manuscripts that span about a millennium. Instead of representing each of these Hebrew texts individually, in CEDAR we create a single text that represents all actual content and potential textual variations of the theoretical Hebrew text of Genesis chapter 1. This single text is called a content pool. The content pool is a network of epigraphic units and discourse units, created as a text in exactly the same way as the texts from RSTI described above.

Here we celebrate the power of the item-based data model. Any item in the content pool can be reused, shared among any number of other specific representations of actual texts, what we call local texts. A local text has its own unique epigraphic and discourse hierarchies, but its epigraphic units are borrowed from the content pool. In other words, a user will pick and choose from among the content represented in the content pool, linking in the epigraphic units needed to articulate the structures of the local text. When a given manuscript – i.e., a local text – attests a variant, that variant is added to the content pool and is *aligned with* the non-variant letter by means of a targeted link, thereby extending the pool of available content. To allow for describing variants, adding annotations, or providing extended scholarly commentary, discourse units in local texts are not reused as-is from a content pool, but are auto-aligned with items in the content pool. This arrangement also allows us to compare and align manuscripts across different languages. See more on this below.

From the point of view of the individual nodes, any given epigraphic unit in the graph maintains a list of the necessary edges that identify the texts wherein it appears. For example, the very first letter of the Hebrew text of Genesis chapter 1 is a ב (*bêt*), a preposition meaning "in," as in the phrase, "In the beginning." This epigraphic unit exists in the content pool. The XML document that represents this item lists twelve paths to identify the twelve texts in which it appears. This letter is one node. The twelve texts are each a node. Similarly, each of the over 5,000 epigraphic units in Genesis chapter 1 maintains its own list of the contexts it which it is used, or rather, *reused.*

This systematic reuse of existing textual data whenever possible, eliminates the need for string manipulation when it comes time to compare multiple manuscripts. Texts are automatically aligned by virtue of sharing the exact same database items. Extensive reuse also eliminates the need to proliferate secondary links to make explicit alignments, thereby dramatically reducing redundancy since there is only one copy of the underlying textual content.

As a practical example, for the dozen Hebrew texts of Genesis chapter 1, we do not create a new letter ב (*bêt*) for each phrase, "In the beginning." We reuse the same ב (*bêt*)

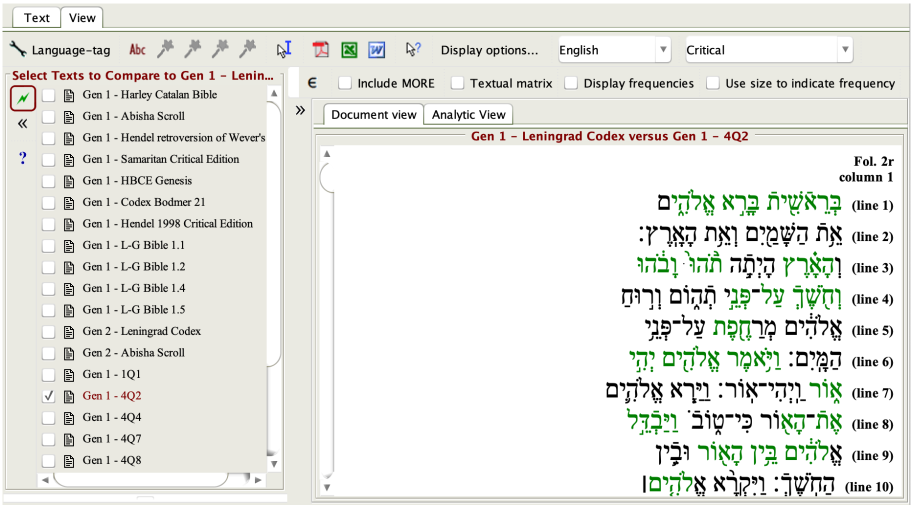


Figure 6. Comparing two Hebrew manuscripts of Genesis chapter 1. Green shows agreement, black shows no overlap.

by adding an edge to point to each text where it is attested. Instead of loading multiple texts, then comparing whether twelve different texts all contain the letter ב (*bêt*) in the opening phrase, the OCHRE network generates subsets of the single content pool to represent any specific text. In this way, because all similarities and differences are defined in the underlying data, comparisons do not need to be performed through a secondary process. Any given Hebrew text is simply the selection of nodes that contain the edges that define those nodes as part of the given text. In the opening phrase, "In the beginning," our ב (*bêt*) is the same ב (*bêt*) in every text.

To compare texts across languages, we use cross-cutting links. In a local text, a discourse unit that represents the Hebrew word for "created" is linked to a discourse unit in the Hebrew content pool that represents that same word. This edge creates the discourse relationship between the local text and content pool in Hebrew. The same is true for local texts in other languages, like Greek and Latin. Each discourse unit in those local texts is linked to the appropriate discourse unit in its respective content pool. Creating relationships across languages is done by aligning their content pools. In the Greek content pool, the Greek words for "in the beginning" are aligned with the Hebrew word for "in the beginning" in the Hebrew content pool.[22] Picture, then, the network of words, from a Hebrew local text, through to the Hebrew content pool,

[22] Relationships between discourse units in content pools of different languages can be one-to-one, one-to-many, or many-to-one.

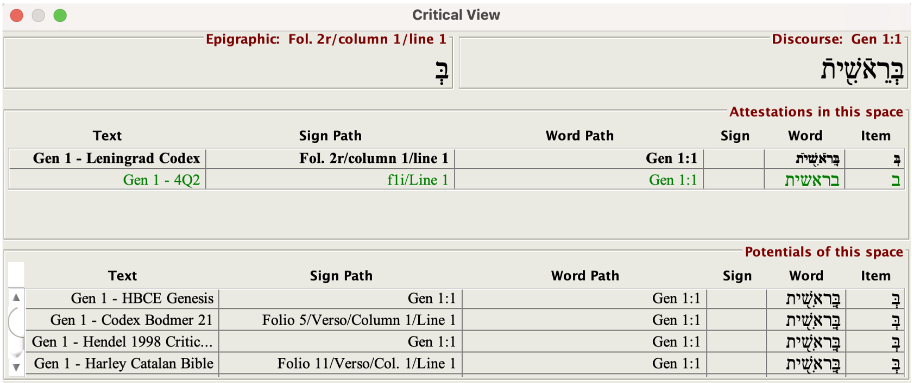


Figure 7. Details of the first letter of the verse, showing all texts where it is attested.

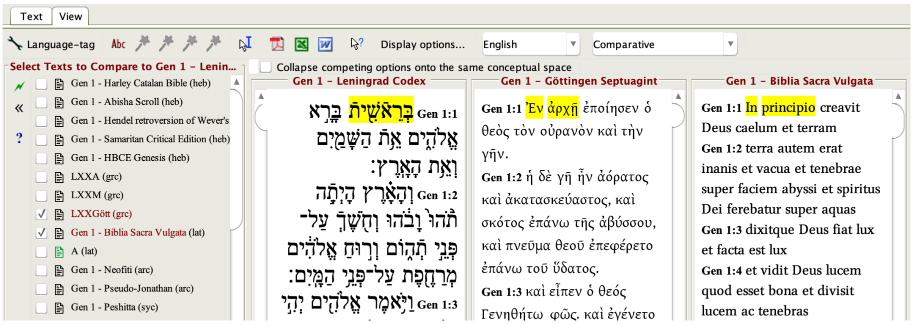


Figure 8. Hebrew, Latin, and Greek comparison with a click and automatic highlighting.

then to the Greek content pool, and finally to a Greek local text. This is the network that allows a user to compare local texts across various languages. Note that with a single mapping between the content of one language pool to the content in another language pool, by following the links from the local texts back to the content pool, *any* text in one language can be compared to *any* text in another language using the already-established links between their respective content pools.

## 4 Implications

The highly granular, item-based, generic model presented by OCHRE for content management will be seen as novel and unfamiliar to many scholars. However, the model is straightforward and easy to understand: items, described by properties, enriched by annotation, organized into hierarchies, and linked to other items. In addition, the undemanding playground of OCHRE's highly generic or *upper* ontology may seem insufficient to the task of competing in the richly-tagged, extensively marked-up world of TEI, or in the universe of the Semantic Web where rival standards – some old, some new – like Dublin Core, FOAF, CIDOC-CRM, FRBR, and so on, vie for adoption, or are constantly being extended. But, allow us to highlight some important implications of OCHRE's approach, not just for textual scholarship, but for scholarship in many academic domains.

First and foremost, the OCHRE approach offers flexibility. The high degree of atomization makes possible the modeling of any type of data. Indulge the following metaphor. If we start with prefabricated walls and floors and roofs, we can build many beautiful and functional structures that conform to the specifications of the original components and the designs of the architects. But if we start with bricks, we can build anything. By atomizing research data to its most minimal meaningful parts, it can be reconstructed in many different ways and for unlimited purposes. The designs are those of the scholar.

The semi-structured, graph data model implemented in OCHRE offers the scholar freedom – freedom not to be locked into a schema designed by someone else or for some other purpose; freedom not to have to decide in advance which schema among many options should be adopted, and then to be trapped in it; freedom not to have to re-tool or transform when standards change or when new best-practices are recommended. OCHRE is, in effect, ontology agnostic. If we need to tag a new concept or feature, we can simply add a new property to the rigorously structured, user-defined taxonomy of our OCHRE project. There is no waiting for an update to an official schema specification to be approved. This is not to say that standards should be ignored. But when using the OCHRE platform, decisions regarding ontological standards are a secondary process, not a primary one. For example, the researcher can map the project-specific taxonomy to the TEI specification and use OCHRE's export or publish function to transform the OCHRE textual items into TEI.[23] Similarly, a different mapping would be used to export OCHRE spatial items to conform to the CIDOC-CRM specification, and so on. A feature to export data as RDF/XML will transform those very same OCHRE items and post them to an RDF triple store. In these transformation processes, OCHRE items, including taxonomic properties, can be mapped onto any number of published ontologies. With tools such as these for exporting and publishing based on semantic mappings, the highly granular OCHRE model gives the user the freedom to play nicely with standards, without first having to store the data in compliance with any single pre-selected ontology.

[23] OCHRE makes it easy to share taxonomies among users or projects, and even to share partial branches of existing taxonomies. In fact, a surprising outcome of the ease of sharing was that, despite OCHRE's extreme flexibility, it had the effect of fostering collaboration rather than encouraging diversity.

The simplicity of the OCHRE data model is matched by the simplicity of a graphical user interface that masks the underlying data format. There are no raw XML files to be edited in oXygen. There is no need to manipulate comma-separated-value (CSV) files. Hundreds, thousands, and even millions of items can be neatly organized into hierarchies within OCHRE's built-in categories of data using a consistent set of tools. The mechanics of the technological tools should not unduly distract the scholar from managing research data.

## 5 Conclusion

In summary, OCHRE's data model, implemented as a semi-structured graph database, is highly flexible and customizable, yet organized according to a general ontological framework. OCHRE has been used for over a decade now for text corpus projects. Managing approximately 20,000 texts in various languages and writing systems, 200,000 images of texts, and tens of thousands of dictionary entries, the system has proven to be a powerful tool for philological study. As it enters its third year of use for text critical projects, OCHRE continues to evolve. It is the underlying graph data model that makes this recent innovation possible: highly atomized data, organized hierarchically into broad ontological classes, supplemented by cross-cutting linking, all the while supporting reuse wherever possible.

## Bibliography

Prosser, Miller, 'Digital Philology in the Ras Shamra Tablet Inventory Project: Text Curation through Computational Intelligence', in *CyberResearch on the Ancient Near East and Neighboring Regions: Case Studies on Archaeological Data, Objects, Texts, and Digital Archiving*, Digital Biblical Studies (Leiden: Brill, 2018), ii <https://doi.org/10.1163/9789004375086_012>

Prosser, Miller C., 'Digging for Data: A Practical Critique of Digital Archaeology', in "*An Excellent Fortress for His Armies, a Refuge for the People": Egyptological, Archaeological, and Biblical Studies in Honor of James K. Hoffmeier*, ed. by R. E. Averbeck and K. L. Younger (University Park: Penn State University Press, 2020), 309–23

———, 'Digital Philology in the Ras Shamra Tablet Inventory Project: Text Curation through Computational Intelligence', in *CyberResearch on the Ancient Near East and Neighboring Regions (Vol. 2): Evaluating New Tools and Methods for Archaeological Data, Objects, Texts, and Digital Archiving*, ed. by Vanessa Bigot Juloux, Tehri Nurmikko-Fuller, and Sveta Matskevich, Digital Biblical Studies (Leiden: Brill, 2018), 314–335.

Robinson, Iam, Jim Weber, and Emil Eifrem, *Graph Databases: New Opportunities for Connected Data*, 2nd edn (Beijing: O'Reilly, 2015)

Schloen, David, and Sandra Schloen, 'Beyond Gutenberg: Transcending the Document

Paradigm in Digital Humanities', *Digital Humanities Quarterly*, 8.4 (2014) <http://digitalhumanities.org:8081/dhq/vol/8/4/000196/000196.html>

Schloen, J. David, and Sandra Schloen, *OCHRE: An Online Cultural and Historical Research Environment* (Winona Lake, IN: Eisenbrauns, 2012)

TEI Consortium, *P5: Guidelines for Electronic Text Encoding and Interchange*, Version 3.6.0 (TEI Consortium, 2019) <https://tei-c.org/guidelines/>

Watson, Wilfred G. E., and Nicolas Wyatt, *Handbook of Ugaritic Studies*, Handbuch Der Orientalistik, 39 (Leiden: Brill, 1999)

Yon, Marguerite, Maurice Sznycer, and Pierre Bordreuil, eds., *Le Pays d'Ougarit Autour de 1200 Av. J.-C.: Histoire et Archéologie : Actes Du Colloque International, Paris, 28 Juin-1er Juillet 1993*, Ras Shamra-Ougarit, 11 (Paris: Éditions Recherche sur les Civilisations, 1995)

# Formal Models

# Projects and Editions

# Appendices

# Biographical Notes

**Thomas Ahrend** (University of Basel, Switzerland – thomas.ahrend@unibas.ch) studied Musicology, Philosophy and Literary Studies in Frankfurt a. M. and Berlin. He received his MA 1996, and his PhD 2005 at Technische Universität Berlin with a dissertation on the instrumental music of Hanns Eisler. 1997–2010 member of the editorial staff of the Hanns Eisler Gesamtausgabe in Berlin. Since September 2010, member of the editorial staff of the Anton Webern Gesamtausgabe at Musikwissenschaftliches Seminar at University of Basel.

**Peter Boot** (Huygens ING, The Netherlands – peter.boot@huygens.knaw.nl) studied mathematics and Dutch language and literature; he wrote his PhD thesis about annotation in scholarly digital editions and its implications for humanities scholarship. He oversaw the creation of the digital edition of the letters of Vincent van Gogh. He is employed as a senior researcher at the Huygens Institute for the History of the Netherlands where he works, among other things, as a consultant in several edition projects.

**Manuel Burghardt** (University of Leipzig, Germany – burghardt@informatik.uni-leipzig.de) is head of the Computational Humanities Group at Leipzig University. He is interested in the use of digital tools and computational techniques to explore new modes of doing research in the humanities. His most recent areas of research are Sentiment Analysis in the Humanities, Drametrics, Computational Intertextuality, Computational Analysis of Movies and Series and Music Information Retrieval.

**Toby Burrows** (University of Oxford, United Kingdom – toby.burrows@oerc.ox.ac.uk) is a Senior Researcher in the Oxford e-Research Centre at the University of Oxford, and a Senior Honorary Research Fellow in the School of Humanities at the University of Western Australia.

**Hugh Cayless** (Duke University, USA - hugh.cayless@duke.edu) is Senior Digital Humanities Developer at the Duke Collaboratory for Classics Computing. Hugh has over a decade of software engineering expertise in both academic and industrial settings. He also holds a Ph.D. in Classics and a Master's in Information Science. He is one of the founders of the EpiDoc collaborative and currently serves on the Technical Council of the Text Encoding Initiative.

**Hans Cools** (University of Basel, Switzerland – 1961-2021) had a master degree in medicine and a specialization in orthopaedic surgery and traumatology (Universities of Ghent and Antwerp, Belgium, 1997), a bachelor's degree in physical therapy, and a standalone degree in informatics (1999). Through various research and project management positions, in both companies and academic institutions, he gained expertise in different aspects of the Semantic Web technologies, focusing particularly on formal data modeling and machine reasoning. Those positions were in internationally collaborative research projects in a biomedical setting, mainly of the 5-7th EU Framework Program. Foremost in these projects were semantic interoperability and reusability of data. Since 2016, he worked in the humanities, as knowledge engineer, ontologist, and Semantic Web technology expert, at the University of Basel, as part of the NIE-INE project, which highlights scholarly editing. He (co-)published several articles, and gave workshops on the implementation of Semantic Web technologies in biomedicine and the humanities. He passed away in April 2021.

**Francesca Giovannetti** (University of Bologna, Italy – francesc.giovannett6@unibo.it) is a second-year PhD student in Digital Humanities at the Department of Classical Philology and Italian Studies, University of Bologna. She received an MA in Digital Humanities from King's College London and a second cycle degree in Digital Humanities and Digital Knowledge from the University of Bologna. She is interested in combining digital scholarly editing with semantic web technologies and in the use of digital technologies in education.

**Matthew Holford** (University of Oxford, United Kingdom – matthew.holford@bodleian.ox.ac.uk) is Tolkien Curator of Medieval Manuscripts at the Bodleian Library, University of Oxford.

**Marijn Koolen** (Royal Netherlands Academy of Arts and Sciences - Humanities Cluster, The Netherlands – marijn.koolen@gmail.coml) studied artificial intelligence and wrote his PhD thesis on using hyperlinks in information retrieval algorithms. He has worked on scholarly annotation for digital humanities research and on annotation-related information behaviour and information systems. He works as a researcher and developer at the Humanities Cluster of the Royal Netherlands Academy of Arts and Sciences, where he leads a project on developing annotation support within the *CLARIAH research infrastructure* project.

**David Lewis** (University of Oxford, United Kingdom – david.lewis@oerc.ox.ac.uk) is a Research Associate in the Oxford e-Research Centre at the University of Oxford.

**Andrew Morrison** (University of Oxford, United Kingdom – andrew.morrison@bodleian.ox.ac.uk) is a Software Engineer in the Bodleian Digital Library Systems and Services, Bodleian Library, University of Oxford.

**Stefan Münnich** (University of Basel, Switzerland – stefan.muennich@unibas.ch) studied musicology and communication science at the Technische Universität Berlin, MA 2011 with a thesis on cantional setting in Heinrich Schütz's Becker-Psalter. 2012 research assistant, 2013–2015 research associate of the Felix Mendelssohn Bartholdy. Sämtliche Briefe edition at University of Leipzig (co-editor of vols. 9 & 12). Since October 2015 research associate of the Anton Webern Gesamtausgabe, Basel; received his Doctorate degree in 2020 at the department of musicology at the University of Basel with a dissertation about music notation and its codes.

**Iian Neill** (Digital Academy of the Academy of Sciences and Literature, University of Mainz - Iian.Neill@adwmainz.de) is a visiting researcher at the Digital Academy of the Academy of Sciences and Literature Department at the University of Mainz, Germany. He is the creator of Codex, a text annotation environment which uses standoff property annotation to generate entities in a graph meta-model. Codex is currently being used to produce a digital edition of the epistles of Hildegard von Bingen at the Digital Academy in Mainz.

**Roberta Padlina** (University of Basel, Switzerland – roberta.padlina@unibas.ch) studied medieval philosophy at the University of Fribourg, Switzerland, obtaining a doctoral degree in June 2020. She has twelve years of professional experience in the field of Digital Humanities, thanks to which she has been able to work closely with different actors involved in the online publication of open access research. Roberta has worked for several years for e-codices –Virtual Library of Manuscripts in Switzerland and currently coordinates the National Infrastructure for Editions (NIE-INE) project. Roberta's main focus is on the opportunities and challenges that the digital shift poses for traditional education and research institutions, including developing semantic web strategies for scholarly publications and cultural goods.

**Kevin Page** (University of Oxford, United Kingdom – kevin.page@oerc.ox.ac.uk) is a Senior Researcher in the Oxford e-Research Centre and Associate Member of Faculty in the Department of Engineering in the University of Oxford.

**Miller C. Prosser** (University of Chicago, USA – m-prosser@uchicago.edu) earned his Ph.D. in Northwest Semitic Philology from the University of Chicago. His academic interests include the social and economic structure of Late Bronze Age Ras Shamra-Ugarit and the use of computational methods for philological and archaeological research. Miller is the Associate Director of the Digital Studies MA program at the University of Chicago where he teaches courses on Data Management and Data Publication for the Humanities. He also works as a researcher at the OCHRE Data Service of the Oriental Institute of the University of Chicago where he consults with and supports research projects using the Online Cultural and Historical Research Environment (OCHRE). He has also worked as a tablet photographer for the Mission de Ras Shamra (Ugarit) and the Persepolis Fortification Archive Project, employing advanced digital photographic methods such as reflectance transformation imaging, photogrammetry, and high-resolution digital scanning.

**Matteo Romanello** (Université de Lausanne, Switzerland - matteo.romanello@unil.ch) is Ambizione SNF Lecturer at the University of Lausanne, where he conducts a project on the commentary tradition of Sophocles' Ajax. Matteo is a Classicist and a Digital Humanities specialist with expertise in various areas of the Humanities, including archaeology and history. After obtaining his PhD from King's College London, he worked as a research scientist at EPFL's DHLAB on the Linked Books and Impresso projects, before moving to his current position. He was also teaching fellow at the University of Rostock, researcher at the German Archaeological Institute, and visiting research scholar at Tufts University.

**Sandra Schloen** (University of Chicago, USA – sschloen@uchicago.edu) is the Manager of the OCHRE Data Service at the Oriental Institute of the University of Chicago, and is the co-designer and developer of the Online Cultural and Historical Research Environment (OCHRE). Trained in computer science and mathematics (B.Sc. University of Toronto; M.Ed. Harvard University), Sandra has spent over 30 years working with technology as a systems analyst, technical trainer, and software developer. A long association with colleagues in the academic community has enabled her to develop a specialty in solving problems in the Digital Humanities where challenges of data capture, data representation and data management abound. Specifically, she has served extensively as a database manager for several archaeological projects in Israel and Turkey, and supports a wide range of research projects at the Oriental Institute and at other universities.

**Desmond Schmidt** (University of Bologna - desmond.allan.schmidt@gmail.com) has a background in classical Greek philology, information security and eResearch. He has worked on several scholarly edition projects, including the Vienna Wittgenstein Edition (1990–2001), Digital Variants (2004–2008), the Australian Electronic Scholarly Editions project (2012–2013), the Charles Harpur Critical Archive (2014-) and a pilot edition of Gianfrano Leopardi's Idilli (2018-). He currently works on developing practical web-based tools for making, visualising and publishing digital scholarly editions.

**Colin Sippl** (University of Regensburg, Germany – colin.sippl@ur.de) is currently a project employee at the University Library of Regensburg. Since 2017, he has been working on extending the open access services of the Electronic Journals Library (EZB). More recently, he has started developing and setting up a digital repository for literature, artefacts and experiments relating to the early life sciences based on the Invenio framework. He specialised in textual data mining and the development of media services in the institutional domain.

**Elena Spadini** (University of Lausanne - elena.spadini@unil.ch) is a postdoctoral researcher at the University of Lausanne. She holds a Ph.D. in Romance Philology from the University of Rome Sapienza (2016) and a M.A. in Digital Humanities from the École nationale des chartes (2014). She was a Marie Curie fellow in the IT Network DiXiT and co-directed the related volume Advances in Digital Scholarly Editing (Sidestone Press, 2017). She published in international journals and taught specialized courses in various European countries in the field of Digital Philology.

**Francesca Tomasi** (University of Bologna - francesca.tomasi@unibo.it) is associate professor in Archival Science, Bibliography and Librarianship at the University of Bologna (Italy). Her research is mostly devoted to digital cultural heritage, with a special attention to documentary digital edition, and a focus on knowledge organization methods in archives and libraries. She is member of different scientific committees of both associations and journals. In particular, she is President of the Library of the School of Humanities in the University of Bologna (BDU - Biblioteca di Discipline Umanistiche), Director of the international second cycle degree in Digital Humanities and Digital Knowledge (DHDK), President of the Italian Association of Digital Humanities (AIUCD – Associazione per l'Informatica Umanistica e la Cultura Digitale), and co-head of the Digital Humanities Advanced Research Center (/DH.ARC). She wrote about 100 papers and 4 monographs related to DH topics. She is editor and scientific director of several digital scholarly environments.

**Athanasios Velios** (University of the Arts London, United Kingdom – a.velios@arts.ac.uk) is Reader in Documentation at the University of the Arts London.

**Georg Vogeler** (University of Graz - georg.vogeler@uni-graz.at) is professor for Digital Humanities at the University of Graz and scientific director of the Austrian Center for Digital Humanities and Cultural Heritage at the Austrian Academy of Sciences. He is a trained historian (Historical Auxiliary Sciences). He spent several years in Italy (Lecce, Venice). In 2011, he became member of faculty at the Centre for Information Modelling at Graz University, where he was nominated full professor for Digital Humanities in 2016 and head of department in 2019. His research interests lie in late medieval and early modern administrative records, diplomatics (digital and non digital), digital scholarly editing and the history of Frederic II of Hohenstaufen (1194–1250). He was and is part in several national and international research projects related to his research interests.

**Christian Wolff** (University of Regensburg, Germany – christian.wolff@ur.de) has been Professor of Media Informatics at the Institute for Information and Media, Language and Culture at the University of Regensburg since 2003. He holds a PhD in information science and is a habilitated computer scientist. His research interests include: human-computer interaction, multimedia and web-based information systems, (multimedia) software engineering and information retrieval (in particular information literacy and social media).