# Knowledge Organization for Digital Humanities

Proceedings of the 15th Conference on Knowledge Organization *WissOrg'17* of the German Chapter of the International Society for Knowledge Organization (ISKO)

30th November – 1st December 2017, Freie Universität Berlin

Edited by Christian Wartena, Michael Franke-Maier and Ernesto de Luca

# Contents

# Knowledge Organization for Digital Humanities

Michel Franke-Maier
Freie Universität Berlin
Universitätsbibliothek
14195 Berlin
franke@ub.fu-berlin.de

Christian Wartena
Hochschule Hannover
D 30539 Hannover
christian.wartena@hs-hannover.de

## ABSTRACT

"Wissensorganisation" is the name of a series of biennial conferences/workshops with a long tradition organized by the German chapter of the International Society of Knowledge Organization (ISKO). The 15th edition of the conference focused on knowledge organization in digital humanities. Structuring and interacting with large data collections has become a major issue in digital humanities. In this proceedings various aspects of knowledge organization in the digital humanities are discussed and the authors of the papers show how projects in digital humanities deal with knowledge organization.

## KEYWORDS

Digital Humanities, Knowledge Organization

**Reference:**
Michel Franke-Maier and Christian Wartena. 2018. Knowledge Organization for Digital Humanities. In *Proceedings of Wissensorganisation 2017: 15. Tagung der Deutschen Sektion der Internationalen Gesellschaft für Wissensorganisation (ISKO) (WissOrg'17).* German Chapter of the ISKO / Freie Universität Berlin , pp. 2-3. https://doi.org/10.17169/FUDOCS_document_000000028863

## 1 OBJECTIVES OF THE WORKSHOP

The field of knowledge organization has always be dealing with highly theoretical and philosophical questions concerning the design of cataloguing systems, thesauri and vocabularies that can be used to describe and classify knowledge. This touches questions about the nature of knowledge and our view on the world. In the paper *Dr. Ingetraut Dahlberg – Eine Pionierin der Wissensorganisation* (Ohly, 2018) impressively shows how Ingetraut Dahlberg dedicated a whole life on finding a perfect knowledge organization system. Regrettably, we have to admit, that the impact of such efforts is very limited. On the other hand a lot of scientist working in different disciplines need to organize their knowledge and use a wide variety of systems and tools for this purpose, that often are very pragmatical and show various degrees of understanding of the main principles of knowledge organization systems. Thus the series of WissOrg conferences want to foster the awareness of the importance of proper knowledge organization in various disciplines and communities.


*WissOrg'17, December 2017, Berlin, Germany*

https://doi.org/10.17169/FUDOCS_document_000000028863

The topic of this years conference was knowledge organization and digital humanities. One invited and five submitted contributions show, how digital humanities projects deal with knowledge organization.

Digital Humanities have a longstanding tradition of visualization of abstract results and developing innovative user interaction designs to enable user access to large collections. It is interesting to see, what role knowledge organization plays in this context. At the same time, people working on knowledge organization systems might learn a lot from the methods digital humanities have developed to access large systems.

## 2 CONTRIBUTIONS

Subsequent to the general meeting of the German Chapter of the International Society of Knowledge Organization (ISKO) on the evening of Tuesday, 30th November, H. Peter Ohly, President of ISKO in the years 2010-2014 and current 1st Vice-President, commemorated Ingetraut Dahlberg (* 20 February 1927, † 24 October 2017). In his paper *Dr. Ingetraut Dahlberg – Eine Pionierin der Wissensorganisation* Ohly summarizes some major milestones of her life and work on classification–especially on her own system, the *Information Coding Classification* (Ohly, 2018). The listing of her name as planned entry in the *Encyclopedia of Knowledge Organization*[1] near S. R. Ranganathan and Paul Otlet stresses out her prominence and expertise in the scientific field of knowledge organization. At this point one idea of Dahlberg should be singled out: 2009 she presented *ten desiderata for knowledge organization* at Wissensorganisation '09. There she claimed that ISKO experts should actively apply their know-how there where it is made public and thus is useful. This not only applies for ISKO experts, but as well for knowledge organization experts in general – as it was proved during the conference for the field of digital humanities, too.

After the keynote – held by Ina Blümel on *Digital Heritage goes Open Science* – Meyer-Sickendiek and Hussein (2018) introduced the audience to the challenges if academics work together with partners beyond the scientific community. In *The New German Copyright Law for Science and Education (UrhWissG): Consequences for DH-Projects Working with Non-Academic Partners* one of the main question they deal with is how to use data which is copyrighted, a key question for all dh-projects working with data sets which are actually created and not already in the public domain. This question is embedded in the presentation of the conception and implementation of a project about the analysis of free verse prosody: *Rhythmicalizer. A digital tool to identify free verse prosody.* The Rhythmicalizer brings the techniques to classify the prosodic

[1] http://www.isko.org/cyclo/

patterns of spoken word and match the results with known rhythmical concepts like i.e. parlando or cut-up-rhythms. It was wonderful to realize that the authors' main question could be solved during the discussion at WissOrg'17 with a reference to the new *Urheberrechts-Wissensgesellschafts-Gesetz* which will come into force in March 2018. So the paper discusses the legal situation before and after this new law, at least pointing to the solution within the German jurisdiction.

The next paper is rather from the field of social sciences as from the humanities. In *Coding Schemes als Wissensorganisationssysteme für Digital Humanities: Mit Political Event Coding über Dynamic Network Analysis zu Ontologie-basierten Dynamic Cognitive Maps* Frank (2018) uses the concept of cognitive maps in conjunction with various appropriate knowledge organization systems to optimize the representation of causal knowledge for classifying political events. By using coding schemes like *Conflict and Mediation Event Observations (CAMEO) Coding Scheme* or its advancement *Political Language Ontology for Verifiable Event Records (PLOVER)* in dynamic or fuzzy networks Frank shows the potentials for optimizing the representation of historical knowledge from perspectives of various actors in dynamic maps. More or less the paper ends with an appeal by citing John Unsworth, librarian and a former member the of *National Council on the Humanities* advisory board (2013-2016): There should be more traineeship in the mixed skills needed in digital humanities "as we become serious about making the known world computable."[2]

Based on *Charles Sanders Peirce's* theory about concepts of signs and iconicity Freyberg (2018) presents in *Density of Knowledge Organization Systems* a new theoretical approach which allows to measure the level of contextualization in knowledge organization systems (KOS). She introduces the term density which originates from network analysis and that she adapts for knowledge organization. She exemplifies this new theoretical concept for understanding and modeling of KOS by pointing to the iconic structures of Europeana and the Linked Open Data Cloud.

Faßhauer (2018) reports in her paper *Accessing and Editing Early Modern Manuscript Collections: Cooperation Possibilities between Holding Institutions and Digital Scholarship* on a typical dh-project in the face of some large obstacles to overcome: the recognition of handwritten text with an idiosyncratic use of special signs, a mix of abbreviated German and Latin words and a huge amount of text of about 40 000 pages. The critical edition of some selected Journals of *Johann Christian Senckenberg*, well-known for his involvement on societal healthcare in Frankfurt on the Main, within a digital framework is a prime example for the interaction between edition philology and librarianship. For editing the textual material the authority data of the *Gemeinsame Normdatei (GND)* was used and even could be improved. Even books without any bibliographic reference in German library catalogues could be identified – a valuable contribution to optimize knowledge organization systems.

In *Ein digitaler Zeichenkatalog als Organisationssystem für die noch nicht entzifferte Schrift der Klassischen Maya* Diehr et al. (2018) report on a project funded until 2029 to develop concepts and tools to handle the complex and in some parts not yet deciphered Classical Maya Hieroglyphics. The result will be a kind of digital dictionary and a complete inventory of Mayan scripture organized by an ontological-based model. Former assignments of meaning to hieroglyphics will be also part of this dictionary, since the model allows to describe former, alternative and even erroneous classifications. The general concept of this kind of modeling is applicable also to other scriptures – especially because of its open way to integrate a not yet consolidated state of knowledge into its conceptual structure.

As editors of these proceedings of *Knowledge Organization for Digital Humanities* we wish an insightful reading in the hope – as Andrea Tatai, the deputy Library Director of the University Library of the Freie Universität Berlin mentioned in her welcoming address – that the nature of Digital Humanities as new knowledge in a new form is inspiring for your own further research activities.

[2] John Unsworth: http://computerphilologie.uni-muenchen.de/jg02/unsworth.html

## REFERENCES


Franziska Diehr, Maximilian Brodhun, Sven Gronemeyer, Katja Diederichs, Christian Prager, and Elisabeth Wagnerand Nikolai Grube. 2018. Ein digitaler Zeichenkatalog als Organisationssystem für die noch nicht entzifferte Schrift der Klassischen Maya. In *Wissensorganisation 2017: 15. Tagung der Deutschen Sektion der Internationalen Gesellschaft für Wissensorganisation (ISKO).* ISKO, Berlin, Germany.

Vera Faßhauer. 2018. Accessing, Editing and Indexing Large Manuscript Collections. In *Wissensorganisation 2017: 15. Tagung der Deutschen Sektion der Internationalen Gesellschaft für Wissensorganisation (ISKO).* ISKO, Berlin, Germany.

Ingo Frank. 2018. Ontologie-basierte kognitive Karten. In *Wissensorganisation 2017: 15. Tagung der Deutschen Sektion der Internationalen Gesellschaft für Wissensorganisation (ISKO).* ISKO, Berlin, Germany.

Linda Freyberg. 2018. Density of Knowledge Organization Systems. In *Wissensorganisation 2017: 15. Tagung der Deutschen Sektion der Internationalen Gesellschaft für Wissensorganisation (ISKO).* ISKO, Berlin, Germany.

Burkhard Meyer-Sickendiek and Hussein Hussein. 2018. The Effect of the New Copyright Law on DH-Projects. In *Wissensorganisation 2017: 15. Tagung der Deutschen Sektion der Internationalen Gesellschaft für Wissensorganisation (ISKO).* ISKO, Berlin, Germany.

H. Peter Ohly. 2018. Dr. Ingetraut Dahlberg - Eine Pionierin der Wissensorganisation. In *Wissensorganisation 2017: 15. Tagung der Deutschen Sektion der Internationalen Gesellschaft für Wissensorganisation (ISKO).* ISKO, Berlin, Germany.

# Dr. Ingetraut Dahlberg - Eine Pionierin der Wissensorganisation

H. Peter Ohly
D-53175 Bonn
Peter.Ohly@gmx.de

## ABSTRACT

Dr. Ingetraut Dahlberg hat wesentlich den Begriff „Wissensorganisation/Knowledge Organization“ eingeführt und geprägt. Auch ist sie der Hauptmotor bei der Gründung der wissenschaftlichen Vereine *Gesellschaft für Klassifikation* und *International Society for Knowledge Organization* sowie der Zeitschriften *International Classification* und *Knowledge Organization* gewesen. 2017 ist Ingetraut Dahlberg im Alter von 90 Jahren verstorben, weshalb hier einige Lebensdaten und wissenschaftliche Beiträge zu ihrer Würdigung dargestellt werden.

## KEYWORDS

Digital Humanities, Knowledge Organization, Wissenorganisation, Klassifikation, Ingetraut Dahlberg

**Reference:**
H. Peter Ohly. 2018. Dr. Ingetraut Dahlberg - Eine Pionierin der Wissensorganisation. In *Proceedings of Wissensorganisation 2017: 15. Tagung der Deutschen Sektion der Internationalen Gesellschaft für Wissensorganisation (ISKO) (WissOrg'17).* German Chapter of the ISKO / Freie Universität Berlin , pp. 4-10. https://doi.org/10.17169/FUDOCS_document_000000028863

## 1 VITA

Dr. Ingetraut Dahlberg wurde am 20. Februar 1928 in Köln als Ingetraut Gessler geboren. Sie starb am 24. November 2017 in Bad König, Odenwald. Der Vater Theodor Gessler stammte aus Wesel, Niederrhein und hatte in Köln Wirtschaftswissenschaften studiert. Die Mutter Luzie, geb. Sauvageot, stammte aus Köln. Die Familie zog nach Frankfurt, wo Ingetraut Gessler mit ihrem Bruder und ihrer Schwester aufwuchs. 1955-1956 war sie mit Reinhard Dahlberg, dem späteren Protagonisten der Wasserstoffumwandlung, verheiratet. Ihr Sohn Wolfgang Dahlberg wurde 1955 geboren. Er war der Autor vieler Bücher, so „Ordnung, Sein und Bewußtsein“ 1984, und verstarb 2012.

Otto Sechser charakterisierte in einer E-Mail vom 30. Oktober 2017 Ingetraut Dahlberg folgend:

> „Dr. Inge Dahlberg was and is going to remain one of great personalities of Classification, Documentation, and Knowledge Organization. Her interests, achievements, and worldwide contacts will be the theme of dissertations. Here I want to write about Dr. Dahlberg as a good-hearted, modest, hard-working, high-principled woman, always ready to help, with enormous social intelligence [...] She will be missed in ISKO.”

### 1.1 Studium und Berufseinstieg

Ingetraut Gessler studierte Philosophie, Katholische Theologie und Anglistik und zeitweise Biologie in Frankfurt und Würzburg. 1948/49 verbrachte sie ein Studienjahr in den USA am Mary Manse College in Toledo, Ohio.

1959 kam Ingetraut Dahlberg ans Gmelin-Institut für Anorganische Chemie, Frankfurt, dessen Direktor Erich Pietsch zu dieser Zeit Präsident der Deutschen Gesellschaft für Dokumentation (DGD) war. Hier bearbeitete sie Bibliographien für die Atomkernenergie-Dokumentation (AED). 1961 wechselte sie zum Rationalisierungs-Kuratorium der Deutschen Wirtschaft (RKW).

### 1.2 Dokumentarische Ausrichtung

Ingetraut Dahlberg begann 1962 mit einer Ausbildung zur wissenschaftlichen Dokumentarin und arbeitete 1963 bei der Deutschen Gesellschaft für Dokumentation an der Erfassung der bibliothekarischen Bestände und einer Dokumentation der Literatur zum Thema Dokumentation inklusive Thesauruserstellung.

Später wurde sie Leiterin Bibliothek und Dokumentationsstelle der DGD. 1964 bis 1965 hatte sie einen Aufenthalt an dem Groth Institute for Crystallographic Data Documentation sowie der Universitätsbibliothek der Florida Atlantic University, Boca Raton. Hier arbeitete sie mit Jean Perreault (Begriffsrelationen; (Perreault, 1994)) zusammen.

### 1.3 Verbands- und Auftragsarbeit

Mit Martin Scheele (Vorsitz) begründete sie (Sekretariat) 1966 das DGD-Komitee *Thesaurusforschung und Klassifikation*. Später resultierte aus diesem Komitee die Veröffentlichung von Dagobert Soergel zu *Indexing languages and thesauri: Construction and maintenance*. Weiter entstand hier ein Deskriptorensystem für die Informationswissenschaften. Von 1967 bis 1969 war Ingetraut Dahlberg Vorsitzende des FID-Revisionskomitees für die Universelle Dezimalklassifikation UDC-03/04 (Common auxiliaries of materials und Common auxiliaries of relations, processes and operations). Hieraus entstanden eine Klassifikation der Dokumentenarten und ihrer speziellen Aspektbegriffe sowie ein Vorschlag zur Revision der UDC.

1967 bis 1974 leitete Ingetraut Dahlberg im DIN-Normenausschuss Terminologie die Revision von DIN 2330 *Begriffe und Benennungen : Allgemeine Grundsätze*, von DIN 2331 *Begriffssysteme und ihre Darstellung* und von DIN 32705 *Klassifikationssysteme : Erstellung und Weiterentwicklung von Klassifikationssystemen*. Bei der International Organization for Standardization arbeitete sie an ISO/TC 37 *Terminology and other language and content resources* und ISO/TC 46 *Information and documentation* mit.


*WissOrg'17, December 2017, Berlin, Germany*

https://doi.org/10.17169/FUDOCS_document_000000028863

1970 war sie Mitglied der Working Group on Indexing and Classification im Rahmen von UNISIST[1] und 1971 im Beratungsgremium *Datenbanksystem für die Bundesrepublik Deutschland* des Innenministeriums der Bundesregierung. 1972–1973 hatte Ingetraut Dahlberg einen DGD-Projektauftrag zur Sammlung der Benennungen von Wissensgebieten. 1972–1974 arbeitete sie im *Subject-Field Reference Code Gremium* vom FID–Committee on Classification Research (FID/CR) an der Erstellung des UNISIST-*Broad System of Ordering* (BSO) mit. 1978–1979 untersuchte sie in der Pilotstudie DB-Thesaurus die Thesaurusfähigkeit von Schlagwörtern. Ab 1979 war sie auch im Committee on Conceptual and Terminological Analysis (COCTA)[2] der *International Political Science Association* (IPSA) und der *International Sociological Association* (ISA) tätig. So führte sie 1981 eine COCTA-Konferenz in Bielefeld durch. 1981–1987 hatte sie den Vorsitz des FID/CR inne. 1983 gab sie die Expertendokumentation Who Is Who in Classification and Indexing heraus.

## 1.4 Promotion

1973 promovierte sie bei Alwin Diemer, Düsseldorf, im Fach Philosophie mit den Nebenfächern Allgemeine Linguistik und Geschichte der Naturwissenschaften. Ihre Dissertation *Das Universale Klassifikationssystem des Wissens, seine ontologischen, wissenschaftstheoretischen und informationstheoretischen Grundlagen* wurde 1974 als *Grundlagen universaler Wissensordnung* im Verlag Dokumentation veröffentlicht (Dahlberg, 1974a). Hierin untersuchte sie verschiedene universelle Klassifikationskonzepte (u.a. DDC, UDC, LCC, Colon Classification) und Probleme von Universal-Klassifikationssystemen und machte Vorschläge für ein neues Universal-Klassifikationssystem.

## 1.5 Lehre und Forschung

1976–1979 führte sie an der Universität Mainz das DFG-Projekt Logstruktur durch, worin inhaltliche, von der Bezeichnung unabhängige Beziehungen zwischen Wissensgebieten gesucht wurden. 1977 entwickelte sie aus den Erkenntnissen ihrer Dissertation die *Information Coding Classification* (ICC), ein facettiertes universales Klassifikationssystem der Wissensgebiete mit ca. 6.500 Begriffen. Es folgten hierzu ausführende Seminarveranstaltungen am *Documentation Research and Training Centre* (DRTC) in Bangalore. Lehrstuhlvertretungen hatte sie 1984/85 an der Universität Saarbrücken, 1985-1987 an der FH Hannover und 1988/89 an der FH Darmstadt.

## 1.6 Verlegerische Tätigkeit, Vereinsgründungen

Mit Alwin Diemer, Jean. M. Perreault, Arashanipalai Neelameghan und Eugen Wuester gründete sie 1974 die Zeitschrift International Classification (IC), die 1993 in Knowledge Organization (KO) umbenannt wurde. Im Heft IC-1974-1 erschien ihr Beitrag *Zur Theorie des Begriffs* (Towards a theory of the concept) (Dahlberg, 1974b), worin auf Begriffsbildung, Begriffssysteme und Begriffsarten eingegangen wird. 1979 gründete sie mit ihrem Sohn Wolfgang das Unternehmen INDEKS für die Erstellung von Registern und Klassifikationssystemen, das schließlich zum INDEKS Verlag wurde.

Ingetraut Dahlberg gründete 1977 mit Robert Fugmann, Martin Scheele, Hanns-Hermann Bock u.a. die deutsche Gesellschaft für Klassifikation (zunächst als GFK, ab 1979 als GfKl abgekürzt). Sie hatte 1977 ihre erste Tagung in Münster. Als dort in der GfKl die Numerische Taxonomie und Datenanalyse überhandnahmen, gründete sie 1989 mit Robert Fugmann, Padmini Raj und Rudolf Ungvary u.a. die International Society for Knowledge Organization (ISKO) mit mehr begrifflicher Ausrichtung. Diese hatte ihre erste Tagung 1990 in Darmstadt. Ihr Publikationsorgan wurde die Zeitschrift International Classification , bzw. deren Nachfolgezeitschrift Knowledge Organization .

Um 1997 erfolgten größere Veränderungen, nicht zuletzt wegen einer Krebsdiagnose. Der INDEKS-Verlag ging an den Ergon-Verlag über. Sie gab die Chefredaktion der Zeitschrift Knowledge Organization ab[3]. Neue Präsidentin der ISKO 1996-1998 wurde Hanne Albrechtsen, Royal School of Librarianship, Kopenhagen.

Ihre umfangreiche Bibliothek zur Klassifikation, Terminologie und den Informationswissenschaften ging an das Maastricht McLuhan Institute (MMI), European Centre for Digital Culture, Knowledge Organisation and Learning Technology , musste später aber wieder zurückgenommen werden und ist nun in Händen des Georg-Eckert-Institut - Leibniz-Institut für internationale Schulbuchforschung, Braunschweig.

## 1.7 Veröffentlichungen und Vorträge

Ingetraut Dahlberg hat eine Vielzahl von Veröffentlichungen herausgebracht, so ergaben sich Ende November 2017 in Google 13.300 Treffer mit ihrem Namen. Im Beitrag in Wikipedia über sie werden 300 Veröffentlichungen genannt und eine Personen-Dokumentation der Association for Information Science and Technology (ASIS&T) von 2014 listet 337 Literaturangaben. Neben den Interviews mit ihr, 2008 in Knowledge Organization 35 (McIlwaine and Mitchell, 2008), 2012 für ASIS&T (Romen, 2012), 2015 mit Claudio Gnoli (Gnoli, 2015) sind für die letzten Jahre folgende Vorträge, bzw. Veröffentlichungen hervorzuheben:

- Deutsche ISKO-Tagung 2009: „Desiderate für die Wissensorganisation“ (Dahlberg, 2013) / „How to improve ISKO's standing: ten desiderata for knowledge organization“ (Dahlberg, 2011)
- Deutsche ISKO-Tagung 2013: „Was ist Wissensorganisation?“ (Dahlberg, 2017) / „Brief communication: What is knowledge organization“ (Dahlberg, 2014a)
- 2014: „Wissensorganisation: Entwicklung, Aufgabe, Anwendung, Zukunft“ 2014 (Dahlberg, 2014b)
- 88. Ernst-Schröder-Seminar, Darmstadt 2015, und Dagstuhl-Workshop Buchprojekt „Corporate Semantic Web“, Wadern 2016: „Warum Universalklassifikation?“
- 2016: „Dokumentenkunde - Dokumentologie: damals - und heute?“ (Dahlberg, 2016)

## 1.8 Ehrungen

Zusammen mit S. D. Boon, Eindhoven bekam Ingetraut Dahlberg 1965 den Preis der International Association of Documentalists and

[1] Unesco Intergovernmental Programme for Co-operation in the Field of Scientific and Technical Information = World Science Information System (Vorsitz: Douglas John Foskett)

[2] Gegründet von Fred W. Riggs und Giovanni Sartori

[3] Nachfolger 1997-1998: Charles Gilreath, The Texas A&M University System

**Abbildung 1: Dr. Ingetraut Dahlberg, 19.05.2014 in Krakau auf der 13. internationalen ISKO Konferenz (Foto von Peter Ohly)**

Information Officers, Paris (A.I.D.). 1996 erhielt sie den International Ranganathan Award für grundlegende Arbeiten in der Klassifikationsforschung. Bei der Deutschen ISKO und der internationalen ISKO wurde sie um 2000 Ehrenmitglied. 2006 erhielt sie den Eugen Wüster Sonderpreis des Internationalen Informationszentrums für Terminologie, Wien (Infoterm).

## 2 DAS FACH WISSENSORGANISATION / KNOWLEDGE ORGANIZATION

In Anlehnung an Bliss' "organization of knowledge" kreierten Dahlberg und die weiteren Mitbegründer der ISKO für ein Gebiet, das früher Ordnungslehre oder Klassifikation benannt wurde, den Begriff Wissensorganisation, der von der Wissenschaftsgemeinschaft schnell übernommen wurde. Nach Dahlberg ist Wissensorganisation die Wissenschaft von der Strukturierung und systematischen Anordnung von Wissenseinheiten (Begriffen) nach den ihnen inhärenten Wissenselementen (Merkmalen) und der Anwendung der so geordneten Begriffe und Klassen von Begriffen zur Beschreibung von wissenswerten Inhalten von Gegenständen jeglicher Art (vgl. Dahlberg (2006a)). Speziell versteht Dahlberg (1998) Wissensorganisation als: Sachgebiet, das sich mit der Ordnung von

a Wissenseinheiten (Begriffen) und
b Objekten aller Art (Mineralien, Pflanzen, Tiere, Dokumente, Bilder, Museumsobjekte, etc.) befasst, die auf entsprechende Begriffe oder Begriffsklassen bezogen werden, um das Wissen über die Welt des Gewussten geordnet festhalten und zur Nutzung weitergeben zu können.

Wissensorganisation umfasst nach ihr die folgenden neun Teilgebiete (1998; vgl. 2017, bzw. die engl. Version 2014a):

(1) die epistemologischen, mathematischen, systemtheoretischen, kognitionswissenschaftlichen und wissenschaftstheoretischen Voraussetzungen der Ordnung von Begriffen wie auch ihr geschichtlicher Hintergrund,
(2) die Kenntnisse der Elemente und Strukturen von Begriffssystemen,
(3) die Methodik der intellektuellen Erstellung, Pflege und Revision dieser Systeme und ihrer Computerisierung einschließlich der Fragen der paradigmatischen und syntagmatischen Relationierung ihrer Elemente und Einheiten sowie die Kompatibilisierung und Evaluierung dieser Systeme,
(4) die Methodik der intellektuellen und maschinellen Anwendung dieser Systeme durch Klassierung und Indexierung,
(5) die Kenntnis der existierenden universalen und
(6) der speziellen Taxonomien und Klassifikations-Systeme wie auch der Dokumentationssprachen (Thesauri),
(7) die Fragen, die sich von Seiten der Einfluss nehmenden Gebiete, Linguistik (~ Linguistik mathematische) und Terminologie, ergeben, einschließlich der Probleme des Retrieval, besonders auch in Online-Zugriff,
(8) die Anwendungen der inhaltlichen Erschließung aller Arten von Dokumenten und in allen Sachgebieten,
(9) das gesamte Umfeld der Organisation von Wissen am Arbeitsplatz, in einzelnen Zentren, Gesellschaften, Ländern und im internationalen Bereich, wie auch die Fragen der Ausbildung, der Ökonomie, der Benutzer, etc.

Später merkte Dahlberg in der Diskussionsliste Wiss-Org an (Dahlberg, 2006b):

> "The concept of ‚organization' however, in its acceptation in German has a wider range than just ‚order', namely ‚planned construction', ‚structure', ‚forming' (Wahrig, 1975), although this does not apply to some other languages where ‚organization' is only used for collectivities like associations or unions, so that in such cases, ‚organization' can only be related to people, not to objects."

Hierzu ist anzumerken, dass fälschlicherWeise hierunter auch 'Wissen in Organisationen' verstanden wurde, was eher dem Gebiet Knowledge Management entspricht und weniger von Dahlberg intendiert wurde.

## 3 DIE ANALYTISCHE, GEGENSTANDSBEZOGENE BEGRIFFSTHEORIE

Mit Bezug auf Frege (Frege, 1969) (Dahlberg, 2014b, S. 36) hat Dahlberg die Gewinnung von Wissenseinheiten auf eine Begriffstheorie (engl.: Concepts) zurückgeführt (Dahlberg, 1974b, 1979, 1987, 2009).

Begriffe sind nach ihr die wesentlichen Elemente jeglicher Wissensordnung aus denen auch die Klassen gebildet werden.

> „Ein Begriff ist eine Wissenseinheit, die dadurch entsteht, dass wesentliche und überprüfbare Aussagen über einen Bezugsgegenstand gemacht werden, die die Begriffsmerkmale erbringen und die diese zum Zwecke der Kommunizierbarkeit in einer Bezeichnung (Benennung oder Code) in kurzer und prägnanter Form zusammenfasst." (Dahlberg, 2014b, S.37ff).

Da in analysierender Form durch einzelne Prädikationen über einen Bezugsgegenstand die Merkmalsmenge eines Begriffs gewonnen wird, spricht sie deshalb auch von einer „gegenstandsbezogenen, analytischen Begriffstheorie". Der sprachwissenschaftliche Aspekt verhindert nach ihrer Meinung dagegen den analytischen Aspekt bei Begriffsbildung und Begriffserkenntnis (z.B. in Dahlberg (2017, S. 13)).

Bezeichnungen (Benennung oder Code) sollten die gewonnenen wesentlichen Merkmale berücksichtigen, um für den Umgang mit dem Begriff Gedächtnishilfe und Anschaulichkeit zu vermitteln, die aber auch möglichst kurz und einprägsam sein sollten, um im sprachlichen Prozess akzeptiert werden zu können. Der Bezug zwischen gemeintem Gegenstand, den Merkmalen und den Bezeichnungen kann als Begriffsdreieck gesehen werden. Die Feststellung der notwendigen Merkmale, die von ihr als Wissenselemente bezeichnet werden, bilden zusammen die Wissenseinheit, was ein begriffskonstituierender Vorgang ist mit der möglichen Folge, dass Begriffe mit gleichen oder ähnlichen Merkmalen in einer Klasse zueinander geführt werden (Begriffskonstruktion). Ein auf diesen Prinzipien erstelltes Klassifikationssystem ist dann ein Definitionssystem, das sich aus sich selbst erklärt.

Die Begriffsrekonstruktion geht von der Sprachverwendung aus, woraus dann abgelesen werden kann, was der Bezugsgegenstand der Benennung gewesen sein mag und welches jeweils die Merkmale dieses Bezugsgegenstandes sind, was ggf. polyseme Benennungen aufdeckt. Beziehungen zwischen Begriffen sind aufgrund der Merkmale der Begriffe feststellbar und darstellbar. Dahlberg unterscheidet formal logische (z.B. Inklusion), formkategoriale (z.B. Facetten) und materiale, inhaltliche Hauptarten von Begriffsbeziehungen. Letztere werden durch die Abstraktionsrelation, Partitionsrelation, Ganzes-Teil-Relation, welche hierarchiebildend wirken, gebildet. Die funktionsbezogene, grammatische oder Syntax-Relation taucht bei der Untergliederung eines Sachgebietes auf. Die Komplementärrelation findet Anwendung in der Gegenüberstellung von Objekten und/oder ihren Eigenschaften (hier verweist sie auf Diemer (1969)).

## 4 KLÄRUNG DES BEGRIFFSFELDES „KLASSIFIKATION"

Da die Bezeichnung 'Klassifikation' umgangssprachlich und fachsprachlich polysem vorkommt, klärt Dahlberg die verschiedenen Bedeutungen durch prägnantere Bezeichnungen Dahlberg (2014b, S. 62-63):

**Klassifikation:** ein Klassifikationssystem

**Klassifizieren:** das Bilden von Klassen von Begriffen nach gemeinsamen Merkmalen (oft in der Umgangssprache mit 'Klassieren' verwechselt)

**Klassifikat:** das Gesamt von Klassen als Resultat des Klassifizierens

**Klassieren:** das Zuordnen von Klassen zu Gegenständen

**Klassat:** das Produkt aus der Zuordnung von Klassen zu Gegenständen und Themen

**Klassifikationswissenschaft (oder Wissensorganisation):** die Lehre über die Klassifikationssysteme, ihre Theorie und Geschichte, ihre Praxis in der Bildung von Klassifikationssystemen, also der Bildung von Klassen aber auch der Anwendung in der Zuordnung von Klassen zu Gegenständen.

**Klasse:** diejenige Menge von (Klassen)-Elementen, die durch ein gemeinsames („klassifikatorisches") Merkmal, auch „Klassem" genannt, zusammengefasst wird.

**Klassem:** das gemeinsame Merkmal, das Klassenelemente in einer Klasse zusammenführt.

**Klassen-Elemente:** Begriffe für Objekte oder Abstrakta, nicht die Ideen, Objekte, Themen an sich.

## 5 DIE INFORMATION CODING CLASSIFICATION (ICC)

1982 beschrieb Dahlberg die von ihr u.a. aus dem DGD-Projekt „Ordnungssystem der Wissensgebiete" und dem DFG-Projekt „Logstruktur" heraus entwickelte Information Coding Classification (ICC) in Knowledge Organization mit dreistufiger Hierarchie (Dahlberg, 1982).

Diese Klassifikation soll die Vorzüge der Universalität, der Facettierung und des Top-down-Ansatzes haben. Sie wurde später auch auf die *International Classification and Indexing Bibliography* (ICIB) und die Bibliografie in der Zeitschrift KO angewendet. Sie geht in seinen Hauptklassen nicht von Disziplinen aus, sondern von neun ontischen Entwicklungsstufen, den Seinsschichten. Sie gliedert diese grob und in den weiteren Hierarchiestufen auch fein nach jeweils neun Kategorien, was auch eine Codierung durch Dezimalzahlen ermöglicht. Es wurden die Lokationen für die Wissensgebiete durch einen Elementstellenplan so festgelegt, dass die erste Hierarchiestufe nach neun Seinsschichten (Objektbereiche als Sachkategorien) und die zweite Hierarchiestufe nach neun funktional ausgerichteten Formkategorien strukturiert sind.

Die möglicherweise 3. und 4. von untergeordneten Wissensgebieten wie auch die 5. und 6. Stufe werden nach den gleichen Sach- und Formkategorien angeordnet. Dadurch wird es möglich, mit den Ziffern der numerischen Codierung eines bestimmten Wissensgebietes immer auf die gleichen Kategorien zuzugreifen, was die Mnemotechnik des Systems verstärkt und auch die Lokalisierung von Inter- und Transdisziplinarität berücksichtigt.

Die Verantwortlichkeit für die ICC wurde 2015 von Dahlberg an das deutsche Chapter der ISKO übertragen. Basierend auf Ergebnissen der Dissertation von 1972 wendete sie hierbei 12 Prinzipien bezüglich der theoretischen Grundlagen und dem Gerüst und Anordnung der gefundenen Wissensgebiete an (Dahlberg, 2014b).

## 6 DIE PRINZIPIEN DER ICC

**Prinzip #1:** Die Inhalte der ICC sind Begriffe und Klassen von Begriffen.

**Prinzip #2:** Als systematisierende Begriffsbeziehungen werden verstanden:

(1) Abstraktionsbeziehung (Genus-Species Beziehung
(2) Partitionsbeziehung (Bestandsbeziehung)
(3) Komplementärbeziehung und
(4) Funktionalbeziehung.

**Prinzip #3:** Die ICC verwendet das Dezimalprinzip in der Anordnung ihrer Hauptklassen und der Aspekte, unter denen diese Hauptklassen eingeteilt werden können, d.h., sie geht von 9 Objektbereichen aus, die durch 9 Aspektbereiche gegliedert werden. Jede einzelne, der somit entstandenen 81 Sachgruppen wird daraufhin wiederum nach den genannten Aspekten in 9 Sachgebiete unterteilt.

**Prinzip #4:** Die Hauptklassen sind Objektbereiche des Seins, also ontische Einheiten, die zu je drei Bereichen gebündelt werden können:
(1) Formen und Strukturen
(2) Energie und Materie unbelebtes Sein
(3) Kosmos und Erde
(4) Bio-Bereich
(5) Mensch-Bereich belebtes Sein
(6) Gesellschaftsbereich
(7) Wirtschaft und Technik
(8) Wissenschaft und Informationsprodukte menschlicher Tätigkeit
(9) Kulturbereich

**Prinzip #5:** Die Untergliederung der Objektbereiche und der Sachgruppen erfolgt nach einem Systemstellenplan, 'Systematifikator' genannt, wobei die ersten drei Aspekte die jeweils konstituierenden einer Sachgruppe oder eines Sachgebietes sind, die zweiten drei die sogenannten Ausprägungen enthalten und die dritten drei die Beziehungen zu Inhalten von „außerhalb" einer Sachgruppe oder eines Sachgebietes:
(1) Allgemeines, Theorien, Prinzipien (Axiomen- und Strukturbezug)
(2) Objektbereich: Gegenstände, Arten, Teile, Eigenschaften (Objektbezug)
(3) Tätigkeitsbereich, Methoden, Prozesse, Aktivitäten, (Aktivitätsbezug)
(4) Eine besondere Eigenschaft oder auch Ausprägung einer Sachgruppe
(5) Personenbezug oder auch Ausprägung einer Sachgruppe
(6) Gesellschaftsbezug oder auch Ausprägung einer Sachgruppe
(7) Einwirkungen von Außen auf Gebiet (instrumenteller Bezug)
(8) Anwendungen der Methoden in anderen Sachgruppen (Ressourcenbezug)
(9) Inf. über das Gebiet und synthetisierende gesellschaftliche Aufgaben (Aktualisierungsbezug)

Diese neun Aspekte werden entsprechend auch bei der Untergliederung der Sachgruppen in Sachgebiete verwendet. Die Anwendung dieses Prinzips in den Sachgruppen und Sachgebieten bewirkt, dass man bei der Suche nach bestimmten Aspekten immer auf die gleichen Zahlen der Notation zugreifen kann.

**Prinzip #6:** Bei der Anordnung der ontischen Objektbereiche von 1-9 handelt es sich um ein Schichtenmodell, das den Vorstellungen von Schichten der Wirklichkeit („Integrationsstufen") entspricht, die einander bedingen, wie dies von Feibleman (1954) und von Hartmann (1964) erläutert und sogar durch Gesetze erklärt wurde. Somit bildet z.B. die Schicht 1 die Voraussetzung für die Schicht 2 und so fort. Auch wurde darauf gesehen, dass sozusagen jede Sachgruppe eine Voraussetzung für die folgende ist.

**Prinzip #7:** Mit dem unter Prinzip #5 genannten Systemstellenplan (bzw. Systematifikator) ist gleichzeitig auch die Möglichkeit von Beziehungen zwischen den Sachgruppen und Sachgebieten verbunden, an folgenden Stellen: 1 Allgemeines und 8 Anwendungen, 9 Wissensvermittlung.

**Prinzip #8:** In der Rasterdarstellung der ICC befindet sich über der Schicht 1 noch die Schicht 0. Ihre mögliche Facettenklassifikation wartet noch auf ihre Realisierung. Ansonsten enthält die Schicht von 01 bis 09 die Bezeichnungen für die Aspekte. (vgl. Dahlberg, 1978)

**Prinzipien #9 und #10:** Sie betreffen die mögliche und benötigte Kombination der Sachgruppen und Sachgebietsbegriffe mit Begriffen von Raum und von Zeit.

**Prinzip #11:** Mnemotechnik im System: Indem Prinzipien gefunden wurden, die das System auf Anhieb überschaubar machen und ein Regelsystem mit mnemotechnischen Eigenschaften besitzt, also die Fähigkeit der leichten Merkbarkeit der Inhalte von Systemstellen.

**Prinzip #12:** Zusammenfassend kann darauf hingewiesen werden, dass die unter den Prinzipien 7, 8 und 9 genannten Kombinationsmöglichkeiten das System zu einem sich selbst vernetzenden macht, mit dem unendlich viele Kombinationen möglich sind.

## 7 ZUR 'BEGRIFFSKULTUR' IN DEN SOZIALWISSENSCHAFTEN

1996 veröffentlichte die Zeitschrift *Ethik und Sozialwissenschaften - Streitforum für Erwägungskultur* in Heft 1 als Hauptartikel von Dahlberg „Zur 'Begriffskultur' in den Sozialwissenschaften: Lassen sich ihre Probleme lösen?" und ließ (sehr lesenswerte) Kritiken hierzu aus dem angesprochenen Fach anfertigen. In der Summe stieß dieser Hauptartikel auf viel Resonanz aber auch Widerspruch, da gerade die Sozialwissenschaften eine sehr vage Begriffsbildung aufweisen, was sich z.T. schon in den Betitelungen der Kritiken widerspiegelt[4]:

- 'Begriffsbildung' aus psychologischer und sprachwissenschaftlicher Sicht: Ein Plädoyer für Vagheit und Pluralität.
- Das Konzil der Lexikographen
- Lässt sich die Unklarheit sozialwissenschaftlicher 'Begriffe' beheben?
- Doch die Begriffe, sie sind nicht so
- Wissenschaftliche Begriffsbildung und das Problem der induktiven Ambiguität
- Der Begriff der "Begriffskultur" aus konstruktivistischer Perspektive
- Begriffstheorie: Basis einer Theorie von Dokumentationssprachen - Basis zur Erklärung kognitiver Informationsverarbeitung

[4] vgl. https://groups.uni-paderborn.de/ewe/index2c6d.html?id=86

- Begriffe, Konzeptionen und Beispiele
- Begriffskultur als Ordnung?
- Eine analytische Begriffstheorie?
- Brauchen wir einen begriffskulturellen DIN-Ausschuß?
- Begriffskultur vs. Termassoziation: Gegen informationswissenschaftliche Problemlösungen durch sozialwissenschaftliche Sprachregelung
- Ordnung und Wahrhaftigkeit versus Realitätsadäquanz und Erkenntnisfortschritt? Sozialwissenschaftliche Begriffskultur zwischen Szylla und Charybdis
- Technisch operationelles versus reflexives philosophisches Denken
- Begriffskultur als Nächstenliebe? Nutzen und Relevanz von Klassifikationen
- Brief
- Zur Rolle von Theorie und Formalisierung bei der Begriffsbildung
- Lässt sich Chancengleichheit durch Begriffe normieren?
- Begriffsklärung als ein metatheoretisches Problem
- Puritanismus der Erkenntnis
- Stabilität und Variabilität von Begriffssystemen
- 'Begriffskultur' - oder: ein Beispiel dafür, wie man es nicht machen sollte
- Vom „Chaos der denkenden Ordnung" zu einer „Ethik statischer Begriffssysteme"
- Bemerkungen zur Begriffstheorie und zur Begriffsethik
- Problemkultur in den Sozialwissenschaften
- Die Problematik wissenschaftlicher Definitionen
- Ordnung als Wissen? (Benseler et al., 1996, S. 3-91)

In der Replik hierzu beklagt Dahlberg u.a., dass sie nicht noch ausführlicher ihre Begriffstheorie vorgetragen habe, denn sie verstehe die Begriffsbildung über aussagenderivierte Merkmale deskriptiv, während die Kritiker wohl eher eine formalisierte Begriffsbildung verstanden haben. Günter Endruweit stellt in seiner Metakritik hierzu u.a. fest:

> „Begriffe, die das Objekt einer Sozialwissenschaft bestimmen, müssten von ihrem Sinn her nicht in Schreibtisch-, sondern in Feldarbeit bestimmt werden. Sie sollen Ausschnitte aus der Wirklichkeit einer konkreten Gesellschaft bezeichnen, und hierbei hat die Gesellschaft – anders als die Objekte der Naturwissenschaften – die Befugnis zur Selbstdefinition. [...] Selbst wenn wir uns sehr um theoriebezogene Systematisierung bemühten, würden wir kein komplettes Begriffssystem, wie Dahlberg es offensichtlich ansteuert, erstellen können, und zwar nicht einmal mehrere parallele, für jede Theorie eines. Das liegt daran, dass wir immer noch, jedenfalls in der Soziologie, mit ‚Theorien mittlerer Reichweite' arbeiten müssen, also immer noch mit einer großen Zahl von Teilsystemen, deren Zusammenhang noch ungeklärt ist." (Endruweit, 1996, s. 88-90)

## 8 DESIDERATE FÜR DIE WISSENSORGANISATION

2009 trug Dahlberg (2013) auf der Deutschen ISKO-Konferenz in Bonn zehn Desiderate für Wissensorganisation vor, welche wesentlich auch die Institutionalisierung des Fachgebietes Wissensorganisation betreffen (in Englisch veröffentlicht später in der Knowledge Organization KO-2011-1 als „How to improve ISKO's standing: ten desiderata for knowledge organization"). Diese Desiderate sind:

1. Erkenne die Einheiten eines Ordnungssystems als Wissenseinheiten/Begriffe und benutze ihre Merkmale zur Herstellung einer Wissensordnung.
2. Die Erstellung von Übersichten der verwendeten Ordnungssysteme, um zu erkennen, wo Schwerpunkte und Präferenzen liegen.

3.1 Es sollte einen WO-Ausbildungslehrplan und eine entsprechende Berufsbezeichnung für die Absolventen erarbeitet werden.

3.2 Eventuell könnte ISKO selbst eine Akademie begründen und Lehrkräfte ausbilden.

4. Die nationalen ISKO-Chapter und das Generalsekretariat sollten sich um Einsetzung und Finanzierung einer Fachkraft bemühen.
5. Die ISKO sollte eine systematische Ordnung aller WO-relevanten Begriffe erarbeiten und als Modell-System für andere Wissensgebiete veröffentlichen.
6. Die Einrichtung von nationalen Instituten der Wissensorganisation sollte beantragt werden. Ihr Aufgabenbereich bezieht sich auf die Erarbeitung von Wissensordnungen, Ausbildungsaktivitäten und Forschungsarbeiten.
7. ISKO Experten sollten aktiv und gezielt ihr „Know-How" dort anbringen, wo es bekannt gemacht werden und damit seine Nützlichkeit erweisen kann. Sie sollten als „Anlaufstellen" erreichbar sein und sowohl beratend als auch statistisch und publizistisch tätig werden.
8. Weltweit sollten die auf dem Gebiet der Klassifikation/Indexierung und Thesauruserstellung tätigen Kollegen auf eine Mitgliedschaft in der ISKO angesprochen werden.
9. ISKO möge das Wissen ihres eigenen Wissensbereiches publizieren, wobei auf professionelle Sachregister geachtet werden müsste.

10.1 Die Wissensorganisation sollte als eigenständige Disziplin erachtet werden, die im Bereich der Wissenschaftswissenschaft anzusiedeln wäre, da sie nur auf diese Weise für ihre vielen möglichen Anwendungsgebiete fungieren kann.

10.2 Eine Wissensordnung „auf einen Blick" könnte durch die ICC ermöglicht werden, da diese von Disziplin-orientierten Universalklassifikationen abweicht und einen einfachen Überblick über alle Wissensgebiete vermittelt. Es wird eine Synthese oder Einheit des zu Wissenden möglich.

## LITERATUR

Frank Benseler, Bettina Blanck, Rainer Greshoff, Reinhard Keil-Slawik, and Werner Loh (Eds.). 1996. *Ethik und Sozialwissenschaften : EuS ; Streitforum für Erwägungskultur.* Vol. 7,1. Westdt. Verl., Opladen.

Ingetraut Dahlberg. 1974a. *Grundlagen universaler Wissensordnung.* Verlag Dokumentation, Pullach.

Ingetraut Dahlberg. 1974b. Zur Theorie des Begriffs. *International Classification* 1, 1 (1974), 12–19.

Ingetraut Dahlberg. 1978. Ontolgical Structures and Universal Classification. In *Sarada Ranganathan Endowment for Library Science*. Bangalore.

Ingetraut Dahlberg. 1979. On the theory of the concept. In *Ordering systems for global information networks. Proc. 3rd FID-CR Intern. Study Conf. Bombay 1975*, A Neelameghan (Ed.). Bangalore, 54–63.

Ingetraut Dahlberg. 1982. ICC Ĭnformation Coding Classification"Principles, Structure and Application Possibilities. *Int. Classif.* 9, 2 (1982), 87–93.

Ingetraut Dahlberg. 1987. Die gegenstandsbezogene, analytische Begriffstheorie und ihre Definitionsarten. In *Beiträge zur Begriffsanalyse. Mannheim: BI Wissenschaftsverlag*, Bernhard Ganter, Rudolf Wille, and Karl Erich Wolff (Eds.). 9–22.

Ingetraut Dahlberg. 1998. *Wissensorganisation.* Oldenbourg Verlag, München.

Ingetraut Dahlberg. 2006a. Definitionen aus dem Begriffsfeld "Wissensorganisation". (2006).

Ingetraut Dahlberg. 2006b. What is Knowledge Organization. (2006). Beiträge im August 2006 zur Diskussionsliste wiss-org@bonn.iz-soz.de sowie zur Diskussionsliste isko-l@lists.gseis.ucla.edu. Nicht mehr im Internet verfügbar.

Ingetraut Dahlberg. 2009. Concepts and Terms: ISKO's Major Challenge. *Knowledge Organization* 36, 2/3 (2009), 169–177.

Ingetraut Dahlberg. 2011. How to improve ISKO's standing: ten desiderata for knowledge organization. *Knowledge Organization* 38, 1 (2011), 69–74.

Ingetraut Dahlberg. 2013. Desiderate für die Wissensorganisation.. In *Wissen – Wissenschaft – Organisation. Proceedings der 12. Tagung der Deutschen ISKO.*, H. Peter Ohly (Ed.). Ergon Verlag, Würzburg.

Ingetraut Dahlberg. 2014a. Brief communication: What is knowledge organization? *Knowledge Organization* 41, 1 (2014), 85–91.

Ingetraut Dahlberg. 2014b. *Wissensorganisation – Entwicklung, Aufgabe, Anwendung, Zukunft.* Textbooks for Knowledge Organization, Vol. 3. Ergon Verlag, Würzburg.

Ingetraut Dahlberg. 2016. Dokumentenkunde - Dokumentologie: damals - und heute? *Information - Wissenschaft & Praxis* 67, 4 (2016), 195–203.

Ingetraut Dahlberg. 2017. Was ist Wissensorganisation?. In *Theorie, Semantik und Organisation von Wissen.*, Wieslav Babik, Peter Ohly, and Karsten Weber (Eds.). Ergon, Würzburg, 12–21.

Alwin Diemer. 1969. Versuch einer Systematik der „allgemeinen Wörter". Arbeitsunterlage für die Sitzung des Komitees Thesaurusforschung der DGD. 9.5.1969. (Mai 1969).

Günter Endruweit. 1996. Probleme sozialwissenschaftlicher Begriffsbildung. *Ethik und Sozialwissenschaften : EuS ; Streitforum für Erwägungskultur* 7, 1 (1996), 88–90.

James K. Feibleman. 1954. Theory of Integrative Levels. *British Journal for the Philosophy of Science* 5, 17 (1954), 59–66.

Gottlob Frege. 1969. *Funktion, Begriff, Bedeutung. Fünf Logische Studien.* (3 ed.). Vandenhoeck and Ruprecht, Göttingen.

Claudio Gnoli. 2015. A place for each toy. An interview with Ingetraut Dahlberg. *AIDAinformazioni. Rivista di scienze dell'informazione* 33, 1-2 (2015), 207–211.

Nicolai Hartmann. 1964. *Der Aufbau der realen Welt. Grundriss der Allgemeinen Kategorienlehre* (3 ed.). de Gruyter, Berlin.

Ia C. McIlwaine and Joan S. (eds.) Mitchell. 2008. Interview with Ingetraut Dahlberg. December 2007. *Knowledge Organization, Special Issue* 35, 2/3 (2008), 82–85.

Jean M. Perreault. 1994. Categories and Relators: A New Schema. *Knowledge Organization* 21, 4 (1994), 189–198.

Gerhard Romen. 2012. Ingetraut Dahlberg - 53 years driving innovation in the science of documentation and knowledge. (2012). https://www.youtube.com/watch?v=NtT5V9e4AsY

Günther Wahrig. 1975. *Deutsches Wörterbuch.* Bertelsmann-Lexikon-Verlag, Gütersloh.

# The New German Copyright Law for Science and Education (UrhWissG): Consequences for DH-Projects Working with Non-Academic Partners

Burkhard Meyer-Sickendiek
Department of Literary Studies, Freie Universität Berlin
Berlin, Germany
bumesi@zedat.fu-berlin.de

Hussein Hussein
Department of Literary Studies, Freie Universität Berlin
Berlin, Germany
hussein@zedat.fu-berlin.de

## ABSTRACT

One of the most challenging aspects of digital humanities is the interaction with data sets from non-academic partners: How can we visualize the knowledge developed during the analysis of non-academic data sets, especially when these sets are copyright protected? And what kind of data analysis in DH-projects will be possible, when the new German Copyright Law for science and education (UrhWissG) will enter into force on 1 March 2018? Our Project is funded since 2017 by the Volkswagen Foundation in the funding line *Mixed methods in the humanities*, our paper reports first activities towards the old and new situation concerning the interaction with data sets from non-academic partners. The aim of our project is to develop a digital tool to identify rhythmical patterns in spoken poetry, focusing the data from *lyrikline* which is the most famous online portal for spoken poetry. The corpus consists of nearly 11, 000 modern as well as postmodern poems read aloud by the original authors. This kind of readout-poetry contains a really new form of knowledge for the literary studies: the prosodic patterns of free verse poetry. We identified and explained a total of 17 prosodic patterns being characteristic for the modern and postmodern German as well as US-American poetry. In this paper, we discuss the new situation before and after the new German Copyright Law with regards to the analysis of this data set, and how to make it accessible to academic teaching.

## KEYWORDS

non-academic data sets in digital humanities, poetry websites and online archives, free verse prosody, rhythmical patterns

**Reference:**
Burkhard Meyer-Sickendiek and Hussein Hussein. 2018. The New German Copyright Law for Science and Education (UrhWissG): Consequences for DH-Projects Working with Non-Academic Partners. In *Proceedings of Wissensorganisation 2017: 15. Tagung der Deutschen Sektion der Internationalen Gesellschaft für Wissensorganisation (ISKO) (WissOrg'17)*. German Chapter of the ISKO / Freie Universität Berlin , pp. 11-16. https://doi.org/10.17169/FUDOCS_document_000000028863


*WissOrg'17, December 2017, Berlin, Germany*

https://doi.org/10.17169/FUDOCS_document_000000028863

## 1 INTRODUCTION

The analysis of poetry usually focuses on printed texts. With regards to metrics, this has meant that the distinction between stressed and unstressed syllables was the most important clue in order to identify the rhythmical shape. Poetic prosody has been analyzed mainly by focusing on metrical schemes such as iambic or trochaic meter (Frank, 1993, Wagenknecht, 1999). In our digital age, these metrical patterns are no longer sufficient because we can not just read, but also listen to poetry today (Bung and Schrödl, 2017). Internet websites for poetry like www.lyrikline.org, www.writing.upenn.edu/pennsound, www.ubu.com, www.poetryfoundation.org, www.poetryarchive.org, or www.poets.org as well as poetry collections from the *Hörverlag* (www.randomhouse.de) offer new insights of how the poets own phrasing creates prosodic features in terms of new rhythmical patterns. The so-called *free verse prosody* which has been created by modern and postmodern poets like Whitman (Gates, 1985)(Gates, 1990), the Imagists (Beyers, 2001, Cooper, 1998, Silkin, 1997), the Black Mountain poets (Golding, 1981, Steele, 1990), and contemporary Hip-Hop poets (Bradley, 2009), might be the most important excample of such rhythmical patterns. The *free verse prosody* means a postmetrical idea of rhythmical patterns, a novel form of prosody, accent, rhythm, and intonation (Andrews, 2017, Finch, 2000, Hartman, 1980).

Our Project, funded since 2017 by the Volkswagen Foundation (german: Volkswagenstiftung) in the funding line *Mixed methods in the humanities*, tries to identify these new rhythmical patterns in modern and postmodern poetry by using digital pattern recognition techniques to analyze a spoken poetry corpus. In this sense, we develop a new knowledge about prosodic patterns, whereas existing tools like *Metricalizer* (Bobenhausen, 2011) and *Sparsar* (Delmonte and Prati, 2014) do not work for these new prosodic features as long as they are trained on metrical patterns like the iambic or trochaic meter. The challenging aspect of our project is the data set: Whereas several projects in the area of digital humanities are dealing with corpora representing literature from the 18th and 19th century, our project is dealing with contemporary Poetry taken from our non-academic partner *lyrikline* (www.lyrikline.org). This website offers over 10, 800 poems by mostly living authors. The analysis of this corpus is somehow ambivalent: It offers totally new insights in various postmetrical prosodic forms and styles. But it also limits the possibilities of analysis, as long as it contains several problems concerning different aspects of copyright. The poems hosted on the *lyrikline* website offer a rich variety of modern and postmodern *free verse prosody*. But the big challenge for research projects dealing with contemporary data sets is to visualize the results of research.

Our project is aiming at classifying the different rhythmical forms and patterns to be gained from analyzing read-out-poetry, and to make use of these classified patterns for academic teaching. This is why we have to invent new ways of organizing digital knowledge gained from our analysis, insofar as the poems representing our classifications to date are not allowed to leave the *lyrikline* server. A further problem is the identification of rhythmical patterns, which is based on a certain philological method including three different steps: a) grammetrical ranking; b) rhythmical phrasing; and c) mapping rubato and prosodic phrasing (Berry, 1997, Finch, 2000, Silkin, 1997). We aim at realizing these steps by creating a software solution. But how can we make use of the digital text/sound corpus in order to develop the software tool without violating copyright issues? The answer is quite easy: We will have to wait until march 2018, when the new law on copyright for the knowledge society comes into effect. In our paper we describe how to get along with this specific situation; and offer further advices for Digital Humanities (DH)-research dealing with non-academic partners and using contemporary artefacts.

The paper is organized as follows: Section 2 gives an overview on the project *Rhythmicalizer*. The database is presented in the Section 3. Section 4 discusses several aspects of the German Copyright Law with special regards to the knowledge to be developed and organized during the analysis of the data set. A special focus will be on the new law on copyright for the knowledge society, which starts on 1st of March 2018 and will reform the regulations on the use of copyright works for education and research. The speech processing tools and data analysis are described in Section 5. Finally, conclusions and future works are presented in Section 6.

## 2 RHYTHMICALIZER

The aim of the project *Rhythmicalizer. A digital tool to identify free verse prosody* (www.rhythmicalizer.net) is to develop a software for the automatic recognition of rhythmical patterns in modern and postmodern poetry. We use the database of the web portal *lyrikline* (see section 3), initiated by the *Literaturbrücke Berlin e.v.* in 1999, which hosts international poetry (currently of 1, 200 poets, almost 10, 800 poems in 70 languages, with approx. 16, 000 translations) by presenting each poem online not only as a text, but also as a unique reading of the poet. This website contains speech and text data of modern and contemporary poetry, giving us access to hundreds of hours of author-spoken poetry. The three-year project is funded by the Volkswagen Foundation in the *Mixed Methods* program. During this period, *lyrikline* offered us the right to use their data-collection.

## 3 DATABASE

The data used in our project is a huge collection of modern and postmodern readout poetry taken from our partner *lyrikline*. *Lyrikline* hosts contemporary international poetry as audio files (read by the authors themselves) and texts (original versions & translations). Users can listen to the poet and read the poems both in their original languages and various translations. The digital material covers more than 10, 800 poems from more than 1, 200 international poets from more than 70 different languages and with approx. 16, 000 translations. Nearly 80% of the *lyrikline* poems are postmetrical poems. In our project, we will focus on those poems written and read out in English and German language (more than 3, 600 poems). We analyze these poems in order to identify rhythmic patterns, based on the theoretical debate of *free verse prosody*. Ideally, this digital analysis would be based on the following interactive steps to be done with our non-academic partner:

(a) the transmission and scientific investigation of the data.
(b) the processing of the data on a server of the Free University Berlin (FU Berlin) corresponding to the Information Technology (IT)-security guidelines of the FU Berlin, that can be used by the project staff.
(c) use of excerpts of poems on a scientific quotation basis for illustration in publications, lectures and further research material, possibly also on the website of the research project.

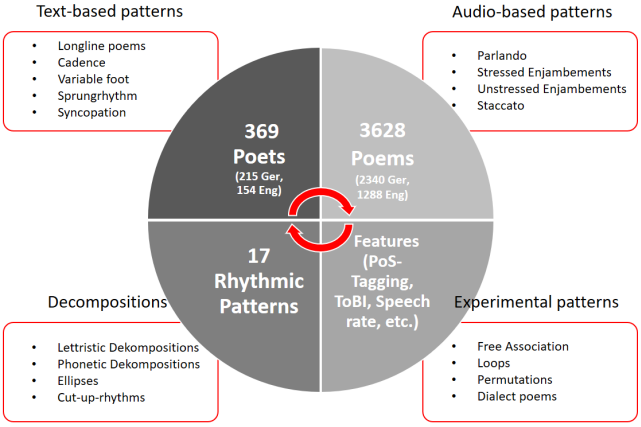



**Figure 1: The main components in the project: database, distribution of rhythmical patterns according to the detection concept based on philological analysis, and the digital tools used for feature extraction.**

## 4 KNOWLEDGE IN DIGITAL HUMANITIES?

The research project *Rhythmicalizer* aims, as described in section 2, to identify the prosody of modern and postmodern poetry by focusing on poetry as read aloud by the original authors. A digital tool for prosody recognition will be developed to structure readout poetry with special regards to the specific rhythm of these readout poems. In his study of *The Scissors of Meters: Grammetrics and Readin*, Donald Wesling distinguished five different prosodic types in the history of *free verse prosody*. In reference to this typology, we distinguish 17 rhythmic patterns from four sample types. Figure 1 shows the database and the distribution of rhythmical patterns. The knowledge that results from this project refers to an US-American scientific discourse, the so-called *free verse prosody*. As long as this theoretical discourse is rather unknown in Germany, we will try to prepare it for academic teaching in the bachelor and master area. With regards to our project, the idea of knowledge means the development of certain rhythmical patterns according to this *free verse prosody*. These rhythmic-prosodic patterns are fundamentally different from the classic patterns of metrical poetry, such as the jamb or the trochee: This is how we offer new knowledge for academic teaching.

## 4.1 Copyright Issues

The key idea of our project is to prepare or illustrate the theory of *free verse prosody* by using digital data. As such, we are a digital humanities project, dealing with digital data by trying to develop digital software being able to structure these digital data. The *Literaturbrücke Berlin e.v.* has basically agreed to provide us with the German and English poems by *lyrikline*. So far, the agreement applies that the data for our purposes will be available on the *lyrikline* website. However, there are a number of difficulties in this respect: Juridically speaking, we are allowed to analyze the *lyrikline* data for our scientific purpose. But to do the full research, we somehow have to wait till next spring: on 1 March 2018, the new German Copyright Law for science and education (german: Urheberrechts-Wissensgesellschafts-Gesetz (UrhWissG)) (UrheberrechtsWissensgesellschaftsGesetz, 2017) will enter into force. From that very moment, the provision of parts of copyright-protected works in (electronic) course reserves for teaching purposes and for the purpose of one's own research will be standardized in § 60 of the new Copyright Law (UrhWissG). This new law will allow us to automatically evaluate a large number of works (source material) for scientific research. A new § 60d of the Copyright Law will give us the permission:

(1) to automatically and systematically reproduce the source material in order to create a corpus to be evaluated, in particular by normalization, structuring and categorization, and
(2) to make the corpus publicly available to a definite group of persons for joint scientific research as well as to individual third parties for the purpose of verifying the quality of scientific research (UrheberrechtsWissensgesellschaftsGesetz, 2017).

But till that date, we do not have the permission for a transmission of these data in terms of processing the data on a server of the FU Berlin, although we can guarantee the IT security guidelines of the FU Berlin. We are only allowed to use excerpts of the poems on a scientific quotation basis for an illustration in publications, lectures and further research material, as well as presenting links on the website of our research project. However, this point is somewhat difficult to grasp under the conditions of the digital humanities standard procedures. Basically, the § 51 of the older Copyright Law (german: Urheberrechtsgesetz (UrhG)) (Urheberrechtsgesetz, 1965) allows the "duplication, distribution and public communication of a published work for the purpose of the quotation, provided that the use in its scope is justified by the particular purpose." This procedure is permissible,

(a) if individual works are included after the publication in an independent scientific work to explain the content,
(b) individual parts of a work are published after publication in a separate language work,
(c) individual passages of a published work of music are cited in an independent work of music.

The § 51 of the traditional Copyright Law only mentions "scientific works", "independent literary works" and "independent works of music". But what about digital databases or digital software? These topics are addressed in the § 53 of the Copyright Law focusing on "Reproductions for private and other own use". These reproductions can be taken or downloaded from digital data-collections, if this serves a scientific purpose. Obviously the download is only legal if the content of that website is legal itself.

Although we will develop a software tool, our scientific analysis of the data will violate neither the moral rights (§ 12-14 UrhG) nor the so-called exploitation rights (§ 15 ff. UrhG) regarding patents, licences, proprietary rights of use and exploitation and other intellectual property rights. Our digital analysis of the *lyrikline* data will lead to the creation of a digital toolset enabling the automatic identification rhythmical patterns. But the text & sound corpus as such will not be a part of the software to be developed. The software will be developed on a training level using that corpus. But the digital result of that training will not include parts of that corpus itself. In other words: The model training uses the data of *lyrikline*, but the model itself contains only insights gained from the corpus.

A further important aspect is the didactic function of our analysis. This concerns in particular our rhythmic-prosodic design of modern and postmodern poems. Our goal is to re-sort parts of *lyrikline*'s collection into lyrical-prosodic patterns in order to make the poetic data accessible to university teaching. The results of our classification will be shown on the website of *lyrikline*, according to § 52a of the old Copyright Law, which allows the "public accessibility for teaching and research". In this regard, it is already now allowed to publish small parts of a work, small-scale works as well as individual contributions from newspapers or magazines for the education in schools or universities and to make quotations publicly accessible for course participants. However, this must not serve a commercial purpose. For this re-designing of the website, we will have to ask permission from the authors of *lyrikline*. But we expect these authors to agree, cause our project offers these authors the opportunity to become more relevant for university events of Bachelor's and Master's degree programs. So far, *lyrikline* was not yet sufficiently sorted for academic teaching. Our theoretical analysis of the *lyrikline* portal will offer a new kind of prosodic knowledge and new examples of rhythmical patterns on that *lyrikline* website itself, which might help these authors to become more canonical. But as mentioned before: The real corpus-based analysis can not begin until March 2017, when the new German Copyright Law comes into force.

## 4.2 Patterns as Knowledge in Digital Humanities

The greatest challenge for our project is to identify each individual poem as a variety of sequential data. Each line of a poem must be transformed into a sequence of prosodic elements. Each poem will be separated into prosodic segments in the first step. Then, the combination of these prosodic segments will be assigned to a particular type of rhythmical patterns. According to (Cooper, 1998)(Andrews, 2017), we expect to identify the following rhythmical patterns (Figure 1 shows the same rhythmical patterns according to the detection concept based on philological analysis):

(1) **longline poem** was developed by Walt Whitman, following the Psalms in the King James Bible. It was taken over by Alan Ginsberg in Howl, became famous in Germany by Walter Höllerers influential theory of the long poem, which

was adapted by Durs Grünbein, Jürgen Becker, Franz Hodjak, Paul Celan, or Stephan Hermlin. Often, the long line is connected to an end-stopped-line;

(2) **cadence** is a sentence-based prosodic repetition coined by the American Imagists (Fletcher, Hulme, Pound, Lowell), which was highly influential for German poets like Gottfried Benn, Hans Magnus Enzensberger, Nicolas Born, and Franz Mon;

(3) **variable foot** is a colon-based prosodic repetition of a "breath-controlled-line". It has been developed by William Carlos Williams in his later poems, and used by German authors like Nicolas Born, Richard Anders, Ernst Jandl, Hans Magnus Enzensberger, and Harald Hartung;

(4) **syncopation** is a placement of rhythmical stresses or accents where they wouldn't normally occur, as used in the prosody of Jazz and Rap poetry by poets such as Hughes, Brown, and Jones, or Members of *The Last Poets* like Umar Bin Hassan, and German Rap-Poets like Bas Böttcher;

(5) **permutation** is a permutative rhythm based on the combination of two (or more) types of rhythms which overlap. The permutation was developed by John Cage and is used in the poems of Ernst Jandl, Franz Joseph Czernin, Jürgen Becker, Eugen Gomringer, and Oskar Pastior;

(6) **parlando**, a litany-like speech-song adopted by the rhythms of ordinary speech, which was very famous during the 1970s. The Parlando is used in poems of Gottfried Benn, Peter Rühmkorf, Hans-Magnus Enzensberger, Franz Hodjak, Michael Krüger, Rolf Haufs, or Uwe Kolbe;

(7) **unstressed enjambement** is an incomplete syntax at the end of a line; the meaning runs over from one poetic line to the next, without terminal punctuation. Many postmodern poets like Heiner Müller, Thomas Kling, Nicolas Born, Jürgen Becker, Elfriede Czurda, Marcel Beyer, Jan Völker Röhnert or Ann Cotten do not stress their enjambments;

(8) **stressed enjambment** is the same kind of an incomplete syntax at the end of a line, which is now emphasized. The stress or the emphasis of the enjambement has a long tradition in the poetry of the former German Democratic Republic (GDR). It was invented by Bertolt Brecht and continued by Karl Mickel or Kerstin Hensel;

(9) **loop** means an unchanged repeated linguistic sequence used in Hip-Hop or the 'rapping' kind of Slam poetry (Edwin Torres, Bob Holman, Sapphire, Saul Williams, Maggie Estep, Dana Bryant, Sekou Sundiata, Amir Sulaimanis, Paul Beatty, Linton Kwesi Johnson, or Bas Böttcher);

(10) **ellipse** is caused by the omission of one or more words in a clause. The elliptic rhythm can be found in intertextual and experimental poetry (Paul Wühr, Friederike Mayröcker, Jürgen Becker, Marcel Beyer, Thomas Kling, and Bert Papenfuß);

(11) **lettristic decomposition** is an art of letters which operates beyond spoken language, the lettristic decomposition was invented by Isidore Isou and adapted by international Sound-Poets like Henri Chopin, Bob Cobbing, Amanda Steward, Jaap Blonk, and Valeri Scherstjanoi as well as German poets like Gerhard Rühm, Ernst Jandl, Hans G. Helms, Franz Mon, Oskar Pastior, and Michael Lentz;

(12) **phonetic decomposition** is a kind of sound poetry which is more based on phonemes and was developed by Kurt Schwitters, Bernard Heidsieck, Helmut Heißenbüttel, Franz Mon, Gerhard Rühm oder Michael Lentz;

(13) **sprung rhythm** is based on a number of stressed syllables in a line and permits an indeterminate number of unstressed syllables. It was developed by Gerard M. Hopkins, William Carlos Williams, and the Black Mountain poets, and was influential for German poetry since the 1960s (Nicolas Born, Rolf-Dieter Brinkmann, Karin Kiwus, etc.);

(14) **free association** forms the prosody of the écriture automatique of Surrealist authors such as André Breton, Robert Desnos, Hans Arp, and Philippe Soupault, and was highly influential for Friederike Mayröcker, Marcel Beyer or Thomas Kling;

(15) **staccato**, which forms an abrubt, detached and choppy poetry like in John Berryman, Thomas Kling or Walter Mehring;

(16) **cut-up-rhythm** is an aleatory literary technique in which a text is cut up and rearranged to create a new text. The concept can be traced to at least the Dadaists of the 1920s, but was popularized in the late 1950s and early 1960s by William S. Burroughs and Brion Gysin; and adapted by Rolf-Dieter Brinkmann;

(17) **dialect poems** are characterized by the fact that they use the specific dialect of a specific region, like in poems from Axel Karner, Ernst Jandl, H. C. Artmann, Wulf Kirsten and Franz Hohler.

A total of 369 poets (215 german and 154 english) will be analyzed (see Figure 1), each of them reading about 12 poems, so we have 2340 german poems and 1288 english poems resulting a total of 3628 poems. By now, we structured and analyzed the data on the project website at the FU Berlin, only mentioning one poem for each pattern. The estimated number of patterns mentioned above is as follows: longline poems (30), cadence (50), variable foot (70), syncopation (20), permutation (10), parlando (60), stressed enjambement (10), unstressed enjambement (100), loop (10), ellipse (50), lettristic decomposition (40), phonetic decomposition (30), sprung rhythm (40), free association (30), staccato (25), cut-up-rhythm (6), dialect (20). To become an example of a rhythmical pattern, the rhythmical pattern has to appear in a certain poem at least 5 times. We expect even more examples during the next two years when analyzing the whole *lyrikline* data set.

## 5 SPEECH PROCESSING AND DATA ANALYSIS

We already created a text-speech alignment for the written poems and spoken recordings. Beside, each poem analyzed was split into prosodic segments; the specific combination of these prosodic segments then was assigned to types of rhythm. Focusing on the 17 rhythmical patterns mentioned in section 4.2, we use the following tools for these analysis and feature extraction tasks:

(1) **PoS-Tagger**: the PoS tagging tool (e.g. (Toutanova et al., 2003)) will identify the linguistic unit dominating each line of a poem. This could be a periodic sentence (Pattern 1), a whole sentence (Pattern 2), a clause (Pattern 3, 7 and 8), an

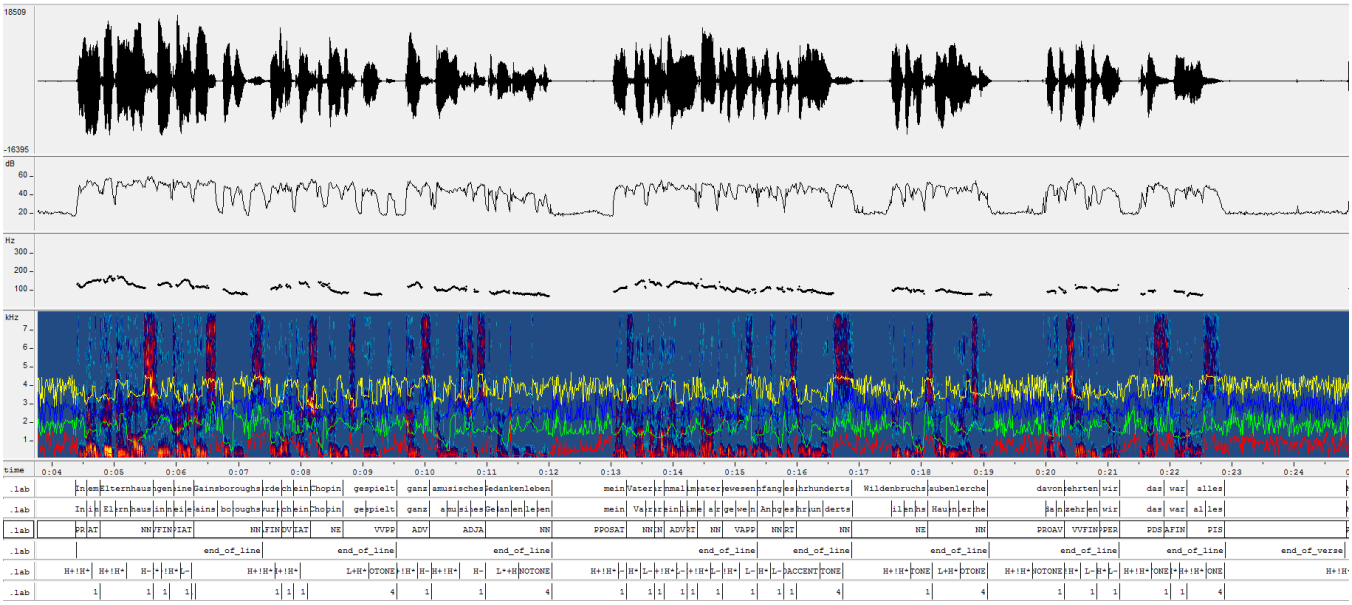


**Figure 2: Analysis of the first lines in the poem "Teils Teils" from the poet "Gottfried Benn" as an example for the "parlando" pattern shown from top to bottom: speech signal, intensity (dB), pitch (Hz), spectrogram, time, word alignment, syllable alignment, PoS-Tagging, End of line, ToBI tones, and ToBI breaks.**

elliptic phrase (7, 8, 10, 16), or an agrammatical expression (11, 12, 15, 16).

(2) **Forced Alignment**: the forced alignment toolkit (Sphinx (Walker et al., 2004)) used to adapt audio and text.

(3) **Intonation and Phrase Annotation**: the Automatic tool for TOnes and Break Indices annotation (AuToBI) (Rosenberg, 2010) toolkit used to identify the intended prosodic grouping of a poem (breaks) as well as the phrasal tones and pitch accents in order to differ between stressed and unstressed enjambements (3, 5, 7, 8).

(4) **Local Speech Rate**: Based on the local speech rate algorithm developed in (Pfitzinger, 1998), we want to create a software solution to identify the local speech rate as a combination of syllable and phone rates.

(5) **Bar and Beat Numbers**: The *Sonic Visualizer* (Visualiser, 2017) tool will help us to annotate the Bar and Beat numbers as well as to map Rubato and Loudness.

(6) **Rhyme Detection**: The automatic detection of rhyme in poetic texts is explained and developed in (Plechàc, 2017).

(7) **MARY-TTS**: The MARY Text-to-Speech (TTS) system, an open-source, multilingual Text-to-Speech Synthesis platform will be used to create a null-model for non-poetic speech to identify the poetic rhythm via contrast.

Figure 2 shows an example of the automatic analysis of pattern number 6 (parlando). The break index indicates that this prosodic pattern is similar to the variable foot (pattern number 3), but lacks the idea of a "breath controlled line". There is no regular stop at the end of each line.

## 6 CONCLUSION AND FUTURE WORK

In this paper, we presented the aim and the concept of the *Rhythmicalizer* project for the identificatiobn of rhythmical patterns in modern and postmodern poetry using audio & text data from non-academic partner. Regarding the problems of copyrights, we had found a juridical solution with regards to the copyright problem. The text-audio-alignments developed as part of the digital analysis of the data will not be published and will not be used any further after completion of our project, in particular not in connection with the software tool which we want to develop. For this reason, we either have to create an interface to visualize our results or have to present all our prosodic patterns on the website of our non-academic partner. Beside, we will write a letter to all the authors in collaboration with the staff of this portal. In this letter, the portal operators themselves will ask for permission to transmit the data to us (FU Berlin). At the same time, we have to develop our project further. We will therefore begin to download individual data from the *lyrikline* server. That's how we can advance our digital pattern classification - not commercial reproduction. This is legal, because *lyrikline* issued us a statement, according to which "the data after release of the project on *lyrikline* are retrievable". Problematic would be the transfer of the data by the operator *lyrikline* itself.

Our project provides to *lyrikline* the classification and pattern analysis of the data generated within the framework of the research project free of charge (if possible in a format compatible with the website of *lyrikline*, for example, XML or HTML). Our project hereby agrees to the inclusion of these classifications on the website of the *lyrikline*.

In the long run, our project will focus on pattern understanding using a supervised form of learning. The philological "teacher" gives

examples for the automatic classification. Our learning technique may be based on machine learning techniques available in common tools such as WEKA (Hall et al., 2009), or deep learning techniques. We will then analyze the computational models' success (in terms of the correctness of the decisions taken on previously unseen data) and their reasoning (in terms of the decision criteria used and how well they align with philological intuition). As one further point, we will make use of unsupervised machine learning techniques (e.g., clustering) to discover further hidden patterns in the audio and textual data, in particular for poems that are not well represented by any of the rhythmical types (1-17) mentioned and which may potentially enhance our understanding of *free verse prosody*. This step will use the machine-learned model by looking at borderline cases, analyzing miscategorizations, examining 'outliers', and using search patterns for prosodic aspects of poetry rather than a laborious manual search.

In accordance with the requirements of the Volkswagen Foundation, our project intends to use the software, which will be developed as part of the research project, as an open-source software for scientific or non-commercial purposes on a free repository such as *CLARIN* or *DARIAH* to make it available for free use or further development. Our project strives to implement the software tool in Java, Python and, if necessary, other programming languages on the basis of established tools in the Linux environment in order to be able to secure the functionality of the software in the long term. The *lyrikline* is free to use the software itself in this framework to the rest own costs.

## ACKNOWLEDGMENTS

This work is funded by the Volkswagen Foundation in the announcement *Mixed Methods in the humanities? Funding possibilities for the combination and the interaction of qualitative hermeneutical and digital methods* (funding code 91926 and 93255).

## REFERENCES

R. Andrews. 2017. *A Prosody of Free Verse: Explorations in Rhythm.* Routledge.
E. Berry. 1997. The Free Verse Spectrum. *College English* 59, 8 (1997), 873–897.
C. Beyers. 2001. *A History of Free Verse.* University of Arkansas Press.
K. Bobenhausen. 2011. The Metricalizer – Automated Metrical Markup of German Poetry. In *Current Trends in Metrical Analysis*, C. Küper (Ed.). 119––131.
A. Bradley. 2009. *Book of Rhymes: The Poetics of Hip Hop.* Basic Books.
S. Bung and J. Schrödl. 2017. *Phänomen Hörbuch: Interdisziplinäre Perspektiven und medialer Wandel.* Transcript Verlag.
G. B. Cooper. 1998. *Mysterious Music: Rhythm and Free Verse.* Stanford University Press.
R. Delmonte and A. M. Prati. 2014. SPARSAR: An Expressive Poetry Reader. Available on http://aclweb.org/anthology/E/E14/E14-2019.pdf. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2014, April 26-30, 2014, Gothenburg, Sweden.* 73–76.
A. Finch. 2000. *The Ghost of Meter: Culture and Prosody in American Free Verse.* University of Michigan Press.
H. J. Frank. 1993. *Handbuch der deutschen Strophenformen.* UTB, Stuttgart.
R. L. Gates. 1985. The Identity of American Free Verse: The Prosodic Study of Whitman's 'Lilacs'. *Language and Style* 18 (1985), 248—-276.
R. L. Gates. 1990. T. S. Eliot's Prosody and the Free Verse Tradition: Restricting Whitman's "Free Growth of Metrical Laws". Available on http://www.jstor.org/stable/1772826. *Poetics Today* 11, 3 (1990), 547–578.
A. Golding. 1981. Charles Olson's Metrical Thicket: Toward a Theory of Free-Verse Prosody. *Language and Style* 14 (1981), 64–78.
M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explorations* 11, 1 (2009), 10–18. https://doi.org/10.1145/1656274.1656278
C. Hartman. 1980. *Free Verse: An Essay on Prosody.* Princeton University Press.
H. R. Pfitzinger. 1998. Local Speech Rate as a Combination of Syllable and Phone Rate. In *Proceedings of 5th International Conference on Spoken Language Processing (ICSLP)*, Vol. 3. 1087–1090.
P. Plechàc. 2017. Collocation-driven Method of Discovering Rhymes in a Corpus of Czech, English, and French Poetic Texts. In *Proceedings of Plotting Poetry. On Mechanically Enhanced Reading.*
A. Rosenberg. 2010. AuToBI - A Tool for Automatic ToBI Annotation. In *Proceedings of INTERSPEECH.* ISCA, Makuhari, Chiba, Japan, 146–149.
J. Silkin. 1997. *The Life of Metrical and Free Verse in Twentieth-Century Poetry.* Palgrave Macmillan UK.
T. Steele. 1990. *Missing Measures: Modern Poetry and the Revolt Against Meter.* University of Arkansas Press.
K. Toutanova, D. Klein, C. D. Manning, and Y. Singer. 2003. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (HLT-NAACL).* Association for Computational Linguistics, Stroudsburg, PA, USA, 252–259.
Urheberrechtsgesetz. 1965. Gesetz über Urheberrecht und verwandte Schutzrechte (Urheberrechtsgesetz). Available on https://www.gesetze-im-internet.de/urhg/UrhG.pdf. (1965). Last accessed at 13. December 2017.
UrheberrechtsWissensgesellschaftsGesetz. 2017. Angleichung des Urheberrechts an die aktuellen Erfordernisse der Wissensgesellschaft (Urheberrechts-Wissensgesellschafts-Gesetz – UrhWissG). Available on http://dip21.bundestag.de/dip21/btd/18/123/1812329.pdf. (2017). Last accessed at 13. December 2017.
S. Visualiser. 2017. Sonic Visualiser. Available on www.sonicvisualiser.org/. (2017). Last accessed at 19. June 2017.
C. Wagenknecht. 1999. *Deutsche Metrik: eine historische Einführung.* Beck.
W. Walker, P. Lamere, P. Kwok, B. Raj, R. Singh, E. Gouvea, P. Wolf, and J. Woelfel. 2004. *Sphinx-4: A Flexible Open Source Framework for Speech Recognition.* Technical Report. Mountain View, CA, USA.

# Ontologie-basierte kognitive Karten

Von Coding Schemes zu Ontologien als Wissensorganisationssysteme für Digital Humanities

Ingo Frank
Leibniz-Institut für Ost- und Südosteuropaforschung
Regensburg, Deutschland
frank@ios-regensburg.de

## ABSTRACT

**English.** Cognitive maps were introduced in order to model complex social systems from the perspective of stakeholders. Common cognitive mapping is problematic because the causal factors are not explicitly specified. Therefore it is often not clear which kinds or types (actor, event, state, etc.) are involved and at what level (micro, meso, macro) the factors are located and to which domain (social, economic, political, historical, etc.) they have to be assigned. This article demonstrates how the representation of causal knowledge can be improved through the use of appropriate knowledge organization systems to build ontology-based fuzzy cognitive maps or dynamic cognitive maps. The concrete example of the advancement of the CAMEO coding scheme towards the PLOVER ontology illustrates the relevance of enhanced interdisciplinary collaboration between Digital Humanities and the Knowledge Organization. It is shown that the construction of cognitive maps requires an ontology as knowledge organizatio n system that allows the modeling of integrative levels to represent causal factors from different domains and their respective theory-dependence.

**Deutsch.** Cognitive Maps wurden eingeführt, um komplexe soziale Systeme aus Sicht beteiligter Akteure zu modellieren. Problematisch dabei ist, dass die kausalen Faktoren nicht explizit spezifiziert werden und daher oft nicht klar ist, um welche Arten oder Typen (Akteur, Ereignis, Zustand, etc.) es sich handelt und auf welcher Ebene (Mikro-, Meso-, Makro-) sich die Faktoren befinden und welchem Bereich (sozial, ökonomisch, politisch, historisch, etc.) sie zuzuordnen sind. Dieser Artikel zeigt, wie die Repräsentation kausalen Wissens in Ontologie-basierten Fuzzy Cognitive Maps oder Dynamic Cognitive Maps durch den Einsatz von geeigneten Wissensorganisationssystemen verbessert werden kann. Am konkreten Beispiel der Weiterentwicklung des CAMEO Coding Schemes zur PLOVER-Ontologie wird die Relevanz verstärkter interdisziplinärer Zusammenarbeit von Digital Humanities und Knowledge Organization verdeutlicht. Es wird gezeigt, dass zum Aufbau von kognitiven Karten eine Ontologie als Wissensorganisationssystem benötigt wird, welche die Modellierung integrativer Ebenen ermöglicht, um kausale Faktoren aus verschiedenen Bereichen und ihre jeweilige Theorie-Abhängigkeit zu repräsentieren.


*WissOrg'17, December 2017, Berlin, Germany*

https://doi.org/10.17169/FUDOCS_document_000000028863

## CCS CONCEPTS

• **Applied computing** → **Arts and humanities**; • **Information systems** → *Ontologies*; *Information extraction*; • **Computing methodologies** → *Vagueness and fuzzy logic*; *Topic modeling*;

## KEYWORDS

Political Event Coding, Topic Modeling, Coding Schemes, Dynamic Network Analysis, Dynamic Cognitive Maps

**Reference:**
Ingo Frank. 2018. Ontologie-basierte kognitive Karten: Von Coding Schemes zu Ontologien als Wissensorganisationssysteme für Digital Humanities. In *Proceedings of Wissensorganisation 2017: 15. Tagung der Deutschen Sektion der Internationalen Gesellschaft für Wissensorganisation (ISKO) (WissOrg'17).* German Chapter of the ISKO / Freie Universität Berlin , pp. 17-24. https://doi.org/10.17169/FUDOCS_document_000000028863

## 1 EINLEITUNG

In den Digital Humanities bestehen derzeit noch Lücken in der interdisziplinären Überschneidung mit relevanten Nachbardisziplinen: Neben Document Engineering (Piotrowski, 2015) und Usability Engineering (Burghardt et al., 2015) besteht auch im Bereich Knowledge Organization Potential für mehr Schnittmengenbildung mit den Digital Humanities (Frank, 2017)[1].

Wissensorganisation kann nach Ingetraut Dahlberg als Metawissenschaft aufgefasst werden (siehe Zitat in Hjørland, 2016). Wissensorganisation ist essentiell für geisteswissenschaftliches Verstehen in den Digital Humanities. – Denn wie schon der (Meta-)Historiker Hayden White schreibt: "the beginning of all understanding is classification" (White, 1978).

Coding Schemes werden beim Aufbau von Ereignisdatenbanken u. a. als Klassifikationssysteme für Ereignisse eingesetzt. Dabei werden sehr große Textmengen mit automatischen Event Coding-Verfahren verarbeitet. Aus den so gewonnen Ereignissen können Ereignisnetzwerke konstruiert werden, die mit Methoden der (dynamischen) Netzwerkanalyse weiterverarbeitet werden können. Die aufbereiteten Ereignisse bzw. Ereignisnetzwerke können außerdem zum Aufbau kognitiver Karten herangezogen werden. Die Aussagekraft von Coding Schemes reicht zur Organisation kausalen Wissens über komplexe Zusammenhänge zwischen kausalen Faktoren aus verschiedenen Bereichen der sozialen, wirtschaftlichen, politischen und historischen Realität allerdings nicht aus. Mir geht es in diesem Artikel daher um einen Ansatz zur Ontologie-basierten Modellierung kognitiver Karten.

[1] Es gibt allerdings neuerdings vermehrt Workshops oder Projekte, die sich konkret mit Wissensorganisation in den Digital Humanities (Flanders and Jannidis, 2015) oder allgemeiner mit formaler Modellierung in den Digital Humanities (Ciula et al., 2016) auseinandersetzen.

Damit die Versprechungen von 'Big Data' aber überhaupt erfüllt bzw. damit die großen Datenmengen überhaupt sinnvoll für weiterführende Verarbeitung zur Beantwortung geistes- und sozialwissenschaftlicher Forschungsfragen herangezogen werden können, müssen die Daten adäquat modelliert und organisiert werden. Im vorliegenden Artikel werde ich in diesem Kontext auf Ereignisdatenbanken, die mit Methoden des automatischen Political Event Coding erstellt worden sind, eingehen. An Beispielen aus der Global Database of Events, Language, and Tone (GDELT) werde ich zeigen, wie Coding Schemes als Wissensorganisationssysteme eingesetzt werden können, um große Datenmengen für explorative Analysen bereitzustellen. Dabei kann das Conflict and Mediation Event Observations (CAMEO) Coding Scheme gemäß Hjørland (2015) als Theorie oder zumindest als Theorie-abhängig betrachtet werden.[2] Noch deutlicher wird diese Auffassung von Wissensorganisationssystemen bei der Weiterentwicklung des Coding Schemes CAMEO zur Ontologie PLOVER (Political Language Ontology for Verifiable Event Records), wie ich anhand von Modellierungsbeispielen für Dynamic Network Analysis und Fuzzy Cognitive Maps bzw. Dynamic Cognitive Maps an der Notwendigkeit einer Theorie der ontologischen bzw. integrativen Ebenen (Kleineberg, 2017) zeigen werde.

## 2 VON POLITICAL EVENT CODING ZU DIGITAL MAPMAKING

Sowohl kognitive Kartierung als auch Netzwerkanalyse kann gemäß Szostak (2004) als Methode des *Mapmaking* betrachtet werden. Mapmaking ist damit nicht auf geografische Karten beschränkt (man denke etwa auch an Dialogue und Argument Mapping (Neil and Ann, 2012)) und im Bereich der Digital Humanities zeichnet sich ein *Digital Mapmaking* dadurch aus, dass als Grundlage für die Erstellung von Karten digitales Quellen- und Archivmaterial verwendet wird.[3] Die zum Digital Mapmaking erforderlichen Forschungsaktivitäten, einschließlich der verwendeten Methoden, Werkzeuge und Forschungsdaten, können im Rahmen von Digital Humanities-Projekten z. B. mit der NeDiMAH Methods Ontology (NeMO) (Constantopoulos et al., 2016, Hughes et al., 2015) – inkl. Wissensorganisationssystemen – formal erfasst und modelliert werden.

Automatisches Political Event Coding hat sich beim Aufbau von Ereignisdatenbanken für die Konfliktforschung als sehr viel schneller und außerdem deutlich zuverlässiger und vor allem auch konsistenter erwiesen als intellektuelles Coding (Schrodt, 2001). Trotzdem gibt es beim Einsatz der Algorithmen Probleme mit der Datenqualität aufgrund systematischer Fehler, was hier aber nicht weiter thematisiert werden soll (Grimmer and Stewart, 2013).

Political Event Coding beruht auf Natural Language Processing und ist durch die Verwendung von Coding Schemes für die Extraktion bestimmter Ereigniskategorien optimiert. Topic Modeling basiert auf Distributional Semantics, benötigt darum keine Coding Schemes und kann daher als Ergänzung zu Political Event Coding herangezogen werden. Darüber hinaus könnte Topic Modeling zum Aufbau von Coding Schemes eingesetzt werden. Structural Topic Modeling (Roberts et al., 2013) eignet sich sehr gut zur inhaltlichen Erschließung großer Nachrichtenkorpora unter Berücksichtigung verschiedener Quellen zur Analyse der Unterschiede zwischen Nachrichten von verschiedenen Nachrichtenagenturen (siehe Beispiel in Abbildung 1). Auf diese Möglichkeiten der Kombination von Methoden der Digital Humanities mit Methoden der Wissensorganisation soll hier aber nicht weiter eingegangen werden.


**Abbildung 1: Topic mit gegensätzlichem Schwerpunkt auf Verteidigung ('defens') und Besatzung ('occupi', 'occup') von Territorium in Meldungen von Nachrichtenagenturen der Konfliktparteien Armenien und Aserbaidschan im Berg-Karabach-Konflikt um den April 2016**

Die Integration verschiedener Quellen zu Konflikten in Informationssystemen ist nicht nur eine technische Herausforderung (Bennett, 2008), sondern erfordert bei der Einbindung und Aufbereitung von Sekundärquellen, wie Meldungen von Nachrichtenagenturen, 'Digital Source Criticism' als neue Form der Quellenkritik (Dulić, 2011). Die automatisch kodierten Ereignisdatensätze bilden die Grundlage für die Generierung von Ereignisnetzwerken (Brandes et al., 2009). Die Ereignisnetzwerke wiederum sind die Basis für explorative Netzwerkanalyse, dynamische Netzwerkanalyse und die Erstellung von dynamischen kognitiven Karten. Im folgenden ein kurzer Auszug aus den Ereignisdaten, die den Beispielen in Abschnitt 3 zugrunde liegen.

```
1980-08-19 VAT    POL    085 08 2  7.0 PL
1980-08-19 CHRCTH POL    010 01 1  0.0 PL
1980-08-20 POLCOP POLOPP 173 17 4 -5.0 PL
1980-08-21 USA    POLLAB 051 05 1  3.4 PL
```

Die Felder repräsentieren (von links nach rechts): Datum des Ereignisses, Code für Akteur 1 (source), Code für Akteur 2 (target), Code für Ereigniskategorie, Code für Basiskategorie, Ereignissklasse (1=Verbal Cooperation, 2=Material Cooperation, 3=Verbal Conflict, 4=Material Conflict), Wert auf der Goldstein-Skala zur Einschätzung des Einflusses des Ereignistyps auf die Stabilität des Landes, geografischer Fokus des Ereignisses. Die Codes für die Akteure repräsentieren in der Regel nationale oder internationale Akteure (z. B. `VAT` für den Vatikan und `POL` für Polen) inkl. vorhandener Typen oder Rollen, wie z. B. `POLCOP` für die polnische Polizei.

[2] "[A] theory is a knowledge organization system (KOS)—and vice versa: Any KOS is, if not a theory, at least theoretically and ideologically loaded." (Hjørland, 2015)

[3] "We are by now well into a phase of civilization when the terrain to be mapped, explored, and annexed is information space, and what's mapped is not continents, regions, or acres but disciplines, ontologies, and concepts." (Unsworth, 2002)

Zur Interaktion mit großen Datenbeständen folge ich im Projekt "Dynamics of Conflict and Cooperation Explorer" Shneidermans Visual Information-Seeking Mantra: "overview first, zoom and filter, then details on demand" (Shneiderman, 1996). Mit synchronoptischer Visualisierung kann man sich zuerst einen Überblick über die aus verschiedenen Konfliktdatenbanken stammenden Ereignisse und Daten aus weiteren Quellen (wie etwa ökonomische Daten aus World Bank Linked Data) verschaffen, anschließend in den Datenstrom hineinzoomen und ihn filtern. Coding Schemes dienen hier als Wissensorganisationssysteme zur Navigation in den großen Datenmengen (z. B. mit Treemap-Visualisierung). Als Alternative können die bereits erwähnten Structural Topic Models als Wissensorganisationssysteme für ein Navigationssystem dienen. Normdatensätze wie GND und JRC-Names können zur Identifikation und Verlinkung von Personen und Organisationen eingebunden werden. Um schließlich von visueller Exploration zu mechanistischer Erklärung überzugehen, wird dynamische Netzwerkanalyse verwendet, um bei Bedarf in die Details zu gehen.

Zur Analyse der Ereignisdaten werden nicht nur statistische Methoden (de Cadenas-Santiago et al., 2015) eingesetzt, sondern auch Methoden der dynamischen Netzwerkanalyse (Brandes and Lerner, 2008, Brandes et al., 2009).[4] Der zeitliche Aspekt darf bei der Analyse von Ereignisnetzwerken nicht vernachlässigt werden (Lemercier, 2015a,b), wobei sich das Relational Event Model (REM) (Butts, 2008) sehr gut für die Analyse von temporalen Netzwerken auf Basis von Ereignisdaten eignet. Besonders interessant für uns ist auch die Analyse von Teilnetzwerken, die in bestimmten Konstellationen auftreten, um z. B. mittels Frequent Subgraph Mining herauszufinden, wie Regierungen typischerweise auf Proteste reagieren (Keneshloo et al., 2014).

An dieser Stelle können wir ein kurzes Zwischenfazit dazu anbringen, was Werkzeuge und Methoden aus den Digital Humanities als Beitrag zur Wissensorganisation und umgekehrt leisten können: Topic Modeling kann zum Aufbau von Wissensorganisationssystemen eingesetzt werden (Hu et al., 2014). Wissensorganisationssysteme dienen als Grundlage für Navigationssysteme zur Navigation in großen Datenmengen.[5]

## 3 AUFBAU ONTOLOGIE-BASIERTER DYNAMIC COGNITIVE MAPS

Geistes- und sozialwissenschaftliche Untersuchungsgegenstände stellen besonders hohe Ansprüche an eine adäquate Modellierung, Wissensorganisation und -repräsentation als Grundvoraussetzung für eine digitale Bearbeitung im Forschungsprozess. Das folgende Zitat bringt diese speziellen Anforderungen besonders schön zum Ausdruck.

> What is at stake is the humanities' unique commitment to wrestle with uncertainty, ambiguity, and complexity; to model incommensurate temporalities and ontologies; to explore not just geographies but psychogeographies and the dark recesses of the self; to attend to nonrepeatable and nonstandard phenomena. (Burdick et al., 2012, S. 108)

In der Einleitung wurde bereits Wissensorganisation als Metawissenschaft charakterisiert – auch das Feld der Digital Humanities kann als Metawissenschaft definiert werden: "The digital humanities study the means and methods of constructing formal models in the humanities" (Piotrowski, 2016). 'Models' bedeutet dabei soviel wie Repräsentation geisteswissenschaftlicher Untersuchungsgegenstände und 'formal' meint logisch kohärent, nicht mehrdeutig und explizit. Das von Axelrod (1976a) eingeführte Cognitive Mapping und insbesondere die Weiterentwicklung zu Fuzzy Cognitive Mapping durch Kosko (1986) bietet dazu einen angemessenen Werkzeugkasten für die Entwicklung diagrammatischer ‚Denkwerkzeuge'[6], weil speziell letzteres durch den Einsatz von Fuzzy Logic den Aspekt der Unsicherheit und Ungenauigkeit formal erfassbar macht.[7]

Fuzzy Cognitive Mapping ermöglicht Einsichten in die Denk- und Sichtweise von an Konflikten beteiligten Akteuren (Dornschneider, 2016, Dornschneider and Henderson, 2016) oder auch Erklärungsansätze für andere komplexe soziale Phänomene, wie Migration (Tezcan, 2014). Diese explanatorische Funktion von Fuzzy Cognitive Maps "is focused on reconstructing the premises behind the behavior of given agents, understanding the reasons for their decisions and for the actions they take and highlighting any distortions and limits in their representation of the situation" (Papageorgiou and Salmeron, 2013) (nach (Codara, 1998)). Die Methode bietet aber auch eine strategische Funktion, um eine explizitere und übersichtlichere Repräsentation einer komplexen Situation zu schaffen (Codara, 1998, Papageorgiou and Salmeron, 2013). Das ist insbesondere auch für die Erforschung von so genannten 'Frozen Conflicts' relevant. Diese Konflikte im post-sowjetischen Raum können als 'Wicked Problems' aufgefasst werden (Eronen, 2016). Daher müssen sie mit Werkzeugen wie Issue-Based Information System (IBIS) (Neil and Ann, 2012, Shum and Okada, 2008) analysiert werden und erfordern wegen ihrer Komplexität eine Modellierung als dynamische Systeme (Coleman et al., 2006). Mit diesem Modellierungsansatz wird im Konfliktmanagement z. B. das Finden von unbeabsichtigten Spätfolgen möglich (siehe Fallstudie Gray and Roos, 2012).

Die Fuzzy Cognitive Map in Abbildung 2 basiert auf dem Causal Loop-Diagramm von Coleman et al. (2006). Die Karte stellt den Konflikt zwischen der polnischen Regierung und der Arbeiterbewegung Solidarność dar, d. h. die kausalen Einflussfaktoren (Knoten) sowie die kausalen Zusammenhänge (Kanten). Positive kausale Einflüsse sind blau dargestellt, negative rot. Die Dicke einer Kante repräsentiert die Stärke des kausalen Einflusses (in linguistischen Termen der Fuzzy-Logik etwa als „schwach", „mittel", oder „stark" zu bezeichnen). Die Größe der Knoten repräsentiert die Stärke der

[4] "[S]tatistical analyses summarize patterns in data, they do not explain them" (Hedström, 2006).

[5] Die Eisberg-Metapher – bekannt aus der Informationsarchitektur (Morville and Rosenfeld, 2002) – veranschaulicht diese Rolle von Wissensorganisationssystemen, die hinter der sichtbaren Benutzeroberfläche einer guten Informationsarchitektur stehen, sehr gut.

[6] Diagrammatisches Denken mithilfe von Diagrammen definiert Peirce als eine Vorgehensweise in drei Schritten: Aufbauen einer Repräsentation, Experimentieren mit der Repräsentation und Beobachten und Analysieren der Ergebnisse (siehe Hoffmann, 2006) – eine Vorgehensweise, die sehr gut zur Modellierung und Simulation komplexer Systeme mit Fuzzy Cognitive Maps passt: Fuzzy Cognitive Maps sind Werkzeuge zur dynamischen Modellierung, womit üblicherweise einfach festgestellt werden kann, welche kausalen Faktoren verändert werden müssen und wie sie angepasst werden müssen (Papageorgiou and Salmeron, 2013). An dieser Stelle tut sich eine weitere Schnittmengenbildung auf – zwischen Semiotik und Digital Humanities …

[7] Nicht polemisch gemeint, kann man dazu anmerken: Soft Computing for Soft Sciences!

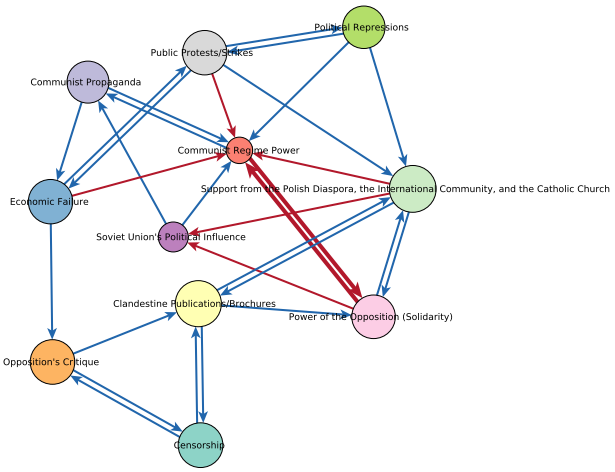



**Abbildung 2: Fuzzy Cognitive Map zum Konflikt zwischen der polnischen Regierung und Solidarność**

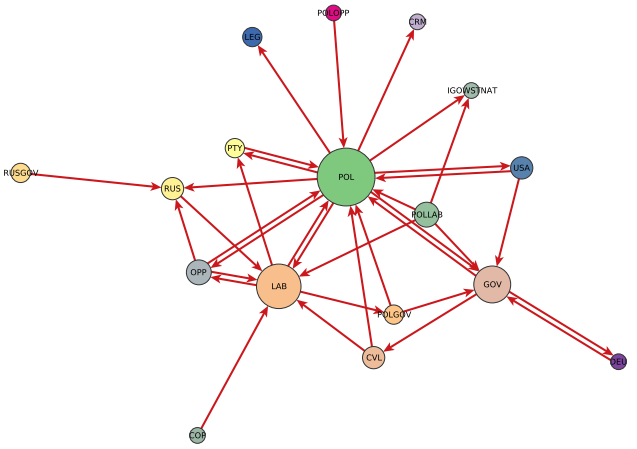



**Abbildung 3: Ereignisnetzwerk für Protestereignisse in Polen um 1980**

kausalen Faktoren innerhalb des modellierten Zustands. Im visualisierten Szenario ist die politische Macht des kommunistischen Regimes in Polen also bereits sehr stark durch den Einfluss der Gegenbewegungen, Proteste und den wirtschaftlichen Zusammenbruch eingeschränkt.

Um komplexe historische Prozesse inkl. Rückkopplungsschleifen wie in der Causal Loop-Analyse von Coleman et al. (2006) oder Gray and Roos (2012) zu modellieren, müssen Fuzzy Cognitive Maps den zeitlichen Aspekt der kausalen Zusammenhänge unterstützen (Carvalho, 2010, 2012, 2013). Damit erhält man Dynamic Cognitive Maps (Carvalho, 2013), die die Analyse der zeitlichen Dynamik von komplexen kausalen Systemen mit Simulationsexperimenten ermöglichen.

Ein Anwendungsfall für die Kombination von Dynamic Network Analysis und Dynamic Cognitive Maps ist die Überprüfung der Plausibilität von kausalen Einflussfaktoren über die Extraktion und Analyse von relevanten Teilnetzwerken. In unserem Beispiel in Abbildung 2 kann etwa das Protestnetzwerk untersucht werden, das sich hinter dem kausalen Konzept "Public Protests/Strikes" verbirgt. In Abbildung 3 wurden dazu Daten zu Protestereignissen (CAMEO Event Code 14) aus GDELT exportiert und mit visone (Brandes and Wagner, 2004) als Ereignisnetzwerk visualisiert. Das Netzwerk stellt ein recht frühes Stadium der Interaktion zwischen den Konfliktparteien dar. Um die für öffentliche Proteste und Arbeiterstreiks nicht relevanten ausländischen Akteure aus dem Interaktionsnetzwerk auszuschließen, können die zugrunde liegenden Ereignisdaten entsprechend nach relevanten lokalen Akteuren und deren Rollen gefiltert werden (siehe Tabelle 1).

Ein weiterer Untersuchungsgegenstand wäre das Unterstützernetzwerk hinter dem kausalen Konzept "Support from the Polish Diaspora, the International Community, and the Catholic Church". Gerade bei diesem kausalen Faktor wird die Integration von Ereignisnetzwerken interessant, da das Konzept aus einer ganzen Reihe von kooperativen kollektiven Akteuren besteht.[8]

[8] An dieser Stelle werde ich wegen der sehr langen Bezeichnung für ein kausalen Begriff im Beispiel oben kurz auf Möglichkeiten zur Verbesserung des 'Labeling' eingehen. Abramova et al. (2010) schlagen ein Kriterium für die Normalform der Begriffe von kausalen Faktoren vor. Demnach ist z. B. der Faktor "Condition of Economics" oder "Social Situation" nicht in Normalform, weil diese Faktoren nicht als Variablen mit numerischem Wert (sofern messbar) oder linguistischen Variablen der Fuzzy Logic (wie „hoch", „niedrig", etc.) aufgefasst werden können. Die Bezeichnung von Begriffen kann demnach rein linguistisch verbessert werden: z. B. mit "Quality of Economics Condition" anstatt "Condition of Economics". Noch besser wäre allerdings die Verwendung von entsprechendem Fachvokabular durch die Einbeziehung von Expertenwissen. Damit erhält man anstatt "Quality of Economics Condition" mit ökonomischer Fachsprache "Stability of Economics Development". – Beim Punkt Einbeziehung von Expertenwissen aus verschiedenen Fachbereichen wird hier auch Ontologie-basierte Modellierung relevant.

| Role Codes | Description |
|---|---|
| **Primary** | |
| COP | Police forces, officers, criminal investigative units |
| GOV | Government: the executive, governing parties |
| JUD | Judiciary: judges, courts |
| MIL | Military |
| OPP | Political opposition: opposition parties, individuals, |
| REB | Rebels: armed and violent opposition groups, individuals |
| SPY | State intelligence services and members |
| UAF | Unaligned armed forces |
| **Secondary** | |
| BUS | Business: businessmen, companies, and enterprises |
| CRM | Criminal: individuals involved in the breaking of law |
| CVL | Civilian: individuals or groups not otherwise specified |
| EDU | Education: educators, schools, students |
| ELI | Elites: former government officials, celebrities, spokespersons |
| LAB | Labor: individuals, organizations concerned with labor issues |
| LEG | Legislature: parliaments, assemblies, lawmakers |
| MED | Media: journalists, newspapers, television stations |
| **Tertiary** | |
| MOD | Moderate: "moderate," "mainstream," etc. |
| RAD | Radical: "radical," "extremist," "fundamentalist," etc. |

**Tabelle 1: Generic Domestic Role Codes im CAMEO Coding Scheme (Auszug aus Event Data Project, 2012)**

Einen guten Überblick über die Möglichkeiten zur Modellierung der Rollen von Akteuren bietet Tabelle 1 aus dem CAMEO Codebook. Mit den primären und sekundären Rollen aus dem CAMEO

Coding Scheme können die in die Ereignisse involvierten Akteure genauer spezifiziert werden.[9]

Die Art der kausalen Faktoren selbst müsste mit einem entsprechend erweiterten Wissensorganisationssystem klassifiziert werden.[10] Hinter "Public Protests/Strikes" steht ein Interaktionsnetzwerk, das als Ereignisnetzwerk modelliert werden kann. Der kausale Faktor kann daher als Ereignis (das aus einer Vielzahl von Ereignissen besteht) typisiert werden. Aber wie konstituiert sich z. B. "Communist Regime Power"? Hier haben wir es nicht mit einem Ereignis zu tun, sondern mit einem Zustand oder vielmehr mit einem strukturellen Faktor – sozialem Kapital – auf Makro-Ebene.

An dieser Stelle wird eine Theorie der ontologischen Ebenen relevant (Gnoli, 2008). Phänomenologische Ontologie nach Husserl[11] oder Hartmann (Hartmann, 2010) bildet den Rahmen für eine Ebenentheorie als Hintergrund für den Aufbau eines Wissensorganisationssystems, das es erlaubt, zwischen kausalen und konstitutiven Abhängigkeiten zu unterscheiden (Poli, 2007, S. 2). Mit GFO (General Formal Ontology) (Herre, 2013) können die Entitäten von kognitiven Karten verschiedenen ontologischen Ebenen zugeordnet werden. GFO unterscheidet ontologische Ebenen in *Strata* und *Level*, was den Schichten und Kategorien bei Hartmann entspricht. Damit lassen sich auch die fehlenden Ebenen in der Region bzw. Schicht des Sozialen wie folgt ergänzen:

```
gfo:Political_level a gfo:Level ;
  gfo:abstract_part_of gfo:Social_stratum .
gfo:Economical_level a gfo:Level ;
  gfo:abstract_part_of gfo:Social_stratum .
```

Dadurch wird auch die Ontologie-basierte Modellierung von sozialen Mechanismen (Hedström and Swedberg, 1998) möglich, was für die Unterstützung mechanistischer Erklärung auf verschiedenen Organisationsebenen durch wissensbasierte Systeme erforderlich ist.

Die Weiterentwicklung von CAMEO zur PLOVER-Ontologie liefert dazu mit dem neuen Kategorienschema zur Modellierung von Ereignissen auch bereits einen ersten Ansatz. In CAMEO gibt es sehr viele Unterkategorien von Ereignissen, um den Bereich anzugeben, in dem ein Ereignis stattfindet. Beispielsweise hat die Ereigniskategorie `07: PROVIDE AID` in CAMEO die Unterkategorien `071 (Provide economic aid)`, `071 (Provide military aid)`, `071 (Provide humanitarian aid)`, usw. In PLOVER hingegen gibt es nur die Kategorie `AID`, die mit dem entsprechenden Kontext kombiniert wird, also z. B. `AID economic` oder `AID military`.

Diese Kodierung von Ereignissen entspricht im Prinzip einer Facettenklassifikation. Damit wird nicht nur eine kombinatorische Explosion von Ereigniskategorien vermieden, sondern außerdem bewegt sich die Entwicklung der PLOVER-Ontologie mit ihren Kontexten, die im Grunde den Ebenen in den Schichten Kultur und Gesellschaft – oder allgemeiner in der Schicht des Geistigen (siehe Tabelle 2 in Gnoli, 2008) – entsprechen, in Richtung einer Theorie der integrativen Ebenen. Die Auflistung der PLOVER Ereignis-Kategorien verschafft einen ersten Eindruck der weiteren Kodierungsmöglichkeiten:

**Verbal cooperation** AGREE, CONSULT, SUPPORT, CONCEDE
**Material cooperation** COOPERATE, AID, RETREAT, INVESTIGATE
**Verbal conflict** DEMAND, DDISAPPROVE, DREJECT, DTHREATEN,SANCTION
**Material conflict** PROTEST, CRIME, MOBILIZE, COERCE, ASSAULT, FIGHT

Gerade für die Schichten Kultur und Gesellschaft, für die die Geistes- und Sozialwissenschaften zuständig sind, ist eine Ontologie-basierte Organisation und Repräsentation von kausalen und konstitutiven Zusammenhängen notwendig, weil sozusagen alles mit allem zusammenhängt:

> A distinctive feature of the social realm is the twofold action of its various domains: on the one hand, each of them operates individually according to its interpretative frame; on the other, all of them operate in parallel, influencing and determining each other [...] (Poli, 2001)

Die Frage ist nun, wie genau diese gegenseitigen Abhängigkeiten beschaffen sind und wie sie formal modelliert werden können. In der Forschung zu kognitive Karten gibt es bereits ein paar wenige Ansätze, die versuchen, kognitive Karten durch die Verwendung von Ontologien besser zu strukturieren. Shayji et al. (2011) versuchen Zusammenhänge zwischen dem politischen und dem wirtschaftlichen Bereich in einer Ontologie zu erfassen, um Daten aus den verschiedenen Sektoren zu integrieren. Chauvin et al. (2009) entwickeln eine Ontologie, um große kognitive Karten besser handzuhaben indem deren Begriffe bzw. kausale Faktoren anhand der Ontologie organisiert werden. Was diese Ansätze allerdings nicht berücksichtigen, ist die Theorie-Abhängigkeit von Begriffen:

> A theory is a form of KOS and theories are the point of departure of any KOS. It is generally understood in KO that concepts are the units of KOS, but the theory-dependence of concepts brings theories to the forefront in analyzing concepts and KOS. The study of theories should therefore be given a high priority within KO concerning the construction and evaluation of KOS. (Hjørland, 2015)

## 4 ZUSAMMENFASSUNG

Als Mehrwert für die digital erweiterte Forschung ergibt sich durch die Verwendung eines einheitlichen Wissensorganisationssystems mit Semantic Web- und Linked Data-Technologie die Möglichkeit der Integration von automatisch generierten Ereignis-Datensätzen zur Konstruktion von Ereignis-Netzwerken für dynamische Netzwerkanalyse und die Erstellung und Zusammenführung kognitiver Karten aus verschiedenen Quellen (Osoba and Kosko, 2017). Im Idealfall können so Dynamic Cognitive Maps zur multiperspektivischen Wissensorganisation und -repräsentation z. B. zur Erschließung historischer Narrative aus Sicht verschiedener historischer Akteure geschaffen werden. Damit würden die Digital Humanities den besonderen Ansprüchen geisteswissenschaftlicher Forschung

[9] Mit den tertiären Rollen schließlich können z. B. auch moderate Rebellengruppen in Syrien erfasst werden, wenn davon in Nachrichtenmeldungen die Rede ist.

[10] In den frühen Cognitive Maps aus (Axelrod, 1976b) erfolgte bereits eine Typisierung von Begriffen unterschieden in P-concepts (policy concepts), C-concepts (cognitive concepts), A-concepts (affective concepts) und V-concepts (value concepts) (siehe Bonham and Shapiro, 1976) und auch eine Kategorisierung nach Policy Domains. Allerdings sind diese Unterscheidungen zugeschnitten auf den konkreten Anwendungsfall der Analyse der Überzeugunssysteme von Entscheidungsträgern aus der Politik.

[11] In den wissenschaftstheoretischen *Ideen III* betont Husserl die Relevanz phänomenologischer Ontologie: „Nur der Phänomenologe wird befähigt sein, die tiefsten Klärungen hinsichtlich der in systematisch konstitutiven Schichten sich aufbauenden Wesenheiten zu vollziehen und so der Begründung der Ontologien vorzuarbeiten, die uns so sehr fehlen." (Husserl, 1952, § 20, S. 105)

an Modellierung, Wissensorganisation und -repräsentation ein deutliches Stück näher kommen.

Vorerst will ich versuchen, kausale Faktoren für den Aufbau von kognitiven Karten durch auf Basis von Ereignisdaten konstruierter Ereignisnetzwerke zu gewinnen: In einem Verfahren zum 'Cognitive Map Mining' könnten Fuzzy Cognitive Maps automatisch aus Forschungsdaten konstruiert werden, um 'big knowledge' zu produzieren (Kosko, 2014, Osoba and Kosko, 2017).[12] Dynamische Netzwerkanalyse auf Basis von automatischem Event Coding anhand Coding Schemes wie CAMEO oder auch PLOVER kann zur Unterstützung des Aufbaus von kognitiven Karten genutzt werden: Durch die Ausnutzung der Typisierung von Ereignissen mit Coding Schemes können die relevanten Elemente eines dynamischen Systems, also die kausalen Faktoren einer Fuzzy Cognitive Map gefunden werden. So kann über die Ereignistypen aus dem Coding Scheme z. B. ein Konfliktnetzwerk (mit den Konfliktparteien) als kausaler Faktor identifiziert werden. Die Rollen der Akteure in einem solchen Teilnetzwerk werden dabei auch über das Coding Scheme bzw. dessen Actor Ontology erfasst (siehe Tabelle 1). Der folgende Auszug aus der Actor Ontology des KEDS (Kansas Event Data System) Balkans Data Set veranschaulicht die Rollen-basierte Modellierung von Akteuren und deren Rollen:

```
NATO OFFICIAL [NAT]
NATO-LED STABILIZATION FORCE IN BOSNIA [NAT]
SERBS IN BOSNIA [BOSSER]
RATKO MLADIC [BOSSER]
MILOSEVIC [SERGOV 890101-971230]
          [FRYGOV 971231-001005]
          [SERSM >001006]
```

Zu den Grenzen des Einsatzes von Coding Schemes und zur Notwendigkeit der Weiterentwicklung zu Ontologien ist in diesem Zusammenhang die folgende Stelle aus dem Handbuch von CAMEO sehr aufschlußreich.

> In the present version of the manual, we are regularly using the phrase "ontology" to refer to what has been variously called in the past a "coding scheme", "coding framework" and probably any number of other things. Over the past couple of years we've taken to calling this an "ontology" since we've been interacting with a lot of folks in the informational sciences communities and they seem to be more comfortable with that phrase. However, [...] the event framework is probably only a taxonomy, though the actor framework is, in fact, heading towards being an ontology.[13]

Der eigentliche Mehrwert ergibt sich erst durch die Entwicklung und den Einsatz einer Ontologie der integrativen Ebenen, also durch den Ausbau der Coding Schemes zu Ontologien, um sowohl Ereignisnetzwerke als auch kognitive Karten – nicht zwangsläufig mit automatischen Verfahren, sondern auch durch intellektuelles 'documentary coding' (Kosko, 1986, Wrightson, 1976) – besser zu strukturieren. Damit können theoretische Hintergrundannahmen explizit gemacht werden, wie beispielsweise die soziologische oder politikwissenschaftliche Theorie, die hinter einer Netzwerkanalyse steht oder die Paradigmen, durch die Perspektiven von Akteuren in kognitiven Karten geprägt sind (siehe Tabelle 1 in Bennett, 2013). Wissensorganisation, die einer Theorie der integrativen Ebenen folgt, wird hier im Kontext von Interdisciplinary Knowledge Organization (Szostak et al., 2016) auch für die Unterstützung des mechanistischen Erklärungsansatzes relevant. "An interesting aspect of the mechanism approach is its interdisciplinarity" (Hedström and Swedberg, 1998): Mechanistische Erklärung auf verschiedenen Ebenen erlaubt die Integration von Erkenntnissen verschiedener Disziplinen (vgl. etwa Keestra, 2011).[14]

Als Ausblick auf die zukünftige Rolle der Wissensorganisation für die Digital Humanities dient abschließend ein historischer Rückblick:

> In some form, the semantic web is our future, and it will require formal representations of the human record. Those representations—ontologies, schemas, knowledge representations, call them what you will—should be produced by people trained in the humanities. Producing them is a discipline that requires training in the humanities, but also in elements of mathematics, logic, engineering, and computer science. Up to now, most of the people who have this mix of skills have been self-made, but as we become serious about making the known world computable, we will need to train such people deliberately. (Unsworth, 2002)

## LITERATUR


Nina Abramova, Zinaida Avdeeva, Svetlana Kovriga, and Dmitry Makarenko. 2010. Subject-formal Methods Based on Cognitive Maps and the Problem of Risk Due to the Human Factor. In *Cognitive Maps*, Karl Perusich (Ed.).

Robert Axelrod. 1976a. The Cognitive Mapping Approach to Decision Making. In *Structure of Decision: The Cognitive Maps of Political Elites*, Robert Axelrod (Ed.). Princeton University Press, 3–17.

Robert Axelrod (Ed.). 1976b. *Structure of Decision: The Cognitive Maps of Political Elites*. Princeton University Press.

Andrew Bennett. 2013. The mother of all isms: Causal mechanisms and structured pluralism in International Relations theory. *European Journal of International Relations* 19, 3 (2013), 459–481.

D. Scott Bennett. 2008. Merging and Meshing Data: Difficulties, Lessons, and Suggestions. In *Building and Using Datasets on Armed Conflicts*, Mayeul Kauffmann (Ed.). NATO Science for Peace and Security Series, Vol. 36. 133–159.

G. Matthew Bonham and Michael J. Shapiro. 1976. Explanation of the Unexpected: The Syrian Intervention in Jordan in 1970. In *Structure of Decision: The Cognitive Maps of Political Elites*, Robert Axelrod (Ed.). Princeton University Press, 113–141.

Ulrik Brandes and Jürgen Lerner. 2008. Visualization of Conflict Networks. In *Building and Using Datasets on Armed Conflicts*, Mayeul Kauffmann (Ed.). NATO Science for Peace and Security Series, Vol. 36. 169–188.

Ulrik Brandes, Jürgen Lerner, and Tom A. B. Snijders. 2009. Networks Evolving Step by Step: Statistical Analysis of Dyadic Event Data. In *2009 International Conference on Advances in Social Network Analysis and Mining*. 200–205.

Ulrik Brandes and Dorothea Wagner. 2004. Visone – Analysis and Visualization of Social Networks. In *Graph Drawing Software*, Michael Jünger (Ed.). Springer, Berlin, 321–340.

Anne Burdick, Johanna Drucker, Peter Lunenfelda, Todd Presner, and Jeffrey Schnapp. 2012. *Digital_Humanities*. MIT Press.

Manuel Burghardt, Christian Wolff, and Christa Womser-Hacker. 2015. Informationsinfrastruktur und informationswissenschaftliche Methoden in den digitalen Geisteswissenschaften. *Information – Wissenschaft & Praxis* 66, 5-6 (Jan. 2015).

Carter T. Butts. 2008. A Relational Event Framework for Social Action. *Sociological Methodology* 38, 1 (2008), 155–200.

João Paulo Carvalho. 2010. On the semantics and the use of Fuzzy Cognitive Maps in social sciences. IEEE, 1–6.


[12] Im Semantic Web können Ontologie-basierte Expertensysteme für die Digital Humanities entwickelt werden (Beispiel im medizinischen Bereich in Papageorgiou et al., 2012). Tsadiras and Bassiliades (2013) verwenden zum Aufbau eines Expertensystems RuleML zur Repräsentation und Prolog zur Simulation von Fuzzy Cognitive Maps. Siehe auch Soergel (2015) und Mayr et al. (2016) zum möglichen Beitrag von Knowledge Organization.

[13] https://github.com/openeventdata/Dictionaries/blob/master/CAMEO.Manual.1.1b3.tex

[14] Dazu wären auch "stacked or multilayered FCMs" zur Repräsentation kausalen Wissens relevant, was Kosko (2014) nur anspricht, aber nicht weiter ausführt.

João Paulo Carvalho. 2012. Rule Based Fuzzy Cognitive Maps in Humanities, Social Sciences and Economics. In *Soft Computing in Humanities and Social Sciences*, Rudolf Seising and Veronica Sanz González (Eds.). Springer, 289–300.

João Paulo Carvalho. 2013. On the Semantics and the Use of Fuzzy Cognitive Maps and Dynamic Cognitive Maps in Social Sciences. *Fuzzy Sets and Systems* 214 (March 2013), 6–19.

Lionel Chauvin, David Genest, and Stéphane Loiseau. 2009. Ontological Cognitive Map. *International Journal on Artificial Intelligence Tools* 18, 05 (2009), 697–716.

Arianna Ciula, Øyvind Eide, Cristina Marras, and Patrick Sahle. 2016. Modelling between Digital and Humanities: Thinking in Practice. In *Digital Humanities 2016: Conference Abstracts*. 762–763. http://dh2016.adho.org/abstracts/421

Lino Codara. 1998. *Le mappe cognitive: Strumenti per la ricerca sociale e l'intervento organizzativo*. Carrocci Editore, Roma.

Peter T. Coleman, Robin R. Vallacher, Andrzej Nowak, and Lan Bui-Wrzosińska. 2006. Protracted Conflicts as Dynamical Systems. In *The Negotiator's Fieldbook: The Desk Reference for the Experienced Negotiator*, Christopher Honeyman Andrea Kupfer Schneider (Ed.). American Bar Association, 61–74.

Panos Constantopoulos, Lorna M. Hughes, Costis Dallas, Vayianos Pertsas, Leonidas Papachristopoulos, and Timoleon Christodoulou. 2016. Contextualized Integration of Digital Humanities Research: Using the NeMO Ontology of Digital Humanities Methods. In *Digital Humanities 2016: Conference Abstracts*. 161–163. http://dh2016.adho.org/abstracts/134

Gonzalo de Cadenas-Santiago, Alicia García Herrero, Álvaro Ortiz Vidal-Abarca, and Tomasa Rodrigo López. 2015. An Empirical Assessment of Social Unrest Dynamics and State Response in Eurasian Countries. (June 2015). Working Paper.

Stephanie Dornschneider. 2016. *Whether to Kill: The Cognitive Maps of Violent and Nonviolent Individuals*. University of Pennsylvania Press.

Stephanie Dornschneider and Nick Henderson. 2016. A Computational Model of Cognitive Maps. *Journal of Conflict Resolution* 60, 2 (2016), 368–399.

Tomislav Dulić. 2011. Peace Research and Source Criticism: Using historical methodology to improve information gathering and analysis. In *Understanding Peace Research: Methods and challenges*, Kristine Höglund and Magnus Öberg (Eds.). Routledge, Chapter 3, 35–46.

Oskari Eronen. 2016. Organising Artisans for Peace: CMI on a Learning Curve. In *Complexity Thinking for Peacebuilding Practice and Evaluation*, Emery Brusset, Cedric de Coning, and Bryn Hughes (Eds.). Palgrave Macmillan, Chapter 6, 141–176.

Event Data Project. 2012. CAMEO Conflict and Mediation Event Observations Event and Actor Codebook. (March 2012). Codebook.

Julia Flanders and Fotis Jannidis. 2015. Knowledge Organization and Data Modeling in the Humanities. https://datasymposium.wordpress.com/

Ingo Frank. 2017. Interdisciplinary Knowledge Organization as Intersection between Information Science and Digital Humanities. In *ISI 2017 Satellite Workshop on the Relationship of Information Science and the Digital Humanities*, Manuel Burghardt and Markus Kattenbeck (Eds.). 25–33. Position Paper.

Claudio Gnoli. 2008. Categories and Facets in Integrative Levels. *Axiomathes* 18, 2 (2008), 177–192.

Stephen Gray and Josefine Roos. 2012. Pride, conflict and complexity: Applying dynamical systems theory to understand local conflict in South Sudan. In *International Association for Conflict Management (IACM) 25th Annual Conference*.

Justin Grimmer and Brandon M. Stewart. 2013. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis* 21, 3 (2013), 267–297.

Nicolai Hartmann. 2010. *Der Aufbau der realen Welt: Grundriß der allgemeinen Kategorienlehre*. De Gruyter.

Peter Hedström. 2006. Explaining Social Change: An Analytical Approach. *Papers: Revista de Sociologia* 80 (2006), 73–95.

Peter Hedström and Richard Swedberg. 1998. Social mechanisms: An introductory essay. In *Social Mechanisms: An Analytical Approach to Social Theory*, Peter Hedström and Richard Swedberg (Eds.). Cambridge University Press, Chapter 1, 1–31.

Heinrich Herre. 2013. Formal Ontology and the Foundation of Knowledge Organization. *Knowledge Organization* 40, 5 (2013), 332–339.

Birger Hjørland. 2015. Theories are Knowledge Organizing Systems (KOS). *Knowledge Organization* 42, 6 (2015), 113–128.

Birger Hjørland. 2016. Knowledge Organization (KO). *Knowledge Organization* 43, 6 (2016), 475–484.

Michael H. G. Hoffmann. 2006. Seeing problems, seeing solutions. Abduction and diagrammatic reasoning in a theory of scientific discovery. (2006).

Zhengyin Hu, Shu Fang, and Tian Liang. 2014. Empirical study of constructing a knowledge organization system of patent documents using topic modeling. *Scientometrics* 100, 3 (Sept. 2014), 787–799.

Lorna M. Hughes, Panos Constantopoulos, and Costis Dallas. 2015. Digital Methods in the Humanities: Understanding and Describing their Use across the Disciplines. In *The New Companion to Digital Humanities*, Susan Schreibman and Ray Siemens (Eds.). Blackwell.

Edmund Husserl. 1952. *Ideen zu einer reinen Phänomenologie und phänomenologischen Philosophie. Drittes Buch: Die Phänomenologie und die Fundamente der Wissenschaften*. Husserliana, Vol. V. Martinus Nijhoff, Den Haag.

Machiel Keestra. 2011. Understanding Human Action: Integrating Meanings, Mechanisms, Causes, and Contexts. In *Interdisciplinary Research: Case Studies of Integrative Understandings of Complex Problems*, Repko Allen, Szostak Rick, and Newell William (Eds.). Sage Publications.

Yaser Keneshloo, Jose Cadena, Gizem Korkmaz, and Naren Ramakrishnan. 2014. Detecting and Forecasting Domestic Political Crises: A Graph-based Approach. In *Proceedings of the 2014 ACM Conference on Web Science (WebSci '14)*. ACM, New York, NY, USA, 192–196.

Michael Kleineberg. 2017. Integrative Levels. *Knowledge Organization* 44 (2017), 349–379.

Bart Kosko. 1986. Fuzzy Cognitive Maps. *International Journal of Man-Machine Studies* 24, 1 (Jan. 1986), 65–75.

Bart Kosko. 2014. Foreword. In *Fuzzy Cognitive Maps for Applied Sciences and Engineering: From Fundamentals to Extensions and Learning Algorithms*, Elpiniki I. Papageorgiou (Ed.). Vol. 54. Springer, vii–ix.

Claire Lemercier. 2015a. Formal network methods in history: why and how? In *Social Networks, Political Institutions, and Rural Societies*. Brepols, 281–310. https://halshs.archives-ouvertes.fr/halshs-00521527

Claire Lemercier. 2015b. Taking time seriously. In *Knoten und Kanten III*, Marten Düring, Markus Gamper, and Linda Reschke (Eds.). Transcript Verlag, 183–211. https://hal.archives-ouvertes.fr/hal-01445932

Philipp Mayr, Douglas Tudhope, Stella Dextre Clarke, Marcia Lei Zeng, and Xia Lin. 2016. Recent applications of Knowledge Organization Systems: introduction to a special issue. *International Journal on Digital Libraries* 17, 1 (March 2016), 1–4.

Peter Morville and Louis Rosenfeld. 2002. *Information Architecture for the World Wide Web* (2nd edition ed.). O'Reilly Media.

Benn Neil and Macintosh Ann. 2012. Making Sense of Macro- and Micro-Argumentation in Policy-Deliberation: Visualisation Techniques and Representation Formats. *Frontiers in Artificial Intelligence and Applications* (2012), 71–82.

Osonde A. Osoba and Bart Kosko. 2017. Fuzzy cognitive maps of public support for insurgency and terrorism. *The Journal of Defense Modeling and Simulation* 14, 1 (2017), 17–32.

Elpiniki I. Papageorgiou, Jos De Roo, Csaba Huszka, and Dirk Colaert. 2012. Formalization of treatment guidelines using Fuzzy Cognitive Maps and semantic web tools. *Journal of Biomedical Informatics* 45, 1 (2012), 45–60.

Elpiniki I. Papageorgiou and Jose L. Salmeron. 2013. A Review of Fuzzy Cognitive Maps Research During the Last Decade. *Transactions on Fuzzy Systems* 21, 1 (Feb. 2013), 66–79.

Michael Piotrowski. 2015. Document Engineering and Digital Humanities. In *NLP for Historical Texts: Computational linguistics and digital humanities*. http://nlphist.hypotheses.org/263 Blog.

Michael Piotrowski. 2016. Digital Humanities, Computational Linguistics, and Natural Language Processing. (March 2016). http://stp.lingfil.uu.se/~nivre/docs/michael_piotrowski_2016.pdf

Roberto Poli. 2001. The Basic Problem of the Theory of Levels of Reality. *Axiomathes* 12, 3 (Sept. 2001), 261–283.

Roberto Poli. 2007. Three obstructions: forms of causation, chronotopoids, and levels of reality. *Axiomathes* 17, 1 (March 2007), 1–18.

Margaret E. Roberts, Brandon M. Stewart, Dustin Tingley, and Edoardo M. Airoldi. 2013. The Structural Topic Model and Applied Social Science. *Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation*.

Philip A. Schrodt. 2001. Automated coding of international event data using sparse parsing techniques. In *Annual Meeting of the International Studies Association*.

Sameera Al Shayji, Nahla El Zant El Kadhi, and Zidong Wang. 2011. Fuzzy Cognitive Map Theory for the Political Domain. In *2011 Federated Conference on Computer Science and Information Systems (FedCSIS)*. 179–186.

Ben Shneiderman. 1996. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In *Proceedings of the 1996 IEEE Symposium on Visual Languages (Visual Languages)*. IEEE Computer Society, 336–343.

Simon Buckingham Shum and Alexandra Okada. 2008. Knowledge Cartography for Controversies: The Iraq Debate. In *Knowledge Cartography*, Alexandra Okada, Simon Buckingham Shum, and Tony Sherborne (Eds.). Springer London, London, 249–265.

Dagobert Soergel. 2015. Unleashing the Power of Data Through Organization: Structure and Connections for Meaning, Learning, and Discovery. *Knowledge Organization* 42, 6 (2015), 401–427.

Rick Szostak. 2004. *Classifying Science: Phenomena, Data, Theory, Method, Practice*. Information Science and Knowledge Management, Vol. 7. Springer Netherlands.

Rick Szostak, Claudio Gnoli, and María López-Huertas. 2016. Phenomenon Versus Discipline-Based Classification. In *Interdisciplinary Knowledge Organization*. Springer International Publishing, Cham, 93–110.

Tolga Tezcan. 2014. The Complex Nature of Migration at a Conceptual Level: An Overlook of the Internal Migration Experience of Gebze Through Fuzzy Cognitive Mapping Method. In *Fuzzy Cognitive Maps for Applied Sciences and Engineering:*

*From Fundamentals to Extensions and Learning Algorithms*, Elpiniki I. Papageorgiou (Ed.). Springer, Berlin, Heidelberg, 319–354.

Athanasios Tsadiras and Nick Bassiliades. 2013. RuleML representation and simulation of Fuzzy Cognitive Maps. *Expert Systems with Applications* 40, 5 (2013), 1413–1426.

John Unsworth. 2002. What is Humanities Computing and What is Not? (2002). http://people.virginia.edu/~jmu2m/mith.00.html

Hayden White. 1978. *Tropics of Discourse: Essays in Cultural Criticism.* Johns Hopkins University Press.

Margaret Tucker Wrightson. 1976. APPENDIX ONE: The Documentary Coding Method. In *Structure of Decision: The Cognitive Maps of Political Elites.* Princeton University Press, 291–332.

# Density of Knowledge Organization Systems

Linda Freyberg
Doctoral Research Group Knowledge Cultures/Digital Media
Leuphana University Lüneburg and
Urban Complexity Lab
University of Applied Sciences Potsdam
linda.freyberg@gmx.de

## ABSTRACT

In this paper the category of density is introduced as a parameter for the measurement of the level of semantic contextualization of Knowledge Organization Systems. Based on previous research regarding the detailedness of description on a metadata level the concept of network density and iconicity of a system as an indicator for the accessibility and usability will be considered. Hence this approach takes the iconic dimension of Knowledge Organization Systems (KOS) in addition to their sentential foundation into account. In particular the metadata structure of the cultural heritage database Europeana and its diagrammatic dimension serves as an example. The theoretical foundation of this research is provided by Charles Sanders Peirce's concepts of signs and his concept of iconicity.

## KEYWORDS

Semiotics, Peirce, Density, Knowledge Organization Systems (KOS), Iconicity, Cultural Heritage Databases, Europeana

**Reference:**
Linda Freyberg. 2018. Density of Knowledge Organization Systems. In *Proceedings of Wissensorganisation 2017: 15. Tagung der Deutschen Sektion der Internationalen Gesellschaft für Wissensorganisation (ISKO) (WissOrg'17).* German Chapter of the ISKO / Freie Universität Berlin , pp. 25-30. https://doi.org/10.17169/FUDOCS_document_000000028863

## 1 INTRODUCTION

Previous research in information sciences concerning the theoretical foundation of knowledge organization has already dealt with the semiotic foundation in general (see Thellefsen/Thellefsen 2004; Friedman/Thellefsen 2011; Friedman/Smiraglia 2013), but a close analysis of the diagrammatic dimension of KOS is a gap, this research might fill. These approaches have mainly focused on the symbolical sign aspect taking spoken or written language and therefore the sentential aspect into account. While almost every knowledge organization system (KOS) is based on scripture (written language), this research focuses on the characteristics and the potential of iconicity in knowledge organization systems. The term knowledge organization systems includes not only the controlled vocabularies or structured languages; it also addresses the databases it selves where (digital) objects and their representations are contextualized. Hence the use of the term KOS in this paper includes those storages or databases for instance cultural heritage databases as a whole. This paper in particular deals with the following aspects of current knowledge organization systems: The density of the semantic description and the iconicity of KOS on an implicit or structural level as a crucial property of density. In traditional KO classifications or thesauri serve as parameters for the description of the (physical) objects. In a digital environment there are also standards such as ontology languages or metadata standards to express the relations of objects. Previous research regarding the quality of semantic description in KOS takes mainly technological standards for instance metadata standards into account. But the use of a description language does not necessarily show the actual level of contextualization of the system, because it is possible to use very few attributes of a potentially very detailed expression language and it is also possible to not link your data to other objects of your data or to other data.

In general this approach refers to the metadata of a system, hence to an information specialist's or system developer's perspective. But it also deals with the question whether those attributes and relations are accessible or searchable? So here the user's perspective comes into sight. Usually there are all fields searchable in the research process, but not so many actual systems use visualization to show the relations and make them explorable.

In the realm of digital KOS the process of organizing knowledge includes the semantic contextualization of information or objects/artifacts and their digital representations expressed in logical relations. Thus the concepts of information, knowledge and contextualization are crucial for this research. Furthermore the relevance of the concept of iconicity for KOS will be elaborated and the concept of density as a parameter for the level of the semantic description in KOS is proposed in this paper.

## 2 INFORMATION

What concept of information is in particular operating in digital environments? What theoretical approaches are effective or fruitful for this area of appliance? In a quite current article on "The Concept of Information in Library and Information Science" (Thellefsen/Thellefsen/Sørensen) from 2015 Shannon's "Mathematical Theory of Communication" from 1948 is referred to as kind of a starting point in terms of information theory. And indeed it still has validity in particular for realms like data processing. Shannon introduced a measurement of information ("binary digits, or more briefly bits", Shannon (1948, p.1)) that depends on probabilities by referring to the decision of which signs of the code are chosen out


*WissOrg'17, December 2017, Berlin, Germany*

https://doi.org/10.17169/FUDOCS_document_000000028863

of the set of all possibilities. Because this model was developed in the context of machine communication, it can be transferred to digital environments where at least the amount of transmitted information of the computer is measurable and where the semantic aspect of information is not the relevant property. But since humans are a significant part in this process of communication the semantic dimension, which could not be easily expressed in bits, is a crucial aspect of information processes. In a semiotic view all communication is expressed in signs. Referring to the universal theory of semeiotic of Charles Sanders Peirce even thoughts and all phanomena are signs (CP.5484). In the French-speaking tradition of semiotics the principle of difference is crucial, for instance for Saussure, Barthes and Derrida. Signs become meaningful in difference to other signs, therefore the concept of contextualization is crucial for information processes. The philosopher Floridi underlines the fact that information is necessarily related to a certain context and depending on this context, which can be defined as "interpretation, power, narrative, message or medium, conversation, construction, acommodity" (Thellefsen et al., 2015, p.59).

The "Data-Information-Knowledge-Wisdom"-model of Ackoff (Ackoff 1989, 3-9) is also based on the level of contextualization. Information therefore is contextualized data and contextualized information is the definition of knowledge. The contextualization refers to the level of organization as Rowley points out: "[I]nformation is defined in terms of data, and is seen to be organized or structured data. This processing lends the data relevance for a specific purpose or context, and thereby makes it meaningful, valuable, useful and relevant."(Rowley, 2007, p.172) Regarding in particular digital environments a even more precise approach is the concept of "unaggregated data", which "have no meaning in themselves". (Rowley, 2007, p.172) This means that unaggregated data is defined as data out of context. In general information is processed by the recipient or user and categorized in his actual state of knowledge. This integration of (new) information in the given knowledge context generates new knowledge, which relates to Bateson's concept of information as a „difference that makes a difference". (Bateson, 1972, p.276) This view is still related to Shannon, but „makes a difference" refers to the pragmatic aspect, that information has to be contextualized to be understood and that an information has to be new to the receiver. In conclusion the meaning of a sign is defined by the difference to other signs, thus contextualization is the main condition of information and the key concept for understanding information and for the formation of knowledge.

The concept of context is considered as interpretive and not as pre-given. Kleineberg (2013) identified pre-given and interpretative as the main categories of context and gave a more detailed description of the contradictive nature of the term contextualization itself. In the realm of Knowledge Organization objects or their digital representations are described and contextualized with other objects and abstract concepts. Hence the core of Knowledge Organization Systems is the semantic description of the objects and the specification of their relations.

## 3 ICONICITY

Peirce's enlargement of sign theory and his various distinctions in three sign categories in particular in his the popular triadic model: Icon, index and symbol directly addresses the "iconic" aspect of signs. Based on the process-oriented definition of information, in which information is involved in semiosis and meaning is gained through contextualization, equivalently the visual expression of information is based on those principles. In particular in a digital environment a diagram could be defined as "a proliferation of manifestly selective packets of dissimilar data correlated in an explicitly process-orientated array that has some of the attributes of a representation, but is situated in the world like an object." (Bender and Marrinan, 2010, p.7). This view could be transferred to the status of objects and their representations in a cultural heritage database where the object itself and its digital representation could be addressed by unified identifiers and therefore both are situated in the environments like objects.

A icon sign shares the properties of its object "by virtue of its being an immediate image, that is to say by virtue of characters which belong to it" (CP 4.447 (Peirce, 1935)). But an icon does not necessarily resembles its object: "Many diagrams resemble their objects not at all in looks", but a similarity "in respect to the relations of their parts [constitutes] that their likeness consists." (CP 2.282 Peirce (1935)) In Frederik Stjernfelt's assumption "similarity is the very source of new ideas", because "some sort of iconic relation to the fact to be explained" (Stjernfelt, 2007, p.77) ) is required. Through the contextualization of all objects new relations can be explored and new information could be discovered. Instead of only resembling the prosperities of the represented object the function of generating new information in the reasoning process is essential for the concept of iconicity as already noted by Peirce: "For a great distinguishing property of the icon is that by the direct observation of it other truths concerning its object can be discovered than those which suffice to determine its construction." (CP, 2.279 Peirce (1935) ) So the concept of iconicity here is operational or functional. This means that in the process of reading or understanding the diagram not only the addition of the properties of all displayed objects are taken into account. This principle is the foundation of visual expressions, for instance digital data visualizations. Visual expressions are therefore even able to "add" meaning and enable the discovery of new information. Or how Smiraglia put it: "Visualization also means synergistically to expand perception by adding understanding beyond that of the textual narrative or data." (Smiraglia, 2015, p.42). The object and their digital representations in KOS build networks therefore they have an iconic dimension at least on a metadata level. Objects, subjects and predicates are organized for instance in RDF (Resource Description Framework) in a triple structure, while the objects can be the subjects of another triple, hence those structures are able to grow at every entity like an infinite semiosis. The iconicity is implicit and does not have a specific functionality in terms of underlining arguments or visualize certain properties of objects. But when you look at these structures objects and their contexts become visible although those structures might not be built to serve a specific purpose, they are more the result of data modeling and topological organization of the objects.

## 4 DENSITY

But how detailed is the contextualization in a certain system? Or referring to the ethnographical method of Geertz (see Geertz, 1973) in

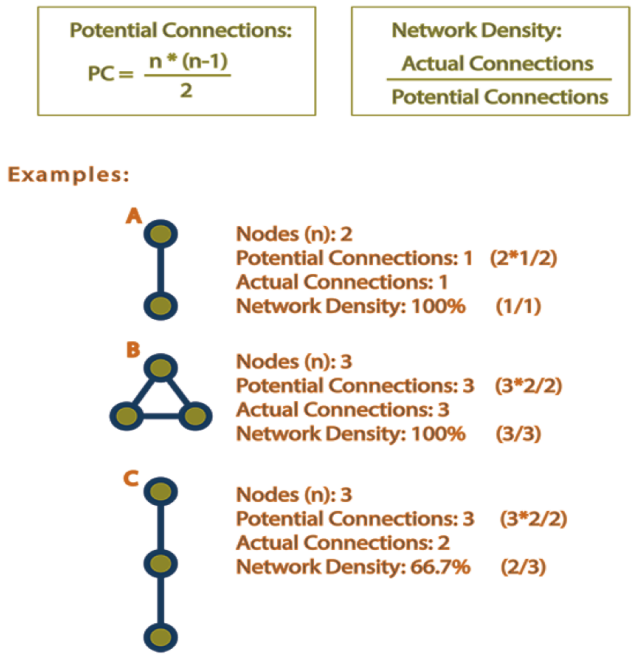



**Figure 1**
**http://www.the-vital-edge.com/what-is-network-density/**

the context of cultural anthropology how thick are the descriptions? And is it possible to measure the "thickness" of those descriptions? If so: What could be parameters for that quality? In following this quality is called density and is proposed as a category for the level of semantic contextualization. This approach can be understood as a reminiscence of Shannon's attempt to propose a measurement of information. Since density is already an established term in network analysis in the following some aspects of network analysis are transferred to the realm of KOS. The density of a network is a measurement of the proportion of potential and actual relations in networks. This concept is related to the mathematical term of a dense graph, in which the number of edges is close to the maximal number of edges. Networks consist of nodes and relations (or edges) and there exists a potential and an actual relation of each entity. The division of those two factors results in the density of a network, which is illustrated in a simple way in Fig. 1. Hence a completely linked network has a factor of 1.

Certainly in existing systems like Knowledge Organization Systems this is mostly not the case, because not all objects are linked to all other objects. Furthermore it is very questionable if the factor of 1 should actually be the indicator of the highest quality of a system. With all objects linked to each other, a single object has no significance in comparison to other objects hence has no significance at all.

In digital social networks, which are a common example of current network analysis research (see Otte/Rousseau 2002), the mere number of the relations for instance of Facebookfriends or Twitterfollowers is not significant for the quality or the type of the specific

**Table 1**
**SPARQL-Query (http://sparql.europeana.eu/), 04/07/17.**

| EDM-type | Number |
|---|---|
| IMAGE | 57.001.905 |
| SOUND | 1.048.582 |
| VIDEO | 1.713.426 |
| TEXT | 44.755.275 |
| 3D[1] | - |

relation. Also the factor of "preferential connectivity" mentioned by Barabasi and Albert (1999) is crucial. Preferential connectivity means that already strongly linked objects will gain more new connections than not that well linked objects. Also new relations will be most probably built to already close objects or referring to the small-world model of Watts and Strogatz of „each vertex being connected to its two nearest and next-nearest neighbors." (Barabasi and Albert, 1999, p.510)

In our example of knowledge organization systems for instance cultural heritage databases this means, that the high number of already existing relations, which represent the high influence on other objects, expresses the significance of an artwork. And in addition to that this object will most probably gain more connections in the future. Another example for these aspects and the attempt to measure quality in general is the analysis of citations. In particular the recent discussion regarding the Impact Factor of Journals, which is a quite common tool in terms of the scientific communication and is even used in the process of hiring in academia reflects the problems those strictly quantitative attempts may cause. In the article "Beat it, impact factor! Publishing elite turns against controversial metric" from 2016 Callaway points out the "inappropriate use" of this measure system (Callaway, 2016, p.210). Even though this article itself is published in the major journal "nature" he elaborates that the impact factor of the journal is not necessarily a indicator of quality of every single article of this journal, because the high impact factor of the journal in general is mostly caused by ONE very highly cited article of an issue. So referring to the introduced terminology of network analysis this article has a preferential connectivity and it could be expected that this specific article will be gain further citations. But the actual impact in terms of meaning is hard to measure, which demonstrates that those metrics and attempts to measure a quality should be contextualized in the meaning of reflected themselves.

## 4.1 Density in Europeana

In general the concept of density could be in particular adapted to the level of semantic description in KOS. Semantic networks are an established expression of relations, which could be for instance expressed in semantic web ontologies. Europeana , which by now includes over 54 million (On 02/15/2017) objects ("artworks, artefacts, books, videos and sounds") from over 6,000 providing institutions, offers a high variety of objects:

[1]The count of 3D-objects was not feasible. Despite of the fact, that the database contains 3D-objects, the query always delivered the value 0.

In terms of network density, if all images in Europeana would be linked to each other, the density of that network would be 1: $\frac{57.001.905}{57.001.905} = 1$. Most certainly this is not the fact, because not all images are linked to another, which also as pointed out earlier referring to their significance would make no sense. As well the possibilities of the semantic description of data in Europeana with their own Europeana Data Model (EDM) (See Europeana) are very detailed and include diverse formats like (DC, SKOS, RDF etc.).

The simplest specification of an artwork or text is the creatorship: "is creator/author of". But for showing this simple relation there is no graph structure or complex network needed. The relations regarding the content of an artwork or text document are the relevant information in particular in terms of understanding the meaning of the object and its significance, which is a direct result of its context. In order to reach a dense semantic description level the connections has to be specified and described in detail. So the number of attributes to describe the relation is relevant here.

## 4.2 Levels of Density

Now we will have a closer look at how to define the levels of density in a knowledge organization system. There have already been attempts to describe or to measure the completeness of records in cultural heritage databases. The "Metadata Quality Assurance Framework"-project of Király (see Király (2015a)), which was conducted in close cooperation with the Europeana Network's Data Quality Committee (see Europeana 3) can provide for a basis. "The starting hypothesis is that it is possible to measure some factors of the data quality with computational tools" For instance the „completeness" of the records as data quality feature which refers to whether the fields are filled and unfilled and if there are mandatory fields. If mandatory fields are empty he claims that the quality goes down. In his metadata quality assurance framework he proposed seven quality metrics: Completeness, accuracy, conformance to expectations, logical consistency and coherence, accessibility, timeliness and provenance.

While this approach focuses on the measurement of the data quality with computational tools (see Király (2015b, p.2)) in terms of the density of the semantic description the following aspects are suggested in addition:

(1) Object Detailedness: How detailed is the description of each object?
(2) Relation Detailedness: How are the relations specified, with how many attributes are the relations described?
(3) Embeddedness: To how many other objects is the object related? What kinds of objects are related to each other?
(4) Iconicity: How are the objects ordered? How are the relations expressed?

In respect to measure the quality of a specific knowledge organization system the following aspects has to be analyzed: First, the types of objects (whether its is a text, an image, a location or a concept) are crucial and then the description of relations and the expression of the relations (description language) should be investigated.

## 4.3 Density as Visual Quality

In digital knowledge organization systems the relations can be formally expressed in semantic web ontologies. For instance with RDF semantic descriptions in graph structures are build hence those structures have a highly iconic dimension on a metadata level. Every entity could be involved in other triples, so it is possible to build infinite networks of predicates like the infinite semiosis mentioned earlier in this paper. Referring to the aforementioned aspects regarding iconicity, the conclusion is, that topological or morphological categorization leads to a higher level of contextualization. But those triples in the metadata you cannot actually see.

Regarding density this concept itself has a visual quality on those levels hence is actually visible. In a database a very detailed description of an object causes a very dense visual expression, because in every field appears a lot of text therefore those database fields seem crowded and so forth.

And by visualizing the objects of a knowledge organization system and their relations for instance in a network an overlap of arcs also causes a very dense visual impression.

This is a visualization of the "Linked Open Data Cloud" (LoD Cloud) of 2017. Diverse dimensions of expressions like colors, sizes, spatial positions and the links are used in this visualization.

Online it can be used dynamically and it divers between incoming and outgoing links. In this diagram all projects, organizations and individuals who publish their data in the Linked Data format are presented. Huge contributors are DBpedia in the hub in a light brown, (national) libraries, universities, museums or cultural heritage databases represented by the size of the bubbles. Some areas in this visualization are denser: The life sciences expressed in red bubbles in the right part seem to be a very active publisher of linked open data and the incoming and outgoing links nearly create a dark grey background where it is not possible anymore to see every single link. This seems like a very high level of density in terms of its visual impression. Other areas on the edges for example the yellow bubbles in the right upper part representing government data does not show any relation at all. Hence the density of these parts is very low or even not there, because some objects lacks of contextualization.

The LoD-Cloud as an unrestrained dynamic network grows every minute and is able to grow at every entity. In general the expression of diverse relations in order to provide for multiple readings is offered here.

# 5 CONCLUSION

As a parameter to measure the level of contextualization in a knowledge organization system, in this paper the term density was introduced. At first the meaning of the term in his original context of network analysis was elaborated and it was analyzed if this concept could be transferred to the realm of digital KO. It has been shown that knowledge organization databases for instance cultural heritage databases are organized in network structures operating with semantic web ontologies. Also the principles of preferential connectivity and the Small World theory could be borrowed from network analysis and transferred to KOS in order to point out how networks like Europeana grow.

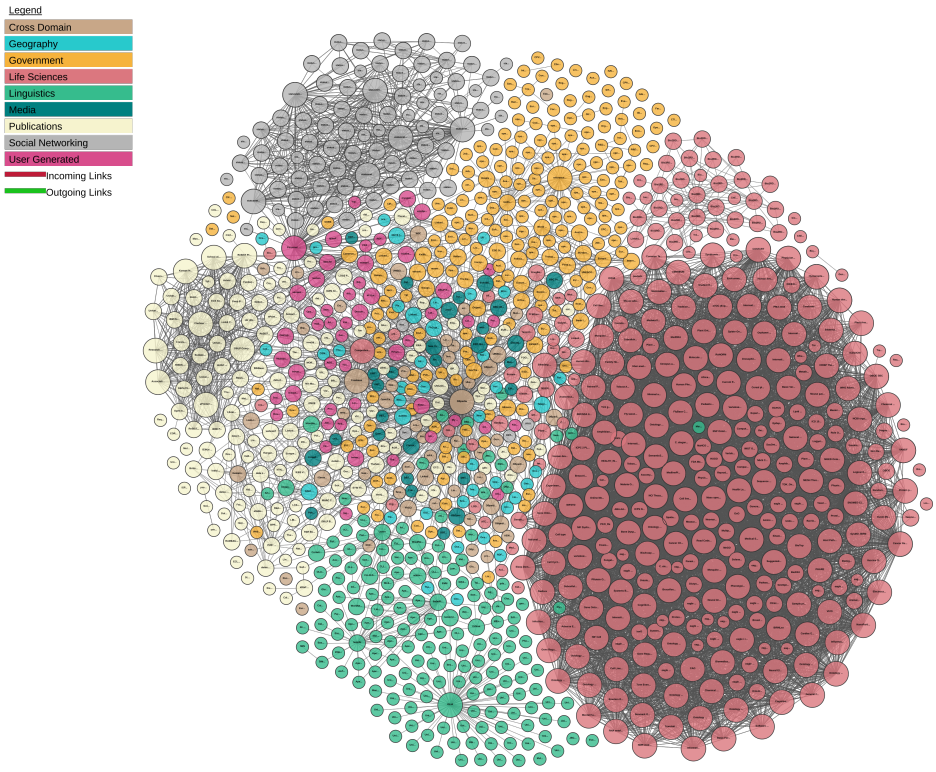



**Figure 2**
**Linking Open Data cloud diagram 2017**

In the theoretical framework mainly based on Peirce's semiotics and his elaborated sign categories the concept of contextualization was identified as the main property of information. In a digital environment the aggregation of data and the specification of its relation leads potentially to information or knowledge. But in this approach contextualization not only includes data processing on a machine level but also involves the semantic and pragmatic dimension for instance the state of knowledge of the interpretant.

Based on the hypothesis that every knowledge organization system has an iconic or diagrammatic dimension, which can be implicit on a structural or topological level or can be explicitly expressed for instance in a graphic data explorer or a graphic interface the relevance of the concept of iconicity was pointed out. In particular Peirce's concept of iconicity, which refers to the idea that visual expressions could be the source of new ideas was added to the concept of density in this paper.

Beside technical aspects in terms of metadata quality, which has been elaborated, the iconicity on a structural level of those systems has been introduced as a tool to discover new relations and therefore to gain new information and knowledge.

To illustrate the specified and detailed description of relations, Europeana and its Data Model (EDM) as well as the Linked Open Data Cloud has served as examples. By visualizing iconic, previously hidden structures between nodes new approaches to data sets could be fostered. In conclusion to specify the relations and make them visible is a huge challenge of digital knowledge organization systems. The relations of a KOS seem to have a complex diagrammatic structure. In particular the elaborated structures on a metadata level were addressed, which already have an iconic character and should be accessible and explorable in order to discover new semantic contexts and in the end to gain knowledge.

The elaborated theoretical framework could enable this process and provide for a theoretical foundation for the understanding and modeling of knowledge organization systems. This approach also provides for parameters to measure the density of specific knowledge organization systems.

In particular for cultural heritage databases like Europeana, where a huge variety of objects is included the visualization of the already existing iconic structures on a metadata level could be very fruitful and offer an different approach to the whole dataset.

This short paper is part of an ongoing PhD-project with the title "Iconicity in Information"

## REFERENCES

Albert-Laszlo Barabasi and Reka Albert. 1999. Emergence of Scaling in Random Networks. *Science* 286, 15 (1999), 509–512.

Gregory Bateson. 1972. Form, Substance, and Difference. *Steps to an Ecology of Mind* (1972), 448–466.

John Bender and Michael Marrinan. 2010. *The Culture of the Diagram.* Stanford University Press, Stanford.

Ewen Callaway. 2016. Beat it, impact factor! Publishing elite turns against controversial metric. *nature* 535 (2016), 210–211. Issue 7611.

Clifford Geertz. 1973. *Thick Description: Toward an Interpretive Theory of Culture. In: The Interpretation of Cultures: Selected Essays.* Basic Books, New York.

Péter Király. 2015a. A Metadata Quality Assurance Framework. (2015). http://144.76.218.178/europeana-qa/ Online; accessed 27-December-2017.

Péter Király. 2015b. Metadata Quality Project Plan. (2015). https://pkiraly.github.io/metadata-quality-project-plan.pdf Online; accessed 27-December-2017.

Michael Kleineberg. 2013. The blind men and the elephant: towards an organization of epistemic contexts. *Knowledge organization* 40 (2013), 340–362. Issue 5.

Charles Sanders Peirce. 1931-1935. *The Collected Papers of Charles Sanders Peirce.* Harvard University Press, Cambridge, MA.

Jennifer Rowley. 2007. The wisdom hierarchy. Representations of the DIKW hierarchy. *Journal of Information Science* 33 (2007), 163–180. Issue 2.

Claude Shannon. 1948. A Mathematical Theory of Communication. *The Bell System Technical Journal* 27 (1948), 379–423, 623–656. Reprinted with corrections from.

Richard P. Smiraglia. 2015. *Domain Analysis for Knowledge Organization: Tools for Ontology Extraction.* Chandos Publishing, Witney, Oxfordshire.

Frederik Stjernfelt. 2007. *Diagrammatology. An Investigation on the Borderlines of Phenomenology, Ontology, and Semiotics.* Springer Netherlands, Dordrecht.

Torkild Thellefsen, Martin Thellefsen, and Bent Sørensen (Eds.). 2015. PThe Concept of Information in Library and Information Science. A Field in Search of Its Boundaries: 9 Short Comments Concerning Information. *Cybernetics and Human Knowing* 22 (2015), 177–187. Issue 1. Comments of Luciano Floridi, Søren Brier, Torkild Thellefsen, Martin Thellefsen, Bent Sørensen, Birger Hjørland, Brenda Dervin, Volkmar Engerer, Ken Herold, Per Hasle and Michael Buckland.

# Accessing, Editing and Indexing Large Manuscript Collections

## The Selected Edition of J. Chr. Senckenberg's Journals

Vera Faßhauer
Goethe University of Frankfurt
Institut für Deutsche Literatur und ihre Didaktik
D 60323 Frankfurt am Main
fasshauer@em.uni-frankfurt.de

## ABSTRACT

Increasing capacities of data storage allow libraries and archives to digitize their manuscript collections and present them online under open access. Yet only few experts are nowadays able to decipher and analyse historical handwriting. In order to grant their users access to the contents, collection-holding institutions need the assistance of scholarly experts. Scholars, on the other hand, depend on the support of the libraries providing digital copies and authority data as well as publication platforms and long-term data storage and maintenance. By example of the compendious and hard-to-decipher Senckenberg journals which served their author as a medium of empirical data collection, the paper will address some major issues concerning the edition and deep indexing of large manuscript collections, present an experiment with the new Handwritten Text Recognition (HTR) technology and consider some basic aspects of a possible cooperation between manuscript-holding institutions and academic specialists.

## KEYWORDS

Historical manuscript collections, Digital edition, Keyword assignment, Cooperation between scholars and libraries, Handwritten Text Recognition, Authority data, Digital copies of historical prints

**Reference:**
Vera Faßhauer. 2018. Accessing, Editing and Indexing Large Manuscript Collections: The Selected Edition of J. Chr. Senckenberg's Journals. In *Proceedings of Wissensorganisation 2017: 15. Tagung der Deutschen Sektion der Internationalen Gesellschaft für Wissensorganisation (ISKO) (WissOrg'17).* German Chapter of the ISKO / Freie Universität Berlin , pp. 31-36. https://doi.org/10.17169/FUDOCS_document_000000028863

## 1 INTRODUCTION

In view of constantly increasing capacities of data storage, more and more libraries take the opportunity to digitize their collections and make them accessible online. The digital copies are regularly equipped with metadata, but rarely with deeper indexing or description of the contents. This strategy has proved sufficient where printed books are concerned: Given an appropriate scan quality, the readability and accessibility of their contents can generally be taken for granted. Moreover, the current stage of pattern recognition technology allows for fairly good results in machine-based recognition of both roman and gothic types, thus making digital reproductions of printed books accessible for full text search and consequently also for distant reading and statistical approaches. Some libraries, as for instance the University and State Library of Saxony-Anhalt (ULB) in Halle, also capture all chapter and paragraph headings in electronic tables of contents and link them to the respective page image.[1] As these headings usually contain all representative names and keywords, they provide an excellent starting point for further book study. Similarly, the page numbers in indexes can be linked to the pages they refer to, as is often done by Google Books.

Digital facsimiles of historical handwriting, however, are subject to wholly different conditions: The Handwritten Text Recognition (HTR) technology is just under development and depends on large amounts of manually created text transcriptions. Yet few people nowadays are able to decipher historical handwriting at all, since the training of palaeographic skills is rare even in university curricula (Cartelli and Palma, 2009, Kamp, 2009). For the greater part of the public, facsimiles mean nothing more than photographic reproductions of historical artefacts with a certain surface structure, the patterns of which can no longer be decoded. The problem intensifies when it comes to diaries and notebooks, which were originally intended for only the author's eyes. Especially the handwriting of erudite authors is often negligent and hard to read. An unrestricted long-term availability of digital facsimiles is therefore not necessarily identical with unlimited accessibility to their contents.

The most profound method of accessing handwritten documents is their full text edition and explanatory annotation. In the case of very large manuscript holdings, however, a complete text transcription does not always appear practicable, and a selection of characteristic parts or the indexing by a larger number of content-related keywords may then seem more reasonable. In order to enable their users to take advantage of their digital collections in either manner, the holding libraries can call on the assistance of scholarly experts familiar not only with historical palaeography but also with the subject matter treated in the manuscripts. Scholars, on the other hand, depend on the libraries' technical and administrative support in order to facilitate and supplement their editorial work. By the example of the Selected Edition of the Senckenberg Journals, which is currently created at the Department of German Studies at Frankfurt University (Faßhauer, 2017a) and funded by the Senckenberg foundation,[2] this paper will propose a possible way of


*WissOrg'17, December 2017, Berlin, Germany*

https://doi.org/10.17169/FUDOCS_document_000000028863

[1] http://digitale.bibliothek.uni-halle.de/vd18/nav/index/all (accessed 5. December 2017).
[2] https://www.uni-frankfurt.de/43891858/projekte (accessed 20. September 2017); http://www.senckenbergische-stiftung.de/start.html (accessed 5. December 2017).

making large manuscript collections accessible in a mixed-methods approach and present some significant instances of cooperation between academic specialists and manuscript-holding institutions.

## 2 SENCKENBERG'S JOURNALS AS A MEDIUM OF EMPIRICAL DATA COLLECTION

Johann Christian Senckenberg (1707–1772), a Frankfurt physician, collector and founder, left handwritten notes on an exceptionally large scale. As a religious dissenter, he refused church attendance as well as the sacraments and pastoral care. Instead, his diaries served him as a means of spiritual and physical self-examination. In the first 18 volumes dating from the time between 1730 and 1742, Senckenberg was chiefly concerned with himself: He noted down all activities and events of the day, including his visits and conversations, his correspondence and readings as well as his sins and moral transgressions. Moreover, he gave detailed accounts of his expenses, his feeding, his physical exercise, the duration and quality of his sleep, his state of health and his changing moods. He also recorded outer circumstances like weather situation, air pressure and temperature. In doing so, he not only aimed at his own moral and dietetic self-improvement: On a professional level, he was strongly influenced by empiricist ideas. He therefore strove to investigate human nature by means of immediate self-experience and captured all available data as immediately and completely as possible. In entitling his early papers "Observationes in me ipso factae", he ascribed them to a typically empiricist textual tradition (Daston and Lunbeck, 2011, Faßhauer, 2017b). After twelve years of self-study, however, Senckenberg increasingly gave his attention to the physical state and moral behaviour of his contemporaries. In his medical notes he meticulously recorded the case histories of all his patients including their symptoms, their diet and moral conduct, the employed treatments and medicine as well as the stages of their convalescence. At the same time, he critically examined the moral defects of his fellow citizens as well as the political grievances of the imperial city of Frankfurt, especially focusing on the members of the city counsil and the socially leading families (Faßhauer, 2018). His journals can therefore be classified in three subgroups, namely (1.) Self-observation, (2.) Medical recordings (3.) Moral and political reflections.

Senckenberg derived the necessity to study human nature thus diligently and patiently from his profound scepticism towards all rationalist approaches which he regarded as irreverent human arrogance. In his opinion, mortal man was incapable of understanding the causalities of nature, which were only known to the creator himself. The naturalist therefore had to be content with the knowledge God revealed to him and compile his own thesaurus of practical experience. His continually and comprehensively collected body data allowed him to conduct comparative evaluations in order to recognize recurrent patterns in human nature: Certain details could be identified as causes or effects of others, and the comparison of similar or differing elements under varying circumstances provided information on the correlations and divergences of factors. A very similar criticism of abstract theoretical models has only recently prompted Chris Anderson to proclaim the "End of Theory" and to call for a quantification of research: In times of Big Data, schematic simplifications of a much more complex reality were superseded by the possibility of searching massive amounts of empirical data allowing for much more realistic statements about man and nature (Anderson, 2008).

**Table 1: Extensive Early Modern Journals in Comparison**

| Autor | Years | Volumes | Max. Words per Entry | Pages in Total |
|---|---|---|---|---|
| S. Pepys | $8\frac{1}{2}$ | 6 | 1,800 | 3,100 |
| Ph. M. Hahn | 20 | 5 | 1,000 | 3,700 |
| Christian II. | 35 | 23 | 500 | 17,400 |
| Senckenberg | 42 | 53 | 5,000 | 40,000 |

Having already been conceived as a huge big data pool by their author, the journals' contents still appear as big data to the modern reader: Senckenberg's notes cover a period of 43 years altogether and extend to 40,000 pages. In the first 13 years of his journal keeping, Senckenberg filled ca. 13,000 quarto pages with 800 to 1,000 words each, reserving the margins for additional annotations and paragraph-related key phrases (1). The entries are dated consecutively and organized according to the sequence of events. At times, he would have written down up to 5000 words per day. In 1733 alone he filled 2,600 pages with approximately two million words. In 1743, he changed his writing habits using narrow strips of paper and filing them in a card index box. The card box system which held about 27,000 sheets was dissolved in the 19th century when the medical and non-medical contents were seperated. At the same time, all the booklets and loose sheets were sorted chronologically and bound into 53 quarto volumes, each of which contains 700 pages on average.

Compared to other extensive early modern diaries, Senckenberg's are unique regarding both their size and the period they cover. Samuel Pepys (1633–1703) for example only filled 3,100 pages within eight and a half years, his daily entries containing between 12 and 1,800 words at the most (Latham and Matthews, 1970, xli). Similarly, the pietist clergyman and engineer Philipp Matthäus Hahn (1739–1790) left five duodecimo volumes; his entries hardly ever exceeded 300 words and amounted to 1,000 words only in exceptional instances. His journals thus came to 3,700 pages in a period of 20 years (Brecht and Paulus, 1979, 36). Not even the journals of Prince Christian II. of Anhalt-Bernburg (1599–1656), which are now being digitally edited in the Herzog August Library of Wolfenbüttel, can compare to Senckenberg's: In 53 years, he filled 17,400 pages, each of which, however, only contained 500 words at the most (Odier et al., 2013). As the perhaps most extensive ego-document of the era, the Senckenberg journals represent an extraordinarily rich source for several disciplines of early modern studies – not only for the histories of church, medicine, science, economy and ideas, but also for cultural, environmental, urban and everyday history as well as literary studies and library science. Moreover, it can provide fresh insights into the history of big data collections and their relation to theoretical approaches.

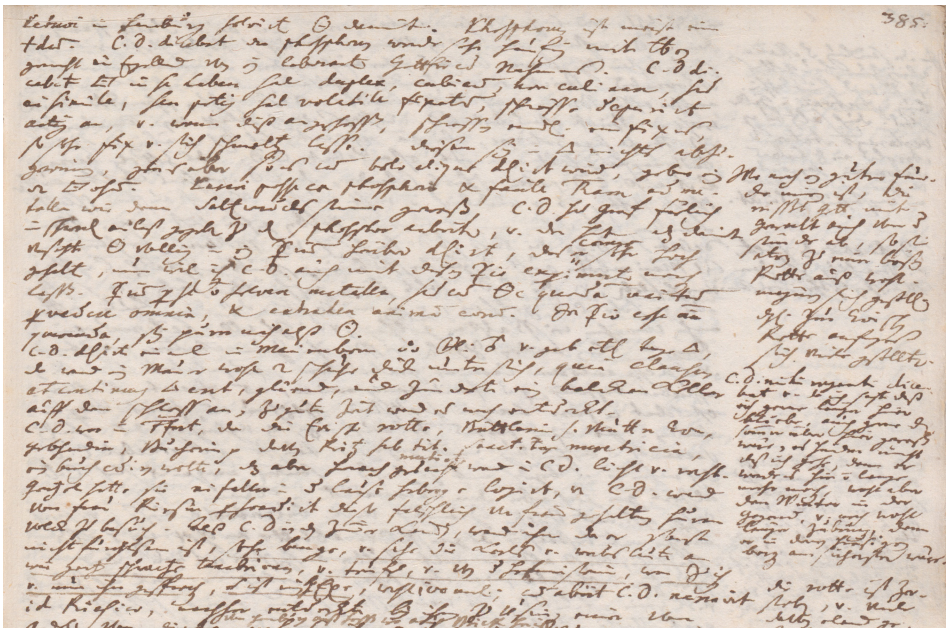


**Figure 1: Page organization and symbol use in Senckenberg's journal of 1732**

## 3 EDITING A HISTORICAL BIG DATA POOL: VOLUME SELECTION AND RECONSTRUCTION OF TEXTUAL MEANING

The Frankfurt University Library (UB) as the holder of the journals has digitized the entire material in high resolution.[3] All volumes are available online under open access-conditions, accompanied by metadata. The headwords, however, refer only to author, work, place and year, but not to the journal's nature and contents. Formerly envisaged editing projects were abandoned due to the enormous and easily underrated amount of material. Similarly, earlier research was impeded by the vast extent of the journals as well as Senckenberg's difficult handwriting (Eulner, 1973, Kriegk, 1869). The ongoing research project is therefore the first genuine endeavour to explore this voluminous and demanding text source and make it available to the scholarly community as well as to the general public.

As a complete edition would hardly be feasible, only selected volumes are being critically edited in full text and equipped with a classical commentary. The critical edition consists of six volumes in total, representing the range of the collection in a content-related and in a biographical respect. On the one hand, this choice comprises two items from each subgroup, namely the self-referential, the medical and the moral observations. Moreover, it characterizes three different decades of Senckenberg's life. Further selection criteria considered the variety of topics, the amount of redundant contents and the possibility of integrating already existing transcription work. The born-digital TEI/XML-based selected edition will be published online under open access.

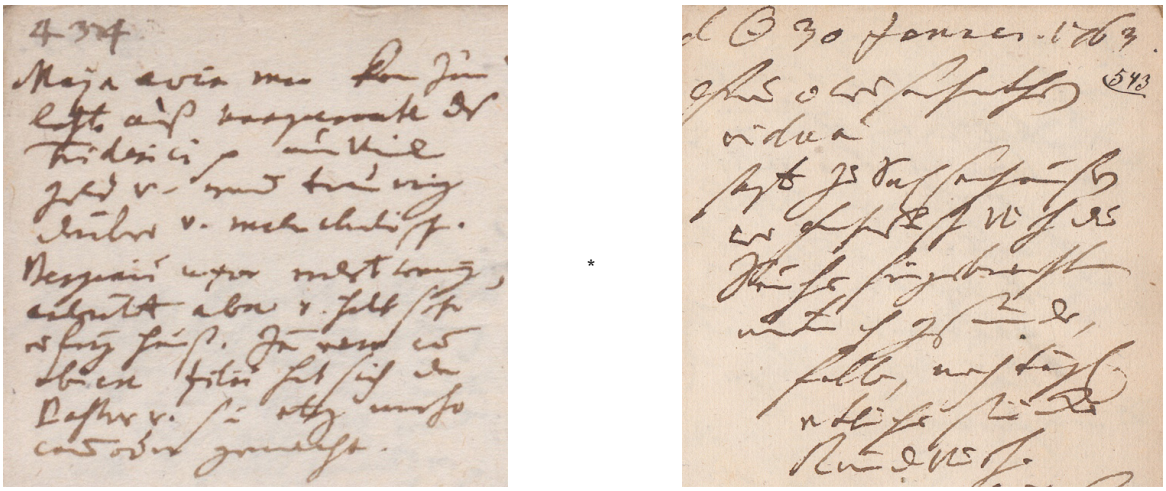
[3]http://sammlungen.ub.uni-frankfurt.de/senckenberg/nav/index/all (accessed 5. December 2017).

Senckenberg's writing routine aimed at taking down a large number of notes in a short time, which only he himself would have to be able to decipher. He kept alternating between German and Latin in one and the same sentence and abbreviated his words in various ways. These abbreviations often cut off the grammatical features at the word endings on which the comprehensibility of his bilingual and complex sentences depends. It is therefore not enough to transcribe the text in a diplomatic manner, but it is also necessary to complete the shortened words and adjust the supplementations to the present grammatical conditions. The same applies to the alchemical and astrological symbols, which are ambiguous both in a grammatical and in a semantic respect. They may either replace a whole word or only a part of it, and represent up to four different meanings along with all possible grammatical forms. The cross (+), for example, can stand for itself, but also for the nouns "acid" and "vinegar," for the adjective "sour," as well as all Latin equivalents in every grammatical case.

Our digital editing work does not only render the text legible for human eyes, but also constitutes its machine readability. The manually transcribed text selection can therefore serve as a basis for computerized pattern recognition, the outcome of which can subsequently be applied to all remaining volumes. By means of this mixed-methods approach, not only the contents of the fully edited volumes are made available for further research, but also the other items of the collection become accessible at least to a limited extent.

## 4 SEMI-AUTOMATIC HANDWRITTEN TEXT RECOGNITION

Consisting of large amounts of text written by one and the same hand, the Senckenberg Journals are an ideal candidate for handwritten text recognition (HTR). In the course of the EU project



a) 1732 b) 1762

**Figure 2: Senckenberg's Handwriting in 1732 vs. 1763**

READ, the Research Center for Pattern Recognition and Human Language Technology at the Polytechnical University of Valencia has developed the HTR tool Transkribus[4] (Mühlberger, 2015, Sánchez et al., 2013). Though still under development, it has already been released as an open-source-software available on the Transkribus web platform at the University of Innsbruck.[5] In order to recognize handwriting automatically, the tool must be trained on the basis of manually generated and precise text transcriptions. In this process, the digital facsimiles must be segmented into text regions, and the text lines have to be defined by baselines. Although Transkribus supplies a tool for automatic segmentation, the process could only be performed semi-automatically on Senckenberg's densely covered journal pages. Most of the automatically drawn baselines had to be corrected manually or supplied completely, especially in places where ascenders and descenders overlapped due to close line spacing. Next, the transcription of each text line was added and linked to its equivalent in the image. In this instance, transcription had to follow the rule "Transcribe only what you see," meaning that all characters must be represented as they appear in the original. Thus, extensions of word endings needed to be removed by Find and Replace commands. Also, paraphrases of symbols were replaced by Unicode characters as far as possible.

My test corpus consisted of 41 pages out of the first subgroup, on which about 16,000 words could be used for training. The training process resulted in an error rate of about 30%, which means that three out of ten characters are not recognized correctly. The fact that Senckenberg's handwriting changed over the years and produced new patterns necessitates separate training processes for all three subgroups, and subsequently the generation of three different HTR models.

The resultant train curve suggests that a continued training with four times the amount of transcription data would bring the error rate down to about 5% or less. Still, one would have to expect at least one faulty character in every third word on average. However, as the remaining volumes are devoted to the same research interest as those selected for the full text edition, they must necessarily treat analogous topics and use a related vocabulary producing similar patterns. It can therefore be expected that the HTR output will allow of fairly good results in full text search with truncated search terms. Moreover, all volumes can then be indexed by means of keyword spotting (Giotis et al., 2017).

## 5 AUTHORITY DATA, DIGITIZED PRINTS AND ONLINE PUBLISHING PLATORMS

Holding libraries can support the scholarly processing of their documents in various respects. Not only are they equipped with the technical facilities for manuscript digitization, but they also manage the authority data of the *Gemeinsame Normdatei* (GND) or *Integrated Authority File*. The Senckenberg edition refers its readers to the GND for all names of persons and places, as far as they are already included. As we are studying hitherto unedited private documents, however, we are bound to come across lesser-known persons whose historical significance is just being established. At the current state of work, this applies to about 25% of the persons Senckenberg mentions. In other instances, already existing entries contain no more detailed information than the respective person's name. Scholars can therefore assist the libraries by supplying their basic research data on historical persons, while the libraries may support the digital editors by setting up new GND entries or completing the existing ones. In the course of the Senckenberg project, about 140 entries have so far been added and corrected with the assistance of the Frankfurt UB. In other instances, scholars can help disambiguate and standardize GND entries. Especially older authors have often been listed several times under different numbers, and can therefore not be tagged unequivocally.

The Senckenberg edition also provides links to digital copies of printed books, which the author discusses or refers to. At present,

[4]https://www.prhlt.upv.es/wp/project/?2016/?recognition-and-enrichment-of-archival-documents (accessed 5. December 2017).
[5]https://transkribus.eu/Transkribus/#scholar-content (accessed 5. December 2017).

```
<p>daß er es auf eine hochzeit geschickt, da es wircklich zu viel gegessen v.
    getruncken<lb/> hat, Dr. <name type="person"
        ref="http://d-nb.info/gnd/172015634">Carl</name><note place="foot"
        resp="editor"> Johann Samuel Carl (1676-1757), Hofarzt in Berleburg </note>
    ei dedit <g>pulveres</g> aliq<ex>uo</ex>t digestivos et <g>pulveres</g> laxanti
    postea sumendum quod<lb/> exequi jussi. <lb/> 5 Kr<ex>euzer</ex>. gab dem
    hammerschmiede zum <g>Branntwein.</g><lb/> der wirth in <name type="place"
        ref="http://d-nb.info/gnd/4005725-2">schwartzenau</name> hat s<ex>ein</ex>
    fieber per <g>vitrio</g>li supprimirt ist aber noch matt,<lb/> war tertiana<note
        place="foot" resp="editor">dreitägiges Fieber</note> die er nicht los
    werden können, Ich commandirte ihm Infusum ex herba <lb/> bellid<ex>is</ex>
        minor<ex>is</ex>, veronic<ex>a</ex> &amp;amp; fl<ex>oribus</ex>
        Chamomill<ex>ae</ex>. Er hat 2 mäddgen Zwillinge,<lb/> die einander so
```

**Figure 3: TEI/XML-encoding with abbreviation and symbol resolution as well as GND references.**

about 80% of these titles have already been digitized and published online, which enables us to link them to the bibliographical references in our edition and thus provide our readers with direct access to the secondary sources. As a religious dissenter, however, Senckenberg was not only interested in widely-read literature, but also in heterodox and esoteric publications of which only very few copies are still extant today. In a journal entry from the August of 1732, for example, Senckenberg lists eleven books he purchased that day at a radical pietist publisher's. Apart from some alchemical titles, he chiefly bought biographical, polemical and devotional literature by dissenting authors, partly translated from French and English. In searching for digital copies of these eleven titles in various online catalogues, I have only had five hits altogether. Four of them I found via the *Bibliography of German Printed Books of the 17th and 18th centuries* (VD 17, VD 18),[6] which referred me to four different German University Libraries. Another work has been digitized by Google Books, however in much poorer quality. Most of the remaining titles were not even recorded in the catalogues and could at best be verified by means of other historical prints mentioning them. For one of the alchemical titles, Senckenberg's bibliographical reference is the only traceable witness of its existence on the whole world wide web to date. In addition, most of the books are no longer part of the Senckenberg collection, as his library was reduced to its strictly medical contents shortly after his death in 1772, while all religious and moral titles were auctioned (Burkhardt, 1992). Apart from showing the relevance of the journals and their edition for historical book and library studies, this example points out another possible field of cooperation: Libraries can support scholarly work by making rare books from their historical holdings accessible online, while scholars can help index and contextualize these titles and perhaps even reconstruct historical libraries in a digital environment.

Finally, thanks to their technical staff and capacities, libraries are able to provide infrastructures and repositories for digital editions. The best-known example is perhaps the virtual research environment TextGrid based at the Göttingen State and University Library (SUB). TextGrid supplies open-source-software comprising various XML-based tools, a geo-browser and dictionaries, and moreover an online publishing portal as well as a repository guaranteeing long-term data storage and maintenance.[7] Some research libraries, like the Herzog August Library of Wolfenbüttel (HAB), carry out in-house editing projects devoted to manuscript collections from their specialist field.[8] But also some smaller libraries, as for instance the Thuringian University and State Library (ThULB) of Jena, provide publication platforms for cultural heritage data and support high-standard scholarly editing.[9]

[6] http://www.vd17.de/;http://www.vd18.de/ (accessed 5. December 2017).

[7] https://textgrid.de/en_US (accessed 5. December 2017).

[8] http://diglib.hab.de/wdb.php?dir=edoc/ed000228&distype=?start&pvID=start (accessed 5. December 2017).

[9] http://www.urmel-dl.de/#portals/1 (accessed 5. December 2017).

## REFERENCES


Chris Anderson. 2008. The end of theory. Will the data deluge make the scientific method obsolete? In *Edge*. http://www.edge.org/3rd_culture/anderson08/anderson08_index.htm (accessed 5.12.2017).

Martin Brecht and Rudolf F. Paulus. 1979. Einleitung. In *Texte zur Geschichte des Pietismus*, Martin Brecht and Rudolf F. Paulus (Eds.). Berlin/New York, 9–41.

Helmut Burkhardt. 1992. Senckenbergische Bibliothek. Bestandsgeschichte. In *Handbuch der historischen Buchbestände in Deutschland*, Bernhard Fabian (Ed.). Vol. 5. Hildesheim, 174–177.

Antonio Cartelli and Marco Palma. 2009. Digistylus – An Online Information System for Palaeography Teaching and Research. In *Kodikologie und Paläographie im digitalen Zeitalter (Schriften des Instituts für Dokumentologie und Editorik 2)*, Malte Rehbein, Patrick Sahle, and Torsten Schaßan (Eds.). Norderstedt, 124–134.

Lorraine Daston and Elizabeth Lunbeck (Eds.). 2011. *Histories of Scientific Observation*. Chicago/London.

Hans-Heinz Eulner. 1973. Johann Christian Senckenbergs Tagebücher als historische Quelle. In *Johann Christian Senckenberg und die Medizin in Frankfurt am Main. Beiträge zum 200. Todestag Johann Christian Senckenbergs am 15. November 1972*, Günter Mann (Ed.). Frankfurt am Main, 1–11.

Vera Faßhauer. 2017a. Nachhaltige Erschließung umfangreicher handschriftlicher Überlieferungen. Ein Fallbeispiel. In *Konferenzabstracts Digitale Nachhaltigkeit. Konferenz der DHd 2017, Universität Bern, 13.–18. Februar 2017*. 162––165. http://www.dhd2017.ch/wp-content/uploads/2017/03/Abstractband_def3_M%C3%A4rz.pdf

Vera Faßhauer. 2017b. "Sacra à Deo in corde discenda, natura ex natura." Die Observationes Johann Christian Senckenbergs als medico-theologische Aufzeichnungspraktik. In *Berichte zur Wissenschaftsgeschichte*. Vol. 40. 225–246.

Vera Faßhauer. 2018. „Franco furtum, Furtum franco". Zur Bedingtheit von J. Chr. Senckenbergs Wahrnehmung der reichsstädtischen Eliten. In *Neue Stadtgeschichte(n). Frankfurt am Main in der Frühen Neuzeit*, Matthias Schnettger and Julia A. Schmidt-Funke (Eds.). Bielefeld, 219–247.

Angelos P. Giotis, Giorgos Sfikas, Basilis Gatos, and Christophoros Nikou. 2017. A survey of document image word spotting techniques. *Pattern Recognition* 68 (2017), 310–332.

Silke Kamp. 2009. Handschriften lesen lernen im digitalen Zeitalter. In *Kodikologie und Paläographie im digitalen Zeitalter (Schriften des Instituts für Dokumentologie und Editorik 2)*, Malte Rehbein, Patrick Sahle, and Torsten Schaßan (Eds.). Norderstedt, 111–122.

Georg Ludwig Kriegk. 1869. *Die Brüder Senckenberg. Eine biographische Darstellung nebst einem Anhang über Goethe's Jugendzeit in Frankfurt a. M.* Frankfurt am Main.

Robert Latham and William Matthews. 1970. Introduction. In *The Diary of Samuel Pepys. A new and complete transcription*, Robert Latham and William Matthews (Eds.). Vol. 1. London, xiv – cxxxvii.

Günter Mühlberger. 2015. Die automatisierte Volltexterkennung historischer Handschriften als gemeinsame Aufgabe von Archiven, Geistes- und Computerwissenschaftlern. Das Modell einer zentralen Transkriptionsplattform als virtuelle Forschungsumgebung. In *Digitalisierung im Archiv. Neue Wege der Bereitstellung des Archivguts. Beiträge des 18. Archivwissenschaftlichen Kolloquiums am 26. und 27. November 2013 (Buchveröffentlichungen der Archivschule Marburg 60)*, Irmgard Christa Becker and Stephanie Oertel (Eds.). Marburg, 87–116.

Antoine Odier, Alexander Zirr, Andreas Herz, and Arndt Schreiber (Eds.). 2013. *Digitale Edition und Kommentierung der Tagebücher des Fürsten Christian II. von Anhalt-Bernburg (1599–1656)*. Wolfenbüttel. http://diglib.hab.de/edoc/ed000228/start.htm (accessed 5.12.2017).

Joan-Andreu Sánchez, Günter Mühlberger, Basilis Gatos, Philip Schofield, Katrien Depuydt, Richard M. Davis, Enrique Vidal, and Jesse de Does. 2013. TranScriptorium. A European Project on Handwritten Text Recognition. In *Proceedings of the ACM Symposium on Document Engineering*. Florence, 227–228.

# Ein digitaler Zeichenkatalog als Organisationssystem für die noch nicht entzifferte Schrift der Klassischen Maya

Franziska Diehr
Maximilian Brodhun
Niedersächsische Staats- und Universitätsbibliothek
Göttingen, Deutschland
diehr|brodhun@sub.uni-goettingen.de

Sven Gronemeyer*
Katja Diederichs
Christian Prager
Elisabeth Wagner
Nikolai Grube
Rheinische Friedrich-Wilhelms-Universität Bonn,
Abteilung für Altamerikanistik
Bonn, Deutschland
sgronemeyer|katja.diederichs|cprager|ewagner|ngrube@
uni-bonn.de

## ABSTRACT

The aim of the research project "Text Database and Dictionary of Classic Mayan" is to create a corpus-based dictionary. In this context, we developed a digital sign catalogue that serves as knowledge organisation system for the Mayan script, which has not yet been fully deciphered. Its modeling is based on the specific characteristics of the script and pursues a new concept for the classification of signs, which takes graph variants into account and makes it possible to deal with uncertain reading hypotheses and undeciphered signs. Based on the ontologies CIDOC CRM and GOLD, the sign catalogue integrates vocabularies modelled in SKOS and a reference system for literature sources. In order to optimally represent the semantic relations, the data model was implemented in RDF. For data acquisition, we use an input mask in the virtual research environment TextGrid, which is specially adapted to project-specific requirements. As an essential part of the project specific data architecture the sign catalogue serves as a basis for the development of the corpus and dictionary.

## KEYWORDS

Digital Humanities, Knowledge Organization, Ontology, Modeling, Signs, Undeciphered Script

**Reference:**
Franziska Diehr, Maximilian Brodhun, Sven Gronemeyer, Katja Diederichs, Christian Prager, Elisabeth Wagner, and Nikolai Grube. 2018. Ein digitaler Zeichenkatalog als Organisationssystem für die noch nicht entzifferte Schrift der Klassischen Maya. In *Proceedings of Wissensorganisation 2017: 15. Tagung der Deutschen Sektion der Internationalen Gesellschaft für Wissensorganisation (ISKO) (WissOrg'17).* German Chapter of the ISKO / Freie Universität Berlin , pp. 37-43. https://doi.org/10.17169/FUDOCS_document_000000028863

*Also with La Trobe University, Department of Archaeology and History.


*WissOrg'17, December 2017, Berlin, Germany*

https://doi.org/10.17169/FUDOCS_document_000000028863

## 1 PROJEKTBESCHREIBUNG

Die über einen Zeitraum von fast 2000 Jahren verwendete Hieroglyphenschrift der vorspanischen Mayakultur ist Gegenstand der Forschung im Projekt "Textdatenbank und Wörterbuch des Klassischen Maya" (TWKM). Unser Ziel ist es, die bisher entdecken ca. 10.000 Textträger und ihre Texte in einem Korpus zu erschließen. Darauf aufbauend erstellen wir ein Wörterbuch, das den gesamten Sprachschatz sowie dessen Verwendung in der Schrift abbildet. Es handelt sich um das erste Projekt, das mittels maschinenlesbarer Erfassung des Ausgangsmaterials alle Vorkommen der Mayahieroglyphen mit Angabe der originalen Schreibung, Umschrift und übersetzung sowie Rahmeninformationen in einer korpusbasierten Datenbank integriert (der Wissenschaften und der Künste, 2013).

Das seit 2014 bis voraussichtlich 2029 laufende Projekt wird von der Nordrhein-Westfälischen Akademie der Wissenschaften und Künste sowie der Union der deutschen Akademien der Wissenschaften gefördert. Die für das Vorhaben an der Abteilung für Altamerikanistik der Universität Bonn eingerichtete TWKM-Arbeitsstelle kooperiert mit der Niedersächsischen Staats- und Universitätsbibliothek Göttingen, um in interdisziplinärer Zusammenarbeit Konzepte und Werkzeuge zur Erschließung und Dokumentation der Texte, ihrer Schriftträger und zum Aufbau des Wörterbuchs zu erarbeiten (Prager, 2014).

## 2 DAS SCHRIFTSYSTEM DES KLASSISCHEN MAYA

Auch wenn das Klassische Maya aufgrund seines ikonischen Charakters als Hieroglyphenschrift bezeichnet wird, handelt es typologisch gesehen um ein logo-syllabisches Schriftsystem, bei dem Logogramme und Syllabogramme die Hauptzeichenklassen bilden. Logogramme bezeichnen konkrete sprachliche Begriffe, wie z.B. PAKAL (Schild). Syllabogramme repräsentieren offene Silben und werden auch als phonetische Komplemente von Logogrammen benutzt. Wörter konnten ausschließlich aus Logogrammen (PAKAL) oder aus Syllabogrammen (pa-ka-la) oder aus einer Kombination aus beiden geschrieben werden (PAKAL-la), siehe Abb. 1 (Montgomery, 2002). Weitere Zeichentypen des Klassischen Maya sind Zahlzeichen und diakritische Zeichen.


**Abbildung 1: Verwendung von Logo- und Syllabogrammen**

## 3 KONZEPTION DES ZEICHENKATALOGS

Als Inventar aller Schriftzeichen dient ein Zeichenkatalog zur Identifizierung der verwendeten Glyphen in einem Text und ist damit ein essentielles Hilfsmittel für die Entzifferungsarbeit. Bisher hat die Mayaschriftforschung etliche Zeicheninventare hervorgebracht.[1] Vor dem Hintergrund, dass zur Zeit der ersten Kataloge erst wenige Zeichen entziffert waren, ist es kaum verwunderlich, dass sie Fehl- und auch Mehrfachklassifikationen aufweisen. Dies ist jedoch nicht ausschließlich dem zeitgenössischen Forschungsstand geschuldet, sondern begründet sich auch in der komplexen Graphembildung der Schrift, die die Zeichenbestimmung zu einer herausfordernden Aufgabe macht. Bis heute konnte die konkrete Anzahl der Schriftzeichen nicht bestimmt werden. Die Angaben bewegen sich zwischen 500 und 1000 Zeichen. Der Entzifferungsgrad bewegt sich dabei zwischen 60 und 80 Prozent.

Mit der Entwicklung unseres digitalen Zeichenkatalogs wollen wir ein neues Konzept der Systematisierung und Klassifikation von Mayaschriftzeichen etablieren. Ein Ziel des Projekts ist es, eine vollständige Neuinventarisierung der Zeichen vorzunehmen und damit auch eine Aussage über die Anzahl der Schriftzeichen zu treffen[2]. Das neue Organisationssystem soll es ermöglichen, die fehlerhaften Klassifikation der bis dato publizierten Kataloge zu korrigieren. Da die Inventare als Konkordanz in unseren Katalog miteinfließen, ist eine umfassende Sicht auf die bisher geleistete Klassifikationsarbeit der Mayaschriftzeichen gegeben. Das Organisationsprinzip des Katalogs ermöglicht weiterhin die Systematisierung der Zeichen unter Berücksichtigung ihrer multiplen Funktionstypen. Auch Hypothesen verschiedener Forscher zur Lesung nicht entzifferter Zeichen, werden in unserem Katalog dokumentiert und qualitativ eingestuft, so dass sie für spätere Analysen zur Verfügung stehen. Der Katalog ist jedoch nicht nur Werkzeug zur Klassifikation, sondern bildet auch den Grundbaustein für die Erstellung des Korpus der ca. 10.000 Texte.

Bei der Konzeption des Katalogs wählten wir einen neuen unorthodoxen Ansatz, der sich bewusst von bisherigen Organisationsprinzipien der Schriftforschung und insbesondere anderer Mayazeicheninventare abhebt. Wir haben gezielt traditionelle Herangehensweisen hinterfragt, um andere Möglichkeiten der Systematisierung und Organisation von Schriftzeichen zu untersuchen. Dazu haben wir bestehende Klassifikationssysteme und linguistische Terminologien untersucht, um passende Konzepte zur Beschreibung von Mayaschriftzeichen zu finden. Dabei stellten wir fest, dass die meisten Konzepte für die Modellierung unseres Katalogs nicht anwendbar sind, da sie bereits zu stark die Anwendbarkeit in einem konkreten linguistischen Zusammenhang fokussieren. Zur Beschreibung, Einordnung und Systematisierung der Schriftzeichen des Klassischen Maya wollen wir jedoch ein Organisationssystem schaffen, dass linguistische Ordnungsprinzipien nur auf einer Metaebene anwendet und weitere Analysestufen und Grammatiken nicht berücksichtigt. Der geringe Entzifferungsgrad der Schrift gibt Anlass für einen aktiven Forschungsdiskurs, der zu vielfältigen Interpretationsmöglichkeiten führt. Diese sind in einem Katalog, der ein Werkzeug zur Klassifikation nicht entzifferter Zeichen dient, zu berücksichtigen. Hypothesen, sowie auch neue Erkenntnisse, müssen in den Katalog aufgenommen werden, um sie für die weitere Analyseprozesse vorhalten zu können. Die Entzifferung der Zeichen des Klassischen Maya kann nur durch die linguistische Analyse des Korpus' erfolgen. Um das Korpus erstellen zu können, müssen die im Text verwendeten Zeichen identifiziert sein. Damit die Prozesse der Zeichenidentifaktion und anschließenden Textanalyse ineinander verschränkt werden können, ist ein Organisationssystem nötig, das flexibel auf Änderungen reagieren kann. Die nötige Flexibilität erreichen wir durch einen ontologisch-basierten Modellierungsansatz. Die Dokumentation von Mayaschriftzeichen in einem ontologisch-basierten, in digitaler Form repräsentierten, Wissensorganisationssystem ist in der Mayaschriftforschung bisher nicht erfolgt und stellt damit einen neuen Ansatz in der Erforschung des Schriftsystems dar.

### 3.1 Ontologisch-basierte Modellierung

Nach der Evaluation geeigneter Basis-Ontologien zur Modellierung des Zeichenkatalogs entschieden wir uns für die Verwendung von CIDOC Conceptual Reference Model (CRM). Trotz des Fokus' auf die Beschreibung von Prozessen zur Dokumentation von Objekten des kulturellen Erbes, enthält das CRM viele geeignete Metakonzepte, die für den Aufbau unseres Katalogs geeignet sind. Die Klassenhierarchie zeigt, dass die meisten Klassen des TWKM-Katalogs als Subklassen des CRM definiert wurden (siehe Abb. 2). Als "identifiable expressions in natural language" wurden *Sign* und *Graph* als spezifische Subkonzepte von *E33 Linguistic Object* gefasst (Group, 2011).

Als linguistische Ontologie, mit dem Ziel grundlegende Kategorien und Relationen für die wissenschaftliche Beschreibung von menschlicher Sprache zu definieren, schien sich die Nutzung von GOLD, General Ontology for Linguistic Description, anzubieten. Nach eingehender Prüfung stellte sich heraus, dass der Fokus von GOLD auf grammatikalische Regeln mit der Betrachtung der Morphosyntax als Ausgangspunkt (Farrar and Langendoen, 2003, 100) in unserem Kontext nur eingeschränkt genutzt werden konnte, da unser Katalogkonzept eine Metaebene abbildet, die zur Organisation von Schriftzeichen dient. Zur Definition der Klasse *SignFunction* konnten wir das Konzept *FeatureStructure* aus GOLD, nachnutzen. Mit seiner Definition als "a kind of information structure, a container or data structure, used to group together qualities or features of some object" (GOL, 2010) ist das Konzept der *FeatureStructure*

[1] Gates (1931), Zimmermann (1956), Evreinov et al. (1961), Thompson (1962), Knorozov (1963), Rendón and Spescha (1965), Grube (1990), Ringle and Smith-Stark (1996), Rodriguez Ochoa et al. (1999), Macri and Vail (2009)

[2] Die tatsächliche Anzahl der Mayaschriftzeichen wird niemals vollständig bestimmbar sein, da viele Textträger im Laufe der Zeit zerstört worden sind und der Forschung nicht zur Verfügung stehen.

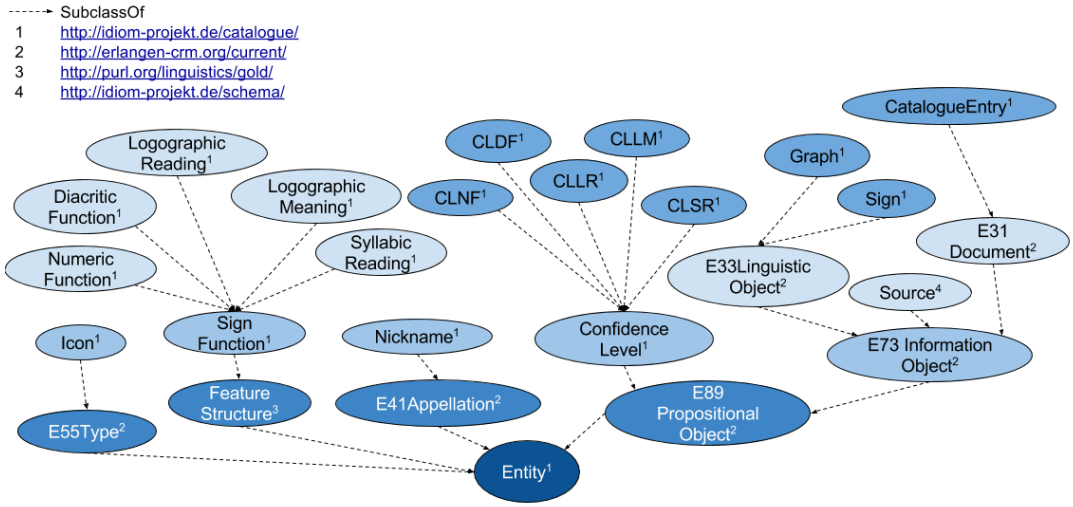



**Abbildung 2: Klassenhierarchie des TWKM-Zeichenkatalogs**

so allgemein gefasst, dass sich unsere Definition der Zeichenfunktion als "a feature assessed to a Sign. The nature of the feature is specified by the subclasses" als Subklasse fassen lässt (siehe Abb. 2) (Database and of Classic Mayan, 2017).

In unserem Zeichenkatalog definieren wir das Schriftzeichen als Entität, die einerseits aus einer Funktion und sprachlichen Expression und andererseits aus einer graphischen Repräsentation besteht. Dazu haben wir die Klassen *Sign* und *Graph* modelliert (siehe Abb. 3). *Graph* repräsentiert alle Varianten (Allographe) eines Graphems. Das *Sign* bestimmt sich durch seine *SignFunction*, welche die Funktion des Zeichens als Logogramm, Syllabogramm, Zahlzeichen oder Diakritikum fasst. Die sprachliche Ebene des Zeichens wird als Transliterationswert bei der jeweiligen Zeichenfunktion erfasst.[3]

## 3.2 Dokumentation von Graphvarianten

Viele der ca. 1000 Zeichen des Klassischen Maya haben mehrere Graphvarianten. Zum Beispiel hat die Silbe /u/ fünf Varianten, die sich zwar diagnostische Merkmale teilen, deren Graphe aber reduziert oder auch segmentiert gestaltet sind (siehe Abb. 4).[4] Dies verdeutlicht, warum die Bestimmung des Graphems, also die Gesamtheit aller Graphe des Zeichens, eine Herausforderung darstellt.

Um dieser schwierigen Aufgabe zu begegnen und eine Möglichkeit der flexiblen Zuordnung und damit andere Art der Graphembestimmung zu ermöglichen, haben wir die Klasse *Graph* modelliert. Sie beschreibt die einzelnen Varianten eines Zeichens. Alle Graphe, die einem Zeichen zugeordnet sind, bilden zusammen das Graphem dieses Zeichens (siehe Abb. 5).[5] Durch die separate Erfassung der einzelnen Graphe ermöglichen wir eine exakte Dokumentation der einzelnen Varianten, wobei wir auch Relationen zwischen den Graphen herstellen, um bspw. auf gemeinsame diagnostische Merkmale hinzuweisen.[6]

Auf Vorarbeiten aus der Forschungscommunity stützend (Houston, 2001, Kelley, 1962) konnten im Projekt erstmals Regeln und Prinzipien zur Bildung von Graphen der Mayaschriftzeichen entwickelt werden. Es wurden 45 Variationstypen definiert, die sich in neun Klassen gliedern (Mono-, Bi-, Tri-, and Variopartite, Division, Animation Head, Animation Figure, Multiplication, Extraction). Jeder Typ ist durch ein Kürzel gekennzeichnet, z.B. "vl" für "variopartite left". Das Graph wird mit der funktionalen und sprachlichen Repräsentationsebene des Zeichens (*Sign*) in Relation gesetzt. Diese Verbindung ist optional, so dass auch Graphe erfassbar sind, die noch keinem Zeichen zugeordnet werden konnten. Das Zeichen wird mit einer Katalognummer versehen, z.B. 123. Wenn ein Graph als Allograph eines Zeichens identifiziert wurde, dann erhält es eine *graphNumber*, die aus der Katalognummer des Zeichens und dem Kürzel des Variationstyps gebildet wird, z.B. 123vl (siehe Abb. 3).

Mayaglyphen weisen einen stark ikonischen Charakter auf. Neben eher abstrakt aussehenden Elementen zeigen sie häufig Gegenständliches wie Menschenköpfe, Tierschädel oder auch Handlungen wie bspw. das Schlagen mit einer Axt. Um die Durchführung paläo- und ikonographischer Studien zu ermöglichen, entwickelten wir ein in Simple Knowledge Organization System (SKOS) modelliertes kontrolliertes Vokabular. Es umfasst 13 Facetten (human and animal, body part, age and sex, pose and gesture, plant and plant part, landscape, artefact physical state, property, orientation), die das Ikon der Graphe formal beschreibbar machen. Das Metadatenschema integriert das Vokabular, indem die einzelnen Konzepte beim Graph erfasst werden können (siehe Abb. 3).

[3] Der Transliterationswert entspricht dabei nicht einem phonemischen Wert des Zeichens, sondern einem graphemischen Wert. Der eigentliche Lesungswert kann erst durch die linguistische Analyse mit dem Korpus bestimmt werden.
[4] Zeichnungen von Christian Prager
[5] Zeichnungen: 126bh und br von Christian Prager, T246 und T126 von Eric S. Thompson, A Catalog of Maya Hieroglyphs, 1962
[6] Für weitere Informationen zu den Relationen siehe auch die Dokumentation des Zeichenkatalogs erreichbar unter http://idiom-projekt.de/catalogue

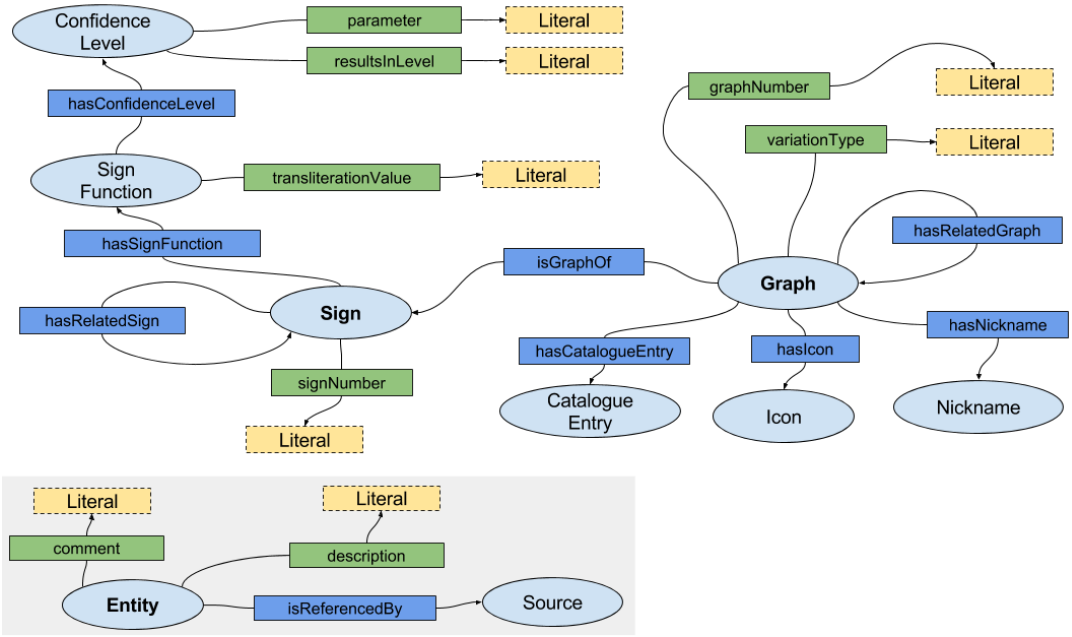



**Abbildung 3: Domain Model des TWKM-Zeichenkatalogs**

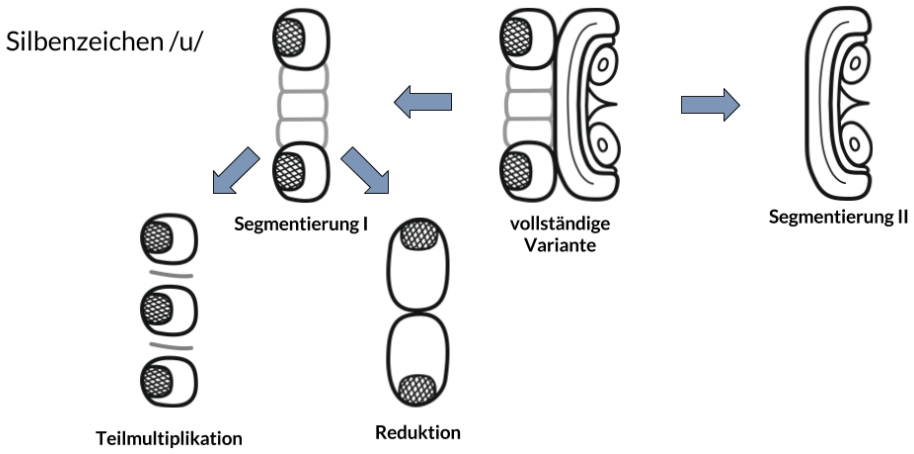



**Abbildung 4: Graphvarianten der Silbe /u/**

## 3.3 Zeichenkatalogkonkordanz

Die noch relativ junge Erforschung des Schriftsystems hat bisher ein knappes dutzend Zeicheninventare hervorgebracht. Aufgrund der herausfordernden Graphembestimmung enthalten sie jedoch häufig Mehrfach- und Falschklassifikationen, die sich an einem Beispiel aus dem etablierten Katalog von J. Eric S. Thompson aus dem Jahre 1962 verdeutlichen lassen: Beim Zeichen "T246" handelt es sich eigentlich um eine Ligatur aus zwei anderen Zeichen, die von Thompson als "T126" und "T136" inventarisiert wurden (siehe Abb. 6) (Thompson, 1962, 450).

Die bisher veröffentlichten Zeicheninventare fließen als Konkordanz in unseren Katalog ein. Um auf die Problematik der fehlerhaften Klassifikation zu reagieren, erfassen wir die konkordanten Katalogeinträge beim jeweiligen Graph. Dies ermöglicht uns auch falsch inventarisierte Zeichen zu dokumentieren, da ein Graph optional einem Zeichen zugeordnet werden kann. Das Zeichen "T246" würde bspw. als Katalogeintrag bei den Graphen "126st" und "136st" vermerkt. Damit ist ein Vergleich zwischen den Katalogen möglich und die Nummern aller Kataloge suchbar. Weiterhin wird durch die Konkordanz ein Überblick über die bisher geleistete Zeichenklassifikationsarbeit der Mayaschriftforschung geschaffen.

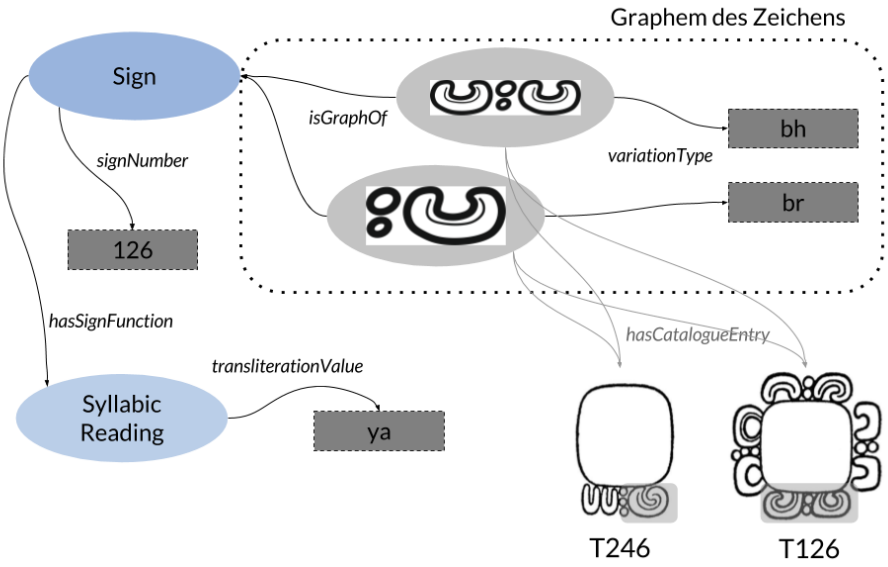



**Abbildung 5: Flexible Graphembestimmung durch separate Erfassung von *Graph***

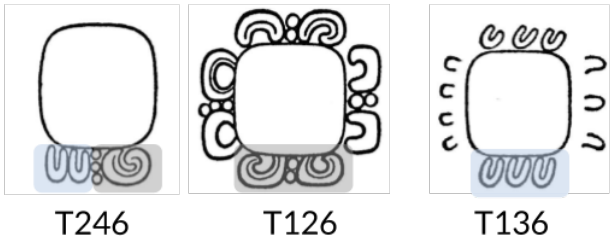



**Abbildung 6: Mehrfachklassifikation im Katalog von Thompson**

## 3.4 Multiple Zeichenfunktionen

Eine weitere Charakteristik der Mayaschriftzeichen besteht darin, dass ein Zeichen mehrere funktionale Ebenen und damit auch sprachliche Lesungen annehmen kann. Das Zeichen mit der Katalognummer "528" kann sowohl als Logogramm TUN (Stein), als Logogramm CHAHUK (der 19. Tagesname) oder auch als Silbe /ku/ gelesen werden. Dabei geben die verwendeten Graphvarianten keinen Hinweis darauf, welche der Lesungen vorliegt. Das Zeichen hat also nicht nur eine, sondern mehrere funktionale und sprachliche Repräsentationen. Für die Zeichen des Klassischen Maya konnten fünf Zeichenfunktionen definiert werden: Zahlzeichen, diakritische Zeichen, Logogramme mit identifizierter Lesung, Logogramme mit unidentifizierter Lesung (hier wird dem Zeichen eine Bedeutung zugeordnet) und syllabische Zeichen mit identifizierter Lesung. Um dieses Phänomen im Zeichenkatalog zu repräsentieren, haben wir die Klasse *SignFunction* mit den jeweiligen Funktionstypen als Unterklassen modelliert (siehe Abb. 2). Die Lesung, besser der Transliterationswert, wird bei der entsprechenden Zeichenfunktion erfasst (siehe Abb. 3). Pro Funktion ist nur ein Wert zulässig, jedoch kann ein Zeichen mehrere Funktionstypen annehmen.

## 3.5 Bewertung von Lesungshypothesen

In der Mayaschriftforschung entfachen noch nicht entzifferte Zeichen einen regen Diskurs, aus dem ständig neue Vorschläge für die Lesung dieser Zeichen hervorgehen. Die Hypothesen müssen in den Zeichenkatalog integriert werden, damit sie für eine Analyse im Korpus zur Verfügung stehen und auf ihre Plausibilität überprüft werden können. Die Lesungsvorschläge weisen unterschiedliche Qualitätsstufen auf. Manche scheinen plausibler als andere. Um die Qualität der Lesungen auch formal bewertbar zu machen, haben wir je Zeichenfunktion ein Kriterienset entwickelt, das sich u.a. am sprachlichen Nutzungskontext (z.B. Wortart, plausibler Text-Bild-Bezug) oder dem Nachweis in modernen Mayasprachen orientiert. Die Kriterien sind mittels Aussagenlogik in Bezug gesetzt, so dass sich, je nach deren Kombination, eine Qualitätsstufe ergibt. Um diese im Modell zu repräsentieren, wurde die Klasse *Confidence-Level* modelliert, die in Relation mit der Klasse *SignFunction* gesetzt wird (siehe Abb. 3). Für den Transliterationswert, der bei der Zeichenfunktion erfasst ist, kann somit eine qualitative Einstufung vorgenommen werden.

Die Einstufung ist insbesondere bei der Plausibilitätsprüfung des Lesungsvorschlags im Korpus relevant. Hier können Lesungen mit besonders hoher Stufe mit denen niedriger Stufe verglichen werden. Für letztere könnten sich bei der Prüfung im Textkontext auch neue Kriterien für deren Plausibilität finden, die dann im Zeichenkatalog ergänzt werden können. Eventuell steigt damit auch deren Qualitätsstufe.

Mit Hilfe der qualitativen Bewertung soll es ermöglicht werden, gesicherte Lesungen für die Schriftzeichen des Klassischen Maya sowie auch neue Entzifferungsvorschläge vorzulegen.

```
<ab xml:id="n2" type="glyph-block">
    <g xml:id="n2G1" n="5017st" ref="textgrid:30gnx.0" rend="left_beside" corresp="#n2G2"/>
    <g xml:id="n2G2" n="16st" ref="textgrid:2skxk.0" rend="left_beside" corresp="#n2S1"/>
    <seg xml:id="n2S1" type="glyph-group" rend="right_beside" corresp="#n2G2">
        <g xml:id="n2G3" n="1010st" ref="textgrid:34rkg.0" rend="above" corresp="#n2G4"/>
        <g xml:id="n2G4" n="116st" ref="textgrid:34rkh.0" rend="beneath" corresp="#n2G3"/>
    </seg>
</ab>
```

**Abbildung 7: Beispiel für Auszeichnung der Graphe im Korpus**

## 4 TECHNISCHE UMSETZUNG IN TEXTGRID

Das erarbeitete Konzept für den digitalen Zeichenkatalog fordert eine Datenstruktur, die es erlaubt, semantische Relationen zwischen eindeutig referenzierbaren Entitäten herzustellen. Ein in RDF realisiertes Datenmodell stellt damit die optimale Form der Wissensrepräsentation dar. Im TWKM-Projekt nutzen wir für die Verwaltung, Erstellung und Präsentation der im Projekt erzeugten Daten die Virtuelle Forschungsumgebung TextGrid.[7] Um die Zeichen in unserem digitalen Katalog zu erfassen, haben wir die RDF-Eingabemaske des TextGrid Labs auf die projektspezifischen Bedürfnisse angepasst.

### 4.1 Dokumentation von Quellennachweisen

Im Zeichenkatalog werden nicht nur Lesungsvorschläge der im Projekt arbeitenden Wissenschaftler dokumentiert, sondern auch publizierte Hypothesen anderer Forscher. Daher ist die Angabe der entsprechenden Quelle unverzichtbar. Um einen lückenlosen Nachweis aller dokumentierten Informationen zu ermöglichen, haben wir ein Nachweissystem für Literaturquellen geschaffen. Zur Erstellung und Verwaltung unserer Projektbibliografie nutzen wir das Literaturverwaltungsprogramm Zotero.[8] Wenn wir einen Quellennachweis erfassen, wird die Zotero-API von der RDF-Erfassungsmaske in TextGrid angesprochen. Es werden URI und die bibliografische Angabe der Quelle bei dem entsprechenden Datensatz gespeichert. Im Metadatenschema wurde die Klasse *Source* modelliert, die in Relation mit allen Entitäten des Katalogs gesetzt werden kann (siehe Abb. 3). Somit können nicht nur die Lesungshypothesen, sondern auch alle anderen Informationen, wie z.B. die konkordanten Katalogeinträge, mit Quellenangaben versehen werden.

## 5 BAUSTEIN FÜR DIE KORPUSERSTELLUNG

Um ein maschinenlesbares Korpus zu erstellen, muss Text vorhanden sein, der ausgezeichnet werden kann. Nun stehen wir vor dem Problem, dass aufgrund der komplexen Kalligrafie der Schriftzeichen und ihrer Graphvarianten auf keinen standardisierten Schriftzeichensatz, wie etwa Unicode, zurückgegriffen werden kann. Zweitens können Zeichen mehrere Funktionstypen haben und diese können wiederum mehrere Lesungsvorschläge haben. Daher können keine phonemisch-transliterierten Werte im Korpus codiert werden. Letzteres würde auch die Möglichkeit ausschließen, die jeweils im Text verwendete Graphvariante auszuzeichnen. Da wir Untersuchungen zu Graphvarianten und deren Verwendungskontext ermöglichen wollen, ist dies jedoch eine notwendige Anforderung. Die einzige Möglichkeit den Korpustext zu erzeugen, besteht darin, auf die graphischen Repräsentationen, also die Graphe zu verweisen. Hier kommt der digitale Zeichenkatalog ins Spiel: Im TEI/XML Code wird jede Glyphe mittels eines Verweises auf die URI des Graphs im Zeichenkatalog erfasst (siehe Abb. 7). Der Text wird somit aus den Instanzen der in RDF-vorliegenden Daten des Zeichenkatalogs gebildet. Mittels eines weiteren Prozessierungsschritts entsteht ein menschenlesbarer Text, der mit den im Zeichenkatalog hinterlegten Transliterationswerte angereichert wird. Auf dieser Basis können nun linguistische Analysen unter Berücksichtigung der multiplen Zeichenfunktion und der verschiedenen Lesungsvorschläge erfolgen.

## 6 VERÖFFENTLICHUNG UND MÖGLICHKEITEN DER DATENNACHNUTZUNG

Die Dokumentation und wissenschaftliche Erschließung der Textträger und des Korpus' sind laufende Arbeiten, die bis Projektende 2029 fertiggestellt werden sollen. Die Inventarisierung der Zeichen mittels des digitalen Katalogs befindet sich derzeit ebenfalls in Bearbeitung und wird voraussichtlich Mitte kommenden Jahres zum Abschluss kommen. Mit seiner Fertigstellung wird der Katalog auf unserem Projektportal veröffentlicht und die RDF-Daten über einen SPARQL-Endpoint zugänglich gemacht. Weiterhin werden die Daten im TextGrid Repository veröffentlicht, wo sie mittels einer OAI-PMH Schnittstelle auch für externe Nutzer abrufbar sind. Die Dokumentation des digitalen TWKM-Zeichenkatalogs ist bereits veröffentlicht und unter http://idiom-projekt.de/catalogue erreichbar.

## 7 FAZIT

Da sich das Konzept des Zeichenkatalogs gezielt von der Einordnung von Schriftzeichen in linguistische Kategorien abhebt, ist es auch für andere noch nicht (vollständig) entzifferte Sprachen übertragbar. Insbesondere die Trennung der graphischen und der funktional-sprachlichen Ebenen, die je nach Erkenntnisstand aufeinander bezogen werden können, bietet eine Flexibilität, die die Klassifikation von Schriftzeichen neu definiert und eine präzise Identifikation anhand bedeutungsunterscheidender Merkmale zulässt. Durch die Einbeziehung bekannter Ergebnisse sowie der Anpassbarkeit an neue Erkenntnisse, ist das Modell dem noch nicht gefestigten Erkenntnisstand der Schriftforschung angepasst. Die

[7] TextGrid - Eine Virtuelle Forschungsumgebung für die Geisteswissenschaften, https://textgrid.de/
[8] Zotero, https://www.zotero.org/

Repräsentation des Wissens im digitalen Raum schafft neue Perspektiven auf das Material und neue Möglichkeiten der Textanalyse und -interpretation. Unser Konzept zur Klassifikation noch nicht vollständig entzifferter Maya-Schriftzeichen und dessen Implentierung in einem digitalen Zeichenkatalog stellt einen Beitrag zur Anwendung von Wissensrepäsentationssystemen in den Digital Humanities dar.

## LITERATUR

2010. General Ontology for Linguistic Description (GOLD). *Department of Linguistics (The LINGUIST List)* (2010). http://linguistics-ontology.org/

Text Database and Dictionary of Classic Mayan. 2017. Sign Catalogue of the TWKM project. (2017). http://idiom-projekt.de/catalogue

Nordrhein-Westfälische Akademie der Wissenschaften und der Künste. 2013. 10,7 Millionen Euro für zwei neue Langzeitforschungsprojekte. (2013). Retrieved May 11, 2017 from http://www.awk.nrw.de/pressemedien/detailansicht-presse/2013-11-25-107-millionen-euro-fuer-zwei-neue-langzeitforschungsprojekte.html

Scott Farrar and D. Terrence Langendoen. 2003. A linguistic ontology for the Semantic Web. *GLOT International* 7, 3 (2003), 97–100.

ICOM/CIDOC CRM Special Interest Group. 2011. CIDOC Conceptual Reference Model, Version 5.0.4. (2011). http://www.cidoc-crm.org/cidoc-crm/

Stephen Houston. 2001. *The Decipherment of Ancient Maya Writing*. University of Oklahoma Press, Norman. 3–19 pages.

David H. Kelley. 1962. Review of A Catalog of Maya Hieroglyphs, by J. Eric S. Thompson. *American Journal of Archaeology* 66 (1962), 436–438.

John Montgomery. 2002. *How to Read Maya Hieroglyphs*. Hippocrene, New York, NY.

Christian Prager. 2014. Zielsetzung: Textdatenbank und Wörterbuch des Klassischen Maya. (2014). Retrieved May 11, 2017 from http://mayawoerterbuch.de

J. Eric S. Thompson. 1962. *A Catalog of Maya Hieroglyphs*.

Christian Wartena, Michael Franke-Maier and Ernesto De Luca. 2018. *Proceedings of Wissensorganisation 2017: 15. Tagung der Deutschen Sektion der Internationalen Gesellschaft für Wissensorganisation (ISKO) (WissOrg'17).* German Chapter of the ISKO / Freie Universität Berlin. ISBN 978-3-96110-163-4
https://doi.org/10.17169/FUDOCS_document_000000028863