Andreas Kuczera,
Thorsten Wübbena,
Thomas Kollatz (Hg.)

Die Modellierung des Zweifels – Schlüsselideen und -konzepte zur graphbasierten Modellierung von Unsicherheiten

4

ZfdG Sonderband

Zeitschrift für digitale Geisteswissenschaften

Beitrag aus:
Sonderband 4 der ZfdG: Die Modellierung des Zweifels – Schlüsselideen und -konzepte zur graphbasierten Modellierung von Unsicherheiten. Hg. von Andreas Kuczera, Thorsten Wübbena und Thomas Kollatz. 2019. DOI: 10.17175/sb004
Titel:
Die Modellierung des Zweifels – Schlüsselideen und -konzepte zur graphbasierten Modellierung von Unsicherheiten. Zur Einführung in diesen Band

Autor/in:
Andreas Kuczera

Kontakt: andreas.kuczera@geschichte.uni-giessen.de
Institution: Akademie der Wissenschaften und der Literatur | Mainz
GND: 1167802993 ORCID: 0000-0003-1020-507X

Autor/in:
Thorsten Wübbena

Kontakt: twuebbena@dfk-paris.org
Institution: Deutsches Forum für Kunstgeschichte Paris
GND: 123312396 ORCID: 0000-0001-8172-6097

Autor/in:
Thomas Kollatz

Kontakt: thomas.kollatz@adwmainz.de
Institution: Akademie der Wissenschaften und der Literatur | Mainz
GND: 1063010942 ORCID: 0000-0003-1904-1841

DOI des Artikels:
10.17175/sb004_013

Nachweis im OPAC der Herzog August Bibliothek:
1046330314

Erstveröffentlichung:
16.01.2019

Lizenz:

Sofern nicht anders angegeben (cc) BY-SA

Medienlizenzen:
Medienrechte liegen bei den Autoren

Letzte Überprüfung aller Verweise:
15.01.2019

GND-Verschlagwortung:
Digital Humanities | Graphdatenbank | Konzeptionelle Modellierung | Zweifel |

Zitierweise:
Andreas Kuczera, Thorsten Wübbena, Thomas Kollatz: Die Modellierung des Zweifels – Schlüsselideen und -konzepte zur graphbasierten Modellierung von Unsicherheiten. Zur Einführung in diesen Band. In: Zeitschrift für digitale Geisteswissenschaften. 2016. PDF Format ohne Paginierung. Als text/html abrufbar unter DOI: 10.17175/sb004_013.

Andreas Kuczera, Thorsten Wübbena, Thomas Kollatz

Die Modellierung des Zweifels – Schlüsselideen und -konzepte zur graphbasierten Modellierung von Unsicherheiten. Zur Einführung in diesen Band

Abstracts

Einführung

Graphdatenbanken werden seit einigen Jahren auch zunehmend in geisteswissenschaftlichen Forschungsvorhaben zur Modellierung von Forschungsdaten und erschließendem Wissen genutzt. Sie ergänzen zunehmend relationale oder auch auf XML beruhende Datenbanksysteme und stellen einen zentralen Punkt der Tagungsreihe »Graphentechnologien« dar, die seit 2017 jährlich an der Akademie der Wissenschaften und der Literatur in Mainz stattfindet. 2018 stand die am 19. und 20. Januar durchgeführte Veranstaltung unter dem Titel:

»Die Modellierung des Zweifels« – Schlüsselideen und -konzepte zur graphbasierten Modellierung von Unsicherheiten

Die Modellierung, das transparente, möglichst objektive Annotieren von Zweifeln ist eine große Herausforderung in den digitalen Geisteswissenschaften. Wenn in vor-digitalen Zeiten sprachliche Mittel für die Abwägung und Beschreibung von Zweifel zur Verfügung standen, die immer auch mit einer gewissen Vagheit verbunden sein konnten, sind digitale Ansätze meist expliziter strukturiert. Darüber hinaus können in verschiedenen Anwendungszusammenhängen heterogene Modellierungsansätze verfolgt werden. Eine große Herausforderung ist dabei die eindeutige und vereinbare Handhabung unsicherer Informationen. Neben der Frage, ob eine Information selbst unsicher ist, kann auch der Grad der Unsicherheit zwischen verschiedenen Beteiligten umstritten sein. Bei dieser Problematik bieten Graphentechnologien mit ihrer Flexibilität interessante Ansätze für eine »Modellierung des Zweifels«, vor allem auch vor dem Hintergrund der stetigen Zunahme und des Wandels wissenschaftlicher Informationen im Laufe der Zeit. In den aus Erhebungen, externen Quellen oder Forschungsprojekten gewonnenen Daten werden Entitäten identifiziert und anschließend Verknüpfungen zwischen ihnen gesucht. Technisch gesehen ergibt sich hier die Herausforderung, die dabei auftretenden Unsicherheiten im Datenmodell im Hinblick auf Transparenz und Interoperabilität zu berücksichtigen, um eine Auswertung der Daten in anderen Vorhaben nicht zu behindern.

Auf der Tagung näherten sich insgesamt 15 Vorträge aus den unterschiedlichsten fachlichen Perspektiven dem Thema an (Tagungsprogramm). Hieraus entwickelten sich sehr interessante, interdisziplinäre Diskussionen, so dass bei den Herausgebern – die zuvor als Programmkomitee für die Auswahl der Beiträge auf der Tagung verantwortlich zeichneten – die Entscheidung reifte, die Tagungsbeiträge inkl. der Diskussionserträge in einem Sammelband zusammenzufassen. So werden nun zwölf Vorträge als vierter Sonderband der Zeitschrift für digitale Geisteswissenschaften erscheinen.

Die erste Session der Graphtagung 2018 stand unter dem Leitthema **Text** und der hiermit verbundene Artikel »Referenzielle Vagheit und Varianz in Texten über Musik« von **Torsten Roeder** unternimmt einen Brückenschlag zwischen XML und Graphentechnologien zur Untersuchung von Referenzen in einem historischen Textkorpus am Beispiel von musikkritischen Texten des 19. Jahrhunderts. Hierbei wird insbesondere die Frage verfolgt, inwieweit Graphenabfragen die Analyse von Referenzialität erleichtern können, insbesondere im Hinblick auf unscharfe Angaben. In ihrem Beitrag »Modellierung von Entzifferungshypothesen in einem digitalen Zeichenkatalog für die Maya-Schrift« schildern **Franziska Diehr et al.** die Entwicklung eines digitalen Zeichenkatalogs für die Maya-Schrift, unter besonderer Berücksichtigung der Modellierung von Entzifferungshypothesen und deren qualitativer Bewertung. **Dominik Kasper** und **Andreas Kuczera** stellen in ihrem Artikel »Modellierung von Zweifel – Vorbild TEI im Graphen« die Auszeichnung von unsicheren Lesarten und editorischer Ergänzungen in Handschriften in den Mittelpunkt, welche sie anhand von TEI-Dokumenten aus dem Deutschen Textarchiv (DTA) untersucht haben, welche eigens dafür in eine Graphdatenbank importiert wurden. Sie verdeutlichen, wie Graphtechnologien hier u.a. die Möglichkeit zur Analyse von Unsicherheit bieten und sich bei einer ausreichend großen Datenmenge persönliche Auszeichnungsprofile der jeweiligen Bearbeiter*innen erstellen lassen.

Die zweite Session trug den Titel **Unsicherheit** und in ihrem Text »Academic Meta Tool – Ein Web-Tool zur Modellierung von Vagheit« stellen **Martin Unold** und **Florian Thiery** eine Methodik zur Modellierung von Vagheit in Graphen vor. Darüber hinaus behandeln sie auch die automatisierte Generierung von implizit gespeichertem Wissen unter Berücksichtigung von Vagheit. Aus musikwissenschaftlicher Perspektive nähert sich **Stefan Münnich** der Problematik der Quellenverluste und der damit verbundenen Unsicherheit an. In seinem Artikel »Quellenverluste (deperdita) als methodologischer Unsicherheitsbereich für Editorik und Datenmodellierung am Beispiel von Anton Weberns George-Lied op. 4 Nr. 5« beleuchtet er die Folgen solcher Fehlstellen für das kulturelle Gedächtnis am Beispiel der editorischen Praxis der Anton Webern Gesamtausgabe.

In der mit **Theorie** betitelten Session Drei findet sich der Beitrag »Accepting and Modeling Uncertainty« von **Michael Piotrowski** wieder. Der Autor zielt hier auf auf die Herausforderung ab, welche Unsicherheit für die Modellierung in den Geisteswissenschaften bedeutet. In den Naturwissenschaften wird eine entsprechende Grundlagenforschung betrieben, aber laut Autor fehlt hier noch eine »Brücke« zu den Geisteswissenschaften, die helfen könnte, die Unsicherheit mit solchen formalen Modellierungsrahmen zu überwinden. **Andreas Wagner** stellte in »Ambiguität und Unsicherheit: Drei Ebenen eines Datenmodells« anhand eines Forschungsprojekts aus dem rechtshistorischen Kontext einen Ansatz vor, wie Unsicherheit abzubilden sein könnte. Er schlägt dazu eine Modellierung des Forschungszusammenhangs auf drei Ebenen vor, welche (a) die historischen Phänomene, (b) die überkommenen Zeugnisse dieser Phänomene und (c) die aktuelle historische Forschung selbst beschreiben.

Technik lautete das Thema der vierten Session und der Beitrag von **Iian Neill** und **Andreas Kuczera** stellte mit SPEEDy einen neuen Ansatz zur Annotation von Texten vor. Grundlage sind Standoff Properties, die indexbasiert mehrdimensionale Annotationen mit Zuordnung zu den jeweiligen Nutzenden ermöglichen. Die Modellierung von Zweifel wird damit über die Möglichkeit konkurrierender Annotationen erleichtert. In »Blockchain – die etwas andere Datenbank« strebt **Katarina Adam** eine Versachlichung der in den Massenmedien stattfindenden Diskussion zur Blockchain-Technologie und den damit verbundenen Themen wie Bitcoin, Risiko und Spekulation an. Die Autorin erläutert, in welchem Umfeld diese Technologie

implementiert wurde, wie der aktuelle Entwicklungsstand ist und wie die Blockchain-Technologie von anderen Ansätzen der Datenspeicherung und -bearbeitung lernen kann.

Der fünfte Block der Tagung stand unter der Frage nach der **Erschließung** und im zugehörigen Beitrag »A Graph Database of Aegean Seals with Uncertain Attributes« von **Martina Trognitz** werden von der Autorin die verschiedenen Quellen der Unsicherheiten im Kontext von mehrseitigen ägäischen Siegeln untersucht und dargestellt, wie diese in einer Graphdatenbank modelliert werden können. In »Genau, wahrscheinlich, eher nicht: Beziehungsprobleme in einem kunsthistorischen Wissensgraph« erörtern **Martin Raspe** und **Georg Schelbert** die »doppelte“ Herausforderung, geisteswissenschaftliche Forschungsdaten in einem digitalen Wissensgraph abzubilden: Wie werden ungewisse Informationen gespeichert und welche Folgen entstehen daraus für Abfrage und Visualisierung? Was drückt Unsicherheit aus und wie beeinflusst sie unser Konzept von Wissen? **Thomas Efers** Beitrag zum Thema »Graphbasierte Modellierung von Faktenprovenienz als Grundlage für die Dokumentation von Zweifel und die Auflösung von Widersprüchen« stellt die Wichtigkeit einer nachvollziehbaren Herkunftsannotation für Wissensbasen in den Digitalen Geisteswissenschaften heraus. Neben der Vorstellung genereller Aspekte bei der Modellierung von Aussagen auf abstrakter und beispielgeleiteter Ebene wird das Konzept einer »Faktenprovenienz« entwickelt und in Aussagemodelle integriert.

Zu danken haben wir dem Verband Digital Humanities im deutschsprachigen Raum (DHd) für die finanzielle Unterstützung dieser Publikation, neo4j für die Finanzierung von Reisestipendien für die Vortragenden, der Akademie der Wissenschaften und der Literatur, Mainz, für die Bereitstellung von Räumen und die Versorgung vor Ort und schließlich den Herausgebern der Zeitschrift für digitale Geisteswissenschaften für die Aufnahme dieses Sammelbandes als Sonderband sowie Lisa Klaffki, Henrike Fricke-Steyer und nicht zuletzt Timo Steyer für das Lektorat und die gute Zusammenarbeit bei der Erstellung dieses Sonderbands.

Andreas Kuczera, Thorsten Wübbena, Thomas Kollatz

Mainz und Paris im Januar 2019

ZfdG

Zeitschrift für digitale Geisteswissenschaften

Beitrag aus:
Sonderband 4 der ZfdG: Die Modellierung des Zweifels – Schlüsselideen und -konzepte zur graphbasierten Modellierung von Unsicherheiten. Hg. von Andreas Kuczera, Thorsten Wübbena und Thomas Kollatz. 2019. DOI: 10.17175/sb004

Titel:
Referenzielle Vagheit und Varianz in Texten über Musik. Annäherungen zwischen Textkodierung und graphbasierten Analysemodellen

Autor/in:
Torsten Roeder

Kontakt:
torsten.roeder@leopoldina.org

Institution:
Leopoldina. Nationale Akademie der Wissenschaften

GND:
1084606364

ORCID:
0000-0001-7043-7820

DOI des Artikels:
10.17175/sb004_001

Nachweis im OPAC der Herzog August Bibliothek:
1037067312

Erstveröffentlichung:
13.03.2019

Lizenz:

Sofern nicht anders angegeben (cc) BY-SA

Medienlizenzen:
Medienrechte liegen bei den Autoren

Letzte Überprüfung aller Verweise:
12.03.2019

GND-Verschlagwortung:
Datenanalyse | Graphdatenbank | Methode | Musikwissenschaft | XML |

Zitierweise:
Torsten Roeder: Referenzielle Vagheit und Varianz in Texten über Musik. Annäherungen zwischen Textkodierung und graphbasierten Analysemodellen. In: Die Modellierung des Zweifels – Schlüsselideen und -konzepte zur graphbasierten Modellierung von Unsicherheiten. Hg. von Andreas Kuczera, Thorsten Wübbena und Thomas Kollatz. 2019 (= Sonderband der Zeitschrift für digitale Geisteswissenschaften, 4). PDF Format ohne Paginierung. Als text/html abrufbar unter DOI: 10.17175/sb004_001.

Torsten Roeder

Referenzielle Vagheit und Varianz in Texten über Musik. Annäherungen zwischen Textkodierung und graphbasierten Analysemodellen

Abstracts

Dieser Beitrag unternimmt einen Brückenschlag zwischen XML und Graphentechnologien zur Untersuchung von Referenzen in einem historischen Textkorpus. Im Hinblick auf sowohl unscharfe als auch unsichere Angaben wird darin die Frage verfolgt, inwieweit Graphenabfragen die Analyse von Referenzialität erleichtern können. Als Beispiel dient hier ein rezeptionsgeschichtliches Forschungsprojekt, das musikkritische Texte des 19. Jahrhunderts zur Messa da Requiem von Giuseppe Verdi untersucht.

This paper builds a bridge between XML and graphing technologies to examine references in a historical body of text, and explores the extent to which graph queries can facilitate the analysis of referentiality, considering fuzzy information in particular. Examples are chosen form a history research project which examines music critiques of the 19th century on the Messa da Requiem by Giuseppe Verdi.

1. Einführung

1.1 Zielrichtung

XML/TEI-P5 bildet in zahlreichen Projekten die Kodierungsgrundlage für die digitale Erfassung historischer Texte. Mithilfe nativer Hypertext-Logik lassen sich aber nicht nur Texte, sondern auch Referenzen zwischen XML-kodierten Texten erfassen. Des Weiteren erleichtert dies den Umgang mit wiederkehrenden semantischen Einheiten: Personen, Orte und Werke lassen sich auf gemeinsame Normdatensätze beziehen, wodurch indirekte Verbindungen zwischen den Texten geschaffen werden.

In einem semantisch angereicherten Textkorpus kann die Untersuchung solch vielfältiger, indirekter Bezüge aufschlussreiche Erkenntnisse – sowohl für die Textsammlung insgesamt als auch für einzelne Texte daraus – befördern. Aufgrund der natürlichsprachlichen Datengrundlage und des damit verbundenen interpretativen Spielraums (der in historischen Sprachstufen oft deutlicher zutage tritt) sind Referenzen allerdings nicht immer eindeutig auf genau eine semantische Einheit abbildbar. Modellierungsansätze, die die Problematik von Unsicherheit, Ambivalenz etc. aufgreifen, existieren zahlreiche. Allerdings geht dies fast zwangsläufig mit einer höheren Komplexität des Datenmodells einher, und die maschinelle Auswertung steht vor der Herausforderung, die in dem Modell repräsentierten Unschärfen auch adäquat interpretieren zu können.

Graphentechnologien können hier in zweierlei Hinsicht eine komplementäre Funktion zu X-Technologien übernehmen. Erstens kann die Analyse von referenziellen Netzwerken – verstanden als Gesamtheit von Hyperlinks zwischen einer Vielzahl von Dokumenten – durch Graphenabfragen profitieren, und zwar ohne dass XML als Datengrundlage vonnöten ist. Zweitens können durch Graphenmodellierung auch Vagheits- und Varianzfaktoren stärker in die Abfragelogik eingebunden werden.[1] Diese beiden Aspekte wird die nachstehende Beschreibung eines Annäherungsprozesses zwischen XML und Graphentechnologien, jeweils in Abschnitt 2 und Abschnitt 3, mit Vorrang verfolgen.

1.2 Thematik des Materials

Das hier zugrundeliegende Material stammt aus einem musikwissenschaftlichen Forschungsprojekt, das sich mit Zeitungstexten zum Verdi-Requiem aus dem deutschsprachigen Raum der Jahre 1874–1878 befasste.[2] Verdis Messa da Requiem, wie sie im Original heißt, stand zeitgeschichtlich im Kontext der nationalstaatlichen Umwälzungen im Europa der 1860er- und 1870er-Jahre: Österreich hatte sich mit Ungarn zur *K. u. K.-Monarchie* verbunden, im Süden Europas waren die Überreste des Kirchenstaates an das junge Königreich Italien gefallen, und seit wenigen Jahren existierte ein Deutsches Kaiserreich. Insbesondere in diesen Ländern strebte man nach einer Konsolidierung des nationalen Selbstverständnisses. 1874 verfasste der italienische Opernkomponist Giuseppe Verdi ein großes, sakrales Werk: die Messa da Requiem, die dem kürzlich verstorbenen italienischen Schriftsteller Alessandro Manzoni – eine der bedeutendsten Figuren des Risorgimento – zugeeignet war, und die am ersten Jahrestag seines Todes in Mailand aufgeführt werden sollte.

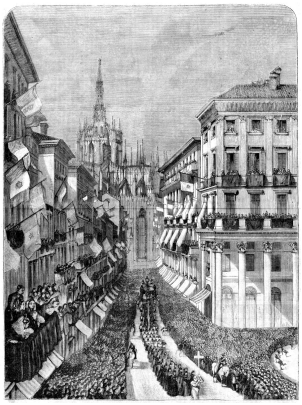


Abb. 1: Begräbnisprozession für Alessandro Manzoni am 29. Mai 1873. In: L'Illustrazione popolare 1873, S. 168–169. [Scan aus dem Besitz des Autors, 29.12.2017.]

[1] Vgl. dazu die Abbildung der TEI-Unschärfe-Semantik in ein Graphenmodell und die dazugehörigen Auswertungen von Kasper / Kuczera 2018 in diesem Sonderband.
[2] Vgl. Roeder 2018a.

Verdis sehr persönliches Widmungswerk – gerne auch Manzoni-Requiem genannt – verbreitete sich nach seiner Uraufführung am 22. Mai 1874 rasch in ganz Europa, insbesondere in Frankreich, in Österreich-Ungarn und auch im Deutschen Reich. Die zeitgenössische Musikkritik ging mit hohen Ansprüchen an die Besprechung des neuen Werkes, das man in dieser Art und Form nicht erwartet hatte: Eine lateinische Totenmesse, ausgerechnet von einem dedizierten Opernkomponisten! Über die Bewertung gingen die Meinungen entsprechend weit auseinander. Begriffen die einen es als durchweg ernsthaftes, dem Anlass des Totengedenkens rundum angemessenes Werk, sahen andere darin hyperemotionale, ja billige Unterhaltungsmusik, die den sakralen Text wie ein profanes Libretto missbrauchte. Dies stand einerseits mit musikästhetischen Fragen und andererseits mit zeitgenössischen Debatten um nationale und religiöse Identität in Verbindung. Diese musikhistorischen Aspekte werden in der folgenden Diskussion um die Modellierung von referenzieller Vagheit und Varianz immer wieder in den Beispielen aufscheinen.

2. Referenzialität

2.1 Textkorpus

Die Grundlage des Forschungsprojektes bildete ein Korpus von ca. 300 deutschsprachigen Texten aus Tageszeitungen und Musikzeitschriften aus den Jahren 1874–1878, welche alle Verdis Messa da Requiem erwähnen, darunter Konzertrezensionen, Werkbesprechungen, Annoncen und andere Textsorten.[3] Das Korpus umfasst insgesamt ca. 1.000.000 Zeichen, von denen die ausführlichen Rezensionen etwa 70 % der gesamten Textmenge ausmachen.[4] Die Presseresonanz war in den Monaten der Erstaufführungen in den jeweiligen Ländern und Städten am stärksten. Dabei war die zeitliche und räumliche Verteilung im Deutschen Reich erkennbar breiter als in Österreich-Ungarn, wo das deutschsprachige Pressewesen klar auf die cisleithanische Hauptstadt Wien zentriert war. Unter den Publikationen im Gebiet des Deutschen Reiches trat indessen die Messestadt Leipzig hervor, da dort fast alle relevanten überregionalen Musikfachzeitschriften des deutschsprachigen Raumes ihren Sitz hatten.[5]

[3] Die Daten sind unter GitHub Torsten Roeder: Verdi Requiem 2018 öffentlich verfügbar.
[4] Für einen Abriss zur Metadatenanalyse vgl. Roeder 2017a sowie Roeder 2018a, S. 83–108.
[5] Darunter zum Beispiel die Allgemeine musikalische Zeitung, das Musikalische Wochenblatt und die Neue Zeitschrift für Musik.

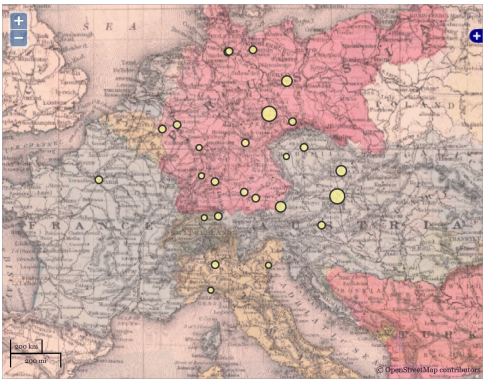


Abb. 2: Geographische Verteilung der Publikationsorte deutschsprachiger Texte zum Verdi-Requiem, 1874–1878. [Grafik: Torsten Roeder. Kartenlayer: Central Europe 1875. In: Mitchell's New Intermediate Geography. Projektion: MapWarper.]

2.2 Datengrundlage

Alle Texte wurden von digitalisierten Vorlagen transkribiert, in TEI-P5 strukturell kodiert und mit semantischen Entitäten mit Normdaten versehen.[6] Letzteres wurde für Personen, Körperschaften, Orte, Werke der Musik, Aufführungen, Bezüge zu anderen Texten innerhalb des Korpus sowie Zitate aus Werken der Literatur vorgenommen. Insgesamt konnten in den Texten etwa 8.000 Bezugspunkte identifiziert werden, die auf 530 unterschiedliche Personen, 142 Aufführungen, 96 Ortschaften und 135 Werke verweisen. Ferner wurden 102 intertextuelle Bezüge identifiziert, von denen sich 68 auf andere Texte im Korpus beziehen. Diese Referenzen können auf ganz verschiedene Art und Weise ausgewertet werden. Dabei stellten sich netzwerk- und graphenbasierte Technologien als außerordentlich hilfreich heraus.[7] Gleichzeitig bewährte sich XML insofern als Datengrundlage, da aus dem TEI-Format, das eben für die Erfassung besonders gut geeignet ist, die relevanten Relationen mithilfe von XQuery oder XSLT ohne nennenswerte Komplikationen extrahiert und in eine Graphenstruktur überführt werden können. Das heißt: die Datengrundlage blieb in diesem Szenario TEI-XML, aber für die Analyse war es hilfreich und zweckdienlich, die Daten in eine andere Struktur zu überführen, da sich mit Graphenabfragen bestimmte Sachverhalte effizienter abfragen ließen. Dies geschah sowohl mit den Metadaten aus dem TEI-Header als auch mit Bezügen in den Texten selbst, wie die beiden nachstehenden Beispiele zeigen werden.[8]

[6] Daraus wurde eine digitale Edition mit Registern und statistischen Visualisierungen hergestellt, die während der Forschungsarbeiten als Arbeitsinstrument verwendet wurde und noch zur Nachnutzung zur Verfügung gestellt werden wird [*Seite ist noch in Vorbereitung*]; vgl. Roeder 2018b.
[7] Einen vergleichbaren Ansatz verfolgt z. B. ein Projekt zur literarischen Inszenierung sozialer Kontakte und Wissenstransfers in Reiseberichten des 19. Jahrhunderts, vgl. Hahn 2015, passim.
[8] Vgl. Roeder 2017b, passim.

2.3 Erstes Analysebeispiel: Geographische Reichweite

Die ersten Beispiele demonstrieren eine Analyse der Metadaten aus dem <teiHeader>. In den Textdateien ist unter <sourceDesc> jeweils erfasst, wo die Texte erschienen sind, und unter <profileDesc>, über welchen Ort die Texte berichteten. Das dichte Netz an Bezügen, das sich dadurch aufspannt – denn besonders in Leipzig und Wien wurde überregional und auch international berichtet – ist zu komplex, als dass man es in Listen- oder Tabellenform analytisch durchdringen könnte. Erst durch eine Umformung der Angaben in Knoten und Kanten nach GraphML und durch Abbildung dieses Graphen mithilfe des Visualisierungstools Gephi wird das Netz der überregionalen Berichterstattung transparent:

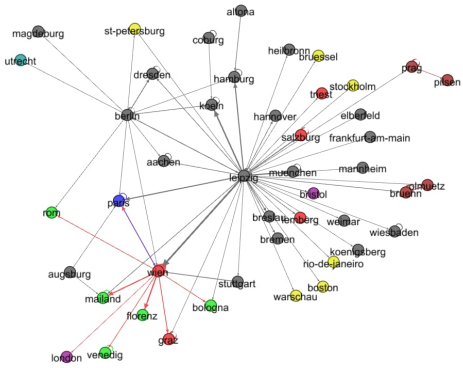


Abb. 3: Publikations- und Aufführungsorte in einer Netzwerkdarstellung. [Grafik: Gephi / Torsten Roeder.]

In dieser Abbildung ist jeder Knoten Ort einer Aufführung oder einer Publikation. Eine Kante zeigt an, dass ein Text, der an Ort A erschienen ist, über eine Aufführung an Ort B berichtete. Es ist erkennbar, dass die Musikfachblätter in Leipzig quasi über alles, auch international, berichteten (sogar über Aufführungen in Nord- und Südamerika wurden kurze Notizen verfasst), während in Wien und Berlin der Fokus auf der nationalen Berichterstattung lag. Auffällig ist außerdem, dass in Wien viele Berichte von Aufführungen in Italien erschienen sind, während die Berliner Kritik lediglich nach Paris blickte: die französische Hauptstadt schien für das Deutsche Reich bedeutsamer zu sein als die meisten italienischen Orte. Dies weist auf eine Segmentierung der Berichterstattung hin: Während Fachblätter international berichteten, beschränkten sich nationale Tageszeitungen auf die Geschehnisse im eigenen Land und ausgewählten Nachbarländern.[9]

[9] Diese These wäre auf allgemeinere Gültigkeit anhand weiterer Analysen mit anderen Textkorpora zu überprüfen. Das Projekt Impresso. Media Monitoring of the Past liefert dafür einen vorstellbaren Rahmen.

2.4 Zweites Analysebeispiel: Intertextualität

Das zweite Beispiel ist indessen aus den Texten selbst generiert und befasst sich mit intertextuellen Beziehungen. Es war üblich, auch die Meinungen anderer Kritiker einzubeziehen und zu diskutieren. In der folgenden Grafik ist jeder Knoten ein einzelner Zeitungsartikel, der die Messa da Requiem thematisiert hat. Jede Kante zeigt eine Bezugnahme an und bedeutet, dass der Artikel am Ausgangsknoten den Artikel am Zielknoten wörtlich zitiert, paraphrasiert oder ausdrücklich erwähnt hat.

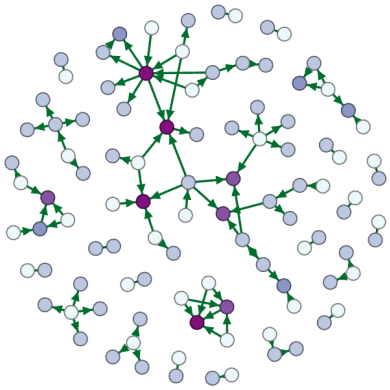


Abb. 4: Intertextuelle Bezüge in einer Netzwerkdarstellung. [Grafik: Gephi / Torsten Roeder.]

Die vier dunkelsten Punkte zeigen die Texte an, auf die am häufigsten Bezug genommen wurde. Diese wurden in Mailand, Paris, Wien, München und Berlin publiziert. Bei näherer Betrachtung wird deutlich, dass die Texte fast immer genannt wurden, um sich von der darin geäußerten Meinung abzugrenzen. Die Diskussion um das Werk ist anhand dieser Texte in ihrer vollen Dynamik nachvollziehbar. Auch die Abfolge der wichtigsten Aufführungen spiegelt sich darin wider: Die Uraufführung in Mailand und die Premiere in Frankreich 1874, sowie die Erstaufführungen in Österreich und im Deutschen Reich 1875. In der darauffolgenden Saison des Jahres 1876 flachte die Intensität der Debatten dann deutlich ab. Entsprechend ist es nicht verwunderlich, dass der chronologisch letzte unter den vielzitierten Texten herangezogen wurde, um sich eben nicht von der darin geäußerten Meinung abzugrenzen, sondern um ihr uneingeschränkt zuzustimmen und damit die Debatten um die Interpretation des Werkes auf sich beruhen zu lassen. Die Extraktion der intertextuellen Referenzen und die Analyse im Graphen beförderten in diesem Fall die Entwicklung einer Diskurshistoriographie.[10]

Es bleibt zu berücksichtigen, dass hier nur explizite Referenzen ausgewertet werden konnten, die vorher durch den Bearbeiter als solche identifiziert worden sind. Nicht vom Autor gekennzeichnete Zitate und vor allem Paraphrasen fließen hier nur in Einzelfällen ein.[11]

[10] Vgl. Roeder 2017b, passim.
[11] Eine solide digitale Unterstützung bei der Paraphrasensuche in Textkorpora stand lange aus, befindet sich derzeit aber in der Entwicklung, vgl. Pöckelmann et al. 2017, passim.

3. Vagheit und Varianz

3.1 Textauswertungen

Anders als die vorhergehenden Analysen orientierte sich das hier gewählte Verfahren zur inhaltlichen Textauswertung an gemeinsamen Bezugspunkten der Texte untereinander. Bezugspunkte bildeten dabei (normierte) Entitäten, wie z.B. Personennamen, Ortsnamen, Werkbezüge etc. (vgl. Abschnitt 2.2). Um beispielsweise alle Kommentare zu dem Beginn des Stückes zu extrahieren, wurden alle Textfragmente aufgelistet, die diese Referenz enthalten. Das Ergebnis sieht man in Abbildung 5, in der alle Kommentare zur Einleitung der Messa da Requiem zusammengestellt sind.

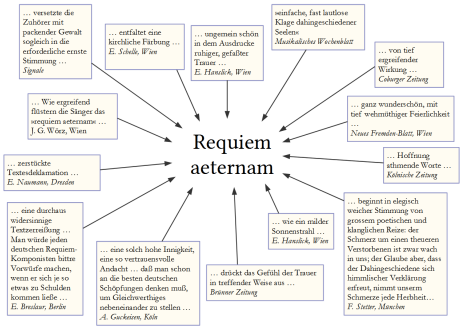


Abb. 5: Kommentare zur Einleitung der Messa da Requiem (Requiem aeternam). [Grafik: Torsten Roeder.]

Die extrahierten Kommentare bilden zusammen ein Spektrum aus allen verfügbaren Meinungen ab. In einer solchen Gesamtdarstellung fällt es leichter, die zahlreichen Einzelmeinungen einzuordnen und zu bewerten, denn ein Hindernis bei einer hermeneutischen Analyse besteht darin, dass von den meisten Autoren nichts über deren Hintergrund bekannt ist, und eine ausgewogene Einschätzung der Einzelmeinungen fast unmöglich ist. Hier indessen geht die Einzelmeinung als Facette in einem großen Meinungsspektrum auf.[12] Ein solches Spektrum ist im Prinzip aus jedem einzelnen Referenzobjekt – einer Person, einem Ort, einem Werk oder einem Teil desselben, einer Aufführung etc. – generierbar.

Der Einsatz von Graphentechnologien wird dann interessant, wenn man mehrere Referenzpunkte in die Betrachtung mit einbeziehen möchte. Innerhalb der hier vorgestellten Textfragmente finden sich nämlich oft weitere Referenzen, etwa Vergleiche zu Werken des klassischen Kanons, wie beispielsweise die Requiem-Kompositionen von Mozart oder Berlioz, aber auch Opern, Oratorien und symphonische Werke, einschließlich der Bezüge zu den jeweiligen Aufführungen, den Beteiligten und den Spielorten. Es wäre von großem Interesse zu

[12]Vgl. Roeder 2017a, passim. und Roeder 2018b, passim.

erfahren, welche Teile der Messa da Requiem wie häufig mit Werken aus den Bereichen Oper, Konzert oder Kirchenmusik verglichen wurden: Dazu sind alle Graphen zu extrahieren, die – innerhalb eines vorher definierten Wortabstandes – eine Referenz zu einem Teil der Messa da Requiem und zu einem anderen Werk der Musik aufweisen. Aus dem Werkregister kann dann leicht die musikalische Gattung in die Abfrage mit einbezogen werden. Bezogen auf diesen Anwendungsfall erscheint das Konzept des *text as a graph* besonders attraktiv: dies würde es ersparen, die Kontexte voneinander abzugrenzen, da eine Abfrage den Textgraphen direkt einbeziehen könnte.[13]

3.2 Unschärfe

Referenzen auf Werke der Musik können sowohl auf ein Werk als Ganzes als auch auf einen Teil oder eine Passage desselben – literarischen Zitaten nicht unähnlich – bezogen sein, aber auch auf eine bestimmte Instrumental- oder Gesangsstimme oder auf einen einzelnen Takt. Sofern aus der Formulierung die gemeinte Stelle unzweifelhaft hervorgeht, verlangt dies lediglich eine Klärung der Referenzpunkte: Technisch ist es auf der Basis etablierter Standards (wie beispielsweise MEI, dem Kodierungsstandard zur Musiknotation) ein Leichtes, sowohl Teile oder Ausschnitte als auch einzelne Elemente eines musikalischen Werkes zu adressieren.[14]

In dem Maße, wie die Präzision der sprachlichen Äußerungen abnimmt, wird eine eindeutige Zuordnung schwieriger. Worauf genau (!) wäre beispielsweise eine vage Referenz wie: »die nachschlagenden Pulse – ein Klangeffect [...] aus Beethovens neunter Symphonie«[15] zu beziehen? Hört (oder liest) man in diesem konkreten Beispiel in die Partitur der ca. 70-minütigen Symphonie hinein, trifft man auf mehrere Stellen, auf welche die Charakteristik der *nachschlagenden Pulse* zutreffen könnte. Hat der Autor nur eine oder alle dieser Stellen gemeint? Durch Interpretation kann immerhin die Plausibilität der verschiedenen Möglichkeiten abgewogen werden, was aber im Datenmodell überhaupt erst einmal abbildbar sein muss.

Etwas anders verhält es sich mit einer Äußerung wie beispielsweise »Der Psalm ist ein herrlicher Satz [...] dessen Harmonik gegen Ende reich und kühn ist« [16]. Diese Aussage ist zunächst nicht mit einem einzelnen Punkt des Musikstückes verbindbar, denn schon der Ausdruck *Harmonik* deutet darauf hin, dass nicht von einem einzelnen Akkord, sondern von einer längeren Passage die Rede ist. Deren Endpunkt ließe sich zwar an den Schluss des Stückes setzen, aber es bleibt unklar, wo genau ihr Beginn anzusetzen wäre. Auch hier ist die Interpretation gefragt, da der harmonische Umschwung im Notentext möglicherweise klar

[13] Vgl. Haentjens Dekker / Birnbaum 2017, passim; Kuczera 2016, passim; Kaufman / Andrews 2016, passim. Die AG Graphentechnologien des deutschsprachigen Verbandes für Digital Humanities 2019 widmete ihre Jahrestagung ausschließlich diesem Thema. Im Übrigen könnte sich bei einem Experiment herausstellen, dass die Abbildung von musikalischer Notation und auch von musikalischer Syntax als Graph viel plausibler ist als die Kodierung in linearer Textform. Im vorliegenden Kontext führt dies leider zu weit.
[14] Im Übrigen könnte sich bei einem Experiment herausstellen, dass die Abbildung von musikalischer Notation und auch von musikalischer Syntax als Graph viel plausibler ist als die Kodierung in linearer Textform. Im vorliegenden Kontext führt dies leider zu weit.
[15] Ambros 1875, S. 3–4.
[16] Anonymus 1875c, S. 1.

erkennbar ist; ist er dies aber nicht, muss der der Rand der Referenz unscharf bleiben. Wäre das hier besprochene Musikstück als Graph referenzierbar, könnte man mit einer variablen Reichweite des Bezuges operieren – also mit einem Graphen vom Ende des Stückes aus, mit offener bzw. variabler Pfadlänge zum Anfang hin, je nachdem, wie viel Spielraum der automatischen Auswertung solch vager Ausdrücke gegeben werden soll.

Ein letztes Beispiel: In zeitgenössischen Aufführungsberichten zur Messa da Requiem wurde oft sehr präzise dokumentiert, welche Teile des Werkes wiederholt worden sind.[17] Dies ist für die Rezeptionsgeschichte der Einzelteile relevant, denn beliebte Stücke wurden später gerne als Einzelausgaben für den Hausgebrauch (insbesondere als Klaviertranskriptionen) herausgegeben. In den Berichten wurde allerdings oft Platz gespart, z.B. wenn Nachrichten aus dem Ausland per Telegramm eintrafen: Dann beschränkte man sich auf summarische Angaben wie »Drei Sätze des Requiems mußten wiederholt werden«[18]. Die möglichen konkreten Bezugspunkte zersplittern hier förmlich in tausende von theoretisch denkbaren Kombinationen. Um dieses Dilemma aufzulösen, ließe sich mithilfe des gesamten Textkorpus ermitteln, welche Zugaben typischerweise gegeben wurden. Sprachlich ausformuliert, klingen die dazu nötigen Abfragen typisch für graphbasierte Queries: Auf welche Stellen beziehen sich andere Texte üblicherweise, die ebenfalls dieses Werk zitieren? Welche Werkteile werden in anderen Texten als Wiederholung genannt, die von derselben Aufführung sprechen?

3.3 Clusterbezüge

Komplexer wird die Situation bei Bezügen auf Werkgruppen, beispielsweise: »wie die Zauberflöte und die Symphonien [Mozarts]«[19]. Im zweiten Teil des Zitats ist kein einzelnes Werk gemeint, sondern ein größerer Gattungskomplex. Ordnet man die Werke im Gesamtregister jeweils einer Gattung zu, kann extrahiert werden, welche Werke des genannten Komponisten in diesem Kontext gemeint sein könnten. Jedoch werden auch hier die Sachverhalte schnell komplexer: Bei Ausdrücken wie »die Requiems von R. Schumann, Brahms, Lachner und Verdi«[20] ist Vorsicht geboten, denn zum Zeitpunkt der Aussage lebten außer Schumann die genannten Komponisten noch, und es war denkbar, dass von diesen noch weitere Requiem-Kompositionen entstehen. Eine Modellierung müsste also mindestens Gattungen und Entstehungszeiten berücksichtigen, da sie mehr Ergebnisse liefern würde, als zu dem historischen Zeitpunkt tatsächlich existieren. Ferner sollte sie – wenn irgend möglich – auch den historischen Kenntnisstand der bis dahin zugeschriebenen Symphonien einbeziehen.[21] Noch abstrakter verhält es sich mit Referenzen, bei denen ein kanonisierter Personalstil gemeint ist. Eine Äußerung »Wie bei Palestrina«[22] wäre demnach sowohl als Referenz auf eine Person anzulegen als auch auf einen musikalischen Stil, der mit dem Werk dieser Person in

[17] Zugaben am Ende eines Konzerts waren nicht üblich, sondern nur direkte Wiederholungen eines gerade gehörten Stückes (ital. ›bis‹, von lat. ›duis‹), bevor es mit dem restlichen Programm weiterging.
[18] Anonymus 1875b, S. 474.
[19] De Lagenevais (Blaze de Bury) 1875, Sp. 478.
[20] Hanslick 1875, S. 2.
[21] Nicht eingeflossen ist hier, dass auch der historische Kenntnisstand der bis dahin Mozart zugeschrieben Symphonien zu berücksichtigen wäre.
[22] Gehring 1874, S. 2.

Beziehung steht und deshalb von diesem abgeleitet wurde.[23] Mit ähnlichen Graphenabfragen wie im vorigen Beispiel (Abschnitt 3.2) wäre dies jedoch zu bewerkstelligen: Besitzt man ein Korpus mit zeittypischen Äußerungen zu Palestrina, ließen sich die kanonischen Werke leicht abfragen.

3.4 Aggregate

Beschreibt man ein Ereignis des Tagesgeschehens, benötigt man dazu oft nur eine vage Angabe, damit der Leser es identifizieren kann. Autoren von Zeitungstexten gehen in dieser Hinsicht aufgrund der meist gebotenen Kürze gern pragmatisch vor und geben nur die Informationen weiter, die der Leser mutmaßlich für das Verständnis benötigt, da sie sich nicht aus dem Kontext ergeben. Beispielsweise benötigt ein Text wie *gestern Abend* in einer Tageszeitung keine explizite Datumsangabe.

Bezieht sich ein Text auf eine musikalische Aufführung, ist zu beachten, dass diese in einem Zusammenspiel von Darbietenden, Handlung, Koordination, Schauplatz und Zuschauern, oder informationstechnisch ausgedrückt: in einem Aggregat aus Personen, Funktionen, Aktionen, Ort und Zeit besteht. Wird ferner eine Aufführung mit derselben Besetzung und am selben Ort wiederholt, entsteht eine Serie von Aufführungen. Allerdings sind Ort, Besetzung und auch das musikalische Programm nicht immer exakt gleich: So können beispielsweise die Mitwirkenden wechseln, oder die Aufführung findet zwar mit denselben Darbietern und in derselben Stadt, aber in einem anderen Konzertsaal statt. Oft verändert sich auch das musikalische Programm. All diese Informationen sind in der Regel nur auf Theaterzetteln vollständig dokumentiert, die oft nicht überliefert sind. Für den Historiker bedeutet dies, dass die Informationen zu einem Ereignis aus verschiedenen Quellen kombiniert werden müssen, die sich aber nur im besten Falle komplementär ergänzen. Bei numerischen Angaben sind zudem oft Fehler im Spiel. Insgesamt besteht bei einer derart fragmentarischen und fehleranfälligen Datenlage die Möglichkeit, dass sich trotz günstiger Quellenlage nicht einmal die Anzahl der Aufführungen genau feststellen lässt.

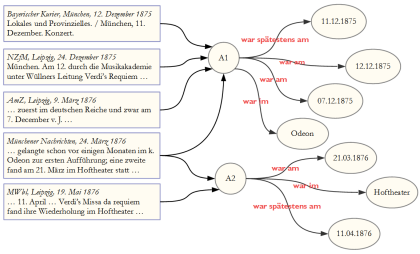


Abb. 6: Bündelung von divergenten Aufführungsdaten. [Grafik: Torsten Roeder.]

[23] Vgl. Kirsch 2013.

Eine Analyse im Graphen kann dabei helfen, die überlieferten Informationen zu einer Aufführung zu aggregieren und gleichzeitig Widersprüche aufzuzeigen. Dazu werden zunächst nur die Referenzen ausgewertet, ohne dass Aufführungen bereits als fester Bezugspunkt definiert sind. Durch eine Abfrage nach Bündeln von Entitäten (z. B. gleiches Datum oder gleicher Ort) entsteht ein klareres Bild, welche Informationen welcher Aufführung zugeordnet werden können und wo eventuell Fehlinformationen vorliegen.[24] In einem zweiten Schritt können dann die Aufführungen als referenzierbare Einheiten angelegt werden.

4. Zusammenfassung

Insgesamt lässt sich bei der Betrachtung der hier aufgeführten Beispiele festhalten, dass zunächst oft Fragen der Modellierung von Referenzen im Zentrum stehen. Der Einsatz einer bestimmten Technologie erscheint daher zunächst nachrangig, solange es um die Datenerfassung geht, und erweist sich bei der späteren Auswertung als umso bedeutsamer. Bei einer großen Vielfalt von möglichen Bezugspunkten, wie sie hier vorliegen, bieten graphbasierte Modelle den Vorteil, dass die Möglichkeiten effizienter auswertbar sind als in einem textbasierten Format.

In den hier angeführten Beispielen trat in XML das Problem der überlappenden Hierarchien nicht auf und braucht deshalb gar nicht erst als Argument gegen XML und für Graphentechnologien ins Feld geführt werden. Vielmehr legen die strukturelle Komplexität der Daten und die zahlreichen semantischen Bezüge es nahe, die analytische Erschließung des Korpus durch anschließende Exportierung in ein Graphenmodell zu erleichtern.

Eine Brücke zwischen XML und Graphentechnologien bildet das Standoff-Markup, mit dem sich Forschende die Möglichkeit schaffen können, eine Vielzahl von wahrscheinlichen oder möglichen Pfaden ohne einen Overload abzubilden und differenziert auszuwerten.[25] Technische Lösungen, die eine Nutzung von graphbasierten Methoden auf der Basis von XML-Markup ermöglichen, sollten deshalb weiter verfolgt und vertieft werden. Dabei besteht glücklicherweise kein Widerspruch darin, XML weiterhin als Textauszeichnungssyntax zu verwenden und für die analytische Auswertung die strukturellen Vorteile der Graphenmodellierung zu nutzen.

[24] Vgl. Efer 2018.
[25] Vgl. Neill / Kuczera 2019.

Abb. 7: Aus der Wiener Satirezeitschrift Kikeriki: [links:] »Was gibt's denn heute im Opernhause? – Ein Requiem«. [rechts:] »Was gibt's denn heute in der Michaeler-Kirche? – Es singen dort [die OpernsängerInnen] der Walter, die Ehnn und die Gindele«. [Bild: Anonymus 1875a, Bearbeitung: Torsten Roeder.]

Dass Grenzen zwischen wohldefinierten Sphären durchlässig sein können (und dass es anders vielleicht gar nicht vorstellbar ist!), zeigt sich auch an der Konzertkultur des 19. Jahrhunderts. Denn die Musikliebhaber strömten sowohl in die Kirchen, um dort Arien und Rezitativen zu lauschen, als auch in die Opernhäuser, um dort Giuseppe Verdis Messa da Requiem zu erleben.

Primärquellen

Anonymus (1875a): Musikalische Verwirrungen. In: Kikeriki 15 (1875), H. 48, S. 3. Karikatur vom 17.06.1875. [online] [Nachweis im GBV]

Anonymus (1875b): Verdi's Requiem. In: Signale für die musikalische Welt 33 (1875), H. 30, S. 474. Artikel vom Juni 1875. [online] [Nachweis im GBV]

Anonymus (1875c): Lokales und Provinzielles, München 11. Dezember. In: Bayerischer Kurier. Zweites Blatt 19 (1875), H. 343–344, S. 1. Artikel vom 12.–13.12.1875. [online] [Nachweis im GBV]

August Wilhelm Ambros: Verdi's Requiem. Aufführung im k. k. Hofoperntheater. In: Wiener Abendpost 132 (1875), S. 3–4. Feuilleton vom 12.06.1875. [online] [Nachweis im GBV]

De Lagenevais (Henri Blaze de Bury): Neuer Pariser Musikbericht. In: Allgemeine musikalische Zeitung 30 (1875), Sp. 477–478. Artikel vom 28.07.1875. [online] [Nachweis im GBV]

Franz Gehring: Verdi's Requiem für Manzoni und seine Kritiker. In: Deutsche Zeitung 910 (1874), S. 1–2. Feuilleton vom 16.07.1874. [online] [Nachweis im GBV]

Eduard Hanslick: Verdi in Wien. In: Neue Freie Presse. Morgenblatt 3889 (1875), S. 1–3. Feuilleton vom 24.06.1875. [online] [Nachweis im GBV]

Sekundärliteratur

Dieser Artikel ist noch in Vorbereitung. Thomas Efer: Graphbasierte Modellierung von Faktenprovenienz als Grundlage für die Dokumentation von Zweifel und die Auflösung von Widersprüchen. In: Die Modellierung des Zweifels – Schlüsselideen und -konzepte zur graphbasierten Modellierung von Unsicherheiten. Hg. von Andreas Kuczera / Thorsten Wübbena / Thomas Kollatz. (Tagung, 19.–20.01.2018, Mainz) Wolfenbüttel 2019. (= Zeitschrift für digitale Geisteswissenschaften / Sonderbände, 4) DOI: 10.17175/sb004_011

Carolin Hahn: Auch ich in Rom! Die literarische Inszenierung sozialer Kontakte und Wissenstransfers in deutschsprachigen Reiseberichten Anfang des 19. Jahrhunderts. In: DHd 2015. Von Daten zu Erkenntnissen. Book of Abstracts. Poster. (DHd: 2, Graz, 23.–27. 02.2015) Graz 2015, S. 134–135. PDF [online] und Poster [online]

Ronald Haentjens Dekker / David J. Birnbaum: It's more than just overlap: Text As Graph. In: Proceedings of Balisage: The Markup Conference 2017. (Balisage: 19, Washington, DC, 01.–04.08.2017) Rockville, MD 2017. (= Balisage Series on Markup Technologies, 19) DOI: 10.4242/BalisageVol19.Dekker01

Dieser Artikel ist noch in Vorbereitung. Dominik Kasper / Andreas Kuczera: Modellierung von Zweifel – Vorbild TEI im Graphen. In: Die Modellierung des Zweifels – Schlüsselideen und -konzepte zur graphbasierten Modellierung von Unsicherheiten. Hg. von Andreas Kuczera / Thorsten Wübbena / Thomas Kollatz. (Tagung, 19.–20.01.2018, Mainz) Wolfenbüttel 2019. (= Zeitschrift für digitale Geisteswissenschaften / Sonderbände, 4) DOI: 10.17175/sb004_003

Sascha Kaufmann / Tara Lee Andrews: Bearbeitung und Annotation historischer Texte mittels Graph-Datenbanken am Beispiel der Chronik des Matthias von Edessa. In: Modellierung, Vernetzung, Visualisierung. Die Digital Humanities als fächerübergreifendes Forschungsparadigma. DHd 2016. Konferenzabstracts. Hg. von Elisabeth Burr. (DHd: 3, Leipzig, 07.–12.03.2016) Leipzig, S. 174–176. PDF. [online] [Nachweis im GBV]

Winfried Kirsch: Kirchenmusikreform, Cäcilianismus und Palestrina-Renaissance. In: Geschichte der Kirchenmusik. Hrsg. von Wolfgang Hochstein und Christoph Krummacher. 4 Bde. Laaber 2011-2014. Bd. 3: Das 19. und frühe 20. Jahrhundert: historisches Bewusstsein und neue Aufbrüche (2013), S. 56–71. (= Enzyklopädie der Kirchenmusik, 1,3) [Nachweis im GBV]

Andreas Kuczera: Digital Editions beyond XML – Graph-based Digital Editions. PDF. [online] In: HistoInformatics 2016 – The 3rd HistoInformatics Workshop. Proceedings of the 3rd HistoInformatics Workshop on Computational History. Hg. von Marten Düring / Adam Jatowt / Johannes Preiser-Kappeller / Antal van Den Bosch. (HistoInformatics: 3, Krakau, 07.11.2016) Aachen 2016, S. 37–46. (= CEUR workshop proceedings, 1632) [online]

Dieser Artikel ist noch in Vorbereitung. Iian Neill / Andreas Kuczera: The Codex – an Atlas of Relations. In: Die Modellierung des Zweifels – Schlüsselideen und -konzepte zur graphbasierten Modellierung von Unsicherheiten. Hg. von Andreas Kuczera / Thorsten Wübbena / Thomas Kollatz. (Tagung, 19.–20.01.2018, Mainz) Wolfenbüttel 2019. (= Zeitschrift für digitale Geisteswissenschaften / Sonderbände, 4) DOI: 10.17175/sb004_008

Begräbnisprozession für Alessandro Manzoni am 29. Mai 1873. In: L'Illustrazione popolare 11 (1873). Artikel vom 13.07.1873, S. 168–169. [Nachweis im GBV]

Marcus Pöckelmann / Jörg Ritter / Eva Wöckener-Gade / Charlotte Schubert: Paraphrasensuche mittels word2vec und der Word Mover's Distance im Altgriechischen. In: Digital Classics Online 3 (2017), H. 3, S. 24–36. DOI: 10.11588/dco.2017.0.40185

Torsten Roeder (2017a): Approaches to the German reception of Verdi's Messa da Requiem through metadata analysis and „horizontal reading". In: Arti musices 48 (2017), H. 2, S. 267–279. DOI: 10.21857/y26kec37k9 [Nachweis im GBV]

Torsten Roeder (2017b): Experimente mit Gephi: Beziehungen im Textkorpus visualisieren. In: Digital Humanities am DHIP. Beitrag vom 28.09.2017. [online]

Torsten Roeder (2018a): Die Rezeption der Messa da Requiem von Giuseppe Verdi im deutschsprachigen Raum 1874–1878. Würzburg 2018. URN: urn:nbn:de:bvb:20-opus-156591

Torsten Roeder (2018b): Horizontales Lesen: Das Verdi-Requiem und die deutsche Kritik. In: Kritik der digitalen Vernunft. DHd 2018. Konferenzabstracts. Hg. von Georg Vogeler. (DHd: 5, Köln, 26.02.–02.03.2018), Köln 2018, S. 232–234. PDF. [online]

Abbildungslegenden und -nachweise

Abb. 1: Begräbnisprozession für Alessandro Manzoni am 29. Mai 1873. In: L'Illustrazione popolare 1873, S. 168–169. [Scan aus dem Besitz des Autors, 29.12.2017.]

Abb. 2: Geographische Verteilung der Publikationsorte deutschsprachiger Texte zum Verdi-Requiem, 1874–1878. [Grafik: Torsten Roeder. Kartenlayer: Central Europe 1875. In: Mitchell's New Intermediate Geography. Projektion: MapWarper.]

Abb. 3: Publikations- und Aufführungsorte in einer Netzwerkdarstellung. [Grafik: Gephi / Torsten Roeder.]

Abb. 4: Intertextuelle Bezüge in einer Netzwerkdarstellung. [Grafik: Gephi / Torsten Roeder.]

Abb. 5: Kommentare zur Einleitung der Messa da Requiem (Requiem aeternam). [Grafik: Torsten Roeder.]

Abb. 6: Bündelung von divergenten Aufführungsdaten. [Grafik: Torsten Roeder.]

Abb. 7: Aus der Wiener Satirezeitschrift Kikeriki: [links:] »Was gibt's denn heute im Opernhause? – Ein Requiem«. [rechts:] »Was gibt's denn heute in der Michaeler-Kirche? – Es singen dort [die OpernsängerInnen] der Walter, die Ehnn und die Gindele«. [Bild: Anonymus 1875a, Bearbeitung: Torsten Roeder.]

ZfdG

Zeitschrift für digitale Geisteswissenschaften

Beitrag aus:
Sonderband 4 der ZfdG: Die Modellierung des Zweifels – Schlüsselideen und -konzepte zur graphbasierten Modellierung von Unsicherheiten. Hg. von Andreas Kuczera, Thorsten Wübbena und Thomas Kollatz. 2019. DOI: 10.17175/sb004

Titel:
Modellierung von Entzifferungshypothesen in einem digitalen Zeichenkatalog für die Maya-Schrift

Autor/in:
Franziska Diehr

Kontakt: diehr@sub.uni-goettingen.de
Institution: Niedersächsische Staats- und Universitätsbibliothek Göttingen
GND: 1155094263

Autor/in:
Sven Gronemeyer

Kontakt: sgronemeyer@uni-bonn.de
Institution: Rheinische Friedrich-Wilhelms-Universität, Abteilung für Altamerikanistik
GND: 1155600487

Autor/in:
Christian Prager

Kontakt: cprager@uni-bonn.de
Institution: Rheinische Friedrich-Wilhelms-Universität, Abteilung für Altamerikanistik
GND: 139962859 ORCID: 0000-0002-7208-9417

Autor/in:
Elisabeth Wagner

Kontakt: ewagner@uni-bonn.de
Institution: Rheinische Friedrich-Wilhelms-Universität, Abteilung für Altamerikanistik
GND: 1169848311

Autor/in:
Katja Diederichs

Kontakt: katja.diederichs@uni-bonn.de
Institution: Rheinische Friedrich-Wilhelms-Universität, Abteilung für Altamerikanistik
GND: 1169862977 ORCID: 0000-0003-3409-9902

Autor/in:
Maximilian Brodhun

Kontakt: brodhun@sub.uni-goettingen.de
Institution: Niedersächsische Staats- und Universitätsbibliothek Göttingen
GND: 1169954073

Autor/in:
Uwe Sikora

Kontakt: sikora@sub.uni-goettingen.de
Institution: Niedersächsische Staats- und Universitätsbibliothek Göttingen
GND: 1155094360

Autor/in:
Nikolai Grube

Kontakt: ngrube@uni-bonn.de
Institution: Rheinische Friedrich-Wilhelms-Universität, Abteilung für Altamerikanistik
GND: 110726553

DOI des Artikels:
10.17175/sb004_002

Nachweis im OPAC der Herzog August Bibliothek:
1031308806

Erstveröffentlichung:
16.01.2019

Lizenz:

Sofern nicht anders angegeben 

Medienlizenzen:
Medienrechte liegen bei den Autoren

Letzte Überprüfung aller Verweise:
16.01.2019

GND-Verschlagwortung:
Maya-Schrift | Wörterbuch | Datenbank | Elektronische Publikation |

Zitierweise:
Franziska Diehr, Sven Gronemeyer, Christian Prager, Elisabeth Wagner, Katja Diederichs, Maximilian Brodhun, Uwe Sikora: Modellierung von Entzifferungshypothesen in einem digitalen Zeichenkatalog für die Maya-Schrift. In: Die Modellierung des Zweifels – Schlüsselideen und -konzepte zur graphbasierten Modellierung von Unsicherheiten. Hg. von Andreas Kuczera, Thorsten Wübbena und Thomas Kollatz. 2019 (= Sonderband der Zeitschrift für digitale Geisteswissenschaften, 4). PDF Format ohne Paginierung. Als text/html abrufbar unter DOI: 10.17175/sb004_002.

Franziska Diehr, Sven Gronemeyer, Christian Prager, Elisabeth Wagner, Katja Diederichs, Maximilian Brodhun, Uwe Sikora

Modellierung von Entzifferungshypothesen in einem digitalen Zeichenkatalog für die Maya-Schrift

Abstracts

Die einzigartigen Eigenschaften der Maya-Schrift stellen die Forschung vor besondere Herausforderungen, aus denen widersprüchliche und zweifelhafte Entzifferungshypothesen hervorgehen. Das Projekt ›Textdatenbank und Wörterbuch des Klassischen Maya‹ verfolgt das Ziel, ein maschinenlesbares Korpus aller Maya-Texte zu erstellen und anhand dessen ein Wörterbuch zu kompilieren. Zur Entzifferung der Schrift ist ein Inventar aller Hieroglyphen ein unverzichtbares Instrument. Dieser Beitrag schildert die Entwicklung des digitalen Zeichenkatalogs unter besonderer Berücksichtigung der Modellierung von Entzifferungshypothesen und deren qualitative Bewertung. Weiterhin thematisiert der Beitrag die Rolle der Wissensrepräsentation in Digital Humanities-Projekten.

The unique characteristics of Maya writing pose special challenges for research, from which contradictory and doubtful deciphering hypotheses emerge. The project ›Text Database and Dictionary of Classic Mayan‹ aims at creating a machine-readable corpus of all Maya texts and compiling a dictionary on this basis. An inventory of all hieroglyphs is an indispensable instrument for further decipherment. This paper describes the development of the digital Sign Catalogue with special consideration of the modelling of decipherment hypotheses and their qualitative evaluation. Furthermore, the article emphasises the role of knowledge representation in Digital Humanities projects.

1. Modellierung des Zweifels im Maya-Wörterbuch-Projekt

Die noch nicht vollständig entzifferte Schrift der vorspanischen Kultur der Maya ist Forschungsgegenstand des Akademievorhabens ›Textdatenbank und Wörterbuch des Klassischen Maya‹. Projektziele sind die Erschließung der bislang rund 10.000 bekannten Texte und ihrer Schriftträger in einem maschinenlesbaren Korpus und die Kompilation eines darauf aufbauenden Wörterbuchs, das den gesamten Sprachschatz und dessen Verwendung in der Schrift abbildet.[1] Gefördert wird das bis voraussichtlich 2028 angelegte Langzeitvorhaben von der Nordrhein-Westfälischen Akademie der Wissenschaften und der Künste sowie der Union der deutschen Akademien der Wissenschaften. Die an der Universität Bonn angesiedelte Arbeitsstelle kooperiert mit der Niedersächsischen Staats- und Universitätsbibliothek Göttingen. Seit 2014 werden in interdisziplinärer Zusammenarbeit die Hieroglyphentexte mit geisteswissenschaftlichen Methoden und informationstechnologischen Werkzeugen aufbereitet, ausgewertet und interpretiert.[2]

[1] Prager 2014a.
[2] Prager 2014b.

Ein Ergebnis dieser Zusammenarbeit und gleichzeitig auch bedeutender Meilenstein des Projekts ist der digitale Zeichenkatalog, dem sich der vorliegende Beitrag widmet. Fu#r die Erstellung des Textkorpus und des Wörterbuchs einer nicht entzifferten Sprache ist ein Zeichenkatalog, ein Inventar aller verwendeten Schriftzeichen, ein unverzichtbares Instrument. Gleichzeitig bildet der Katalog auch die Kernkomponente zur Generierung eines maschinenlesbaren Texts. Das Projekt beschäftigt sich mit einer nur teilweise entzifferten Schrift, deren Erforschung bisher zahlreiche Vermutungen und Interpretationen über Lesung der Schriftzeichen hervorbrachte. Für eine heuristische Entzifferungsarbeit sind diese Lesungshypothesen bei Untersuchung der Schrift einzubeziehen. Bei der Entwicklung des Zeichenkatalogs entstand ein System zur qualitativen Bewertung von Entzifferungshypothesen und Lesungsvorschlägen, welches unter anderem Hilfestellung bei der linguistischen Analyse der Texte bietet (siehe Abschnitt 5). Die Herausforderung bei der Modellierung dieses Systems bestand darin, sowohl mehrere Lesungsvorschläge abzubilden als auch jene Faktoren zu bestimmen, die zur Formulierung der jeweiligen Entzifferungshypothese geführt haben. Ziel der Wissensrepräsentation ist auch die explizite Beschreibung der dabei angewandten Methoden. Dies ermöglicht dem Wissenschaftler anhand des Modells Analysen durchzuführen und Schlussfolgerungen zu ziehen.

Die Modellierung unsicherer Entzifferungen und deren Bewertung fungieren als konkrete Beispiele für die Repräsentation komplexer Wissensobjekte mit interpretatorischem Charakter, wie sie typisch für geisteswissenschaftliche Forschungsdaten sind. In diesem Zusammenhang hebt der Beitrag insbesondere auf die Rolle der Wissensrepräsentation in Vorhaben ab, die sich in den digitalen Geisteswissenschaften verorten und berücksichtigt dabei vor allem die Herausforderungen, die sich bei der Modellierung von vagen und unsicheren Informationen stellen.

Um die Schwierigkeiten aufzuzeigen, die sich bei der Entwicklung eines digitalen Zeichenkatalogs für Maya-Schriftzeichen ergeben, beschreibt der folgende Abschnitt den Aufbau und die Charakteristika des Schriftsystems. Anschließend setzen wir uns mit den bisher publizierten Zeichenkatalogen als forschungsgeschichtliche Quellen auseinander und eruieren dabei insbesondere die Probleme, die sich in der Vergangenheit bei der Zeichenklassifikation stellten. Der vierte Teil widmet sich der Entwicklung des Zeichenkatalogs. Die jeweiligen Unterabschnitte beschreiben die einzelnen Schritte des Modellierungsprozesses. Dabei werden Konzepte und Methoden der Wissensrepräsentation definiert und deren Anwendung geschildert. Im Zusammenhang mit der Modellierung unsicherer Lesungen schließt Abschnitt 5 in der Beschreibung des Systems zur qualitativen Bewertung von Entzifferungshypothesen an. Während der 6. Teil die genutzten Techniken aufzeigt, um mittels Zeichenkatalog und TEI-Kodierung ein maschinenlesbares Korpus zu erzeugen, schließt sich der 7. mit der linguistischen Analyse des Texts an und zeichnet den Weg zum Wörterbuch des Klassischen Maya. Weitere Features des Zeichenkatalogs, welche die Untersuchungen zum Schriftsystem unterstützen, werden in Abschnitt 8 vorgestellt. Im vorletzten Abschnitt geben wir Informationen rund um die Möglichkeiten der Nachnutzung der Daten und Forschungsergebnisse. Im Fazit erläutern wir anhand des Zeichenkatalogs wie Modellierung als eine Forschungsmethode der Digital Humanities den Umgang mit Unsicherheiten und Vagheiten in geisteswissenschaftlichen Forschungsdaten ermöglicht.

2. Charakteristika der Mayaschrift

Im Vergleich zu den anderen mesoamerikanischen Schriftsystemen wie Isthmisch oder Aztekisch weist die Mayaschrift einen überdurchschnittlich langen Anwendungszeitraum von rund 2000 Jahren auf. Erste präklassische Textzeugen entstanden bereits im 3. Jh. v. Chr. Ihren schöpferischen Höhepunkt erreichte die Schrifttradition in der Klassischen Periode (100–810 n. Chr.). Durch die Ankunft der spanischen Konquistadoren erfuhren die Maya einen tiefen Einschnitt in ihrer Kultur, was auch die Verwendung ihrer Schrift betraf. So war es ihnen nur noch möglich, ihre Schrift im Verborgenen zu gebrauchen. Im Untergrund schrieben sie noch bis in das späte 17. Jh. weiter, danach bricht die Verwendung der Maya-Schrift vollständig ab und wurde nicht weiter aktiv gepflegt.

Erst die Entdecker und Gelehrten des 18. und insbesondere des 19. Jh. brachten die Maya-Texte wieder ans Licht, wodurch sie immer weiter in den Fokus der Forschung rückten. Bis heute haben sich schätzungsweise 10.000 Textträger erhalten (Abbildung 1), vor allem Monumentalinschriften, die an Gebäuden und auf Steinmonumenten wie Altäre und Stelen angebracht wurden. Auch auf Keramiken und Kleinartefakten wie Schmuck und Alltagsgegenstände sind zahlreiche Texte zu finden. Eine Besonderheit stellen die auf papierähnlichem Material geschriebenen Handschriften dar, von denen leider nur drei einwandfrei als authentisch geltende Exemplare überliefert wurden. Eine vierte, deren Echtheit bisher umstritten war, ist erst jüngst als authentisch bestätigt worden.

Abb. 1: Beispiele für Maya-Hieroglyphen, links La Corona Panel 1, rechts polychrome Keramik unbekannter Herkunft; Fotos: Sven Gronemeyer.

Das Schriftsystem des Klassischen Maya zeichnet sich durch einen ikonischen Charakter aus, weshalb es auch als Hieroglyphenschrift bezeichnet wird. Typologisch handelt es sich um ein logo-syllabisches System, das durch die zwei grundsätzlichen Zeichenklassen der Logo- und Syllabogramme gekennzeichnet ist. Logogramme stehen für konkrete sprachliche Begriffe, wie z.B. **PAKAL** das Wort für ›Schild‹, und verweisen dabei, mit nur wenigen Ausnahmen, immer nur auf ein Denotat. Vokal- und Silbenzeichen repräsentieren zum einen Silbenkomponenten und dienen zum anderen zur Schreibung lexikalischer und grammatischer Morpheme. Sie werden aber auch als phonetische Komplemente von Logogrammen verwendet, wie in **PAKAL-la** . Dies ermöglichte es, Wörter ausschließlich mit Logogrammen oder Silbenzeichen (**pa-ka-la**) zu schreiben, meist wurde aber beides miteinander kombiniert (Abbildung 2).[3]

[3] Diehr et al. 2017, S. 1186.

Abb. 2: Schreibung von Hieroglyphen. John Montgomery: How to Read Maya Hieroglyphs. Hippocrene, New York, NY, 2002.

Die Zeichen wurden derart angeordnet, dass sie nahezu rechteckige Blöcke bildeten. Dabei entspricht ein solcher Hieroglyphenblock wahrscheinlich der emischen Vorstellung eines Wortes. Innerhalb eines Blocks konnten die Graphe auf verschiedenste Weise arrangiert werden: Je nach Platzbedarf und Ästhetik konnten sie miteinander verschmelzen, sich überlappen, infigiert oder gedreht werden. Die Blöcke wurden in der Regel in Doppelkolumnen angeordnet und von links nach rechts und von oben nach unten gelesen. Die so konstruierten Sätze fügen sich zu komplexen Texten zusammen, deren Syntax und Struktur jener der modernen Maya-Sprachen ähneln.[4]

Die Inschriften weisen eine ausgeprägte kalligraphische Komplexität auf, die vor allem durch die Variation der Schreibung von Schriftzeichen entsteht. Dies erlaubte den Schreibern, Texte ästhetisch anspruchsvoll und ohne die Wiederholung von Zeichen zu verfassen. Für einen sprachlichen Ausdruck gibt es nicht nur eine eindeutige graphische Entsprechung, sondern meist mehrere, auch sehr unterschiedliche, Schreibvarianten. Das Syllabogramm **u** konnte entweder in seiner Vollform gebraucht werden oder alternativ nur jeweils eins seiner beiden Segmente. Die Segmente konnten aber auch reduziert oder multipliziert dargestellt werden. Als weiteres Beispiel zeigt das Silbenzeichen **yi** wie einfache Formen zu sogenannten Kopfvarianten transformieren. Selbst die diagnostischen Merkmale (in Abbildung 3 markiert) sind nicht in allen Graphvarianten vorhanden. Es war sogar möglich, andere Graphe einzuschieben bzw. zu umklammern, wie das Beispiel **ma** zeigt (Abbildung 3).

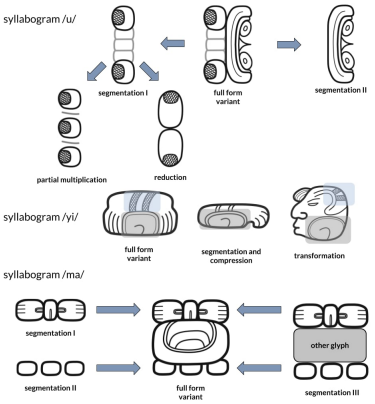


Abb. 3: Graphvarianten. Konzept: Franziska Diehr; Zeichnungen: Christian Prager.

[4] Prager 2014b.

Die Art und Weise wie Varianten gebildet werden und somit die Komplexität der Mayaschrift erst ermöglicht wurde, war bisher nicht Gegenstand systematischer Untersuchungen. Es existieren einige wenige Einzelstudien,[5] die etwa Substitutionen, Rotationsprinzipien oder Symmetrien untersuchen. Wir konnten im Rahmen der Vorarbeiten für den digitalen Zeichenkatalog erstmalig erkennen, dass es allgemeingültige Regeln und Prinzipien für die Bildung von Graphen gibt, die sich über graphotaktische Analysen ableiten und definieren ließen.[6] Insgesamt haben wir 45 einzelne Variationsmöglichkeiten identifiziert, die sich in neun Klassen einteilen lassen (Mono-, Bi-, Tri-, & Variopartite, Division, Animation Head, Animation Figure, Multiplication, Extraction).

Eine weitere Besonderheit der Schrift und Sprache ist die Polyvalenz von Zeichen. Gleich aussehende Glyphen können sowohl als Logogramme oder als Silbenzeichen gelesen werden. Das Zeichen 528 kann als Logogramm **TUN** (›Stein‹) und **CHAHUK** (Name des 19. Tags im Maya-Kalender) oder als Silbe **ku** verwendet werden. Die verwendete Graphvariante gibt dabei keinen Hinweis darauf, welche der Lesungen vorliegt.

Dies verdeutlicht eindrucksvoll, wie frei die Maya-Schreiber bei der Gestaltung ihrer Texte waren und erklärt auch die Herausforderungen, die sich bei der Entzifferung der Schrift stellen, die bis heute noch nicht vollständig entschlüsselt werden konnte. Insbesondere die Variation der Graphe in synchroner wie diachroner Hinsicht macht die Bestimmung des Zeichengraphems und seiner Zuordnung zu einem sprachlichen Ausdruck schwierig. Bis heute konnte die genaue Anzahl von Zeichen und ihren Variationen nicht genau bestimmt werden. Eine Durchsicht publizierter Zeicheninventare weist zwischen 700 bis 1000 sprachlicher Zeichen aus, denen nach eigener Schätzung etwa 3000 Graphe zugeordnet werden können.

3. Zeichenkataloge als Klassifikationsinstrumente

Die Erforschung der Schrift erfolgt auf Grundlage bisher entdeckter Textträger. Viele Textzeugen wurden jedoch im Laufe der Zeit zerstört, sind in Depots von Archiven und Museen verschollen, in Privatsammlungen verborgen oder wurden schlicht noch nicht entdeckt. Es wird also zum einen immer eine geringe Zahl an Zeichen geben, die nicht inventarisiert sind. Zum anderen ist davon auszugehen, dass es in den kommenden Jahren zu weiteren Neuentdeckungen kommen wird. So wurden Anfang 2018 die Ergebnisse einer großangelegten Lidar-Prospektion (Lidar: Light detection and ranging) in Guatemala vorgestellt, die eine wesentlich dichtere Besiedlung im tropischen Tiefland aufzeigen als bislang angenommen,[7] was auch eine größere Anzahl an Textträgern impliziert.

[5] vgl. Beyer 1934a, passim; Beyer 1934b, passim; Beyer 1936, passim; Beyer 1937, passim; Lacadena 1995, S. 204f.
[6] Die Systematik wurde erstmals am 8. April 2016 im Rahmen der Tagung ‚Ägyptologische »Binsen« Weisheiten III' an der Akademie der Wissenschaften und der Literatur in Mainz vorgestellt und wird demnächst in der Tagungsakte publiziert.
[7] Clyens 2018.

Die Neuentdeckung von Inschriften und damit auch der Fund bisher unbekannter Zeichen stellt eine Herausforderung für die Klassifikation der Maya-Schriftzeichen dar. Bisherige Zeicheninventare konnten selbstverständlich immer nur die zur jeweiligen Zeit bekannten Zeichen für den Aufbau ihrer Kataloge berücksichtigen.

In unserem digitalen Zeichenkatalog wollen wir den Fall der Neuentdeckung antizipieren und die damit einhergehende Neuklassifikation von Zeichen als mögliches Szenario berücksichtigen. Daher benötigen wir einen Datenverarbeitungsprozess, der flexibel neue Erkenntnisse einbinden und verarbeiten kann.

Die Erforschung der Maya-Schrift ist, verglichen mit der antiker Sprachen und Schriften, noch relativ jung. Trotz der bereits im späten 19. Jh. gewonnenen Erkenntnis, dass viele Texte mit kalendarischen Daten durchsetzt sind,[8] wurde den Texten ein historischer Charakter und damit eine sprachliche Grundlage abgesprochen.[9] Erst in den 1950er Jahren erkannte Juri Knorosow den logo-syllabischen Charakter der Mayaschrift[10] und konnte die ersten sprachlich abgesicherten Lesungen präsentieren. Bedauerlicherweise blieb seine Arbeit aufgrund des Eisernen Vorhangs noch längere Zeit von der Forschung unbeachtet, obwohl sie 1958 in englischer Übersetzung erschien.[11]

Der Hieroglyphenkatalog von J. Eric S. Thompson[12] enthält etliche Mehrfach- und Falschklassifikationen, aber er dient der Forschung noch heute als Standardreferenz zur Zeichenidentifikation und anschließenden Transliteration als Vorbereitung für die linguistische Untersuchung der Texte, auch wenn Thompson in Ablehnung von Knorosows Arbeiten keine Lesungen angab.[13] Insgesamt wurden bisher neun andere Zeicheninventare vorgelegt. Sie alle weisen fehlerhafte Zuordnungen auf. Besonders problematisch sind dabei die Mehrfachklassifikationen von Zeichen, bei denen mehrere Allographe eines Graphems als verschiedene Zeichen inventarisiert wurden.[14] Die Kataloge sind zumeist einfache Graphinventare. Erst mit Knorosows »Compendio Xcaret«, posthum 1999 herausgegeben,[15] wurden erstmals Lesungen mit den entsprechenden Graphemen verknüpft. Diese basieren allerdings auf dem Forschungsstand der 1960er. Selbst die beiden neuesten Kataloge[16] bilden Lesungen nur unreflektiert ab.

Mit unserem digitalen Zeichenkatalog wollen wir beide Expressionsebenen eines Schriftzeichens, die funktional-sprachliche und die graphemische, berücksichtigen und so modellieren, dass die Zuordnung beider Ebenen zueinander feingranular und flexibel ist.

[8] Morley 1915, passim.
[9] Thompson 1956, S. 169.
[10] Knorosow 1952, passim.
[11] vgl. Knorosow 1958, passim.
[12] vgl. Thompson 1962, passim.
[13] Prager 2014c.
[14] vgl. Grube 1990, passim; Kelley 1962, passim; Kurbjuhn 1989, passim; Riese 2006, passim; Ringle / Smith-Stark 1996, passim.
[15] Knorosow 1999, passim.
[16] vgl. Macri / Looper 2003, passim; Macri / Vail 2009, passim.

Ein weiterer Nachteil traditioneller Zeichenkataloge ist, dass sie durch ihre gedruckte Fassung unveränderbar und somit nicht dynamisch erweiterbar sind. Dies verhindert, dass Fehlklassifikationen korrigiert werden oder neue Relationen zwischen Zeichen etabliert werden können. Auch hier kann ein Zeichenkatalog in digitaler Form Abhilfe schaffen, indem er flexibel auf Änderungen reagieren kann und dabei gleichzeitig persistente Identifikationsmöglichkeiten bietet.

4. Modellierung des digitalen Zeichenkatalogs

Wir entwickeln den digitalen Zeichenkatalog mit dem Ziel, eine vollständige Neuinventarisierung der Maya-Zeichen vorzunehmen und damit eine gesicherte Aussage über die Anzahl bisher bekannter Schriftzeichen zu treffen. Mit dem Zeichenkatalog etablieren wir ein neuartiges Konzept zur Systematisierung und Klassifikation von Schriftzeichen. Die spezifischen Charakteristika des komplexen Schriftsystems des Klassischen Maya werden im Modell explizit repräsentiert: Graphvarianten, multifunktionale Zeichen und multiple Transliterationswerte werden definiert und miteinander in Beziehung gesetzt. Ein besonderes Augenmerk liegt dabei auf den Lesungshypothesen, die im Katalog nicht nur dokumentiert, sondern auch objektiv bewertet und qualitativ eingestuft werden, so dass sie in aufbereiteter Form für spätere Analysen zur Verfügung stehen. Der Katalog dient nicht nur als Werkzeug zur Klassifikation und zur Erforschung der Schrift, weiterhin bildet er auch die Kernkomponente zur Erstellung des Inschriftenkorpus.

Zur Entwicklung eines digital vorgehaltenen Zeichenkatalogs müssen die Schriftzeichen und ihre Spezifika in einem maschinenlesbaren Modell abgebildet werden. Wir verstehen Modellierung als eine Methode der digitalen geisteswissenschaftlichen Forschung, die darauf abzielt, Objekte und das Wissen über sie in einem Datenmodell zu repräsentieren. Im Sinne von Sowa bedeutet dies, die Semantik von Wissensobjekten explizit zu machen und in ein Datenmodell zu übertragen.[17] Als ›Wissensobjekt‹ bezeichnen wir die epistemische Vorstellung von konkreten Entitäten, die sich in einem spezifischen Wissenszusammenhang konstituieren. Wissensobjekte existieren nicht von sich aus. Sie werden durch wissensgenerierende Prozesse erzeugt; sie formen sich aus den Aussagen, Analysen und Interpretationen über konkrete Entitäten, die Gegenstand des Erkenntnisinteresses sind. Als ›Domäne‹ bzw. ›Gegenstandsbereich‹ bezeichnen wir den spezifischen Wissenszusammenhang aus dem sich Fragestellungen an das Objekt ergeben.

Der Prozess der Modellierung von Wissensobjekten und des spezifischen Gegenstandsbereichs lässt sich in folgende Schritte untergliedern: (a) Analyse domänenspezifischer Anforderungen, (b) Wissensrepräsentation mittels konzeptioneller Modellierung und (c) Konstruktion eines maschinenlesbaren Modells.

[17] Sowa 2000, S. 132.

4.1 Analyse domänen-spezifischer Anforderungen

Die Anforderungen an unseren Katalog für die Schriftzeichen des Klassischen Maya wurden mittels intensiven Wissensaustauschs und -transfers zwischen Domänenexperte und Modellierer ermittelt. Dabei wurden konkrete Bedarfe an das Modell, dessen Implementierung in einem Erfassungssystem sowie an die technische Umgebung eruiert. Zur Ermittlung der Anforderungen wurde ein Vorgehen gewählt, das sich an die Methode des Experteninterviews anlehnt. Laut Reinhold stellt diese »insbesondere im Kontext der Modellierung von Forschungsprozessen und Forschungsdaten in den Digital Humanities [...] eine adäquate Methode der Informationsbedarfsanalyse dar, da aufseiten der Forschenden ein hohes Maß an implizitem Wissen zu erwarten ist«.[18] Es sei dabei aber zu beachten, dass von Experten nur Wissen mitgeteilt wird, wenn der Fragende bereits eine hohe Expertise in dem jeweiligen Thema besitzt.[19] Der Prozess fordert damit ein hohes Maß an Einarbeitung in das Domänenwissen auf Seite des Modellierers, der in der Lage sein muss das jeweilige Fachgebiet und den betreffenden Gegenstandsbereich aus disziplinärer Sicht zu beschreiben.

Der Prozess der Anforderungsermittlung stellt den am stärksten interdisziplinär geprägten Bereich in der gemeinsamen Projektarbeit dar. Einer explorativ-hermeneutischen Arbeitsweise entsprechend, gingen wir dabei wie folgt vor: In einem ersten Schritt erfolgte eine Einarbeitung in die grundlegenden Konzepte des Gegenstandsbereichs, also der Linguistik und Sprachforschung. Dabei analysierten wir auch den Gegenstand, den Zeichenkatalog, auf seine Bedeutung als unverzichtbares Hilfsmittel für die Entzifferungsarbeit und seine projekt- und auch disziplinspezifische Funktion. Weiterhin erfolgte eine intensive Auseinandersetzung mit dem Aufbau des Schriftsystems des Klassischen Maya mit dem Ziel dessen Schriftzeichen und vor allem deren Funktion im Schriftsystem explizit im Modell zu repräsentieren.

Die dabei entstandenen Gesprächsprotokolle bildeten die Grundlage für einen Anforderungskatalog, der als Arbeitspapier dazu dient weitere Bedarfe zu ermitteln und vorhandene zu konkretisieren. Anhand dieses Katalogs wurden in einer Phase des intensiven und regelmäßigen Wissensaustauschs einzelne Konzepte besprochen und Fragen zur Funktionsweise der Zeichen und des Schriftsystems erörtert.

4.2 Wissensrepräsentation mittels konzeptioneller Modellierung

Nach unserer Überzeugung stellt der Prozess der Wissensrepräsentation eine hermeneutische Methode dar, die darauf abzielt, ein maschinenlesbares Modell zu konstruieren. Mittels der konzeptionellen Modellierung wird definiert was Sowa ›ontologische Kategorien‹ nennt. Sie bestimmen alles, was in einer Computeranwendung dargestellt werden kann.[20] Die Erstellung eines ontologischen Modells zielt darauf ab, Wissensobjekte,

[18] Reinhold 2015, S. 330.
[19] Flick 2007, S. 218–219.
[20] Sowa 2000, S. 51.

ihre Beziehung zueinander und zu ihrer Domäne explizit zu beschreiben. Die Definition dieser Kategorien ist besonders schwierig, wenn es um vage und unsichere Informationen geht: »any incompleteness, distortions, or restrictions in the framework of categories must inevitably limit the generality of every program and database that uses those categories«.[21] Da ›Wissen‹ über Objekte hinterfragt oder unterschiedlich interpretiert werden kann, ist es notwendig, die verschiedenen Wissensstände im Modell abzubilden, um solchen Verzerrungen entgegenzuwirken und die Wissensbasis genau im Sinne der definierten ontologischen Kategorien zu begrenzen.

Zunächst recherchierten wir in fachspezifischer Literatur und linguistischen Terminologien wie bspw. SIL Glossary of Linguistic Terms[22] nach Definitionen und Konzepten zur Beschreibung von Schriftsystemen und Schriftzeichen. Die Analyse dieses Materials ergab jedoch, dass die meisten Konzepte für unser Modell nicht anwendbar sind, da sie bereits zu stark die Anwendbarkeit in einem konkreten linguistischen Zusammenhang fokussieren. Unser Ziel ist es aber, ontologische Kategorien zur Repräsentation von Schriftzeichen und ihrer Funktion in einem Schriftsystemen zu definieren. Linguistische Kategorien sind dabei nur auf einer Metaebene anwendbar.[23]

Auf der Literaturrecherche und den Ergebnissen der Anforderungsanalyse aufbauend, definierten wir Konzepte und modellierten deren Beziehungen zueinander und zum Gegenstandsbereich in einer Ontologie, die wir als OWL-Schema verfassten. Die Abbildung 4 zeigt das Domain-Model der Ontologie und veranschaulicht die Kernkonzepte und ihre Beziehungen zueinander.

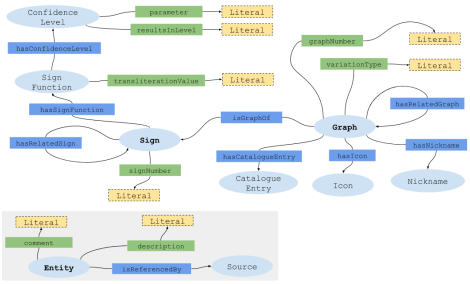


Abb. 4: Domain-Model der Zeichenkatalog-Ontologie. Konzept: Franziska Diehr.

In unserem Verständnis konstituiert sich ein Schriftzeichen aus der Vereinigung zweier Ebenen: einer sprachlich-funktionalen Ebene, die zum einen in Anlehnung an Ferdinand de Saussure[24] die Vorstellung und das Lautbild enthält und zum anderen die konkrete Funktion des Zeichens im Schriftsystem beschreibt, und einer Ebene der graphischen Repräsentation, die alle möglichen Darstellungsformen beinhaltet, welche das Konzept der sprachlich-funktionalen Ebene repräsentieren.

[21] Sowa 2000, S. 51.
[22] Loos et al. 2003, passim.
[23] Diehr et al. 2018, S. 38.
[24] Saussure 1931, S. 28–29.

Betrachten wir als Beispiel ein Maya-Schriftzeichen: Es hat die sprachliche Expression **yi** und erfüllt damit die Funktion eines Silbenzeichens. Für das Syllabogramm **yi** existieren mindestens drei unterschiedliche graphische Repräsentationsformen (siehe Abbildung 3).

Diese Repräsentationsform bezeichnen wir als Graph.[25] Ein Graph ist eine abstrahierte, typisierte Form eines konkret realisierten Individuums eines Schriftzeichens. Das im Katalog erfasste Graph von **yi** in der Variation ›anthropomorphic head variant‹ stellt einen Typ dar, der alle individuellen Schreibvarianten und damit alle konkreten Vorkommen für **yi** in Form einer ›Kopfvariante‹ prototypisch abbildet.

Alle Graphe, die einer gemeinsamen sprachlichen Expression zugeordnet werden, stehen in einer allographen Beziehung zueinander und bilden in ihrer Gesamtheit Varianten des Graphems des Schriftzeichens.[26]

Ein Graph kann nur genau einer sprachlich-funktionalen Expression (idiomcat:Sign) zugeordnet werden. Diese Relation (idiomcat:isGraphOf) ist optional, so dass auch Graphe inventarisiert werden können, die bisher noch keinem ›Sign‹ zugeordnet werden konnten.

Die Funktion eines Zeichens innerhalb des Schriftsystems bezeichnen wir als Zeichenfunktion (idiomcat:SignFunction). Für das Klassische Maya konnten neben den beiden Hauptklassen Logogramm und Syllabogramm (idiomcat:SyllabicReading, siehe Abbildung 4) zwei weitere Funktionstypen identifiziert werden, da wir nach Rogers[27] auch Numerale (idiomcat:NumericFunction) und Diakritika (idiomcat:DiacriticFunction) einbeziehen. Logogramme werden in ihrer Funktion noch weiter unterteilt: Hier unterscheiden wir solche, die einen identifizierten Lautwert haben (idiomcat:LogographicReading) und jene, für die lediglich ein semantisches Feld eingegrenzt werden kann (idiomcat:LogographicMeaning). Zur Klassifikation von Schriftzeichen mag diese Unterscheidung irrelevant sein, da die Funktion des Zeichens für das Schriftsystem ein und dieselbe ist. Jedoch modellieren wir mit der Ontologie einen konkreten Anwendungsfall: Der Zeichenkatalog dient als Hilfsinstrument zur Entzifferung des Schriftsystems. Im Katalog dokumentieren wir auch Lesungshypothesen und bewerten diese qualitativ (idiomcat:ConfidenceLevel). Daher ist eine Differenzierung der Logogramme notwendig, weil jeweils unterschiedliche Bewertungskriterien (idiomcat:parameter) anzusetzen sind. Auch bei der linguistischen Analyse des Korpus ist es sinnvoll zwischen konkreten Lautwerten und einer Bedeutung zu unterscheiden. Der Lautwert bzw. die Bedeutung wird als Transliterationswert (idiomcat:transliterationValue) repräsentiert.

[25] Die vom Projekt gewählte Bezeichnung ›Graph‹ als Konzept der abstrahierten, typisierten Form eines konkret realisierten Schriftzeichens, steht projektintern noch zur Diskussion. Da gemeinhin im linguistischen Diskurs das konkret realisierte Schriftzeichen als ›Graph‹ bezeichnet wird, stellt die abstrahierte Form eine Art Prototyp des Graphs dar. Hier ist noch zu überlegen, wie eine derartige Typisierung in einem Zwischenschritt zwischen dem konkret realisierten Graph und dem Graphem betrachtet werden kann. So könnte etwa die typisierte Form in Abgrenzung zum Graph als Meta-Graph oder Proto-Graph betrachtet werden.
[26] vgl. Bussmann 2002, S. 294.
[27] Rogers 2005, S. 10.

4.3 Konstruktion eines maschinenlesbaren Modells

Nachdem die Objekte definiert, ihre Strukturierung und Beziehungen zueinander in einem konzeptionellen Modell erfasst wurden, erfolgte die Entwicklung des Datenmodells. Konzepte und Strukturen werden in eine maschinenlesbare Form übertragen indem sie in einer entsprechenden Syntax formuliert werden, die sich zur Abbildung des konzeptionellen Modells eignet. Da der vorliegende Zeichenkatalog als Ontologie konzipiert wurde, wird eine Datenstruktur benötigt, die semantische Relationen zwischen eindeutig referenzierbaren Entitäten abbilden kann.

Fu#r die Verwaltung, Erstellung und Präsentation der im Projekt erzeugten Daten nutzen wir die virtuelle Forschungsumgebung TextGrid.[28] Zur Datenerfassung nutzen wir die RDF-Eingabemaske des TextGrid-Labs nach, die wir an unsere projektspezifischen Bedürfnisse angepasst haben. Die Maske rendert anhand eines in Turtle-Syntax hinterlegten RDF-Schemas HTML-Formulare, die zur Erfassung der Daten dienen. Um die Eingabemaske für den Zeichenkatalog nutzbar zu machen, muss das als Ontologie konzipierte Modell in ein RDF-Schema übertragen werden. Während die in OWL repräsentierte Ontologie vor allem zur Dokumentation des konzeptionellen Modells dient, stellt das RDF-Schema die Maschinenlesbarkeit her.

Dies veranschaulicht, dass bei der Datenmodellierung oft auch pragmatische Entscheidungen getroffen werden müssen. Da die Funktionsweise der Eingabemaske auf ein RDF-Schema in Turtle-Syntax ausgerichtet ist, war es nicht möglich die in OWL verfasste Ontologie direkt zu nutzen. Im vorliegenden Fall ist dies jedoch unproblematisch, da das in OWL verfasste Schema verlustfrei in ein RDF-Schema übertragen werden konnte. Die erzeugten Daten werden in einem Triple Store gespeichert. Die Triple-Struktur repräsentiert die Komplexität der ontologischen Beziehungen, die im konzeptionellen Modell definiert wurden. Mittels der Query Language SPARQL sind dementsprechend anspruchsvolle Abfragen möglich.

Neben dem Ziel die Wissensobjekte und den Gegenstandsbereich maschinenlesbar abzubilden, spielt die Herstellung von Interoperabilität zu anderen Systemen und Schemata eine wichtige Rolle bei der Entwicklung von Datenmodellen. Die Übernahme bereits definierter Konzepte ermöglicht die Ausschöpfung des Potenzials vorhandener Metadatenstandards: Ihre Integration verbessert die Auffindbarkeit von Ontologien und erhöht sowohl deren Qualität, als auch die der darauf zugreifenden Anwendungen, da die Wissensbasis kontinuierlich angereichert wird.[29] Für die Nachnutzung und Integration von Ontologien benennt Simperl folgende Schritte: (1) Recherche nach nachnutzbaren Ontologien, (2) integrationsorientierte Evaluation und (3) Integration der Ontologien in das eigene Modell.[30]

[28] vgl. Neuroth et al. 2015, passim.
[29] Gradmann et al. 2013, S. 275–276.
[30] Simperl 2010, S. 246.

Die Auffindbarkeitschancen bereits existierender Ontologien sind dann gering, wenn es sich um fachspezifische Schemata handelt, die häufig nur in den jeweiligen Domänen bekannt sind. Die dort entwickelten Ontologien sind naturgemäß sehr spezifisch und auf einen konkreten Anwendungsfall ausgerichtet, was ihre unmittelbare Nachnutzung und die Integration von Konzepten erschwert.[31] Daher ist es sinnvoll, neben fachspezifischen Ontologien, auch sogenannte Top-Level-Ontologien nachzunutzen. Letztere ermöglichen die Verständigung über allgemeingültige Konzepte, die dementsprechend im eigenen Schema nicht neu definiert werden müssen.[32]

Bei der Recherche nach nachnutzbaren Ontologien für das Zeichenkatalog-Schema fanden wir u.a. die General Ontology for Linguistic Description (GOLD). Mit dem Ziel grundlegende Kategorien und Relationen für die wissenschaftliche Beschreibung von menschlicher Sprache zu definieren, schien GOLD für unsere Zwecke nachnutzbar zu sein. Bei der integrationsorierentierten Evaluation stellte sich heraus, dass GOLD den Fokus auf grammatikalische Regeln mit der Betrachtung der Morphosyntax als Ausgangspunkt legt.[33] In unserem Kontext konnten die in GOLD definierten Konzepte nur eingeschränkt genutzt werden, da unser Konzept eine Metaebene abbildet, die zur Organisation von Schriftzeichen dient.

Dennoch boten einige Konzepte einen guten Anhaltspunkt für die Definition unseres Schemas. Mit der Definition von gold:FeatureStructure als »a kind of information structure, a container or data structure, used to group together qualities or features of some object«[34] ist das Konzept so allgemein gehalten, dass sich unsere Definition von idiomcat:SignFunction als »a feature assessed to a Sign. The nature of the feature is specified by the subclasses«[35] als Subklasse fassen lässt.

Die Evaluation geeigneter Top-Level-Ontologien zeigte, dass das CIDOC Conceptual Reference Model (CIDOC CRM) trotz dessen Fokus' auf die Beschreibung von Prozessen zur Dokumentation von Objekten des kulturellen Erbes viele Metakonzepte beinhaltet, die fu#r den Aufbau unseres Katalogs geeignet sind. Die Klassenhierarchie des Zeichenkatalog-Datenmodells zeigt, dass die meisten Klassen unseres Schemas als Subklassen des CIDOC CRM definiert wurden (Abbildung 5). Als »identifiable expressions in natural language« wurden Sign und Graph als spezifische Subkonzepte von crm:E33_Linguistic_Object gefasst.[36]

[31] So auch bei der vorliegenden Zeichenkatalog-Ontologie, die spezifisch für den Projektkontext entwickelt wurde, auch wenn sie prinzipiell auch auf andere Anwendungsfälle abstrahierbar ist.
[32] Milton / Smith 2004, S. 85.
[33] Farrar / Langendoen 2003, S. 100.
[34] Farrar 2010, passim.
[35] Text Database and Dictionary of Classic Mayan 2017.
[36] ICOM/CIDOC CRM Special Interest Group 2011.

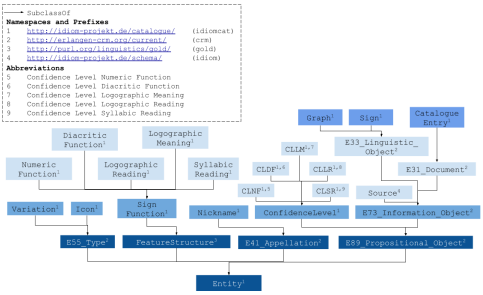


Abb. 5: Klassenhierarchie der Zeichenkatalog-Ontologie. Konzept: Franziska Diehr.

5. Entwicklung eines Systems zur qualitativen Bewertung von Entzifferungshypothesen

In unserem Zeichenkatalog wollen wir publizierte wie auch eigene Lesungshypothesen berücksichtigen, vor allem da die Lesungen für die linguistische Analyse des Textkorpus zur Erstellung des Wörterbuchs verwendet werden.

Auch wenn die Forschung mit fehlerbehafteten Katalogen arbeitet, konnte die Entzifferung der Maya-Schrift in den vergangenen Jahrzehnten deutlich vorangetrieben werden. So legte insbesondere David Stuart in den 1980er Jahren eine bedeutende Studie mit Entzifferungen vor,[37] die dem Prozess neue Impulse brachte. Doch es herrscht keinesfalls Einigkeit über die Lesung aller Zeichen. Die Gründe, warum Zeichen nicht gleichermaßen entziffert sind, können vielfältig sein, wenn etwa ein Zeichen nur ein einziges Mal attestiert ist, oder keine Indikatoren vorliegen, die auf die Phonemik hindeuten. Da Entzifferungen bisher meist an isolierten Beispielen in ›handverlesenen‹ Texten durch den einzelnen Epigraphiker vorgenommen wurden, wurden sie bisher nie anhand eines umfassenden Textkorpus überprüft und abgeglichen, da dieser bisher nicht besteht und erst durch das Projekt ›Textdatenbank und Wörterbuch des Klassischen Maya‹ aufgebaut wird. Nicht nur aus diesen Gründen wurden und werden für etliche Schriftzeichen unterschiedliche Lesungshypothesen vorgelegt. Individuelle Interpretationen des Kontexts oder linguistische Grundlagen tragen ebenso hierzu bei.

5.1 Entstehung und Plausibilität von Lesungshypothesen

Grammatologisch können verschiedene Kategorien von Entzifferungen definiert werden, die sprachlich und semantisch sind, erweitert nach der Zeichenfunktion bei Riese[38] und Entzifferungskriterien bei Houston.[39] Der Lautgehalt von sprachlich gesicherten Zeichen kann in einer Reihe von Kontexten verifiziert werden, häufig mit semantischer und lexikalischer Übereinstimmung bei Morphographen, teilweise auch mit polyvalenten Lesungen. Bei

[37] vgl. Stuart 1987, passim.
[38] vgl. Riese 1971, S. 20–23.
[39] vgl. Houston 2001, S. 9–10.

operationalen Lesungen kann aufgrund bestimmter Indizien auf den Lautwert geschlossen werden, etwa durch vokalharmonische Regeln, phonemische Komplemente oder das semantische Feld. Allerdings können hier auch multiple Lesungsvorschläge auftreten, die keine Polyvalenz aufweisen und eine unterschiedliche Plausibilität haben, basierend auf der individuellen Interpretation des Kontexts oder der Kenntnis des Materials in Ermangelung eines Korpus. Teilweise können aufgrund fehlender Indizien auch nur Teile des Lautgehalts isoliert werden, etwa der Auslaut durch Komplemente. Von nur teilweise oder gar nicht lesbaren Zeichen kann aber oftmals die Semantik und fast immer die Wortklasse eingegrenzt werden, entweder aufgrund des Kontexts oder auch der Ikonizität des Graphs. Etwa ein Drittel des Zeicheninventars widersetzt sich bisher jeder vernünftigen Interpretation, meist sind dies Morphographe und unter diesen überwiegend Substantive.

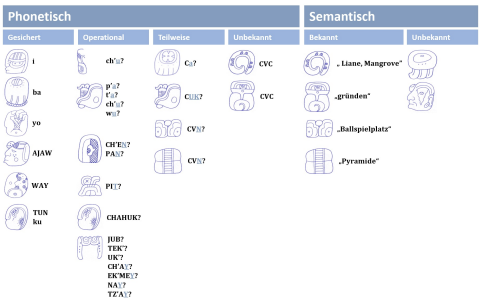


Abb. 6: Entzifferungskategorien. Konzept: Sven Gronemeyer; Zeichnungen der Glyphen: Matthew Looper.

5.2 Definition von Aussagelogiken zur Bestimmung eines Konfidenzlevels

Da jede Zeichenfunktion eine unterschiedliche linguistische Funktion erfüllt, erfordert jede die Zusammenstellung eines eigenen Kriterien-Sets anhand derer eine Hypothese zur Entzifferung eines Zeichens überprüft werden kann. Beispielsweise kann ein Logogramm nicht in einer stammmedialen Position eines Wortes auftauchen, ein Silbenzeichen dagegen schon (Kriterium ›m‹). Die Bewertung einer Entzifferung geschieht qualitativ über die Verknüpfung bestimmter Kriterien in Aussagenlogiken, die in disjunkten numerischen Werten die Plausibilität anzeigen. Jede Zeichenfunktion hat eine unterschiedliche Anzahl an Plausibilitäten, und kommen durch neue Texte neue Vorkommen mit bisher nicht berücksichtigten Kriterien hinzu, können diese ergänzt werden, wodurch gegebenenfalls die Plausibilität um ein oder mehrere Level steigen kann.

Für die Silbenzeichen etwa sind folgende Kriterien definiert worden:

Diese werden in vier Aussagenlogiken zur Festlegung von vier Plausibilitätsstufen kombiniert:

1 (excellent) = d # u # (k # (s # q)) # (a # (f # m) # (s # q)) # ((o # f # m # y) # (g # r # c) # (s # q)) 2 (good) = ((t # o # c) # v # (s # q)) # (a # o # f # m) # (g # r # c)) # (g # r # c) # (s # q)) # (f # m) # (s # q)) 3 (partial) = t # c # l # r ¬ v 4 (weak) = g # m # r

Die Kriterien und Aussagenlogiken wurden auch unter kritischer Betrachtung der bisherigen Entzifferungspraxis konzipiert. Mit einer Rechtfertigungsschrift für einen Inquisitionsprozess des Bischofs von Yucatan, Diego de Landa, haben wir tatsächlich eine zeitgenössische Beschreibung der Maya-Schrift[40] etwa aus dem Jahre 1566, auch wenn Landa von einem Alphabet ausging. Trotzdem finden sich im Manuskript Annotationen zu bestimmten Hieroglyphen mit einem syllabischen Wert, die dann den Schlüssel zur Entzifferung brachten. Und zu den Silbenzeichen gehören natürlich auch die Zeichen, die reine Vokale repräsentieren.

Schon vor Knorosow war bekannt, dass es in den erhaltenen Kodizes einen starken Text-Bild-Bezug gibt. So konnte Paul Schellhas strukturalistisch die Eigennamen verschiedener Götter isolieren.[41] Und dieser Bezug stellte sich auch als hilfreich bei der Kontrolle der Entzifferungen heraus, so wie sie durch Knorosow den Anfang nahmen.[42] Unter den Belegen mit einer syllabischen Annotation im Landa-Manuskript finden sich auch <cu> und <ku> (**ku** und **k'u** in heutiger Transliteration). Diese Beispiele sind also nur durch das Kriterium ›d‹ entziffert, da sie durch einen Zeitzeugen belegt sind, weitere Kriterien wären ausschließlich komplementär.

Beide Zeichen finden sich in initialer Position (Kriterium ›f‹) in Hieroglyphenblöcken, die Tiere benennen, die in der entsprechenden Vignette ebenfalls abgebildet sind: ein Truthahn und ein Geier (Abbildung 7). Ein Abgleich mit kolonialzeitlichen Wörterbüchern ergab die Bezeichnungen *kutz* bzw. *k'uch* für die beiden Tiere. Damit haben wir jeweils auch eine Hypothese zum Anlaut des jeweils zweiten Silbenzeichens (**tzV** und **chV**), ohne jedoch den Vokal zu kennen, denn beide Hieroglyphen sind nicht in Landas Liste vorhanden. Damit wäre das Kriterium ›r‹ erfüllt und Stufe 3 erreicht.

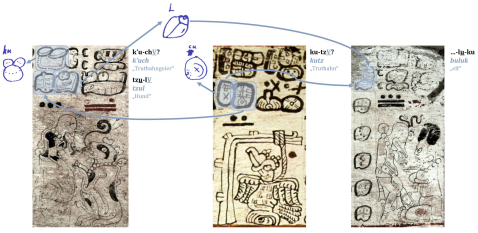


Abb. 7: Entzifferung von Silbenzeichen. Konzept: Sven Gronemeyer; Zeichnungen der Silbenzeichen: Diego de Landa, Relación de las cosas de Yucatán. Fray Di[eg]o de Landa: MDLXVI. Unveröffentlichtes Manuskript. Madrid, Biblioteca de la Real Academia de Historia, 1566; Faksimile Codex Dresden: SLUB online]; Faksimile

[40] Landa 1959, S. 104–106.
[41] vgl. Schellhas 1897, passim.
[42] Knorosow 1958, S. 288–289.

Codex Madrid: Ferdinand Anders Codex Tro-Cortesianus (Codex Madrid): Museo de America Madrid. Codices Selecti Phototypice Impressi 8. Akademische Druck- u. Verlagsanstalt, Graz, 1967.

Das Silbenzeichen **tzV** taucht aber in initialer Wortposition bei der Abbildung eines Hundes auf, der in den Wörterbüchern als *tzul* belegt ist. Damit ist der syllabische Wert **tzu** bestätigt und das Zeichen steigt auf Stufe 2, wobei es in Verbindung mit der Vignette sogar in Stufe 1 (Kriterium ›s‹) aufrückt. Das zweite Zeichen ist somit **lV** ; der Anlaut wird durch die Annotation <L> bei einem sehr ähnlichen Zeichen im Landa-Manuskript bestätigt. Die Beispiele haben bisher Vokalharmonie (**KV1-KV1**) gezeigt, so dass man als Hypothese die Lesung **lu** aufstellen kann, die mit den entsprechenden Kriterien auf Stufe 2 gestellt werden kann.

Das Zeichen taucht an anderer Stelle in einem Block in einer stammmedialen Schreibung auf. Aufgrund struktureller Analogien muss hier eine Zahl stehen, die aus der arithmetischen Struktur dieses Almanachs berechnet werden kann. Statt des entsprechenden Zahlzeichens »11« wurde hier also phonemisch **bu-lu-ku** für *buluk* geschrieben (wobei das zerstörte Silbenzeichen **bu** rekonstruiert wurde). Der angenommene Vokal kann also hier bestätigt werden und das Zeichen ist auf Stufe 1 entziffert.

6. Techniken zur Erzeugung eines maschinenlesbaren Textkorpus

Ein Ziel des Projekts ist es, ein maschinenlesbares Textkorpus aller Inschriften zu erstellen. Aufgrund der komplexen Graphemik, der Zeichenpolyvalenz und des geringen Entzifferungsstatus ist es jedoch nicht möglich den Text in phonemisch transliterierten Werten zu erfassen. Daher ist es auch wenig verwunderlich, dass derzeit kein standardisierter maschinenlesbarer Schriftsatz, wie etwa Unicode, für die Maya-Schrift vorhanden ist. Es gibt zwar Bestrebungen in diese Richtung, [43] diese genügen in ihrer jetzigen Ausführung aber nicht den klassifikatorischen Anforderungen der Maya-Schrift.

Für die Erstellung des Textkorpus möchten wir auf die jeweils konkret verwendeten Graphvarianten verweisen, um Untersuchungen zur Verwendung der Schrift in ihrer räumlich-zeitlichen Entwicklung und ihres Gebrauchs in Abhängigkeit vom Textträger, dessen Aufstellungsort sowie des Textinhalts durchzuführen.

Das Textkorpus kodieren wir in XML unter Benutzung der TEI-P5-Richtlinien.[44] Um die Struktur, die Anordnung der Glyphen zueinander und weitere inschriftenspezifische Phänomene auszuzeichnen, haben wir ein TEI-konformes, anwendungsspezifisches Schema erarbeitet. Statt phonemisch transliterierte Werte zu nutzen, verwenden wir Semantic-Web-Technologie um im XML-Dokument auf die in RDF gespeicherte Ressource zu verweisen. Im Zeichenkatalog ist jede Graphvariante als eigenständige Ressource erfasst und verfügt dadurch über einen URI. Im TEI-kodierten Text erfassen wir jede Glyphe mit dem Element <g> (character

[43] vgl. Pallan Gayol / Anderson 2018, passim.
[44] Text Encoding Initiative 2018, passim.

or glyph) und nutzen das Attribut @ref (»«) um auf den jeweiligen URI zu verweisen (Abbildung 8). Der Text selbst besteht somit aus externen Ressourcen und bildet zusammen mit dem Zeichenkatalog ein ontologisch-verlinktes System.

Abb. 8: Exemplarische Auszeichnung eines Hieroglyphenblocks in TEI/XML. Screenshot aus dem Projektbereich von ›Textdatenbank und Wörterbuch des Klassischen Maya‹ im TextGrid Lab.

Mit dieser Vorgehensweise tragen wir auch der Anforderung Rechnung, verschiedene Entzifferungshypothesen zu berücksichtigen. Da im Textkorpus auf das Graph verwiesen wird, welches im digitalen Katalog potenziell mit mehreren möglichen Lesungsvorschlägen verbunden ist, besteht nun die Möglichkeit die Inschrift unter Berücksichtigung verschiedener Hypothesen zu analysieren. Ein weiterer Vorteil dieser Vorgehensweise ist, dass neue Entzifferungen flexibel eingebunden werden können. Auch im Falle neu formulierter, gesicherter Aussagen müssen diese nicht aufwändig in das Korpus eingearbeitet werden. Durch die Verwendung von URIs kann eindeutig auf die entsprechende Ressource im Zeichenkatalog referenziert werden. An der Kodierung des Korpustexts selbst ändert sich nichts, sie bleibt stabil.

7. Auf dem Weg zum Wörterbuch: linguistische Analyse des Textkorpus

Die linguistische Analyse der Texte ist kein Bestandteil des Korpus und wird durch ein separates Tool realisiert, einer weiteren Komponente unserer virtuellen Arbeitsumgebung (Abbildung 9).[45] Um einen ›lesbaren‹ Text aus dem in TEI kodierten und aus URI-Referenzen bestehenden Korpus zu generieren, ist ein weiterer Prozessierungsschritt notwendig. Dazu liest das Annotationsprogramm die XML-Dokumente ein und fragt die im Zeichenkatalog hinterlegten Transliterationswerte mittels API (Application Programming Interface) ab.

[45] ALMAH (Annotator for the Linguistic Analysis of Maya Hieroglyphs) wird derzeit in Kooperation mit Dr. Cristina Vertan (Universität Hamburg) entwickelt. Vgl. Grube et al. 2018, S. 5–7.

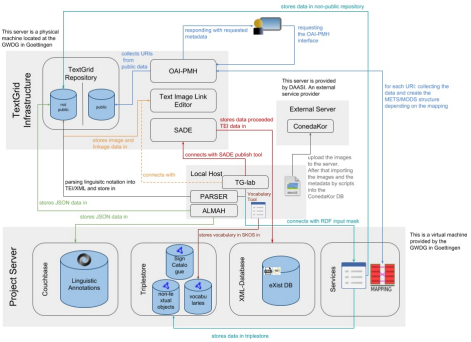


Abb. 9: Systemarchitektur des Projekts ›Textdatenbank und Wörterbuch des Klassischen Maya‹. Konzept: Maximilian Brodhun.

Zur Vorbereitung der linguistischen Analyse wird die Zeichennummer (idiomcat:signNumber, siehe Abbildung 4) als Grundlage einer numerischen Transliteration benutzt, und diese zunächst in eine graphematische Transliteration (idiomcat:transliterationValue) überführt, wofür die Nummer mit dem der Zeichenfunktion zugeordneten Transliterationswert ersetzt wird, z. B. 1.[[528.116]:713] in **u.[[TUN.ni]:K'AL]** . Dies kann wegen Polyvalenz nur semi-automatisch erfolgen, da kontextabhängig die Zeichenfunktion bzw. die korrekte Lesung eruiert werden muss. So kann Zeichen 528 hier nicht die Silbe **ku** (kein sinnvolles Wort) oder den Tagesnamen **CHAHUK** (keine Kalenderangabe) repräsentieren.

An dieser Stelle wird auch das Konfidenzlevel einer Lesungshypothese relevant. Jeder für ein Zeichen getroffene Entzifferungsvorschlag kann im Textkorpus in der Gesamtheit der Belege analysiert werden, ebenso kann jedes Kriterium überprüft werden. Im Idealfall ergeben sich dadurch neue Kriterien, welche die Plausibilität einer Hypothese steigern und vielleicht zu einer vollständigen Entzifferung führen.

Mit den Transliterationswerten kann nun auch in der folgenden graphemischen Transliteration auf Zeichentranspositionen reagiert werden, aus **u.[[TUN.ni]:K'AL]** wird nun **u-K'AL-TUN-ni** . Als kritisch für den weiteren Analyseverlauf ist die folgende phonemische Transliteration.

Hier werden die quasi als Container genutzten Transliterationswerte aus dem Zeichenkatalog kontextuell der korrekten sprachlichen Lesung angepasst. So besitzt das Zeichen 561 **CHAN** üblicherweise die Lesung *chan* - ›Himmel‹. Syllabische Substitutionen **ka-na** in Nordwest-Yucatan zeigen aber, dass das Zeichen dort in einem vernakularen Kontext *káan* ausgesprochen wurde. Der Einfluss lokaler Maya-Sprachen (relevant sind die drei Sprachfamilien Ch'olan, Yukatekan und Tzeltalan) auf die Schriftsprache ist noch nicht systematisch erforscht und mit Ergebnissen der historischen Linguistik abgeglichen worden, Einzelstudien weisen aber auf eine gewisse Permeabilität der Hochsprache

hin.[46] Ab diesem Schritt ist auch die Zeichenfunktion nicht mehr von Bedeutung, sondern die Zeichenverwendung, also ob ein Zeichen zur Schreibung einer Wortwurzel, eines grammatikalischen Morphems oder als phonemisches Komplement gebraucht wird, was auch durch spezifische Operatoren ausgedrückt wird: **u=k'al=tuunni** .

Die Anlage paralleler Analysestränge ist aber nicht nur bei verschiedenen, zu überprüfenden Entzifferungsvorschlägen eine Notwendigkeit. Bestimmte unentzifferte Logogramme tauchen stets mit einem nachgestellten Silbenzeichen auf, bei dem nicht ganz klar ist, ob es ein phonemisches Komplement ist oder ob tatsächlich ein einzelnes Digraph vorliegt, das vielleicht eine andere Lautstruktur hat, als das phonemische Komplement anzeigen mag. Weiter können bestimmte Schreibungen aufgrund der Morphosyntax des Klassischen Maya auf mehr als eine Art analysiert werden. So kann der Krönungs-Ausdruck **HUN-K'AL-ja tu-BAH** als (1) **k'al=huun=ja tu=baah** > *k'al-Ø+huun-[a]j-Ø t.u-baah* oder (2) **k'ahl=ja huun tu=baah** > *k'ahl-[a]j-Ø huun t.u-baah* analysiert werden. Im ersten Fall finden wir ein inchoatives Kompositum mit Ellipse des Agens vor: ›es bekam Stirnband-gehalten an seinen Kopf‹, im zweiten Fall ist das Verb passiviert und *huun* ist das syntaktische Agens: ›es wurde gehalten das Stirnband an seinen Kopf‹. Derartige Fälle können im Korpus jeweils mit anderen syntaktisch ähnlichen Konstruktionen abgefragt und einer weiteren Studie unterzogen werden.

Durch die Verbindung von Zeichenkatalog, Textkorpus und linguistischer Analyse entsteht letztendlich ein dynamischer Text, der je nach Forschungsfrage individuell generiert werden kann. Dieser Ansatz der ontologischen Vernetzung der Komponenten dürfte auch für die Erforschung weiterer nicht entzifferter Schriften von Interesse sein.

Die analytische Annotation erfasst Wortwurzeln, Wortstämme und grammatikalische Morpheme. Damit wird bereits in der Analyse lemmatisiert und die Grundlage für eine semi-automatische Erstellung des Wörterbuches geschaffen, das damit auch alle Belegstellen eines Lemmas listen kann. Die bei der Analyse entstehenden Texte schaffen auch ein dynamisches Wörterbuch, etwa wenn neue Wortbedeutungen im Kontext isoliert werden.

8. Erforschung der Schrift anhand des Zeichenkatalogs

Der Zeichenkatalog ist nicht nur ein Repositorium der Zeichen und Graphe der Mayaschrift oder Hilfsmittel zur Kodierung des Textkorpus, sondern es werden auch weitreichende Informationen zur Ikonizität der Graphe erfasst. So wird der Zeichenkatalog selbst zu einer Infrastruktur für grammatologische Forschungsfragen.

[46] vgl. Lacadena / Wichmann 2002, passim; Lacadena / Wichmann 2005, passim.

8.1 Ikonizität

Wie andere hieroglyphische Schriftsysteme weist auch die Mayaschrift einen hohen Grad an Ikonizität auf. Diese wird im Zeichenkatalog durch das in SKOS [47] modellierte kontrollierte Vokabular ›Icon‹ abgebildet (siehe Abbildung 3), das auf der Basis der fünf Kategorien intuitiver Ontologie (Person, Tier, Pflanze, Artefakt, natürliches Objekt) nach Pascal Boyer[48] entwickelt und verfeinert wurde. Es soll nicht nur das konkrete Designat beschrieben werden, sofern dies Grundlage für die Gestaltung des Graphs ist, wie etwa der Schild für das Logogramm **PAKAL** . Weitere Facetten, die zur Zeichenidentität beitragen, basieren auf ikonographischen Standards, die natürliche Existenz des Designats zu abstrahieren oder immaterielle wie dingliche Eigenschaften sichtbar zu machen. So stellt etwa das Logogramm **MAM** ›Großvater‹ einen menschlichen Kopf dar, den wir weiter als männlich und senil beschreiben können. Oder das Logogramm **EB** ›Treppe‹ ist nicht wie zu erwarten mit der ikonographischen Konvention für die Darstellung von ›Stein‹ markiert, sondern mit der für ›Holz‹. Mit einer feingranularen Charakterisierung der Ikonizität eines Graphs können u.a. auch kognitionswissenschaftliche Fragestellungen an den Katalog gerichtet werden, die dann auch verwendet werden können, um darzustellen, wie die klassischen Maya ihre Welt wahrgenommen haben.

8.2 Graphemrelationen

Zwangsläufig werden bestimmte ikonographische Facetten bei mehr als einem Graph auftauchen. Im Laufe der Schriftgeschichte bildeten bestimmte Ikone die Grundlage für die Bildung neuer Schriftzeichen. Dadurch teilen die Schriftzeichen sich zumeist diagnostische Merkmale, die aber auch unabhängig von deren Ableitung bei verschiedenen Graphen bestehen können. Relationen wie diese können im Zeichenkatalog auf der Graphebene ebenfalls erfasst werden. Neben der Möglichkeit, die Evolution des Schriftsystems damit zu erforschen, ergeben sich konkrete Mehrwerte bei der Entzifferungsarbeit, etwa wenn im Fall einer erodierten Textstelle auch nach anderen Graphen mit einer bestimmten Facette im Ikon gesucht werden kann und so sichergestellt werden kann, dass keine Möglichkeit der Lesung und Rekonstruktion unberücksichtigt bleibt.

Die Einbettung eines Ikons (idiomcat:containsIconicElementsFrom) in ein anderes Graph (Abbildung 10) kann auf konzeptionelle Verwandschaften bei den Graphen hinweisen. In der Maya-Ikonographie wird ›Wind‹ durch ein T-förmiges Element dargestellt, dass sich auch in anderen Graphen inkorporiert wiederfindet.

[47] Miles / Bechhofer 2009, passim.
[48] Boyer 2009, passim.

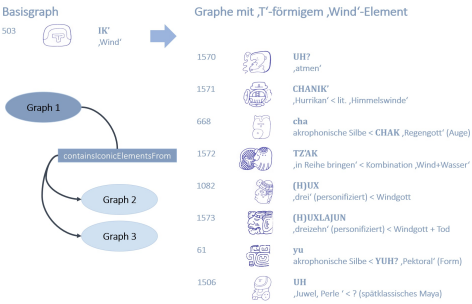


Abb. 10: Graphemrelationen: ikonische Elemente. Konzept: Sven Gronemeyer; Zeichnungen der Glyphen: Ian Graham, Matthew Looper, John Montgomery, David Mora-Marín.

Wie man sehen kann, taucht es häufiger in Graphen auf, die direkt in Verbindung zu Luft und Wind stehen. Im ersten Beispiel für das Logogramm **UH?** findet sich das Wind-Element im oberen Bogen einer mäandernden Linie wieder, die allgemein für einen Luftstrom steht, der den Mund verlässt. In anderen Fällen markiert das Ikon Körperteile von übernatürlichen Akteuren, die in Verbindung zu Wind und atmosphärischen Phänomenen stehen, wie auch das Auge des Regengottes, das akrophonisch für die Silbe **cha** gebraucht wird. Der Fall von **TZ'AK** ist besonders interessant. Die Semantik ist ›etwas in Reihe(nfolge) bringen, ordnen' und das Verb wird etwa in der Zeitzählung benutzt. Es existieren mehrere Digraphe, die zwei Ikone mit einem gegensätzlichen, konsekutiven oder kasaulistischem Konzept kombinieren,[49] hier Wind und Wasser, so wie im Maya-Tiefland häufig auf einen kurzfristig anschwellenden Wind der nachmittäglichem Regen folgt.

Eine andere Methode ist die Derivation von Graphen (Abbildung 11). Hierbei wird die Form des Basisgraphs beibehalten, und entweder Binnenelemente verändert oder weitere Elemente an das Basisgraph angefügt (idiomcat:isDerivedFrom). Häufig besitzt das zugehörige Zeichen, sofern es sich um ein Logogramm handelt, eine verwandte semantische Domäne.

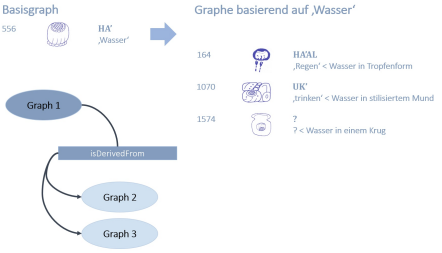


Abb. 11: Graphemrelationen: Derivation. Konzept: Sven Gronemeyer; Zeichnungen der Glyphen: Matthew Looper.

[49] Stuart 2003, passim.

Werden dem Logogramm **HA'** für ›Wasser‹ Reihen von Tropfen angefügt, so wird hieraus das Logogramm **HA'AL** ›Regen‹.[50] Fügt man das Graph an die stilisierte Repräsentation eines Mundes, so entsteht hieraus das Logogramm **UK'** für ›trinken‹. Einen besonders anschaulichen Fall stellt das Logogramm **PAS** ›öffnen, dämmern‹ dar, welches das Ikon von ›Sonne‹ zwischen die von ›Himmel‹ und ›Erde‹ stellt.

In einer Studie zur Innovation von Zeichen, speziell Silbenzeichen mit Lautwert **bV** und **mV**, konnte Alfonso Lacadena[51] nachweisen, dass mit der Übernahme des Vorläufers der Mayaschrift, der eine Sprache aus der Familie des Mije-Sokean kodierte, u.a. neue Zeichen für die dort nicht vorkommenden Laute /b/ und /C'/ geschaffen werden mussten, ein Prozess der im Prinzip nach aktueller Notwendigkeit die gesamte Frühklassik andauerte. Auch wurden nach ähnlichem Prinzip in der Spätklassik weitere Allographe für bestehende Silbenzeichen generiert (Abbildung 12). Hier führen häufig nur Detailänderungen im Graph zu einer neuen Zeichenidentität, so dass wir (nicht nur) für diese Fälle eine weitere Relation geteilter Merkmale anzeigen können (idiomcat:sharesDiagnosticElementsWith).

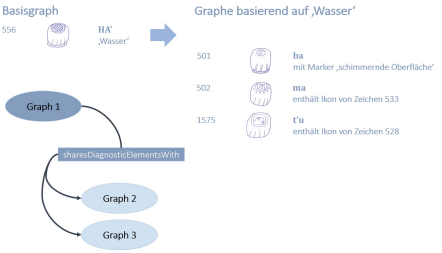


Abb. 12: Graphemrelationen: Diagnostische Elemente. Konzept: Sven Gronemeyer; Zeichnungen der Glyphen: Matthew Looper.

Schriftgeschichtlich können wir beobachten, dass das Logogramm **HA'** bereits in den ältesten frühklassischen Texten auftaucht. Die offenbar daraus abgeleiteten Silbenzeichen teilen sich mit dem Basisgraph die diagnostischen Merkmale einer innen liegenden Kartusche (in der jeweils die Zeichenidentität festgelegt wird), einer darunterliegenden Punktereihe und einer Anzahl von (Doppel-)Bögen am unteren Rand. Wir können also erst einmal rein deskriptiv Ähnlichkeiten abbilden, bevor wir tatsächlich nachweisen können, ob die Relation ›idiomcat:isDerivedFrom‹ tatsächlich zu setzen ist.

[50] Lacadena 2004, S. 88–93.
[51] Lacadena 2010, passim.

8.3 Zeichenrelationen

Für die meisten Zeichen gibt es jeweils mindestens eine Graphvariante, die auf denselben diagnostischen Merkmalen basiert, und welche die Zeichenidentität bestimmt, unbeschadet der Segmentation oder Transformation. Eine geringe Anzahl an Lautwerten wird dabei mit Graphen realisiert, die über keine Verwandtschaft verfügen und auf unterschiedlichen Graphikonen aufbauen (Abbildung 13). Eine Durchsicht der beiden neuesten Kataloge, die auch eine Liste der Silbenzeichen umfassen, zeigt, dass hier die Zahl dieser sogenannten diskreten Allographe üblicherweise sehr beschränkt ist und Variabilität zumeist mit Graphvarianten eines Zeichens erzeugt wird.

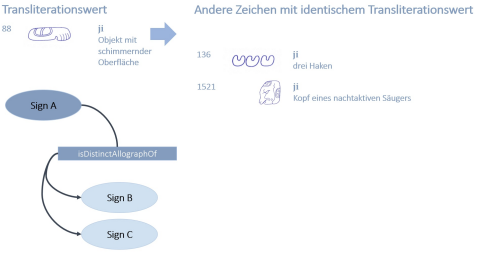


Abb. 13: Allographe Zeichen. Konzept: Sven Gronemeyer; Zeichnungen der Glyphen: Matthew Looper.

Wir finden bei Logogrammen das Phänomen von Homonymen, wobei jede unterschiedliche semantische Domäne üblicherweise heterographisch realisiert wird[52], es aber teilweise zu rebusartigen Schreibungen kommt,[53] was noch einmal die inhaltliche Komponente bei Logogrammen betont. Wie bei den Silbenzeichen auch scheint aber die Anzahl von diskreten Allographen bei gleicher Wortbedeutung recht begrenzt zu sein. Es können dabei unterschiedliche Konnotationen eines Wortes repräsentiert werden, z. B. **AJAW** ›Herrscher‹ abstrakt durch eine Thronbank oder konkret durch eine Kopfvariante mit Stirnband als Zeichen der Autorität, oder *pars pro toto*, etwa **AK** ›Schildkröte‹ einmal durch den Panzer oder den Kopf.

Die dafür im Zeichenkatalog verwendete Relation (idiomcat:isDistinctAllographOf) erlaubt also die Suche nach anderen Zeichen mit demselben Transliterationswert. So können etwa Abfragen zur Innovation von Zeichen in Verbindung mit dem Textkorpus gestellt oder Aussagen über regionale Varianten getätigt werden, wie etwa über eine spezielle Variante von **bi** auf Keramiken aus Xultun.[54]

[52] Gronemeyer 2015, S. 110–112.
[53] Houston 1984, passim.
[54] Krempel / Matteo 2012, S. 145.

8.4 Einbindung forschungsgeschichtlicher Ergebnisse

Die Neuinventarisierung der Schriftzeichen durch das Projekt stellt im Wesentlichen eine Revision des Zeichenkatalogs von J. Eric S. Thompson aus dem Jahre 1962 dar, der als Quasi-Standard in der Forschung weite Verbreitung gefunden hat. So werden etliche ›Thompson-Nummern‹ als Zeichennummern in unserem Katalog weiterverwendet, was ihren Wiedererkennungswert bei Fachkollegen steigert. Neben Thompson steht unser Zeichenkatalog natürlich auch in einer forschungsgeschichtlichen Tradition und Reflexion der anderen Katalogisierungsbemühungen.

Daher wird auf Graphebene eine Konkordanz (idiomcat:hasCatalogueEntry, siehe Abbildung 4) zu den bisherigen Werken angelegt. In dieser werden die Katalognummern, Kommentare und Abbildungen der bereits publizierten Inventare hinterlegt. Damit ist auch erstmalig ein Vergleich aller Kataloge möglich, was auch die Entscheidungen der Revision und die Entscheidung für Graphvarianten in unserem Katalog transparenter macht.

Überhaupt möchten wir mit dem Zeichenkatalog eine Wissensbasis zur Schriftforschung schaffen. Gemäß guter wissenschaftlicher Praxis werden wir dokumentierte Aussagen mit entsprechenden Quellenverweisen versehen. Damit sind sämtliche Informationen nachvollziehbar und überprüfbar. Insbesondere bei den Lesungsvorschlägen ist eine Nennung der entsprechenden Studie unverzichtbar. Alle Entitäten des Schemas können mit Quellen verbunden werden (dct:isReferencedBy, siehe Abbildung 4). Die Bibliographie wird über die Open-Source-Plattform Zotero verwaltet und frei zugänglich zur Verfügung gestellt und sowohl mit dem URI der entsprechenden Ressource als auch mit einer formatierten bibliographischen Angabe bei der entsprechenden Instanz gespeichert.

9. Veröffentlichung der Daten und Möglichkeiten der Nachnutzung

Die Inventarisierung und Klassifikation der Maya-Schriftzeichen ist aktuell in Bearbeitung und wird voraussichtlich im Laufe des Jahres 2018 zum Abschluss kommen. Sobald alle Graphvarianten und Zeichen erfasst wurden, wird konsequent mit der Auszeichnung der Inschriften und der Erstellung des Korpus begonnen. Dokumentation und wissenschaftliche Erschließung der Texte und der Textträger werden sich über die verbleibende Projektlaufzeit bis 2028 erstrecken. Die Daten werden sukzessive auf unserem, sich derzeit in der Konzeption befindlichen, Projektportal zugänglich gemacht. Des Weiteren werden die Korpusdaten auch im TextGrid Repository (TG Rep) veröffentlicht, wo sie mittels OAI-PMH Schnittstelle auch fu#r externe Nutzer abrufbar sind. Die RDF-Daten des Zeichenkatalogs werden ebenfalls über das Portal, das TG Rep und auch mittels eines SPARQL-Endpoints abrufbar. Sämtliche

im Projekt entstandenen Schemata sind im öffentlichen Bereich unseres Git-Repositoriums[55] einsehbar und können unter einer CC BY-4.0 Lizenz genutzt werden. Die Dokumentation der Zeichenkatalog-Ontologie steht auch als Webseite zur Verfügung.[56]

10. Fazit: Modellierung als Forschungsmethode der Digitalen Geisteswissenschaften

Im vorliegenden Beitrag wurde am Beispiel des Zeichenkatalogs für die Maya-Schrift beschrieben, wie Methoden der Wissensrepräsentation genutzt werden können, um vage und unsichere Informationen in einem maschinenlesbaren Modell abzubilden und sie somit für weitere Analysen aufzubereiten. Um Vagheiten und Unsicherheiten zu beschreiben und bewertbar zu machen, wurde ein qualitativer Ansatz gewählt. Die Angabe eines Bewertungslevels erfolgt durch sprachspezifische Kriterien, die mittels Aussagenlogik miteinander kombiniert sind. Durch die Kombination bestimmter Kriterien kann die Plausibilität einer Hypothese geschlussfolgert werden. Es wäre von Interesse diesen Ansatz auch auf andere Anwendungsfälle zu übertragen und zu explorieren, inwieweit die kriterienbasierte Bewertung den Umgang mit zweifelbehafteten Wissensobjekten in der digitalen geisteswissenschaftlichen Forschung ermöglicht.

Modellierungsprozesse bewirken eine Reflexion der Definitionen und Methoden des jeweiligen Fachbereichs, der Fragen an die betreffenden Wissensobjekte und ihren Gegenstandsbereich stellt. Die Repräsentation dieses Wissens in einem maschinenlesbaren Modell erfordert die präzise Definition der Objekte sowie deren Beziehungen zueinander und zu ihrer Wissensbasis. Fachtraditionen werden dabei hinterfragt und angewandte Methoden auf ihre Grundlagen überprüft. So wurde im konkreten Fall der Modellierung der Entzifferungshypothesen deutlich, dass Aussagen über die Plausibilität von Lesungsvorschlägen nur anhand formaler Bewertungskriterien getroffen werden können. Ohne Reflexion der intuitiv-pragmatischen Bewertungsmechanismen hätte kein Evaluationssystem entstehen können, welches auf Grundlage formalisierter und logisch-kategorisierter Parameter operiert.

Modellierung leistet damit einen zentralen Beitrag zu den Forschungsmethoden der Digitalen Geisteswissenschaften, indem sie durch bewusstes Hinterfragen zur Reflexion disziplinspezifischer Wissensgenerierung anregt. Durch den Modellierungsprozess entstehen explizit definierte Objekte und Beziehungen, die in maschinenlesbare Daten und Datenstrukturen transformiert werden. Durch diesen Transformationsprozess stehen die Wissensobjekte und ihre Wissensbasis für weitere Analysen zur Verfügung, seien es hermeneutische oder auch quantitative Verfahren.

[55] Text Database and Dictionary of Classic Mayan 2018.
[56] Text Database and Dictionary of Classic Mayan 2017.

Bibliographische Angaben

Hermann Beyer (1934a): The Position of the Affixes in Maya Writing I. In: Maya Research 1 (1934), H. 1-2, S. 20–29. [Nachweis im GBV]

Hermann Beyer (1934b): The Position of the Affixes in Maya Writing II. In: Maya Research 1 (1934), H. 1-2, S.101–108. [Nachweis im GBV]

Hermann Beyer: The Position of the Affixes in Maya Writing III. In: Maya Research 3 (1936), H. 1, S.102–104. [Nachweis im GBV]

Hermann Beyer: Studies on the Inscriptions of Chichen Itza. In: Contributions to American Anthropology and History 4 (1937), H. 21, S. 29–175. [Nachweis im GBV]

Pascal Boyer: Functional Origins of Religious Concepts: Ontological and Strategic Selection in Evolved Minds. In: The Journal of the Royal Anthropological Institute 6 (2009), H. 2, S. 195–214. [Nachweis im GBV]

Hadumod Bussmann: Lexikon der Sprachwissenschaft. 3. aktualisierte und erweiterte Auflage. Stuttgart 2002. [Nachweis im GBV]

Tom Clyens: Exclusive: Laser Scans Reveal Maya »Megalopolis« Below Guatemalan Jungle. A vast, interconnected network of ancient cities was home to millions more people than previously thought. In: National Geographic. Beitrag vom 01.02.2018. [online]

Franziska Diehr / Sven Gronemeyer / Christian Prager / Maximilian Brodhun / Elisabeth Wagner / Katja Diederichs / Nikolai Grube: Modellierung eines digitalen Zeichenkatalogs für die Hieroglyphen des Klassischen Maya. DOI: 10.18420/in2017_120 In: Informatik 2017. Hg. von Maximilian Eibl / Martin Gaedke. (Informatik 2017, Chemnitz, 25.-29-09-2017) Bonn 2017, S. 1185–1196. Handle: 20.500.12116/3568

Franziska Diehr / Sven Gronemeyer / Christian Prager / Elisabeth Wagner / Maximilian Brodhun / Katja Diederichs / Nikolai Grube: Ein digitaler Zeichenkatalog als Organisationssystem für die noch nicht entzifferte Schrift der Klassischen Maya. In: 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. Hg. von Christian Wartena / Michael Franke-Maier / Ernesto De Luca. (ISKO: 15, Berlin, 30.11.-01.12.2017) Berlin 2018, S. 37–43. Handle: fub188/20535

Scott Farrar: General Ontology for Linguistic Description (GOLD). Department of Linguistics (The LINGUIST List), Indiana University, 2010. [online]

Scott Farrar / D. Terrence Langendoen: A Linguistic Ontology for the Semantic Web. In: GLOT International 7 (2003), H. 3, S. 97–100. [Nachweis im GBV]

Uwe Flick: Qualitative Sozialforschung: Eine Einführung. 7. Auflage, vollständig überarbeitet und erweiterte Neuausgabe. Reinbek 2007. [Nachweis im GBV]

Stefan Gradmann / Evelyn Dröge / Julia Iwanowa / Violeta Trkulja / Steffen Hennicke: Wege zur Integration von Ontologien am Beispiel einer Spezifizierung des Europeana Data Model. In: Informationswissenschaft zwischen virtueller Infrastruktur und materiellen Lebenswelten. Hg. von Hans-Christoph Hobohm. (Internationales Symposiums für Informationswissenschaft. Proceedings des 13. Internationalen Symposiums für Informationswissenschaft. (ISI: 13, Potsdam, 19.-22.03.2013) Glückstadt 2013, S. 273–284. [Nachweis im GBV]

Sven Gronemeyer: Class Struggle: Towards a Better Understanding of Maya Writing Using Comparative Graphematics. In: On Methods: How We Know What We Think We Know About the Maya. Hg. von Harri Kettunen / Christophe Helmke. (European Maya Conference: 17, Helsinki, 09-15.12.2012) München 2015, S.101–117. # (= Acta Mesoamericana, 28) [Nachweis im GBV]

Nikolai Grube / Christian Prager / Katja Diederichs / Sven Gronemeyer / Antje Grothe / Céline Tamignaux / Elisabeth Wagner / Maximilian Brodhun / Franziska Diehr: Annual Report 2017. Bonn 2018. (= Textdatenbank und Wörterbuch des Klassischen Maya/Project Report, 5) DOI: 10.20376/IDIOM-23665556.18.pr005.de

Nikolai Grube: Die Entwicklung der Mayaschrift: Grundlagen zur Erforschung des Wandels der Mayaschrift von der Protoklassik bis zur spanischen Eroberung. Berlin 1990. (= Acta Mesoamericana, 3)[Nachweis im GBV]

Stephen Douglas Houston: Introduction. In: The Decipherment of Ancient Maya Writin *g*. Hg. von Stephen D. Houston / Oswaldo Chinchilla Mazariegos / David Stuart. Norman, OK. 2001, S. 3–19. [Nachweis im GBV]

Stephen Douglas Houston: An Example of Homophony in Maya Script. In: American Antiquity 49 (1984), H. 4, S. 790–805. [Nachweis im GBV]

ICOM/CIDOC CRM Special Interest Group: CIDOC Conceptual Reference Model. Hg. von Martin Doerr / Matthew Stiff / Nick Crofts / Stephen Stead / Tony Gill. Version 5.0.4. 2011. [online]

David Humiston Kelley: Review. John Eric Sidney Thompson: A Catalog of Maya Hieroglyphs. American Journal of Archaeology 66 (1962), H. 4, S. 436–438. [Nachweis im GBV]

Juri Valentinovich Knorosow: Drevnyaya pis'mennost' Tsentral'noy Ameriki. In: Sovetskaja Etnografija 3 (1952), H. 2, S. 100–118. [Nachweis im GBV]

Juri Valentinovich Knorosow: The Problem of the Study of the Maya Hieroglyphic Writing. In: American Antiquity 23 (1958), H. 3, S. 284–291. [Nachweis im GBV]

Juri Valentinovich Knorosow: Compendio Xcaret de la Escritura Jeroglífica Maya. 3 Bde. Chetumal 1999. [Nachweis im GBV]

Guido Krempel / Sebastián Matteo: Painted Styles of the North-Eastern Peten from a Local Perspective: The Palace Schools of Yax We'en Chan K'inich, Lord of Xultun. In: Proceedings of the 1st Cracow Maya Conference: Archaeology and Epigraphy of the Eastern Central Maya Lowlands. Hg. von Christophe Helmke / Jarosław Źrałka. (Cracow Maya Conference: 1, Krakau, 25.-27.02.2011) Kraków 2012, S. 135–171. [Nachweis im GBV]

Kornelia Kurbjuhn: Maya: The Complete Catalogue of Glyph Readings. Kassel 1989. [Nachweis im GBV]

Alfonso Lacadena: Evolución formal de las grafías escriturarias mayas: Implicaciones históricas y culturales. Madrid 1995. PDF. [online]

Alfonso Lacadena: On the Reading of Two Glyphic Appelatives of the Rain God. In: Continuity and Change: Maya Religious Practices in Temporal Perspective, Hg. von Daniel Graña-Behrens / Nikolai Grube / Christian M. Prager / Frauke Sachse / Stefanie Teufel / Elisabeth Wagner. (European Maya Conference: 5, Bonn, 12.2000) Markt Schwaben 2004, S. 87–98. (= Acta Mesoamericana, 14) [Nachweis im GBV]

Alfonso Lacadena: Historical Implications of the Presence of Non-Mayan Linguistic Features in the Maya Script. In: The Maya and Their Neighbours: Internal and External Contacts Through Time. Hg. von Laura van Broekhoven / Frauke Sachse / Benjamin Vis / Rogelio Valencia Rivera. (European Maya Conference: 10, Leiden, 09.-10.12.2005) Markt Schwaben 2010, S. 29–39. (= Acta Mesoamericana, 22) [Nachweis im GBV]

Alfonso Lacadena / Søren Wichmann: The Distribution of Lowland Maya Languages in the Classic Period. In: La organización social entre los mayas. Hg. von Vera Tiesler Blos / Rafael Cobos / Merle G. Robertson. 2 Bde. (Mesa Redonda de Palenque: 3, Palenque, Chiapas, Mexico, 27.06.-01.07.1999) México (u.a.) 2002. Bd. 2, S. 275–319. [Nachweis im GBV]

Alfonso Lacadena / Søren Wichmann: The Dynamics of Language in the Western Lowland Maya Region. [online] In: Art for Archaeology's Sake: Material Culture and Style across the Disciplines, Hg. von Andrea Waters-Rist / Christine Cluney / Calla McNamee / Larry Steinbrenner. (Conference of the Archaeological Association of the University of Calgary : 33, Calgary, 2005) Calgary, Alta 2005, S. 32–48.

Diego de Landa: Relación de las cosas de Yucatán. 8. edition. México 1959. (= Biblioteca Porrúa, 13) [Nachweis im GBV]

Eugene E. Loos / Susan Anderson / Dwight H. Day / Paul C. Jordan / J. Douglas Wingate: Glossary of Linguistic Terms. Hg. von SIL International. Dallas, TX. 2003. [online]

Martha Jane Macri / Matthew George Looper: The New Catalog of Maya Hieroglyphs. 2 Bde. Norman, OK. 2003. Bd. 1: The Classic Period Inscriptions. (= Civilization of the American Indian Series, 247) [Nachweis im GBV]

Martha Jane Macri / Matthew George Looper: The New Catalog of Maya Hieroglyphs. 2 Bde. Norman, OK. 2009. Bd. 2: The Codical Texts.(= Civilization of the American Indian Series, 264) [Nachweis im GBV]

SKOS Simple Knowledge Organisation System. Reference. Hg. von Alistair Miles / Sean Bechhofer. In: W3C. Recommendation vom 18. August 2009. [online]

Simon K. Milton / Barry Smith: Top-level Ontology: The Problem with Naturalism. In: Formal Ontology in Information Systems. Hg. von Achille Carlo Varzi / Laure Vieu. (FOIS 2004, Turin, 04.-06.11.2004) Amsterdam (u.a) 2004, S. 85–94. [Nachweis im GBV]

Sylvanus Griswold Morley: An Introduction to the Study of the Maya Hieroglyphs. Washington, D.C. 1915. (= Bulletin/Bureau of American Ethnology/Smithsonian Institution, 57) [Nachweis im GBV]

TextGrid: Von der Community – für die Community. Eine Virtuelle Forschungsumgebung für die Geisteswissenschaften. Hg. von Heike Neuroth / Andrea Rapp / Sibylle Söring. Göttingen 2015. DOI: 10.3249/webdoc-3947 [Nachweis im GBV]

Carlos Pallan Gayol / Deborah Anderson: Achieving Machine-Readable Mayan Text via Unicode: Blending »Old World« Script-encoding with Novel Digital Approaches. [online] In: Digital Humanities 2018: Puentes-Bridges. Book of Abstracts. Hg. von Jonathan Girón Palau / Isabel Galina Russell. (DH 2018, Mexico City, 26.-30.06.2018) Mexico 2018, S. 256–261. PDF. [online]

Christian Prager (2014a): Überblick über das Projekt. In: Textdatenbank und Wörterbuch des Klassischen Maya. 16. Dezember 2014. [online]

Christian Prager (2014b): Zielsetzung. In: Textdatenbank und Wörterbuch des Klassischen Maya. 4. Dezember 2014. [online]

Christian Prager (2014c): Zeichentypen und Kataloge. In: Textdatenbank und Wörterbuch des Klassischen Maya. 16. Dezember 2014. [online]

Anke Reinhold: Das Experteninterview als zentrale Methode der Wissensmodellierung in den Digital Humanities. In: Information Wissenschaft Praxis 66 (2015), H. 5-6, S. 327–333. [Nachweis im GBV]

Berthold Riese: Grundlagen zur Entzifferung der Mayahieroglyphen, dargestellt an den Inschriften von Copan. München 1971. (= Beiträge zur mittelamerikanischen Völkerkunde, 11) [Nachweis im GBV]

Berthold Riese: Drei neue Maya-Hieroglyphen Kataloge. [online] In: Anthropos 101 (2006), H. 1, S. 238–246. [online] [Nachweis im GBV]

William M. Ringle / Thomas C. Smith-Stark: A Concordance to the Inscriptions of Palenque, Chiapas, Mexico. New Orleans, LA. 1996. (= Publication/Middle American Research Institute, 62) [Nachweis im GBV]

Henry Rogers: Writing Systems: A Linguistic Approach. Oxford (u.a.) 2005. (= Blackwell Textbooks in Linguistics, 18) [Nachweis im GBV]

Ferdinand de Saussure: Cours de Linguistique Générale. 3. Auflage. Paris 1931. [Nachweis im GBV]

Paul Schellhas: Die Göttergestalten der Mayahandschriften. Ein mythologisches Kulturbild aus dem alten Amerika. Dresden 1897. [Nachweis im GBV]

Elena Simperl: Guidelines for Reusing Ontologies on the Semantic Web. DOI: 10.1142/S1793351X10001012 In: International Journal of Semantic Computing 4 (2010), H. 2, S. 239–283. [online] [Nachweis im GBV]

John F. Sowa: Knowledge Representation. Logical, Philosophical, and Computational Foundations. Pacific Grove, CA. 2000. [Nachweis im GBV]

David Stuart: On the Paired Variants of TZ'AK. Cambridge, MA. 2003. PDF. [online]

David Stuart: Ten Phonetic Syllables. Washington, D.C. 1987. (= Research Reports on Ancient Maya Writing, 14) [online] [Nachweis im GBV]

Text Encoding Initiative, P5: Guidelines for Electronic Text Encoding and Interchange, Version 3.4.0., 23. July 2018. [online]

Text Database and Dictionary of Classic Mayan: Ontology of the Sign Catalogue for Classic Mayan, 2017. [online]

Text Database and Dictionary of Classic Mayan: Gitolite Repository, 2018. [online]

John Eric Sidney Thompson: The Rise and Fall of Maya Civilization. Norman, OK. 1956. (= The Civilization of the American Indian Series, 39) Siehe auch: [Nachweis im GBV]

John Eric Sidney Thompson: A Catalog of Maya Hieroglyphs. Norman, OK. 1962. (= The Civilization of the American Indian Series, 62) [Nachweis im GBV]

Abbildungslegenden und -nachweise

Abb. 1: Beispiele für Maya-Hieroglyphen, links La Corona Panel 1, rechts polychrome Keramik unbekannter Herkunft; Fotos: Sven Gronemeyer.

Abb. 2: Schreibung von Hieroglyphen. John Montgomery: How to Read Maya Hieroglyphs. Hippocrene, New York, NY, 2002.

Abb. 3: Graphvarianten. Konzept: Franziska Diehr; Zeichnungen: Christian Prager.

Abb. 4: Domain-Model der Zeichenkatalog-Ontologie. Konzept: Franziska Diehr.

Abb. 5: Klassenhierarchie der Zeichenkatalog-Ontologie. Konzept: Franziska Diehr.

Abb. 6: Entzifferungskategorien. Konzept: Sven Gronemeyer; Zeichnungen der Glyphen: Matthew Looper.

Abb. 7: Entzifferung von Silbenzeichen. Konzept: Sven Gronemeyer; Zeichnungen der Silbenzeichen: Diego de Landa, Relación de las cosas de Yucatán. Fray Di[eg]o de Landa: MDLXVI. Unveröffentlichtes Manuskript. Madrid, Biblioteca de la Real Academia de Historia, 1566; Faksimile Codex Dresden: SLUB [online]; Faksimile Codex Madrid: Ferdinand Anders Codex Tro-Cortesianus (Codex Madrid): Museo de America Madrid. Codices Selecti Phototypice Impressi 8. Akademische Druck- u. Verlagsanstalt, Graz, 1967.

Abb. 8: Exemplarische Auszeichnung eines Hieroglyphenblocks in TEI/XML. Screenshot aus dem Projektbereich von ›Textdatenbank und Wörterbuch des Klassischen Maya‹ im TextGrid Lab.

Abb. 9: Systemarchitektur des Projekts ›Textdatenbank und Wörterbuch des Klassischen Maya‹. Konzept: Maximilian Brodhun.

Abb. 10: Graphemrelationen: ikonische Elemente. Konzept: Sven Gronemeyer; Zeichnungen der Glyphen: Ian Graham, Matthew Looper, John Montgomery, David Mora-Marín.

Abb. 11: Graphemrelationen: Derivation. Konzept: Sven Gronemeyer; Zeichnungen der Glyphen: Matthew Looper.

Abb. 12: Graphemrelationen: Diagnostische Elemente. Konzept: Sven Gronemeyer; Zeichnungen der Glyphen: Matthew Looper.

Abb. 13: Allographe Zeichen. Konzept: Sven Gronemeyer; Zeichnungen der Glyphen: Matthew Looper.

ZfdG

Zeitschrift für digitale Geisteswissenschaften

Beitrag aus:
Sonderband 4 der ZfdG: Die Modellierung des Zweifels – Schlüsselideen und -konzepte zur graphbasierten Modellierung von Unsicherheiten. Hg. von Andreas Kuczera, Thorsten Wübbena und Thomas Kollatz. 2019. DOI: 10.17175/sb004

Titel:
Modellierung von Zweifel – Vorbild TEI im Graphen

Autor/in:
Andreas Kuczera

Kontakt: andreas.kuczera@adwmainz.de
Institution: Akademie der Wissenschaften und der Literatur, Mainz
GND: 1167802993 ORCID: 0000-0003-1020-507X

Autor/in:
Dominik Kasper

Kontakt: dominik.kasper@adwmainz.de
Institution: Akademie der Wissenschaften und der Literatur, Mainz
GND: 1018231137 ORCID: 0000-0002-6587-381X

DOI des Artikels:
10.17175/sb004_003

Nachweis im OPAC der Herzog August Bibliothek:
1037067967

Erstveröffentlichung:
18.07.2019

Lizenz:

Sofern nicht anders angegeben (cc) BY-SA

Medienlizenzen:
Medienrechte liegen bei den Autoren

Letzte Überprüfung aller Verweise:
03.06.2019

GND-Verschlagwortung:
Edition | Graphdatenbank | Semantische Modellierung | Text Encoding Initiative | Ungewissheit |

Zitierweise:
Andreas Kuczera, Dominik Kasper: Modellierung von Zweifel – Vorbild TEI im Graphen. In: Die Modellierung des Zweifels – Schlüsselideen und -konzepte zur graphbasierten Modellierung von Unsicherheiten. Hg. von Andreas Kuczera, Thorsten Wübbena und Thomas Kollatz. 2019 (= Sonderband der Zeitschrift für digitale Geisteswissenschaften, 4). PDF Format ohne Paginierung. Als text/html abrufbar unter DOI: 10.17175/sb004_003.

Andreas Kuczera, Dominik Kasper

Modellierung von Zweifel – Vorbild TEI im Graphen

Abstracts

Im Fokus der hier ausgewerteten Integration von TEI-Dokumenten aus dem Deutschen Textarchiv (DTA) in eine Graphdatenbank (Neo4j) steht die Auszeichnung von unsicheren Lesarten und editorischen Ergänzungen in Handschriften. In diesem Zusammenhang legen wir auch die TEI-Richtlinien zum Umgang mit Zweifel und Unsicherheiten dar. In Editions- oder Transkriptionsvorhaben arbeiten zumeist mehrere Personen. Deshalb sehen wir responsibility-Angaben an zweifelhaften Stellen als zentral für die Interoperabilität der Daten und die intersubjektive Nachvollziehbarkeit von Einzelentscheidungen an. Dies gilt insbesondere dann, wenn zusätzlich Angaben zum Grad der Sicherheit einer Auflösung möglich sind. Graphtechnologien bieten hier Möglichkeiten zur Modellierung, Visualisierung und Analyse von Unsicherheit und Verantwortung. Bei einer ausreichend großen Datenmenge lassen sich beispielsweise persönliche Auszeichnungsprofile der jeweiligen Bearbeiter erstellen.

The focus of the imports of TEI documents from the German Text Archive (DTA) into a graph database (Neo4j) evaluated here is the marking of uncertain readings and editorial additions in manuscripts. In this context, we also briefly present the TEI guidelines for dealing with doubts and uncertainties. Since several people usually work in editing or transcription projects, we regard responsibility information at dubious points as central to the interoperability of data and the intersubjective traceability of individual decisions. This applies in particular if additional information on the degree of security of a resolution is possible. Graph technologies offer possibilities for modeling, visualization and analysis of uncertainty and responsibility. With a sufficiently large amount of data, personal labeling profiles of the respective editors can be created, for example.

1. Modellierung von Zweifel in der TEI

1.1 Codierung von Unsicherheiten in der Lesart und von Lücken im Text

Die Text Encoding Initiative (TEI) bietet bei der Transkription von Texten verschiedene Möglichkeiten, zweifelhafte Lesarten auszuzeichnen und diese sowie damit in Zusammenhang stehende Informationen umfangreich zu dokumentieren. In diesem Beitrag werden zunächst die Elemente und Attribute zur Modellierung von Unsicherheit und Zweifel in der TEI dargestellt. Zentral ist in diesem Zusammenhang unseres Erachtens das *Kapitel 11.3.3.2* Use of the gap, del, damage, unclear, and supplied Elements in Combination der TEI-Richtlinien,[1] wo deren kombinierte Verwendung erklärt wird. Im Folgenden liefern wir hier eine Zusammenfassung dieser speziellen Codierungsmöglichkeiten. Die Links in den

[1] TEI-Guidelines 2018 Kapitel 11.3.3.2. Use of the gap, del, damage, unclear, and supplied Elements in Combination.

Fußnoten am Ende jeder Einzelerklärung führen zum Grundeintrag des Elements in der Online-Dokumentation der TEI, worin dessen allgemeine Bedeutung und Verwendung beschrieben wird. Im Einzelnen gehen wir hier ein auf:

- <gap> ist ein leeres Element, das eine Lücke im Text kennzeichnet. Unter einer Lücke wird eine Stelle verstanden, an der durch Tilgung oder Schaden alles komplett unlesbar ist.[2]
- <supplied> umschließt einen editorisch ergänzten Teil, bei einer durch Tilgung oder Schaden komplett unlesbaren Stelle im Text (siehe <gap>).[3]
- <unclear> umschließt einen transkribierten Buchstaben oder Textteil an einer Stelle, wo noch etwas unsicher lesbar ist, aber ein Teil durch Tilgung oder Schaden unlesbar wurde.[4]
- <@cert> ist ein Attribut, das den Grad an Sicherheit bzw. Gewissheit bei der unsicheren Lesung (als Attribut von unclear) oder der ergänzten Stelle (als Attribut von <supplied>) beinhaltet.[5] In der Regel werden fixe Werte vorgegeben (high, medium, low, unknown).[6]
- <@resp> ist ein Attribut, das einen Verweis auf den oder die Edierende beinhaltet, der oder die für die Auflösung der unsicheren Lesart oder die Ergänzung verantwortlich zeichnet.[7]
- <@reason> ist ein Attribut, indem die Ursache für Schäden, Tilgungen oder Lücken dokumentiert werden kann.[8] Auch hier werden in der Regel fixe Werte vorgegeben.[9]

Da die TEI-Richtlinien die Verwendung von <gap>, <supplied> und <unclear> für den hier betrachteten Modellierungszusammenhang in Kombination mit weiteren transkriptionstypischen Elementen erklären,[10] seien diese hier ebenfalls kurz erläutert:

- <damage> umschließt einen beschädigten Teil, bei einer Stelle, an der noch etwas lesbar ist, aber ein Teil durch Schaden unleserlich wurde.[11]
- <subst> umschließt einen Ersetzungsvorgang, der beispielsweise mit add und del näher beschrieben werden kann.[12]
- <del> umschließt den getilgten Teil bei einer Stelle, an der noch etwas lesbar ist, aber ein Teil durch Tilgung unleserlich wurde.[13]
- <add> umschließt den ergänzten Teil bei einer Stelle, an der etwas verbessert bzw. ersetzt wurde.[14]

[2] Siehe auch TEI Guidelines 2018, Kapitel 3.4.3 Additions, Deletions, and Omissions <gap> .
[3] Siehe auch TEI Guidelines 2018, Kapitel 11.3.3.1 Damage, Illegibility, and Supplied Text <supplied> .
[4] Siehe auch TEI Guidelines 2018, Kapitel 11.3.3.1 Damage, Illegibility, and Supplied Text <unclear> .
[5] Siehe auch TEI Guidelines 2018, att.global.responsibility <@cert> .
[6] Siehe auch TEI Guidelines 2018, teidata.certainty.
[7] Siehe auch TEI Guidelines 2018, att.global.responsibility <@resp> .
[8] Siehe auch TEI Guidelines 2018, teidata.certainty <@reason> .
[9] Siehe auch TEI Guidelines 2018, Kapitel 3.4.3 Additions, Deletions, and Omissions <@reason> .
[10] Vgl. TEI Guidelines 2018, Kapitel 11.3.3.2 Use of the gap, del, damage, unclear, and supplied Elements in Combination.
[11] Siehe auch TEI Guidelines 2018, Kapitel 11.3.3.1 Damage, Illegibility, and Supplied Text <damage> .
[12] Siehe auch TEI Guidelines 2018, Kapitel 11.3.1.5 Substitutions <subst> .
[13] Siehe auch TEI Guidelines 2018, Kapitel 3.4.3 Additions, Deletions, and Omissions <del> .
[14] Siehe auch TEI Guidelines 2018, Kapitel 3.4.3 Additions, Deletions, and Omissions <add> .

1.2 Codierung von allgemeiner Unsicherheit, Verantwortlichkeit und Genauigkeit

Neben der Möglichkeit, die Sicherheit einer editorischen Entscheidung, deren Begründung und Verantwortlichkeit zu dokumentieren, beinhaltet die TEI auch ein Modul, um Zweifel und Unsicherheit ob der richtigen Verwendung von TEI-Elementen selbst zu kodieren. Ebenfalls abgedeckt werden darin die Auszeichnung von Unsicherheiten bei der (vermeintlichen) Identifikation einer Entität, einer Textstruktur oder auch bei der Angabe und Auflösung numerischer Werte.

Erwähnt werden müssen hier daher auch die im Modul certainty auftretenden Elemente <precision>, <respons> und das mit dem Modul gleichnamige <certainty>:[15]

- <certainty> dient der Kodierung von Unsicherheiten bei der Verwendung von Elementen und Attributen bzw. damit ausgezeichneter Bereiche und zielt damit prinzipiell auf die Inhaltsebene. So können beispielsweise Zweifel daran dokumentiert werden, ob es sich um einen Orts- oder Personennamen handelt oder auch ob ein Absatz mit der Seite endet oder sich noch auf die nächste erstreckt.[16]
- <precision> ermöglicht die Exaktheit bzw. Genauigkeit von numerischen Aspekten des Markups (Datierungen, Einheiten, sonstige Zahlenwerte) graduell zu codieren bzw. näher zu beschreiben.[17]
- <respons> identifiziert den oder die Edierende, welche für bestimmte Aspekte von Inhalt und Auszeichnung verantwortlich zeichnet. Es ist gegenüber dem oben vorgestellten resp-Attribut angesichts verschiedener Attributionsmöglichkeiten deutlich feingranularer.[18]

Die genannten Elemente können mit zahlreichen Attributen versehen werden und bieten insgesamt sehr detaillierte Optionen, um Zweifel und Unsicherheit auf verschiedenen Ebenen in TEI-XML abzubilden. In der Praxis von mit TEI arbeitenden Projekten wie der Carl-Maria-von-Weber-Gesamtausgabe (WeGa),[19] dem Deutschen Textarchiv (DTA) und der noch in der Entwicklung befindlichen PROPYLÄEN. Forschungsplattform für Goethes Biographica spielt das Modul certainty keine Rolle. Die Ergebnisse einer 2010 durchgeführten und 2012 veröffentlichten Studie zur TEI-Benutzung[20] bei der Handschriftenkodierung lassen annehmen, dass die scientific community der TEI-Benutzenden die in den Modulen core (<unclear>, <gap>, <add>, core) und transcr (<supplied>, <damage>, <subst>) befindlichen Elemente für ausreichend hält, um das gewünschte Maß an Dokumentation von Zweifel und Unsicherheit zu erreichen. Die Attribute @resp und @cert gehören zur TEI-Infrastruktur, im Basis-Modul tei.

[15] Vgl. TEI Guidelines 2018, Kapitel 21 Certainty, Precision, and Responsibility und besonders TEI Guidelines 2018 den Abschnitt 21.1.2 Structured Indications of Uncertainty zur hier folgenden Zusammenfassung.
[16] Siehe auch TEI Guidelines 2018, Kapitel 21.1.2 Structured Indications of Uncertainty <certainty> .
[17] Siehe auch TEI Guidelines 2018, Kapitel 21.2 Indications of Precision <precision> .
[18] Siehe auch TEI Guidelines 2018, Kapitel 21.3 Attribution of Responsibility <respons> .
[19] Vgl. Carl-Maria-von-Weber-Gesamtausgabe 2018.
[20] Vgl. Burghart / Rehbein 2012.

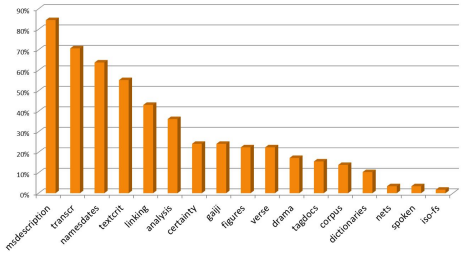


Abb. 1: TEI-Modules used in manuscript encoding projects (apart from the four basic ones: core, tei, header, textstructure). [Burghart / Rehbein 2012 , Fig. 11. CC BY-ND 3.0.]

Wie die obige Grafik zeigt, kommt das Modul certainty nur bei ca. 22 % der Vorhaben, die Handschriften auszeichnen, zum Einsatz. Gleichzeitig sehen nur sehr wenige TEI-Anwenderinnen und -Anwender Bedarf für eine Ausweitung oder Verbesserung der Richtlinien zur Verwendung der Elemente dieses Moduls, wie die folgende Grafik deutlich macht.

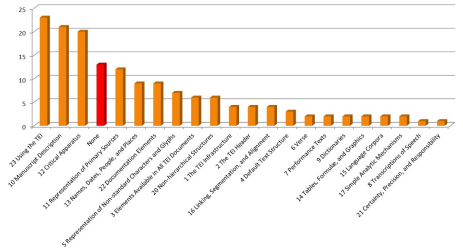


Abb. 2: In what areas do you wish the Guidelines to be improved? Figures are given in absolute numbers. [Burghart / Rehbein 2012, Fig. 18. CC BY-ND 3.0.]

Das entsprechende Kapitel der Richtlinien steht an letzter Stelle. Damit muss dieser Bereich als gut dokumentiert gelten. Interpretiert man dieses Ergebnis aber in Kombination mit der Aussage von Abbildung 1, so liegt die Annahme eines nur geringen Interesses der Fachwelt an einer sehr detaillierten Dokumentation von Zweifel und Unsicherheit nah. Zweifellos werden dabei arbeitsökonomische Aspekte eine Rolle spielen, aber auch Fragen nach dem Mehrwert solcher Informationen.

Nach dem kurzen Exkurs über die theoretische Tiefe der TEI hinsichtlich der Modellierung von Zweifel und Unsicherheit, ist der starke Eindruck entstanden, dass nur ein geringer Teil der Möglichkeiten genutzt wird. Im Folgenden richten wir den Blick auf zwei Beispiele aus der Praxis zum Umgang mit unsicheren Lesarten und Lücken im Text.

2. Zwei Beispiele aus der Praxis

2.1 DTA-Basisformat-Realisierung

Das Deutsche Textarchiv (DTA) hat ein eigenes, sehr umfangreich dokumentiertes Datenformat etabliert, das DTA-Basisformat, welches ein subset [21] der TEI ist. Zu unsicheren Lesarten bzw. schwer lesbaren Zeichen heißt es dort:

> »Ist die Leserlichkeit der Quelle eingeschränkt, sodass der Text rekonstruiert werden muss bzw. die Lesung des Editors nicht gesichert ist, kann dies durch die Elemente <unclear> und <supplied> wiedergegeben werden. Dabei wird <unclear> verwendet, wenn in der Quelle vorhandenes Material nur undeutlich lesbar ist. Der Grund für die Verwendung des <unclear>-Elements wird mit dem @reason-Attribut, der Grad der Sicherheit der Lesung kann im @cert-Attribut wiedergegeben werden. [...] Die Verwendung des Attributs @reason in <unclear> ist dabei obligatorisch, die Verwendung von @cert ist fakultativ.[...] Wenn in der Quelle wahrscheinlich oder möglicherweise vorhandenes Material rekonstruiert wird, so ist dies mit dem Element <supplied> wiederzugeben. Der Grund für die Unleserlichkeit wird im @reason-Attribut wiedergegeben, die Sicherheit der Rekonstruktion steht im @cert-Attribut.«[22]

> »Lassen sich die Zeichen nicht erkennen und nicht mehr rekonstruieren, wird das Tag <gap/> gesetzt, um die Lücke anzuzeigen. Innerhalb des <gap>-Tags kann mittels der Attribute @unit, @quantity und @reason der Bezug angezeigt werden, wie viele Zeichen die Lücke umfasst, so wie der Grund der Fehlstelle[.]«[23]

2.1.1 Code-Beispiel

Das folgende XML-Fragment stammt aus der Transkription von Gotthilf Patzigs Mitschriften von Humboldts Vorträgen über physische Geographie.[24]

```
1) <xml>
2) [...] So wie die Geogno&#x017F;ie durch die Auf-
3) <lb /> findung u. nähere Beachtung der thieri&#x017F;chen Ver-
4) <lb /> &#x017F;teinerungen aufgeklärt wurde: &#x017F;o hat
5) <lb /> der phy&#x017F;i&#x017F;che Theil der A&#x017F;tronomie durch
6) <lb /> die Entdeckungen im Gebiet der Optik gewon&#x0303;<supplied
7) reason="damage" resp="#BF">en;</supplied>
8) <lb /> u. die Cometen &#x017F;ind be&#x017F;onders näher
```

[21] Der Begriff ist hier im engeren mathematischen Sinne als echte Teilmenge zu verstehen. Das heißt, das DTABf ist eine reduzierte Fassung der TEI-P5-Richtlinien ohne Erweiterungen durch eigene Elemente.
[22] DTA-Basisformat 2011–2018, Unsichere Lesarten. Dort finden sich auch mögliche Werte für die genannten Attribute.
[23] DTA-Basisformat 2011–2018, Schwer bzw. nicht entzifferbare Zeichen und Auslassungen.
[24] Patzig 2007 (1827/1828), S. 13. Zur HTML-Ansicht des Beispiels: Deutsches Textarchiv 2007–2019, Patzig.

```
9) <lb /> beobachtet werden. Jn be&#x017F;tändiger
10) <lb /> Bewegung kan&#x0303; man die&#x017F;e eine perio-
11) <lb /> di&#x017F;ch o&#x017F;cilirende nen&#x0303;en. Die&#x017F;e kan&#x0303;
12) <lb /> gehem&#x0303;t, ge&#x017F;töhrt werden auf viele Wei&#x017F;e;
13) <lb /> den&#x0303; welch ein geringer Stoß von auß<unclear
14) reason="illegible" cert="high" resp="#CT">en</unclear>
15) <lb /> dürfte dazu gehören &#x017F;ie in Bewegung
16) <lb /> zu &#x017F;etzen, da die Dün&#x0303;igkeit der&#x017F;elben
17) <lb /> Alles über&#x017F;teigt was wir &#x017F;elb&#x017F;t von
18) <lb /> Gas-Arten auf der Erde ken&#x0303;en. Die&#x017F;e
19) <lb /> Dün&#x0303;igkeit i&#x017F;t 5000 mal geringer als die
20) <lb /> Dichtigkeit der Erde. &#x2013; Kom&#x0303;en wir jetzt
21) <lb /><note place="left"><hi rendition="#u">Telluri&#x017F;che
22) Verhältni&#x017F;&#x017F;e</hi>
23) <lb /></note><hi rendition="#u">zu den telluri&#x017F;chen
24) Verhältni&#x017F;&#x017F;en, <subst><del rendition="#erased"><gap
25) reason="illegible" /></del><add place="across">&#x017F;o
26) werd</add></subst>en
27) <lb /> wir die Form, Größe u. Dichtigkeit des
28) <lb /> Planeten betrachten &#x2013;</hi>[...]
29) </xml>
```

In Zeile 6 und 7 ist eine editorische Textergänzung vorgenommen worden. Hier hat der oder die Bearbeitende mit dem Kürzel »BF« aufgrund einer Beschädigung der Vorlage die Zeichen »en;« ergänzt. In Zeile 13 und 14 dokumentiert der Bearbeitende »CT« die unsichere Lesart der Zeichen »en« am Ende einer Zeile. Die Auflösung erfolgte mit hoher Gewissheit. Der in den Zeilen 24 bis 26 in einem <subst> codierte Vorgang lässt sich mit Bezug auf die Dokumentation in natürlicher Sprache so formulieren: Eine Tilgung durch Radieren, Auskratzen o. ä. hat hier eine Lücke im Text entstehen lassen, eine nicht mehr lesbare Stelle. Direkt darüber wurde nun die Zeichen »ſo werd» geschrieben. Wer diese Annotation vorgenommen hat, ist hier nicht ersichtlich.

2.2 Aus der Entwicklung – PROPYLÄEN. Goethes Biographica

Im Projekt PROPYLÄEN werden die textkritischen Anmerkungen nicht direkt im Text codiert, sondern in einem per Referenz verknüpften Apparatbereich (in TEI-Code ausgedrückt: <variantEncoding method="location-referenced" location="external" />). Im Apparat und konstituierten Text werden in der Handschrift *nicht eindeutig entzifferbare* Buchstaben oder Zahlen ebenfalls mit <unclear reason="illegible"> codiert. In der Handschrift *nicht entzifferbare* Zeichen werden mit <gap reason="illegible" extent="ANZAHL DER ZEICHEN SOFERN ERMITTELBAR"> ausgezeichnet.

2.2.1 Code-Beispiele

Die Codierung wurde zu Ansichtszwecken vereinfacht, alle Kommentarreferenzen wurden entfernt.

Unsichere Lesart (Codierung im konstituierten Text):[25]

```
1) <div type="entry" xml:id="GT01_1782_007">
2) <head>
3) <origDate when="1782-01-07" rendition="#fraktur">
4) 7 Mont. Isidorus
5) </origDate>
6) </head>
7) <note ana="metadaten">
8) <placeName type="uebernachtungsort">Weimar</placeName>
9) </note>
10) <p>Ackten und verschiedne Besorgungen. Mittags Crone.
11) um halb 5 zur reg. H. dann zu Seckend. wo # war und über
12) Aufzüge gesprochen wurde p zur Waldner
13) war # dasel<unclear reason="illegible">b</unclear>st und
14) Stein. kam #. Ging mit ihm auf Zimmer, ihm die Erfindung
15) zu erzählen.</p>
16) </div>
```

In Zeile 13 wird die unsichere Lesart des Buchstabens »b« vermerkt.

Nicht mehr lesbare Stelle oder Lücke (Codierung im Apparat):[26]

```
1) <app xml:id="app_05">
2) <rdg>
3) <subst>
4) <del rendition="#sofortkorrektur">
5) <gap reason="illegible" extent="1 char" />
6) </del>
7) <add>den Fus.</add>
8) </subst>
9) </rdg>
10) </app>
```

Der Apparateintrag codiert, dass ein nicht mehr lesbares Zeichen zu »den Fus.« korrigiert wurde. Die Korrektur erfolgte sofort und nicht später, bspw. durch einen Schreiber.

[25] Bisher unveröffentlichte und in Modellierungs- und Auszeichnungsarbeit befindliche Daten aus der Retrodigitalisierung, s. Goethe Tagebücher 1998, T I,1, S. 130, Z. 1–5.
[26] Goethe Tagebücher 1998, T I,1, S. 130, Z. 14.

3. Modellierung im Graphen

3.1 Die Kosmos-Vorträge von Alexander von Humboldt

Im zweiten Teil des Beitrags werden die Elemente <unclear> und <supplied> aus dem DTA-Basisformat näher betrachtet. Datengrundlage sind hierbei die Kosmos-Vorträge von Alexander von Humboldt, die dieser 1827/28 in Berlin einmal an der Universität und einmal an der Sing-Akademie gehalten hat. Zu diesen Vorträgen liegen Mitschriften vor, die im Rahmen eines Forschungsprojekts im Deutschen Textarchiv transkribiert wurden.[27] Humboldts eigene Manuskripte zu den Vorträgen sind nicht erhalten. Es gibt aber für beide Vortragsreihen Mitschriften von Zuhörenden, die aber voneinander abweichen. Vereinfacht gefragt, geht es also darum, was Humboldt wirklich gesagt hat.

Im Folgenden werden fünf der Mitschriften gemeinsam in eine Graphdatenbank eingespielt, die Verwendung der Elemente <unclear> und <supplied> untersucht und schließlich in Relation zu den Edierenden gebracht, die sie in der Transkription verwendet haben (also jenen, die im @resp-Attribut genannt sind). Graphdatenbanken sind sehr gut für die Darstellung stark vernetzter Daten geeignet und in diesem Fall wäre es ein Versuch wert, die Rolle von Editorinnen und Editoren über die Grenzen der Dateien hinweg auszuwerten.

3.2 Import von TEI-XML in eine Graphdatenbank

Für die Analyse von TEI-Unsicherheitsannotationen im Graphen müssen die XML-Daten zunächst in die Graphdatenbank Neo4j importiert werden. Hierfür hat Stefan Armbruster[28] die apoc-Bibliothek von Neo4j um die procedure apoc.load.xml erweitert. Prinzipiell können XML-Dateien ohne größere Probleme in einen Graphen importiert werden, da sie einen geerdeten, gerichteten azyklischen Graphen darstellen, der vielfache Elternbeziehungen verhindert. Damit stellen sie ein Ordered Hierarchy of Content Objects (OHCO) dar.

Der folgende query importiert eine der fünf Vorlesungsmitschriften:[29]

call

apoc.xml.import('http://www.deutschestextarchiv.de/book/download_xml/
parthey_msgermqu1711_1828',{createNextWordRelationships:true})

[27] Vgl. Thomas et al. 2016, S. 287–318.
[28] Stefan Armbruster ist bei neo4j tätig.
[29] Mit dem Befehl wird die Vorlesungsmitschrift von Gustav Parthey importiert. Zur Partheymitschrift vgl. Deutsches Textarchiv 2007–2019, Parthey 1827/1828. Informationen zu den weiteren Mitschriften sind zu finden unter Deutsches Textarchiv 2007–2019, Titeldaten. Der Query für den Import aller fünf Mitschriften steht als Download zur Verfügung. Der Import von Texten des Deutschen Textarchivs in die Graphdatenbank Neo4j wird ausführlich erklärt im Kapitel *XML Text im Graphen* in Kuczera 2018.

```
yield node return node;

// URL von Dokument auf alle Wort-Knoten kopieren:

match (d:XmlDocument)-[:NEXT_WORD*]->(w:XmlWord)

set w.url = d.url;

// Knoten durchzählen

MATCH p =

(start:XmlDocument)-[:NEXT*]->(end:XmlTag)

WHERE NOT (end)-[:NEXT]->() AND start.url =

'http://www.deutschestextarchiv.de/book/download_xml/parthey_msgermqu1711_1828'

WITH nodes(p) as nodes, range(0, size(nodes(p))) AS indexes

UNWIND indexes AS index

SET (nodes[index]).DtaID = index;
```

Für den Import der weiteren Mitschriften muss in dem Befehl die DTA-URL entsprechend geändert werden.

Beim Import werden die XML-Knoten in Graphknoten umgewandelt und verschiedene Arten von Kanten erstellt, welche die Baum-Hierarchie des XMLs im Graphen abbilden. Mit der Option createNextWordRelationships:true wird darüber hinaus festgelegt, dass die im XML vorhandenen Textknoten über NEXT_WORD-Kanten miteinander verknüpft werden. Zu beachten ist hierbei, dass es in TEI-XML zwei verschiedene Arten von Elementen gibt. Die eine Klasse dient der Klassifizierung von Text, die zweite Art bringt Varianten und zusätzlichen Text mit, der beim Import in seiner Serialität eingelesen und mit NEXT_WORD-Kanten verbunden wird. Dies kann dann zur Sinnentstellung des Textes führen.[30] Mit dem zweiten cypher-Befehl wird jedem XmlWord-Knoten die URL des XML-Dokuments als property mitgegeben. Damit behält man im Graphen beim Betrachten der Ergebnisse den Überblick und kann die XmlWord-Knoten einem XML-Dokument zuordnen. Der dritte query nummeriert die Knoten pro Datei durch und macht sie damit innerhalb des Dokuments eindeutig referenzierbar.

[30] Zum Import von XML-Text in die Graphdatenbank Neo4j vgl. insbesondere das Kapitel zu *XML-Text im Graphen* in Kuczera 2018.

3.3 Das XML-Element <unclear> im Graphen

Mit dem folgenden query wird eine Stelle im importierten XML aufgerufen, an der ein <unclear>-Element verwendet wurde:

// unclear-Beispiel

MATCH

(t1:XmlTag {_name:'lb'})<-[:NEXT_SIBLING]-(t2:XmlTag {_name:'unclear'})<-[:NEXT]-(w3:XmlWord {text:'auß'}),

(w1:XmlWord)-[:NEXT_WORD]->

(w2:XmlWord)-[:NEXT_WORD]->

(w3:XmlWord)-[:NEXT_WORD]->

(w4:XmlWord)-[:NEXT_WORD]->

(w5:XmlWord)

RETURN *;

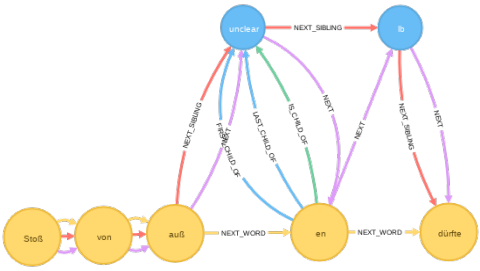


Abb. 3: Beispiel zur Graphmodellierung eines unclear-Elements. [Kasper / Kuczera 2019.]

Die entsprechende Stelle sieht in XML wie folgt aus:[31]

```
Wei&#x017F;e;<lb/>
den&#x0303; welch ein geringer Stoß von auß<unclear
reason="illegible" cert="high" resp="#CT">en</unclear>
<lb/>
dürfte dazu gehören &#x017F;ie in Bewegung<lb/>
zu &#x017F;etzen, da die Dün&#x0303;igkeit
```

Abb. 4: Das unclear-Beispiel in der XML-Ansicht des DTA. [Kasper / Kuczera 2019.]

[31] Vgl. Deutsches Textarchiv 2007–2019, Patzig 1827/1828.

Das für den Grad des Zweifels maßgebliche @cert-Attribut befindet sich in den properties des unclear-Knotens.

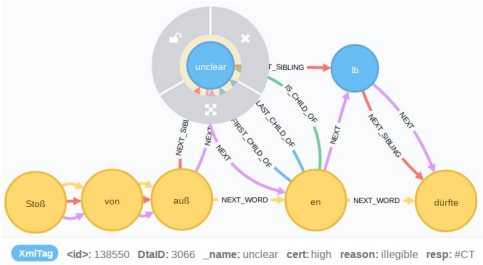


Abb. 5: Die properties des unclear-Knotens. [Kasper / Kuczera 2019.]

Wie im Beispiel aufgezeigt, können mit dem Import alle Informationen des XMLs verlustfrei in den Graph überführt und abgebildet werden.

3.4 Die Zweifel der Edierenden

Die objektive Gewichtung von Zweifeln im Hinblick auf die Interoperabilität ist schwierig. Ermöglicht man den Edierenden feingranularere Abstufungen, um Zweifel zum Ausdruck zu bringen (z. B. in 10er-Schritten von 0 % bis 100 %) führt das oft zu Verunsicherung. Gibt es nur zwei Stufen, wie im DTA-Basisformat mit *high* und *low*, bleibt die Gewichtung grob, Vergleiche fallen aber leichter. Der hier vorgestellte Ansatz verzichtet auf eine objektive Vergleichbarkeit und ordnet die von den Edierenden vergebenen gewichteten Zweifeln den Personen zu. Stehen genügend Daten zur Verfügung, könnte aus den Annotationen ein persönlicher Fingerabdruck des jeweiligen Edierenden erstellt werden.

Zunächst wird mit folgendem cypher query abgefragt, welcher Edierende in welchem Dokument welche XML-Elemente genutzt hat, wobei die XML-Elemente sowohl das @cert- als auch das @resp-Attribut haben müssen:

```
// Zweifelsattribute in der TEI pro Dokument

MATCH (n:XmlTag)

WHERE n.resp IS NOT NULL

AND n.cert IS NOT NULL

RETURN n.url, n._name AS Element, n.resp AS Person, n.reason, n.cert, count(n.resp) AS Anzahl

ORDER BY Anzahl DESC
```

Name	Element	Person	n.reason	n.cert	Anzahl
patzig_msgermfol841842_1828	unclear	#BF	illegible	high	137
patzig_msgermfol841842_1828	unclear	#BF	illegible	low	97
parthey_msgermqu1711_1828	unclear	#CT	illegible	high	34
nn_msgermqu2345_1827	unclear	#BF	illegible	high	27
parthey_msgermqu1711_1828	unclear	#CT	illegible	low	25
parthey_msgermqu1711_1828	unclear	#CT	covered	high	22
nn_msgermqu2345_1827	unclear	#BF	illegible	low	19
patzig_msgermfol841842_1828	unclear	#BF	covered	low	19
patzig_msgermfol841842_1828	unclear	#BF	covered	high	16
parthey_msgermqu1711_1828	unclear	#CT	covered	low	13

Abb. 6: Die gekürzt wiedergegebene Tabelle zeigt die häufigsten Ergebnisse des obigen Querys. Die Angaben in der ersten Spalte der Tabelle wurden aus Gründen der Übersichtlichkeit um den URL-Teil gekürzt, der bei allen Mitschriften gleich ist. [Kasper / Kuczera 2019.]

Mit Abstand am häufigsten wurde bei der Transkription der Vorlesungsmitschriften das <unclear>-Element verwendet, mit einigem Abstand gefolgt vom <supplied>-Element.

3.4.1 Die Identifizierung des Edierenden

Die Spalte Person in der Tabelle gibt den Inhalt der resp-property an, in der die Person des Edierenden mit einem Kürzel wiedergegeben wird. Im XML-Header werden diese Kürzel auf folgende Personen aufgelöst:

Kürzel	Name
CT	Christian Thomas
BF	Benjamin Fiechter
TK	Tina Krell

Abb. 7: Aufschlüsselung der Edierendenkürzel. [Kasper / Kuczera 2019.]

Es sind Personen, die mit der Transkription der Humboldt-Vorlesungsmitschriften befasst waren. Im nächsten Schritt werden nun alle in den resp-properties genannten Edierenden explizit als Personenknoten erstellt und mit jenen unclear- und supplied-Knoten verknüpft, für die sie verantwortlich sind. Mit folgendem query werden die Personen erstellt:

// Edierende Personenknoten erstellen

```
MATCH (n:XmlTag)

WHERE n.resp IS NOT NULL

AND n.cert IS NOT NULL

MERGE (p:Person {name:n.resp})

RETURN *;

// Alle XML-Elemente mit resp-Attribut den erstellten Personen zuordnen

MATCH (n:XmlTag), (p:Person {name:n.resp})

WHERE n.resp IS NOT NULL

AND n.cert IS NOT NULL

MERGE (p)<-[:RESPONSIBLE {cert:n.cert}]-(n)

RETURN *;
```

und mit den entsprechenden unclear- und supplied-Knoten im Graphen über RESPONSIBLE-Kanten verknüpft.

Abb. 8: unclear- und supplied-Knoten, die von #CT erstellt worden sind. [Kasper / Kuczera 2019.]

3.4.2 Statistik zur Zweifel im Graphen

Mit dem folgenden cypher query fragen wir die Häufigkeit der unclear- und supplied-Knoten im Graphen ab und ordnen sie den Edierenden zu:

// 1 Elementanzahl pro Person

MATCH (n:XmlTag)-[:RESPONSIBLE]->(p:Person)

RETURN n._name AS Elementname, p.name AS Editorname,

count(n._name) AS Elementanzahl ORDER BY Elementanzahl DESC;

Elementname	Editorname	Elementanzahl
"unclear"	"#BF"	337
"unclear"	"#CT"	134
"supplied"	"#CT"	7
"supplied"	"#TK"	5
"supplied"	"#BF"	2

Abb. 9: Häufigkeit der unclear- und supplied-Knoten der jeweiligen Edierenden. [Kasper / Kuczera 2019.]

Der Editor #BF hat in den fünf in der Graphdatenbank enthaltenen Dokumenten insgesamt 337 <unclear>- und nur zwei <supplied>-Elemente eingefügt, während der Editor #CT nur für 134 <unclear>- und für 7 <supplied>-Elemente verantwortlich ist.

Der folgende cypher query nimmt noch den Inhalt des @cert-Attributs hinzu.

// 2 Zweifel pro Person über alles

MATCH (n:XmlTag)-[:RESPONSIBLE]->(p:Person)

RETURN n._name AS Elementname, n.cert AS Zweifel, p.name AS Editorname, count(n._name) AS Elementanzahl ORDER BY Elementanzahl DESC;

Elementname	Zweifel	Editorname	Elementanzahl
"unclear"	"high"	"#BF"	190
"unclear"	"low"	"#BF"	147
"unclear"	"high"	"#CT"	76
"unclear"	"low"	"#CT"	58
"supplied"	"high"	"#CT"	7
"supplied"	"high"	"#TK"	5
"supplied"	"high"	"#BF"	2

Abb. 10: Häufigkeit der unclear- und supplied-Knoten mit Angabe des cert-Attributs. [Kasper / Kuczera 2019.]

Damit differenziert sich das Bild etwas, jedoch sind die Anteile von high- und low-Werten bei den jeweiligen Editoren im Durchschnitt gleich.[32]

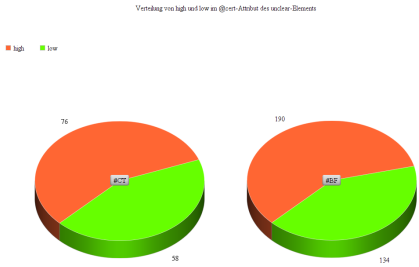


Abb. 11: Verteilung von high- und low-Werten sind beim cert-Attribut bei den Edierenden etwa gleich verteilt. [Kasper / Kuczera 2019.]

Mit dem nächsten query wird das Raster auf Dokumentenebene verfeinert:

```
// 5 Dokumente mit Bearbeitern, Elementen und Zweifeln

MATCH (n:XmlTag)-[:RESPONSIBLE]->(p:Person)

RETURN n.url AS Dokument, p.name AS Editorname, n._name AS Elementname, n.cert AS Zweifel, count(n._name) AS Elementanzahl ORDER BY Dokument, Elementname, Zweifel, Elementanzahl;
```

[32] Die Grafik wurde erstellt mit dem Onlinetool Diagrammwerkzeug.

Dokument	Elementname	Zweifel	Editorname	Elementanzahl
"http://www.deutschestextarchiv.de/book/download_xml/patzig_msgermfol841842_1828"	"unclear"	"high"	"#BF"	153
"http://www.deutschestextarchiv.de/book/download_xml/patzig_msgermfol841842_1828"	"unclear"	"low"	"#BF"	116
"http://www.deutschestextarchiv.de/book/download_xml/parthey_msgermqu1711_1828"	"unclear"	"high"	"#CT"	56
"http://www.deutschestextarchiv.de/book/download_xml/parthey_msgermqu1711_1828"	"unclear"	"low"	"#CT"	38
"http://www.deutschestextarchiv.de/book/download_xml/nn_msgermqu2345_1827"	"unclear"	"high"	"#BF"	36
"http://www.deutschestextarchiv.de/book/download_xml/nn_msgermqu2345_1827"	"unclear"	"low"	"#BF"	31
"http://www.deutschestextarchiv.de/book/download_xml/hufeland_privatbesitz_1829"	"unclear"	"high"	"#CT"	9
"http://www.deutschestextarchiv.de/book/download_xml/hufeland_privatbesitz_1829"	"unclear"	"low"	"#CT"	8
"http://www.deutschestextarchiv.de/book/download_xml/patzig_msgermfol841842_1828"	"unclear"	"low"	"#CT"	7
"http://www.deutschestextarchiv.de/book/download_xml/hufeland_privatbesitz_1829"	"supplied"	"high"	"#TK"	5
"http://www.deutschestextarchiv.de/book/download_xml/nn_msgermqu2345_1827"	"unclear"	"high"	"#CT"	5
"http://www.deutschestextarchiv.de/book/download_xml/nn_msgermqu2124_1827"	"unclear"	"high"	"#CT"	4
"http://www.deutschestextarchiv.de/book/download_xml/patzig_msgermfol841842_1828"	"supplied"	"high"	"#CT"	3
"http://www.deutschestextarchiv.de/book/download_xml/nn_msgermqu2345_1827"	"unclear"	"low"	"#CT"	3
"http://www.deutschestextarchiv.de/book/download_xml/nn_msgermqu2345_1827"	"supplied"	"high"	"#CT"	2
"http://www.deutschestextarchiv.de/book/download_xml/nn_msgermqu2124_1827"	"unclear"	"low"	"#CT"	2
"http://www.deutschestextarchiv.de/book/download_xml/patzig_msgermfol841842_1828"	"supplied"	"high"	"#BF"	2
"http://www.deutschestextarchiv.de/book/download_xml/patzig_msgermfol841842_1828"	"unclear"	"high"	"#CT"	2
"http://www.deutschestextarchiv.de/book/download_xml/parthey_msgermqu1711_1828"	"supplied"	"high"	"#CT"	1
"http://www.deutschestextarchiv.de/book/download_xml/parthey_msgermqu1711_1828"	"unclear"	"high"	"#BF"	1
"http://www.deutschestextarchiv.de/book/download_xml/nn_msgermqu2124_1827"	"supplied"	"high"	"#CT"	1

Abb. 12: Verteilung der unclear- und supplied-Knoten mit Angaben zum Zweifel auf Dokumentebene. [Kasper / Kuczera 2019.]

Es ist zu erkennen, dass #BF die Vorlesungsmitschrift von Patzig ediert hat und dabei die meisten <unclear>-Elemente vergeben hat. #CT hat in der Vorlesungsmitschrift von Parthey dagegen nur 94 <unclear>-Elemente verwendet. Dies könnte daran liegen, dass die Parthey-Vorlesungsmitschrift besser lesbar ist. Eine kurze Nachfrage beim Projekt Humboldt-Kosmos ergab aber, dass #BF wissenschaftliche Hilfskraft, #CT aber wissenschaftlicher Mitarbeiter ist. Die Ergebnisse der Tabelle könnten also auch zeigen, dass #CT die Handschriften besser lesen kann und deshalb weniger <unclear>-Elemente vergeben hat.

3.5 Verfeinerung des Profils

Im Folgenden wird die o. a. Auswertung in Kreisdiagrammen dargestellt. Die Farben der Legenden sind über alle Grafiken gleich. Im äußersten Ring werden die Anteile der Bearbeitenden an den in einem Dokument vergebenen <supplied>- und <unclear>-Elemente gezeigt.

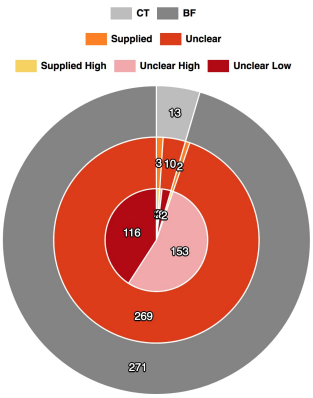


Abb. 13: Unsicherheitsverteilung auf Editorenebene bei Patzig. [Kasper / Kuczera 2019.]

Patzig

Bei der Transkription der Mitschrift von Patzig hat Benjamin Fiechter (BF) den größten Teil der <supplied>- und <unclear>-Elemente erstellt, ein kleinerer Teil wurde von Christian Thomas eingegeben. Im zweiten Ring ist zu erkennen, dass überwiegend <unclear>-Elemente vergeben wurden und dass Christian Thomas von den (wenigen) <supplied>-Elementen im Vergleich den größeren Teil eingetragen hat. Schließlich bleibt anzumerken, dass Benjamin Fiechter bei der Vergabe der <unclear>-Elemente im Verhältnis wesentlich öfter Zweifel hatte.

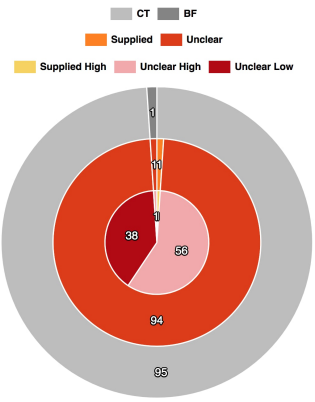


Abb. 14: Unsicherheitsverteilung auf Editorenebene bei Parthey. [Kasper / Kuczera 2019.]

Parthey

In der Mitschrift von Parthey zeichnet ganz überwiegend Christian Thomas für die <supplied>- und <unclear>-Elemente verantwortlich, verwendet aber fast nur das <unclear>-Element. Die folgenden Kreisdiagramme zeigen noch die Verteilungen der Vorlesungsmitschriften NN1, NN2 und Hufeland.

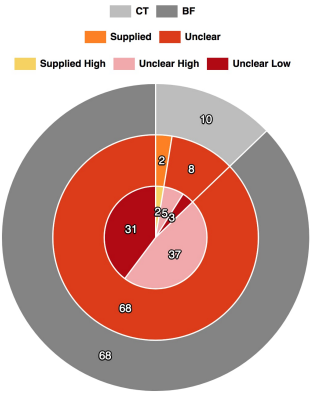


Abb. 15: Unsicherheitsverteilung auf Editorenebene in NN2. [Kasper / Kuczera 2019.]

NN1

Die Mitschrift NN1 wurde überwiedend von Benjamin Fiechter ausgezeichnet, der kein unclear-Element verwendet. Christian Thomas annotiert dagegen zwei Stellen mit supplied-Elementen.

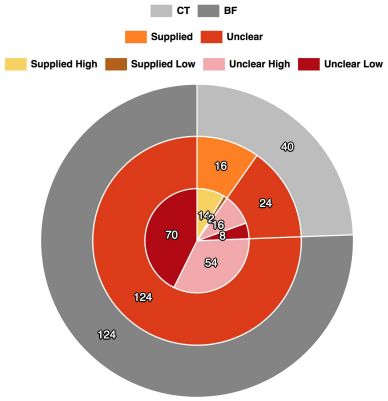


Abb. 16: Unsicherheitsverteilung auf Editorenebene in NN2. [Kasper / Kuczera 2019.]

NN2

Ein ähnliches Bild ergibt sich für die Mitschrift NN2. Auch hier verwendet Benjamin Fiechter nur unclear-Elemente, während Christian Thomas auch einen Anteil an supplied-Elementen vergibt.

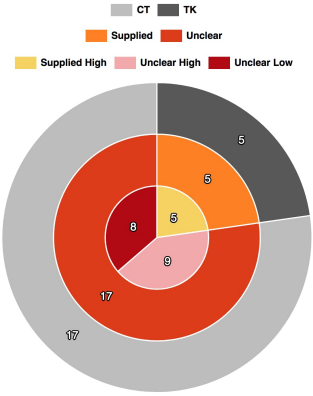


Abb. 17: Unsicherheitsverteilung auf Editorenebene bei Hufeland. [Kasper / Kuczera 2019.]

Hufeland

Interessant scheint vor allem die Mitschrift Hufeland (Abbildung 17), bei der Christian Thomas fast alle <unclear>-Elemente eingetragen hat, während Benjamin Fiechter alle <supplied>-Elemente vergeben hat.[33] Eine sehr interessante Art von Arbeitsteilung, die sich so in keiner anderen Transkription findet. Ein kleiner Anteil der Bearbeitung wurde bei dieser Mitschrift auch von Tina Krell vorgenommen.

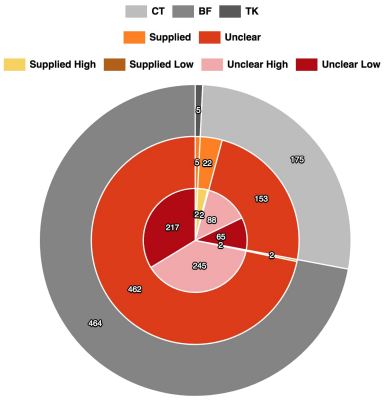


Abb. 18: Unsicherheitsverteilung auf Editorenebene insgesamt. [Kasper / Kuczera 2019.]

[33] Zur Mitschrift von Otto Hufeland vgl. Deutsches Textarchiv 2007–2019, Hufeland 1829.

Gesamt

In der letzten Grafik wurden alle Angaben noch einmal über alle Handschriften zusammengefasst. Es zeigt sich, dass Benjamin Fiechter einen großen Teil der <unclear>-Elemente vergeben hat, Christian Thomas einen kleineren, dafür aber fast alle <supplied>-Elemente.

Für die Erstellung eines persönlichen Auszeichnungsprofils von Edierenden wäre es am besten, verschiedene Edierende unabhängig voneinander die gleiche Quelle annotieren zu lassen und die Ergebnisse zu vergleichen. Liegen ausreichend Daten vor, wäre es denkbar, über Dokumentengrenzen hinweg persönliche Auszeichnungsprofile der Edierenden zu erstellen. Mit diesen Profilen könnten die verschiedenen, in der TEI möglichen Werte für die Attribute von Unsicherheit, näher bestimmt und möglicherweise auch vereinheitlicht werden.

4. Zusammenfassung

Die Nähe der Richtlinien von TEI einsetzenden Vorhaben wie DTA, PROPYLÄEN, und anderen, z. B. der Carl-Maria-von-Weber-Gesamtausgabe (WEGA), zu den TEI-Guidelines macht die Daten dieser Editionen im Bereich des allgemeinen Umgangs mit Textlücken und unsicheren Lesarten (auch mit XML-Mitteln) vergleichbar. Gleichzeitig erleichtert diese Nähe auch die Entwicklung von spezielleren TEI-Import-Routinen für Neo4j.

Schwieriger ist dies jedoch für den Vergleich von Gewichtungen in der Sicherheit (Attribut @cert) der Auflösung von unsicheren Lesarten oder Textergänzungen. Hier spielt die subjektive Entscheidung des Edierenden eine zentrale Rolle. Die Angaben im Attribut @resp lassen sich hier allerdings heranziehen, um einen Eindruck zu bekommen, wie Edierende in welchen Fällen gewichtet.[34] Stehen genügend Daten zur Verfügung, könnte aus den Annotationen ein persönliches Auszeichnungsprofil des Bearbeitenden erstellt werden.

[34] Zugleich wird damit auch eine Anforderung an digitale Editionen erfüllt, nämlich: »Every act of editing in a digital edition should be attributed explicitly to the person who did it.« Robinson 2013.

Bibliographische Angaben

Marjorie Burghart / Malte Rehbein: The Present and Future of the TEI Community for Manuscript Encoding. In: Journal of the Text Encoding Initiative (2012), H. 2. Artikel vom 03.02.2012. DOI: 10.4000/jtei.372

Carl-Maria-von-Weber-Gesamtausgabe. Digitale Edition. Editionsrichtlinien zur Ausgabe der Briefe, Tagebücher und Dokumente Webers. Hg. von Gerhard Allroggen. Version 3.2.1 vom 08.01.2018. [online]

Deutsches Textarchiv. Grundlage für ein Referenzkorpus der neuhochdeutschen Sprache. Hg. von der Berlin-Brandenburgischen Akademie der Wissenschaften, Berlin 2007–2019. [online]

DTA-Basisformat. Das von CLARIN-D und der DFG empfohlene TEI-Format für historische Texte. Hg. vom Zentrum Sprache an der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW). Berlin 2011–2018. [online], hier besonders DTA-Basisformat Manuskript: [online]

Johann Wolfgang Goethe. Tagebücher. Hg. von Wolfgang Albrecht / Andreas Döhler. Band I,1. 1775–1787. Bisher unveröffentlichte retrodigitalisierte Datenfassung. Druckfassung: Stuttgart 1998.

Andreas Kuczera: Graphentechnologien in den digitalen Geisteswissenschaften. Modellierung – Import – Analyse. Github Pages. August 2018–. [online]

Andreas Kuczera (2017a): Graphentechnologien in den Digitalen Geisteswissenschaften. In: ABI Technik 37 (2017) H. 3. 15.09.2017. DOI: 10.1515/abitech-2017-0042

Andreas Kuczera (2017b): Das Deutsche Textarchiv in der Graphenwelt. In: Mittelalter. Interdisziplinäre Forschung und Rezeptionsgeschichte. Blogbeitrag vom 04.04.2017, aktualisiert am 06.06.2017. [online]

Gotthilf Patzig: Vorträge über physische Geographie des Freiherrn Alexander von Humbold: gehalten im großen Hörsaale des Universitäts-Gebäudes zu Berlin im Wintersemester 1827/28 vom 3ten Novbr. 1827. bis 26 April 1828. Aus schriftlichen Notizen nach jedem Vortrage zusammengestellt vom Rechnungsrath Gotthilf Friedrich Patzig. Berlin 1827/28 (= Nachschrift der ‚Kosmos-Vorträge' Alexander von Humboldts in der Berliner Universität, 3.11.1827–26.4.1828), S. 9. In: Deutsches Textarchiv. Grundlage für ein Referenzkorpus der neuhochdeutschen Sprache. Hg. von der Berlin-Brandenburgischen Akademie der Wissenschaften. Berlin 2007–2019. [online]

Peter Robinson: Five Desiderata for Scholarly Editions in Digital Form. In: Proceedings of Digital Humanities (University of Nebraska–Lincoln, 16.–19.07.2013). Long Paper vom 19.07.2013. [online]

TEI Guidelines. P5: Guidelines for Electronic Text Encoding and Interchange. Version 3.4.0. Revision 1fa0b54 vom 23.07.2018. [online]

Christian Thomas / Benjamin Fiechter / Marius Hug: Methoden und Ziele der Erschließung handschriftlicher Quellen zu Alexander von Humboldts Kosmos-Vorträgen. Das Projekt Hidden Kosmos der Humboldt-Universität zu Berlin. In: Horizonte der Humboldtforschung: Natur, Kultur, Schreiben. Hg. von Ottmar Ette / Julian Drews. Hildesheim u. a. 2016, S. 287–318. (= Potsdamer inter- und transkulturelle Texte (Pointe), 16). [Nachweis im GBV] Siehe auch Preprint PDF [online]

Abbildungsverzeichnis

ZfdG

Zeitschrift für digitale Geisteswissenschaften

Beitrag aus:
Sonderband 4 der ZfdG: Die Modellierung des Zweifels – Schlüsselideen und -konzepte zur graphbasierten Modellierung von Unsicherheiten. Hg. von Andreas Kuczera, Thorsten Wübbena und Thomas Kollatz. 2019. DOI: 10.17175/sb004

Titel:
Academic Meta Tool – Ein Web-Tool zur Modellierung von Vagheit

Autor/in:
Martin Unold

Kontakt: martin.unold@hs-mainz.de
Institution: Hochschule Mainz University of Applied Sciences, Institut für Raumbezogene Informations- und Messtechnik
GND: 1084131374 ORCID: 0000-0003-2913-2421

Autor/in:
Florian Thiery

Kontakt: thiery@rgzm.de
Institution: Römisch-Germanisches Zentralmuseum – Leibniz-Forschungsinstitut für Archäologie
GND: 1169955746 ORCID: 0000-0002-3246-3531

Autor/in:
Allard Mees

Kontakt: mees@rgzm.de
Institution: Römisch-Germanisches Zentralmuseum – Leibniz-Forschungsinstitut für Archäologie
GND: 124281400 ORCID: 0000-0002-7634-5342

DOI des Artikels:
10.17175/sb004_004

Nachweis im OPAC der Herzog August Bibliothek:
1037071964

Erstveröffentlichung:
28.02.2019

Lizenz:

Sofern nicht anders angegeben (cc) BY-SA

Medienlizenzen:
Medienrechte liegen bei den Autoren

Letzte Überprüfung aller Verweise:
28.02.2019

GND-Verschlagwortung:
konzeptionelle Modellierung | Vagheit | Softwaresystem |

Zitierweise:
Martin Unold, Florian Thiery, Allard Mees: Academic Meta Tool – Ein Web-Tool zur Modellierung von Vagheit . In: Die Modellierung des Zweifels – Schlüsselideen und -konzepte zur graphbasierten Modellierung von Unsicherheiten. Hg. von Andreas Kuczera, Thorsten Wübbena und Thomas Kollatz. 2019 (= Sonderband der Zeitschrift für digitale Geisteswissenschaften, 4). PDF Format ohne Paginierung. Als text/html abrufbar unter DOI: 10.17175/sb004_004.

Martin Unold, Florian Thiery, Allard Mees

Academic Meta Tool – Ein Web-Tool zur Modellierung von Vagheit

Abstracts

In diesem Artikel stellen wir eine Methodik zur Modellierung von Vagheit in Graphen vor. Neben der Modellierung behandeln wir auch die automatisierte Generierung von implizit gespeichertem Wissen unter Berücksichtigung von Vagheit. Diese wendet Verfahren aus dem Gebiet der Beschreibungslogik auf graphbasierte Daten an. Ebenfalls präsentieren wir in diesem Artikel unsere Softwareentwicklungen, welche die beschriebene Methodik umsetzen und zeigen deren Nutzen anhand von drei Fallbeispielen in den Geistes- und Kulturwissenschaften auf.

In this article, we introduce a methodological proposal for modelling vagueness in graphs. In addition to the modelling, we also deal with the automatic generation of implicitly stored knowledge when considering vagueness. We use the ideas of algorithms designed for description logics and apply them on graph data. We also present our software development that implements the proposed methodology. We will demonstrate the use of our applications based on three use cases in humanities and cultural studies.

1. Einleitung

Graphdatenbanken und Triplestores stellen bei der Modellierung von Forschungsdaten eine Alternative zu relationalen Datenbanken dar. In den letzten Jahren entwickelten sich in den digitalen Geistes- und Kulturwissenschaften umfangreiche Communities wie Pelagios Commons, welche graphbasierte Netzwerke in Graphdatenbanken und Triplestores modellieren und Forschungsdaten zur Verfügung stellen. Im Vergleich zu einer Tabellenstruktur eignet sich eine Modellierung der Daten in einem Graphen besser um hochvernetzte Forschungsdaten interoperabel zur Verfügung zu stellen, indem man zwei Knoten aus verschiedenen Ressourcen miteinander über eine Kante verbindet. Dieser Technik bedienen sich viele Forschungsprojekte durch die Verwendung von Linked Open Data (LOD).

Dabei ist in den meisten Fällen jedoch keine Modellierung von Vagheit oder Unsicherheit möglich. Speziell in der Archäologie und hier insbesondere bei der Zuweisung von Darstellungen zu übergeordneten Konzepten tritt das Problem der Modellierung des Zweifels in hohem Maße auf. Werden zum Beispiel Motive auf Münzen oder Darstellungen auf südgallischer Terra Sigillata detektiert und gespeichert, so trifft die Übereinstimmung häufig nur zu einem gewissen Grad zu. Diese Zweifel werden traditionell und aus historischen Gründen, wie bei der Samian Research-Datenbank des Römisch-Germanischen Zentralmuseums (RGZM), in einer relationalen Tabellenstruktur gespeichert und beinhalten das Zeichen ›?‹ oder andere Kombinationen, wie beispielsweise ›15/17R or 18/31R‹. Abgesehen davon gibt es auch Bestrebungen, diese Datenbanken mittels LOD zur Verfügung zu stellen und somit auch die grundlegenden Daten transparent und nachvollziehbar bereitzustellen.

In den Digital Humanities werden zur Verschlagwortung häufig Fachthesauri und Taxonomien verwendet, die zumeist als LOD via ›Simple Knowledge Organisation System‹, kurz SKOS, im Web zur Verfügung stehen oder gestellt werden. Hierbei werden jedoch bewusst nur vage Aussagen über den Grad der Verbindung zweier Knoten zugelassen (A skos:related B) und die Transitivität eingeschränkt, da sonst ungewollte Schlussfolgerungen auftreten.

Neben der Herausforderung der Verschlagwortung spielen in den Geistes- und Kulturwissenschaften auch Personennetzwerke eine Rolle. Hier bestehen Verbindungen zwischen den jeweiligen Personen-Instanzen oft nur zu einem gewissen Grad, welche sich spezifisch semantisch modellieren lassen – zum Beispiel Verwandtheitsgrad vs. lockere Bekanntschaft. Hier bietet das Semantic Web die Friend of a Friend (FOAF) Ontologie an, in welcher der Grad der Verbindung mittels foaf:knows jedoch nur mit null- oder hundertprozentiger Intensität angegeben werden kann. Es gibt also mit FOAF keine Möglichkeit etwas über die Intensität der Beziehung aussagen zu können. Diese Modellierungen, welche zum Beispiel in Social-Media-Netzwerken (Twitter, Facebook, Instagram, etc.) genutzt werden, bieten keine adäquaten Verfahren für die Anwendung in wissenschaftlichen Personennetzwerke an.

In allen Modellierungen von Graphen, ob nun in einer Graphdatenbank oder in einem Triplestore, tritt ein wie zuvor beschrieben häufiges Problem auf: die Vagheit von Kanten bzw. Aussagen. Das bedeutet, dass eine Verbindung zwischen zwei Knoten bzw. Ressourcen nur zu einem gewissen Grad besteht. Dies ist nicht zu verwechseln mit Unsicherheit, bei der unbekannt ist, ob die Verbindung überhaupt besteht. Bei Personen-Netzwerken wäre ein Beispiel, wenn die Beziehung nicht besonders intensiv ist, es sich also subjektiv eher um eine Bekanntschaft als eine Freundschaft handelt. Sind in einem Datensatz viele Freundschaftsbeziehungen vorhanden, die aber eine unterschiedliche Intensität bedeuten, müsste man entweder alle auf gleiche Weise verknüpfen oder sehr viele verschiedene Verknüpfungen erfinden, die aber im Wesentlichen das Gleiche bedeuten.

Sowohl gängige Graphdatenbanken als auch gebräuchliche Triplestores bieten jedoch keine Möglichkeit, Unsicherheiten oder Vagheiten zu modellieren. Das ›Academic Meta Tool‹ (AMT) greift dieses Problem auf und bietet dem oder der Nutzer*in an, Kantengewichte einzufügen und darauf Inferenz unter Berücksichtigung von Vagheit vorzunehmen. AMT bietet also die Chance, sämtlichen Kanten eine Gewichtung hinzuzufügen, um dadurch die Vagheit dieser Kante auszudrücken. Eine Beziehung zwischen zwei Knoten besteht also nur zu einem gewissen Grad. Dieser Grad, d.h. das Kantengewicht, wird üblicherweise in Prozent angegeben. AMT beinhaltet zusätzlich ein Verfahren, mit dem – unter Zuhilfenahme einer vordefinierten Ontologie – aus vorhandenen graphbasierten Daten automatisch Schlussfolgerungen gezogen werden können (Reasoning).

Dieser Artikel ist wie folgt gegliedert: Zunächst stellen wir Arbeiten vor, die mit unserer Studie verwandt sind. Wir gehen ebenfalls auf ähnliche Softwarelösungen ein und grenzen diese gegenüber unserer Entwicklungen ab. Danach folgt ein einleitendes Kapitel zum Thema Vagheit in Graphen. Es fasst kurz die wichtigsten theoretischen Informationen rund um das Thema Vagheit zusammen. Im vierten Kapitel beschreiben wir die von uns

entwickelte Methodik und gehen auch bereits auf mögliche Implementierungen in Form von Programmcode ein. Im fünften Kapitel präsentieren wir konkrete Web-Anwendungen und deren Einsatz in Projekten der Digital Humanities. Zum Schluss fassen wir die wichtigsten Punkte des Artikels zusammen und ziehen ein Fazit.

2. Verwandte Arbeiten

Karsten Tolle und David Wigg-Wolf[1] beschreiben in ihrer Arbeit ›Uncertainty Handling for Ancient Coinage‹ einen Vorschlag zur semantischen Modellierung von Linked Data – genauer geht es um die Beschreibung von Unsicherheiten bei der Bestimmung von Münzdarstellungen. Hier wird insbesondere die W3C Uncertainty Ontology[2] genutzt.

Die W3C Uncertainty Ontology (un) basiert darauf, dass eine Aussage (un:Sentence) mit einer Unsicherheit behaftet ist (un:Uncertainty), welche unterschiedliche Ausprägungen besitzt: un:UncertaintyType (Klassifikation der Unsicherheit, wie Ambiguity, Empirical, Vagueness, Inconsistency, Incompleteness), un:UncertaintyNature (Aleatory oder Epistemic), un:UncertaintyDerivation (Angaben, wie die Unsicherheit entstanden ist, z.B. objektiv oder subjektiv) und un:UncertaintyModel (mathematische Theorien für Uncertainty Types wie Probability oder RandomSets).

Die in den Editionswissenschaften weit verbreitete Auszeichnungssprache Text Encoding Initiative (TEI) nutzt das Element ›certainty‹ und ›precision‹ zur Beschreibung einer Unsicherheit.[3]

Aussagen oder Annotationen ohne genaue Angabe eines Grades können im Semantic Web mit der ›Open Annotation Ontologie‹ verarbeitet werden. Hier werden zwei Ressourcen über eine Annotation und Body- und Target Attributen miteinander verknüpft.[4] Die ›Pelagios Commons Initiative‹ nutzt z.B. diese Ontologie zur Verknüpfung von Datensätzen und Ressourcen des Gazetteers Pleiades.[5]

Das ›Simple Knowledge Organisation System‹, kurz SKOS, ist eine formale Sprache zur Kodierung von Schlagworten in Thesauri und Klassifikationen oder anderen kontrollierten Vokabularen mit Hilfe des Resource Description Framework (RDF) und RDFS-Schemas.[6] SKOS bietet die Möglichkeit über semantische Relationen und mapping properties vage Beziehungen zwischen skos:Concepts auszudrücken. Hierbei stellt sich jedoch das Problem der (nicht ermöglichten) Transitivität sowie die generelle Problematik ungenauer Aussagen der Relationen, die nicht quantitativ messbar und auswertbar sind:

[1] vgl. Tolle / Wigg-Wolf 2015.
[2] vgl. Laskey et al. 2008.
[3] vgl. Text Encoding Initiative Consortium 2018.
[4] vgl. Sanderson et al. 2017.
[5] vgl. Muccigrosso 2018.
[6] vgl. Miles / Bechhofer 2009.

> »The property skos:related is used to assert an associative link between two SKOS concepts.«[7]

> »A skos:closeMatch link indicates that two concepts are sufficiently similar that they can be used interchangeably in some information retrieval applications. A skos:exactMatch link indicates a high degree of confidence that two concepts can be used interchangeably across a wide range of information retrieval applications.«[8]

> »Note that skos:related is not a transitive property.«[9]

> »<A>skos:exactMatch<B>. <B> skos:exactMatch <C>. entails <A> skos:exactMatch <C> . All other SKOS mapping properties are not transitive.«[10]

Aus der SKOS Ontologie haben sich bereits einige SKOS-Editoren wie ›‹ oder ›‹ im Web etabliert. Darüber hinaus arbeitet das Mainzer Zentrum für Digitaltität in den Geistes- und Kulturwissenschaften (mainzed) an einer eigenen, für geisteswissenschaftliche Belange angepassten, Web-App: dem Labeling System[11].

Ein Beispiel für eine Erweiterung des CIDOC Conceptual Reference Model (CRM) um Attribute der Unsicherheit haben bereits Bruhn et al.[12] in ihren Arbeiten mit dem Arches Project System vorgestellt. Dabei werden die Unsicherheiten mit dem Typ E55 in den Zeitangaben, sowie die in den Fundzuweisungen modelliert.[13]

Die Behandlung von Vagheit in formal aufbereiteten Wissensbeständen, insbesondere für Beschreibungslogiken, ist bereits in einer Vielzahl an Programmen umgesetzt.[14] Eine Implementierung, die unserer sehr nahe kommt, ist die Web-Anwendung THRILL on SWISH.[15] Allerdings verwendet sie Unsicherheit statt Vagheit. Wir werden im nächsten Kapitel noch näher auf die Unterschiede zwischen diesen beiden eng verwandten Theorien eingehen.

3. Vagheit in Graphen

In diesem Kapitel stellen wir grundlegende Begriffe und Ideen zur Behandlung von Vagheit in Graphen vor. Zunächst grenzen wir den Begriff der Vagheit gegenüber dem Begriff der Unsicherheit ab. Anschließend befassen wir uns mit der Modellierung von

[7] vgl. Miles / Bechhofer 2009, Kapitel 8.1.
[8] vgl. Miles / Bechhofer 2009, Kapitel 10.6.8.
[9] vgl. Miles / Bechhofer 2009, Kapitel 8.6.4.
[10] vgl. Miles / Bechhofer 2009, Kapitel 10.6.3.
[11] vgl. Thiery / Engel 2016 und Piotrowski et al. 2014.
[12] vgl. Bruhn et al. 2015, S. 345–346.
[13] vgl. Kohr 2014a, Kohr 2014b.
[14] vgl. Stoilos et al. 2005, Bobillo et al. 2008, Bobillo et al. 2013, Tsatsou et al. 2014.
[15] vgl. Bellodi et al. 2017.

Vagheit in graphbasierten Daten, sowie deren Verarbeitung. Dabei legen wir ein besonderes Augenmerk auf die automatisierte Generierung von implizit gespeichertem Wissen mit Hilfe von Regelwerken.

Vagheit ist ein Maß für die Präzision einer Aussage. Eine vage Aussage trifft also nur zu einem gewissen Grad zu. Trifft beispielsweise der Wetterbericht die Aussage »Morgen wird es Niederschlag geben«, so könnte morgen ein leichtes Nieseln, ein mäßiger Regen oder ein schweres Gewitter stattfinden. Abhilfe könnte hier beispielsweise die Angabe der Niederschlagsmenge leisten. Doch nicht zu verwechseln ist eine solche vage Aussage mit einer unsicheren Aussage. Bei Unsicherheit ist gänzlich unbekannt, ob die getroffene Aussage überhaupt wahr ist. Trifft der Wetterbericht beispielsweise die Aussage »Morgen regnet es mit 75 %-iger Wahrscheinlichkeit«, dann handelt es sich um eine unsichere Aussage. Sie sagt aus, dass in drei von vier Fällen die Kernaussage wahr ist, es also morgen regnet und in einem von vier Fällen falsch ist, es also morgen nicht regnet. Dubois und Prade beschreiben eine ausführlichere Klarstellung der Unterschiede zwischen Vagheit und Unsicherheit.[16] In diesem Artikel behandeln wir nur vage (und keine unsicheren) Aussagen und wir gehen davon aus, dass alle vagen Aussagen mit Werten zwischen 0 und 1 ausgedrückt werden können. Beispielsweise könnte ein schwacher Regen mit dem Wert 0.25 (25 %) zur Aussage »Morgen wird es Niederschlag geben« angegeben werden.

Vagheit kann theoretisch an verschiedenen Stellen in einem Graphen auftreten, entsprechend komplex kann auch die Speicherung von Vagheit in Graphdatenbanken werden. Der am häufigsten auftretende Fall ist die Zuordnung eines Gewichts zu einer vagen Kante, das ausdrückt, zu welchem Grad oder mit welcher Intensität die Verbindung zwischen den zwei Knoten, besteht. Man könnte analog auch andere Informationen in einem Graphen mit einem Vagheitswert versehen, zum Beispiel die Typisierung der Knoten. Wir beschränken uns hier allerdings auf die Verwendung von Vagheit als Kantengewicht, genau genommen erlauben wir sogar nur Werte zwischen 0 und 1 für die Gewichtung. Ein solches Kantengewicht kann relativ einfach in Graphdatenbanken gespeichert werden, da diese Werte nicht besonders außergewöhnlich sind. Interessanter ist allerdings die Verarbeitung der Kantengewichte, wenn regelbasiert Schlussfolgerungen getroffen werden sollen, das heißt, wenn automatisiert neue Kanten entstehen, die ebenfalls Vagheitswerte tragen.

Dazu bedienen wir uns den Techniken, die üblicherweise für Beschreibungslogiken Anwendung finden und wenden diese auf graphbasierte Daten an. Das hat den Vorteil, dass die dadurch entstehenden Graphen unmittelbar anschlussfähig an andere Ontologien und Linked Open Data sind. Eine in Aussagenlogik interpretierte Transformation einer vagen Beschreibungslogik erfolgt durch die Verwendung einer mehrwertigen Logik.[17] Der Nachteil mehrwertiger Logiken ist, dass innerhalb dieser nicht alle Gesetze der klassischen Aussagenlogik gelten können, wie zum Beispiel das Gesetz von De Morgan oder das der doppelten Negation. Da dieser Nachteil ohnehin besteht, haben wir uns entschieden, dem Benutzer unserer Software verschiedene mehrwertige Logiken anzubieten, die er beliebig

[16] vgl. Dubois et al. 2001.
[17] vgl. Lukasiewicz / Straccia 2008.

kombinieren kann. Dadurch ist es möglich, jeder Regel eine individuelle Interpretation zuzuweisen. Die folgende Grafik veranschaulicht die Verknüpfung zweier vager Aussagen durch Konjunktion und Disjunktion.

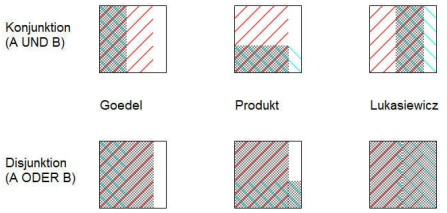


Abb. 1: Verknüpfung zweier vager Aussagen durch Konjunktion und Disjunktion. [Eigene Darstellung, CC BY 4.0].

Für weitere Ausführungen zum Thema Vagheit in Beschreibungslogiken verweisen wir auf das Literaturverzeichnis. Für die Verarbeitung von Vagheit in graphbasierten Datenbanken empfehlen wir insbesondere die Werke von Borzooei et al., Akram et al. und Castelltort et al..[18] Im folgenden Kapitel erläutern wir unsere Implementierungen näher.

4. Implementierung ›Academic Meta Tool‹

Nachdem wir im vorangegangenen Kapitel die theoretischen Grundlagen dargelegt haben, beschäftigen wir uns nun mit der Umsetzung unserer Überlegungen. Dabei beschreiben wir in diesem Kapitel vor allem die konzeptuellen Lösungsideen. Zunächst stellen wir die für das Academic Meta Tool entwickelte Meta-Ontologie vor. Diese Meta-Ontologie beschreibt eine Sprache zur Erstellung konkreter Ontologien für Anwendungsszenarien, stellt aber selbst kein Anwendungsszenario dar. Im Anschluss präsentieren wir die Umsetzung des Reasoning-Programms als JavaScript-Bibliothek. Auch diese Bibliothek ist kein vollständiges Programm, sondern lediglich ein Framework, das bei der Implementierung von konkreten Anwendungen für das Academic Meta Tool sehr nützlich sein kann. Die Implementierung einer Software, die direkt in Fallbeispielen eingesetzt werden kann, ist also erst Bestandteil des nächsten Kapitels.

4.1 Meta-Ontologie

Um das Academic Meta Tool zu nutzen, ist es zunächst erforderlich, eine Ontologie[19] zu entwickeln, welche das Schema und die Axiome für ein gewisses Anwendungsszenario beschreibt. Um eine solche Ontologie für das Academic Meta Tool zu erstellen, stehen bisher vier Typen von Aussagen zur Verfügung, die wir im Folgenden genauer erläutern werden.

[18] vgl. Borzooei et al. 2017, Akram et al.2014, Castelltort et al. 2014.
[19] Ontologie vgl. Unold / Thiery 2018c und Unold / Thiery 2018f, Vokabular vgl. Unold / Thiery 2018d und Unold / Thiery 2018e.

Wir demonstrieren die einzelnen Typen von Aussagen anhand einer Beispiel-Ontologie zur Modellierung von Orten, die in verschiedenen Himmelsrichtungen zueinander liegen. Außerdem ziehen wir jeweils einen Vergleich zu entsprechenden Ausdrücken in der Web Ontology Language (OWL).

Zunächst einmal ist es möglich, Kategorien für Knoten vorzugeben. Wir nennen solche Kategorien auch Konzepte. In OWL entspricht dies dem Prädikat owl:Class. Jedem Konzept kann ein Name und eine Kurzbeschreibung zugeordnet werden. In unserer Beispiel-Ontologie gibt es nur das Konzept ›Place‹. Analog können auch Kategorien für Kanten vorgegeben werden, die wir als ›Rollen‹ bezeichnen. In OWL entspricht dies dem Prädikat owl:ObjectProperty. Jeder Rolle kann ein Name zugeordnet werden sowie ein Konzept für Quellknoten (entspricht rdfs:domain) und Zielknoten (entspricht rdfs:range). In unserer Beispiel-Ontologie gibt es die Rollen northOf, eastOf, southOf und westOf. Sie haben alle jeweils Place sowohl als Quell- als auch als Zielknoten. Die folgende Grafik illustriert die Konzepte und Rollen.

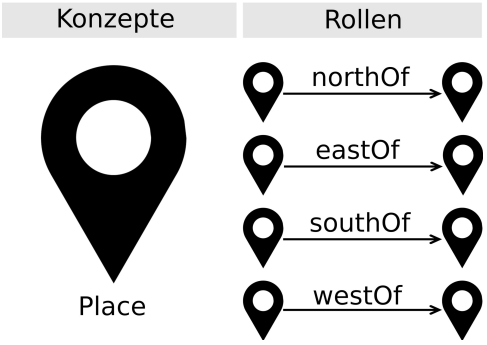


Abb. 2: Das Konzept Place und die Rolen northOf, eastOf, southOf, westOf. [Eigene Darstellung, CC BY 4.0].

Zusätzlich können zwei Typen von Axiomen formuliert werden. Der eine Typ ist die Rollen-Kettenregel. Sie entspricht ungefähr owl:ObjectPropertyChain in OWL 2. Allerdings ist im Academic Meta Tool neben der Angabe der Rollen in der Kette direkt die daraus resultierende Rolle anzugeben. Zusätzlich muss ebenfalls festgelegt werden, nach welcher mehrwertigen Logik (Lukasiewicz, Produkt oder Goedel) das Reasoning erfolgen soll. In unserem Beispiel formulieren wir das Axiom, dass sämtliche Rollen transitiv sind, unter Verwendung der optimistisch agierenden Goedel-Logik. In Abbildung 3 ist dies illustriert.

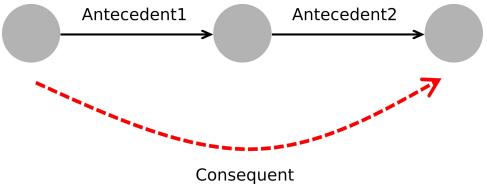


Abb. 3: Schematische Darstellung des Role-Chain-Axioms (Rollen-Kettenregel). [Eigene Darstellung, CC BY 4.0].

Der andere Typ von Axiomen ist die Inverse (vgl. Abbildung 4). Sie entspricht dem Prädikat owl:inverseOf in OWL. Hier sind eine Rolle und ihre Inverse anzugeben. In unserem Beispiel wären dies die Axiome, dass northOf die Inverse von southOf ist und dass eastOf die Inverse von westOf ist.

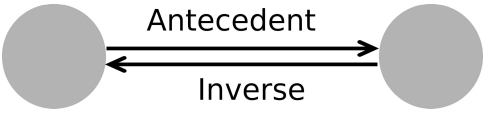


Abb. 4: Schematische Darstellung des Inverse-Axioms. [Eigene Darstellung, CC BY 4.0].

4.2 JavaScript Bibliothek

Zur Implementierung von Web-Editoren für konkrete Anwendungsfälle haben wir eine JavaScript-Bibliothek entwickelt. Diese bietet Funktionalitäten zur Datenverwaltung, zur Kommunikation mit einem Datenbank-Server (hier: RDF4J Triplestore), sowie ein Reasoning-Programm. An dieser Stelle möchten wir die Funktionalitäten nicht im Detail ausführen und verweisen auf die Veröffentlichung der Bibliothek.[20]

Die genannte JavaScript-Bibliothek dient dazu, eigene Ontologien, die sich des Academic Meta Tools bedienen, im Web zu veröffentlichen und Benutzern für die Dateneingabe zur Verfügung zu stellen. Jede Beispielontologie benötigt die Implementierung eines individuellen Webviewers zur Anzeige und zum Editieren der Daten. Ein generischer Ansatz wäre zwar auch denkbar, es hat sich allerdings herausgestellt, dass die Benutzerführung mit angepassten Oberflächen besser funktioniert.

5. Anwendungsbeispiele

Anhand von drei Anwendungsbeispielen werden wir in den folgenden Unterkapiteln die konkrete Implementierung der AMT-Ontologie in geisteswissenschaftliche Fragestellungen erörtern. Kapitel 5.1 befasst sich mit einer Ontologie zu einem Expert*innen-Netzwerk des mainzed.[21] In Kapitel 5.2 wird ein Töpfer-Netzwerk südgallischer Terra Sigillata-Punzen mittels einer AMT-Ontologie beschrieben.[22] Im dritten Anwendungsbeispiel beschäftigen wir uns mit der Beschreibung von Darstellungen auf archäologischen Kleinfunden mit Hilfe des Academic Meta Tools.[23]

[20] vgl. Unold / Thiery 2018a.
[21] vgl. Unold / Thiery 2018g.
[22] vgl. Thiery / Mees 2018c.
[23] vgl. Thiery / Mees 2018d.

5.1 mainzed Expert*innen-Netzwerk

Das mainzed ist eine Verbundinitiative und ein offenes Netzwerk zur Sammlung von digitalen Kompetenzen in Mainz.[24]

Mainzed bietet Wissenschaftlerinnen und Wissenschaftlern am Standort Mainz eine gemeinsame Plattform zum Austausch von Wissen, der Entwicklung von Projekten und dem Ausbau von eigenen mainz(ed)-spezifischen Forschungsschwerpunkten. Alle Mitglieder des mainzed-Netzwerks stehen in fachlichen und auch hierarchischen Beziehungen zueinander. Zudem haben die Mitglieder des sehr heterogenen Netzwerks viele unterschiedliche Interessen, die in einer SKOS-ähnlichen Taxonomie (z.B. skos:broader, skos:narrower, skos:related) modellierbar sind. An einem Community Day haben die Mitglieder des Interessennetzwerks jeweils ihre Sicht auf das Netzwerk dargestellt, das heißt ihre Beziehungen und Interessen, sowie deren Vagheitsgrad angegeben. AMT eignet sich nun dazu, dieses Netz zu visualisieren und weitere Schlussfolgerungen zu ziehen – das heißt neue Beziehungen zwischen mainzed-Mitgliedern und mögliche neue Interessensgebiete aufzeigen, welche bei der Eingabe nicht erkennbar waren. Dieses Verfahren kann zu einem intelligenten Personen-Interessen-Netzwerk des mainzed führen, was einen Mehrwert für jedes Individuum des Netzwerks darstellt.

Zur Implementierung dieses mainzed Personen-Interessen-Netzwerks muss eine eigene mainzed-Ontologie nach der bereits in Kapitel 4 vorgestellten Academic Meta Tool-Ontologie entwickelt werden. Diese Ontologie besteht aus zwei Konzepten, fünf Rollen und zwölf Axiomen.[25]

Die mainzed-Ontologie beinhaltet das Konzept Person (P) und das Konzept Interest (I). Zur Verknüpfung dieser Konzepte sind Rollen implementiert, welche sowohl assoziative Beziehungen wie auch inverse Beziehungen zwischen den Personen und Interessen beinhalten: P connectedWith P, P interestedIn I, I interestOf P, I2 subInterestOf I1, P1 superInterestOf P2, vgl. Abbildung 5.

[24] vgl. Klammt 2018.
[25] vgl. Unold / Thiery 2018b.

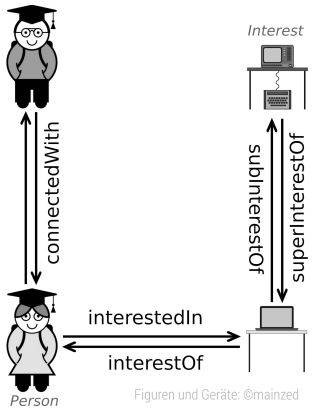


Abb. 5: Konzepte und Rollen der mainzed-Ontologie. [Eigene Darstellung, Figuren und Geräte: © mainzed, CC BY 4.0].

Zur Bildung neuen Wissens im Netzwerk werden fünf Rollen-Kettenregeln (vgl. Abbildung 6) eingeführt, welche durch Wahl einer geeigneten mehrwertigen Logik Schlussfolgerungen ziehen lassen. Diese sind:

- Axiom 01: connectedWith connectedWith connectedWith(ProductLogic)
- Axiom 02: subInterestOf subInterestOf subInterestOf(GoedelLogic)
- Axiom 03: superInterestOf superInterestOf superInterestOf(GoedelLogic)
- Axiom 04: interestedIn interestOf connectedWith(LukasiewiczLogic)
- Axiom 05: interestedIn subInterestOf interestedIn(GoedelLogic).

Zudem werden fünf inverse Axiome und zwei disjunkte Axiome hinzugefügt:

- Axiom 06: interestOf interestedIn
- Axiom 07: interestedIn interestOf
- Axiom 08: subInterestOf superInterestOf
- Axiom 09: superInterestOf subInterestOf
- Axiom 10: connectedWith connectedWith
- Axiom 11: selfdisjoint(subInterestOf)
- Axiom 12: selfdisjoint(superInterestOf).

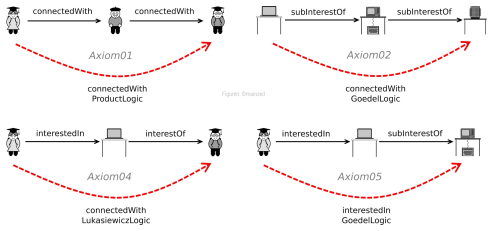


Abb. 6: Rollen-Kettenregeln der mainzed-Ontologie. [Eigene Darstellung, Figuren und Geräte: © mainzed, CC BY 4.0].

Aus der Ontologie und den Rollen-Kettenregeln entstehen folgende Schlussfolgerungen: Beziehungen zwischen Personen sind invers und transitiv. Somit ergeben sich hier Beziehungsketten, vgl. Axiom 01. Hierarchische Beziehungen zwischen Interessen sind ebenfalls invers und transitiv aufgebaut, vgl. Axiom 02 / 03. Interessieren sich zwei Personen für das Gleiche, so stehen sie miteinander in Verbindung, vgl. Axiom 04. Interessiert sich eine Person für etwas, so interessiert sie sich auch für das ›Super-Interesse‹, vgl. Axiom 05.

Führen wir dies nun an einem konkreten Beispiel aus: Im mainzed-Netzwerk existieren die virtuellen Personen Emma, Ben und Fynn. Aus einer Vielzahl von Interessen im Netzwerk und der genannten Personen entnehmen wir Informatik, Programmieren und Java. Auf dem zuvor erwähnten Community Day wurden die persönlichen Beziehungen wie folgt angegeben: (Person:Ben)-[connectedWith:0.8]->(Person:Fynn) und (Person:Fynn)-[connectedWith:0.6]->(Person:Emma). Die Interessen stehen in einer Kette wie folgt in Verbindung:

(Interest:Informatik)-[superInterestOf:0.6]->(Interest:Programmieren)

und

(Interest:Programmieren)-[superInterestOf:0.5]->(Interest:JAVA).

Zum Interesse JAVA stehen Ben und Fynn mit einem gewissen Grad in Verbindung:

(Person:Ben)-[interestedIn:0.7]->(Interest:JAVA)

und

(Person:Fynn)-[interestedIn:0.9]->(Interest:JAVA).

Die mainzed-Ontologie erlaubt es nun, Schlussfolgerungen aus diesen Eingaben zu ziehen: Axiom 01 ermöglicht durch die Produkt-Logik die Verknüpfung zwischen Ben und Emma via Fynn zu 48 %, Axiom 03 die Verknüpfung durch die Goedel-Logik zwischen Informatik zu JAVA via Programmieren zu 50 %, Axiom 04 die Verknüpfung durch die Lukasiewicz Logik zwischen Ben und Fynn via JAVA zu 60 % und Axiom 05 die Verknüpfung durch die Goedel Logik zwischen Fynn und Programmieren via JAVA zu 50 %.

Nach einer Eingabe der zuvor beschrieben Daten in den prototypischen Academic Meta Tool Playground (vgl. Abbildung 7) bietet dieser ein Reasoning der Daten, einen Download als RDF, cypher, Knoten und Kanten als CSV, sowie als JSON-Objekt.

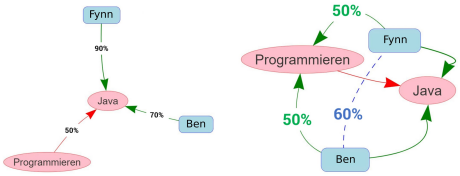


Abb. 7: Schlussfolgerungen der mainzed-Ontologie im Web-Viewer. [Eigene Darstellung, CC BY 4.0].

5.2 Töpfer-Netzwerk südgallischer Terra Sigillata

Betrachten wir eine konkrete archäologische Problematik, die es zu modellieren gilt: In den römischen Sigillata-Manufakturen wurde vom 1. Jahrhundert v. Chr. bis zum 3. Jahrhundert n. Chr. sehr hochwertiges, reliefverziertes Tafelgeschirr hergestellt, das flächendeckend im Römischen Reich vermarktet wurde. Darüber hinaus wurden die einzelnen Gefäße mit dem Namen des Töpfers gestempelt, was die namentliche Zuweisung der Zierzonen ermöglicht. Die Online-Datenbank Samian Research des Römisch-Germanischen Zentralmuseums beinhaltet zurzeit mehr als 245.000 Töpferstempel aus ganz Europa und auch ein Katalog der Reliefverzierungen befindet sich momentan im Aufbau. Das Ziel dieser Katalogerstellung ist es, einzelne Figurenpatrizen (z.B. Bogenschützen), womit die Gefäße innerhalb einer Sigillata-Zierzone dekoriert wurden, einem Töpfer zuordnen zu können. Da diese Figurenpunzen von den Töpfern nicht nur miteinander geteilt, sondern die Figurenstempel auch voneinander abgeformt wurden, spielt in diesem Bestimmungsprozess eine genaue Identifikation eine sehr große Rolle.

Abb. 8: Darstellung von Bogenschützen-Punzen. [Eigene Darstellung, © RGZM / Mees, CC BY 4.0].

So müssen z.B. im Zuge des Forschungsprojektes dutzende Bogenschützen-Punzen (vgl. Abbildung 8) miteinander abgeglichen werden um die einzelnen Patrizen zu benennen. Jedoch gibt es zwischen der originalen Punze aus Ton und einer möglichen Abformung einen transitiven Schwund, der durch Trocknung entsteht. Dieser Schwund beträgt ca. 10 % pro Abformungsschritt.[26]

[26] vgl. Hoffmann 1983.

Analog zur mainzed-Ontologie muss hierzu eine eigene samian ontology implementiert werden. Diese folgt ebenfalls der Academic Meta Tool-Ontologie. Die samian ontology besteht aus zwei Konzepten, sechs Rollen und ebenfalls zwölf Axiomen.[27]

Die samian ontology beinhaltet hier das Konzept Töpfer (T) und das Konzept Punze (P). Zur Verknüpfung dieser Konzepte sind Rollen implementiert, welche sowohl assoziative Beziehungen, wie auch inverse Beziehungen zwischen den Töpfern und den Punzenabformungen beinhaltet. Dabei ist zu beachten, dass aus archäologischer Perspektive die Rollen ›isCreatorOf‹ und ›wasCreatedBy‹ mit einem Grad von 1.0 zu versehen sind und die Abformungen ›istMutterpunzeVon‹ und ›istTochterpunzeVon‹ mit einem Grad von 0.9 notiert werden. Die Rollen sind: T isConnectedWith T, T isCreatorOf P, P wasCreatedBy P, P1 istMutterpunzeVon P2, P2 istTochterpunzeVon P1, T arbeitetInWerkstattMit P.

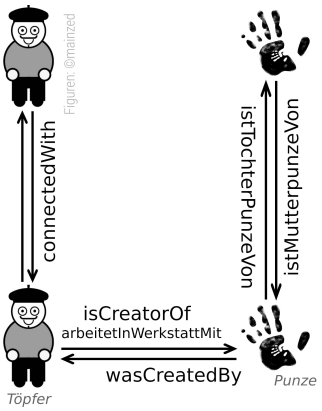


Abb. 9: Konzepte und Rollen der samian ontology. [Eigene Darstellung, Figuren und Geräte: © mainzed, CC BY 4.0].

Zur Verdeutlichung von Beziehungen zwischen Töpfern und Punzen werden nun sechs Rollen-Kettenregeln mit geeigneten Logiken eingeführt (vgl. Abbildung 9), welche Schlussfolgerungen zulassen. Diese sind:

- Axiom 01: isConnectedWith isConnectedWith isConnectedWith(ProductLogic)
- Axiom 02: istTochterpunzeVon istTochterpunzeVon istTochterpunzeVon(GoedelLogic)
- Axiom 03: istMutterpunzeVon istMutterpunzeVon istMutterpunzeVon(GoedelLogic)
- Axiom 04: isCreatorOf wasCreatedBy isConnectedWith(LukasiewiczLogic)
- Axiom 05: isCreatorOf istTochterpunzeVon arbeitetInWerkstattMit(GoedelLogic).

Zudem werden fünf inverse Axiome und zwei disjunkte Axiome hinzugefügt:

- Axiom 06: istTochterpunzeVon istMutterpunzeVon
- Axiom 07: istMutterpunzeVon istTochterpunzeVon

[27] vgl. Thiery / Mees 2018a.

- Axiom 08: isCreatorOf wasCreatedBy
- Axiom 09: wasCreatedBy isCreatorOf
- Axiom 10: isConnectedWith isConnectedWith
- Axiom 11: selfdisjoint(istTochterpunzeVon)
- Axiom 12: selfdisjoint(istMutterpunzeVon).

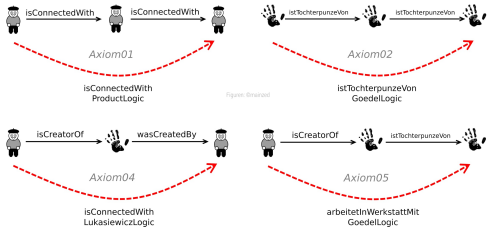


Abb. 10: Rollen-Kettenregeln der samian ontology. [Eigene Darstellung, Figuren und Geräte: © mainzed, CC BY 4.0].

Die samian ontology zeigt durch die Rollen-Kettenregeln (vgl. Abbildung 10) auf: Beziehungen zwischen Töpfern sind, wie bei den Personen in der mainzed-Ontologie, invers und transitiv. Somit ergeben sich hier Beziehungsketten, vgl. Axiom 01. Hierarchische Beziehungen zwischen Punzen aufgrund der Abformungen sind ebenfalls invers und transitiv aufgebaut, vgl. Axiom 02 / 03. Haben zwei Töpfer eine Beziehung zu einer Punze, so stehen sie miteinander in Verbindung, vgl. Axiom 04. Hat ein Töpfer eine Punze erstellt, so kann aus einer Tochterpunze geschlossen werden, dass der Töpfer wahrscheinlich in der gleichen Werkstatt gearbeitet hat, vgl. Axiom 05. Somit ergibt sich ein Töpfer-Personen-Netzwerk, welches zum Beispiel eine Grundlage für die wissenschaftliche Untersuchung von Pachtverhältnissen in den antiken Töpfereien bildet.

Ein ebenfalls für diesen Fall prototypisch entwickelter Academic Meta Tool Punzen-Viewer zeigt nun neue Verknüpfungen durch ein Reasoning auf und ermöglicht den Export in gängige Formate zur Weiterbearbeitung in externer Software (vgl. Abbildung 11).

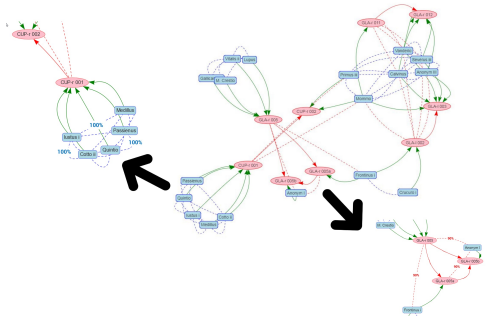


Abb. 11: Schlussfolgerungen der samian ontology im Web-Viewer. [Eigene Darstellung, CC BY 4.0].

5.3 Darstellungen auf archäologischen Kleinfunden

Bereits in den Ausführungen von Karsten Tolle und David Wigg-Wolf auf der CAA 2014 in Paris zum ›Uncertainty Handling for Ancient Coinage‹ wurde versucht, eine Lösung zur semantischen Modellierung mittels Linked Data für das folgende Problem zu finden: Auf einem archäologischen Kleinfund, hier einer Münze, ist ein Portrait abgebildet. Wichtig ist hier die eindeutige Identifizierung der Person zur weiteren Bearbeitung. Bei einer Befragung von Expert*innen konnte eine einhundertprozentige Sicherheit nicht gegeben werden: »Ich bin mir zu 80 % sicher, dass es sich bei der portraitierten Person um Titus handelt, aber es könnte auch 60 % Titus und 40 % Nero sein.«[28]

Eine ähnliche Problematik ergibt sich bei den NAVIS Schiffsdatenbanken des Römisch-Germanischen Zentralmuseums. In NAVIS II werden Darstellungen von Schiffen auf Mosaiken, Monumenten etc. im Web zur Verfügung gestellt, in NAVIS III sind Schiffsdarstellungen auf Münzen für die Scientific Community verfügbar. In beiden Datenbanken werden analog zum Fall von Tolle und Wigg-Wolf die Darstellungen einem Attribut zugeordnet, z.B. Titus und Nero, aber auch Handel und Krieg oder Paddeln und Rudern. Bislang werden diese Verknüpfungen 1:1 mit einer im Datenmodell 100 % möglichen Sicherheit modelliert. Um dieser sehr subjektiven Wahrnehmung eine Objektivität zu geben, wäre eine vage Verknüpfung, die nur zu einem gewissen Grad existiert, transparent und nachvollziehbar. Darüber hinaus wird heutzutage zur Standardisierung der Verschlagwortung (Keywords) von Objekt-Darstellungen zu Thesauri-Konzepten der Linked Data Cloud gelinkt (z.B. Getty AAT, English Heritage, etc.). In diesen Thesauri bestehen jedoch wiederum Abhängigkeiten zu einem gewissen Grad, welche zumeist mittels der benutzte SKOS Ontologie nicht exakt abgebildet werden kann. Zur inhaltlichen Erschließung ist dies jedoch nötig. Das Academic Meta Tool ist hervorragend dazu geeignet, den Prozess von der Darstellung auf dem Objekt zur Verschlagwortung in einem Keyword bis hin zur Verlinkung in ein Thesaurus-Konzept semantisch zu modellieren.

Für diesen Fall muss eine eigene navis ontology implementiert werden. Sie besteht aus drei Konzepten, sechs Rollen und ebenfalls 18 Axiomen.[29]

Die navis ontology enthält die Konzepte Object (O), Keyword (K) und Concept (C). Zur Verknüpfung zwischen Objekt und Keyword werden die Rollen ›O hasDepiction K‹ und ›K isDepictionOf O‹ genutzt. Zur Verknüpfung zwischen Keyword und Thesaurus-Konzept gibt es die die Rollen ›K matchesWith C‹ und ›C matchedBy K‹, sowie zur hierarchischen Ordnung im Thesaurus die Rollen ›C broaderThan C‹ und ›C narrowerThan C‹. Wir gehen hier davon aus, dass der Grad der Verknüpfung zunimmt, je weiter es in Richtung des Top-Level-Konzepts geht – der Grad der anderen Verbindungen ist von dem oder der Wissenschaftler*in selbst zu bestimmen. Abbildung 12 zeigt die Konzepte und Rollen auf.

[28] vgl. Tolle / Wigg-Wolf 2015, S.173.
[29] vgl. Thiery / Mees 2018b.

Abb. 12: Konzepte und Rollen der navis ontology. [Eigene Darstellung, CC BY 4.0].

Hier werden nun drei Rollen-Kettenregeln (einschließlich der jeweiligen Inverse) mit geeigneten Logiken eingeführt (vgl. Abbildung 13): Diese sind:

- Axiom 01: hasDepiction matchesWith matchesWith(ProductLogic)
- Axiom 02: matchesWith broaderThan broaderThan(ProductLogic)
- Axiom 03: broaderThan broaderThan broaderThan(ProductLogic).

Zudem werden sechs inverse Axiome und sechs disjunkte Axiome hinzugefügt.

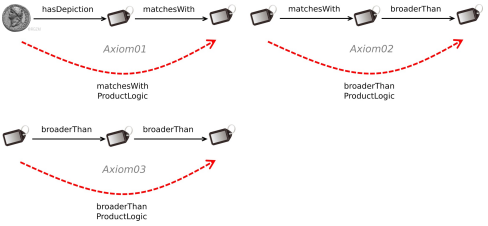


Abb. 13: Rollen-Kettenregeln der navis ontology. [Eigene Darstellung, CC BY 4.0].

Durch die Rollen-Kettenregeln in der navis ontology ergeben sich folgende Schlussfolgerungen: Wird ein Objekt mit einem Keyword verschlagwortet und dieses mit einem Konzept in einem Thesaurus verknüpft, besteht zu einem gewissen Grad auch eine Verknüpfung zwischen Objekt und dem Thesaurus-Konzept, vgl. Axiom 01. Ist dieses Thesaurus-Konzept ein broader-Konzept, wird auch das Keyword mit einem gewissen Grad damit verknüpft, vgl. Axiom 02. Zudem besitzen alle hierarchisch organisierten Keywords in den Thesauri Beziehungen in einem bestimmteren Grad zueinander, vgl. Axiom 03.

Wir zeigen hier zwei konkrete Beispiele aus den NAVIS Schiffsdatenbanken (vgl. Abbildung 14). Eine Darstellung zeigt ein Schiff – es stellt sich die Frage: Ist es ein Ruder- oder ein Segelschiff? Hier kann der Wissenschaftler oder die Wissenschaftlerin sich für ≥ 50 % Segelschiff bzw. ≥ 50 % Ruderschiff entscheiden. Eine weitere Darstellung zeigt ein Relief. Das darauf abgebildete Schiff könnte ein Transport- oder Militärschiff darstellen, da sowohl Weinfässer, als auch Soldaten abgebildet sind. Auch hier kann sich der Wissenschaftler oder die Wissenschaftlerin nun entscheiden, wohl ≥ 40 % Transportschiff bzw. ≥ 60 % Millitärschiff.

Abb. 14: links: Darstellung einer römischen Münze mit Schiffsdarstellung (O41650 aus NAVIS III) und eines Monuments (NeumagenMonument1 aus NAVIS II). [Eigene Darstellung, © RGZM, CC BY 4.0].

Ein kleiner Einblick kann auch hier über einen prototypischen Viewer angesehen werden.

6. Ausblick

In diesem Artikel haben wir das von uns entwickelte Academic Meta Tool (AMT) vorgestellt, mit dem man Vagheit in Graphen modellieren kann. Es bietet die Möglichkeit, eine Ontologie mit vagen Inferenzregeln zu erstellen. Diese Regeln sind auf die in Kapitel 4 beschriebenen limitiert – wir planen jedoch weitere Regeln in neueren Versionen von AMT hinzuzufügen. Allerdings können die AMT-Regeln nicht bis zur Ausdrucksstärke von OWL erweitert werden, da hier Beschränkungen in der Berechenbarkeit vorliegen. Dennoch wird insbesondere die Kettenregel, eine speziell bei Graphen sehr wichtige Regel, von vielen OWL-Reasonern nicht unterstützt – weder mit noch ohne Berücksichtigung von Vagheit.

Durch die Verwendung von Web-Standards wie RDF und OWL ist eine unmittelbare Anbindung an andere Linked Open Data (LOD) problemlos möglich. Somit können die mit AMT erstellten Informationen mit anderen Ressourcen verknüpft werden und zur Anreicherung des Giant Global Graphs beitragen. Leider ist die Modellierung von Vagheit im Semantic Web noch nicht vom W3C standardisiert, daher nutzen wir zur Repräsentation von Vagheit eine Eigenentwicklung, da das W3C bisher nicht geplant hat Vagheit in das RDF-Format aufzunehmen.

Wie in Kapitel 3 bereits erläutert, möchten wir noch einmal darauf hinweisen, dass das Academic Meta Tool lediglich Vagheit unterstützt und keine Unsicherheit. Die Software ist also zur Modellierung solcher geisteswissenschaftlichen Fragestellungen geeignet, in denen viel Wissen vorhanden ist, aber eine klassische Modellierung (ohne Vagheit) an zu feinteiliger Kategorisierung scheitert – Beispiele hierfür haben wir in Kapitel 5 behandelt. Durch die Verwendung von AMT ist die Datenmodellierung nicht auf eine binäre Entscheidung (ja oder nein) beschränkt und es ist möglich, eine Information so abspeichern, dass sie nur zu einem gewissen Grad zutrifft.

Mehr Informationen sind auch über die GitHub Repositorien des mainzed verfügbar.

7. Danksagung

Wir möchten an dieser Stelle bei allen bedanken, die uns bei der Erstellung der Idee und der Use-Cases unterstützt haben. Für den Bereich der Informatik geht der Dank an Christophe Cruz (Université Bourgogne Franche-Comté) und für den Bereich der Digitalen Geisteswissenschaften an Prof. Dr. Kai-Christian Bruhn (Hochschule Mainz). Wir bedanken uns ebenfalls bei Katharina Kiefer (Studentin des Studiengangs Digitale Methodik in den Geistes- und Kulturwissenschaften an der Universität Mainz und der Hochschule Mainz) für das Korrekturlesen dieses Artikels.

Bibliographische Angaben

Elena Bellodi / Evelina Lamma / Fabrizio Riguzzi / Riccardo Zese / Giuseppe Cota: A web system for reasoning with probabilistic OWL. In: Software: Practice and Experience 47 (2017), H. 1, S. 125–142. [Nachweis im GBV]

Fernando Bobillo / Umberto Straccia: fuzzyDL: An expressive fuzzy description logic reasoner. In: IEEE International Conference on Fuzzy Systems. 5 Bde. (FUZZ-IEEE 2008, Hong Kong, 01.-06.06.2008) Piscataway, NJ 2008. Bd. 2, S. 923–930. [Nachweis im GBV]

Fernando Bobillo / Miguel Delgado / Juan Gómez-Romero: Reasoning in fuzzy OWL 2 with DeLorean. In: Uncertainty reasoning for the semantic Web II. Hg. von Fernando Bobillo / Paulo C. G. Costa / Claudia d'Amato / Nicola Fanizzi / Kathryn B. Laskey / Kenneth J. Laskey / Thomas Lukasiewicz / Matthias Nickles / Michael Pool. Berlin u.a. 2013, S. 119–138. [Nachweis im GBV]

Rajab Ali Borzooei / Hossein Rashmanlou: New concepts of vague graphs. In: International Journal of Machine Learning and Cybernetics 8 (2017), H. 4, S. 1081–1092. [Nachweis im GBV]

Kai-Chrstian Bruhn / Thomas Engel / Tobias Kohr / Detlef Gronenborn: Integrating Complex Archaeological Datasets from the Neolithic in a Web-Based GIS. In: CAA2014. 21st Century Archaeology. Concepts, methods and tools. Proceedings of the 42nd Annual Conference on Computer Applications and Quantitative Methods in Archaeology. Hg. von François Giligny / François Djindjian / Laurent Costa / Paola Moscati / Sandrine Robert. (CAA: 42, Paris, 22.–25.04.2014) Oxford 2015, S. 341–348. [Nachweis im GBV]

Arnaud Castelltort / Anne Laurent: Fuzzy queries over NoSQL graph databases: perspectives for extending the cypher language. In: International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems. Hg. von Anne Laurent / Oliver Strauss / Bernadette Bouchon-Meunier / Ronald R. Yager. 3 Bde. (IPMU: 15, Montpellier, 15.–19.07.2014) Cham 2014. Bd. 3, S. 384–395. [Nachweis im GBV]

Didier Dubois / Henri Prade: Possibility theory, probability theory and multiple-valued logics: A clarification. In: Annals of mathematics and Artificial Intelligence 32 (2001), H. 1–4, S. 35–66. [Nachweis im GBV]

Bettina Hoffmann: Die Rolle handwerklicher Verfahren bei der Formgebung reliefverzierter Terra Sigillata. München 1983. [Nachweis im GBV]

Thomas Lukasiewicz / Umberto Straccia: Managing uncertainty and vagueness in description logics for the semantic web. In: Web Semantics: Science, Services and Agents on the World Wide Web 6 (2008), H. 4, S. 291–308. [Nachweis im GBV]

Michael Piotrowski / Giovanni Colavizza / Florian Thiery / Kai-Christian Bruhn: The Labeling System: A new approach to overcome the vocabulary bottleneck. In: DH-CASE II: Collaborative Annotations on Shared Environments: Metadata, Tools and Techniques in the Digital Humanities. Hg von Patrick Schmitz / Laurie Pearce / Quinn Dombrowski. (DH-CASE: 2, Fort Collins, CO, 16.09.2014) New York, NY 2014. [Nachweis im GBV]

Giorgos Stoilos / Giorgos B. Stamou / Vassilis Tzouvaras / Jeff Z. Pan / Ian Horrocks: Fuzzy OWL: Uncertainty and the Semantic Web. In: Proceedings of the OWLED*05 Worshop on OWL: Experiences and Directions. Hg. von Bernardo Cuenca Grau / Ian Horrocks / Bijan Parsia / Peter Patel-Schneider. (OWLED: 5, Galway, 11.–12.11.2005) Aachen 2005. (= CEUR workshop proceedings, 188) [online]

Florian Thiery / Thomas Engel: The Labeling System: The Labelling System: A Bottom-up Approach for Enriched Vocabularies in the Humanities. In: CAA2015. Keep the Revolution Going. Proceedings of the 43rd Annual Conference on Computer Applications and Quantitative Methods in Archaeology. Hg. von Stefano Campana / Roberto Scopigno / Gabriella Carpentiero / Marianna Cirillo. (CAA: 43, Siena, 30.03.–03.04.2015) 2. Bde. Oxford 2016. Bd. 1, S. 259–268. [Nachweis im GBV]

Karsten Tolle / David Wigg-Wolf: Uncertainty Handling for Ancient Coinage. In: CAA2014. 21st Century Archaeology. Concepts, methods and tools. Proceedings of the 42nd Annual Conference on Computer Applications and Quantitative Methods in Archaeology. Hg. von François Giligny / François Djindjian / Laurent Costa / Paola Moscati / Sandrine Robert. (CAA: 42, Paris, 22.–25.04.2014) Oxford 2015, S. 171–178. [Nachweis im GBV]

Dorothea Tsatsou / Stamatia Dasiopoulou / Ioannis Kompatsiaris / Vasileios Mezaris: LiFR: a lightweight fuzzy DL reasoner. In: The semantic web: ESWC 2014 satellite events. Hg. von Valentina Presutti / Eva Blomqvist / Raphael Troncy / Harald Sack / Ioannis Papadakis / Anna Tordai. (ESWC: 11, Anissaras, 25.–29.05.2014) Cham u.a. 2014, S. 263–267. (= Lecture notes in computer science, 8798) [Nachweis im GBV]

Online-Quellen

Muhammad Akram / Feng Feng / Shahzad Sarwar / Youne Bae Jun: Certain types of vague graphs. In: Scientific Bulletin / Series A / University Politehnica of Bucharest 76 (2014), H. 1, S. 141–154. PDF. [online] [Nachweis im GBV]

Anne Klammt: FAQ Flyer des Mainzer Zentrums für Digitalität in den Geistes- und Kulturwissenschaften (mainzed) in deutscher Sprache. In: zenodo.org. Version 4 vom 18.09.2018. DOI: 10.5281/zenodo.1324187

Kenneth J. Laskey / Kathryn B. Laskey / Paulo C. G. Costa / Mieczyslaw M. Kokar / Trevor Martin / Thomas Lukasiewicz: Uncertainty Reasoning for the World Wide Web. In: w3.org. W3C Incubator Group Report vom 31.03.2008. [online]

Tobias Kohr (2014a): Component Certainty Type Authority Document. In: github.com/i3mainz. Arches Import Best Practices, Revision f001e53. 06.05.2014. [online]

Tobias Kohr (2014b): Phase Type Assignment Certainty Type Authority Document. In: github.com/i3mainz. Arches Import Best Practices, Revision f001e53. 06.05.2014. [online]

Alistair Miles / Sean Bechhofer: Simple Knowledge Organization System Reference. In: w3.org. W3C Recommendation vom 18.08.2009. [online]

John Muccigrosso: Joining Pelagios. In: github.com/pelagios. Pelagios Cookbook, Revision 2ee585f. 22.06.2018. [online]

Robert Sanderson / J. Paul Getty Trust / Paolo Ciccarese / Benjamin Young: Introduction. Web Annotation Data Model. In: w3.org. W3C Recommendation vom 23.02.2017. [online]

TEI Guidelines. Certainty, Precision, and Responsibility. Hg. von Text Encoding Initiative Consortium. Version 3.4.0. 23.07.2018. [online]

Florian Thiery / Allard Mees (2018a): Academic Meta Tool - samian ontology. In: zenodo.org. Version 1 vom 19.01.2018. DOI: 10.5281/zenodo.1341109

Florian Thiery / Allard Mees (2018b): Academic Meta Tool - navis ontology. In: zenodo.org. Version 1 vom 22.03.2018. DOI: 10.5281/zenodo.1341111

Florian Thiery / Allard Mees (2018c): Putting Samian pots together – modelling ceramic service family roots – connecting figure types. Wie Graphen bei der Modellierung des Zweifels helfen können. (Graphentechnologien 2018, Mainz, 19.–20.01.2018) In: zenodo.org. Präsentation vom 19.01.2018. DOI: 10.5281/zenodo.1155747

Florian Thiery / Allard Mees (2018d): Taming Ambiguity - Dealing with doubts in archaeological datasets using LOD. (Computer Applications and Quantitative Methods in Archaeology, Tübingen, 30.04.–03.04.2015) In: zenodo.org. Präsentation vom 22.03.2018. DOI: https://doi.org/10.5281/zenodo.1200111

Martin Unold / Florian Thiery (2018a): Academic Meta Tool – amt.js. In: zenodo.org. Version 1.0 vom 19.01.2018. DOI: 10.5281/zenodo.1342311

Martin Unold / Florian Thiery (2018b): Academic Meta Tool - mainzed ontology. In: zenodo.org. Version 1 vom 19.01.2018. DOI: 10.5281/zenodo.1341107

Martin Unold / Florian Thiery (2018c): Academic Meta Tool – Ontology. Leonard Edition vom 19.01.2018. [online]

Martin Unold / Florian Thiery (2018d): Academic Meta Tool – Vocabulary. Penny Edition vom 19.01.2018. [online]

Martin Unold / Florian Thiery (2018e): Academic Meta Tool – Vocabulary. In: zenodo.org. Penny Edition vom 19.01.2018. DOI: 10.5281/zenodo.1342530

Martin Unold / Florian Thiery (2018f): Academic Meta Tool – Ontology. In: zenodo.org. Leonard Edition vom 19.01.2018. DOI: 10.5281/zenodo.1342536

Martin Unold / Florian Thiery (2018g): Academic Meta Tool – Ein Web-Tool zur Modellierung des Zweifels. (Graphentechnologien 2018, Mainz, 19.–20.01.2018) In: zenodo.org. Präsentation vom 19.01.2018. DOI: 10.5281/zenodo.1155726

Abbildungslegenden und -nachweise

Abb. 1: Verknüpfung zweier vager Aussagen durch Konjunktion und Disjunktion. [Eigene Darstellung, CC BY 4.0].

Abb. 2: Das Konzept Place und die Rolen northOf, eastOf, southOf, westOf. [Eigene Darstellung, CC BY 4.0].

Abb. 3: Schematische Darstellung des Role-Chain-Axioms (Rollen-Kettenregel). [Eigene Darstellung, CC BY 4.0].

Abb. 4: Schematische Darstellung des Inverse-Axioms. [Eigene Darstellung, CC BY 4.0].

Abb. 5: Konzepte und Rollen der mainzed ontology. [Eigene Darstellung, Figuren und Geräte: © mainzed, CC BY 4.0].

Abb. 6: Rollen-Kettenregeln der mainzed ontology. [Eigene Darstellung, Figuren und Geräte: © mainzed, CC BY 4.0].

Abb. 7: Schlussfolgerungen der mainzed ontology im Web-Viewer. [Eigene Darstellung, CC BY 4.0].

Abb. 8: Darstellung von Bogenschützen-Punzen. [Eigene Darstellung, © RGZM / Mees, CC BY 4.0].

Abb. 9: Konzepte und Rollen der samian ontology. [Eigene Darstellung, Figuren und Geräte: ©mainzed, CC BY 4.0].

Abb. 10: Rollen-Kettenregeln der samian ontology. [Eigene Darstellung, Figuren und Geräte: © mainzed, CC BY 4.0].

Abb. 11: Schlussfolgerungen der samian ontology im Web-Viewer. [Eigene Darstellung, CC BY 4.0].

Abb. 12: Konzepte und Rollen der navis ontology. [Eigene Darstellung, CC BY 4.0].

Abb. 13: Rollen-Kettenregeln der navis ontology. [Eigene Darstellung, CC BY 4.0].

Abb. 14: links: Darstellung einer römischen Münze mit Schiffsdarstellung (O41650 aus NAVIS III) und eines Monuments (NeumagenMonument1 aus NAVIS II). [Eigene Darstellung, © RGZM, CC BY 4.0].

ZfdG

Zeitschrift für digitale Geisteswissenschaften

Beitrag aus:
Sonderband 4 der ZfdG: Die Modellierung des Zweifels – Schlüsselideen und -konzepte zur graphbasierten Modellierung von Unsicherheiten. Hg. von Andreas Kuczera, Thorsten Wübbena und Thomas Kollatz. 2019. DOI: 10.17175/sb004

Titel:
Quellenverluste (Deperdita) als methodologischer Unsicherheitsbereich für Editorik und Datenmodellierung am Beispiel von Anton Weberns George-Lied op. 4 Nr. 5

Autor/in:
Stefan Münnich

Kontakt:
stefan.muennich@unibas.ch

Institution:
Universität Basel, Departement Künste, Medien, Philosophie

GND:
1068032472

ORCID:
0000-0002-0744-5374

DOI des Artikels:
10.17175/sb004_005

Nachweis im OPAC der Herzog August Bibliothek:
1031300139

Erstveröffentlichung:
03.07.2019

Lizenz:
Sofern nicht anders angegeben

Medienlizenzen:
Medienrechte liegen bei den Autoren

Letzte Überprüfung aller Verweise:
03.06.2019

GND-Verschlagwortung:
Quellenforschung | Konzeptionelle Modellierung | Musikphilologie | Semantic Web | Wissensrepräsentation |

Zitierweise:
Stefan Münnich: Quellenverluste (Deperdita) als methodologischer Unsicherheitsbereich für Editorik und Datenmodellierung am Beispiel von Anton Weberns George-Lied op. 4 Nr. 5. In: Die Modellierung des Zweifels – Schlüsselideen und -konzepte zur graphbasierten Modellierung von Unsicherheiten. Hg. von Andreas Kuczera, Thorsten Wübbena und Thomas Kollatz. 2019 (= Sonderband der Zeitschrift für digitale Geisteswissenschaften, 4). PDF Format ohne Paginierung. Als text/html abrufbar unter DOI: 10.17175/sb004_005.

Stefan Münnich

Quellenverluste (Deperdita) als methodologischer Unsicherheitsbereich für Editorik und Datenmodellierung am Beispiel von Anton Weberns George-Lied op. 4 Nr. 5

Abstracts

Aus drei Perspektiven nähert sich der Beitrag der Problematik der Quellenverluste (*Deperdita*) an: In einem ersten Abschnitt werden diese als methodologischer Unsicherheitsbereich für die historische Forschung identifiziert und mögliche Kategorisierungen vorgeschlagen. Das darauffolgende Fallbeispiel aus der editorischen Praxis der Anton Webern Gesamtausgabe beleuchtet die Folgen solcher Fehlstellen für das kulturelle Gedächtnis, bevor im letzten Teil Vorschläge zur graphbasierten Modellierung von Negationen und Deperdita vorgelegt werden. Ziel der Diskussion ist die Schärfung des Problembewusstseins für die Herausforderungen im Umgang mit nicht (mehr) vorhandenem bzw. erschlossenem historischen Material sowie für einen konstruktiven und produktiven Zugang zu Unsicherheiten.

The paper approaches the problem of the loss of sources (*deperdita*) from three different perspectives: In a first section, deperdita are identified as a methodological area of uncertainty for historical research and possible categorisations are proposed. The following case study from the work of the Anton Webern Gesamtausgabe illustrates the consequences of such defects in cultural memory caused by missing sources, before proposals for graph-based modeling of negations and deperdita are presented in the last part. The aim of the discussion is to raise awareness of the challenges involved in dealing with historical material that is no longer available and for a constructive and productive approach to uncertainties.

1. Einleitung

Quellenverluste erzeugen Bruchstellen im Überlieferungskontext einer Komposition oder eines Textes und erschweren, wenn nicht gar verunmöglichen somit eine Feststellung der Quellenabhängigkeiten. Um dabei überhaupt von einem Quellenverlust sprechen zu können, bedarf es zunächst des expliziten Erkennens eines Nicht(mehr)-Vorhandenen, das zuvor existiert hat. Eine solche Erkenntnis muss aus dem Zweifel an einem historiographischen Narrativ, aus Unstimmigkeiten in der Überlieferungssituation der verfügbaren Quellen selbst oder aus oft nur vagen Hinweisen in Kontextmaterialien wie Briefen, Tagebüchern und Ähnlichem gewonnen werden. Unsicherheiten und Zweifel stellen so mitunter die Grundbausteine für Hypothesenbildungen dar.

Worüber aber sprechen wir, wenn wir von Quellenverlusten reden? Welche Varianten von Zweifel an einem Überlieferungskontext, welche Modi von Unsicherheit über den Verbleib von Quellen lassen sich beobachten? Obwohl Quellenverluste gemeinhin als methodologisches Problem erkannt und benannt werden, mangelt es bislang an einer grundlegenden Reflexion und expliziten Beschreibung dieses philologischen Unsicherheitsbereichs.

Dies gilt umso mehr für die Beschreibung und Modellierung von nicht (mehr) vorhandenen Quellen im digitalen Kontext. Denn im Rahmen maschinensprachlicher Verarbeitung spielt Explizität, das Explizitmachen der beschriebenen Objekte und ihrer Eigenschaften, eine grundlegende Rolle. Zwar kann eine Maschine[1] die notwendigen Berechnungen im Zusammenhang mit Wahrscheinlichkeiten und Annahmen sehr schnell und genau ausführen, dazu müssen aber komplexe Aufgabenstellungen in berechenbare bzw. schaltbare Einzeloperationen zerlegt werden. Dies gilt auch für den Fall logischer Ableitungen (Inferenz), für welche entsprechende Ableitungsregeln expliziert werden müssen, oder selbst maschinelle Lernprozesse, bei denen Inputs, Verhaltensstrategien oder Belohnungskriterien für Lern(fort)schritte ausdrücklich formuliert werden. Was dahingehend nicht explizit modelliert ist, kann von der Maschine nicht verarbeitet werden, oder schärfer ausgedrückt: es existiert nicht.

Wie können nun aber nicht (mehr) vorhandene musikalische Quellen und Schriften so modelliert werden, dass sie im Zusammenhang digitaler Editionen, Kataloge oder Repertorien zugleich als *Quelle*, aber eben auch als *Verlust* greifbar und verarbeitbar sind? Wie kann ein digitales Objekt, das existiert, Instanz von etwas sein, dass nicht (mehr) existiert? Wie also kann dieser Unsicherheitsbereich mit seinen impliziten Unschärfen im Rahmen einer graphbasierten Modellierung explizit gemacht werden?

Am Beispiel verschollener Materialien zu Anton Weberns George-Lied op. 4 Nr. 5 soll in vorliegendem Beitrag aufgezeigt werden, welche Herausforderungen und Möglichkeiten bei einer solchen Modellierung grundsätzlich bestehen und wie sich eine an der Anton Webern Gesamtausgabe (Basel, Schweiz) entwickelte Ontologie für die damit zusammenhängenden editorischen Modelle im Speziellen dazu verhalten kann.

2. Unsicherheit & Verlust

Mit dem Thema »*Die Modellierung des Zweifels*«. *Schlüsselideen und -konzepte zur graphbasierten Modellierung von Unsicherheiten* haben die Organisatoren der Tagung Graphentechnologien 2018 in Mainz dankenswerterweise den Blick auf einen Themenkomplex gerichtet, der den Kern geisteswissenschaftlichen Nachdenkens berührt, sind hierbei doch durch Argumentation und Kontext verhandelte Unsicherheiten und Zweifel zentrale Kategorien. Unsicherheit ist mithin so elementar, dass sie als »eine zentrale Bedingung menschlichen Lebens, deren Auswirkungen alle Bereiche des Lebens durchziehen«,[2] beschrieben werden kann. Sei es, dass man die Ursachen für eine solche Zentralität von Unsicherheit für das menschliche Dasein im Anschluss an Heidegger, Kierkegaard oder schon Augustinus in einer existenzial-ontologischen, christlich-religiös verorteten Angst durch die Erkenntnis der eigenen Endlichkeit sehen möchte.[3] Sei es, dass es sich hierbei um

[1] Unter Maschine seien hier und im Folgenden universell programmierbare (im Sinne von Turing-Vollständigkeit) Computer verstanden.
[2] Wulf / Zirfas 2015, S. 9.
[3] Letztendlich ist das mit dem christlichen Sündenfall verbundene Narrativ einer Verbannung aus dem Paradies (»Urstand«) zu einem »Leben auf Erden« – also der Übergang aus der Sicherheit einer zeitlosen Unendlichkeit des Raumes in die Unsicherheit einer durch Zeit sanktionierten Endlichkeit – nicht viel

eine Art kultivierten tierischen Fluchtinstinkt handelt, der uns als gejagtem Jäger evolutionär noch in den Knochen steckt. Oder sei es, dass man hier im Anschluss an Freud eher einen psychisch-mentalen Entwicklungsprozess zu Grunde legen möchte, bei dem sich die Sicherheit eines angeborenen Urvertrauens zunehmend stärkeren Verlustängsten, Zweifeln und Unsicherheiten ausgesetzt sieht, deren Überwindung oder Verarbeitung letztendlich zur Reifung eines Individuums als emanzipiertes Subjekt beiträgt.[4] Unsicherheit, so scheint es, bedingt der Schablone einer ursprünglichen Sicherheit – hier eines paradiesischen Urstands oder kindlichen Urvertrauens. Eine solche Kausalität erkennt auch Ludwig Wittgenstein an, wenn er in seinen späten Notizen Über Gewißheit (1949–1959) im Zusammenhang mit Glaubenssätzen und deren Anzweiflung formuliert: »114. Wer keiner Tatsache gewiß ist, der kann auch des Sinnes seiner Worte nicht gewiß sein. 115. [...] Das Spiel des Zweifelns selbst setzt schon die Gewißheit voraus.«[5] Das Unsicherheitsspiel lässt sich nur auf einem sicheren Untergrund, einem stabilen Spielplan durchführen. Egal, ob es sich dabei um die ›Unsicherheit des Gestern‹ handelt, die als Folge einer zurückliegenden Verlusterfahrung hervorgerufen wird, oder um die ›Unsicherheit des Morgen‹, die sich der Unvorhersehbarkeit einer zukünftigen Entwicklung ausgesetzt sieht.[6] Für den Umgang mit Unsicherheiten sind Bewältigungs- und Planungsstrategien vonnöten – sei es durch ohnmächtiges Ertragen oder aktives Handeln – die mithin den Rahmen für die »Aufrechterhaltung und Erweiterung von Handlungsfähigkeit bei Unsicherheit«[7] bzw. trotz oder gerade wegen (?) Unsicherheit abstecken. Wenn nunmehr Unsicherheit durch die hier angeführten Fälle als eine unhintergehbare Bedingung menschlichen Daseins greifbar wird, so bilden Verluste – als ›Unsicherheiten des Gestern‹ – in dem hier in aller Kürze aufgespannten Verstehensraum eine mögliche zentrale Kategorie, die eine solche Unsicherheit motivieren.[8]

2.1 Unsicherheitsbereiche historischer Forschung: nicht (mehr) vorhandene Quellen

Im Umgang mit historischen Informationsträgern (z. B. Manuskripte, Drucke, Tonaufnahmen, Scherben) sehen sich Forscher, Editoren, Archivare, Bibliothekare permanent mit der Vergänglichkeit des Materials konfrontiert, während dieses konsultiert, katalogisiert, inventarisiert, digitalisiert oder aufgrund anderweitigen spezifischen Interesses untersucht

weniger als eine Metapher für eben diese uns nicht angeborene, sondern meist schmerzhaft ›erlernte‹ Vergänglichkeitserkenntnis. Zur philosophischen Auseinandersetzung mit dem Komplex ›Angst‹ und den teils kontradiktischen Bezugnahmen der genannten Autoren vgl. Muñoz Criollo 2013, insb. S. 137ff., sowie Geier 2017, S. 124ff.

[4] Zur Entwicklung der Angsttheorie bei Freud und nachfolgenden psychoanalytischen Positionen vgl. Meyer 2005, passim. Zu einer musik- und entwicklungstherapeutischen Perspektive vgl. Renz 2009, passim.

[5] Wittgenstein 1970, S. 39.

[6] Der Umgang mit erst noch stattfindenden Ereignissen und den daraus resultierenden Unsicherheiten ist besonders in marktökonomischen Prozessen und unternehmerischen Entscheidungen relevant. Zunehmend bilden sich dabei Konzepte heraus, die – frei nach dem Alan Curtis Kay nachgesagten Diktum »The best way to predict the future is to invent it« – Unwägbarkeiten aktiv einzubeziehen statt auszuschließen suchen, wie der von Saras D. Sarasvathy vorgeschlagene *Effectuation*-Ansatz (vgl. u. a. Sarasvathy 2001; Sarasvathy 2008). In der Kunst und speziell der Musikpraxis ist der Umgang mit Unsicherheiten, Unvorhergesehenem, Zufall spätestens seit den 1960er Jahren mit den »Experimenten« eines John Cage oder der Fluxus-Bewegung etabliert. Ansätze zu einem kreativen und ergebnisoffenen Einbeziehen von Ungewissheit in der Musik reichen aber mindestens bis in die Anfänge der frühen Neuzeit zurück, wie etwa musikalische Würfelspiele à la Mozart oder Rätselkanons zeigen; vgl. u. a. Schiltz 2015.

[7] Böhle 2013, S. 281.

[8] Zu weiteren Dimensionen des so mehrschichtigen wie komplexen Begriffs ›Unsicherheit‹ vgl. die übrigen Beiträge dieses Sonderbands sowie Jeschke et al. 2013, passim.

werden soll. Erst durch dieses Interesse, durch das Befragen des Materials, wird es zu einer Quelle. Eine Quelle zu sein, ist somit keine charakteristische Eigenschaft, die irgendein Material (oder eine Person) an sich hat oder besitzt, man könnte sogar sagen, Quellen existieren per se gar nicht:

> »Es *gibt* keine Quellen, Quellen *sind* nicht, Quelle zu sein ist keine einem Ding oder einer Person eignende Eigenschaft, sondern sie werden zu Quellen *gemacht* – nämlich von demjenigen, der sie als solche für die Konstruktion seiner jeweiligen Interpretation von Vergangenheit nutzt.«[9]

Durch die Möglichkeit, die in den überlieferten Materialien gefundenen Hinweise mit plausiblen Interpretationen oder Narrativen von Vergangenheit zu verknüpfen, tragen die in diesem Vorgang als solche deklarierten Quellen zu einer Wissensstruktur bei, die der Kulturanthropologe Jan Assmann als »kulturelles Gedächtnis« bezeichnet hat. Das heißt, sie tragen bei zu einem Repertoire von Narrativen, von Erklärungsmodellen, Glaubenssätzen und Bewältigungsstrategien, das unsere Sichtweisen und unser Verständnis von Welt und uns selbst in Bezug auf unsere Vergangenheit und Gegenwart prägt.[10] Aber jegliche historiographische Erzählung – und damit auch das kulturelle Gedächtnis – beruhen zu einem großen Teil eben nicht auf der Gewissheit harter Fakten, sondern sind aus unsicheren Indizien, Hinweisen oder Anhaltspunkten konstruiert. Auch wenn alle je existenten Materialien berücksichtigt werden könnten, ist noch nichts über einen möglichen Wahrheitsgehalt oder eine wie auch immer geartete »Objektivität« einer aus Zusammenspiel und Kontextualisierung dieser Materialien hervorgegangenen Interpretation ausgesagt. Zugleich muss angenommen werden, dass nur ein Bruchteil des ehemals vorhandenen historischen Materials überhaupt überliefert ist. Es lässt sich so nicht einmal auf der Materialebene der relative Umfang der überlieferten Teilmenge mit letztgültiger Sicherheit einschätzen, da nicht klar ist, zu welcher Gesamtgröße diese Teilmenge in Beziehung zu setzen ist.

Terminologisch in seiner Totalität als *Quellenausfall* und in Anwendung auf das spezifische Fehlen eines individuellen Informationsträgers als *Quellenverlust* adressierbar, erzeugt dieses Phänomen massive Unsicherheiten, denn es unterbricht den Überlieferungskontext der historischen Objekte – sei es eine Komposition, ein Text, ein Gemälde oder ein Bauwerk – und verkompliziert dessen adäquate Rekonstruktion bzw. Beschreibung. Folgerichtig wurde es von dem Musikwissenschaftler Georg von Dadelsen, der an verschiedenen Editionsprojekten, u. a.der Neuen Bach Ausgabe, maßgeblichen Anteil hatte, als ein generelles Problem aller historischen Disziplinen bezeichnet, und als der Ort, an dem Hypothesen – als gedanklich und sprachlich modellierte Unsicherheitsaussagen – notwendig seien, um das »quellenlose Vakuum«[11] zu füllen. Derartige Leerstellen lassen sich als Gedächtnislücke im kulturellen Gedächtnis beschreiben, oder noch ärger als Gedächtnisverlust; der Verlust des Materials bedingt einen Verlust im kulturellen Gedächtnis. Es braucht zum einen Anstrengungen, um diesem Gedächtnisverlust entgegenzuwirken, so wie es bewusster Bemühungen bedarf, historisches Material zu bewahren und zu erhalten. Zum anderen bedarf es eines

[9] Kümper 2014, S. 18. Vgl. auch Kirn 1968, S. 30: »Quellen nennen wir alle Texte, Gegenstände oder Tatsachen, aus denen Kenntnis der Vergangenheit gewonnen werden kann.«
[10] Vgl. Assmann 1988, passim; und Assmann 2013, passim.
[11] Von Dadelsen 1988, S. 127.

konstruktiven, offensiven Umgangs mit diesem verlustbedingten Unsicherheitsbereich, um auch hier eine »Handlungsfähigkeit mit Ungewissheit«[12] – und Handlungsfähigkeit im doppelten Sinne sowohl des Aktivwerdenkönnens als auch der narrativen Darstellungsmöglichkeit – aufrechtzuerhalten bzw. herzustellen.

Wie aber können wir wissen, dass etwas als Quelle befragt werden könnte, wenn es verloren ist? Können wir bei nicht (mehr) vorhandenem Material überhaupt von »Quellen«, und somit auch »Quellenverlust« oder »Quellenausfall«, sprechen? Dies ließe sich insofern bejahen, als dass zweifel- oder dokumentbasierte Hinweise aus anderen Quellen auf die (vormalige) Existenz nicht (mehr) vorhandenen Materials schließen lassen können und durch die explizite Benennung und Einbindung in ein plausibles Narrativ als konstruierter »Interpretation von Vergangenheit« sich dieses neu manifestieren kann. In seiner Dissertation über Die Kriegsverluste der Musiksammlungen deutscher Bibliotheken 1942–1945 hat Nicola Schneider dafür den Begriff der *Deperdita* aus der Mediävistik übernommen, der dort für nicht überlieferte Dokumente verwendet wird, deren Inhalt aber aus anderen Quellen oder Kontexten rekonstruierbar ist.[13] Eine Anwendung des Begriffs erscheint mir auch im vorliegenden Kontext durchaus hilfreich und sinnvoll.

2.2 Mögliche Kategorisierungen von Deperdita

Bislang scheint sich keine Studie oder Arbeit explizit mit einem systematischen Modell von Deperdita befasst zu haben, obwohl diese in der Philologie, der Provenienzforschung und von Gedächtnisinstitutionen, vornehmlich Bereichen und Einrichtungen, die permanent mit nicht (mehr) vorhandenen Quellen konfrontiert sind, weithin als methodologisches Problem identifiziert wurden.[14] In der einschlägigen Fachliteratur findet sich nicht viel mehr als die sehr verbreitete (und offensichtliche) Unterscheidung, dass Quellen (oder Kulturgüter in der Provenienzforschung) »entweder verschollen (unwiederbringlich verloren) oder in Privatbesitz«[15] sind. In Bezug auf den Begriff »Kriegsverlust« hat Nicola Schneider darauf hingewiesen, dass dieser ein Container für mindestens drei verschiedene Verlustarten sei: Er umfasse »vernichtete« (endgültig zerstörte), »verschollene« (aktueller Standort und Erhaltungszustand nicht bekannt) und »dem ursprünglichen Besitzer entzogene« (enteignete, meist im Ausland lagernde) Quellen.[16]

Bei dem Versuch, die vorgenannten Kategorien von Deperdita in einer Übersichtstabelle anzuordnen, zeigt sich schnell, dass noch weitere Optionen zu berücksichtigen sind (vgl. Abbildung 1). Neben den erwähnten gängigen Optionen (in der Tabelle mit einem * gekennzeichnet) gibt es noch mindestens fünf weitere, insbesondere im Bereich der physisch nicht existierenden (»non-existent«) Quellen. Diese stellen methodisch eine besondere

[12] Böhle 2013, S. 290.
[13] Schneider 2013, S. 8. Sebastian Wedler gilt der Dank, mich freundlicherweise auf diese Arbeit aufmerksam gemacht zu haben. Grundsätzlich zu Bedeutung und Nutzen von Deperdita in der Mediävistik vgl. Hartmann 2014, passim.
[14] Vgl. u. a.Schwindt 2016, passim.
[15] Scheideler 2017, S. 180.
[16] Schneider 2013, S. 8.

Herausforderung dar, markieren sie doch keine materiellen Verluste im eigentlichen Sinn. Dennoch müssen sie in diese Betrachtung einbezogen werden, können sie doch aus einem bestimmten Anlass innerhalb eines gewissen Zeithorizonts als eine mögliche reale Quelle gehandelt werden und sich so in ein interpretatorisches Narrativ einschreiben. Obwohl ein entsprechender Inhalt niemals rekonstruiert oder erschlossen werden kann, wird der Überlieferungskontext gleichsam gestört, so dass sich dabei durchaus von »sekundären« Deperdita sprechen lässt. Hierunter fallen beabsichtigte, aber nie begonnene Werke (»unrealized, intended«), Quellen, die von Personen oder in dokumentarischen Hinweisen falsch deklariert wurden (»falsely declared«) oder deren Existenz von den bearbeitenden Forschern fälschlicherweise angenommen wurde (»falsely assumed«). Zwei sehr spezielle Fälle sind 1.) Quellen, die einem falschen Autor zugeschrieben wurden (»falsely attributed«; hier sorgt die falsche Zuordnung für die Nicht-Existenz der Quelle, in Bezug auf den eigentlichen Autor aber handelt es sich wiederum um eine reale, physisch existierende Quelle), und 2.) Quellen, die durch keinerlei dokumentarisch bezeugte Evidenz, sondern nur aufgrund einer auf professionelles Fachwissen gestützten Intuition oder Annahme als wahrscheinlich gelten (»hypothetical, rumored«).[17]

Physicality	Access	Status	Evidence	Location
real	none	access blocked*	Owner Documents (= DOC)	known
		irretrievably lost* / extinct*	DOC	none (last known provenance)
	none [pot. possible]	falsely attributed	DOC editor	none [known]
	potentially possible	private hands*	DOC	(un)known (last known provenance)
		missing*		
real / non-existent		hypothetical rumored	intuition rumor NODOC	unknown
non-existent	none	unrealized intended (but not started)	DOC plans intentions	none
		falsely declared	DOC author source	
		falsely assumed	DOC editor	

Abb. 1: Zusammenstellung verschiedener Kategorien von Deperdita, unterschieden nach physischer Existenz, Zugangsmöglichkeit, Status, Evidenz und Aufbewahrungsort (DOC = dokumentarisch bezeugte Evidenz). [Eigene Grafik 2018. CC-BY 4.0.]

Nach diesen methodologischen Vorüberlegungen soll im Folgenden der Blick auf ein Fallbeispiel aus der editorischen Praxis der Anton Webern Gesamtausgabe gerichtet werden.

[17] Ein Beispiel für diesen Spezialfall findet sich in der *Neuen Bach Ausgabe* III/2.2,2 unter den Quellen zu den Bach-Chorälen (Rempp 1996, S. 85): Da fast 200 Choräle nur aus Sekundärquellen bekannt sind und sich für diese auch keinerlei Vorlagen in Kantaten oder Passionen finden lassen, nimmt der Herausgeber Frieder Rempp (und mittlerweile auch ein Großteil der Bach-Forschung) hier eine unbekannte Quelle [Y] an, die eine persönliche Choralsammlung Johann Sebastian Bachs gewesen sein muss. Mein Dank gilt Christoph Wolff und Luke Dahn für Hinweise und weiterführende Diskussionen zu diesem Fall.

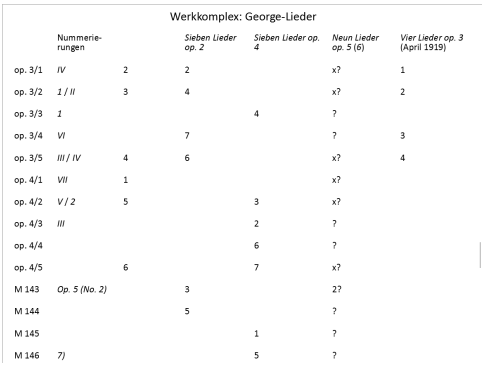
3. Webern, George und die verlorene Quelle

Würde man in einem Raum voller Musikwissenschaftler die Aufgabe stellen, Komponisten zu benennen, bei denen eine komplexe und schwierige Quellenlage vorliegt, wäre vermutlich Anton Webern (1883–1945) nicht darunter. Es gibt andere Komponisten wie Johann Sebastian Bach (1685–1750), Joseph Haydn (1732–1809) oder Edgard Varèse (1883–1965), für die die Situation viel prekärer ist. Doch so wie Weberns Musik zuweilen als Kleinod an Subtilität und Verdichtung beschrieben wird, so scheint dies auch für das Quellenmaterial seiner Kompositionen zu gelten. Es birgt eine gewisse Subtilität, welche die Komplexität und die Schwierigkeiten in einzelne Brennpunkte verdichtet. Die Situation ist mitunter weniger offensichtlich als bei anderen Komponisten, aber keineswegs weniger diffizil.[18]

3.1 Anton Weberns George-Lieder

Zunächst aber einige Bemerkungen zum Werkkomplex der George-Lieder[19] op. 3 und 4: Vermutlich zwischen 1908 und 1909 komponierte Webern insgesamt 14 Lieder für Singstimme und Klavier nach Gedichten von Stefan George (1868–1933). Der Impuls zur Beschäftigung mit der Lyrik Georges und zu deren Vertonung geht allen bekannten Hinweisen nach auf Weberns Lehrer, Mentor und Freund Arnold Schönberg (1874–1951) zurück, der zwischen Dezember 1907 und Mai 1908 selbst mehrere George-Vertonungen (darunter das Lied Ich darf nicht dankend op. 14/1 sowie sechs Nummern aus Fünfzehn Gedichte aus Das Buch der hängenden Gärten von Stefan George op. 15) fertiggestellt hatte.[20] Musikhistoriographisch markieren die Lieder den Übergang von den – noch tonalen Prinzipien verpflichteten – frühen Kompositionen zur sogenannten Atonalität in Weberns Schaffen (abgesehen von der wohl frühesten George-Vertonung Erwachen aus dem tiefsten Traumesschoße verzichtet er hier u. a. konsequent auf sämtliche Tonartvorzeichnungen).[21] Wie sehr Webern dabei offenbar nicht nur an einer radikaleren Tonsprache, sondern auch an einer großformalen Anordnung und Zusammenstellung der Einzellieder gelegen war, tritt in zahlreichen Streichungen und Änderungen zumeist auf den Manuskripttitelseiten zutage, die Weberns verschiedene Überlegungen und Planungen erahnen lassen:

[18] Die folgenden Ausführungen beruhen auf den Arbeiten von Thomas Ahrend, Mitglied der Editionsleitung der Anton Webern Gesamtausgabe und Herausgeber des ersten Bandes der Gesamtausgabe, welcher die Klavierlieder Weberns, darunter auch die Gruppe der George-Lieder op. 3 und 4, enthalten wird. Für die Bereitstellung von Materialien sowie anregende Diskussionen sei hiermit herzlich gedankt. Weiterführende Informationen zur Anton Webern Gesamtausgabe, die als historisch-kritische Edition das »gesamte kompositorische Schaffen des österreichischen Komponisten Anton Weberns und Mitglied des engsten Kreises der sog. Zweiten Wiener Schule der Öffentlichkeit in wissenschaftlich angemessener und der musikalischen Praxis dienender Form zugänglich machen will« auf der Projektwebseite.
[19] Ausführlich zu unterschiedlichen editorischen, textlichen oder werkgenetischen Aspekten von Weberns George-Vertonungen vgl. Ahrend 2011, passim; Ahrend 2016, passim; Budde 1971, passim; Brinkmann 1973, passim.
[20] Vgl. dazu Ahrend 2011, S. 66ff. und S. 72f.
[21] Vgl. Ahrend 2011, S. 53; Budde 1971, S. 11.



Werkkomplex: George-Lieder							
	Nummerierungen		*Sieben Lieder op. 2*	*Sieben Lieder op. 4*	*Neun Lieder op. 5* (6)	*Vier Lieder op. 3* (April 1919)
op. 3/1	*IV*	2	2		x?	1
op. 3/2	*1 / II*	3	4		x?	2
op. 3/3	*1*			4	?	
op. 3/4	*VI*		7		?	3
op. 3/5	*III / IV*	4	6		x?	4
op. 4/1	*VII*	1			x?	
op. 4/2	*V / 2*	5		3	x?	
op. 4/3	*III*			2	?	
op. 4/4				6	?	
op. 4/5		6		7	x?	
M 143	*Op. 5 (No. 2)*		3		2?	
M 144			5		?	
M 145				1	?	
M 146	*7)*			5	?	

Abb. 2: Übersicht über nachweisbare Versuche Weberns, die George-Lieder als Sammlungen zusammenzufassen (Sp. 4–7). Sp. 1: Angabe der heutigen Opus- bzw. Moldenhauernummer (M) eines Liedes. Sp. 2: Dokumentarisch nachweisbare Nummerierungen. Sp. 3: An der Uraufführung 1910 beteiligte Stücke. [Anton Webern Gesamtausgabe 2018. CC-BY-NC-SA 4.0.]

In Abbildung 2 sind exemplarisch diese Nummerierungen (Spalte 2) und die verschiedenen Zusammenstellungen von zunächst zweimal sieben Liedern (als op. 2 und 4 vorgesehen), dann neun (als op. 5 bzw. 6) bzw. vier Liedern (als op. 3) vermerkt. Hinzu kommen eine bereits am 8. Februar 1910 erfolgte Uraufführung von sechs dieser Lieder (vgl. Abbildung 2, Spalte 3) in einem Konzert des *Vereins für Kunst und Kultur* in Wien mit der Sängerin Martha Winternitz-Dorda (1880–1958) sowie Weberns Pläne, 1911 eine Auswahl von neun bzw. zehn Liedern bei den Verlagen Dreililien in Berlin bzw. Tischer & Jagenberg in Köln publizieren zu lassen, und – nachdem sich diese zerschlugen – ein geplanter Privatdruck von wiederum neun Liedern (als op. 2) im Jahr 1912 (in Abbildung 2 nicht vermerkt). Zu einer endgültigen Drucklegung kam es aber erst 1919 (Verein für musikalische Privataufführungen, Wien; 1921 Copyright von Universal Edition übernommen) bzw. 1923 (Universal Edition, Wien) als Opus 3 und 4 mit jeweils fünf Liedern.[22]

Tatsächlich wurde allerdings ein einziges der George-Lieder Weberns bereits über zehn Jahre früher publiziert: »Ihr tratet zu dem Herde«, nach einem Text aus Georges Gedichtband Das Jahr der Seele [23] wurde 1912 neben Stücken von Arnold Schönberg und Alban Berg (1885–1935) in den Almanach Der Blaue Reiter [24] von Wassily Kandinsky (1866–1944) und Franz Marc (1880–1916) aufgenommen (vgl. Abbildung 3).

[22] Zur Problematik der chronologischen Reihenfolge und Ordnungsprinzipien vgl. Ahrend 2011, S. 60–66. Die Frage nach einer Modellierung der in diesem Komplex enthaltenen Abhängigkeiten, Bezugnahmen und Verweise wird die Anton Webern Gesamtausgabe noch intensiv beschäftigen.
[23] Das von Webern verwendete Exemplar der 1904 im Verlag J. Bondi in Berlin publizierten dritten Auflage des Gedichtbands (Erstveröffentlichung 1897) befindet sich heute in Basel, Paul Sacher Stiftung, Sammlung Anton Webern.
[24] Kandinsky / Marc 1912.

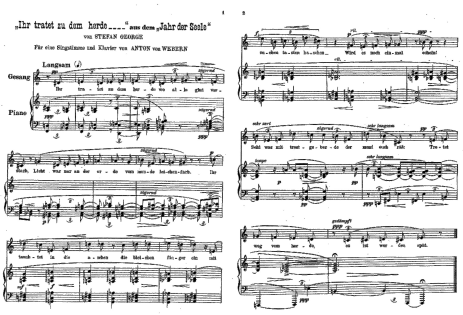


Abb. 3: Abdruck von Anton Weberns George-Vertonung »Ihr tratet zu dem Herde« als Beilage in: Der Blaue Reiter. Kandinsky / Marc 1976 (1912). [Public Domain Marked.]

Im Vergleich dazu die als Opus 4 Nr. 5 im Jahr 1923 gedruckte Version von »Ihr tratet zu dem Herde« als Hörbeispiel[25] (Audio 1):

Audio 1: Anton Webern, »Ihr tratet zu dem Herde« op. 4 Nr. 5 (Druckfassung 1923). Audio-Aufnahme anlässlich einer Aufführung der Fünf Lieder nach Gedichten von Stefan George op. 4 mit Kate Maroney (Mezzosopran) und Daniel Schlosberg (Klavier) im Spectrum, New York City (NY), Mai 2016 [Kate Maroney und Daniel Schlosberg 2016. Mit freundlicher Genehmigung von Kate Maroney und Daniel Schlosberg.] [online].

Dieses Lied bzw. dessen früher Druck sollen im Folgenden als Fallbeispiel dafür dienen, wie aus einer scheinbar unproblematischen Quellenlage sich ein »Quellenvakuum« (vgl. den ersten Abschnitt dieses Beitrags) bilden kann.

3.2 Frühfassung(en) gesucht – Deperdita von »Ihr tratet zu dem Herde« (op. 4 Nr. 5)

Bei einem ersten Blick auf die vorläufige Quellenübersicht zu Weberns Opus 4 (vgl. Abbildung 4) könnte die Situation nicht besser sein. Es sind fast alle Autographe überliefert, es gibt verschiedene Fassungen in verschiedenen Korrekturstufen und Abschriften sowie gedruckte Partituren.

[25] Eine Einspielung der im *Blauen Reiter* publizierten Fassung liegt bislang nicht vor.

Fünf Lieder nach Gedichten von Stefan George op. 4

A Autograph von Nr. I (New York, NY, The Morgan Library, Dept. of Music Manuscripts and Books, Robert Owen Lehman Collection, W376.L716)

B Autograph von Nr. II, III und IV (Basel, Paul Sacher Stiftung, Sammlung Anton Webern)

C Abschrift fremder Hand von Nr. I und II (Basel, Paul Sacher Stiftung, Sammlung Anton Webern)

[D] Autograph von Nr. V. Stichvorlage für **E**. **Verschollen**.

E Druck von Nr. V. Beilage in: *Der blaue Reiter*, München: Piper, 1912

E¹ Handexemplar von **E** mit Korrekturen Weberns (New York, NY, The Morgan Library, Dept. of Music Manuscripts and Books, PMM 21)

F Autograph (Basel, Paul Sacher Stiftung, Sammlung Anton Webern)

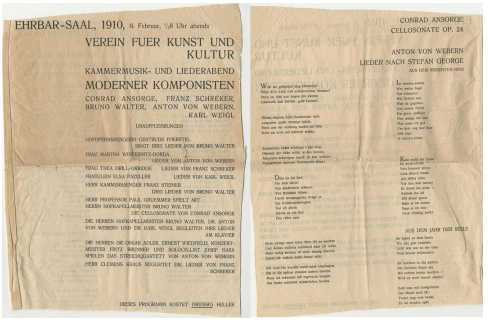
G Abschrift fremder Hand von **F**. Stichvorlage für **H** (US-NYpm, Dept. of Music Manuscripts and Books, Robert Owen Lehman Collection, W376.L716)

H Druck. Wien: Universal Edition, 1923

Abb. 4: Vorläufige Quellenübersicht zu Weberns *Fünf Liedern nach Gedichten von Stefan George* op. 4. [Anton Webern Gesamtausgabe 2018. CC-BY-NC-SA 4.0.]

Und dennoch gibt es einen ›Unruhestifter‹, einen Unsicherheitsauslöser: Denn während der genannte Druck von »Ihr tratet zu dem Herde« im Blauen Reiter (Quellensigle E) überliefert ist, fehlt ein entsprechendes Manuskript, das als Stichvorlage gedient haben könnte. Der Druck ist somit die einzige bekannte existierende Quelle des Liedes in dieser frühen Fassung. Wie aber kann das Stück 1912 in München publiziert worden sein, wenn es keine Stichvorlage gab? Zwar hielt sich Webern im November 1911 anlässlich der Uraufführung von Gustav Mahlers Das Lied von der Erde unter Bruno Walter nachweislich in München auf, es ist aber kaum davon auszugehen, dass er das Lied einem Stecher vor Ort diktiert hätte. Vielmehr lässt sich mit großer Sicherheit annehmen, dass ein vormals existierendes Manuskript verschollen ist. In der Quellenübersicht ist dieses unter der eingeklammerten Quellensigle [D] vermerkt. Das ist aber erst der Anfang einer philologischen Spurensuche, an deren Ende sich diese Fehlstelle vervierfacht haben wird.

Die wichtigsten Belege für eine solche Aufgliederung sind verschiedene Briefpassagen Weberns, in denen er sich zu dem Stück äußert, sowie das Konzertprogramm der Uraufführung 1910 in Wien, das 2016 in den Karl Weigl Papers der Yale University wiederentdeckt wurde (vgl. Abbildung 5).


EHRBAR-SAAL, 1910, 8. Februar, ½8 Uhr abends

VEREIN FUER KUNST UND KULTUR

KAMMERMUSIK- UND LIEDERABEND MODERNER KOMPONISTEN

CONRAD ANSORGE, FRANZ SCHREKER, BRUNO WALTER, ANTON VON WEBERN, KARL WEIGL

CONRAD ANSORGE
CELLOSONATE OP. 24

ANTON VON WEBERN
LIEDER NACH STEFAN GEORGE

Abb. 5: Konzertprogramm für den »Kammermusik- und Liederabend moderner Komponisten« am 8. Februar 1910 im Verein für Kunst und Kultur, Wien (New Haven, CT, Yale University, Irving S. Gilmore Music Library, The Papers of Karl Weigl, MSS 73). [Public Domain Marked.]

Aus dem Konzertprogramm wird ersichtlich, welche sechs der insgesamt 14 überlieferten George-Lieder an der Uraufführung beteiligt waren (in der Aufführungsreihenfolge: op. 4/1, op. 3/1, op. 3/2, op. 3/5, op. 4/2, op. 4/5), darunter als letztes auch »Ihr tratet zu dem Herde«. Interessanterweise existieren für die anderen fünf beteiligten Lieder Abschriften von unbekannter Hand, deren aufführungspraktische Eintragungen darauf schließen lassen, dass sie anlässlich dieser Uraufführung angefertigt und verwendet wurden. In der Quellenübersicht (Abbildung 4) sind unter der Quellensigle C die Abschriften der Lieder Op. 4/1 und Op. 4/2 aufgelistet; entsprechende Abschriften existieren auch für die aus dem späteren Opus 3 an der Uraufführung beteiligten Lieder. Aber wo ist die Kopie des letzten Liedes »Ihr tratet zu dem Herde«? Sollte man nicht annehmen, dass auch hier eine aufführungspraktische Einrichtung existiert haben muss? Die vermisste Quelle [D] zerfällt somit sehr wahrscheinlich in ein vermisstes Autograph [Da] und eine vermisste Abschrift [Dc]. Erfreulicherweise wird diese bislang doch recht vage Annahme durch einen Brief Weberns an seinen Schwager Paul Königer vom November 1911 untermauert:

> »Sie wollen dieses Lied von mir; ja ich habe die zwei Exemplare, die ich hatte, verschickt an den ›blauen Reiter‹ und an den Verleger. Ich würde es aber, wenn das noch möglich ist, aus der Skizze neuerdings abschreiben und Ihnen schenken.«[26]

Der genannte »Verleger« ist Gerhard Tischer (Tischer & Jagenberg) in Köln, bei dem Webern insgesamt zehn George-Lieder publizieren wollte (was es sehr wahrscheinlich macht, dass Webern das Autograph nach Köln und die Abschrift nach München geschickt hatte). Zumindest hat es zwei Manuskripte (»Exemplare«) gegeben, von denen Webern im Herbst 1911 keines mehr in seinen Händen hielt. Da Königer aber offensichtlich sehr eindringlich um eine Kopie des Stückes gebeten haben muss, fährt Webern im selben Brief fort, dass er im Bedarfsfall eine Abschrift »aus der Skizze« anfertigen und Königer schenken werde. Diese Aussage Weberns ist äußerst mysteriös, denn es gibt für keines seiner George-Lieder auch nur einen Hinweis auf so etwas wie Skizzenmaterial. Entweder muss dieses als vermisst bzw. vernichtet für den gesamten Werkkomplex gelten; oder aber die Skizzen haben tatsächlich nie existiert, wobei es sich um eine fälschliche Deklaration Weberns handeln würde (»falsely declared« in Abbildung 1). Möglicherweise wollte er Königer nur versichern, dass dieser in jedem Fall eine Kopie bekäme, und spekulierte auf die zwischenzeitliche Rückkehr seiner Manuskripte aus Köln oder München.

Im Januar 1912 überschlagen sich dann nach langer Wartezeit die Ereignisse: Webern erhält tatsächlich sein Manuskript aus Köln zurück – zusammen mit der für ihn unerfreulichen Absage des Verlegers Tischer[27] – und zwei Wochen später ein Korrekturexemplar aus München, wie aus einem Brief Weberns an Alban Berg hervorgeht.[28] Er kann so immerhin die Korrekturen für den

[26] Brief Weberns an Paul Königer, 23. November 1911 (Österreichische Nationalbibliothek, Sammlung von Handschriften und alten Drucken, Autogr. 975/7–6).
[27] »Mein Lied bekommst Du bald. Es erscheint übrigens im ›blauen Reiter‹ […] Dr Tischer hat mir meine Noten wieder zurückgeschickt. Ich bin also zum 3. Male abgewiesen worden.« (Brief Weberns an Paul Königer, 11. Januar 1912; Österreichische Nationalbibliothek, Sammlung von Handschriften und alten Drucken, Autogr. 975/7–8).
[28] »Ich habe vorgestern die Korrektur meines Liedes das im ›blauen Reiter‹ erscheint bekommen. Du auch die Deines Liedes?« (Brief Weberns an Alban Berg, 25. Januar 1912; Österreichische Nationalbibliothek, Musiksammlung, L6. Alban-Berg-Stiftung.202).

Blauen Reiter anhand des aus Köln zurückerfolgten Manuskripts ausführen. Aber für die hier aufgestellte Fehlliste heißt das, dass nicht nur die möglicherweise fälschlich erklärten Skizzen [Dsk], sondern auch noch ein nur in letztgenanntem Brief erwähntes Korrekturexemplar [Dpc] hinzugefügt werden müssen.

Man kann davon ausgehen, dass Webern das Korrekturexemplar nach München zurückgeschickt und wenig später Königer das behaltene Manuskript des Liedes als Geschenk überreicht hat, womit er ein zweites Mal kein Manuskript mehr zur Verfügung hatte. Was dies zu einer plausiblen Annahme und einem plausiblen historiographischen Narrativ macht, ist die für das gesamte Oeuvre Weberns sehr auffällige Tatsache, dass Webern im weiteren Kompositionsprozess seine Änderungen und Korrekturen nicht wie üblich direkt in ein bestehendes Manuskript eintrug, sondern – offensichtlich in Ermangelung eines solchen – gezwungen war, sein Belegexemplar des Blauen Reiters (vgl. Abbildung 4, Quellensigle E1), das er Anfang Juni 1912 erhielt, zu verwenden.[29]

Wie sich zeigt, musste die erste Annahme, dass es sich nur um eine einzige fehlende Stichvorlage handelt, korrigiert werden zu der plausiblen Vorstellung, dass in diesem Fall mindestens drei, wenn nicht vier Deperdita involviert sind. Wenn man davon ausgeht, dass alle dieser Deperdita mehr oder weniger stark modifizierte Fassungen des Liedes »tratet zu dem Herde« enthalten haben können (was im Hinblick auf Weberns Arbeitsweise sehr wahrscheinlich ist), lässt sich das Ausmaß der Lücke im kulturellen Gedächtnis erahnen, die durch das »quellenlose Vakuum« zwischen Sigle C und E erzeugt wird. Durch Befragen des überlieferten Kontextmaterials konnte jedoch der durch den Verlust der Manuskripte eröffnete Unsicherheitsraum mithilfe von ineinander verwobenen Indizien, Referenzen und Hinweisen produktiv aufgefüllt werden. Am praktischen Beispiel wird hier deutlich, dass sich eine Handlungsfähigkeit – wie bereits in Abschnitt 2.3 dargelegt – in solch einem Fall nur durch ein plausibles historiographisches Narrativ aufrechterhalten lässt, soll heißen: Die bloße Information unterrepräsentiert den eigentlichen Sachverhalt, es bedarf eines interpretatorischen, fachwissenschaftlich informierten und begründeten Narrativs, einer kontextuellen Erläuterung und Einbettung, um die Grenzen des Unsicherheitsbereichs abzustecken und die entworfene Hypothese darin zu verorten und einzubetten.

Während der Verlust von Quellen also bereits eine deutliche Herausforderung für die »traditionelle« (Musik-)Philologie darstellt, gilt dies umso mehr für die Beschreibung und Modellierung von Deperdita in einem digitalen, computergestützten Kontext. Da die Anton Webern Gesamtausgabe als sogenannte Hybrid-Ausgabe[30] konzipiert ist, sieht auch sie sich mit dieser Problemstellung konfrontiert. Der letzte Abschnitt des Beitrags soll sich daher diesem Aspekt widmen.

[29] Handexemplar in US-NYpm, Dept. of Music Manuscripts and Books, PMM 21.

[30] Neben Druckbänden, die bei der Universal Edition in Wien erscheinen werden, sollen die digitalen Anteile der AWG im Rahmen des am DHLab der Universität Basel entwickelten Software-Frameworks Knora (mit dem User-Interface Salsah) online zugänglich gemacht werden. *Knora* ermöglicht es, anwendungsspezifische Modelle um projektspezifische Ontologien zu ergänzen, die innerhalb von *Knora / Salsah* erstellt, bearbeitet und verknüpft werden können. Zudem lassen sich Faksimiles IIIF-konform einbinden, anzeigen und annotieren, was auch im Bereich von Notengrafiken einen möglichen tragfähigen Ansatz darstellt. Seit längerem wird Salsah bereits als Archiv-Datenbank für Kontextmaterialien und für die Dokumentensammlung der AWG produktiv eingesetzt, ein Bereich für die konkrete editorische Arbeit, wobei

4. Digitale Gedächtnislücken: graphbasierte Modellierung von Deperdita

Das kulturelle Gedächtnis wandelt sich zwangsläufig unter den Paradigmen von Digitalität:[31]

> »Individuen und Kulturen bauen ihr Gedächtnis interaktiv durch Kommunikation in Sprache, Bildern und rituellen Wiederholungen auf. Beide, Individuen und Kulturen, organisieren ihr Gedächtnis mit Hilfe externer Speichermedien und kultureller Praktiken. Ohne diese läßt sich kein generationen- und epochenübergreifendes Gedächtnis aufbauen, was zugleich bedeutet, daß sich mit dem wandelnden Entwicklungsstand dieser Medien auch die Verfaßtheit des Gedächtnisses notwendig mitverändert.«[32]

So fließen zunehmend maschinell aufbereitete und vor allem verbreitete Informationen in das kulturelle Gedächtnis ein und transformieren es in ein ›digital geprägtes‹. Wenn man so will, wird es um die Fähigkeit – aber auch Verpflichtung – erweitert, »kulturelles Wissen über digitale und vernetzte Medien dauerhaft und unverfälscht über Generationen hinweg weiterzugeben«[33] (vgl. Video 1).

Video 1: Interview mit Ellen Euler während der 4. »Das ist Netzpolitik!«-Konferenz am 1. September 2017 in Berlin zum Thema »Freier Zugang zum digitalen Gedächtnis!?« [netzpolitik.org 2017. CC-BY-NC-SA 4.0.] [online].

Soll auch ein ›digitales Gedächtnis‹ angeblich niemals vergessen, ist es doch vor Fehlstellen und ›Gedächtnislücken‹ genauso wenig gefeit wie sein rein ›analoges‹ Pendant: sei es durch unwiderrufliches Überschreiben, durch Abschaltung, durch veraltete Hardware oder schlichtweg Desinteresse. Zudem erbt es die ›analogen‹ Fehlstellen wie auch ›analogen‹ Vorurteile.[34] Es bedarf also auch hier vermehrter Anstrengungen, um Gedächtnislücken in digitalen Wissensstrukturen zu verhindern.

4.1 Ontologien als Zugriffsmöglichkeiten auf Weltwissen

Wenn es um die computerbasierte Modellierung von Wissensstrukturen geht, sei vorangestellt, dass jedes Modell nur ein reduziertes, vereinfachendes und unvollkommenes »Surrogat« des betrachteten Weltausschnitts und weder allumfassend noch abschließend sein kann.[35] Es gibt

---

die Editoren direkt innerhalb der Forschungsumgebung ihre Kritischen Berichte und Editionen erstellen können, befindet sich, auch in Kooperation mit dem schweizweiten Projekt Nationale Infrastruktur für Editionen (NIE-INE), im Aufbau. Vgl. den Editionsprototyp der Anton Webern Gesamtausgabe sowie die NIE-INE-Projektwebsite.

[31] Zu einer Problematisierung des Begriffs vgl. Reichert 2017, passim.

[32] Assmann 2006, S. 19.

[33] So die Definition Ellen Eulers für ein ›digitales Gedächtnis‹ in Lebert 2017. Vgl. auch den vollständigen Vortrag Eulers während der 4. »Das ist Netzpolitik«-Konferenz (Berlin, 1. September 2017). Was hierbei ›dauerhaft‹ und ›unverfälscht‹ konkret bedeutet bzw. bedeuten kann, bleibt wohl noch für einige Zeit Gegenstand der fachwissenschaftlichen und -politischen Diskussionen.

[34] Vgl. Caliskan et al. 2017, passim; Zillien 2009, passim.

[35] Vgl. Davis et al. 1993, S. 18f. Zum Modellbegriff vgl. grundlegend Stachowiak 1973, insb. S. 128–133, sowie den Beitrag von Michael Piotrowski im vorliegenden Sonderband, Piotrowski 2019.

somit weniger ›richtige‹ Wege, als vielmehr bestmögliche Pfade, auf denen es – im Zweifelsfall bis zur erfolgreichen Feststellung einer Sackgasse – den jeweiligen Untersuchungsgegenstand, die darauf anzuwendenden Fragestellungen sowie die eigene Perspektive zu erproben gilt. Einen möglichen Pfad stellen dabei semantische Technologien dar, wie sie in der Vision eines Web of Data (besser bekannt als *Semantic Web*) formuliert sind.[36] Dahinter steht die Idee, das World Wide Web um Technologien zu bereichern, die in der Lage sind, semantisch aufbereitete Informationen und Kontextualisierungen zu integrieren und zu verarbeiten. Dabei bilden Ontologien gleichsam das Herzstück, den zentralen Block, mit deren Hilfe Informationen über einen betreffenden Weltausschnitt formalisiert, strukturiert, kontextualisiert und semantisch vernetzt werden und so zu einer Systematisierung und Modellierung von Wissensstrukturen beitragen können. Laut der gängigsten Definition handelt es sich bei diesen semantischen Geflechten – nichts anderes sind im informationstheoretischen Kontext ›Ontologien‹ – um »explicit, formal specification[s] of a shared conceptualisation«,[37] explizite und formale Ausführungen eines gemeinsam genutzten und verstandenen Konzepts. Dass derartige Denkmodelle und Zugriffsmöglichkeiten auf ein bestimmtes Verständnis von Welt nicht erst ein Produkt des Computerzeitalters darstellen, zeigt ein Blick auf ein mehr als 800 Jahre altes Diagramm des Scholastikers Petrus von Poitiers (um 1125/1130–1215):

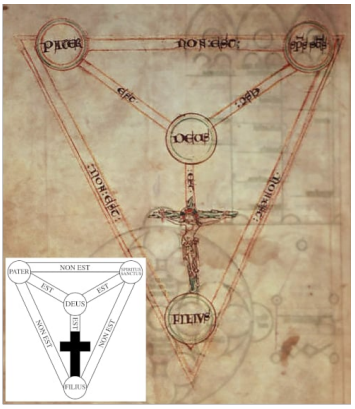


Abb. 6: Hintergrund: Früheste bekannte Fassung des »Scutum Fidei«-Diagramms von Petrus von Poitiers, um 1210 (British Library, Cotton Faustina manuscript B. VII, folio 42v). [British Library, via Wikimedia Commons. Public Domain Marked.] Vordergrund: Transkription und schematische Übertragung des »Scutum fidei«-Diagramms. [AnonMoos 2018, via Wikimedia Commons. Public Domain Marked.]

Dieses *scutum fidei* (»Schild der Dreieinigkeit«) erfüllt vollständig die Anforderungen der obigen Definition von Ontologien: Es gibt eine gemeinsame Konzeptualisierung (das christliche Gottesbild und die Dreieinigkeit des Vaters, des Sohnes und des Heiligen Geistes) sowie dessen explizite Formalisierung in Form eines scholastischen Diagramms. Darüber hinaus, und das ist der eigentlich entscheidende Punkt, werden die Konzepte durch Relationen in Beziehung zueinander gesetzt, damit kontextualisiert und verknüpft: eine »est«-Beziehung

[36] Vgl. Berners-Lee 1998, passim; Berners-Lee et al. 2001, passim. Speziell zum Potenzial für die geisteswissenschaftliche Forschung vgl. Oldman et al. 2014, passim; Oldman et al. 2016, passim.
[37] Definition nach Studer et al. 1998, S. 184, die eine Präzisierung derjenigen von Gruber 1993, S. 199, darstellt. Vgl. allgemein zu Ontologien auch Rehbein 2017, passim; Stuckenschmidt 2011, passim.

(Vater, Sohn und Heiliger Geist sind Gott) und gleichzeitig eine »non est«-Beziehung (Vater, Sohn und Heiliger Geist sind nicht identisch). In der Terminologie von Ontologien ließe sich dies als Disjunktheit (*disjointness*) bezeichnen, d. h., verschiedene Klassen können nicht dieselbe Instanz teilen.[38] Auf einer abstrakten Ebene lassen sich in dieser frühen Ontologie somit Vater, Sohn und Heiliger Geist als disjunkte Subklassen von Gott auffassen. Der Schritt zu einer maschinensprachlich verarbeitbaren Darstellung ist nur noch ein kleiner: Abbildung 7 und Abbildung 8 zeigen den »Schild der Dreieinigkeit« modelliert in der Web Ontology Language *OWL*.[39]

```
@prefix : <http://www.semantic-example.org/ontology/2018/07/shieldoftrinity#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@base <http://www.semantic-example.org/ontology/2018/07/shieldoftrinity> .

<http://www.semantic-example.org/ontology/2018/07/shieldoftrinity> rdf:type owl:Ontology .

:God rdf:type owl:Class .

:Father rdf:type owl:Class ;
        rdfs:subClassOf :God .

:Son rdf:type owl:Class ;
     rdfs:subClassOf :God .

:HolySpirit rdf:type owl:Class ;
            rdfs:subClassOf :God .

[ rdf:type owl:AllDisjointClasses ;
  owl:members ( :Father
                :HolySpirit
                :Son)
] .
```

Abb. 7: Modellierung des »Scutum fidei« in OWL (Turtle-Notation). [Eigene Grafik 2018. CC-BY 4.0.]

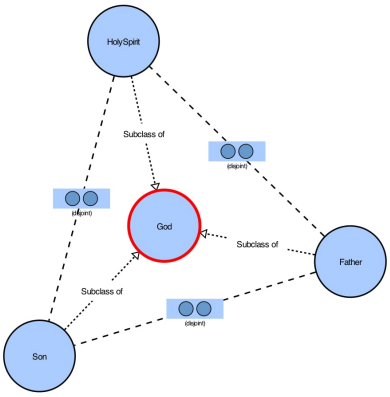


Abb. 8: Modellierung des »Scutum fidei« in OWL (Graph-Visualisierung). [Eigene Grafik 2018, generiert mit WebVowl 1.1.1. CC-BY 4.0.]

4.2 Ein Ding der Unmöglichkeit? Zur Modellierung von Negationen

Wie gesehen, können Ontologien Zugang zu Wissen über die Welt sowohl in für Menschen verständlicher als auch maschinenverarbeitbarer Weise ermöglichen. Dabei stellt die Modellierung von Negation oder Nicht-Existenz jedoch eine Herausforderung dar, denn sie

[38] Grundsätzlich zur Frage von Disjunktheit in Ontologien vgl. Stevens / Sattler 2014, passim.
[39] Spezifikation unter OWL 2 Web Ontology Language. Structural Specification and Functional-Style Syntax (Second Edition).

ist nicht trivial – wenn man, was nicht nur im Bereich von Semantic Web-Modellierung zu empfehlen ist, eine offene Welt voraussetzt (*Open World Assumption*).[40] Unter dieser Annahme stellen nicht getroffene Aussagen über eine Ressource kein Votum zu deren Nicht-Existenz dar, wie unter Annahme einer geschlossenen Welt (*Closed World Assumption*), sondern bedeuten eben nur, dass momentan noch keine Aussage über diesen Sachverhalt getroffen wurde, was aber jederzeit nachgeholt werden könnte. Eine Nicht-Existenz im eigentlichen Sinne (ein logisches NOT) existiert hier nicht. Aber es braucht Negation, um fehlende Quellen zu modellieren. Zwar ließe sich eine negative Zuweisung
<A> <hasNoSource> <Source>
problemlos erstellen, allerdings offenbart sich deren negierende Bedeutung nur einem menschlichen Anwender. Für eine Maschine ist der entsprechenden Klasse ein Attribut zugewiesen, das zwar mit der Zeichenkette »hasNoSource« bezeichnet ist, sie besitzt aber dennoch ein Attribut. Am Beispiel des »Schildes der Dreieinigkeit« lässt sich dies erneut gut verdeutlichen: Dessen Schöpfer benutzte eine vermeintliche Nicht-Beziehung (»non est«), aber gerade durch diese treten die betreffenden Entitäten (Vater, Sohn, Heiliger Geist) in eine tatsächliche Beziehung, die nämlich besagt, dass sie nicht gleich sind. Es bleibt immer eine existierende Beziehung, unabhängig davon, ob man sie »nicht vorhanden« oder »keine Beziehung« nennt. Paradoxerweise muss offensichtlich eine Beziehung etabliert werden, um ihre Nicht-Existenz zu modellieren. Bei einer entsprechenden Suchabfrage (z.B. in SPARQL), welche Instanzen etwas NICHT haben, müsste man also Instanzen finden, die ein Attribut haben, das eine Verneinung beschreibt. Mittlerweile wurden mit *NOT EXISTS* bzw. *MINUS* der Spezifikation von SPARQL zwei Modifikatoren hinzugefügt, die zumindest in der Datenabfrage eine Suche nach nicht vorhandenen Datenmustern bzw. das Entfernen bestimmter Muster aus dem Suchergebnis ermöglichen.[41]

Doch auch auf der Modellierungsseite gibt es bereits bestehende Ansätze zum Umgang mit Negationen: OWL zum Beispiel kennt neben der bereits erwähnten Disjunktheit (*disjointWith*) noch *complementOf* (Klassen in komplementärem Verhältnis) und *negativePropertyAssertion* (ein Subjekt ist mit einem Objekt nicht durch die benannte Property verbunden).

owl:disjointWith

```
:Book rdf:type owl:Class ;
      owl:disjointWith :Writer .
```

owl:complementOf

```
:Book rdf:type owl:Class ;
      rdfs:subClassOf [
            owl:complementOf :Writer
] .
```

[40] Zu den Implikationen einer *Open World Assumption* vgl. u. a. Allemang / Hendler 2011, S. 10f.
[41] Vgl. den Abschnitt zu Negation in der SPARQL-Spezifikation und die entsprechende Diskussion der SPARQL Working Group des W3-Consortiums. Vgl. auch das Kapitel *7.4.3 Default Negation in SPARQL* in Isaac 2011, S. 123f., sowie DuCharme 2011, S. 57–59; Walsh 2004, passim.

owl:negativePropertyAssertion

```
:npa rdf: type owl:negativePropertyAssertion ;
      owl:sourceIndividual :A ;
      owl:assertionProperty :has Source ;
      owl:targetIndividual :Source .
```

Einen anderen Ansatz verfolgt das Argumentationsmodell CRMinf, das im Orbit von CIDOC CRM entwickelt wurde.[42] Dieses Modell verwendet ein Pattern mit einem Glaubenssatz (»I2 Belief«), der einer bestimmten Aussage (»I4 Proposition Set«) einen entsprechenden Wahrheitswert (»I6 Belief Value«) zuordnet (vgl. Abbildung 9 und Video 2).[43]

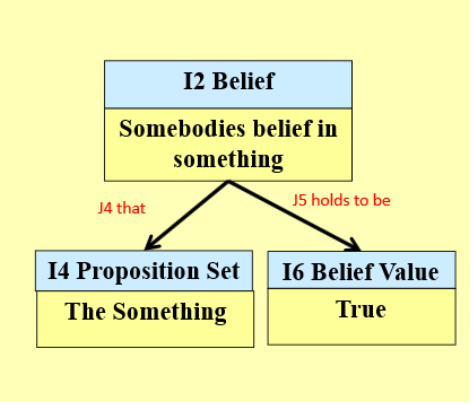


Abb. 9: Belief-Pattern im CRMinf-Modell (aus: Stead 2015. [CIDOC CRM 2015. CC-BY 4.0.]
Video 2: Stephen Stead: CRMinf: the Argumentation Model. Video-Aufnahme anlässlich einer Präsentation beim CIDOC CRM Special Interest Group-Meeting im OeRC (University of Oxford's e-Research Centre), Oxford, im Februar 2015. [CIDOC CRM 2015. CC-BY 4.0.] [online]

Dieses Modell hat das Potenzial, besonders im Bereich der (digitalen) Geisteswissenschaften von großem Nutzen sein zu können, erlaubt es doch den Umgang mit Unsicherheit, Zweifel, Hypothesen oder jeglicher Art von argumentativen Schlussfolgerungen. Allerdings sind bislang nur wenige Projekte bekannt, welche dieses Modell umfassend auf eine reale Fragestellung angewandt hätten.[44] Wenn man sich noch einmal die Problematik der möglichen Skizzen für Weberns »Ihr tratet zu dem Herde« (vgl. Abschnitt 3.2) vor Augen führt, wird ziemlich schnell klar, warum: Es wird sehr, sehr schnell sehr komplex. Abbildung 10 zeigt nur den »einfachen« Fall, dass die Überzeugung eines Forschers, Webern hätte diese Skizzen falsch deklariert, von

[42] Spezifikation unter CRMinf: the Argumentation Model. Mein Dank gilt Dominic Oldman vom British Museum, der mich auf dieses Modell aufmerksam gemacht hat, sowie Stephen Stead für weiterführende Informationen.
[43] Weiterführend zu Belief-Patterns vgl. auch den Beitrag von Michael Piotrowski im vorliegenden Sonderband, Piotrowski 2019.
[44] Unter anderem wird CRMinf in der vom British Museum betreuten Plattform ResearchSpace implementiert.

einem zweiten Forscher übernommen und akzeptiert wird. Es lässt sich leicht die zunehmende Komplexität vorstellen, wenn Widersprüche oder wissenschaftliche Kontroversen in die Modellierung mit einbezogen werden.[45]

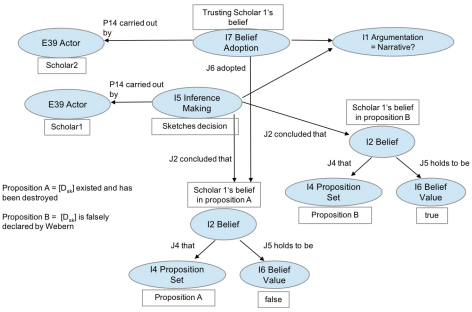


Abb. 10: Anwendung des Argumentations-Modells CRMinf: ein Glaubenssatz von Forscher 1 wird von Forscher 2 übernommen (nach: Stead 2015. [Eigene Grafik 2018. CC-BY 4.0.]

4.3 Deperdita im Modell

Abschließend sollen die vorstehend angerissenen Problemfelder in einem Modellierungsvorschlag zusammengeführt werden. Ein entsprechendes Datenmodell sollte nunmehr

1. der Definition und dem Begriff von »Quelle« genügen,
2. in der Lage sein, sowohl existierende Quellen als auch Deperdita abzubilden,
3. die unterschiedlichen Kategorien von Deperdita umfassen,
4. Informationen über Herkunft, möglichen Standort, Erreichbarkeit und Verfügbarkeit aufnehmen können,
5. den Aspekt des Narrativs, der Hypothesenbildung einschließen und deutlich machen.

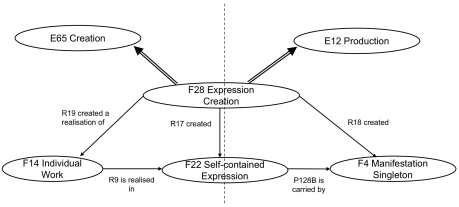


Abb. 11: Doppelt dreistufiges Pattern in FRBRoo, nach: IFLA Working Group on FRBR/CRM Dialogue, S. 21. [Eigene Grafik 2018. CC-BY 4.0.]

[45] Prinzipiell ist eine solche Komplexität durchaus zu begrüßen. Denn allzu oft zwingen doch argumentativ beschränkende bzw. beschränkte Gegebenheiten »digitaler« Anwendungen gegenüber »analogen« wissenschaftlichen Standards und *Best Practices* zurückzustecken. Unterkomplexität oder Unterspezifiziertheit können auch eine Form digitaler Gedächtnislücken provozieren.

Als Grundlage des Modells wird auf ein doppelt dreistufiges Pattern aus *FRBRoo* [46] zurückgegriffen (vgl. Abbildung 11), in welchem ein abstraktes Werkkonzept (*F14 Individual Work*) und dessen individuelle Realisierung (*F22 Self-contained Expression*) in einem Entstehungsprozess (F28 Expression Creation) erschaffen werden (konzeptuelle Ebene). Zugleich entsteht dabei ein manifester Zeichenträger (*F4 Manifestation Singleton*), der den Inhalt der *F22 Self-contained Expression* aufnimmt (physische Ebene). Hierdurch lassen sich differenzierte Aussagen sowohl über den Entstehungs- als auch den Produktionsprozess z.B. einer Komposition treffen. Hinzu kommt, dass dieses Pattern die Anbindung und Nachnutzung weiterer Modelle ermöglicht: So wird es z.B. auch vom französischen DoReMus-Projekt (*Doing Reusable Musical Data*)[47] verwendet, das die momentan wohl umfassendste Ontologie im Bereich der Katalogisierung musikalischer Daten entwickelt hat unter ausgiebiger Nachnutzung von *Music Ontology*, *CIDOC CRM*, *FRBRoo* und *Europeana Data Model*.

Die in Abbildung 12 farbig markierten Konzepte und Relationen sind Erweiterungsvorschläge dieses Grundmodells (schwarz), die vor allem die Rolle des tatsächlichen Zeichenträgers (*Manifestation Singleton*) betreffen. Zum einen soll die ausführliche Beschreibung eines Zeichenträgers während einer editorischen Tätigkeit als Quelle definiert und verwendet werden können. Zum anderen sollte durch einen Bewertungsprozess (*Existence Status Assignment*) der Überlieferungsstatus des Zeichenträgers sowie Zugangsmöglichkeiten und Aufbewahrungsort festgehalten werden – unter Einbezug des CRMinf-Argumentationsmodells.

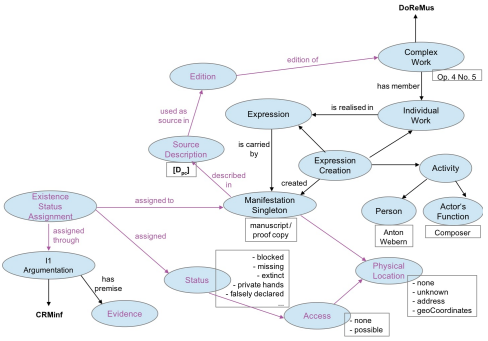


Abb. 12: Modellierungsvorschlag für Deperdita unter Rückgriff auf FRBRoo, CRMinf und DoReMus-Ontologie (schwarz). Erweiterungsvorschläge farbig. [Eigene Grafik 2018. CC-BY 4.0.]

[46] Eine objekt-orientierte Definition der »Functional Requirements for Bibliographic Records« (FRBR). Vgl. IFLA Working Group on FRBR/CRM Dialogue 2015.
[47] Vgl. die DoReMus-Projektwebsite.

5. Schluss

»Zweifel [Unsicherheit] ist kein sehr angenehmer Zustand, aber Versicherung [Sicherheit] ist lächerlich«, so schrieb es Voltaire an Friedrich den Großen.[48] Dies kann mithin als Trost gelten, wenn die Herausforderungen und Schwierigkeiten, die sich bei der Modellierung von Wissensstrukturen im Allgemeinen und Nicht-Existenzen im heutigen digitalen Kontext ergeben, in einem Artikel wie diesem nicht aufzulösen sind, und möglicherweise mehr Fragen als Antworten hinterlassen. Als ›Surrogat‹ des bisherigen Diskurses kann er selbigen wieder anreichern, weitertragen und neue Wissensrepräsentationen herausfordern. Er könnte aber genauso verloren gehen, falsch deklariert, zitiert oder falsch zugeschrieben werden, schließlich gar der Zugriff verweigert werden, zu einem Gerücht verkümmern. Der ›deperditären‹ Verlustmöglichkeiten sind viele und sie haben Auswirkungen auf das ›digitale Gedächtnis‹. Unabhängig vom Schicksal des vorliegenden Beitrags – und das ist das Beruhigende – tragen unzählige digitale Wissensrepräsentationen zu diesem erweiterten kulturellen Gedächtnis bei, wie es handschriftliche oder gedruckte Materialien lange Zeit getan haben und weiterhin tun. Und es liegt in unserer Verantwortung, in der Verantwortung von Domainexperten, Fachwissenschaftlern, Gedächtnisinstitutionen, Datenmodellierern und IT-Spezialisten, sowie allen, die an den Schnittstellen und der Peripherie involviert sind, dieses kulturelle (digitale) Gedächtnis zu pflegen und anzureichern, unsere Argumente und Zweifel zur Verfügung zu stellen, damit wir unsere Narrative mit, trotz und gerade wegen Unsicherheit auch in diesem jungen, digitalen Medium zu konstruieren und handlungsfähig zu halten und Gedächtnislücken zu minimieren in der Lage sind. Alles andere wäre (mit Voltaire gesprochen) lächerlich.

[48] »Le doute n'est pas un état bien agréable, mais l'assurance est un état ridicule« (Brief Voltaires an Friedrich II. von Preußen, 28. November 1770 – oft fälschlich deklariert als 6. April 1767 –, zit. nach: Voltaire 2017 [1770]).

Bibliographische Angaben

Thomas Ahrend: Zu Anton Weberns George-Vertonung »Erwachen aus dem tiefsten Traumesschosse«. Eine Spurensuche. In: Jahrbuch 2011 des Staatlichen Instituts für Musikforschung Preußischer Kulturbesitz. Hg. von Simone Hohmaier. Mainz 2011, S. 53–73. [Nachweis im GBV]

Thomas Ahrend: Editorische Probleme des vertonten Textes in Anton Weberns George-Vertonungen. Beispiele zur aktuellen Arbeit an der Anton Webern Gesamtausgabe. In: Schweizer Jahrbuch für Musikwissenschaft. Neue Folge 33 (2013). Hg. von Luca Zoppelli. Bern 2016, S. 199–211. [Nachweis im GBV]

Dean Allemang / James Hendler: Semantic Web for the Working Ontologist. Effective Modeling in RDFS and OWL. 2. Auflage. Amsterdam u. a. 2011.[Nachweis im GBV]

Aleida Assmann: Erinnerungsräume: Formen und Wandlungen des kulturellen Gedächtnisses. 3. Auflage. München 2006. [Nachweis im GBV]

Jan Assmann: Kollektives Gedächtnis und kulturelle Identität. In: Kultur und Gedächtnis. Hg. von Jan Assmann / Tonio Hölscher. Frankfurt/Main 1988, S. 9–19. PDF. [online] [Nachweis im GBV]

Jan Assmann: Das kulturelle Gedächtnis. Schrift, Erinnerung und politische Identität in frühen Hochkulturen. 7. Auflage. München 2013. [Nachweis im GBV]

Tim Berners-Lee: What the Semantic Web can represent. In: w3.org. Design Issues. Beitrag vom 17.09.1998. [online]

Tim Berners-Lee / James Hendler / Ora Lassila: The Semantic Web. In: Scientific American 284 (2001), H. 5, S. 34-43. Artikel vom 17.05.2001. [online] [Nachweis im GBV]

Fritz Böhle: Handlungsfähigkeit mit Ungewissheit – Neue Herausforderungen und Ansätze für den Umgang mit Ungewissheit. Eine Betrachtung aus sozioökonomischer Sicht. In: Exploring Uncertainty. Ungewissheit und Unsicherheit im interdisziplinären Diskurs. Hg. von Sabina Jeschke / Eva-Maria Jakobs / Alicia Dröge. Wiesbaden 2013, S. 281–293. [Nachweis im GBV]

Reinhold Brinkmann: Die George-Lieder 1908/9 und 1919/23 – Ein Kapitel Webern-Philologie. In: Webern-Kongress. Hg. von der österreichischen Gesellschaft für Musik. Kassel u. a. 1973, S. 40–50. (= Beiträge / Österreichische Gesellschaft für Musik, [4]. 1972/73) [Nachweis im GBV]

Elmar Budde: Anton Weberns Lieder op. 3. Untersuchungen zur frühen Atonalität bei Webern. Wiesbaden 1971. (= Beihefte zum Archiv für Musikwissenschaft, 9) [Nachweis im GBV]

Aylin Caliskan / Joanna J. Bryson / Arvind Narayanan: Semantics Derived Automatically from Language Corpora Contain Human-Like Biases. In: Science 356 (2017), H. 6334, S. 183–186. [Nachweis im GBV]

Georg von Dadelsen: Über Quellenausfall und Hypothesenbildung. In: Das musikalische Kunstwerk. Geschichte. Ästhetik. Theorie. Festschrift Carl Dahlhaus zum 60. Geburtstag. Hg. von Hermann Danuser / Helga de la Motte-Haber / Silke Leopold / N. Miller. Laaber 1988, S. 127–130. [Nachweis im GBV]

Randall Davis / Howard Shrobe / Peter Szolovits: What is a Knowledge Representation? In: AI Magazine 14 (1993), H. 1, S. 17–33. [online]

Bob DuCharme: Learning SPARQL. Querying and Updating with SPARQL 1.1. Beijing u. a. 2011. [Nachweis im GBV]

Manfred Geier: Wittgenstein und Heidegger. Die letzten Philosophen. Reinbek/Hamburg 2017. [Nachweis im GBV]

Thomas Gruber: A Translation Approach to Portable Ontology Specifications. In: Knowledge Acquisition 22 (1993), H. 5, S. 199–220. [Nachweis im GBV]

Martina Hartmann: Die Edition von Quellen, die es nicht mehr gibt. Die merowingischen und karolingischen Deperdita. In: Pourquoi éditer des textes médiévaux au XXIe siècle? Hg. von Olivier Canteaut / Rolf Große. (Rencontre de la Gallia Pontificia: 8, Paris, 08.05.2013) Paris 2014. (= Discussions, 9) [online]

IFLA Working Group on FRBR/CRM Dialogue: Definition of FRBROO: A Conceptual Model for Bibliographic Information in Object-Oriented Formalism. Hg. von Chryssoula Bekiari / Martin Doerr / Patrick Le Boeuf / Pat Riva. Version 2.4 von November 2015. PDF. [online]

Antoine Isaac / Simon Schenk / Ansgar Scherp: Semantic Web Languages. In: Multimedia Semantics. Metadata, Analysis and Interaction. Hg. von Raphael Troncy / Benoit Huet / Simon Schenk. u. a. 2011, S. 99–128. [Nachweis im GBV]

Exploring Uncertainty. Ungewissheit und Unsicherheit im interdisziplinären Diskurs. Hg. von Sabina Jeschke / Eva-Maria Jakobs / Alicia Dröge. Wiesbaden 2013. [Nachweis im GBV]

Der blaue Reiter. Hg. von Wassily Kandinsky / Franz Marc. München 1912. [Nachweis im GBV]

Der blaue Reiter. Hg. von Wassily Kandinsky / Franz Marc. Faksimile-Druck der Ausgabe München 1912. München 1976. [Nachweis im GBV]

Paul Kirn: Einführung in die Geschichtswissenschaft. Hg. von Joachim Leuschner. 5., bearbeitete und ergänzte Auflage. Berlin 1968. [Nachweis im GBV]

Hiram Kümper: Materialwissenschaft Mediävistik. Eine Einführung in die Historischen Hilfswissenschaften. Paderborn 2014. [Nachweis im GBV]

Yannick Lebert: Interview mit Ellen Euler: Freier Zugang zum digitalen Gedächtnis!? In: netzpolitik.org. Beitrag vom 23.10.2017. [online]

Guido Meyer: Konzepte der Angst in der Psychoanalyse. 2 Bde. Frankfurt/Main 2005. Bd. 1: 1895–1950. (= Wissen & Praxis, 131) [Nachweis im GBV]

Ivan Alexander Muñoz Criollo: Die Rezeption der Philosophie Søren Kierkegaards bei Karl Jaspers und Martin Heidegger. Zürich 2013. DOI: 10.5167/uzh-137922

Dominic Oldman / Martin Doerr / Gerald de Jong / Barry Norton / Thomas Wikman: Realizing Lessons of the Last 20 Years: A Manifesto for Data Provisioning & Aggregation Services for the Digital Humanities. A Position Paper. In: D-Lib Magazine 20 (2014), H. 7-8. DOI: 10.1045/july2014-oldman

Dominic Oldman / Martin Doerr / Stefan Gradmann: Zen and the Art of Linked Data: New Strategies for a Semantic Web of Humanist Knowledge. In: A New Companion to Digital Humanities. Hg. von Susan Schreibman / Raymond George Siemens / John Unsworth. Chichester u. a. 2016, S. 251–273. (= Blackwell Companions to Literature and Culture, 93) [Nachweis im GBV]

Michael Piotrowski: Accepting and Modeling Uncertainty. In: »Die Modellierung des Zweifels – graphbasierte Modellierung von Unsicherheiten«. Hg. von Andreas Kuczera / Thorsten Wübbena / Thomas Kollatz. Wolfenbüttel 2018. (= Zeitschrift für digitale Geisteswissenschaften / Sonderbände, 4) DOI: 10.17175/sb004_006

Malte Rehbein: Ontologien. In: Digital Humanities. Eine Einführung. Hg. von Fotis Jannidis / Hubertus Kohle / Malte Rehbein. Stuttgart 2017, S. 162–176. [Nachweis im GBV]

Ramón Reichert: Theorien digitaler Medien. In: Digital Humanities. Eine Einführung. Hg. von Fotis Jannidis / Hubertus Kohle / Malte Rehbein. Stuttgart 2017, S. 19–34. [Nachweis im GBV]

Johann Sebastian Bach: Choräle und geistliche Lieder, Teil 2. Choräle der Sammlung C.P.E. Bach nach dem Druck von 1784–1787. Kritischer Bericht. Hg. von Frieder Rempp. Kassel u. a. 1996. (= Johann Sebastian Bach. Neue Ausgabe sämtlicher Werke [NBA], III/2.2) [Nachweis im GBV]

Monika Renz: Zwischen Urangst und Urvertrauen. Aller Anfang ist Übergang. Musik, Symbol und Spiritualität in der therapeutischen Arbeit. 2., erweiterte und aktualisierte Neuauflage. Paderborn 2009. [Nachweis im GBV]

Saras D. Sarasvathy: Causation and Effectuation: Toward a Theoretical Shift from Economic Inevitability to Entrepreneurial Contingency. In: Academy of Management Review 26 (2001), H. 2, S. 243–263. [Nachweis im GBV]

Saras D. Sarasvathy: Effectuation: Elements of Entrepreneurial Expertise. Cheltenham u. a. 2008. [Nachweis im GBV]

Ullrich Scheideler: Editorische Grundlagenarbeit. In: Musikphilologie. Hg. von Bernhard R. Appel / Reinmar Emans. Laaber 2017, S. 177–196. (= Kompendien Musik, 3) [Nachweis im GBV]

Katelijne Schiltz: Music and Riddle Culture in the Renaissance. Cambridge u. a. 2015. [Nachweis im GBV]

Nicola Schneider: Die Kriegsverluste der Musiksammlungen deutscher Bibliotheken 1942–1945. Zürich 2013. PDF. [online]

Nicole Schwindt: Quellen, II.1.. In: MGG Online. Hg. von Laurenz Lütteken. Kassel u. a. 2016 [1997]. [online]

Herbert Stachowiak: Allgemeine Modelltheorie. Wien u. a. 1973. [Nachweis im GBV]

Stephen Stead: CRMinf: the Argumentation Model. In: slideplayer.com. Oxford 2015. CIDOC CRM:SIG-Präsentation von 02.2015. [online]

Robert Stevens / Uli Sattler: Disjointness Between Classes in an Ontology. In: Ontogenesis. Artikel vom 8. Mai 2014. [online]

Heinz Stuckenschmidt: Ontologien. 2. Auflage. Berlin u. a. 2011. (= Informatik im Fokus) [Nachweis im GBV]

Rudi Studer / Richard Benjamins / Dieter Fensel: Knowledge Engineering: Principles and Methods. In: Data & Knowledge Engineering 25 (1998), H. 1-2, S. 161–197. [Nachweis im GBV]

Voltaire [François Marie Arouet]: Letter to Friedrich II., King of Prussia. 28. November 1770. Letter-Nr. D16792. In: e-enlightenment.com. Electronic Enlightenment Scholarly Edition of Correspondence. Digital Correspondence of Voltaire. Hg. von Nicholas Cronk. Transkription von Theodor Besterman. Oxford 2017 [2008]. DOI: 10.13051/ee:doc/voltfrVF1210104a1c

Norman Walsh: Not in RDF. In: Norman.Walsh.name 54 (2004), H. 7. Blogbeitrag vom 02.04.2004. [online]

Ludwig Wittgenstein: Über Gewißheit. Hg. von Gertrude Elizabeth Margaret Anscombe und Georg Henrik von Wright. Frankfurt/Main 1970. (= Bibliothek Suhrkamp, 250) [Nachweis im GBV]

Christoph Wulf / Jörg Zirfas: Editorial. In: Paragrana 24 (2015), H. 1, S. 9–10. [Nachweis im GBV]

Nicole Zillien: Digitale Ungleichheit. Neue Technologien und alte Ungleichheiten in der Informations- und Wissensgesellschaft. 2. Auflage. Wiesbaden 2009. [Nachweis im GBV]

Medienverzeichnis

Abb. 1: Zusammenstellung verschiedener Kategorien von Deperdita, unterschieden nach physischer Existenz, Zugangsmöglichkeit, Status, Evidenz und Aufbewahrungsort (DOC = dokumentarisch bezeugte Evidenz). [Eigene Grafik 2018. CC-BY 4.0 .]

Abb. 2: Übersicht über nachweisbare Versuche Weberns, die George-Lieder als Sammlungen zusammenzufassen (Sp. 4–7). Sp. 1: Angabe der heutigen Opus- bzw. Moldenhauernummer (M) eines Liedes. Sp. 2: Dokumentarisch nachweisbare Nummerierungen. Sp. 3: An der Uraufführung 1910 beteiligte Stücke. [Anton Webern Gesamtausgabe 2018. CC-BY-NC-SA 4.0.]

Abb. 3: Abdruck von Anton Weberns George-Vertonung »Ihr tratet zu dem Herde« als Beilage in: Der Blaue Reiter. Kandinsky / Marc 1976 (1912). [Public Domain Marked.]

Abb. 4: Vorläufige Quellenübersicht zu Weberns *Fünf Liedern nach Gedichten von Stefan George* op. 4. [Anton Webern Gesamtausgabe 2018. CC-BY-NC-SA 4.0.]

Abb. 5: Konzertprogramm für den »Kammermusik- und Liederabend moderner Komponisten« am 8. Februar 1910 im Verein für Kunst und Kultur, Wien (New Haven, CT, Yale University, Irving S. Gilmore Music Library, The Papers of Karl Weigl, MSS 73). [Public Domain Marked]

Abb. 6: Hintergrund: Früheste bekannte Fassung des »Scutum Fidei«-Diagramms von Petrus von Poitiers, um 1210 (British Library, Cotton Faustina manuscript B. VII, folio 42v). [British Library, via Wikimedia Commons. Public Domain Marked.] Vordergrund: Transkription und schematische Übertragung des »Scutum fidei«-Diagramms. [AnonMoos 2018, via Wikimedia Commons. Public Domain Marked.]

Abb. 7: Modellierung des »Scutum fidei« in OWL (Turtle-Notation) [Eigene Grafik 2018. CC-BY 4.0.]

Abb. 8: Modellierung des »Scutum fidei« in OWL (Graph-Visualisierung) [Eigene Grafik, generiert mit WebVowl 1.1.1, 2018. CC-BY 4.0.]

Abb. 9: Belief-Pattern im CRMinf-Modell (aus: Stead 2015. [CIDOC CRM 2015. CC-BY 4.0.]

Abb. 10: Anwendung des Argumentations-Modells CRMinf: ein Glaubenssatz von Forscher 1 wird von Forscher 2 übernommen (nach: Stead 2015. [Eigene Grafik 2018. CC-BY 4.0.]

Abb. 11: Doppelt dreistufiges Pattern in FRBRoo, nach: IFLA Working Group on FRBR/CRM Dialogue 2015, S. 21. [Eigene Grafik 2018. CC-BY 4.0.]

Abb. 12: Modellierungsvorschlag für Deperdita unter Rückgriff auf FRBRoo, CRMinf und DoReMus-Ontologie (schwarz). Erweiterungsvorschläge farbig. [Eigene Grafik 2018. CC-BY 4.0.]

Audio 1: Anton Webern, »Ihr tratet zu dem Herde« op. 4 Nr. 5 (Druckfassung 1923). Audio-Aufnahme anlässlich einer Aufführung der *Fünf Lieder nach Gedichten von Stefan George* op. 4 mit Kate Maroney (Mezzosopran) und Daniel Schlosberg (Klavier) im Spectrum, New York City (NY), Mai 2016. [Kate Maroney und Daniel Schlosberg 2016. Mit freundlicher Genehmigung von Kate Maroney und Daniel Schlosberg.] [online]

Video 1: Interview mit Ellen Euler während der 4. »Das ist Netzpolitik!«-Konferenz am 1. September 2017 in Berlin zum Thema »Freier Zugang zum digitalen Gedächtnis!?« [netzpolitik.org 2017. CC-BY-NC-SA 4.0.] [online]

Video 2: Stephen Stead: CRMinf: the Argumentation Model. Video-Aufnahme anlässlich einer Präsentation beim CIDOC CRM Special Interest Group-Meeting im OeRC (University of Oxford's e-Research Centre), Oxford, im Februar 2015. [CIDOC CRM 2015. CC-BY 4.0.] [online]

ZfdG

Zeitschrift für digitale Geisteswissenschaften

Beitrag aus:
Sonderband 4 der ZfdG: Die Modellierung des Zweifels – Schlüsselideen und -konzepte zur graphbasierten Modellierung von Unsicherheiten. Hg. von Andreas Kuczera, Thorsten Wübbena und Thomas Kollatz. 2019.
DOI: 10.17175/sb004

Titel:
Accepting and Modeling Uncertainty

Autor/in:
Michael Piotrowski

Kontakt:
michael.piotrowski@unil.ch

Institution:
Université de Lausanne, Section des sciences du langage et de l'information

GND:
139045368

ORCID:
0000-0003-3307-5386

DOI des Artikels:

10.17175/sb004_006a

Nachweis im OPAC der Herzog August Bibliothek:
1037072987

Erstveröffentlichung:
31.07.2019

Lizenz:

Sofern nicht anders angegeben (cc) BY-SA

Medienlizenzen:
Medienrechte liegen bei den Autoren

Letzte Überprüfung aller Verweise:
30.07.2019

GND-Verschlagwortung:
Computerunterstütztes Verfahren | Digital Humanities | Modellierung | Unsicherheit |

Zitierweise:
Michael Piotrowski: Accepting and Modeling Uncertainty. In: Die Modellierung des Zweifels – Schlüsselideen und -konzepte zur graphbasierten Modellierung von Unsicherheiten. Hg. von Andreas Kuczera, Thorsten Wübbena und Thomas Kollatz. 2019 (= Sonderband der Zeitschrift für digitale Geisteswissenschaften, 4). PDF Format ohne Paginierung. Als text/html abrufbar unter DOI: 10.17175/sb004_006a.

Michael Piotrowski

Accepting and Modeling Uncertainty

Abstracts

Unsicherheit ist eine Herausforderung für die Erstellung informatischer Modelle in den Geisteswissenschaften: wir können sie weder ignorieren, noch eliminieren, daher müssen wir sie modellieren – und benötigen dafür entsprechende formale Modelle. Solche Modelle existieren und werden in verschiedenen Bereichen praktisch eingesetzt. Ebenso gibt es aktive Grundlagenforschung zu Unsicherheit und ihrer Repräsentation in Mathematik, Philosophie und Informatik. Manche der dort entwickelten Ansätze könnten auch für die Modellierung von Unsicherheit in den Geisteswissenschaften interessant sein – aber es fehlt bisher eine Brücke, um diese Ansätze auf ihre Eignung in den Geisteswissenschaften überprüfen zu können. Wir vertreten die Ansicht, dass die digitalen Geisteswissenschaften Unsicherheit über projektspezifische Modelle hinaus betrachten muss; insbesondere müssen wir ein besseres Verständnis von Unsicherheit in den Geisteswissenschaften anstreben, um allgemeinere Ansätze für ihre Behandlung zu entwickeln.

This article aims to outline the challenge of uncertainty for the construction of computational models in the humanities. Since we can neither ignore nor eliminate uncertainty, we need to model it, and so we need computational models of uncertainty. Such models already exist and are being used in practical applications. There is ongoing fundamental research on uncertainty and its representation in mathematics, philosophy, and computer science. Some of these approaches may be suitable for modeling uncertainty in the humanities—but we are still lacking the »bridge« that could relate the uncertainty encountered in the humanities to such formal modeling frameworks. We argue that DH needs to go beyond project-specific models of uncertainty and consider uncertainty more generally; in particular, we must closely examine various types of uncertainty in the humanities and seek to develop more general frameworks for handling it.

1. Introduction

As the saying goes, »nothing can be said to be certain, except death and taxes.« Uncertainty is an unavoidable aspect of life and thus we have an intuitive understanding of uncertainty, but coming up with a strict definition of uncertainty is hard. Uncertainty is generally considered to be related to a lack of information (or *ignorance*) or imperfect information. Ignorance is used here in a non-pejorative sense; Smithson remarks that ignorance is usually treated »as either the absence or the distortion of »true« knowledge, and uncertainty as some form of incompleteness in information or knowledge.«[1] He notes that to »some extent these commonsense conceptions are reasonable, but they may have deflected attention away from ignorance by defining it indirectly as *non*knowledge.«[2]

Predictions, such as about the weather, are generally uncertain, as we only have limited information about the future. But uncertainty does not only concern the future; the following statements could all be said to express uncertainty:

[1] Smithson 1989, p. 1.
[2] Smithson 1989, p. 1, emphasis original.

1. I know Bob is married, but I don't know the name of his spouse.

2. I know Bob is married, but I don't remember whether his wife's name was Alice or Alicia (or was it Alison?).

3. Jack told me that Bob's wife is called Alice, but John said it was Alicia.

4. Bob is about 30.

5. Bob is 30 or 31.

6. Bob is between 30 and 35 years old.

7. Bob is a player.

Upon closer inspection, the uncertainty in these statements not only relates to different pieces of information, but it also takes different forms. For example, in statement 1, the name of Bob's wife is completely unknown, whereas in 2 it is one from a set of names; statement 4 could be called vague, whereas in 5 a set, and in 6 a range of possible ages is given. In addition, not all options may be equally likely; in statement 2, for example, the speaker may consider Alice or Alicia more likely than Alison. In statement 3, uncertainty stems from conflicting information; whether the speaker considers one of the two options more likely may also depend on whether Jack or John is considered (or believed to be) more trustworthy. Finally, in statement 7, uncertainty about the meaning of the statement is caused by the lack of context and by the semantic ambiguity of the word »player.«

These examples are not intended to be exhaustive but rather to illustrate that uncertainty can have different causes, take different forms, and is related to other phenomena such as imprecision, vagueness, and ambiguity; it may also involve issues of belief and trust. These different types of uncertainty may thus have different consequences and may need to be addressed in different ways. Some cases of uncertainty may be resolved, or the uncertainty may at least be reduced; for example, we may be able to ask Bob about the name of his wife, or new information may allow us to narrow the range of ages in statement 6. Predictions about the future, on the other hand, will remain uncertain until the prediction can be compared to the actual outcome. Yet other cases of uncertainty are effectively unresolvable because the required information is inaccessible (such as other people's thoughts, lost documents, or perfectly precise measurements) or nonexistent (e.g., when considering counterfactual questions such as »If John Wooden were alive and coaching in the NCAA today, would he be as successful?«), or because the criteria for deciding an issue are unknown or subjective.

In fact, one may argue that in the general case, uncertainty can never fully be resolved, as we will never have perfect knowledge when we are dealing with the real world; Parsons notes that »any intelligent system, human or otherwise, is constrained to have finite knowledge by virtue of its finite storage capacity, so it will always be possible to find some fact that is unknown by a particular system.«[3]

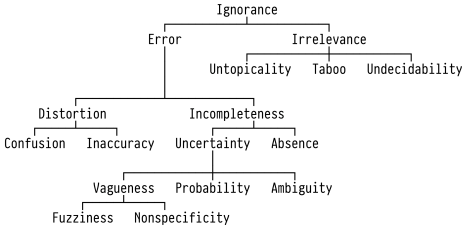


Fig. 1: Smithson's taxonomy of ignorance. [Piotrowski 2019, redrawn after Smithson 1989, p. 9.]

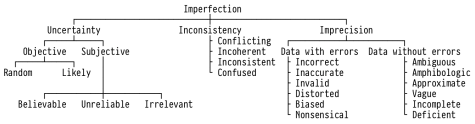


Fig. 2: Smets's taxonomy of imperfection. [Piotrowski 2019, drawn after Smets 1997.]

Various taxonomies have been proposed that aim to systematize uncertainty and related concepts, for example by Smithson (see Figure 1), who considers it a sub-type of a broader concept of *ignorance*,[4] or by Smets, who uses *imperfection* as the general concept (see Figure 2).[5] Parsons discusses further taxonomies that have been proposed,[6] eventually coming to the conclusion that »while it is far from clear that the taxonomies that they provide are of fundamental importance, they do help to outline what uncertainty is.«[7]

As uncertainty is so pervasive, it obviously also affects research and scholarship; Pollack writes: »The uncertainty arises in many ways, and the nature of the uncertainty may change through time, but the scientific endeavor is never free of uncertainty.«[8] Humans are generally quite good at dealing with uncertainty in everyday life, for example with respect to the weather or to public transit. However, these estimations of uncertainty heavily depend on individual knowledge and previous experience, and are generally hard to communicate. For example, what should an out-of-town visitor make of a statement like »usually the train's on time, but sometimes it's delayed quite a bit in the morning, but I think you're gonna be fine«?

[3] Parsons 2001, pp. 7–8.
[4] Smithson 1989.
[5] Smets 1997.
[6] Parsons 2001, pp. 9–15. Smithson originally proposed a diagram Smithson 1989, p. 9, whereas Smets presented his »structured thesaurus of imperfection« in list form Smets 1997, pp. 246–247. In the overview of taxonomies of uncertainty, Parsons presents Smet's taxonomy also in graphical form, but completely omits inconsistency, one of the three main types.
[7] Parsons 2001, p. 10.
[8] Pollack 2005, p. 5.

As scholarly research strives for intersubjectivity, it requires transparency with respect to uncertainty; appealing to »common sense,« experience, or intuition is clearly insufficient. Mathematics and the natural sciences have developed intricate formal methods for dealing with (particular types of) uncertainty, which is probably one reason why »people who are not scientists often equate science with certainty, rather than uncertainty.«[9]

In the humanities, uncertainty is usually described verbally; one domain that describes (again, particular types of) uncertainty in a relatively systematic fashion is that of critical scholarly editions: critical apparatuses record illegible passages, uncertain readings, uncertain identifications of persons and places, and other cases of uncertainty pertaining to the edited sources. In addition, research in the humanities does not only need to deal with uncertain, vague, incomplete, or missing information, but also with an irreducible variety of positions (points of view, values, criteria, perspectives, approaches, readings, etc.), resulting in what Paul Ricœur calls »le conflit des interprétations.«[10]

Now, what about *digital* humanities? If digital humanities is the intersection (or at the intersection?) of computer science and the humanities, as is often said, what does this mean for the handling of uncertainty? For example, Burdick et al. argue that computing »relies on principles that are […] at odds with humanistic methods,«[11] and assert that »ambiguity and implicit assumptions are crucial to the humanities.«[12] »What is at stake,« they conclude, »is the humanities' unique commitment to wrestle with uncertainty, ambiguity, and complexity«.[13] We believe that one cannot really answer this question without defining what one means by »digital humanities«. In the next section we will therefore first present our definition of digital humanities. As we will see, the concept of models is central to our definition; we will thus, in the subsequent section, outline our notion of models. In the following section, we will give a brief overview of the modeling of uncertainty in computer science and in digital humanities. In the final section we will conclude our discussion by outlining what we belief to be the specific challenges for digital humanities and what should be the next steps to advance the state of the art.

2. Defining digital humanities

2.1 Challenge

Kirschenbaum has argued that the »network effects« of blogs and Twitter have turned the term *digital humanities* into a »free-floating signifier«.[14] Perhaps it is not unique to digital humanities, but it is still a rather strange situation that a field tries to constitute itself around a marketing term rather than the other way round. The multifariousness of its understandings is succinctly summarized by Ramsay in the volume *Defining Digital Humanities*:[15]

[9] Pollack 2005, p. 6.
[10] »Cette difficulté – celle-là même qui a mis en mouvement ma recherche – la voici: il n'y a pas d'herméneutique générale, pas de canon universel pour l'exégèse, mais des théories séparées et opposées concernant les règles de l'interprétation.« (Ricœur 1965, p. 37).
[11] Burdick et al. 2012, p. 15.
[12] Burdick et al. 2012, p. 16.
[13] Burdick et al. 2012, p. 108.
[14] Kirschenbaum 2012.
[15] Terras et al. 2013.

> »[...] the term can mean anything from media studies to electronic art, from data mining to edutech, from scholarly editing to anarchic blogging, while inviting code junkies, digital artists, standards wonks, transhumanists, game theorists, free culture advocates, archivists, librarians, and edupunks under its capacious canvas.«[16]

Some claim that a definition is no longer needed:

> »I don't think many people in DH care about definitions too much now. Thankfully the debates have moved on.«[17]

Others even go as far as to argue that it is *impossible* to know what digital humanities is:

> »In closing, I will be as plain as I can be: we will never know what digital humanities ›is‹ because we don't want to know nor is it useful for us to know.«[18]

We should pause briefly at this point and remind ourselves that the question is, in fact, neither what digital humanities *is* ontologically, nor how to exhaustively describe »the disparate activities carried on under its banner.«[19] The question is rather how we *want* to define it—what Carnap[20] called an *explication*. We also contend that it is not only »useful« to explicate, but crucial: the development of a research program, as well as the creation of academic positions, departments, and programs require a consensus around an explicit definition. How would one otherwise ensure the relevance and quality of research, the comparability of degree programs (and thus student mobility), or the adequate evaluation of research programs and thus their financing? And how would one want to cooperate with researchers from other fields?

2.2 Approach

We think the problem of defining digital humanities is unnecessarily exacerbated by confounding a number of related, but actually distinct issues. In short, we posit that any coherent field of research (regardless of whether one wants to consider it a discipline or not) is ultimately defined by a unique combination of (1) a research *object* and (2) a research *objective*.[21] Research *methods* constitute a third aspect, but only play

[16] Ramsay 2013, p. 239f.
[17] Berry 2017.
[18] Kirschenbaum 2014, p. 59.
[19] Kirsch 2014.
[20] Carnap 1950, p. 3.
[21] A more detailed discussion is to be found in Piotrowski 2018.

a secondary role: research methods depend on the research object and the research objective, as well as on technical and scientific progress, which requires them to adapt and, at the same time, permits them to evolve. The research object and the research objective, however, remain relatively stable over time. We would also like to point out that disciplines have never used a single method: they always use a variety of methods. For example, while qualitative methods may certainly be considered »typical« for many humanities disciplines, quantitative methods have always been used as well.[22] This means that it is not useful to attempt to define digital humanities (or any other field or discipline for that matter) by way of the methods it (currently) happens to use, such as: »Digital Humanities is born of the encounter between traditional humanities and computational methods.«[23]

Despite the hype currently surrounding digital humanities, it is neither the humanities' first nor only encounter with computer science. One notable example is computational linguistics. Linguistics is the study of human language; like any other field of research, it studies its research object by creating models of it. Computational linguistics has the same research object and the same research objective as »traditional« linguistics—the essential difference is that it creates *computational* models of language. Computational models have the important advantages that they are 1) formal and 2) executable, and can thus—among other things—be automatically tested against large amounts of actual linguistic utterances.

The construction of computational models of human language poses, however, a number of specific challenges that differ substantially from other modeling tasks in computer science, including related ones, such as the study of formal languages. Computational linguistics thus actually consists of *two* fields: *applied* computational linguistics, which creates formal models of particular languages, and *theoretical* computational linguistics, which serves as a kind of »metascience« for the former, studying the means and methods of constructing computational models in linguistics *in general* and providing the »building materials.« One could thus argue that applied computational linguistics is essentially linguistics, whereas theoretical computational linguistics is essentially computer science: it does not study human language, but rather computational issues in modeling human language.[24]

If we apply these considerations to digital humanities, we can define *digital humanities* in the following precise fashion:

We thus understand construction of formal models as the core of digital humanities, a view we notably share with authors such as McCarty, Meunier, and Thaller.[25] This is not surprising, as historically speaking, »computers came into existence for the sake of modeling.«[26] Consequently, models are the foundation for any serious computational work in digital humanities. To name just a few topics, the production of digital critical editions, visual clustering of artworks, historical network analysis, virtual archaeological reconstruction, or authorship attribution: all of this is only secondarily a question of computing power. Primarily, it is a question of modeling texts, artworks, buildings, relationships, authors, etc., and the findings about them in a way that

[22] In fact, it was the desire to automate quantitative analyses that motivated pioneers such as Roberto Busa or David Packard to use computers for research in the humanities.
[23] Burdick et al. 2012, p. 3.
[24] This brief outline of computational linguistics is, of course, a simplification. Nevertheless, we think it does capture the essence, including the recent trend to almost exclusively use machine learning approaches.
[25] Cf. McCarty 2014, passim; Meunier 2017, passim; Thaller 2017, passim.
[26] Mahoney 2000.

can be meaningfully processed by computers. The models can take many forms, but *some* kind of formal model is the precondition for any type of computational processing; we will say a bit more about models in the following section.

3. Models

The construction of models is an everyday task; we construct models all the time. In his influential 1971 paper »Counterintuitive behavior of social systems,«[27] Forrester points out:

> »Each of us uses models constantly. Every person in private life and in business instinctively uses models for decision making. The mental images in one's head about one's surroundings are models. One's head does not contain real families, businesses, cities, governments, or countries. One uses selected concepts and relationships to represent real systems. A mental image is a model. All decisions are taken on the basis of models. All laws are passed on the basis of models. All executive actions are taken on the basis of models.«

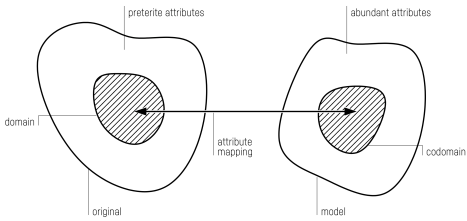


Fig. 3: Original – model mapping. [Piotrowski 2019, redrawn after Stachowiak 1973, p. 157.]

The construction of models in the humanities is thus not per se new: all research, whatever the domain, is based on models. As in the case of *uncertainty*, we have an intuitive understanding of the term *model*, but it is surprisingly hard to come up with a good definition. We use the term in the sense of Stachowiak's *Allgemeine Modelltheorie*.[28] The basic assumption is that arbitrary objects can be described as *individuals* characterized by a finite number of *attributes*.[29] Attributes can be characteristics and properties of individuals, relations between individuals, properties of properties, properties of relations, etc.[30] Modeling is then a mapping of attributes from the original (which can itself be a model) to the model, as illustrated in Figure 3. According to Stachowiak, models are characterized by three fundamental properties:[31]

Mapping property (Abbildungsmerkmal) Models are always models *of something*, namely mappings from, representations of natural or artificial originals, which can themselves be models.

Reduction property (Verkürzungsmerkmal) Models generally *do not capture all* attributes of the original they represent, but only those that the model creators and / or model users deem relevant.

[27] Forrester 1995, p. 4. This is a revised version of Forrester 1971.
[28] Stachowiak 1973.
[29] Stachowiak notes that there is no intrinsic difference between individuals and attributes, individuals are merely introduced for convenience, cf. Stachowiak 1973, p. 134.
[30] Stachowiak 1973, p. 134.
[31] Stachowiak 1973, pp. 131–133.

Pragmatic property (pragmatisches Merkmal) Models are not per se uniquely assigned to their originals. They fulfill their replacement function

1. for particular subjects that use the model,

2. within particular time intervals,

and 3. restricted to particular mental or actual operations.

With respect to the mapping of attributes, three interesting cases shall be briefly mentioned: *preterition*, *abundance*, and *contrasting*. Preterite attributes are attributes that are *not mapped* from the original to the model; abundant attributes are attributes that *do not exist* in the original. Contrasting refers to the exaggeration of certain attributes in the model, typically to highlight certain aspects of the original.

Now, if all disciplines create models, the choice is not »whether to build models; it's whether to build explicit ones.«[32] In contrast to much of the natural and engineering sciences, which tend to use—explicit and formal—mathematical models, models in the humanities are traditionally often only partially explicit and tend to be expressed informally using natural language.[33] The word *formal* means nothing more than »logically coherent + unambiguous + explicit.«[34] While there are different degrees of formalization, it should be clear that in the context of digital humanities we are ultimately interested in a degree of formalization that allows models to be processed and manipulated by computers, i.e., *computational models*. Traditional—informal—models in the humanities do not lend themselves to computational implementation as directly as mathematical models. Furthermore, research questions in the humanities are primarily qualitative rather than quantitative, which, too, has held back the full adoption of the computer as a *modeling tool* rather than just as a writing tool and a »knowledge jukebox.«[35]

4. Modeling uncertainty

If we accept that »being uncertain is the natural state of things,«[36] it follows that we need to consider uncertainty when constructing models. Furthermore, if we define digital humanities as the construction of *formal* models in the humanities, we thus also need to reflect upon how to *formally* (i.e., in a logically coherent, unambiguous, and explicit fashion) represent uncertainty as it occurs in the humanities. The formal representation of uncertainty in digital humanities should have two main objectives:

[32] Epstein 2008.
[33] As the humanities study the human-made world, they are based on a shared understanding of humanity. It is thus only natural that in the communication of models among humans information presumed to be known is not explicated.
[34] Gladkij / Mel'#uk 1969, p. 9.
[35] McCarty 2014, p. 27.
[36] Parsons 2001, p. 9.

1. to make uncertainty explicit,

and 2., to allow reasoning under and about uncertainty.

The following brief example may serve to illustrate these two objectives. Consider a database of historical persons—a common computational model. In some cases, there will be uncertainty with respect to the dates of birth and death of these persons. The uncertainty may take different forms; for example, only one of the dates may be known with some certainty. But it may also happen that both dates are unknown and the existence of a particular person is only inferred from some evidence indicating that they were alive at some point, for example by a document such as a contract.

Suppose the (database) model represented birth and death by a single date each (e.g., of the SQL
DATE
type): dates would all be represented in the same way, whether certain or uncertain, exact or approximate. Even if the procedure for mapping uncertain dates to a date in the model were documented somewhere, the uncertainty would not be represented *formally* and thus be inaccessible to the computer. The computer may thus respond to a query with an exact date such as 1291-08-01, but there would be no information about the certainty (nor the precision) of this date. The computer would also be able to carry out operations such as date arithmetic without any problems—but one date may in fact represent a range and the other just a guess. Taken at face value, this may lead to an »illusion of factuality,« which is obviously problematic, in particular if this information were to be used as a basis for further work, as the uncertainty would propagate.

If, however, the database used unstructured text fields to represent the dates, users could enter something like »between 1291 and 1295,« »late 13th century«, »probably 1291«, etc., or even describe how an approximate date was inferred, and thus preserve the uncertainty. The obvious downside is that such informal representations cannot be processed by the computer: neither could we perform queries or arithmetic on the dates (reasoning under uncertainty), as searching for »1291« will not find »late 13th century« and vice versa, nor could we perform operations on the uncertainty itself, such as a query for all dates of birth that are known to be exact to at least one year (reasoning about uncertainty).

4.1 Uncertainty in computer science

None of this is new. The problem of managing uncertain data has received much attention in computer science, motivated by numerous real-world applications that need to deal with uncertain data. For example, in information retrieval, the actual relevance of a document depends on the information needs of the user, which are only vaguely expressed in the query. Uncertainty also occurs with measurement data of all sorts, environmental sensors, or RFID systems due to technical limitations, noise, or transmission errors. It also occurs with biomedical data, e.g., protein–protein interactions, or when working with anonymized data. Data is obviously uncertain when it has been constructed using statistical forecasting, or when it is based on

spatiotemporal extrapolation, e.g., in mobile applications.[37] The two most widespread approaches for uncertain databases are fuzzy databases and probabilistic databases;[38] uncertain graphs are increasingly used to represent noisy (linked) data in a variety of emerging application scenarios and have recently become a topic of interest in the database and data mining communities.[39]

Classic (i.e., symbolic) artificial intelligence (AI) is a second area in computer science in which extensive research on the representation of uncertainty and on reasoning under and about uncertainty has been carried out. This research was primarily motivated by the need to model uncertainty in knowledge representation (KR), which was in turn driven by the development of *expert systems*, starting in the early 1970s;[40] the objective of these systems was »to capture the knowledge of an expert in a particular problem domain, represent it in a modular, expandable structure, and transfer it to other users in the same problem domain.«[41] To accomplish this goal, research needed to address knowledge acquisition, knowledge representation, inference mechanisms, control strategies, user interfaces, common-sense reasoning—and dealing with uncertainty. A large number of formal methods have been proposed for managing uncertainty in AI systems.[42] Issues of uncertainty are obviously not limited to expert systems, but concern all artificially intelligent agents, from dialog systems to self-driving cars. »Modern« AI research now mostly focuses on deep learning;[43] however, symbolic representations recently gained renewed interest in the context of the Semantic Web,[44] where there is no central authority for resolving contradictions,[45] and issues relating to the trustworthiness of information.[46]

Computational approaches for dealing with uncertainty draw heavily on research in mathematics—in particular probability theory, statistics, and information theory—and logic, in particular epistemic logic; at the same time, many modern formal approaches in mathematics and logic are computational and directly motivated by requirements from KR.[47] Two specific examples are McBurney's dialectical argumentation framework for qualitative representation of epistemic uncertainty in scientific domains, or the fuzzy model for representing uncertain, subjective, and vague temporal knowledge in ontologies by Nagypal.[48] Shannon's information theory[49] measures information by the reduction of uncertainty by receiving a message, and

[37] For overviews, see Motro 1995; Aggarwal / Yu 2009; Aggarwal 2009; Suciu et al. 2011; Khan et al. 2018.
[38] For an overview of fuzzy databases, see Kacprzyk et al. 2015; for an overview of probabilistic databases, see Singh et al. 2008; Suciu et al. 2011.
[39] For an overview, see Khan et al. 2018.
[40] One historically important expert system was MYCIN, cf. Buchanan / Shortliffe 1984.
[41] Bonissone / Tong 1985, p. 241.
[42] Parsons / Hunter 1998 give a concise overview of the most important formalisms for handling uncertainty. See also Ng / Abramson 1990.
[43] Cf. Ghahramani 2015, passim for an overview on probabilistic machine learning for handling uncertainty in this framework.
[44] Cf., e.g., Cai et al. 2012b.
[45] Cf., e.g., Novácek / Smrž 2006.
[46] Cf., e.g., Hartig 2009, Hartig 2017.
[47] For overviews see, e.g., Cai et al. 2012a; Ma et al. 2014; Hunter / Parsons 1998.
[48] McBurney / Parsons 2001; Nagypál / Motik 2003.
[49] Shannon 1948.

uncertainty (i.e., a lack of information) by entropy. Even though very abstract, this theory has had a huge impact on the development of concrete information and communication systems. Information theory is still a very active field of research; of particular interest in this context is work on *uncertain* information,[50] where it is *known* that some piece of information is valid under certain assumptions, but it is *unknown* whether these assumptions actually hold. Also relevant is the extensive work on fuzzy sets and related concepts;[51] Zadeh later outlined a Generalized Theory of Uncertainty[52] aiming to integrate a number of different approaches. Another strand of research is based on the Dempster–Shafer theory (DST),[53] which also continues to produce a number of interesting approaches, such as the theory of hints.[54]

In epistemic logic, there are several approaches that explicitly take uncertainty and source trust into account, e.g., *subjective logic*, a probabilistic logic for uncertain probabilities. Uncertainty is preserved throughout the analysis and is made explicit in the results so that it is possible to distinguish between certain and uncertain conclusions. One interesting recent development is justification logic.[55] Justification logics are epistemic logics that allow knowledge and belief modalities to be made explicit in the form of so-called *justification terms*. There are also a number of extensions to basic justification logic; particularly interesting in this context is Milnikel's logic of uncertain justifications, a variant in which one can formally express statements like »I have degree *r* of confidence that *t* is evidence for the truth of *X*«,[56] and generalizations to multi-agent justification logic, where multiple agents share common knowledge.[57]

4.2 Uncertainty in digital humanities

In the preceding section we have briefly (and eclectically) reviewed some of the research on the representation of uncertainty in computer science; the point of this overview was not to give readers a complete summary of the state of the art but rather to highlight that computer science is far from oblivious to the issue of uncertainty and that its modeling remains an area of active research in computer science as well as in mathematics and logic.

What is the state of the art in digital humanities? Searching the archives of the journal *Digital Scholarship in the Humanities* (DSH), the leading journal in the field (founded in 1986 under the name *Literary and Linguistic Computing*) for the term *uncertainty* yields (at the time of this writing) 87 articles.[58] We obviously do not claim this to be a thorough literature review, but we nevertheless consider this result to reflect, at least to some extent, the state of research concerning uncertainty in DH: on the one hand, there is

[50] Cf., e.g., Kohlas / Eichenberger 2009.
[51] Introduced by Zadeh 1965.
[52] Zadeh 2006.
[53] Dempster 1967; Shafer 1976.
[54] Kohlas / Monney 2008; Pouly et al. 2013.
[55] Artemov 2001.
[56] Milnikel 2014.
[57] Artemov 2006; Bucheli et al. 2012.
[58] The related terms *ambiguity* and *vagueness* yield 169 and 41 results, respectively.

definitely an awareness of the problem, also witnessed, for example, by this special issue and the preceding workshop at the Academy of Sciences and Literature in Mainz that prompted this paper. On the other hand, over a period of 32 years, the number of articles explicitly touching on the topic is not very high—and most only mention it in passing. Just to give an impression, here are some examples of digital humanities publications explicitly dealing with uncertainty (not limited to LLC/DSH):

– In *The Virtual Monastery: Re-Presenting Time, Human Movement, and Uncertainty at Saint-Jean-des-Vignes, Soissons*, Bonde et al. discuss the representation of uncertainty in visual archaeological reconstructions.[59]

– In *Artefacts and Errors: Acknowledging Issues of Representation in the Digital Imaging of Ancient Texts*, Terras studies the additional uncertainty introduced in paleography and epigraphy by digitization. She concludes that uncertainty is »an important issue to address when building computational systems to aid papyrologists. Unfortunately, encapsulating uncertainty in computational systems is not a straightforward process.«[60]

– The paper *Digitizing the act of papyrological interpretation: negotiating spurious exactitude and genuine uncertainty* by Tarte discusses similar issues. The author stresses that an important aspect of digital papyrology, beyond the mere digitization of the artifact, is »to enable the digital capture of the thought process that builds interpretations of ancient and damaged texts«, i.e., to document the choices made by the editor in the face of uncertainty.[61]

– *Known Unknowns: Representing Uncertainty in Historical Time* and *On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges* are examples of papers discussing the visualization of temporal and spatiotemporal uncertainty in timelines and maps.[62]

– *Uncertain about Uncertainty: Different Ways of Processing Fuzziness in Digital Humanities Data* focuses on practical aspects, such as how to »ensure that such information [on persons and places] is added in a coherent way, while allowing the data to be vague or apparently contradictory.«[63]

Our general impression is that most papers discuss uncertainty as it occurs in the context of a particular research project, and in most cases aim to also solve the issues in this context. This is not surprising, as currently most publications in digital humanities belong to the applied digital humanities. One notable

[59] Bonde et al. 2009, passim.
[60] Terras 2010, p. 50.
[61] Tarte 2011, p. 352.
[62] Kräutli / Boyd Davis 2013, passim; Jänicke et al. 2015, passim.
[63] Binder et al. 2014, p. 96.

exception is the paper by Tarte already mentioned above, as it also discusses the modeling of uncertainty in more general terms. The author notes that »[c]apturing uncertainty is vital to the recording process«,[64] and explicitly bases her approach on theoretical frameworks, namely argumentation theory and theory of justification to »provide a formal, yet invisible, epistemological framework that allows us to point out inconsistencies without forbidding them.«[65]

The state of the art in digital humanities is perhaps best illustrated by the Guidelines of the Text Encoding Iniative (TEI), the de facto standard for digital critical editions. There are some provisions for modeling uncertainty in Chapter 21, *Certainty, Precision, and Responsibility*, which define »several methods of recording uncertainty about the text or its markup«.[66] For example, editors can use the elements and attributes provided to indicate that some aspects of the encoded text are problematic or uncertain, and to record who is responsible for making certain choices. The
certainty
element defined in this chapter may be used to record the nature and degree of uncertainty in a structured way, which allows encoders to express quite complex phenomena. To take an example from the chapter, one could express that in the passage »Elizabeth went to Essex; she had always liked Essex,« one thinks that there is a 60 percent chance that »Essex« refers to the county, and a 40 percent chance that it refers to the earl, and furthermore that the occurrences of the word are not independent: if the first occurrence refers to the county, one may decide that it is excluded that the second refers to the earl. However, we are not aware of any TEI application using these facilities, for which there are probably two main reasons: the complexity of the markup and, perhaps even more importantly, the fact that the TEI Guidelines leave it open how one may determine the »probability,« »chance,« or »certainty« of an interpretation. As Tarte points out, »quantifying uncertainty is always risky and usually presupposes that problems are complete, i.e. that all the alternatives to a given situation are known […], which is far from being the case in a papyrological context«,[67] and, one may add, neither in most other contexts in the humanities. Binder et al. also mention the additional challenge of communicating the uncertainty recorded in this way to human users.[68]

We are only aware of one framework in the wider field of digital humanities aiming for a more general modeling of uncertainty, the *Integrated Argumentation Model* (IAM) by Doerr, Kritsotak, and Boutiska.[69] The original motivation for IAM comes from the domains of archaeology and cultural heritage; consequently, IAM is intended as an addition to the CIDOC Conceptual Reference Model.[70] Instead of trying to quantify »certainty« (or quietly assuming that it can be quantified), the IAM approaches the problem as an argumentation process, during which hypotheses may be strengthened or weakened, similar to the approach used by Tarte.[71]

[64] Tarte 2011, p. 357.
[65] Tarte 2011, p. 355; for an overview of argumentation theory, cf. Parsons 2001, pp. 150–155; Toulmin 2003; for theory of justification, Tarte specifically refers to the approach of Haack 2001.
[66] TEI Guidelines, Chapter 21.
[67] Tarte 2011, p. 355.
[68] Binder et al. 2014, p. 96.
[69] Doerr et al. 2011, passim.
[70] Crofts et al. 2011.
[71] Cf. Tarte 2011, passim.

5. Conclusion

In this article we have tried to show the challenge of uncertainty for the construction of computational models in the humanities—digital humanities. It is clear that we cannot ignore uncertainty, and we cannot eliminate it either: we need to model it, and we thus need computational models of uncertainty.

Computational models of uncertainty already exist in various research disciplines and are being used in commercial and industrial applications. There is ongoing fundamental research on uncertainty and its representation in mathematics, philosophy, and computer science. Some of these approaches may also be suitable frameworks for computational modeling of uncertainty in the humanities—what is lacking, however, is the »bridge« that could relate the uncertainty encountered in humanities research to these formal modeling frameworks. What is missing in particular, is a *systematic account of uncertainty in humanities research*, which would aim to document causes for uncertainty, as well as its behavior, i.e., questions such as: Could this type of uncertainty be (in principle) resolved? What information would be required to resolve it? What happens when new information becomes available? How is it taken into account? And so on.

We would like to stress that the goal is not to come up with an answer to the ontological question of the »true nature« of uncertainty, but rather to find ways of modeling this omnipresent but still elusive phenomenon. We would also like to stress that it is unlikely that anyone will ever come up with a »grand theory of uncertainty« allowing for a single universal model—we recall Stachowiak's pragmatic property of models. What may very well be possible, though, are more general modeling frameworks that provide researchers with the building materials required for constructing models of particular types of uncertainty in a particular domain.

As we have noted above, there are numerous approaches for the formal modeling of uncertainty, which differ in many respects. Smets notes: »Newcomers to the area of uncertainty modeling are often overwhelmed by the multitude of models. One frequent reaction is to adopt one of the models, and use it in every context. Another reaction is to accept all the models and to apply them more or less randomly. Both attitudes are inappropriate.«[72] We are not sure whether the (digital) humanities have even reached this point; in any case, there is no catalog of uncertainty modeling frameworks from which humanities scholars could pick a framework according to a set of criteria. Neither are there best practices in the humanities that could guide the selection; whether an approach that was successfully used in one digital humanities project can be transferred to another is not much easier to answer than the question of whether the methods used for dealing with uncertain environmental measurement data could be adapted to the domain of medieval manuscripts.

Intuitively, the kinds of uncertainty encountered in the humanities tend to differ in some respects from many more »traditional« applications, such as in engineering, meteorology, or demography; for example,

- humanities data is often more like expert knowledge than measurements;
- the amount of available data tends to be relatively small and often concerns singular, non-repeatable events;
- the reasoning is rather evidential than predictive;

[72] Smets 1997, p. 226.

– intercausal, counterfactual, and deductive reasoning may be of more importance;

– phenomena similar to those known as *selectively reported data* and *missing data* in statistics may be more frequent;[73]

– uncertainty is rather qualitative than quantitative, or at least hard to quantify;

– belief is likely to play an important role with respect to conflicts of interpretation.

This is obviously just a very superficial assessment. All of these phenomena may occur in other fields as well —knowledge representation in classical AI comes to mind—, and there are certainly cases in the humanities that do not exhibit them. What we are trying to say is that uncertainty in humanities research likely has »typical« characteristics, just like uncertainty in, say, gambling and structural engineering is caused and influenced by different factors and thus exhibits particular characteristics; the two fields are also driven by different concerns, so some modeling approaches will be more pertinent than others.[74]

As Bradley notes, »[w]hen dealing with uncertainty, it often seems like the right approach is to try to quantify that uncertainty,« but, he stresses, »there are limits to what we can (and should) quantify.«[75] This certainly applies to many types of uncertainty in the humanities; mathematical and logical approaches to uncertainty are important as potential modeling and reasoning frameworks to target, but there is still the open problem of how (if at all) to quantify this uncertainty—which is not incidentally the main issue in the approach suggested by the TEI Guidelines (see above).

From this we conclude that we need theoretical digital humanities and theory formation to study questions such as *what types of uncertainty do we find in the humanities? What are the specifics? What can be generalized?* Building on our insights we can then evaluate approaches from mathematics, logic, and computer science and develop methods and recommendations for applied digital humanities. In other words: It is the task of the theoretical digital humanities to develop, in close dialog with the humanities disciplines, modeling frameworks and methods that can specifically address these challenges. Theoretical digital humanities is crucial for laying the theoretical groundwork for the development of models and

methods *appropriate* for the humanities—instead of using »second-hand« methods originally developed for completely different purposes. This is, we believe, what Meister refers to when he characterizes digital humanities as »a methodology that cuts across disciplines, systematically as well as conceptually«.[76]

And this way, digital humanities can in fact play an important role for the transformation of the humanities within the larger digital transformation of society as a whole, and mean more than just »contemporary humanities«.

[73] Dawid / Dickey 1977; Rubin 1976.
[74] For an interesting overview of uncertainty in structural engineering, see Bulleit 2008.
[75] Bradley 2018, p. 31.
[76] Meister 2012, p. 84.

Bibliographic References

Charu Chandra Aggarwal / Philip S. Yu: A survey of uncertain data algorithms and applications. In: IEEE Transactions on Knowledge and Data Engineering 21 (2009), no. 5, pp. 609–623. [Nachweis im GBV]

Managing and mining uncertain data. Ed. by Charu Chandra Aggarwal. Boston, MA 2009. [Nachweis im GBV]

Sergei Artemov: Explicit provability and constructive semantics. In: Bulletin of Symbolic Logic 7 (2001), no. 1, pp. 1–36. [Nachweis im GBV]

Sergei Artemov: Justified common knowledge. In: Theoretical Computer Science 357 (2006), no. 1–3, pp. 4–22. DOI: 10.1016/j.tcs.2006.03.009 [Nachweis im GBV]

David M. Berry. twitter.com. @berrydm. June 2017. https://twitter.com/berrydm/status/877532752735764480. Tweet no longer available.

Frank Binder / Bastian Entrup / Ines Schiller / Henning Lobin: Uncertain about uncertainty. Different ways of processing fuzziness in digital humanities data. In: Proceedings of Digital Humanities 2014. (DH2014, Lausanne, 07.–11.07.2014) Lausanne 2014, pp. 95–98. [online]

Sheila Bonde / Clark Maines / Elli Mylonas / Julia Flanders: The virtual monastery. Re-presenting time, human movement, and uncertainty at Saint-Jean-des-Vignes, Soissons. In: Visual Resources 25 (2009), no. 4, pp. 363–377. [Nachweis im GBV]

Piero P. Bonissone / Richard M. Tong: Editorial: Reasoning with uncertainty in expert systems. In: International Journal of Man-Machine Studies 22 (1985), no. 3, pp. 241–250. [Nachweis im GBV]

Seamus Bradley: Uncertain reasoning. In: The Reasoner 12 (2018), no. 4, pp. 31–32. PDF. [online]

Rule-based expert systems: The MYCIN experiments of the Stanford Heuristic Programming Project: Ed. by Bruce G. Buchanan / Edward Hance Shortliffe. Reading, MA 1984. [online] [Nachweis im GBV]

Samuel Bucheli / Roman Kuznets / Thomas Studer: Justifications for common knowledge. In: Journal of Applied Non-Classical Logics 21 (2012), no. 1, pp. 35–60. [Nachweis im GBV]

William M. Bulleit: Uncertainty in structural engineering. In: Practice Periodical on Structural Design and Construction 13 (2008), no. 1, pp. 24–30. [Nachweis im GBV]

Anne Burdick / Johanna Drucker / Peter Lunenfeld / Todd Presner / Jeffrey Schnapp: Digital Humanities. Cambridge, MA 2012. [Nachweis im GBV]

Yi Cai / Ching-man Au Yeung / Ho-fung Leung (2012a): Fuzzy computational ontologies in contexts. Berlin et al. 2012. [Nachweis im GBV]

Yi Cai / Ching-man Au Yeung / Ho-fung Leung (2012b): Modeling uncertainty in knowledge representation. In: Fuzzy computational ontologies in contexts. Berlin et al. 2012, pp. 37–47. [Nachweis im GBV]

Rudolf Carnap: Logical foundations of probability. Chicago, IL 1950. [Nachweis im GBV]

A. P. Dawid / James M. Dickey: Likelihood and Bayesian inference from selectively reported data. In: Journal of the American Statistical Association 72 (1977), no. 360a, pp. 845–850. [Nachweis im GBV]

Definition of the CIDOC Conceptual Reference Model. Ed. by Nick Crofts / Martin Doerr / Tony Gill / Stephen Stead / Matthew Stiff. Paris 2011. PDF. [online]

Arthur P. Dempster: Upper and lower probabilities induced by a multivalued mapping. In: The Annals of Mathematical Statistics 38 (1967), no. 2, pp. 325–339. DOI: 10.1214/aoms/1177698950 [Nachweis im GBV]

Martin Doerr / Athina Kritsotaki / Katerina Boutsika: Factual argumentation - a core model for assertions making. In: Journal on Computing and Cultural Heritage 3 (2011), no. 3, pp. 8:1–8:34. [Nachweis im GBV]

Joshua M. Epstein: Why model? In: Journal of Artificial Societies and Social Simulation 11 (2008), no. 4. [online] [Nachweis im GBV]

Jay Wright Forrester: Counterintuitive behavior of social systems. 1995. PDF. [online]

Jay Wright Forrester: Counterintuitive behavior of social systems. In: MIT Technology Review 73 (1971), no. 3, pp. 52–68. [Nachweis im GBV]

Zoubin Ghahramani: Probabilistic machine learning and artificial intelligence. In: Nature 521 (2015), no. 7553, pp. 452–459. [Nachweis im GBV]

Aleksej Vsevolodovi# Gladkij / Igor Aleksandrovi# Mel'#uk: Elementy matemati#eskoj lingvistiki. Moskva 1969. [Nachweis im GBV]

Susan Haack: Evidence and inquiry towards reconstruction in epistemology. Reprinted. London 2001. [Nachweis im GBV]

Olaf Hartig: Foundations of RDF* and SPARQL*: An alternative approach to statement-level metadata in RDF. In: Proceedings of the 11th Alberto Mendelzon International Workshop on Foundations of Data Management and the Web (AMW 2017). Ed. by Juan Reutter / Divesh Srivastava. Montevideo 2017. PDF. [online]

Olaf Hartig: Querying trust in RDF data with tSPARQL. In: The Semantic Web: Research and applications. Proceedings of the 6th European Semantic Web Conference. Ed. by Lora Aroyo / Paolo Traverso / Fabio Ciravegna / Philipp Cimiano / Tom Heath / Eero Hyvönen / Riichiro Mizoguchi / Eyal Oren / Marta Sabou / Elena Simperl. (ESWC: 6, Heraklion, 31.05.-04.06.2009) Berlin et al. 2009, pp. 5–20. DOI: 10.1007/978-3-642-02121-3_5 [Nachweis im GBV]

Applications of uncertainty formalisms. Ed. by Anthony Hunter / Simon Parsons. Berlin et al. 1998. [Nachweis im GBV]

Stefan Jänicke / Greta Franzini / Muhammad F. Cheema / Gerik Scheuermann: On close and distant reading in digital humanities. A survey and future challenges. In: Eurographics Conference on Visualization (EuroVis) – STARs. Ed. by Rita Borgo / Fabio Ganovelli / Ivan Viola. (EuroVis: 17, Cagliari, 25.–29.05.2015) Geneve 2015. PDF. [online] [Nachweis im GBV]

Janusz Kacprzyk / S#awomir Zadro#ny / Guy De Tré: Fuzziness in database management systems. Half a century of developments and future prospects. In: Fuzzy Sets and Systems 281 (2015), pp. 300–307. [Nachweis im GBV]

Arijit Khan / Yuan Ye / Lei Chen: On uncertain graphs. In: Synthesis Lectures on Data Management 10 (2018), no. 1, pp. 1–94. DOI: 10.2200/s00862ed1v01y201807dtm048

Adam Kirsch: Technology is taking over English departments. In: New Republic. Article from 02.05.2014. [online]

Matthew G. Kirschenbaum: What is Digital Humanities, and why are they saying such terrible things about it? In: Differences 25 (2014), no. 1, pp. 46–63. [Nachweis im GBV]

Matthew G. Kirschenbaum: What is digital humanities and what's it doing in English departments? In: Debates in the digital humanities. Ed. by Matthew K. Gold. Minneapolis, MN 2012, pp. 3–11. [online] [Nachweis im GBV]

Jürg Kohlas / Paul-André Monney: An algebraic theory for statistical information based on the theory of hints. In: International Journal of Approximate Reasoning 48 (2008), no. 2, pp. 378–398. DOI: 10.1016/j.ijar.2007.05.003 [Nachweis im GBV]

Jürg Kohlas / Christian Eichenberger: Uncertain information. In: Formal theories of information. Lecture notes in computer science. Ed. by Giovanni Sommaruga. Berlin et al. 2009, pp. 128–160. [Nachweis im GBV]

Florian Kräutli / Stephen Boyd Davis: Known unknowns: Representing uncertainty in historical time. In: Electronic visualisation and the arts. Ed. by Kia Ng et al. (EVA 2013, London, 29.–31.07.2013) Swindon et al. 2013, pp. 61–68. [online] [Nachweis im GBV]

Zongmin Ma / Fu Zhang / Li Yan / Jingwei Cheng: Fuzzy knowledge management for the Semantic Web. Heidelberg et al. 2014. [Nachweis im GBV]

Michael S. Mahoney: Historical perspectives on models and modeling. In: Scientific Models: Their Historical and Philosophical Relevance. (DHS-DLMPS: 13, Zürich, 19.–22.10.2000) Zürich 2000. [online]

Peter McBurney / Simon Parsons: Representing epistemic uncertainty by means of dialectical argumentation. In: Annals of Mathematics and Artificial Intelligence 32 (2001), no. 1–4, pp. 125–169. [Nachweis im GBV]

Willard McCarty: Humanities computing. Paperback. Basingstoke 2014. [Nachweis im GBV]

Jean-Guy Meunier: Humanités numériques et modélisation scientifique. In: Questions de communication 31 (2017), no. 1, pp. 19–48. [Nachweis im GBV]

Jan Christoph Meister: DH is us or on the unbearable lightness of a shared methodology. In: Historical Social Research, 37 (2012), no. 3, pp. 77–85. [online]

Robert S. Milnikel: The logic of uncertain justifications. In: Annals of Pure and Applied Logic 165 (2014), no. 1, pp. 305–315. DOI: 10.1016/j.apal.2013.07.015 [Nachweis im GBV]

Amihai Motro: Management of uncertainty in database systems. In: Modern database systems. The object model, interoperability, and beyond. Ed. by Won Kim. New York, NY 1995, pp. 457–476. [Nachweis im GBV]

Gábor Nagypál / Boris Motik: A fuzzy model for representing uncertain, subjective, and vague temporal knowledge in ontologies. In: On the move to meaningful Internet systems 2003: CoopIS, DOA, and ODBASE. Lecture notes in computer science. Ed. by Robert Meersman / Zahir Tari / Douglas C. Schmidt. (Conference, Catania, 03.–07.11.2003) Berlin et al. 2003, pp. 906–923. [Nachweis im GBV]

Keung-Chi Ng / Bruce Abramson: Uncertainty management in expert systems. In: IEEE Expert, 5 (1990), no. 2, pp. 29–48. DOI: 10.1109/64.53180.

Vít Nová#ek / Pavel Smrž: Empirical merging of ontologies—a proposal of universal uncertainty representation framework. In: The Semantic Web: Research and applications. Ed. by York Sure / John Domingue. (ESWC: 3, Budva, 11.–14.06.2006) Berlin et al. 2006, pp. 65–79. DOI: 10.1007/11762256_8 [Nachweis im GBV]

Simon Parsons / Anthony Hunter: A review of uncertainty handling formalisms. In: Applications of uncertainty formalisms. Ed. by Anthony Hunter / Simon Parsons. Berlin et al. 1998, pp. 8–37. [Nachweis im GBV]

Simon Parsons: Qualitative approaches for reasoning under uncertainty. Cambridge, MA 2001. [Nachweis im GBV]

Michael Piotrowski: Digital humanities: An explication. In: Workshop der GI Fachgruppe „Informatik und Digital Humanities": Im Spannungsfeld zwischen Tool-Building und Forschung auf Augenhöhe – Informatik und die Digital Humanities. Workshop Proceedings. Ed. by Manuel Burghardt / Claudia Müller-Birn. Gesellschaft für Informatik e.v. (INF-DH, Berlin, 25.09.2018) Bonn 2018. DOI: 10.18420/infdh2018-07

Henry N. Pollack: Uncertain science … uncertain world. Cambridge 2005. [Nachweis im GBV]

Marc Pouly / Jürg Kohlas / Peter Y. A. Ryan: Generalized information theory for hints. In: International Journal of Approximate Reasoning 54 (2013), no. 1, pp. 228–251. DOI: 10.1016/j.ijar.2012.08.004 [Nachweis im GBV]

Stephen Ramsay: Who's in and who's out. In: Defining digital humanities. Ed. by Melissa Terras / Julianne Nyhan / Edward Vanhoutte. Farnham et al. 2013, pp. 239–241. [Nachweis im GBV]

Paul Ricœur: De l'interprétation: Essai sur Freud. Paris 1965. [Nachweis im GBV]

Donald B. Rubin: Inference and missing data. In: Biometrika 63 (1976), no. 3, pp. 581–592. DOI: 10.1093/biomet/63.3.581 [Nachweis im GBV]

Glenn Shafer: A mathematical theory of evidence. Princeton, NJ 1976. [Nachweis im GBV]

Claude E. Shannon: A mathematical theory of communications. In: The Bell System Technical Journal 27 (1948), pp. 379–432. [Nachweis im GBV]

Sarvjeet Singh / Chris Mayfield / Rahul Shah / Sunil Prabhakar / Susanne Hambrusch / Jennifer Neville / Reynold Cheng: Database support for probabilistic attributes and tuples. In: 2008 IEEE 24th International Conference on Data Engineering. 3 Vol. (ICDE: 24, Cancun, 07.–12.04.2008) Piscataway, NJ 2008. Vol. 2, pp. 1053–1061. [Nachweis im GBV]

Michael Smithson: Ignorance and uncertainty. New York, NY 1989. [Nachweis im GBV]

Philippe Smets: Imperfect information: Imprecision and uncertainty. In: Uncertainty management in information systems. Ed. by Amihai Motro / Philippe Smets. Boston, MA et al. 1997, pp. 225–254. [Nachweis im GBV]

Herbert Stachowiak: Allgemeine Modelltheorie. Wien et al. 1973. [Nachweis im GBV]

Dan Suciu / Dan Olteanu / Christopher Ré / Christoph Koch: Probabilistic databases. San Rafael, CA 2011. DOI: 10.2200/s00362ed1v01y201105dtm016 [Nachweis im GBV]

Ségolène M. Tarte: Digitizing the act of papyrological interpretation: Negotiating spurious exactitude and genuine uncertainty. In: Literary and Linguistic Computing 26 (2011), no. 3, pp. 349–358. DOI: 10.1093/llc/fqr015

Melissa Terras: Artefacts and errors: Acknowledging issues of representation in the digital imaging of ancient texts. In: Kodikologie und Paläographie im digitalen Zeitalter 2 / Codicology and palaeography in the digital age. Ed. by Franz Fischer / Christiane Fritze / Georg Vogeler. 4 Vol. Norderstedt 2010. Vol. 2, pp. 43–61. URN: urn:nbn:de:hbz:38-43429 [Nachweis im GBV]

Defining digital humanities. Ed. by Melissa Terras / Julianne Nyhan / Edward Vanhoutte. Farnham 2013. [Nachweis im GBV]

Manfred Thaller: Between the chairs: An interdisciplinary career. In: Historical Social Research Supplement 29 (2017), pp. 7–109. DOI: 10.12759/hsr.suppl.29.2017.7-109 [Nachweis im GBV]

Stephen E. Toulmin: The uses of argument. Updated edition. Cambridge et al. 2003. [Nachweis im GBV]

Lotfi A. Zadeh: Fuzzy sets. In: Information and Control 8 (1965), no. 3, pp. 338–353. DOI: 10.1016/s0019-9958(65)90241-x [Nachweis im GBV]

Lotfi A. Zadeh: Generalized theory of uncertainty (GTU) - principal concepts and ideas. In: Computational Statistics & Data Analysis 51 (2006), no. 1, pp. 15–46. DOI: 10.1016/j.csda.2006.04.029 [Nachweis im GBV]

List of Figures with Captions

ZfdG

Zeitschrift für digitale Geisteswissenschaften

Beitrag aus:
Sonderband 4 der ZfdG: Die Modellierung des Zweifels – Schlüsselideen und -konzepte zur graphbasierten Modellierung von Unsicherheiten. Hg. von Andreas Kuczera, Thorsten Wübbena und Thomas Kollatz. 2019. DOI: 10.17175/sb004

Titel:
Ambiguität und Unsicherheit: Drei Ebenen eines Datenmodells

Autor/in:
Andreas Wagner

Kontakt:
andreas.wagner@adwmainz.de

Institution:
Akademie der Wissenschaften und der Literatur, Mainz

GND:
136680607

ORCID:
0000-0003-1835-1653

DOI des Artikels:
10.17175/sb004_007

Nachweis im OPAC der Herzog August Bibliothek:
1037073762

Erstveröffentlichung:
30.01.2019

Lizenz:

Sofern nicht anders angegeben (cc) BY-SA

Medienlizenzen:
Medienrechte liegen bei den Autoren

Letzte Überprüfung aller Verweise:
30.01.2019

GND-Verschlagwortung:
Digital Humanities | Datenmodell | Graphdatenbank | Kirchenrecht 1567–1681 | Referenzmodell | Ungewissheit |

Zitierweise:
Andreas Wagner: Ambiguität und Unsicherheit: Drei Ebenen eines Datenmodells. In: Die Modellierung des Zweifels – Schlüsselideen und -konzepte zur graphbasierten Modellierung von Unsicherheiten. Hg. von Andreas Kuczera, Thorsten Wübbena und Thomas Kollatz. 2019 (= Sonderband der Zeitschrift für digitale Geisteswissenschaften, 4). PDF Format ohne Paginierung. Als text/html abrufbar unter DOI: 10.17175/sb004_007.

Andreas Wagner

Ambiguität und Unsicherheit: Drei Ebenen eines Datenmodells

Abstracts

Dieser Beitrag stellt anhand von Forschungen im Rahmen der Max-Planck-Nachwuchsgruppe ›Die Regierung der Universalkirche nach dem Konzil von Trient‹ einen Ansatz vor, Phänomene von Unsicherheit in geisteswissenschaftlichen, genauer: rechtshistorischen Zusammenhängen in einem Datenmodell abzubilden. Die Kernthese besteht im Vorschlag einer Modellierung des Forschungszusammenhangs auf drei Ebenen, welche (a) die historischen Phänomene, (b) die überkommenen Zeugnisse dieser Phänomene und (c) die aktuelle historische Forschung selbst beschreiben. Während Ambiguitäten und Unklarheiten zwar auf allen dreien dieser Ebenen entstehen, werden sie allein auf der dritten Ebene modelliert, denn sie können nicht von der Beobachtung in der Forschung getrennt werden.[1] Im Folgenden wird das Forschungsprojekt ›Die Regierung der Universalkirche nach dem Konzil von Trient‹ im Hinblick insbesondere auf Unsicherheiten in der Datenmodellierung vorgestellt (1.). Im Anschluss daran wird ein vorläufiges, bisweilen noch informelles, Datenmodell beschrieben (2.) und schließlich ein Ausblick vorgenommen, der die weitere Entwicklung des Modells ebenso wie die Herausforderungen im Projektkontext, v.a. in der Dateneingabe und -prozessierung betrachtet (3.).

Using a specific research project, this article presents an approach for displaying phenomena of uncertainty in humanities contexts – here, legal history – in a data model. In the proposal for modeling a research context, the core thesis depends on three levels that describe a) the historical phenomena, b) the traditional evidence for this phenomena, and c) the latest historical research. Although ambiguities and uncertainties will be present on all three levels, they will be modeled on the third level only since they can not be separated from the historical research. The following paper presents the research project ›Die Regierung der Universalkirche nach dem Konzil von Trient‹ with regard to uncertainties in data modeling (1.). The preliminary, at times unofficial, data model will then be described (2.); finally, this article will discuss future prospects for the further development of the model as well as challenges within the context of the project, such as in data entry and processing (3.).

1. Der Projektzusammenhang und unsichere Informationen

1.1 Ein rechtshistorisches Forschungsprojekt

Das Konzil von Trient (1545–1563) gilt als wichtiger Meilenstein in der doktrinären und institutionellen Modernisierung der katholischen Kirche. Mit den Konzilsbeschlüssen reagierte diese auf theologische, juristische und moralische Herausforderungen, die sich unter anderem im Kontext der Reformation, durch ihre eigene Rolle in einer nun globalisierten politischen Landschaft und durch eine virulent gewordene Pluralität und Dynamik sozialer, kultureller

[1] Um genau zu sein, werden auch Informationen, die im Forschungsprozess den ersten beiden Ebenen zugeschrieben werden, zwar eben dort festgehalten, aber mit einer Information auf der dritten Ebene verknüpft, die eben die Unsicherheit der Zuschreibung festhält. Eine unleserliche Abkürzung ist in diesem Sinne zwar ein Phänomen, das in der Quelle selbst verortet ist, aber *unleserlich* ist sie eben immer nur *für eine Leserin* oder *einen Leser*.

und politischer Verhältnisse ergeben hatten. Dies erforderte aber auch handlungsfähige und zugleich flexible, institutionelle Mechanismen, durch die diese Beschlüsse auf die vielfältigen Handlungskontexte und die jeweils anfallenden Probleme anzuwenden oder mit ihnen in Dialog zu bringen wären. Mit dem Auftrag einer autoritativen Interpretation und je zu aktualisierenden Umsetzung wurde 1567 die Konzilskongregation eingerichtet, deren Beratungsprotokolle bis heute im Vatikanischen Geheimarchiv lagern. Dieses Gremium nahm Petitionen und Fragen zur Auslegung der Konzilsbeschlüsse entgegen, die aus kirchlichen Einrichtungen und von Funktionsträgern aus der ganzen Welt nach Rom gesandt wurden. Die Fragen wurden in ggf. mehreren Sitzungsterminen beraten, weitere Informationen wurden eingeholt bzw. Stellungnahmen Dritter wurden eingereicht, bevor eine Entscheidung getroffen und in den Beschluss-Protokollen festgehalten wurde.[2]

Um die Weise zu untersuchen, in der das Recht als Instrument des beständigen Abgleichs zwischen der universalen Institution der Kirche und den jeweiligen lokalen, konkreten Regulierungsbedarfen eingesetzt wurde oder seine Wirkung entfaltete, wurde 2013 unter der Leitung von Benedetta Albani die Max-Planck-Forschungsgruppe ›Die Regierung der Universalkirche nach dem Konzil von Trient‹ am Max-Planck-Institut für europäische Rechtsgeschichte in Frankfurt/Main eingerichtet. Durch verschiedene Einzelfragen erschließen und beforschen die Mitglieder der Gruppe die Akten der Konzilskongregation — eine Datenmenge, die kaum ohne die Hilfe digitaler Methoden intellektuell durchdrungen werden kann:

- Die Forschung betrifft einen Zeitraum von 114 Jahren (1567–1681) und 17 Pontifikaten,
- an den Beratungen waren über 100 Kardinäle beteiligt,
- die Eingaben stammen aus ca. 750 weltweit verteilten Diözesen.
- Die Beratungsprotokolle der Konzilskongregation umfassen 270 Bände Archivmaterialien,
- und sie geben über 33.000 Entscheidungen wieder.

Jede einzelne dieser Entscheidungen wird durch Datenfelder wie z.B. das Datum der Sitzung, den Petenten, den infrage stehenden Konzilsbeschluss, die beteiligten Kardinäle und Sekretäre und ggf. den Beschluss der Kongregation beschrieben –nach Erfahrungen der Projektgruppe, die in der einen oder anderen Form bereits seit fünf Jahren an diesem Datenbestand arbeitet, wurden über 100 mögliche Datenfelder identifiziert.[3]

1.2 Unsicherheiten, Unklarheiten und Ambiguitäten

Die erste Quelle von Zweifeln bezüglich der Informationen, mit denen im Projekt gearbeitet wird, liegt auf der Hand: Denn wie es in solchen Projekten zumeist der Fall ist, ist nicht für jede Entscheidung jede Information in den Protokollen festgehalten. Die formalen Anforderungen

[2] Vgl. Albani 2009.
[3] Diese Zahl beinhaltet allerdings auch in gewisser Weise redundante Felder wie z.B. durch die Projektgruppe zu recherchierende Normdaten-Identifikatoren für die relevanten Personen, eine Standard-Ansetzungsform und die jeweilige Schreibung bzw. Abkürzung ihres Namens im erfassten Protokoll.

sorgen zwar dafür, dass z.B. das Datum der Sitzung in allen Fällen aufgeschrieben wurde, so dass dieses Feld – von Beschädigungen der Archivalien oder unklaren Formulierungen abgesehen – immer befüllt werden kann. Der einschlägige Konzilsbeschluss oder gar formelle wie informelle Einflussnahmen Dritter sind jedoch nicht in allen Fällen dokumentiert und wir haben einen unvollständigen und insofern inkonsistenten Datenbestand. Weitere Probleme ergeben sich durch die Manuskript-Charakteristik der Materials: So liegen häufig unleserliche Schreibungen und – vor allem bei Personennamen – Abkürzungen vor, die z.T. mehrdeutig sind. In anderen Fällen werden Personen eher durch ihre Rolle oder Funktion bezeichnet. Spätestens in diesen Fällen kann eine Differenzierung vorgenommen werden zwischen lückenhafter oder unleserlicher Vorlage auf der einen Seite und unsicherem, bloß wahrscheinlichen Wissen der heutigen Leserinnen auf der anderen. Denn in einem Ausdruck wie »der Domkapitular von Assisi« kann einerseits z.B. der Ortsname unleserlich sein, wodurch nicht klar ist, von welcher Person die Rede ist, es kann aber andererseits auch der Ausdruck ohne weiteres lesbar, jedoch unser Wissen darüber, wer zur betreffenden Zeit diese Rolle ausfüllte, unsicher sein.

Es ist aber noch auf eine dritte Quelle von Unsicherheiten hinzuweisen, die gar nicht direkt mit dem archivalischen Quellenmaterial oder den aktuellen Forschungsprozessen zusammen hängt: Denn schon diesseits ihrer Dokumentation sind in den historischen Phänomenen selbst diffuse Phänomene wichtiger Teil des Praxiszusammenhangs. Schon für einen Teilnehmer der Kongregationsberatungen selbst wäre in vielen Fällen nicht zweifelsfrei entscheidbar, ob z.B. eine dritte Partei Einfluss ausgeübt hat. Und dies nicht erst wegen des unvollkommenen Wissens dieses vorgestellten Teilnehmers, sondern schon wegen der Vagheit der Kategorie der ›Einflussnahme‹. Gewiss ist das Projekt darauf angelegt, formelle und insofern objektivierte und eindeutige Eigenschaften und Prozesse zu erfassen. Allerdings ist erstens nicht *a priori* überschaubar, ob es sich bei einzelnen zu erfassenden Datenfeldern nicht doch um Informationen handeln könnte, die erst *ex post* vereindeutigt wurden (und sei es durch die Selbst-Interpretation der Teilnehmer) und die in den Dokumenten lediglich präzise scheinen, in der Praxis aber diffuser waren. Und zweitens kann es sich im Zuge der Überlegungen zu einer möglichen Verallgemeinerbarkeit unseres Datenmodells durchaus als produktiv erweisen, über diese Fragen nachzudenken.

Diffuse historische Phänomene, lückenhaftes oder unleserliches Quellenmaterial und unsicheres Kontextwissen der heutigen Forscherinnen sind somit drei Ebenen, auf denen mitunter Zweifel und selbstkritische Vorbehalte bezüglich der Informationen, mit denen im Projekt gearbeitet wird, angebracht sind. Offenkundig würde der wissenschaftliche Prozess ganz wesentlich verarmt, wenn man Informationen, die in einer der genannten Hinsichten fraglich wären, überhaupt nicht mehr berücksichtigen würde. Somit muss es im Rahmen guter wissenschaftlicher Praxis auch in einer solchen Datenbank darum gehen, die entsprechenden Unsicherheiten möglichst transparent und nachvollziehbar zu transportieren und zu kommunizieren, d.h. sie müssen flexibel und genau dokumentiert und eng an die durch sie beschriebenen Daten angelagert werden.[4]

[4] Hier unterscheiden sich die Anforderungen an Transparenz und systematischer Selbstkritik nicht von denen, die etwa an wissenschaftliche Editionen von Texten gerichtet werden. Vgl. etwa Clement 2011.

In der bisherigen Projektarbeit erfassen die Forscherinnen Informationen aus den Archivmaterialien in einer Office-Tabellenkalkulation und markieren Ergänzungen und Vermutungen wie üblich typographisch, z.B. durch Kursivierungen, geschweifte oder eckige Klammern (›[…]‹) usw. (Dazu gibt es ein im Projekt erstelltes Richtliniendokument, in dem aufgeschlüsselt wird, welche typographische Hervorhebung in welchem Datenfeld was zum Ausdruck bringt.). Über die Notiz hinaus, dass eine bestimmte Information durch die heutigen Forscherinnen in das Material getragen wurde, befassen wir uns mit Möglichkeiten, die Gewissheit oder Zuversicht der Forscherin bezüglich der Richtigkeit dieser Eintragung festzuhalten, eventuell neben einer Bemerkung dazu, auf welche (›externen‹) Quellen sie sich dabei stützt. Das hier vorgestellte Datenmodell räumt diese Möglichkeiten ein, es wird aber auch kritisch zu überprüfen sein, ob sie in praktikabler Weise in den Arbeitsprozess integriert werden können.

2. Die drei Ebenen des Datenmodells

Aktuell wird das Datenmodell in einem iterativen Verfahren weiterentwickelt, das sich einerseits (kurz- bzw. mittelfristig) an einer flexiblen und explorativen Beforschung des Materials und andererseits (langfristig) an einer interoperablen Veröffentlichung der Daten orientiert. Konkret bedeutet das eine konzeptuelle Orientierung am CIDOC-CRM-Standard[5] und zugleich eine umgehende Implementierung und kontinuierliche Anpassung in einer Property-Graph-Datenbank (Neo4j). Die konsequente Konversion des gesamten Datenbestandes in ein CIDOC-CRM-kompatibles Format und ihre Publikation als RDF / Linked Open Data werden zu einem späteren Zeitpunkt erfolgen, während bereits jetzt projektintern mit den Werkzeugen der Graphdatenbank gearbeitet wird. So können, auch während kontinuierlich noch weiter Daten erfasst werden, bereits im aktuellen Stand der Datenbank z.B. über einfache Graphanalysen Dubletten identifiziert oder Hypothesen über Problemkonjunkturen oder Personennetzwerke gebildet werden. Eine weitere Herausforderung im Prozess der Datenmodellierung besteht darin, den Besonderheiten spezifisch juridischer Phänomene durch die Integration von Einsichten aus historischen Projekten (z.B. *Trials of the Late Roman Republic* [6]) oder von Erfahrungen und etablierten Standards in der digitalen Erfassung geltenden Rechts und juridischer Deliberationen (z.B. *LKIF-core* [7]) Rechnung zu tragen.

2.1. Historische Vorgänge: Der Fall

In den Akten sind die Aufzeichnungen zuerst nach Beratungsterminen organisiert und werden in den ersten Schritten der Datenerfassung weiter aufgegliedert, so dass in der vorläufigen Datenerfassung durch die Forscherinnen in einer Office-Tabellenkalkulation einzelne Zeilen einzelnen Gesprächs-›Runden‹ an diesen Terminen entsprechen (vgl. Abbildung 1).

[5] vgl. Ore et al. 2017.
[6] vgl. Sperberg-McQueen 2016.
[7] vgl. Hoekstra et al. 2008.

Abb. 1: Datenerfassung im Spreadsheet. ©Forschungsprojekt ›Die Regierung der Universalkirche nach dem Konzil von Trient‹ 2018.

Allerdings beziehen sich Erkenntnisinteresse und wissenschaftliche Fragestellung in erster Linie nicht auf diese Archivalien oder ihre Organisation selbst, sondern auf die in ihnen dokumentierten historischen Prozesse. Van Ruymbeke et al. schlagen eine Differenzierung zwischen der ›historischen‹ oder ›phänomenalen‹ und der ›informationellen‹ Dimension vor, also zwischen den beobachtbaren Phänomenen und dem Diskurs darüber.[8] In ähnlicher Weise konzentrieren wir uns auf dieser ersten Ebene zunächst auf die historischen Phänomene: auf den oder die Petenten (in der Abbildung orange hervorgehoben), auf die Diözese, aus der die Eingabe erfolgt ist oder die von dem Problem betroffen ist (gelb), auf die in dieser Angelegenheit interpretationsbedürftigen Beschlüsse des Trienter Konzils (grün), sowie auf die an der Beratung beteiligten Kardinäle (dunkelblau), und schließlich auf das Datum (hellblau) und die Kurzfassung ihres Beschlusses (rot).

So ist der hinter der Erfassung und eben auch schon hinter dem Archiveintrag stehende historische (und genuin juridische) Zusammenhang eines ›Falls‹ die zentrale Entität der Modellierung des historischen Phänomens, ihm sind die Besprechungstermine dann erst zugeordnet. Diese Intuition wird dadurch bestätigt, dass die Protokolle durchaus innerhalb der Beratungssitzungen eine Gliederung nach den verschiedenen am jeweiligen Termin behandelten Fällen aufweisen, und sie wird weiter bekräftigt durch andere Projekte und Forschungen, die vergleichbare gerichtliche oder gerichts-analoge Vorgänge modellieren und den ›Fall‹ ebenso ins Zentrum stellen.[9] An den Fall werden nun weitere Informationen angelagert, teils als Eigenschaften des Falls, teils als mit ihm verbundene, weitere Entitäten: Ein Fall wird eröffnet dadurch, dass eine Person eine Eingabe an die Kongregation macht und die Auslegung eines Beschlusses des Trienter Konzils erbittet. Dies wird dadurch wiedergegeben, dass Personen als eigenständige Entitäten modelliert und dann mit einer qualifizierten n:m-Relation (»Petent_in«) mit dem Fall verknüpft werden. So wird die Unterscheidung zwischen der *Rolle* der Petentin und der physischen Person in der getrennten Notation von Personen- und Relations-Eigenschaften aufgezeichnet. Eine gegebene Person kann in mehreren Fällen als Petent und in anderen Fällen in einer anderen Rolle auftreten, während umgekehrt ein Fall mehrere Petenten haben kann.[10] Ebenso wird der angesprochene Konzilsbeschluss

[8] Vgl. Van Ruymbeke et al. 2017, passim. Dabei gewinnen sie ihre Definitionen in Auseinandersetzung mit der CIDOC-CRM-Erweiterung CRMsci (Doerr et al. 2018) und behalten leider ein (hier allerdings nicht weiter zu diskutierendes) physikalistisches Verständnis der ›Realität‹ bei.
[9] Z.B. Wyner / Hoekstra 2012, Sperberg-McQueen 2016.
[10] Vgl. erneut Hoekstra et al. 2008. Ob eine Person in diesem Kontext in ein- und demselben Fall in verschiedenen Rollen auftreten kann, ist erst noch wissenschaftlich zu diskutieren und kann ggf. später in einem Schema oder anderen konsistenzgewährenden Mechanismen sichergestellt werden. Zu thematischen Rollen allgemeiner vgl. Goy et al. 2018. Dort wird auch diskutiert, dass, anders als für soziale Rollen (z.B.

als ›Materie‹ mit den üblichen Indikatoren Sitzung und Dekret bzw. Canon festgehalten. Analog verhält es sich mit den anderen an dem Verfahren beteiligten Personen, mit den getroffenen Beschlüssen, den anderen Fällen, auf die die Beratungen mitunter verweisen und auch den einzelnen Anhörungsterminen. Abbildung 2 gibt eine vorläufige Konzeption dieser Zusammenhänge wieder.

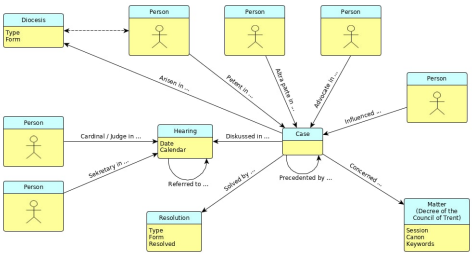


Abb. 2: Datenmodell der Ebene der historischen Phänomene. ©Forschungsprojekt ›Die Regierung der Universalkirche nach dem Konzil von Trient‹ 2018.

Dabei wird das Datenmodell, werden also die Entitätstypen, ihre Attribute und Relationen (sowie einstweilen auch Attribute von Relationen) im Ausgang von der über Jahre ausdifferenzierten und bewährten Datenerfassung zunächst intuitiv formuliert und dann in der schemafreien Graphdatenbank *Neo4j* implementiert. Eine Abbildung auf stärker standardisierte Modelle wie CIDOC CRM bzw. RDF wird später evaluiert.[11] Zwar ist diese Abbildung als Exportschnittstelle in jedem Fall vorgesehen, allerdings ist es aktuell durchaus auch vorstellbar, dass der Datenbestand im Projekt allein im Rahmen der aktuell aufgebauten (und ggf. weiterentwickelten) Graphdatenbank beforscht und im Datenexport z.B. auf eine SPARQL-Schnittstelle verzichtet wird.

2.2. Schriftliche Quellen: Positiones

Die historischen Fälle wurden, wie oben beschrieben, in Berichten, sog. *Positiones* protokolliert, die sich in den Archiven finden lassen. In diesem Zusammenhang können z.B. der schriftführende Sekretär und Informationen über Siegel- und Pergament-Beschaffenheiten erfasst werden. Vor allem aber ist auf dieser Ebene der protokollarische Bericht (eine ideelle Entität, die sich auf die Phänomene der ersten Ebene bezieht) mit einem schriftlichen Dokument (einem symbolischen Ausdruck) und dieses wiederum mit einem materiellen

Äbtissin, Kanzler), thematische Rollen eine Modellierung als eigenständige Entität i.d.R. nicht erfordern und es nahe liegt, sie ausschließlich in Verbindung mit dem jeweiligen Fall zu modellieren — im vorliegenden Projekt eben als Typ der Relation zwischen Person und Fall.

[11] Vor allem auf dieser Ebene konzentriert sich im Übrigen die rechtliche Besonderheit des Forschungsgegenstands und -projekts. Die im Weiteren beschriebenen zweite und dritte Ebene sind auf historische Forschung bzw. auf geisteswissenschaftliche Forschung als solche zugeschnitten und weisen jenen spezifischen juridischen Charakter nicht mehr auf. Das legt auch nahe, auf dieser ersten Ebene sorgfältig zu überlegen, ob neben dem aus dem Bedarf von Kulturerbe-Institutionen stammenden CIDOC CRM Standard nicht auch Taxonomien, Vokabularien oder Ontologien aus dem Bereich des geltenden Rechts und der Rechtsinformatik integriert werden können. Zu denken wäre etwa an *LKIF-Core* (Hoekstra et al. 2008), *UNDO* (Peroni et al. 2017) oder die verschiedenen Elemente aus dem *European Legal Taxonomy Syllabus* (Ajani et al. 2016).

Träger verknüpft, der seinerseits in bestimmten archivalischen Fundzusammenhängen (Medieneinheit, Regal, Fundus, Gebäude, Adresse) eingeordnet ist. Damit ist auf dieser Ebene auch eine wichtige Möglichkeit angesiedelt, einige der angesprochenen Unsicherheiten in der Datenbank abzubilden. Denn neben der allgemeinen Relation zwischen Fall und Bericht werden auch die speziellen Einträge von Personen und deren Relationen auf der Ebene der historischen Vorgänge den Namen bzw. den Symbolen und Zeichenfolgen in den *Positiones* abgelesen. Wir möchten also idealerweise zweierlei Dinge explizit festhalten: Erstens, dass sich z.B. unter den Zeichenfolgen im Dokument eine befindet, die wir als eine Namensnennung identifiziert haben; zweitens, dass diese Namensnennung unserer Ansicht nach eine bestimmte historische Person bezeichnet. Streng genommen sind die historischen Personen also nur über Instanzen von Namensnennungen in den Dokumenten mit den Fällen assoziiert.

Informationen dieser Art sind in der aktuellen Datenerfassung nur sehr implizit und grundsätzlich enthalten, so dass diese uns für die Modellierung auf dieser Ebene keine ausreichende Orientierung bietet. Die oben in Anschlag gebrachte Begrifflichkeit von ideeller Entität, symbolischem Ausdruck und materiellem Träger legt allerdings nahe, sich an den CIDOC CRM Standard (und seine Erweiterungen FRBRoo bzw. CRMtex)[12] anzulehnen. In einem ersten Anlauf verwenden wir also auf dieser Ebene Namen für Entitäts- und Relationstypen, die aus jenem Kontext stammen, wie in Abbildung 3 ersichtlich:[13] Die *Positio* ist dem Wesen nach ein CIDOC CRM »E31_Document«, welches den Fall dokumentiert (»P70_documents«). Im Besonderen enthält das Dokument (»P165_incorporates«) einen symbolischen Ausdruck, genauer: die Namensnennung eines Akteurs (»E82_Actor_Appellation«, eine Unterklasse von »E41_Appellation« bzw. »E90_Symbolic_Object«). Diese Namensnennung wird in der Quelle in einer Kette von Glyphen (»TX7_Written_Text_Fragment«) physisch abgebildet, und sie identifiziert (»P131_identifies«) einen Akteur (»E39_Actor«, entweder eine einzelne »E21_Person« oder eine ganze »E74_Group«) auf der historischen Ebene 1.

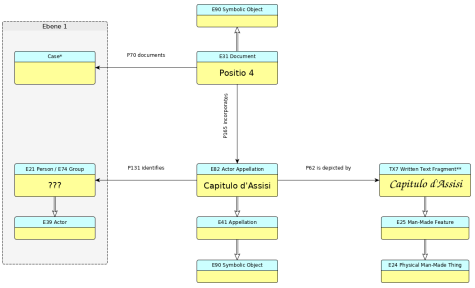


Abb. 3: Datenmodell der Ebene der archivalischen Quellen. ©Forschungsprojekt ›Die Regierung der Universalkirche nach dem Konzil von Trient‹ 2018.

[12] vgl. Bekiari et al. 2017 und Murano / Felicetti 2017.
[13] An dieser Stelle sei noch einmal darauf hingewiesen, dass es sich um einen aktuell noch andauernden, iterativen Entwicklungsprozess handelt und die hier vorgestellte Konzeption sich erst noch in Implementierung und praktischer Datenerfassung und -analyse bewähren muss.

Im aktuellen Entwicklungsstand ist noch offen, in welcher Weise diese Information mit den Aspekten der verschiedenen Rollen auf der Ebene 1 der historischen Vorgänge und Entitäten integriert wird. Denn es geht einerseits nicht allein um die Nennung einer Person im Zusammenhang mit einem Fall, sondern auch um die Zuweisung einer bestimmten Rolle innerhalb des Falles an diese Person, und andererseits wird die Person oft nicht mit persönlichem Namen, sondern mit ihrer sozialen Rolle (z.B. »der Domkapitular von Assisi«) genannt, die Namensnennung (»E82_Actor_Appellation«) identifiziert also zunächst gar keine konkrete Person. Ein erster Ansatz unter Rückgriff auf den Vorschlag von Martin Doerr[14] versucht, dies über Attribute von Attributen (»P14.1_in_the_role_of« als Attribut von »P14_performed«) abzubilden. Dies würde uns aber zwingen, eine Aktivität in das Datenmodell hinein zu modellieren, über welche die genannten Attribute mit der Person verbunden werden könnten. Zudem ist zweifelhaft, ob diese Konstruktion sich in der Implementierung und im späteren Export in RDF leicht umsetzen lässt. Eine andere Möglichkeit wäre, die Rollen von Petenten, beratenden Kardinälen usw. doch zu eigenständigen Entitäten zu machen. So lassen sie sich zu Endpunkten weiterer Relationen, etwa mit den Namensnennungen, machen. Aber auch diese Strategie ist nicht unproblematisch: Wir würden — entgegen der ursprünglichen, oben dargestellten Intuition und entgegen der Empfehlung in[15] — die Entitäten in uneleganter Weise vervielfachen, was nicht zuletzt die Abfragen komplizierter macht; die Namensnennung (oder die Nennung der sozialen Rolle, z.B. des »Domkapitular von Assisi«) ist ja selbst gar nicht auf die Rolle »Petent« bezogen; schließlich wird womöglich die spätere Integration mit CIDOC CRM erschwert. Kurzum: die Überlegungen hierzu sind zum gegenwärtigen Zeitpunkt noch nicht abgeschlossen.

2.3. Forschungsprozesse: Assignments

Auf der und durch die dritte Ebene wird es schließlich konkret möglich, die im Forschungszusammenhang auftretenden Zweifel festzuhalten. Dabei führt kein Weg daran vorbei, die Forscherinnen selbst ins Datenmodell mit einzubauen. Denn sie sind es, die Unklarheiten des Quellenmaterials feststellen oder selbst fallible Hypothesen aufstellen. Die Attribute und Relationen auf den ersten beiden Ebenen gehen zurück auf Interpretationen der Forscherinnen und müssen ggf. mit einer Art ›Zweifels-Koeffizienten‹ versehen werden. Dies abzubilden bringt jedoch beträchtliche Komplikationen mit sich: In der Implementierung in der Graphdatenbank ist es zwar möglich, Relationen mit Eigenschaften zu versehen, nicht aber Attribute. So könnte man zwar z.B. eine ungewisse Einmischung des spanischen Königs als eine Relation zwischen König und Fall modellieren, an die man die Eigenschaft ›ungewiss‹ und eventuell noch weitere Eigenschaften wie Quellen notiert, auf die sich die Annahme stützt oder solche, die sie umgekehrt gerade in Zweifel ziehen. Es ist aber schon nicht mehr einfach möglich, die Form eines Beschlusses, wenn sie als Attribut des Beschlusses notiert ist, zum Gegenstand einer vergleichbaren Qualifikation zu machen. Auf der Seite des uns in vielen Fragen zur Orientierung dienenden CIDOC CRM Ökosystems werden die entsprechenden Probleme von Alexiev[16] beschrieben.

[14] vgl. Martin Doerr 2015, passim.
[15] vgl. Goy et al. 2018.
[16] vgl. Alexiev 2012.

Wir erproben in diesem Zusammenhang die Umsetzung eines Vorschlages von Niccolucci und Hermon: In zweifelhaften Fällen wird die Feststellung eines Wertes selbst als ein Ereignis (»E13_Attribute_Assignment« oder eine seiner Unterklassen) modelliert, das selbst zum Gegenstand einer Einschätzung der Zuverlässigkeit werden kann (vgl. Abbildung 4). [17] Auf diese Weise ist es möglich, konkurrierende Interpretationen verschiedener Forscherinnen hinsichtlich bestimmter Attribute oder Relationen vollständig zu erfassen. Ebenso können verschiedene Forscherinnen einer jeden solchen Interpretation unterschiedliche Gewissheits- und Zuverlässigkeits-Werte (Abschnitt 2.3) zuschreiben.

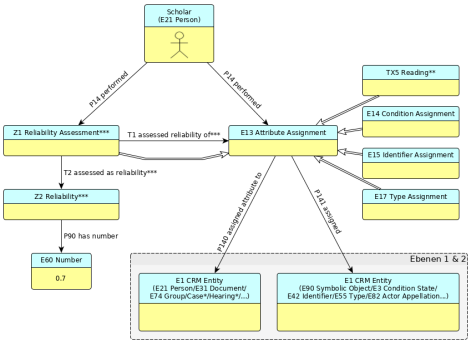


Abb. 4: Datenmodell der Ebene des Forschungsprozesses. ©Forschungsprojekt ›Die Regierung der Universalkirche nach dem Konzil von Trient‹ 2018.
[18]

Als Alternative zu diesem Vorschlag wäre die explizite Modellierung von ganzen Interpretationszusammenhängen als ›Überzeugungssystemen‹ denkbar, wie sie die CRMinf Erweiterung[19] definiert, die mit neuen Entitäten (»I1_Argumentation«, »I2_Belief«, »I4_Proposition_Set« u.ä.) den Prozess des wissenschaftlichen Interpretierens und Schlussfolgerns insgesamt abbildet.[20] Das würde einerseits eine deutlichere und weitergehende Modellierung der wechselseitigen Stützung und Erklärung einzelner Aspekte jenes Wissens, idealerweise auch ausführliches Schlussfolgern über hier erst noch implizit enthaltenes Wissen erlauben. Aus Gründen der pragmatischen Implementierung und Anwendung folgen wir jedoch dem einfacheren Ansatz von Niccolucci und Hermon, der von diesen als mit CRMinf zwar kompatibel, jedoch mit den genannten Vorzügen qualifiziert beschrieben wird.[21]

[17] vgl. Niccolucci / Hermon 2016.
[18] Bezeichnungen der Entitäten und Relationen nach CIDOC CRM v6.2. Doppelte Pfeile geben Unterklassen-Relationen an. (*) Für die Entitäten ›Case‹ und ›Hearing‹ ist noch keine CIDOC-CRM-kompatible Definition gefunden. (**) Die Entität ›TX5 Reading‹ ist definiert in der CIDOC CRM Erweiterung CRMtex; vgl. Murano / Felicetti 2017. (***) Die Entitäten ›Z1 Reliability Assessment‹ und ›Z2 Reliability‹ sowie die Relationen ›T1 assessed reliability of‹ und ›T2 assessed as reliability‹ sind definiert in Niccolucci / Hermon 2016.
[19] vgl. Stead et al. 2015.
[20] vgl. dazu auch erneut Van Ruymbeke et al. 2017.
[21] vgl. Niccolucci / Hermon 2016, S. 286. Im Speziellen für die Fragen der graphologischen Unsicherheiten vgl. Murano / Felicetti 2017.

Die Zuverlässigkeitsschätzung nach Niccolucci und Hermon[22] erlaubt es, neben einer quantitativen Bestimmung der Zuverlässigkeit auch Faktoren festzuhalten, die die Einschätzung beeinflussen (über »P15_was_influenced_by« oder »P33_used_specific_technique«, die etwa die Klassifikation über einen definierten Kriterienkatalog aufnehmen kann) und auf weitere Dokumentation, z.B. einen Aufsatz, in dem die Forscherin zur Frage Stellung bezieht, hinzuweisen (über »P70_is_documented_in«). Gründe wie die Tatsache, dass einer der möglichen Werte der historischen Realität wahrhaft entspricht, dass diese vergangene historische ›Realität‹ aber eben nicht mehr objektiv überprüfbar ist, führen Niccolucci und Hermon dazu, die Zuverlässigkeitswerte als numerische Werte zwischen 0 und 1 im Rahmen eines ›Fuzzy‹ Kalküls zu verstehen. Damit können die Alternativen nicht nur mit Vorbehalten versehen und verglichen werden, sondern es können im Prinzip auch weitergehende Berechnungen nach einem formalen Kalkül angestellt werden.[23]

3. Ausblick

In diesem Ausblick werden einige Herausforderungen genannt, die in der Modellierung und im praktischen Projektkontext noch offen sind oder sich als Herausforderungen durch die bisherigen Schritte erst deutlich herauskristallisiert haben.

Erstens gilt es, die oben genannten Vorläufigkeiten aufzusuchen, d.h. Punkte wie die Verankerungen der Relationen zwischen den Ebenen auf der jeweils tieferen Ebene zu klären und das Datenmodell konsequenter an CIDOC CRM auszurichten. Entitäts- und Relationstypen sollten ebenso wie Attribut-Bezeichnungen den durch CIDOC CRM definierten Bezeichnungen entsprechen. Gegebenenfalls kann auf offizielle Erweiterungen wie CRMinf zurückgegriffen werden. Um die Komplexität in der Eingabe und Bearbeitung der Daten zu beschränken oder zu maskieren, könnten auch eigene Bezeichnungen gewählt werden, solange diese möglichst eindeutig auf CIDOC CRM (und Erweiterungen) gemappt werden können. Sollte dies schließlich aus Gründen der ökonomischen und intuitiven Arbeit mit dem Datenbestand nötig werden, könnte es sich anbieten, einzelne Phänomene nach einer eigenen Konzeption zu modellieren und sie erst in einem letzten Export-Schritt durch eine wohldefinierte Transformation in RDF bzw. CIDOC CRM zu überführen. Dabei ist vor allem an die Möglichkeiten zu denken, bei der projektinternen Erfassung und Beforschung der Daten mit einem *Property Graph* zu arbeiten, also den Relationen ihrerseits Attribute zuordnen zu können. Dies lässt sich nicht direkt in RDF Tripel abbilden, sondern nur über die (automatische) Erzeugung zusätzlicher Entitäten anstelle der Relationen, die dann einerseits mit den relationierten Entitäten verknüpft werden und andererseits die Relations-Attribute aufnehmen können.

[22] vgl. Niccolucci / Hermon 2016.
[23] Der Begriff der unscharfen, ›fuzzy‹ Mengen, Systeme und Methoden wurde von Zadeh eingeführt, vgl. Zadeh 1965. Zu seiner Verwendung im geisteswissenschaftlichen Zusammenhang vgl. Termini 2012 und Thaller 2018. In unserem Fall werden beim Import aus dem Spreadsheet Daten, die als Interpretation gekennzeichnet sind, in einem ersten Schritt mit einem Standard-Gewissheitswert von 0,95 versehen. Durch eine Abfrage von solchermaßen qualifizierten *Assignments* können sie in einem zweiten Durchgang in Fällen, wo es sich um eine tatsächlich ungesicherte Interpretation handelt, nach unten korrigiert werden. Welche Workflows und Werte sich hier optimal verwenden lassen, muss sich aber erst noch in der Praxis ergeben.

Zweitens liegt auf der Hand, dass mit der Entwicklung des Datenmodells nur der allererste (oder jedenfalls nur ein sehr früher) Schritt getan ist, dass es also in einer Datenbank (a) implementiert und mit den bereits vorliegenden und weiter angelieferten Daten (b) befüllt werden muss. Dabei gehen wir im Projekt parallel vor: eine auf der Basis der bereits vorliegenden Erfassungs-Tabellen intuitiv erzeugte und befüllte Graphdatenbank (vgl. Abbildung 5) wird zunehmend und durch systematische Transformationen an das reflektierte Datenmodell angepasst, während umgekehrt das Datenmodell nicht zuletzt durch die Arbeit mit der bereits bestehenden Graphdatenbank verfeinert wird. Durch die systematische Definition und Redefinition der Datenbank kann diese jederzeit mit unterschiedlichen Datenmodellen aus den ursprünglichen (CSV-)Tabellen neu erzeugt werden.

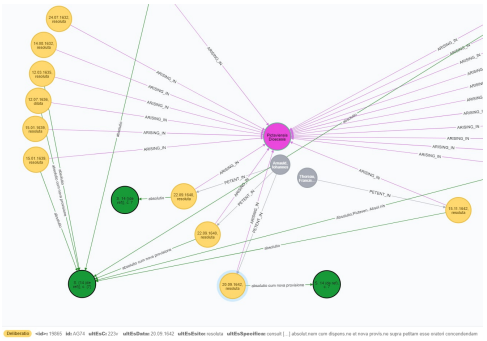


Abb. 5: Ausschnitt der Graphdatenbank. ©Forschungsprojekt ›Die Regierung der Universalkirche nach dem Konzil von Trient‹ 2018.

Daran muss drittens die Entwicklung von Oberflächen-Elementen anschließen, die zum einen auch technisch weniger versierten Projektbeteiligten die Erfassung, Bearbeitung und Abfrage von Daten erlaubt. Beispielsweise sollte die Komplexität der Einschätzung von Zuverlässigkeiten aus arbeitsökonomischen Gründen durch technische Vorkehrungen wie z.B. ein- und ausblendbare Eingabefelder, Default-Werte, und möglicherweise grafische Elemente wie Repräsentation und Auswahl der möglichen Werte auf einer Farbskala (anstelle von numerischen Werten und Komma- oder Prozentzahlen) reduziert bzw. ausgeblendet und nur bei Bedarf erweitert und expliziert werden. Auch für die Dokumentation der Forschung und als Angebot an andere an dem Datensatz interessierte Besucher kann darüber nachgedacht werden, bestimmte Ansichten und Analysen der Daten aufzubereiten und über den Datenbank-Server frei zugänglich zu machen. Dazu eignen sich die sogenannten *Browser Guides* in Neo4j hervorragend (vgl. Abbildung 6).[24]

[24] In den Neo4j Browser Guides sind der Mechanismus und das gleichnamige Format dokumentiert.

Abb. 6: Beispiel eines *Browser Guides. ©Forschungsprojekt ›Die Regierung der Universalkirche nach dem Konzil von Trient‹ 2018.*

Viertens und letztens schließlich muss geklärt werden, in welcher Weise die hier im Fokus stehenden Informationen zu Ungewissheiten und Zweifeln eigentlich ausgewertet und in den Suchen und Analysen berücksichtigt werden sollen. Unter welchen Bedingungen und in welcher Form sollen z.B. Diskussionsteilnehmer, deren Identität unsicher ist, in Abfragen angezeigt werden? Nur, wenn die Abfragenden dies in irgendeiner Weise explizit anfordern oder immer? Sollen sie chronologisch bzw. alphabetisch in die Ergebnismenge einsortiert werden oder anhand der Gewissheit, mit der sie zur korrekten Ergebnismenge gehören? Gibt es einen praktikablen Weg, den ›Zuverlässigkeitswert‹ eines Feldes in der grafischen Anzeige zu repräsentieren? Auch den einer Relation? Wie kann abgebildet werden, dass mit einer Bezeichnung zwei Personen im Datenbestand gemeint sein könnten? Und noch vor jeder Überlegung zu Fragen der Darstellung: Wie verändert sich die Gewissheit, wenn nach zwei Feldern gefragt ist und beide ungewiss sind, wie wenn ein Feld sehr ungewiss ist?[25]

Die weitere Projektarbeit wird zeigen, ob sich auf alle diese Fragen und für alle genannten Herausforderungen befriedigende, praktikable und zugleich wissenschaftlich fortschrittliche Lösungen und Antworten finden lassen.

[25] Zu Ansätzen, Kalküle aus der *fuzzy logic* in historischer Forschung oder allgemeiner zur Integration verschiedener Expertenmeinungen heranzuziehen vgl. Thaller 2018 bzw. Yager et al. 2012.

Bibliographische Angaben

Gianmaria Ajani / Guido Boella / Luigi di Caro / Livio Robaldo / Llio Humphreys / Sabrina Praduroux / Piercarlo Rossi / Andrea Violato: The European Legal Taxonomy Syllabus: A Multi-Lingual, Multi-Level Ontology Framework to Untangle the Web of European Legal Terminology. In: Applied Ontology 11 (2016), H. 4, S. 325–375. [Nachweis im GBV]

Benedetta Albani: In universo christiano orbe: la Sacra Congregazione del Concilio e l'amministrazione dei sacramenti nel Nuovo Mondo (secoli XVI–XVII). DOI: 10.3406/mefr.2009.10577 In: Mélanges de l'École française de Rome / Italie et Méditerranée 121 (2009), H. 1, S. 63–73. [online] [Nachweis im GBV]

Vladimir Alexiev: Types and Annotations for CIDOC CRM Properties. Report at Digital Presentation and Preservation of Cultural and Scientific Heritage (DiPP2012, Veliko Tarnovo 18.-21.09.2012). In: ontotext.com. Documents. Publications. Veliko Tarnovo 18.09.2012. PDF. [online]

Chryssoula Bekiari / Martin Doerr / Patrick Le Bœuf / Pat Riva: FRBRoo: object-oriented definition and mapping from FRBRER, FRAD and FRSAD. In: cidoc-crm.org. Frbroo. Version 3.0. von September 2017. PDF. [online]

Tanya Clement: Knowledge Representation and Digital Scholarly Editions in Theory and Practice. DOI: 10.4000/jtei.203 In: Journal of the Text Encoding Initiative 1 (2011). DOI: 10.4000/jtei.125

Martin Doerr: How to model Roles in the CIDOCCRM RDF encoding. In: cidoc-crm.org. Frequently Asked Questions. Beitrag vom 12.02.2015. [online]

Martin Doerr / Athina Kritsotaki / Yannis Rousakis / Gerald Hiebel / Maria Theodoridou et al.: Definition of the CRMsci. An Extension of CIDOC-CRM to support scientific observation. In: cidoc-crm.org. CRMsci. Version 1.2.5. von Mai 2018. PDF. [online]

Anna Goy / Diego Magro / Marco Rovera: On the role of thematic roles in a historical event ontology. In: Applied Ontology 13 (2018), H. 1, S. 19–39. [Nachweis im GBV]

Rinke Hoekstra / Joost Breuker / Marcello Di Bello / Alexander Boer: The LKIF Core Ontology of Basic Legal Concepts. PDF. [online] In: Proceedings of LOAIT 07: II Workshop on Legal Ontologies and Artificial Intelligence Techniques. Hg. von Pompeu Casanovas / Maria Angela Biasiotti / Enrico Francesconi / Maria Teresa Sagri. (LOAIT: 2, Stanford, CA, 04.06.2007) Aachen 2008, S. 43–63. URN: urn:nbn:de:0074-321-1

Francesca Murano / Achille Felicetti: Definition of the CRMtex. An Extension of CIDOC CRM to Model Ancient Textual Entities. In: cidoc-crm.org. CRMtex. Version 0.8. von Januar 2017. PDF. [online]

Franco Niccolucci / Sorin Hermon: Expressing Reliability with CIDOC CRM. In: International Journal on Digital Libraries 18 (2016), H. 4, S. 281–87. Artikel vom 07.10.2016. [Nachweis im GBV]

Definition of the CIDOC Conceptual Reference Model. Hg. von Christian Emil Ore / Martin Doerr / Patrick LeBœuf / Stephen Stead. In: cidoc-crm.org. Version 6.2.2. von September 2017. PDF. [online]

Silvio Peroni / Monica Palmirani / Fabio Vitali: UNDO: The United Nations System Document Ontology. In: International Semantic Web Conference 2017. Hg. von Claudia D'Amato. 2 Bde. (ISWC 2017, Wien, 21.-25.10.2017) Cham 2017. Bd. 2, S. 175–183. [Nachweis im GBV]

Muriel Van Ruymbeke / Pierre Hallot / Roland Billen: Enhancing CIDOC-CRM and Compatible Models with the Concept of Multiple Interpretation. DOI: 10.5194/isprs-annals-IV-2-W2-287-2017 In: ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences 2017 IV-2/W2. Hg. von J. Hayes / C. Ouimet / M. Santana Quintero / S. Fai / L. Smith. (Ottawa, Canada 28.08.- 01.09.2017) Red Hook 2017, S. 287–94. [online] [Nachweis im GBV]

Michael Sperberg-McQueen: Trials of the Late Roman Republic: Providing XML infrastructure on a shoe-string for a distributed academic project. In: Proceedings of Balisage: The Markup Conference 2016. (Balisage: 17, Washington, DC, 02-05.-08.2016) Rockville, MD 2016. (= Balisage Series on Markup Technologies, 17) [online]

Stephen Stead / Martin Doerr et al.: CRMinf: the Argumentation Model. An Extension of CIDOC-CRM to support argumentation. In: cidoc-crm.org. CRMinf. Version 0.7. von Februar 2015. [online]

Settimo Termini: On some ›Family Resemblances‹ of Fuzzy Set Theory and Human Sciences. In: Soft Computing in Humanities and Social Sciences. Hg. von Rudolf Seising / Veronica Sanz. Berlin u.a. 2012, S. 39–54 (= Studies in Fuzziness and Soft Computing, 273) [Nachweis im GBV]

Manfred Thaller: On Information in Historical Sources. In: A Digital Ivory Tower. Prolegomena for a computer science for historical studies. Blogbeitrag vom 24. April 2018. [online]

Adam Wyner / Rinke Hoekstra: A Legal Case OWL Ontology with an Instantiation of Popov v. Hayashi. In: Artificial Intelligence and Law 20 (2012), H. 1, S. 83–107. [Nachweis im GBV]

Ronald Yager / Henri Prade / Didier Dubois: Merging Fuzzy Information. In: Fuzzy Sets in Approximate Reasoning and Information Systems. Boston, MA u.a. 2012. (= The Handbook of Fuzzy Sets Series, 5) [Nachweis im GBV]

Lotfi Zadeh: Fuzzy Sets. In: Information and Control 8 (1965), H. 3, S. 338–353. [Nachweis im GBV]

Abbildungslegenden- und nachweise

Abb. 1: Datenerfassung im Spreadsheet. ©Forschungsprojekt ›Die Regierung der Universalkirche nach dem Konzil von Trient‹ 2018.

Abb. 2: Datenmodell der Ebene der historischen Phänomene. ©Forschungsprojekt ›Die Regierung der Universalkirche nach dem Konzil von Trient‹ 2018.

Abb. 3: Datenmodell der Ebene der archivalischen Quellen. ©Forschungsprojekt ›Die Regierung der Universalkirche nach dem Konzil von Trient‹ 2018.

Abb. 4: Datenmodell der Ebene des Forschungsprozesses. ©Forschungsprojekt ›Die Regierung der Universalkirche nach dem Konzil von Trient‹ 2018.

Abb. 5: Ausschnitt der Graphdatenbank. ©Forschungsprojekt ›Die Regierung der Universalkirche nach dem Konzil von Trient‹ 2018.

Abb. 6: Beispiel eines *Browser Guides*. ©Forschungsprojekt ›Die Regierung der Universalkirche nach dem Konzil von Trient‹ 2018.

ZfdG

Zeitschrift für digitale Geisteswissenschaften

Beitrag aus:
Sonderband 4 der ZfdG: Die Modellierung des Zweifels – Schlüsselideen und -konzepte zur graphbasierten Modellierung von Unsicherheiten. Hg. von Andreas Kuczera, Thorsten Wübbena und Thomas Kollatz. 2019. DOI: 10.17175/sb004

Titel:
The Codex – an Atlas of Relations

Autor/in:
Iian Neill

Kontakt: Iian.Neill@adwmainz.de
Institution: Akademie der Wissenschaften und der Literatur, Mainz
GND: 1082253677

Autor/in:
Andreas Kuczera

Kontakt: andreas.kuczera@adwmainz.de
Institution: Akademie der Wissenschaften und der Literatur, Mainz
GND: 1167802993 ORCID: 0000-0003-1020-507X

DOI des Artikels:
10.17175/sb004_008

Nachweis im OPAC der Herzog August Bibliothek:
1037074203

Erstveröffentlichung:
08.05.2019

Lizenz:

Sofern nicht anders angegeben (cc) BY-SA

Medienlizenzen:
Medienrechte liegen bei den Autoren

Letzte Überprüfung aller Verweise:
08.05.2019

GND-Verschlagwortung:
Graphdatenbank | Semantisches Datenmodell | Textanalyse | Wissensrepräsentation |

Zitierweise:
Iian Neill, Andreas Kuczera: The Codex – an Atlas of Relations. In: Die Modellierung des Zweifels – Schlüsselideen und -konzepte zur graphbasierten Modellierung von Unsicherheiten. Hg. von Andreas Kuczera, Thorsten Wübbena und Thomas Kollatz. 2019 (= Sonderband der Zeitschrift für digitale Geisteswissenschaften, 4). PDF Format ohne Paginierung. Als text/html abrufbar unter DOI: 10.17175/sb004_008.

Iian Neill, Andreas Kuczera

The Codex – an Atlas of Relations

Abstracts

This paper looks at how deep integration between text and data is attempted in The Codex project. Standoff properties are used to mediate between the plain text stream and entities modelled in the Neo4j graph database. A dynamic standoff property text editor was constructed to enable real-time changes to text and annotations without invalidating standoff property indexes. An examination of the multidimensional affordances offered by standoff properties is explored, with reference to how annotations and graph entities can combine to construct an ›atlas of history‹ using Codex.

In diesem Beitrag wird The Codex vorgestellt, ein Projekt, in dem basierend auf Standoff Properties Texte multidimensional annotiert und die Annotationen in eine Graphdatenbank eingebettet werden können. Darüberhinaus sind Basistext und Annotationen im Unterschied zu vielen anderen Standoff-Markup-Systemen editierbar. Auch Annotationen selbst können wieder annotiert werden, was den Forschungsdiskurs leichter nachvollziehbar machen könnte.

1. Introduction

The Codex is a project that aims to achieve deep integration between text and structured data.[1] Although data can be ›extracted‹ from texts and stored in a database – whether that data be persons or places referenced, events recounted, numerical quantities reported, etc. – the non-sequential nature of databases can make it difficult to put the data back into context. While becoming amenable to computational analysis, the crucial narrative or argumentative structure is usually lost. Text itself can be considered to be a kind of database, one that is on the one hand constrained by its sequential presentation but on the other hand makes capable the modelling of thought in its multidimensional complexity. XML is a powerful tool for the modelling of text by allowing regions of text to be marked up with semantic tags or elements, and languages such as XPATH can be used to query the XML document model.[2] However, the use of markup itself introduces a discontinuity between the text and the data annotated. Markup changes the very text it marks up by creating a new marked-up document. Also, overlapping annotations – such as are commonly required in manuscript presentation annotations – cannot be directly expressed in XML's tree-like hierarchical structure, necessitating workarounds such as standoff markup which further degrade the readability of the XML document. Without freely overlapping annotations the most essential multidimensional aspects of text (its multitude of meanings) cannot be adequately marked up.

[1] We would like to thank Elena Spadini and Joris van Zundert for their invaluable comments in preparing this article.

[2] XPATH is arguably not a substitute for database query languages such as SQL or Cypher, however. Although XML bears some similarities to a database, in that its elements can be used to represent entities, a database can represent entities from thousands of documents and allow complex querying of them, whether with set-like operations such as SQL JOINs, filtering across tables, ordering and grouping, etc. Fundamentally, XML is a document markup format while a database is a relational system (whether the entities related are tables as in SQL databases or nodes and edges as in graph databases).

The Codex aims to bridge the divide between database and text – to achieve ›deep integration‹ – by eschewing markup entirely. Standoff properties are used, instead, to represent annotations. As defined by Desmond Allan Schmidt, the use of standoff properties is a technique for recording textual properties that do not conform to a context-free grammar, and can freely overlap.[3] While standoff properties have been proposed in the digital humanities several times,[4] Codex offers a practical solution to adding annotations to text, allowing the user to *make changes to the text in real-time without breaking existing annotations*. In addition, Codex offers a selection of stylistic, presentation, and semantic annotation types, as well as tools like named entity recognition (NER) and pronoun selection. The user can easily link annotations to entities in the graph database, or create new entities and their dependencies, entirely within modal windows, allowing the user to capture and construct data within the editor. They can also add footnotes, marginalia, etc., within a modal window editor, offering the same features as the editor in the window before. Annotations *themselves* can be annotated with editorial commentary entered via the modal editor.[5] In sum, through the use of a real-time standoff properties editor, a powerful modal-window entity management system, and the Neo4j graph database, Codex aims to provide an environment suitable for annotating all varieties of text, and linking annotations back to the text on a character-level.

One of the main goals of Codex, building on the deep integration of text and data, is to construct an ›atlas of history‹ from primary source texts. This phrase is a metaphor for both the *kinds* of annotations used as well as the graphical tools employed to *visualise* connections: ›atlas‹ here refers to a network of relations that are amenable to projection in various ways.[6] In other words, the purpose of Codex is not simply to annotate entities in text, but to represent these entities as nodes and relationships in the Neo4j graph database. The end result is both an annotated document *and* a graph dataset, which means that the more entity annotations are added to texts the richer the network of relations *between* texts becomes. As a basic example, if the corpus were the collected letters of Michelangelo[7] and references to Michelangelo's father Lodovico Buonarroti were annotated, it would be trivial to query the database to find all other texts referring to Lodovico.[8] But to make clearer what can be annotated and represented in the Codex system, we propose to give an overview of the main annotation types.

[3] Schmidt 2016a, pp. 63–69; Schmidt 2016b.
[4] Schmidt 2016a, p. 64.
[5] For example, if you wished to question the attribution of an agent annotation (such as a person) without deleting or changing the annotation, you could add comments against the annotation itself.
[6] Perhaps a more accurate metaphor for Codex would be an ›atlas of relations‹ as data doesn't have to contain spatial or temporal information to be visualised.
[7] An ongoing Codex corpus, along with the diary entries of Luca Landucci, a 15th century Florentine citizen, see del Badia 1883. Imported into Codex was the English translation by de Rosen Jervis 1927, see del Badia 1927. Additional Letters from Michelangelo where transcribed and annotated in Codex, see Carden 1913.
[8] With the proper nouns also the pronouns are annotated, one can get an accurate idea of the referential density in these letters.

2. Annotations

The types of annotations in Codex currently fall into three main categories: stylistic; presentation; and semantic.[9] Stylistic annotations include commonly used typographical styles like italics, bold, underline, strikethrough, subscript, superscript, and forms of emphasis like spaced and uppercased text. The size, colour, and font type can also be set.

Fig. 1: Stylistic, layout, and semantic annotation buttons in the Codex editor. [Neill / Kuczera 2019]

However, the more interesting categories are presentation and semantic. These are broken down as follows.

2.1 Presentation

2.1.1 Page, line, paragraph, sentence, column, etc.

These annotations denote regions of text corresponding to the presentation (or layout) of a page in a manuscript or a publication. Because standoff properties can overlap freely, presentation annotations in Codex pose no danger of truncating text with presentation markup.

2.1.2 Hyphens as a zero point annotation

Hyphens present a challenge for annotating medieval manuscripts; while the editor wishes to record the location of the *hyphen*, it is not desirable to intersperse the plain text with hyphens as this will confound literal text searches. The hyphen annotation in Codex avoids this problem by representing hyphenation with *zero-dimensional annotations*. A zero-dimensional annotation is a special case of a standoff property that has a *start index* value but no *end index* value; in this way, an annotation effectively refers to a position in the text *between characters*. The hyphen itself is not stored in the text (leaving the words unhyphenated), but the annotation indicates the location of the hyphen in the original. Zero point annotation is a generalizable feature of standoff properties and can be used for other cases as well.

[9] There is a fourth category in use in Codex that is out of the scope of this paper: machine-generated syntaxical annotations. These are annotations generated by Natural Language Processing (NLP) libraries such as Google's Cloud Natural Language API which essentially add a syntaxical substrate to standoff property documents. Their potential, when combined with other annotation layers, may be explored in a future paper.

2.2 Semantic

Before proceeding we should note that for each semantic annotation there is a corresponding semantic entity in the system. The types of annotations and entities available in Codex is defined by the application code and not by an existing standard, meaning that the system can be configured by a programmer with more annotations and entities as desired.[10] Each entity is modelled as a combination of nodes and edges in the graph database, sometimes expressed in hypernode structures (in other words, an entity modelled over a cluster of nodes).[11] Creating a semantic annotation for a text is tantamount to either selecting a pre-existing entity, or creating a new, corresponding entity in the modal window interface. Conversely, entities can also be created and managed outside of the Codex editor in sections of the Codex interface specific to that entity type. Therefore, entities can be considered as an independent data-set from texts but capable of integration with texts via semantic annotations. This means that entities can be exported or imported irrespective of texts in which they may or may not be mentioned: entities don't have to be mentioned in (or inferred from) texts to exist in the system.

2.2.1 Agents

An agent refers to any kind of entity that is mentioned in the text.[12] A non-exhaustive list of agent types includes people, places, and objects (natural and man-made).[13] Collective agents like organisations, families and other groups can be represented,[14] as well as *aspects* of an agent at a point in time.[15] Metadata about an agent can be recorded in associated property nodes which are dynamically defined in the interface as needed. For example, one can record the gender of a person, their height, weight, etc., and create new property types as required.[16]

Fig. 2: A record of Lorenzo de' Medici's gender as a property record. If desired, this can even be linked back to a statement in the text via a property annotation. [Neill / Kuczera 2019]

[10] Extending the standoff property editor requires some (arguably basic) knowledge of JavaScript, while creating new types of entities requires grounding in the C# and Cypher programming languages and the ASP.NET MVC framework.
[11] According to Wikipedia, a hypernode is: A kind of graph whose node set can contain other graphs as well as basic nodes. (Wiktionary: hypernode.)
[12] Although an agent was initially defined in Codex as just a person or group that causes events (i.e., an ›historical agent‹), this definition proved too restrictive to capture the variety of entities actually referred to in texts, and which serve agent-like functions in parts of speech.
[13] An agent is stored as an (:Agent) node in the graph database. Additional information specific to the agent type (for example, latitude and longitude for place agents) is recorded by storing the data in (:Property) nodes, and relating them to the (:Agent) node they pertain to.
[14] A collective is a group of agents, such as an organisation, a family, etc. Third-party pronoun parties are often represented as agent collectives in Codex.
[15] Agent aspects can be temporal or quantitative; for example, ›the Christ Child‹ agent is a temporal aspect of ›the Christ‹ agent, while an amount of ›3,000 moggia‹ is a quantitative aspect of (a portion of) ›7,000 moggia‹ in Landucci's diary entry of April 6th, 1484 (del Badia 1883, p. 39).
[16] Property data can also be time-specific for variable properties, such as height and weight to name two.

Agents can also be related to each other via dynamically-created relations. (In Codex, these are called meta-relations, for reasons which are discussed later in the section on this entity.) The example below shows some of Lorenzo de' Medici's genealogical relationships, as well as his connection to transient agent collectives like his presence in a group of six Florentine ambassadors to Rome in 1471, and in another embassy to Rome in 1483.

RELATIONSHIP ID	RELATION	AGENT 2
2e0a3d13-8c5a-4762-903f-2d83df11ca2b	Part of	Six ambassadors [1471/09/23]
46a56b46-f962-4735-ba82-b7ff7dae6922	Brother of	Giuliano de' Medici
980feab9-ed6f-4b7d-b13d-0518333e1283	Part of	The Medici family
993428e1-efb9-43b3-a299-7ba23460f730	Child of	Cosimo de' Medici
cb1ad5d7-e02e-4393-8ca9-2ffa997b607e	Child of	Madonna Lucrezia [de' Medici]
fbc97384-90df-4a2e-8fbb-d0917cca45dd	Part of	Three Florentine ambassadors [1483/11/10]
2afb8754-1705-456a-ae32-97087991fe27	Married to	Madonna Clarice
00700f9b-5369-4a5e-a426-d1dead67d710	Parent of	Giovanni de' Medici

Fig. 3: Screenshot from ›Codex‹ of a list of Lorenzo de' Medici's relationships in the system. [Neill / Kuczera 2019]

2.2.2 Claims

A claim refers to a statement concerning one or more agents, usually with respect to a place and a time. A claim is essentially a statement that usually takes the form of an event (an event claim), but can also represent a thought or an opinion. A claim entity in Codex is not taken as a statement of fact, substantiated or otherwise, but is rather a data-structure resembling a verb-phrase with prepositional agents. For example, the statement that »Lorenzo de' Medici died on his estate at Careggi« made by Luca Landucci in his diary entry of April 8th, 1492, is modelled in Codex as a ›(Subject) Lorenzo de' Medici, ›(Event) died‹, ›(At) Careggi‹, visualised in the Codex interface and Neo4j database browser in Figure 4 and Figure 5.[17] Our approach is not to assert whether or not the statement reports a fact, merely to allow the editor to annotate the statement.

	ROLE	TYPE	NAME	AGENTS	TIME
View	Event	died	Lorenzo de' Medici died on his estate at Careggi	(According To) Luca Landucci (b.1436 - d. 1516; apothecary) (At) Careggi (Subject) Lorenzo de' Medici (b. 1449/01/01 - d. 1492/05/08; active from 1469/12/02)	On 1492/April/8 - -:--:--

Fig. 4: The event claim by Luca Landucci about Lorenzo de' Medici's death in the Codex interface. [Neill / Kuczera 2019]

[17] del Badia 1883, p. 53-54.

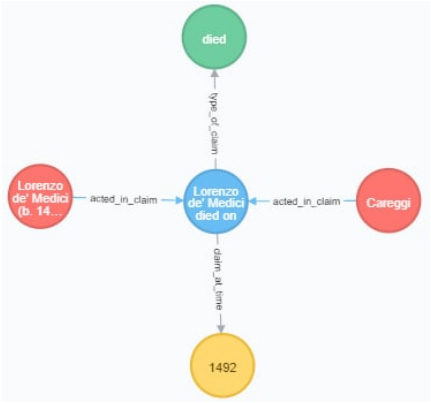


Fig. 5: A representation of the above event claim as nodes and edges in the Neo4j browser. The blue node is a (:Claim); the red nodes are (:Agent)s; the green node is a (:Concept); and the yellow node is the (:Time). [Neill / Kuczera 2019]

2.2.3 Texts

A text entity in Codex is composed of plain text and a collection of standoff-properties.[18] For convenience, texts can be assigned a ›type‹ indicating their function (such as ›body‹, ›footnote‹, ›margin note‹, etc.) but there are no limitations about the kind of text stored. Therefore, a text can contain as much of or as little of the source text as appropriate. In the case of the Luca Landucci Diary, each diary entry (of which there are several per page in the source) is stored in a separate text node, whereas each Michelangelo letter as a whole is stored in a text node.[19] Presentation annotations are used to mark sections of the text corresponding to the source (e.g., pages, columns, etc.), and the structure annotation can be used to link texts to structure entities which represent arbitrary sections of a publication (e.g., the chapters the text belongs to).

A text annotation in Codex is an annotation that relates a region of text in a text entity to a different text entity. There are two annotation topologies available to text annotations: they can either be applied to a region of text, like most annotations; or they can be inserted ›invisibly‹ between characters.[20] We can think of these topologies as one-dimensional and zero-

[18] Entity means here the node or cluster of nodes and edges that model the entity in the graph. For example, a text entity consists of a (:Text) node and a collection of (:StandoffProperty) nodes connected via the [:text_has_standoff_property] edge. We use the term standoff property document (SPD) to refer to the combination of plain text and standoff properties that makes up an annotated text in Codex.
[19] There is currently a practical limitation to the length of an individual text node in Codex of about ten kilobytes per node, due to a technical restriction related to how many HTML elements a browser can render on the screen at one time without noticeable lag. Other web applications have implemented solutions to this rendering issue, however, which can also be applied to Codex at a later date.
[20] A zero-dimensional annotation is ›invisible‹ with respect to the plain text stream (output) but is displayed within the Codex editor. The editor freely supports text characters that are either ›in-stream‹ or ›out-of-stream‹. ›In-stream‹ characters appear within the plain text stream, ›out-of-stream‹ characters do not; but both kinds can be annotated.

dimensional annotations.[21] It will be remembered that zero-dimensional annotations are also used to represent hyphen annotations; in the case of text annotations they could function as footnote numbers that the editor wishes to position in a text, but doesn't want to be included in the text per se.

2.2.4 Meta-Relations

A meta-relation entity is a relationship between agents with a number of features distinguishing it from simple relationships or edges in a graph database:

1. It is dynamically definable in the Codex interface. We have found the ability to create new relationship types in an *ad hoc* way to be invaluable for capturing the fluidity of relationships in texts. When the selection of relationship types is hard-wired into a program it becomes burdensome to create new ones; enabling the user to create them freely in the interface encourages the spontaneous creation of relationship types that are more fit-for-purpose;
2. It is bidirectional, meaning the user is able to specify both directions of the relationship (e.g., ›parent of‹/›child of‹) rather than being forced to adopt a single direction (e.g., ›parent of‹) imposed by graph edges. This encourages the user to think in terms of the overall relationship – e.g., ›parentage‹ – rather than forcing them to arbitrarily choose a single relationship to represent by implication a bidirectional relationship.[22] One advantage of this in constructing Cypher queries is that an agent's participation in a meta-relation can be found *without needing to consider their role in the relationship*, although that role is recorded and can still be explicitly queried if desired;[23]
3. They are composable within a *hierarchy*, effectively allowing *relationships themselves to be treated as a graph*. For example, one can define an overarching relation type like ›interpersonal relationships‹ and nest subordinate types beneath it, like ›social relationships‹, ›family relationships‹, ›professional relationships‹, etc., allowing one to query relationships between agents (e.g., people) at an abstract level. Rather than being limited to finding simply the ›friends of‹ a person, one can expand a query to also retrieve ›associates of‹, ›acquaintances of‹, ›confidantes of‹, etc.

A meta-relation annotation, therefore, is an annotation that refers to a meta-relation entity. Its practical purpose is to allow the editor to annotate agent relationships from texts, extending the network of relationships between agents. Ultimately, such networks allow the user to *find indirect connections between texts* on the basis of the relationships between agents

[21] A one-dimensional annotation contains both start index and end index values equivalent to a line segment, whereas a zero-dimensional annotation contains only a start index equivalent to a point.

[22] Whether or not one models ›parentage‹ with a ›()-[:parent_of]->() ‹ edge or a ›()-[:child_of]->() ‹ edge is to some extent an arbitrary choice; and with inherently unidirectional relationships such as ›married to‹ either two relationships need to be created at all times -- (a)-[:married_to]->(b), (b)-[:married_to]->(a) -- or the read query needs to accommodate either possibility, i.e., (a)-[married_to]-(b).

[23] Because a meta-relation contains more information than a unidirectional edge, meta-relation hypernodes can be programmatically decomposed into unidirectional edges as a final projection, or as a Cypher optimisation step.

established in the network. In Figure 6, the orange line beneath ›son of Antonio‹ is a meta-relation annotation indicating the source of the statement that Luca Landucci was the son of Antonio Landucci.

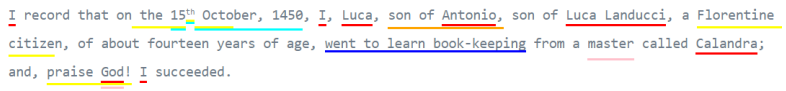


Fig. 6: An example of a meta-relation annotation (orange underline) in the Codex interface. [Neill / Kuczera 2019]

2.2.5 Concepts

A concept in Codex is a class or type that, taken as a whole, is part of a common ontology in the system. Note that the ontology is *not* common in the sense of being a ›universal‹ or ›world‹ ontology, but merely functions as a subgraph that is shared among other entity types (such as agents, claims, meta-relations, etc.) for the purposes of a common, reusable reference. Rather than constituting a universal, top-down ontology, the concept subgraph is in practise composed of any number of open or idiolectic ontologies defined by the user.[24] Codex already contains a number of idiolectic ontologies, such as ontologies for types of events, places, relationships, professions, etc. For example, claim entities reference the ›Events‹ subset of the concept subgraph. The concept ›Event‹ is the root node of the Events ontology and contains all types (and subtypes) of events that occur in the project domain.[25] Note that concepts can have more than one parent, if desired, as a graph is not bound by the limitations of a tree. Changing the structure of the ontology (e.g., moving a child concept to a different parent) is easily done through the Codex interface, meaning that ontological structures can be kept fluid to suit the evolving understanding of the project domain.

A concept annotation, therefore, is an annotation that refers to a concept entity in the common ontology. In Figure 7, the green underlines represent concept annotations on the words ›flautist‹ and ›prodigy‹.

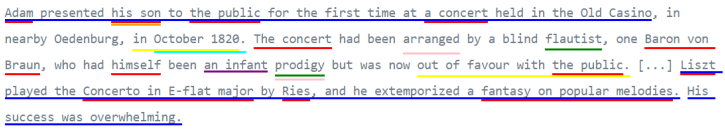


Fig. 7: Concept annotations (green) from Alan Walker's biography of Franz Liszt in the Codex interface, Walker 1983. [Neill / Kuczera 2019]

[24] It is planned to extend the system with a feature to import ontologies – such as OWL ontologies – into the concept subgraph to save the user the labour of entering them manually.
[25] A *project* in Codex is a collection of texts that belong together. As a practical example, the Diary of Luca Landucci (del Badia 1883) and the collected letters of Michelangelo (Carden 1913) are two projects in Codex. The project domain is the set of all texts and their associated entities (e.g., agents, claims, concepts, etc.) in the project. Because more than one project can share the same graph database, associated entities across project domains may overlap. For example, the ›Landucci Diary‹ and ›Michelangelo letters‹ projects have some agents in common, such as Lorenzo de' Medici and his family, various 15th century popes, Italian cities, etc. If overlap is not desired, associated entities *can* be segregated – but in our experience project domain overlap is one of the most exciting features of Codex as it can lead to the discovery of obscure connections between domains. Of course, the easiest way to segregate project domains is simply to reserve one Neo4j database instance per project.

2.2.6 Data Sets and Data Points

A data point in Codex is defined as a quantity with a unit of measurement that is attributable to a place and a time. A data set is a collection of data points. In the example below, the statement that »three people fell dead on this day« translates to a data-point where the value is »3«, the unit of measure is »people«, the place is »Florence«, and the time is »April 23rd, 1483« (the date of Landucci's diary entry).[26]

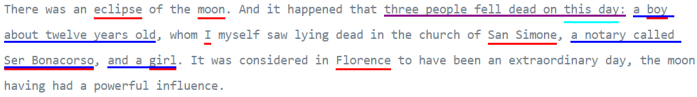


Fig. 8: The data point annotation (dark purple underline) indicates a statement that can be represented as an independent numerical quantity in the Codex interface. [Neill / Kuczera 2019]

The idea of a data point is to enable the editor to extract numerical data from a text that may be of statistical interest.[27] As indicated in the above Figure 8, a data point annotation links the text to a data point entity.

Some practical examples of data sets that can be extracted from historical sources include epidemiological data, weather records, census figures, crime statistics, etc.

2.2.7 Times

A time entity in Codex is a representation of a date and time in various degrees of precision. The entity is composed of nine components, which are all optional. Options for c. include ›on‹, ›before‹, ›after‹, and ›circa‹; options for Section include ›early‹ and ›late‹; and options for Season include ›Winter‹, ›Summer‹, ›Autumn‹, ›Spring‹. The variety of options is meant to reflect the realities of date representation in historical texts. (We intend to review the W3C Time Ontology for guidance on refining this model.)

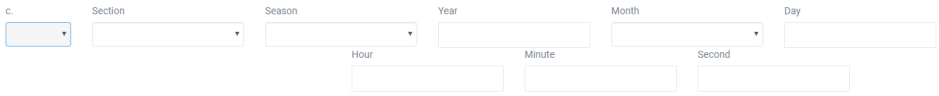


Fig. 9: The data entry modal window for a time entity in the Codex interface. [Neill / Kuczera 2019]

[26] del Badia 1883, p. 37.
[27] Data points can also store text values as well as numbers, so they can be used to represent state values. An example would be states of weather, such as rain, snow, floods, etc.

A time annotation is applied to any part of the text with an identifiable date / time, even if the text does not state a numerical date. For example, in Figure 8 above the time annotation (blue underline) links the text »this day« to the stated date of the diary entry (April 23rd, 1483).

Now that an overview of the main annotation types in Codex has been given, it is necessary to examine the standoff properties model as it forms the basis of Codex's approach to text-as-graph.

3. Standoff Properties

Aside from markup formats, word chains present another approach to annotation. A word chain is a graph model of a text where each word is treated as a token node and structure is indicated by relating each node to its next sibling in the sentence. Lexical and presentational annotations can also be linked to token nodes to model the structure of paragraphs and larger units.

Fig. 10: Text as a chain of word nodes in the Neo4j browser. [Neill / Kuczera 2019]

Like standoff properties, word chains represent a markup-free alternative to XML document formats, and offer a solution to the overlapping annotations problem. However, at present updates to word chains are managed through graph database queries, which requires some programming expertise. The Codex standoff property editor offers a trade-off between the multidimensional affordances of the graph and the technical simplicity, endurance, and sustainability of the text stream. Another distinction between word chains and standoff properties is that word chains take the word as the smallest token, which poses challenges for annotations inside of words (let alone how one chooses to define word boundaries).

The simplest solution from an annotation standpoint is conceivably to treat the *character* and not the *word* as the smallest token unit; however, managing a chain of character nodes as a graph data-structure would be exponentially more unwieldy than a chain of word nodes. However, this assumes that the characters themselves need to be represented as nodes; in fact, what defines an annotation is that it is a region of text with a certain intention (whether stylistic, presentational, semantic, etc.). If one moves away from the token node concept to an annotation node concept, then the text of the document can be stored in a plain text format (sans markup) and annotations can annotate the text by using *start* and *end* character indexes.

The removal of embedded markup makes the text stream easily readable to both humans and machines. It also solves the overlapping annotation problem because the properties are stored apart from the text, and not subject to hierarchical encoding conflicts. Multiple properties can refer to the same regions of text – or overlapping regions – by virtue of the start and end character indexes. Standoff properties are inherently discrete objects which

coexist in a ›flat‹ hierarchy, that is to say, with no imposed hierarchy at all. If a standoff property references a linked entity in the database, it can be easily connected via an edge, allowing full traceability from entities to the regions of text they are referenced by.

A standoff property, then, is essentially a data structure (tuple) that models the following attributes:

1. Type. A string representing the name (i.e., the type) of the annotation.
2. StartIndex. An integer representing the index position of the first character of the annotation: *0* <= *x* < *n*, where *x* is the index and *n* is the length of the text.
3. EndIndex. An integer representing the index position of the last character of the annotation within the length (same rule as the StartIndex).

In practise one would wish to extend this with a fourth attribute:

4. Value. A string representing data specific to the annotation, such as the unique identifier of a referenced entity, or alternatively a colour value, text size, font, etc.

At this point, one might object that standoff properties are not practical in the context of a text editor due to the likelihood of ›breaking‹ standoff property indexes upon text changes: deleting or adding a single character would necessitate the real-time recalculation of potentially every standoff property index in memory. Although this may not even be an issue in practise with texts of a certain size and number of standoff properties (given the efficiency of JavaScript interpreters in modern browsers), the Codex standoff property editor takes an approach that effectively sidesteps this issue. Text in the Codex editor is represented as a NodeList of HTML SPAN elements connected via reference pointers to an array of standoff property JavaScript objects.[28] Because a linked list is used to represent the text,[29] and because the standoff properties are linked to the text with pointers rather than indexes, characters can be freely added to or removed from the text without the need to recalculate indexes while editing.[30] Only when the user is ready to export the data from the editor (e.g., for saving) are the dynamic node pointers converted to static index numbers. As this step happens *after* all changes have been made to the text in the editing session, there is no danger of the indexes being out of alignment. The plain text and properties are exported as a JSON object which can be saved to a text file or converted to a data storage format of the user's choice. Codex translates the JSON export into a standoff property graph, linked to entity nodes as directed by the annotation type. These standoff property nodes can be converted to various formats such as text as a chain of word nodes.[31]

[28]The JavaScript NodeList data type combines a linked list with a tree structure, in that each element is connected to its previous and next siblings, and also connected to its children and parent nodes.
[29]Incidentally, this is basically the same principle as using a chain of character-token nodes to model a text in a graph database, except that it is done in memory in JavaScript in a web application.
[30]Deletions are a special case where the deleted text selection traverses the start or the end of a standoff property text range. In this case, affected properties are updated to ensure that the start and end pointers refer to the correct (remaining) characters.
[31]Because standoff properties are equivalent to character nodes in resolution, there is no difficulty in converting them to word nodes, which are lower in resolution.

Building on our technical definition of a standoff property, Codex extends this model further to include attributes that aid with database integration.

5. GUID. A 32-character string functioning as a unique identifier of the standoff property. This is required for saving the property to the database.
6. UserGUID. A GUID (see above) representing the user who created the annotation.
7. Index. An optional integer representing the order in which the standoff property was created.
8. Text. An optional string representing the source text referred to by the annotation. This is an optional attribute to make it easier to view standoff properties in the database. The StartIndex and EndIndex attributes are the source of truth with respect to the location of the annotation in the text.
9. Layer. An optional string representing arbitrary groups (layers) that the annotation is assigned to. For example, if the user wanted to group several agent annotations referencing artists, they could assign the annotations to an ›artist‹ layer. This grouping could be used for filtering either in the database, or in the editor itself.
10. IsZeroPoint: A boolean value indicating whether the standoff property is a zero-point annotation; that is, an annotation that refers to an invisible index point in the text. This can be thought of as an annotation that refers to the space between two characters.
11. IsDeleted: A boolean value indicating whether the standoff property has been marked as deleted.

To give a real-world example, below are images of a text in the Codex editor as well as a representation of a portion of it in as a JSON export.

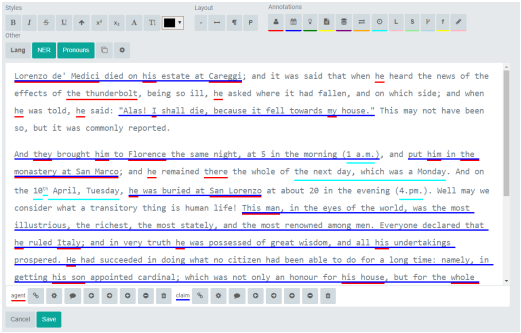


Fig. 11: Luca Landucci's diary entry for April 8th, 1492 on the death of Lorenzo de' Medici in the Codex interface, del Badia 1883, p. 53–54. [Neill / Kuczera 2019]

Following is an extract from the JSON export of the above text, with yellow highlights showing the typical parts of the standoff property data structure. The green parts show which text the annotation covers. The blue part shows the text itself.

{

"text": "Lorenzo de' Medici died on his estate at Careggi; and it was said that when he heard the news of the effects of the thunderbolt, being so ill, he asked where it had fallen, and on which side; ... ",

"properties": [{

"index": 23,

"guid": "cde24f38-81cb-4110-b368-b5b1f4ed4d53",

"type": "agent",

"layer": null,

"text": "Lorenzo de' Medici",

"value": "e45fed44-17a0-4c2c-9c00-858667a17904",

"startIndex": 0,

"endIndex": 17,

"isZeroPoint": false,

"isDeleted": false,

"userGuid": "fb067f75-a121-47c1-8767-99271c75cfc0"

}, {

"index": 25,

"guid": "5e33d2a8-dea0-4bbf-b4f6-8ccf40dac4d9",

"type": "agent",

"layer": null,

"text": "Careggi",

"value": "d8eda97c-79d7-43ad-b43f-fd3e3a05c68e",

"startIndex": 41,

"endIndex": 47,

"isZeroPoint": false,

"isDeleted": false,

"userGuid": "fb067f75-a121-47c1-8767-99271c75cfc0"

}, {

"index": 46,

"guid": "94eebe67-0136-4f54-a4c6-6e7072dfb3de",

"type": "claim",

"layer": null,

"text": "Lorenzo de' Medici died on his estate at Careggi",

"value": "175742e2-a342-455d-a1b9-3fe3a96e3fd9",

"startIndex": 0,

"endIndex": 47,

"isZeroPoint": false,

"isDeleted": false,

"userGuid": "fb067f75-a121-47c1-8767-99271c75cfc0"

}]

}

4. The Modelling of Doubt

Before proceeding to reflect on the possibilities of deep integration between text and data, it is important to review how the above discussion bears on the subject of the January 2018 graph-conference, »The modelling of doubts«,[32] hosted by the Akademie der Wissenschaften und der Literatur, Mainz. Although Codex wasn't designed with the intention of modelling doubt in the strict sense of quantifying it, it can be said that it aims at *modelling interpretation* in the following ways:

1. The ability to freely overlap annotations means that the same text regions can be annotated multiple times, allowing multiple interpretations to be captured. For instance, if two editors disagree about the identity of a person in the text, they could *each* add their own agent annotation to the same text region as overlaps are permitted;
2. Comments can be added to *annotations themselves*, enabling editorial discussions about the annotation validity to be recorded;
3. The ability to annotate agent properties, event claims, and meta-relations leads to *full transparency* around statements that are often just presented as established fact. For example, rather than simply accepting as fact the claim that Lorenzo de' Medici died at Careggi on April 8th, 1492, the claim annotation in Codex is traceable back to the precise section of the Luca Landucci text it occurs in.

5. Implications

The full potential of the standoff property model on the integration between text and graph data structures has yet to be documented. Aside from the convenience of plain text free of markup, of overlapping annotations, and of annotations whose sources are traceable back to precise text ranges, there are two features of standoff properties that suggest possibilities for computational analysis:

1. Standoff properties can be *grouped into layers*, where a layer is defined either implicitly by the annotation type or explicitly by a value stored in the Layer attribute;[33]
2. The StartIndex and EndIndex attributes offer the possibility of *combining annotations* that are contained by or overlap the same text range.

[32] »Die Modellierung des Zweifels« – Schlüsselideen und -konzepte zur graphbasierten Modellierung von Unsicherheiten am 19. und 20. Januar 2018 in der Akademie der Wissenschaften und der Literatur, Mainz.

[33] Incidentally, these layers could be exported into the standoff property JSON format, or a module could be written to export them into an XML format of choice. It should also be technically possible to convert a Codex standoff property document into an XML format such as TEI-XML, but this has not been actively developed yet.

These qualities of layering and combination have already led to useful features in the Codex editor (such as modal windows for managing named-entity and pronoun annotation candidates), but they offer computational insights as well. It is trivial, for example, to write a Cypher query to find all annotations in a text (or in all texts) that overlap each other in various ways. Overlapping annotations have the potential to enrich the text they annotate with their combined meanings. One approach that we are exploring in Codex is the comparison of manually-entered annotations (such as agents and event claims) with machine-generated syntaxical annotations, providing a natural language analysis of texts. This is an example of how standoff properties can support a multidimensional analysis of the text, allowing human-generated and machine-generated annotations to exist together.

6. Conclusion

The Codex uses standoff properties to integrate annotations with text in a graph database. Annotations map back to text at the character level and can be overlapped without constraint. The ability to overlap and comment annotations offers a convenient system for capturing discussions around doubt. The variety of semantic annotations – including event claims, meta-relations, agent properties, etc. – leads to a system where ›factual data‹ can be easily traced back to its text sources. Beyond the modelling of interpretations, Codex seeks to enable project editors to build an ›atlas of relations‹ from their source texts, integrating graph entities with text on a character level. The intended result is not so much a marked-up document (although this is a given) but a graph dataset with ›deep roots‹ in its constituent source texts, mediated through layers of standoff property annotations. Codex's real-time standoff property editor and modal-window entity management system are tools that we hope will assist editors in exploring the connections between structured data and text in their own digital editions.

Bibliographic References

Luca Landucci: Diario fiorentino dal 1450 al 1516. Continuato da un anonimo fino al 1542. Pubblicato sui codici della Comunale di Siena e della Marucelliana. Publ. by Iodoco del Badia. Florenz 1883. [Nachweis im GBV]

Luca Landucci: A Florentine Diary from 1450 to 1516. Publ. by Iodoco del Badia. Transl. by Alice de Rosen Jervis. New York 1927.

Michelangelo Buonarroti: Michelangelo: A Record of His Life as Told in His Own Letters and Papers. Publ. by Robert Walter Carden. London 1913.

Desmond Allan Schmidt (2016a): Using standoff properties for marking-up historical documents in the humanities. In: it – Information Technology 58 (2016), H. 2, S. 63–69. DOI: 10.1515/itit-2015-0030

Desmond Allan Schmidt (2016b): Standoff properties as an alternative to XML for digital historical editions. 2016. PDF. [online]

Alan Walker: Franz Liszt. 3 Vol. New York 1983. Vol. 1: The virtuoso years: 1811–1847. [Nachweis im GBV]

List of Figures with Captions

ZfdG

Zeitschrift für digitale Geisteswissenschaften

Beitrag aus:
Sonderband 4 der ZfdG: Die Modellierung des Zweifels – Schlüsselideen und -konzepte zur graphbasierten Modellierung von Unsicherheiten. Hg. von Andreas Kuczera, Thorsten Wübbena und Thomas Kollatz. 2019. DOI: 10.17175/sb004

Titel:
Blockchain – die etwas andere Datenbank

Autor/in:
Katarina Adam

Kontakt:
katarina.adam@htw-berlin.de

Institution:
Hochschule für Technik und Wirtschaft Berlin

GND:
1082051918

DOI des Artikels:
10.17175/sb004_009

Nachweis im OPAC der Herzog August Bibliothek:
1037074483

Erstveröffentlichung:
24.04.2019

Lizenz:

Sofern nicht anders angegeben (cc) BY-SA

Medienlizenzen:
Medienrechte liegen bei den Autoren

Letzte Überprüfung aller Verweise:
24.04.2019

GND-Verschlagwortung:
Blockchain | Netzwerkdatenbanksystem | Transaktion |

Zitierweise:
Katarina Adam: Blockchain – die etwas andere Datenbank. In: Die Modellierung des Zweifels – Schlüsselideen und -konzepte zur graphbasierten Modellierung von Unsicherheiten. Hg. von Andreas Kuczera, Thorsten Wübbena und Thomas Kollatz. 2019 (= Sonderband der Zeitschrift für digitale Geisteswissenschaften, 4). PDF Format ohne Paginierung. Als text/html abrufbar unter DOI: 10.17175/sb004_009.

Katarina Adam

Blockchain – die etwas andere Datenbank

Abstracts

Trotz vieler Artikel, Berichte und Aufmerksamkeit in den Massenmedien ist die Blockchain-Technologie für die meisten verbunden mit Begriffen wie Bitcoin, Risiko, Spekulation, hohe Volatilität und auch der dunklen Seite des Internets. Das Potential der Technologie wird bei diesen teilweise emotional geführten Diskussionen außer Acht gelassen. Um einen kleinen Beitrag zur Versachlichung zu leisten, ist dieser nachfolgende Aufsatz geschrieben. Es wird erläutert, in welchem Umfeld diese Technologie implementiert wurde, welcher geniale Ansatz hinter der Technologie steht, wie der aktuelle Entwicklungsstand ist und wie die Blockchain Technologie von anderen Ansätzen der Datenspeicherung und -bearbeitung lernen kann. Hierzu wird die Graphdatenbank herangezogen und es werden beide Ansätze verglichen.

Despite many articles, reports and attention in the meantime in the mass media, blockchain technology is for most associated with bitcoin, risk, speculation, high volatility and also the dark side of the Internet. The potential of the technology is ignored in these sometimes emotional discussions. In order to make a small contribution to objectification, this following essay is written. It explains the environment in which this technology has been implemented, the ingenious approach behind the technology, the current state of development and how the blockchain technology can learn from other approaches to data storage and processing. For this the graph database is used and both approaches are compared.

1. Siegeszug der Blockchain-Technologie!?

Spätestens mit dem Siegeszug der wohl berühmtesten Krypto-Währung Bitcoin hat die dahinter liegende Blockchain-Technologie den sogenannten Mainstream erreicht. Fast tagtäglich überbieten sich die Medien darin um, mal mit Schreckens-, mal mit Erfolgsgeschichten, über diese Technologie zu schreiben. Blockchain-Technologie gehört als ein Bestandteil zu dem großen Thema Digitalisierung. Und unbestritten ist das dieser Technologie zugesprochene disruptive Potential. Bis sich dieses Potenzial jedoch in Gänze entfalten kann, wird es noch einige Zeit dauern. Ein Grund mehr, sich mit dem Potenzial und den vielfältigen Anwendungsmöglichkeiten auseinanderzusetzen und auch den Vergleich zu anderen Datenbanken nicht zu scheuen. Es kann und darf lebhaft diskutiert werden, ob es sich mit der Blockchain-Technologie um einen Paradigmenwechsel handelt.

Nachfolgend wird zunächst darauf eingegangen, in welchem Umfeld diese Technologie entstanden ist und was die Einzigartigkeit der Erfindung Blockchain ausmacht. Anschließend werden die Eigenschaften der Technologie am Beispiel der ersten Applikation, dem Bitcoin, dargelegt. Die technische Weiterentwicklung macht auch vor der Blockchain-Technologie nicht halt. Daher werden die Ansätze der nachfolgenden Generationen besprochen. Die Blockchain-Technologie ist ein sehr mächtiges Werkzeug, dessen Anwendung mit einigen Beispielen illustriert wird, um die extreme Bandbreite der Anwendungsmöglichkeiten und damit das Potential dieser Technologie erahnen zu können. Die berechtigte Frage, warum

sich weltweit die Projekte noch immer in bestenfalls der Beta-Testing Phase befinden, wird im Fazit beleuchtet. Zuvor wird jedoch der Vergleich zwischen graphbasierten Datenbanken und blockchainbasierten Datenbanken gewagt.

2. Volkswirtschaftliche Retroperspektive – oder, wie alles begann

Um die Blockchain-Technologie richtig einordnen zu können, ist es hilfreich, sich noch einmal zurückzuversetzen in das Jahr 2008, als die Finanzkrise, die ihren Ursprung in den USA nahm, wie ein Flächenbrand um den Globus zog. Ökonomen ziehen hier gern die Parallele zur Großen Depression Ende der 20er Jahre des letzten Jahrhunderts.[1]

Vor der großen Finanzkrise gab es auf dem amerikanischen Häusermarkt eine Flut von – im Nachhinein unverantwortlichen – Hypothekenkrediten, begleitet von einem Versagen der Finanzaufsicht und -regulierung.

Um das Ausmaß der Krise zu verstehen, bedarf es eines weiteren Schritts zurück in die Vergangenheit. Das 9/11-Attentat löste eine Schwemme an Geldern aus, um sowohl die Wunden aus den Attacken als auch denen aus dem Platzen der Dotcom-Blase Anfang der 2000er Jahre zu überwinden. Amerika befand sich in einer Rezession und der damalige Vorsitzende der amerikanischen Zentralbank (FED), Alan Greenspan, entschied sich aus Furcht vor einer Deflation zur Politik des billigen Geldes. Dieses billige Geld ermöglichte Geringverdienern (sog. Sub-Prime-Kreditnehmer), sich Wohnungseigentum anzuschaffen. Die kreditbasierte Nachfrage nach Immobilien war in den Folgejahren des Attentats ungezügelt und Häuserpreise schossen in die Höhe.[2] Durch geschickte Bündelung der Kredite (sogenanntes Pooling) entstanden wiederum auf Seiten von Banken völlig neuartige und bestenfalls bedingt einer Regulierung unterworfene Investitions-Produkte. Produkte, die vielfach nicht mehr von den eigentlichen Bankern verstanden wurden, weil algorithmusgetriebene Berechnungen neue Standards und Methoden ermöglicht hatten.[3] Grob gesagt führte die Annahme, dass die Häuserpreise so schnell nicht sinken können, die Niedrigzinspolitik, die den Hunger der Investoren auf Rendite in immer risikoreichere Segmente führte, das Pooling von verschiedenen Krediten sowie das Versagen der Regulierungs- und Aufsichtsbehörden zum Zusammenbruch des Finanzmarktes.[4] Dieses eigentlich lokal auf die USA begrenzte Phänomen konnte sich deshalb in einer Art Dominoeffekt um den Globus bewegen, weil die neu geschaffenen Finanzprodukte weltweit an sowohl institutionelle als auch vermögende Einzelinvestoren veräußert worden waren. Und damit ist die amerikanische Sub-Prime-Krise verantwortlich für eine bis heute anhaltende Finanzkrise, die im Herbst 2008 als prominentes Opfer die Investmentbank Lehman Brothers in die Insolvenz trieb.

[1] Vgl. Sinn 2010, S. 19.
[2] Vgl. Mallaby 2016, S. 596, 597.
[3] Vgl. Bloss et al. 2009, z. B. S. 69 ff.
[4] Vgl. Bloss et al. 2009, z. B. S. 69 ff.

Zu diesem Zeitpunkt mussten weltweit die Zentralbanken intervenieren und in die Märkte eingreifen, um den völligen Kollaps zu verhindern. Nicht verhindern konnten die Zentralbanken, dass zuvor Gewinne privatisiert wurden, nun aber in der Krisenzeit Steuergelder herhalten mussten, um das Bankensystem und damit auch die Wirtschaft am Leben zu erhalten.

In diesem wirtschaftlichen Umfeld ist erstmalig die Blockchain-Technologie auf den Markt gekommen. Die Intention des Erfinders Satoshi Nakamoto (bis heute ist ungeklärt, wer hinter dem Pseudonym steht) war und ist die Abschaffung von Mittelsmännern, die ein System wie das Bankensystem verlangsamen, verteuern und ineffizient halten. Mit einem sogenannten Peer-to-Peer-Netzwerk (p2p) sind diese Mittelsmänner nicht mehr nötig, Zahlungen können digital ebenso abgewickelt werden, wie es bei Barzahlungen mit den herkömmlichen Währungen, z. B. Dollar, Euro, Yen (sogenanntes Fiat-Geld)[5] möglich ist.[6]

Dabei ist die Schaffung einer digitalen Währung kein neuer Ansatz, denn bereits in den 90er-Jahren haben u. a. Nick Szabo (Bit Gold, 1998) und Wei Dai (B-Money, 1998) die ersten Ansätze zu digitalen Währungen veröffentlicht:

Kurz nachdem das B-Money-Whitepaper veröffentlicht wurde, startete Nick Szabo ein sehr ähnliches Projekt namens Bit Gold. Dieses elektronische Währungssystem hatte sein eigenes Proof-of-Work-System, das sich nicht allzu sehr von dem unterscheidet, wie Bitcoin heute geprägt wird. Alle Lösungen wurden kryptographisch zusammengestellt und für die Öffentlichkeit zugänglich gemacht, ähnlich dem Blockchainansatz, den wir heute kennen. Noch wichtiger ist, dass Bit Gold das erste Konzept war, das sich von den zentralisierten Behörden entfernt hat, um doppelte Ausgaben (Double Spending-Problem) der Währung zu vermeiden. Es ist einfach, digital Vorliegendes zu kopieren und ein weiteres Mal zu verwenden; dies kann ein Dokument, ein Song, ein Foto und auch eine digitale Währungseinheit sein, die kopiert und ein weiteres Mal genutzt wird. Dies gilt es jedoch zu verhindern, um Vertrauen in eine digitale Währung zu schaffen. Betrugspräventionen und Fälschungssicherheit ermöglichen dieses Vertrauen, das durch die gesellschaftlich-staatliche Ordnung der herkömmlichen Fiat-Währungen und die dahinterliegende Unabhängigkeit der Geld schaffenden Zentralbanken in der Realwelt vorhanden ist. [7]

Szabos Ziel war es, die Eigenschaften von Gold wiederherstellen, indem er den Zwischenhändler ganz ausschaltete. Man könnte sagen, dass Bit Gold und Bitcoin nicht allzu unterschiedlich sind, obwohl sie mehr als zehn Jahre auseinander liegen.[8]

Digicash, entworfen von David Chaum, war die erste Art von elektronischem Geld, die durch die Verwendung von kryptographischen Protokollen Anonymität bot. Zur Zeit der Entwicklung war Digicash ein revolutionäres Konzept. Durch die Verwendung von

[5] Fiat ; lateinisch für ›*es werde*‹, Fiat-Währung sind Währungen ohne intrinsischen Wert.
[6] Vgl. Nakamoto 2008, S. 1.
[7] Vgl. Chohan 2017, S. 1
[8] Vgl. Szabo 2005.

Kryptographie mit öffentlichem und privatem Schlüssel konnte jeder seine eigene Bank werden und seine Gelder ohne Aufsicht durch Dritte kontrollieren. Ausgegebene Zahlungen wären für Banken und Regierungen nicht nachvollziehbar gewesen. Leider hat das Unternehmen hinter Digicash 1998 Konkurs angemeldet und ist letztlich 2002 verkauft worden. Es ist offensichtlich, dass Digicash ein spannendes Konzept war, aber eben seiner Zeit weit voraus.[9]

All diese Arbeiten an digitaler Währung haben dazu beigetragen, dass der Bitcoin entworfen werden konnte. Mit Hilfe eines Proof-of-Work-Algorithmus für die Generierung und Verteilung neuer Münzen sind viele Funktionen der Vorgänger in das von Satoshi Nakamoto entwickelte Bitcoin-Protokoll eingeflossen. Was die Besonderheit des Ansatzes von Satoshi Nakamoto aus- und damit überlebensfähig macht, ist zum einen die Lösung von Problemen des Double Spending und zum anderen die Vorgabe, nur valide Transaktionen, die die Zustimmung der Mehrheit haben, auf der Blockchain zu speichern.

Da die Blockchain-Technologie nicht mehr nur als interessanter Anwendungsfall für digitale Währungen verstanden wird, und mittlerweile jede Industrie sich bemüht zu verstehen, wie sie mittels dieser Technologie Prozesse verschlanken und vereinfachen kann, wird in diesem Zusammenhang auch von der Distributed-Ledger-Technologie (DLT) gesprochen. Beide Begriffe werden gern synonym verwendet und in beiden Fällen handelt es sich um eine Art Kassenbuch, in dem sämtliche Transaktionen festgehalten werden.

Nachfolgend werden zunächst die Eigenschaften der Bitcoin-Blockchain erläutert. Das Hauptaugenmerk der Bitcoin-Blockchain gilt dem Finanzsektor und den dazugehörigen Finanztransaktionen. Die darauf aufbauenden Versionen oder Generationen adressieren weiterführende Konzepte.

3. Eigenschaften einer Blockchain

> » *Die* Blockchain ist zunächst eine große, jedoch dezentrale Datenbank. Zu den Besonderheiten, die die Technik bzw. Datenbank auszeichnet, gehört, dass eine valide Transaktion nicht rückgängig gemacht werden kann, sämtliche Transaktionen transparent und schnell sind, keine zentrale Instanz benötigt wird und alle Transaktionen durch kryptografische Verschlüsselung sicher sind. Bei den Transaktionen, die mittels Blockchain abgewickelt werden können, kann es sich praktisch um jede Form der Übertragung handeln.«[10]

Zu den Details: Viele der Komponenten der Bitcoin-Blockchain existierten schon vorher (s. o. digitale Währung), doch mit der Bitcoin-Blockchain wurden die relevanten Schwachstellen überwunden. Digitales lässt sich einfach kopieren und immer wieder verwenden. Ein fataler

[9] Vgl. Anonymus 1999.
[10] Vgl. Adam 2017, S. 74 ff.

Fehler, wenn es sich um eine digitale Währung handeln soll, denn genau wie eine Fiat-Währung soll die digitale Währung die Tausch-, Rechen- und Aufbewahrungsfunktion erfüllen, um Verbreitung zu erlangen. Diesen Anforderungen genügt der Bitcoin im gewissen Maße.[11]

Warum diese Funktionen trotz aller Einschränkungen so gut sind, liegt vielleicht auch daran, dass es Satoshi Nakamoto u. a. mit seinem Bitcoin-Protokoll gelang, das in der IT-Welt bekannte Phänomen der Byzantine Fault Tolerance zu lösen. Das Bitcoin-Protokoll basiert auf Hashcash für den Proof-of-Work (PoW),[12] Byzantine Fault Tolerance für verteilte Netzwerke sowie der Blockchain als Kette miteinander verketteter Blöcke.

Adam Back empfahl im Mai 1997 Hashcash als Mechanismus, um den systematischen Missbrauch von nicht gemessenen Internet-Ressourcen wie E-Mail-Spams zu drosseln. Dieser Mechanismus basiert auf der Idee, dass jeder, der eine E-Mail verschickt, zuvor in eine Rechenleistung investieren und darüber einen Nachweis ablegen muss (Proof-of-Work). Nur wenn dies erfolgt ist, wird die Mail vom Empfänger akzeptiert. Die zu erbringende Rechenleistung ist beim Versenden einer Mail für den Absender kaum merkbar. Bei Massen- und Spam-Mails kann dieser Mechanismus jedoch unerwünschte Zeitverzögerung mit sich bringen. Auch wenn sich dieses Konzept für E-Mails nie so recht durchgesetzt hat, ist dieser Proof-of-Work beim Bitcoin wichtige Voraussetzung des sogenannten Minings, bei dem die Miner die Transaktion sowohl verifizieren als auch neue Coins schürfen [13] und diese in den Umlauf bringen.

Hashcoin verwendet ebenso wie die Blockchain kryptografische Hash-Funktionen,[14] Hashfunktionen wandeln eine Information beliebiger Länge in einen Hashwert[15] mit festgelegter Länge um (siehe z. B. Abb. 1).

Abb. 1: Umwandlung eines Textes in einen hexadezimalen Hashwert. [Katarina Adam 2019]

[11] Zwar erfüllt der Bitcoin alle Funktionen, jedoch ist er höchst volatil und damit ist die Aufbewahrungsfunktion eingeschränkt. Auch nehmen Händler diese Währung nur bedingt im Tausch gegen Waren an.
[12] Vgl. Back 2002.
[13] Beim Mining liegt die kryptografische Aufgabe darin, mit einer zweifachen SHA-256-Berechnung einen Hash-Wert zu finden, der unterhalb eines gewissen Grenzwertes liegt.
[14] Vgl. Güting / Dieker 2018, S. 128-129.
[15] Oft wird der Hashwert als eine hexadezimale Zeichenkette codiert, d.h. der Hashwert besteht aus einer Kombination aus Zahlen und Buchstaben zwischen 0 und 9 sowie A bis F (als Ersatz für die Zahlen 10 bis 15).

Die Hashfunktion SHA-256 (Secure Hash Algorithm 256) funktioniert nur in eine Richtung. Sie ist somit kollisionsfrei und das bedeutet, dass zwei verschiedene Inputwerte nicht auf denselben Outputwert kommen können. Gemäß dem obigen Beispiel ›*Es war ein großartiger Sonnenuntergang*‹ mit dem Hashwert

8e412b754db9d57983d22b2a4b0ef882a1798f7ef099f42bc576b162ef9d766b

kann nicht der gleiche Hashwert für den sinnverwandten Satz ›*Der Sonnenuntergang war großartig*‹ geschaffen werden. Im zweiten Falle lautete der Hashwert

fb15586a3ac8885b47582348996064954eea1b1fb6be62a35f1164ff6a5e686.

Dieses Prinzip ist die Grundlage des Proof-of-Work-Konzepts und wird in ähnlicher Form bei sämtlichen digitalen Währungen genutzt.

Eine weitere interessante Komponente bei der Ursprungs-Blockchain ist, wie Satoshi Nakamoto das sogenannte *Byzantine General Problem* für seine Anwendung genutzt hat. In der Informationstechnik wird mit dem *byzantinischen Problem* der Fehler beschrieben, bei dem sich ein System falsch bzw. völlig unerwartet verhalten kann.[16]

Das Problem beschreibt ein Szenario, in dem sich mehrere Generäle auf einen Zeitpunkt einigen müssen, an dem sie den gemeinsamen Feind angreifen wollen. Die Schwierigkeit dabei ist, dass einer oder mehrere der Generäle ein Verräter sein könnte, was in der Konsequenz bedeutet, dass er oder sie falsche Angaben über ihr Vorgehen machen könnten. Zum Beispiel könnte sich eine Gruppe von Generälen entschließen, eine Stadt anzugreifen und alle Generäle sind mit ihren Soldaten quasi sternförmig um die Stadtmauer verteilt. Von fünf Generälen sind zwei für Angriff und zwei für Rückzug. Der fünfte General sichert sowohl den Generälen, die für den Angriff stimmen, als auch den Generälen, die für den Rückzug sind, seine Unterstützung zu. Der Angriff scheitert, da dem Angriff kein Konsens des Handelns unterliegt.

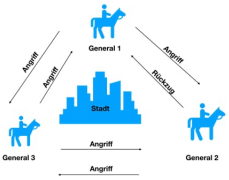


Abb. 2: Byzantine Generals Problem. [Katarina Adam 2019.]

[16] Vgl. Erstmalig wurde dieses Problem 1975 von Akkoyunly et al. 1975 aufgeworfen, detailliert von Lamport et al. 1982 beschrieben.

Die angreifenden Kräfte können die Stadt nicht erobern, weil keine zentrale Instanz das Vorhandensein von Vertrauen unter allen Generälen überprüfen kann. In diesem Szenario kann die sogenannte byzantinische Fehlertoleranz[17] dann erreicht werden, wenn die loyalen Generäle effektiv miteinander kommunizieren können, um die unbestrittene Einigung über ihre gemeinsame Strategie zu erzielen.

In der Bitcoin-Blockchain ist diese byzantinische Fehlertoleranz implementiert und das Problem wird über das Peer-to-Peer-Netzwerk (p2p) und ein Ledgersystem gelöst. Ein Algorithmus legt fest, welcher Miner den nächsten Block bestimmt und damit einigt sich das Netzwerk auf eine gemeinsame Wahrheit, die für und gegen jede*n Teilnehmer*in im Netzwerk gilt. Die Teilnehmenden an einem Netzwerk müssen sich nicht vertrauen, da alle Transaktionen transparent und nachvollziehbar sind. Dieses p2p-Netzwerk zeichnet Transaktionen nicht nur auf, sondern verifiziert diese. Ist die Mehrheit der Teilnehmenden des Netzwerkes der Ansicht, dass diese Transaktion korrekt ist, dann wird diese im Block gespeichert. Fehlerhafte Transaktionen werden identifiziert und entsprechend nicht weiter bearbeitet. Da sämtliche geprüfte und in den Blöcken gespeicherte Transaktionen nicht mehr veränderbar sind, ohne das dies auffällt (siehe veränderter Hashwert), wird Verlässlichkeit in die Transaktion gebracht. Es bedarf keines Mittelmanns mehr, der durch seine Position dafür bürgt, dass die durchgeführte Transaktion korrekt ist; das Netzwerk übernimmt diese Aufgabe.

Die Blockchain ist in erster Linie ein Aufzeichnungs-Ledger, der allen Beteiligten eine sichere und synchronisierte Aufzeichnung der Transaktionen von Anfang bis Ende bietet. Eine Blockchain kann Hunderte von Transaktionen sehr schnell aufzeichnen und nutzt hierfür mehrere kryptographische Mechanismen, um Datensicherheit zu gewährleisten.

Ähnliche Transaktionen auf der Blockchain werden zu einer funktionalen Einheit, einem Block, zusammengefasst und dann mit einem Zeitstempel (einem kryptographischen Fingerabdruck) versiegelt. Es können Daten nur in Blöcken angehängt werden, nicht jedoch wie bei herkömmlichen Datenbanken verändert oder gelöscht werden.

Die Bitcoin-Blockchain ist geschaffen worden, um Finanztransaktionen ohne Intermediäre direkt, schnell und zuverlässig abzuwickeln. Blockchains aber können viel mehr und dienen als Katalysatoren. Sie werden die Art und Weise, wie wir heutzutage Prozesse abwickeln, verändern und dies wird sich auf Regierung, Lebensweise, traditionelle Unternehmensmodelle, Gesellschaft und globale Institutionen auswirken.

[17]Die byzantinische Fehlertoleranz ist jenes Merkmal, das ein System definiert, welches die Klasse der Fehler toleriert, die zum Problem der byzantinischen Generäle gehören.

4. Blockchains der Generation 2.0

Mit der Bitcoin-Blockchain ist es möglich, finanzielle Transaktionen digital, schnell und direkt abzuwickeln. Jedoch ist es schwer, mit dieser Blockchain andere Werte zu transferieren. Dazu ist die Bitcoin-Blockchain über einen sogenannten Fork (Abzweigung) in die Ethereum-Blockchain durch u. a. Vitalik Buterin erweitert worden.

Ethereum nutzt die Grundlagen der Bitcoin-Blockchain, indem hier ebenfalls ein dezentrales Netzwerk agiert, jedoch ist Ethereum eine Plattform für sogenannte Distributed Apps (Dapps) und Distributed decentralized Organisation (DAO), die jeder Entwickler, jedes Unternehmen und / oder jede Privatperson auf der Plattform entwickeln und betreiben kann. Verteilte Knotenpunkte des Netzwerkes führen sogenannte Smart Contracts aus. Ein Smart Contract (kluger Vertrag) ist zunächst *nur* eine bindende Wenn-Dann-Abfolge. Es handelt sich hierbei um Programme, die auf einem Ledger / einer Blockchain abgelegt werden und deren Variablen-Werte durch Transaktionen geändert werden können. Smart Contracts sind jedoch nicht in der Lage, selbständig Transaktionen auszulösen. Vielmehr muss ein Benutzer-Account dieses initiieren.[18]

Mit dieser Erweiterung, nicht nur finanzielle Transaktionen über ein dezentrales Netzwerk abzubilden und abzuwickeln, ergeben sich neue Anwendungsmöglichkeiten. Zentral organisierte Systeme sind immer der Gefahr einer Attacke, eines Angriffs im Sinne einer Manipulation ihrer Daten ausgesetzt. Dezentralisierte Blockchain-Systeme hingegen legen dieses unmittelbar offen und sind in der Lage, entsprechend zu reagieren (beispielsweise, in dem die Transaktion als nicht valide in der Bearbeitung ignoriert wird).

Als vorteilhaft wird dabei angesehen, dass menschliches Versagen ausgeschlossen werden kann Anwälte nicht mehr zwingend für eine Vielzahl von Verträgen benötigt werden dieses zu Kosten- und Zeiteinsparungen führt die Verträge automatisch ausgeführt werden.

Nachteilig ist – zumindest aus juristischer Sicht – die Frage, wie eine Willenserklärung in einen Code gepackt werden kann. Eine Vielzahl an Jurist*innen beschäftigt sich daher sehr intensiv mit Fragestellungen dieser Art, um Smart Contract und das intrinsische Potential im juristischen Sinne zu kanalisieren.

Neben der erweiterten Nutzungs- / Anwendungsmöglichkeit können auf der Ethereum-Plattform ca. 15 Transaktionen pro Sekunde (TPS) durchgeführt werden. Auf der Bitcoin-Blockchain sind es dagegen nur ca. vier bis sieben Transaktionen pro Sekunde (TPS).

Auch wenn die Ethereum-Blockchain mehr als eine Verdopplung der Transaktionenabwicklungen pro Sekunde ermöglicht, reicht es nicht, um den Anforderungen der Realwirtschaft gerecht zu werden. Ziel sind bis zu einer Million TPS – und die Generation 3.0 arbeitet an entsprechenden Lösungen.

[18] Vgl. Ethereum.

5. Blockchains der Generation 3.0

Die Wirkungsweise von Blockchain-Lösungen und das große Veränderungspotential sind mittlerweile unbestritten. Was fehlt, ist der Übertrag in reale Prozessanforderungen und Prozessgeschwindigkeiten.

Weltweit arbeiten eine Vielzahl von Teams / Startups an Lösungen, um die TPS massiv zu erhöhen. Vitalik Buterin hat beispielsweise im Sommer 2018 bekräftigt, auch für die Ethereum-Plattform das Problem der Skalierbarkeit und der Netzwerküberlastung beheben zu wollen.[19]

Insgesamt geht es den auf dem Markt sichtbaren Lösungsansätzen im Prinzip darum, dass nicht alle Teilnehmer*innen des Netzwerkes jede Transaktion verifizieren müssen, sondern dass über Side-Chains oder Off-Chains verschiedene Transaktionen des Netzwerkes bearbeitet werden. Dieses *Ausgliedern* ermöglicht eine höhere Transaktionsgeschwindigkeit.[20] Viele Ansätze sind denkbar und fördern einen interdisziplinären Austausch, u. a. mit der Graphentechnologie.

6. Graphen

Netzwerke umgeben uns – nicht nur in der digitalen Welt. Beziehungen zu Familienmitgliedern, Freund*innen, Verwandten und Kolleg*innen sind ein solches (soziales) Netzwerk. In der IT existieren ebenfalls Netzwerke, gebildet aus Computern, Programmen und User*innen. Genau wie in der analogen Welt werden in der digitalen Welt Informationen ausgetauscht. Die Art des Austausches kann zentral oder dezentral organisiert sein. Die Beziehungen, die zwischen den Daten und Informationen bestehen, sind von Interesse und aus dem *richtigen* Herausfiltern der relevanten Informationen lassen sich Geschäftsmodelle kreieren.

Ein Graph stellt Objekte wie beispielsweise Knoten und Beziehungspunkte als Kanten dar und man unterscheidet sogenannte gerichtete (Darstellung einer Beziehung) und ungerichtete (symmetrische Beziehung) Graphen.[21]

»Graphdatenbanken sind dafür prädestiniert, relevante Informationsnetzwerke transaktional zu speichern und besonders schnell und effizient abzufragen. Das Datenmodell besteht aus Knoten, die mittels gerichteter, getypter Verbindungen miteinander verknüpft sind. Beide können beliebige Mengen von Attribut-Wert-Paaren (Properties) enthalten. Daher wird dieses Datenmodell auch als „ *Property-Graph* “ bezeichnet.«[22]

[19] Bei Ethereum werden das Sharding und Plasma als Möglichkeit genannt: Das Sharding erlaubt eine parallele Nutzung der Blockchain. Plasma wiederum verdichtet den Zugriff auf die Blockchain durch die Bündelung von vielen Transaktionen.
[20] Vgl. z. B. plasma, zk systems, constellation.
[21] Vgl. Güting / Dieker 2018, S. 201 ff.
[22] Vgl. Hunger 2014, S. 10.

Graphdatenbanken erzielen dank agilerer und flexiblerer Modelle vielfach eine bessere Performanz als herkömmliche, relationale Datenbanken, denn sie unterstützen systembedingt Datenstrukturen ohne sonst übliche Beschränkungen. Zusätzlich können Graphendatenbanken problemlos erweitert werden, ohne das bestehende Applikationen darunter leiden. Sie sind heute bereits häufig im Einsatz bei sozialen Netzwerken, Netz- und Cloudmanagement und auch bei der Betrugserkennung. Die Fähigkeiten von Graphdatenbanken sind daher auch für Blockchain-Entwickler*innen von hohem Interesse.

7. Blockchain vs. Graphdatenbank

Wie erläutert, ist die Blockchain eine Verkettung von Blöcken, die die getätigten und verifizierten Transaktionen enthalten. Diese Blöcke sind in ihrer Größe beschränkt (z. B. ermöglicht die Bitcoin-Blockchain maximal ein Megabyte) und damit auch in der Anzahl der Transaktionen. Weiterhin können nur Miner diese Blöcke erzeugen, was Kritiker*innen veranlasst, dieses als Machtvakuum anzusehen.

Graphen könnten eine Alternative sein, da sie diese Limitationen aufbrechen. Ledger können graphenbasiert organisiert werden. Jede*r Teilnehmer*in müsste dann schon im Netz vorhandene Transaktionen bestätigen, wenn er oder sie selbst eine Transaktion hinzufügen möchte. Blöcke wie bei der Blockchain sind dann nicht mehr nötig, sondern die Kanten des Graphen bilden die Transaktionsbestätigungen ab. Diese Transaktionsbestätigungen könnten sich anders verhalten als die bei der Blockchain. Ein graphenbasiertes Netzwerk ist ebenfalls dezentralisiert, jedoch existieren keine Rollen für das Erzeugen von Knoten.

Plattformen wie IOTA[23] oder Byteball[24] nutzen bei ihren Blockchain-Ansätzen dieses System der Graphen. Jedoch ist noch nicht sichergestellt, ob und wie dieses System sich verhält, welche Parameter benötigt und wie genutzt werden müssen, um ein stabiles und effektives Netzwerk zu haben.

Das Zusammenführen von Blockchain-Datenbanken und Graphdatenbanken ist ein vielversprechender Ansatz, um die großen Datenmengen von Heute und die der Zukunft so zu managen, dass sie die Vorteile beider Systeme vereinen und damit dezentrale, schnelle, transparente und effiziente Transaktionen ermöglichen. Aus Autorinnensicht ist es daher kein Blockchain versus Graphdatenbank, sondern vielmehr ein Miteinander.

[23] Vgl. IOTA.
[24] Vgl. Byteball White Paper Churyumov 2016, S. 4 ff.

8. Fazit / Herausforderungen

In der digitalen Ökonomie spielen Daten eine Schlüsselrolle. Wie sie gesichert, geschützt, aufbewahrt und genutzt werden, ist für den und die Einzelne*n von ebenso großer Bedeutung wie für Unternehmen. Jedoch ist die schiere Masse von Daten, die im Zuge der immer weiter fortschreitenden Digitalisierung produziert werden, nicht der eigentliche Wert. Der eigentliche Wert erschließt sich aus den Verknüpfungen von Daten.

Beide Datenbankenarten haben ihre Existenzberechtigung und wie immer gilt es zu prüfen, für welchen Anwendungszweck welches Instrument, welche Datenbank benötigt wird. Das Verschmelzen von Graphdatenbanken mit dem Ansatz der Blockchain-Technologie scheint sehr vielversprechend, denn viele ungelöste Fragen und Beschränkungen innerhalb der Blockchain-Technologie können, ergänzt um den Umgang mit Daten aus der Graphendatenbank, neu durchdacht werden.

Ein großes Problem und daher auch der Grund, warum noch keine massentaugliche Anwendung, die auf der Blockchain-Technologie basiert, aus dem Beta-Testing-Modus heraus ist, ist die Skalierbarkeit. Solange nur bis zu 15 TPS abgewickelt werden können, reicht dieses nicht aus, um in der Realwirtschaft zu bestehen. Auch, dass die gesamte Historie herunterzuladen ist, entwickelt sich bei Fortschreibung der Transaktion zu einem Hindernis. Die Bitcoin-Blockchain weist eine Größe von 179 MB auf (Stand Mitte August 2018) und allein diese Historie herunterzuladen dauert geschätzte 24 Stunden und mehr (je nach CPU). Ein weiteres, aber lösbares Hindernis ist die Interoperabilität zwischen sowohl den verschiedenen Blockchain-Arten als auch zu den klassischen (z. B.) Zahlungssystemen. Die neue Datenschutzgrundverordnung (DSGVO), die im Mai 2018 endgültig in Kraft getreten ist, stellt die Welt der Blockchains ebenfalls vor neue Herausforderungen.

Auch wenn es die (Bitcoin-)Blockchain seit gut zehn Jahren gibt, so sind noch viele Fragen unbeantwortet. Behördliche Fragen stehen ebenso im Raum wie rechtliche. Auf den ersten Blick vermag es so zu sein, dass die technischen Probleme leichter zu lösen sind, aber schon in dieser kurzen Zusammenfassung ist festzustellen, dass auch in diesem Zusammenhang noch viel Forschungs- und Entdeckungspotential vorhanden ist.

Daher, zum Vergleich: Es hat gut 20 Jahre gedauert, ehe das HTML-Protokoll das Internet-Protokoll ergänzte. Jedoch erst mit der Einführung des HTML-Protokolls schossen die Anwendungsmöglichkeiten in die Höhe und das World Wide Web wurde überhaupt erst möglich. In Bezug auf die Blockchain-Technologie stehen wir vor diesem Durchbruch. Jedoch vermag keiner zu sagen, wann er in welcher Form kommt. Sicher ist aber, der Durchbruch wird kommen!

Bibliographische Angaben

Katarina Adam: Build to last: Blockchain on the cutting edge of the real. In: Industrie von morgen. Hg. von Matthias Knaut. Berlin 2017, S. 74–82. [Nachweis im GBV]

Eralp Abdurrahim Akkoyunly / Kattamuri Ekanadham / R. V. Huber: Some Constraints and Tradeoffs in the Design of Network Communikation. In: Proceedings of the fifth ACM symposium on Operating systems principles. (SOSP: 75, Austin, TX, 19.–21.11.1975) New York, NY 1975, pp. 67–74. [Nachweis im GBV]

Anonymus: How DigiCash Blew Everything. Hg. von Ian Grigg. In: cryptome.org. Beitrag vom 10.02.1999. [online]

Adam Back: Hashcash – A Denial of Service Counter Measure. 2002. PDF. [online]

Michael Bloss / Dietmar Ernst / Joachim Häcker / Nadine Eil: Von der Subprime-Krise zur Finanzkrise. München 2009. [Nachweis im GBV]

Usman Waqqas Chohan: The Double-Spending Problem and Cryptocurrencies, Discussion Paper. In: SSRN Electronic Journal (2017). Artikel vom 19.12.2017. DOI: 10.2139/ssrn.3090174

Anton Churyumov: Byteball: A Decentralised System for Storage and Transfer of Value. 2016. PDF. [online]

Henning Diedrich: Ethereum. Blockchains, digital assets, smart contracts, decentralized autonomous organizations. Lexington, KY 2016. [Nachweis im GBV]

Introduction to Smart Contracts. Hg. von Ethereum. Version v0.4.24. In: solidity.readthedocs.io. 2016–2018. [online]

Ralf Hartmut Güting / Stefan Dieker: Datenstrukturen und Algorithmen. 4., erweiterte und überarbeitete Auflage. Wiesbaden 2018. [Nachweis im GBV]

Michael Hunger: Neo4j 2.0. Eine Graphendatenbank für alle. Frankfurt / Main 2014. [Nachweis im GBV]

What is IOTA? A permissionless distributed ledger for a new economy. Hg. von IOTA Foundation. In: iota.org. Berlin 2018. [online]

Leslie Lamport / Robert Shostak / Marshall Pease: The Byzantine Generals Problem. In: Transactions on programming languages and systems 4 (1982), no. 3, pp. 382–401. [Nachweis im GBV]

Sebastian Mallaby: The Man Who Knew. The Life & Times of Alan Greenspan. London u. a. 2016. [Nachweis im GBV]

Satoshi Nakamoto: Bitcoin: A Peer-to-Peer Electronic Cash System. 2008. PDF. [online]

Hans-Werner Sinn: Kasino-Kapitalismus. Wie es zur Finanzkrise kam. Berlin 2010. [Nachweis im GBV]

Nick Szabo: Bit Gold. In: nakamotoinstitute.org. Hg. von Satoshi Nakamoto Institute. Beitrag vom 29.12.2005. [online]

Abbildungslegenden und -nachweise

Abb. 1: Umwandlung eines Textes in einen hexadezimalen Hashwert. [Katarina Adam 2019.]

Abb. 2: Byzantine Generals Problem. [Katarina Adam 2019.]

ZfdG

Zeitschrift für digitale Geisteswissenschaften

Beitrag aus:
Sonderband 4 der ZfdG: Die Modellierung des Zweifels – Schlüsselideen und -konzepte zur graphbasierten Modellierung von Unsicherheiten. Hg. von Andreas Kuczera, Thorsten Wübbena und Thomas Kollatz. 2019. DOI: 10.17175/sb04

Titel:
A Graph Database of Aegean Seals with Uncertain Attributes

Autor/in:
Martina Trognitz

Kontakt:
martina.trognitz@oeaw.ac.at

Institution:
Österreichische Akademie der Wissenschaften, Austrian Centre for Digital Humanities

GND:
116996270X

ORCID:
0000-0003-0485-6861

DOI des Artikels:
10.17175/sb004_010

Nachweis im OPAC der Herzog August Bibliothek:
1037074726

Erstveröffentlichung:
10.04.2019

Lizenz:

Sofern nicht anders angegeben (cc) BY-SA

Medienlizenzen:
Medienrechte liegen bei den Autoren

Letzte Überprüfung aller Verweise:
09.04.2019

GND-Verschlagwortung:
konzeptionelle Modellierung | Graphdatenbank | Netzwerkanalyse | Ägäische Kultur | Siegel |

Zitierweise:
Martina Trognitz: A Graph Database of Aegean Seals with Uncertain Attributes. In: Die Modellierung des Zweifels – Schlüsselideen und -konzepte zur graphbasierten Modellierung von Unsicherheiten. Hg. von Andreas Kuczera, Thorsten Wübbena und Thomas Kollatz. 2019 (= Sonderband der Zeitschrift für digitale Geisteswissenschaften, 4). PDF Format ohne Paginierung. Als text/html abrufbar unter DOI: 10.17175/sb004_010.

Martina Trognitz

A Graph Database of Aegean Seals with Uncertain Attributes

Abstracts

Um die Anwendung von Netzwerkanalyse bei der Untersuchung von mehrseitigen ägäischen Siegeln zu ermöglichen, wurde eine Graphdatenbank in Neo4j erstellt. Viele der die Siegel beschreibenden Attribute enthalten unsichere Werte, die besonderer Aufmerksamkeit bedürfen, um sie in die Graphdatenbank zu integrieren. Der Aufsatz untersucht die verschiedenen Quellen der Unsicherheiten und präsentiert, wie diese in der Graphdatenbank modelliert werden können. Schließlich wird ein Anwendungsbeispiel vorgestellt, das auf die Darstellungen von Lebewesen auf den Siegeln fokussiert und die erstellte Graphdatenbank verwendet.

To facilitate the application of network analysis for the study of multi-sided Aegean seals, a graph database was implemented in Neo4j. Many of the seal's attributes contain uncertainties, which require special attention if they are incorporated into the graph database. The article examines where the uncertainties originate from and presents how they can be modelled in the graph database. Finally, a practical example is presented, which focuses on depictions of creatures on seals and makes use of the created graph database.

1. Introduction

Aegean seals are small stone, bone or ivory objects of varying shapes, including discs, cylinders, rectangular blocks or triangular prisms. They originate mostly from Bronze Age Crete (Minoan seals) and mainland Greece (Mycenaean seals), thus dating from 3000 to 1100 BCE.

The seals were used for labelling, sealing or securing other objects, by producing relief impressions in soft materials like clay with their engraved faces.[1] About ten percent of the approximately 10,000 seals known today have more than one such seal face, i.e. are multi-sided. An example is depicted in figure 1. For this group, it still remains unclear if the choice of the motifs engraved on the different faces of a seal follows specific rules or was haphazard.[2] The work presented here is part of a project that aims to find answers to this question by means of computational methods.

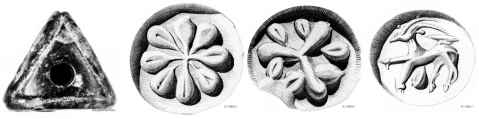


Fig. 1: The three-sided seal CMS II,1 085. [Graphic by courtesy of the CMS Heidelberg.]

[1] Kryszkowska 2005, p. 2.
[2] Trognitz 2017, p. 184.

Network analysis and its visualisation is an exploratory tool that can help in understanding how motif components such as depicted creatures are combined with each other. The method comes from network theory, a field concerned with the study of graphs. Therefore, storing information about the multi-sided seals in a graph database was a natural choice. While developing the data model, special attention had to be paid to unclear or uncertain attributes describing a seal, originating from e. g. a lack of distinct features of the depicted creatures for their unambiguous identification.

The data source used for this work is described in the following section, as well as the way information about the seals is organised. Section 3 focuses on the different kinds of uncertain attributes and explores why uncertainties in the dataset exist. Section 4 presents how the information was transferred into a graph database and introduces the data model.

2. Data Source

All seals considered in this work are recorded in the ›Corpus der Minoischen und Mykenischen Siegel‹ (CMS), a project established in Marburg in 1958, which moved to Heidelberg in 2011. Aim of the CMS is to document and publish all known Aegean seals. In 2007, all seal descriptions were included in the freely accessible object database Arachne. In Arachne, each seal is described in a relational database model with about fifty attributes, such as number of seal faces, dimensions, material, ornaments, and figurative motifs. These attributes are organised into eleven thematic groups: ›identification‹, ›provenience‹, ›shape‹, ›material & technique‹, ›measurements & preservation‹, ›general information about decoration‹, ›stylistic classification‹, ›ornaments‹, ›characters‹, ›figurative motifs excepting creatures‹, and ›creatures‹.

For this work, a total of 1033 pieces was taken into account, forming a dataset that consists of 1033 seals with 301 two-sided, 637 three-sided, 87 four-sided seals, and a few seals (3, 3, 1, and 1 pieces) with more faces (5, 6, 8, and 14 respectively).

Information about the seals was imported from Arachne via its Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)[3] endpoint and then processed into two tables. The first contains all attributes describing one seal face in each row and for the second table all attributes of seal faces belonging to the same seal were merged into a single row. An additional table, containing information about depicted creatures was generated via web scraping. Finally, a fourth table with geospatial information was compiled from the second table with one seal per row and manually enriched with the coordinates of the seal's place of discovery. These tables constitute the data basis for all applications in this project, including the graph database described here.

[3] Lagoze et al. 2015, passim.

3. Uncertainties in the Dataset

Uncertainties in this dataset originate from three different causes, which are linked to the seal's provenience, condition, and ultimately to a human component. These will be briefly examined in the next three subsections. The last subsection describes how values in the dataset are marked as uncertain and which attributes actually contain uncertain values.

3.1 Uncertainties from Provenience

Provenience of a seal, i.e. the archaeological context in which it was found, is important to be able to examine the meaning of a seal considering its place of deposit and other objects associated with it. A secure find context also provides certainty for geographic localisation and dating of a seal. For a few seals, the lack of a secure context might even lead to question their genuineness. If a seal was acquired under dubious circumstances, which in most of the cases means that it was not found during an archaeological excavation, information about its geographic origin has to be taken with a pinch of salt and context dating is not possible at all. Geographic information is available for a total of 583 seals, although for only 289 pieces a secure context is known.

3.2 Uncertainties from a Seal's Condition

The overall condition of a seal plays a crucial role in the identification of its engravings. Determining the used ornaments, hieroglyphic signs or a creature's species is much easier when studying a well preserved and complete seal as opposed to a battered, broken or even fragmentary one. The seals in different preservation conditions in figure 2 shall serve as a visual example.

Fig. 2: A fragment of CMS IS 038a, the slightly damaged CMS I 287b, and the well preserved CMS XII 135b. The first seal side depicts either a bovine or a goat, while the others show a goat. [Graphic by courtesy of the CMS Heidelberg.]

3.3 Uncertainties by Human Action

An unpredictable cause for uncertain descriptions of a seal is human action, which is manifold. To begin with, the seals are man-made objects and thus already in their making the seal engraver may have made mistakes when cutting the stone or may even have obfuscated

the motifs on purpose, thus resulting in ambiguous compositions. Additionally, the carved material, the tools used, the style, and the craftsmanship of the engraver led to very different depictions of the same motif. This can be seen by comparing the seal sides depicted in figures 2 and 3.

All this might then contribute to an uncertain denomination of motifs depicted on a seal by modern scholars while cataloguing the seals. Depending on a scholar's background additional uncertainties may arise due to insufficient knowledge in other fields of expertise, as e.g. the wrong attribution of a specific bird species. In some cases, further vagueness is introduced when a seal is studied by more than one scholar, who might disagree in their interpretations.

When the information is then transferred into the database Arachne further uncertainties can be introduced by input errors or by leaving some fields empty. The latter is ambiguous, because it could mean that either the field was forgotten or that there actually is no information available for it.

3.4 Uncertain Values in the Data Set

Three different markers are used in Arachne to denote uncertain, but probable values in the database: a question mark, giving the different options, and a combination of the first two variants. When options are indicated, only two are given. In figure 3 for example, the values for the attribute ›Lebewesen‹ (creature) of CMS X 322c, CMS IS 038a, and CMS III 504a are ›Ziege?‹ (goat?), ›Rind oder Ziege‹ (bovine or goat), and ›Rind oder Ziege?‹ (bovine or goat?). In the dataset no value has more than two different options.

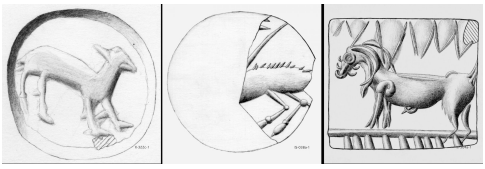


Fig. 3: Three seal sides with uncertainly identified creatures on them. The animal on CMS X 322c is designated as ›Ziege?‹ (goat?), the one on CMS IS 038a as ›Rind oder Ziege‹ (bovine or goat), and the creature on CMS III 504a ›Rind oder Ziege?‹ (bovine or goat?). [Graphic by courtesy of the CMS Heidelberg.]

For supposedly present values which could not be identified, the terms ›undefinierbar‹ or ›undef.‹ (undefinable, indeterminate) were used. The term ›Intagliolücke‹ (gap in the intaglio) is also used for similar cases.

Not all of the sources for uncertainties discussed above are represented in the data set. Some uncertainties caused by human action are not denoted as such, because they are implicit, such as those caused by errors or lack of knowledge. An additional attribute to state, whether a seal came from a secure context or not was added to the data set by consulting the introductory sections of the printed CMS volumes.[4]

Markers for uncertainty can be found in values for almost all attributes in the data set. From the eleven thematic groups mentioned in section 2 only ›identification‹ does not contain any vague values.

Besides the additionally introduced attribute for secure contexts, some attributes about the provenience of a seal such as places of origin are marked with a question mark when deemed doubtful. Uncertainties in the description of a seal's shape are present if the seal is in a bad condition or only a fragment is preserved. This is why seemingly distinct properties such as the shape of the whole seal or of individual seal faces, the type of perforation or in one case – CMS V 256 – even the number of seal faces cannot always be indicated without doubt.

Attributes in the thematic group ›material & technique‹ include information about a seal's material, the technique employed to make the seal, and any other details worth cataloguing. Uncertain values are present for all of those attributes in the dataset.

As can be expected, attributes in the group ›measurements & preservation‹ containing dimensions do not contain any uncertain values. In contrast, the preservation condition of a seal does so, especially if it is not possible to securely indicate that a seal shows tool marks or signs of burning.

All attributes describing the engravings a seal is bearing have values with uncertainties. They are part of the thematic groups ›general information about decoration‹, ›stylistic classification‹, ›ornaments‹, ›characters‹, ›figurative motifs excepting creatures‹, and ›creatures‹. In addition to the example for uncertain identification of creatures given in figure 3 above, two more cases shall be provided.

The attribute ›Standardornament‹ (standard ornament), part of the thematic group ›ornaments‹, can contain multiple values, which in turn can all individually be marked as uncertain. In figure 4, CMS II,1 136a is shown. ›Standardornament‹ contains the value ›Hakenspirale?(2), Punkt, undefinierbar‹ (spiral hook?(2), dot, indeterminate). It indicates that possibly two spiral hooks, one dot and and a further, indeterminate element, are depicted on the seal.

[4] Corpus der minoischen und mykenischen Siegel. Ed by. Akademie der Wissenschaften und der Literatur Mainz 1966–, passim.

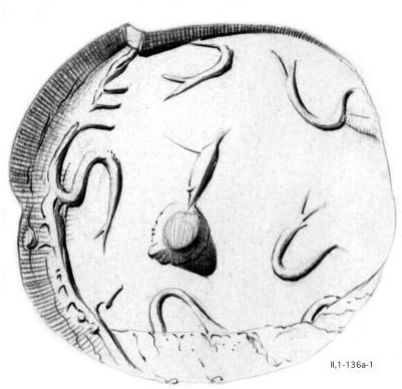


Fig. 4: CMS II,1 136a. The engravings on the seal's face are described with › Hakenspirale?(2), Punkt, undefinierbar‹ (spiral hook?(2), dot, indeterminate). [Graphic by courtesy of the CMS Heidelberg.]

The attribute ›Schrift‹ (script), which belongs to the thematic group ›characters‹, may also contain multiple values in order to list all characters present on a seal's face. The script used on Aegean seals is in most cases Cretan hieroglyphic. The hieroglyphs are specified by using the acronym ›CHIC‹ and a number which relates to the Corpus Hieroglyphicarum Inscriptionum Cretae.[5] For example, the four-sided seal CMS II,2 316 in figure 5 has script on all sides, of which one character, CHIC 056, on face c is marked as uncertain because of the seal face's damage.

Fig. 5: The four sides of CMS II,2 316 with Cretan hieroglyphs: CHIC 044, 049 | CHIC X, 029, 077, 049 | CHIC X, 057, 034, 056? | CHIC X, 044, 005. [Graphic by courtesy of the CMS Heidelberg.]

4. The Graph Database

For the task at hand, which is concerned with the interrelations of motifs, not all attributes available are relevant. Thus the data model which is described in subsection 4.1 is simpler than the one used in Arachne. The graph database management system used was Neo4j, which offers a freely available Community Edition. The import of data into Neo4j is outlined in subsection 4.2 and modelling of uncertainties in the graph database is detailed in the last subsection.

[5] Olivier et al. 1996, passim.

4.1 Data Model

A major difference between the data model described here and the one used by Arachne is that for the graph database seals and their seal faces were modeled as distinct nodes. This allows to separate attributes describing features of the whole seal like material, shape or geographic information, from those only describing a single face, such as depicted motifs. This distinction was not made in Arachne.

Besides the two node types for a seal and for a seal side, 13 further node types were included and are either related to the whole seal or to a side: ›Siegeltyp‹ (seal shape), ›Materialgruppe‹ (material group), ›Material‹ (material), ›Ort‹ (place)‹ ›Stilgruppe‹ (style group), ›Epoche‹ (period), ›Makroornament‹ (macro ornament), ›Standardornament‹ (standard ornament), ›Symbole‹ (symbols), ›Objekte‹ (objects), ›Pflanzen‹ (plants), ›Schrift‹ (script), and ›Lebewesen‹ (creatures).

Nodes can be connected with each other by edges which in Neo4j are called relationships. For this dataset, a total of 17 different relationship types were created. The relationship type defines how and to which other node types a node can be connected. For example, the relationship type ›hat_Siegelseite‹ (has seal side) connects the node type ›Siegel‹ (seal) with the node type ›Siegelseite‹ (seal side). Four different relationship types can be used to connect the node type ›Siegelseite‹ (seal side) with ›Epoche‹ (period), in order to state whether the period represents the beginning of a context date (›hat_EpocheAnfKontext‹), its end (›hat_EpocheEndKontext‹), the beginning of a style date (›hat_EpocheAnfStil‹) or its end (›hat_EpocheEndStil‹).

The node types and their relationships are depicted in figure 6. The node type ›Siegel‹ (seal) is further described with additional properties, including the CMS number, the calculated volume, and the numbers of empty seal faces and depicted creatures.

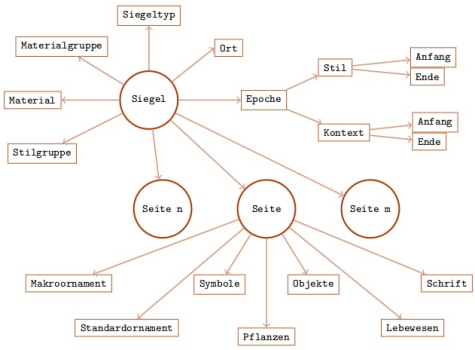


Fig. 6: The data model for Aegean seals used in Neo4j. [Graphic by Martina Trognitz, Vienna.]

4.2 Import into Neo4j

As described in section 2, the dataset was imported into two tables with further information provided in two additional tables. All those tables were imported into Neo4j with Cypher, a query language for graph databases. The Cypher script created for the import does not only import the dataset, but also cleans and filters it. This means that attributes with empty values are not imported into the graph, as well as attributes containing ›None‹ or ›nein‹.

In Arachne multiple equal values for an attribute are not listed separately, but the number of repetitions is set in parentheses, as already seen for CMS II,1 136a exemplified in section 3.4 and figure 4, where ›Standardornament‹ contains the value ›Hakenspirale?(2), Punkt, undefinierbar‹ (spiral hook?(2), dot, indeterminate). If the number is greater than six a greater-than sign is used, thus resulting in something like e.g. ›Hakenspirale(>6)‹. This was accounted for during import by creating an edge for each repetition. Unfortunately, it was not possible to determine the exact number for ›>6‹, which is why a maximum of six edges between a seal side and e.g. a specific standard ornament is created.

In figure 7, a part of the Cypher script for the import of ›Standardornament‹ demonstrates how data is converted into the graph data model. The resulting graph for the seal CMS II,1 085 from Figure 1 is shown in figure 8. Overall a total of 4637 nodes was created with 18306 relationships interconnecting them.

```
// (ssso:Standardornamente {})
FOREACH ( ignoreMe IN CASE WHEN NOT(line.`STANDARDORN` CONTAINS "nein") THEN [1] ELSE [] END |
FOREACH ( j in split(line.`STANDARDORN`, ", ") |
FOREACH ( ignoreMe IN CASE WHEN NOT(j CONTAINS "(") THEN [1] ELSE [] END |
MERGE (ssso:Standardornamente {standardornamente: replace(j, "?", "")})
MERGE (ss)-[sshso:hat_Standardornament {count: 0}]->(ssso)
    SET sshso.weight=
    CASE WHEN (j CONTAINS '?')
        THEN toFloat(0.8)
    ELSE  toFloat(1)
    END
)
FOREACH ( ignoreMe IN CASE WHEN (j CONTAINS "(") THEN [1] ELSE [] END |
FOREACH ( ignoreME IN CASE WHEN (j CONTAINS ">") THEN [1] ELSE [] END |
FOREACH (i in range(0, toInteger(replace(replace(split(j, ">")[1], ")", ""), "?", ""))-1) |
MERGE (ssso:Standardornamente {standardornamente: replace(split(j,">")[0], "?", "")})
MERGE (ss)-[sshso:hat_Standardornament {count: i+1, more: ">6"}]->(ssso)
    SET sshso.weight =
    CASE WHEN (j CONTAINS '?')
        THEN toFloat(0.8)
    ELSE  toFloat(1)
    END
))
FOREACH ( ignoreME IN CASE WHEN NOT (j CONTAINS ">") THEN [1] ELSE [] END |
FOREACH (i in range(0, toInteger(replace(replace(split(split(j, "(")[1], ")")[0], ")", ""), "?",
""))-1) |
MERGE (ssso:Standardornamente {standardornamente: replace(split(j,"(")[0], "?", "")})
MERGE (ss)-[sshso:hat_Standardornament {count: i+1}]->(ssso)
    SET sshso.weight =
    CASE WHEN (j CONTAINS '?')
        THEN toFloat(0.8)
    ELSE  toFloat(1)
    END
))
)))
```

Fig. 7: A section of a Cypher script, which imports standard ornaments depicted on seal sides and sets a weight on the edges depending on the certainty of their identification. [Graphic by Martina Trognitz, Vienna.]

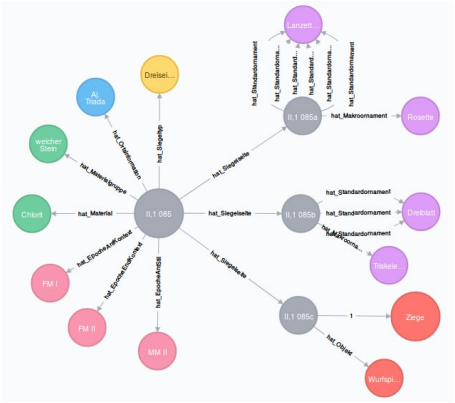


Fig. 8: CMS II,1 085 as a graph in Neo4j. [Graphic by Martina Trognitz, Vienna.]

4.3 Modelling Uncertainties

There are different ways to represent uncertainties in graph databases that depend on the application and its planned queries. For instance, different nodes can be created for certain and uncertain values, like a node for ›Ziege‹ (goat) and another one for ›Ziege?‹ (goat?). In the worst case this would lead to the double amount of nodes. Adding a property to indicate uncertainty to a node would also lead to an increase in the number of nodes. Another way is to use the edges, either by providing different kinds of relationships (e.g. ›has‹ and ›may-have‹) or by including weights. The former leads to an increase in the number of relationships in the graph. The latter approach keeps the number of nodes and edges at a minimum and also allows to include a measure for the uncertainty such as percent. In the data model presented here uncertainties are modelled by applying edge weights ranging from 0 (0 % sure) to 1 (100 % sure).

Not all of the uncertainties mentioned in section 3 can be represented in the graph database, this is only possible for those which are marked as such in the data source. The weights are also set with the Cypher script, based on the value provided in Arachne. If a value comes without any uncertainty marker (e.g. ›Ziege‹ (goat)), the weight of the edge is set to 1. If the value is marked as uncertain with a question mark (e.g. ›Ziege?‹ (goat)), the weight of the edge is $1 > x < 0.5$ which in this case was set to 0.8. For those values where two options are given, two edges are created. The weight of those edges depends on the presence of a question mark. A value like ›Rind oder Ziege‹ (bovine or goat) leads to two edges with a weight of 0.5, i.e. a 50 % chance that one of the two options is actually depicted. The value ›Rind oder Ziege?‹ results in two edges with the weight of $0.5 > y < 0$, which in this case was set to 0.3. Furthermore, a count was introduced in order to be able to distinguish between two uncertain edge pairs belonging to one seal side.

In figure 9, seal CMS IS 038 is shown as a graph in Neo4j. Three edges, connecting the seal sides with creatures are uncertain with weights of 0.5 and 0.8.

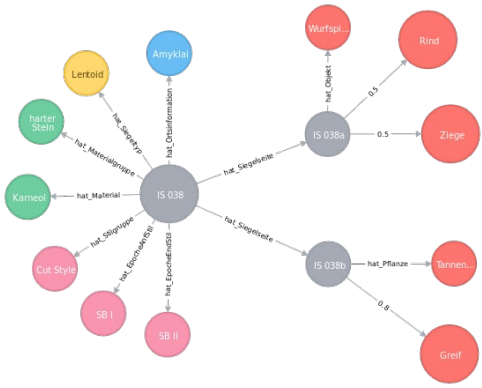


Fig. 9: CMS IS 038 with all its attributes in Neo4j. The edges connecting side A with creatures are weighted with 0.5, due to the value being ›Rind oder Ziege‹ (bovine or goat). Side B is connected with ›Greif‹ (griffin) with a weight of 0.8, because a question mark was present in the value. [Graphic by Martina Trognitz, Vienna.]

5. Use Case

After data import, the graph database can be queried with Cypher. In this way the amount of all seals bearing creatures on them can be queried (724 seals in total). Also the amount of seals with uncertain edges to the node type ›Lebewesen‹ (creatures) can easily be identified and counted with a single query (329 seals). When viewing the database in the browser provided by Neo4j, query results can also be visualised and explored, which is something most relational database management systems do not provide out of the box.

As mentioned in the first section, the graph database was set up to facilitate doing network analysis on the dataset. This shall be demonstrated on the set of creatures depicted on a seal, where a specific research objective is to find out which creatures are combined with each other on seals.

For this task a set of nodes containing all creatures and the connections between them is needed. Since the connections are not contained in the database, they have to be created with the query shown in figure 10.

```
MATCH (c1:Lebewesen)--()--(:Siegel)--()--(c2:Lebewesen)
WHERE id(c1) < id(c2)
MERGE (c1)-[r:mit]-(c2)
ON CREATE SET r.weight = 1
ON MATCH SET r.weight = r.weight +1
```

Fig. 10: The Cypher query establishes links between two different creatures depicted on the same seal and increases the link count (the weight of the edge) for every occurrence of this pair on other seals. [Graphic by Martina Trognitz, Vienna.]

This can then be exported in order to be processed with a network analysis software, such as visone. The dataset can be provided in graphML format, which can be exported from Neo4j with a single command by using the neo4j-shell-tools. Here another major advantage of using a graph database becomes clear, because the list of nodes and edges does not have to be tediously produced from the four tables introduced in section 2.

Two different datasets were exported as graphML, imported into visone, analysed and visualised. In figure 11, the resulting graph of the dataset containing all co-occurrences of different creatures, including certain and uncertain ones, is displayed. When only considering creatures with a certainty equal to 0.8 or above, the graph results in the image in figure 12.

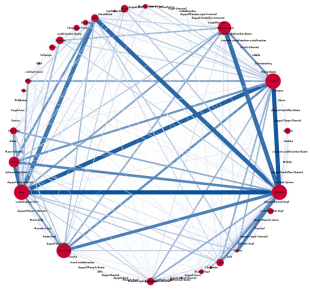


Fig. 11: A network visualisation showing which creatures are often depicted together on a seal. The underlying data takes into account all existent depictions, regardless if they are certain or uncertain. [Graphic by Martina Trognitz, Vienna.]

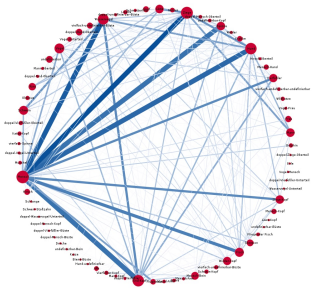


Fig. 12: A network visualisation showing which creatures are often depicted together on a seal. The underlying data only takes into account existent depictions with a certainty equal or greater than 0.8. [Graphic by Martina Trognitz, Vienna.]

6. Conclusion

The implementation of a graph database for Aegean seals with more than one side for sealing requires special attention when uncertainties are present in the dataset. It is not possible to include uncertainties implicit to the dataset into the data model, but for those made explicit, various ways exist. In the case displayed here, modelling uncertainties as weighted edges proved to be most suitable.

Working with a graph database in practice proves to be intuitive and fast, especially because data is not only displayed in tabular form, but also as a graph with expandable nodes. Thus the user is allowed to further explore queried results.

Exporting data for further examination with network analysis software feels almost natural and by including or excluding uncertain values this process can be further tuned. The use case presented does only scratch the surface of the possibilities, as e.g. the result displayed in figure 11 does include both possible values for uncertain values with options. In further experiments the data could be analysed by only including one option and then be compared to an analysis including the other option.

By expanding the graph database model with further node types in order to e.g. introduce broader categories for the creatures, such as ›wild animal‹ or ›domestic animal‹, the underlying rules inherent to the use of motifs on the seals might be uncovered.

Bibliographic References

Corpus der minoischen und mykenischen Siegel. Ed. by Akadademie der Wissenschaften und der Literatur Mainz. 13 Vol. Berlin et al. 1966–. [Nachweis im GBV]

Olga Kryszkowska: Aegean Seals: An Introduction. London 2005. [Nachweis im GBV]

Carl Lagoze / Herbert Van de Sompel / Michael Nelson / Simeon Werner: The Open Archives Initiative Protocol for Metadata Harvesting. Document Version 08.01.2015. [online]

Jean-Pierre Olivier / Louis Godart / Jean-Claude Poursat: Corpus hieroglyphicarum inscriptionum cretae. Paris 1996. Siehe auch [Nachweis im GBV]

Martina Trognitz: Approaching Multi-Sided Aegean Seals with Machine Learning Techniques. In: Archaeological Approaches to Breaking Boundaries: Interaction, Integration & Division. Proceedings of the Graduate Archaeology at Oxford Conferences 2015–2016. Ed. by Rebecca O'Sullivan / Christina Marini / Julia Binnberg. Oxford 2017, pp. 183–198. [Nachweis im GBV]

List of Figures with Captions

ZfdG

Zeitschrift für digitale Geisteswissenschaften

Beitrag aus:
Sonderband 4 der ZfdG: Die Modellierung des Zweifels – Schlüsselideen und -konzepte zur graphbasierten Modellierung von Unsicherheiten. Hg. von Andreas Kuczera, Thorsten Wübbena und Thomas Kollatz. 2019. DOI: 10.17175/sb004

Titel:
Genau, wahrscheinlich, eher nicht: Beziehungsprobleme in einem kunsthistorischen Wissensgraph

Autor/in:
Martin Raspe

Kontakt: raspe@biblhertz.it
Institution: Bibliotheca Hertziana – Max Planck Institut für Kunstgeschichte, Rom
GND: 139144145 ORCID: 0000-0003-0861-0412

Autor/in:
Georg Schelbert

Kontakt: georg.schelbert@hu-berlin.de
Institution: Humboldt-Universität zu Berlin, Institut für Kunst- und Bildgeschichte
GND: 133448231 ORCID: 0000-0002-7314-8589

DOI des Artikels:
10.17175/sb004_012

Nachweis im OPAC der Herzog August Bibliothek:
1037075307

Erstveröffentlichung:
13.02.2019

Lizenz:

Sofern nicht anders angegeben 

Medienlizenzen:
Medienrechte liegen bei den Autoren

Letzte Überprüfung aller Verweise:
13.02.2019

GND-Verschlagwortung:
Datenmodell | Graphdatenbank | Konzeptionelle Modellierung | Kunstwissenschaft | Wissensrepräsentation |

Zitierweise:
Martin Raspe, Georg Schelbert: Genau, wahrscheinlich, eher nicht: Beziehungsprobleme in einem kunsthistorischen Wissensgraph. In: Die Modellierung des Zweifels – Schlüsselideen und -konzepte zur graphbasierten Modellierung von Unsicherheiten. Hg. von Andreas Kuczera, Thorsten Wübbena und Thomas Kollatz. 2019 (= Sonderband der Zeitschrift für digitale Geisteswissenschaften, 4). PDF Format ohne Paginierung. Als text/html abrufbar unter DOI: 10.17175/sb004_012.

Martin Raspe, Georg Schelbert

Genau, wahrscheinlich, eher nicht: Beziehungsprobleme in einem kunsthistorischen Wissensgraph

Abstracts

Geisteswissenschaftliche Forschungsdaten in einem digitalen Wissensgraph abzubilden, bringt eine doppelte Herausforderung mit sich: Wie werden ungewisse Angaben in Form elektronischer Daten gespeichert und welche Folgen entstehen daraus für Abfrage und Visualisierung? Was drückt Unsicherheit aus und wie beeinflusst sie unser Konzept von Wissen? Das kulturhistorische Informationssystem ZUCCARO der Bibliotheca Hertziana ist ein Beispiel für einen komplexen Wissensgraph. Darin sind drei Fälle von Unsicherheit oder Vagheit zu unterscheiden: Genauigkeit, Plausibilität und negative Aussagen. Der Beitrag untersucht diese drei Aspekte im Hinblick auf die geschichtliche Wirklichkeit, den wissenschaftlichen Forschungsstand und die Benutzbarkeit der Daten für den Wissenschaftler.

To represent research data from the realm of the humanities in a digital knowledge graph presents a double challenge: How is uncertain data to be stored, and what does that imply for retrieval and presentation? Which notions are conveyed through uncertain data, and how do they influence our concept of knowledge? The cultural history information system ZUCCARO created by the Bibliotheca Hertziana is a good example for a complex knowledge graph. Three cases of uncertainty or vagueness are to be distinguished: Precision, plausibility and positively negative assertions. Our contribution studies the three aspects with regard to the historic reality, the state of scientific research and the usability of the content data for scholars.

»Uncertainty is the lack of information«[1]

»Uncertainty is the prerequisite to gaining knowledge and frequently the result as well«[2]

1. Digitalität, Wissensgraph und Unsicherheit

1.1 Wissen

Mit dem Begriff Wissen wird – sowohl in den Natur- als auch in den Geisteswissenschaften – ein von einem größeren Personenkreis geteilter Bestand von begründeten Aussagen bezeichnet.[3] Dieser Wissensbestand zeichnet sich im besten Fall durch einen »größtmöglichen Grad an Gewissheit«[4] aus. So wird erreicht, dass die gemeinsam geteilten Auffassungen als gültig beziehungsweise ›wahr‹ angenommen werden können. Eine absolute Verlässlichkeit kann nicht erreicht werden: Wissen hat prinzipiell und stets den Charakter des Vorläufigen.

[1] Bonneau et al. 2014.
[2] Hamilton 1936.
[3] Einleitend hierzu Wikipedia: Wissen.
[4] Wikipedia: Gewissheit.

Wissenschaftliche Tätigkeit besteht in der Erzeugung, der Anreicherung und der Vertiefung von Wissen. In der Praxis geschieht dies durch das Aufstellen und das Weitergeben beziehungsweise Mitteilen von Aussagen, die durch Argumente begründet sind. Erst durch die argumentative Begründung kommt die Wissenschaftlichkeit zustande. Die Begründung ist nicht nur deshalb notwendig, weil jeder Aussage über die Wirklichkeit ein Quantum an Distanz, Subjektivität oder Unzuverlässigkeit anhaftet, sondern auch, weil zur Definition der wissenschaftlichen Mitteilung gehört, dass der Adressat überprüfen kann, wie das an ihn weitergegebene Wissen entstanden ist. Nur so ist eine ›Rückkopplung‹ zwischen Empfänger und Sender möglich; nur so wird sichergestellt, dass der Adressat die Aussage des Urhebers für sich übernehmen kann. Er muss die Aussage in ihrem Entstehungskontext verstehen und ihren Grad an Gewissheit beurteilen können. Unsicherheit – oder besser Mangel an Gewissheit – ist also auch in der Praxis geisteswissenschaftlicher Forschung ein Aspekt, der jedwedem Wissen innewohnt.

Diese Ungewissheit in einem komplexen digitalen Wissensspeicher abzubilden, bringt eine doppelte Herausforderung mit sich. Einerseits eine praktische: Wie können und sollen ungewisse oder unsichere Daten modelliert und gespeichert werden, und welche Folgen entstehen daraus für die Darstellung und die Abfrage dieser Daten? Andererseits eine theoretische: In welcher Weise wird die in ein Datenformat gebrachte Unsicherheit vom Benutzer inhaltlich verstanden und wie beeinflusst sie unser Konzept von Wissen?

1.2 Das Informationssystem ZUCCARO als Wissensgraph

Bei einem komplexen Wissensgraph wie dem Informationssystem *ZUCCARO* der Bibliotheca Hertziana, das wir als Beispiel heranziehen,[5] wird kulturgeschichtliches Wissen grundsätzlich in Form von zeitbasierten Relationen zwischen Akteuren beziehungsweise Entitäten dargestellt. Anhand von Einzelfällen wollen wir anschaulich demonstrieren, welche Möglichkeiten es gibt, Unsicherheiten im Rahmen historischer Angaben zu modellieren. ZUCCARO ist ein Informationssystem für die historischen Kulturwissenschaften. Es wird an der Bibliotheca Hertziana – Max-Planck-Institut für Kunstgeschichte – in Rom seit 2003 von den Verfassern konzipiert und entwickelt. Seit 2005 ist es in einem relationalen Datenbanksystem prototypisch realisiert und seit 2008 online zugänglich. Das zu Grunde liegende Datenmodell ist so generisch und zugleich erweiterbar ausgelegt, dass es Forschungsprojekte aus unterschiedlichen Bereichen der Kulturwissenschaften unterstützt. Das System kann als universelles Repositorium dienen; prinzipiell ist es an alle gängigen Standardformate in den historischen Disziplinen anpassbar.

Derzeit kommt das System vor allem zur Sammlung von Informationen und Bildmaterial zur italienischen, besonders römischen Kunstgeschichte zum Einsatz. In erster Linie bilden Materialien aus den Forschungsprojekten Lineamenta (italienische Architekturzeichnungen des 18. Jahrhunderts) und ArsRoma (römische Malerei um 1600 im gesellschaftlichen Umkreis von Caravaggio) die Datenbasis. Das System enthält aber auch zahlreiche weitere Bestände,

[5] Zugang und Dokumentation: Zuccaro.

speziell zur Topographie der Stadt Rom (Bauwerke, Institutionen, Stadtpläne, Veduten), zu Künstleraufenthalten in Italien im 19. Jahrhundert sowie zu vielen Rara-Digitalisaten aus der Institutsbibliothek. Die Projektdaten sind durch *tags* gekennzeichnet, aber technisch nicht weiter separiert. Der Datenbestand ist also grundsätzlich offen und erweiterbar: Er enthält gemeinsame Stammdaten und zahllose Querverbindungen und kann insofern niemals als ›abgeschlossen‹ gelten. Durch jede neue Eingabe wird das Wissensnetz dichter geknüpft und dadurch nützlicher.

Kulturhistorische Forschung besteht nicht allein im Katalogisieren von materiellen Gegenständen, Artefakten und Bauwerken, sondern bezieht auch den historischen, politisch-gesellschaftlichen und ideell-konzeptionellen Kontext mit ein. Über die Werke hinaus widmet sie sich den Personen, Institutionen, sozialen Gruppen, Berufen und gesellschaftlichen Funktionen. Diese üblicherweise als ›Metadaten‹ bezeichneten Inhalte sind oft selbst wichtige Gegenstände der Forschung. Deshalb dienen sie als Rahmen unseres Informationssystems nicht vorrangig zur Objekterschließung, sondern werden ebenfalls als Datenobjekte erster Ordnung behandelt. Darüber hinaus untersucht die Forschung formale, inhaltliche und topographische Zusammenhänge und berücksichtigt Archivalien, Dokumente und Fachpublikationen. Alle diese Gegenstände werden in ZUCCARO als gleichwertig angesehen und können sowohl mit gezielten Abfragen als auch durch exploratives Browsen studiert werden.

Kulturhistorisches Wissen entsteht durch die Vernetzung von Informationen. Es bildet sich durch die Dokumentation geschichtlicher Ereignisse, an denen Personen, Objekte, Orte und immaterielle Konzepte beteiligt sind und die durch Quellen und Literatur historisch belegt sind. ZUCCARO trägt dieser Struktur des Wissens Rechnung, indem es jeden statischen Gegenstand – sei es ein handelnder ›Akteur‹ oder ein passives ›Objekt‹ – als sogenannte *Entität* behandelt. Die Entität ist lediglich ein digitaler Stellvertreter des Gegenstands. Der Datensatz wird mit Hilfe eines eindeutigen *identifiers* adressiert und kann durch Eigenschaften (*properties*) näher bestimmt werden, zum Beispiel durch Namen oder Bezeichnungen in verschiedenen Sprachen.

Zwischen derartigen abstrakten Entitäten können konkrete *Relationen* angelegt werden. In diesen Beziehungen ist das eigentliche historische Wissen enthalten. Generell sind diese Beziehungen zwischen jeweils zwei Entitäten in der Form von einfachen Aussagesätzen (*semantischen Tripeln*) nach dem Muster Subjekt-Prädikat-Objekt darstellbar, zum Beispiel: Die Nachtwache wurde gemalt von Rembrandt, oder: Der Mann mit dem Goldhelm wurde in der Technik Öl auf Leinwand ausgeführt.[6]

Nimmt man zu solch einer auf das Allerwesentlichste reduzierten Aussage noch den Zeitaspekt hinzu, und zwar in Gestalt einer Datierung, so kommt man zu einer Datensatzform, die man als nicht weiter reduzierbares mikrohistorisches Element ansehen kann, gewissermaßen als das kleinstmögliche geschichtliche Ereignis, wie zum Beispiel:

[6] Zur Definition Wikipedia: Semantic triple.

›Rembrandt kaufte im Jahre 1639 das Haus in der Jodenbreestraat‹. Dabei sind Rembrandt und das Haus feststehende Entitäten, während das erweiterte Prädikat ›kaufte im Jahre 1639‹ eine historische Beziehung zwischen beiden Entitäten herstellt.

Auf diese Weise steht das historische Ereignis im Zentrum unseres Datenmodells (Abbildung 1). Dieses Konzept, bei dem eine Beziehung zwischen Entitäten durch ein Datum oder einem Zeitraum erweitert wird, bezeichnen wir als ›Ereignis‹ oder *event*. Mit diesem Datenformat können alle inhaltlichen Entitäten auf einfache, generische Weise miteinander verbunden werden. Natürlich muss eine Beziehung nicht zwingend mit einem Datum versehen werden; sie kann außerdem durch zusätzliche Attribute genauer qualifiziert werden. Das kann die Spezifizierung des Beziehungstyps betreffen, zum Beispiel [Rembrandt] – war Schüler von – [Pieter Lastman], oder auch die Angabe einer Quantität, zum Beispiel [das Rembrandthaus] – besitzt eine Anzahl von 4 – [Fensterachsen].

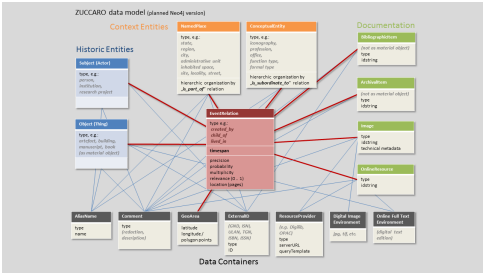


Abb. 1: Schematisches Datenmodell des Informationssystems ZUCCARO (Stand 2013). CC-BY-NC-SA 4.0.

Gewiss sind zahlreiche kulturhistorische Aussagen denkbar, die mit diesem generischen Format nicht ohne weiteres zu erfassen sind. Das Konzept kommt aber der technischen Umsetzung sehr entgegen und ermöglicht eine vielseitige Durchsuchbarkeit der Datenbestände unter beliebigen Aspekten und Fragestellungen. Von entscheidender Bedeutung ist noch ein zweiter Aspekt des Konzepts: Nicht nur Entitäten können miteinander in Beziehung gesetzt werden; Beziehungsdatensätze können ihrerseits durch eigene Beziehungen mit weiteren Datensätzen verbunden werden. Die Beziehungen werden dadurch *reifiziert*,[7] sie werden also selber als statische Objekte behandelt. Auf diese Weise ist es möglich, mit einer Beziehung zusätzliche Umstände zu verknüpfen wie etwa den Kaufpreis oder den Verkäufer, aber auch den historischen oder fachwissenschaftlichen Beleg, etwa ein Archivstück oder eine Publikation mit Seitenangabe. Insbesondere diese Funktionalität ist von besonderer Wichtigkeit, da durch sie das Element der wissenschaftlichen Begründung abgebildet werden kann. In dem man eine Beziehung mit Belegen verknüpft, werden die historischen Angaben, die die Datenbank macht, nachprüfbar. Erst dann bekommen die Relationen wissenschaftlichen Charakter und unterscheiden sich etwa von einfachen Verknüpfungen im Rahmen von *linked open data*, bei denen nicht deutlich wird, auf welcher Grundlage die Aussage zustande gekommen ist. Beziehungen werden somit ihrerseits zu Objekt-Instanzen,

[7] Definition Wikipedia: Reifikation.

die in einem Graphen eigentlich durch Knoten repräsentiert werden müssten. Dadurch entsteht eine hybride Struktur, die nicht ohne Weiteres mit der klassischen Graphentheorie kompatibel ist.[8]

Betrachtet man viele solcher ›Mikroereignisse‹ zusammen, so wird klar, wie das Datenmodell komplexere historische Zusammenhänge darstellen kann, etwa die Biographie eines Künstlers anhand der Orte, an denen er sich nacheinander aufgehalten hat, oder seine Karriere anhand der Kontakte zu Förderern und Auftraggebern, oder auch die Abfolge der Personen, die ein Amt ausübten, oder die Kunstwerke, die zu einer historischen Sammlung gehörten, die heute zerstreut ist. Natürlich ist die Voraussetzung dafür, dass zu der gegebenen Fragestellung ein dichter und konsistenter Datenbestand vorliegt.

Wissenschaftliche Anfragen aus derart unterschiedlichen Blickwinkeln sind mit den meisten herkömmlichen Datenbank-Modellen kaum abzubilden. Eine solche ›polyfokale‹ Organisation von Forschungsdaten, die auf unterschiedliche Forschungsfragen reagiert, findet jedoch erfreulicherweise im Bereich der Digital Humanities zunehmend Anklang. Das Datenmodell von ZUCCARO ist angeregt durch die Datenbank zur Antikenrezeption *Census* [9] und die kulturwissenschaftliche Ontologie *CIDOC-CRM*. [10] Es ist jedoch grundsätzlich generisch konzipiert und legt den Benutzer nicht auf ausgewählte Blickwinkel oder Forschungsgegenstände fest. Im informatischen Sinn stellt das Datenmodell einen sogenannten *property graph* dar. Eine solche netzartige Datenstruktur besteht aus *nodes* (Knoten, bei uns Entitäten) und *edges* (Kanten, bei uns Beziehungen), die beide *properties* (Eigenschaften, also Felder bzw. Attribute) besitzen können.[11] Seit einigen Jahren ist ein wachsendes Interesse an derartigen Datenmodellen zu beobachten, was uns hinsichtlich der Richtigkeit des eingeschlagenen Weges bestärkt.

Mit Hilfe eines derartigen Datenmodells, das auf dem Prinzip *property graph* beruht, können sämtliche Daten im Idealfall redundanzfrei als erweiterte Tripelstrukturen dargestellt werden. Zurzeit ist ZUCCARO als Prototyp in einem proprietären, relationalen Datenbanksystem implementiert und mit einem Web-Frontend auf der Basis von XML-Ausgabedaten und Rendering mit Hilfe von XSLT-Templates ausgestattet. Alle Beziehungen müssen durch *joins* zwischen Tabellen abgebildet werden, was sehr komplex werden kann und die Performance belastet. Mit dem Aufkommen der Graphentechnologien steht dem Datenmodell erstmalig eine adäquate Software-Basis gegenüber. Seit einiger Zeit ist die Umsetzung auf ein Graphdatenbanksystem und ein auf modernen Web-Technologien basierendes Framework in Vorbereitung. Vorgesehen ist *Neo4j* als Datenbank-Software und *Mojolicious* oder *Phoenix* [12] als Management Interface. Besonderer Wert wird auf

[8] Um sich auf ein Datentripel als Ganzes bzw. auf die Instanz einer Relation beziehen zu können, verwendet man in dem Datenformat RDF sogenannte N-Quads – um ein viertes Element erweiterte Tripel.
[9] Census of Antique Works of Art and Architecture.
[10] CIDOC Conceptual Reference Model.
[11] Frisendal 2017.
[12] Mojolicious; Phoenix Framework.

Anschlussfähigkeit und Schnittstellen im Bereich *semantic web* gelegt. Sie soll durch standardisierte Formate und Schnittstellen erreicht werden, beispielsweise durch eine in GraphQL formulierte API.[13]

1.3 The modelling gap: Unsicherheit durch Abstraktion

Überall dort, wo historische Sachverhalte und Zusammenhänge in ein abstraktes Datenbankformat übertragen werden sollen, ergibt sich das Problem des Übergangs vom analogen Kontinuum zur digitalen Fragmentierung. Die fließende raum-zeitliche Entwicklung der traditionellen, textbasierten Geschichtsdarstellung muss in Ereignisse, Zeitabschnitte, Raumelemente und Sinneinheiten zerlegt werden, um sie in digitaler Form zu speichern, zu vergleichen und je nach Forschungsinteresse neu aggregieren zu können. Entscheidungen zur Zusammenfassung oder Teilung von Dingen werden notwendig: Wo sind natürlich gegebene Grenzen? Wie sehr fragmentieren wir die Welt? Ein Bauwerk bildet beispielsweise eine Einheit – es datentechnisch von seinen Nachbargebäuden zu unterscheiden, leuchtet sofort ein. Andererseits besteht es vielleicht aus Haupt- und Nebengebäuden und hat verschiedene Bauphasen – wie weit geht man bei der Aufteilung?

Nicht alle so entstehende Datenfragmente sind gleichermaßen umfangreich oder wichtig. Im Vergleich zur Wirklichkeit ergibt sich das Problem der Verhältnismäßigkeit der digitalen Granularität: Kleine Dinge erhalten ein überproportionales Gewicht im Datenspeicher, wenn sie einzeln modelliert werden. Der Datensatz kennt keine Dimension: Das Formular für das Fürstenschloss hat gleich viele Felder wie dasjenige für das Gärtnerhaus. Besonders sinnfällig zeigt sich die Diskrepanz im Bereich der CAD-Modelle: Der gesamte Baukörper des aus schlichten, quaderförmigen Elementen bestehenden Bauhaus in Dessau kann beispielsweise mit weitaus weniger Polygonen modelliert werden als ein einziges korinthisches Kapitell mit seinen naturnahen Blattformen und geschwungenen Linien.

Aber es ist nicht nur die Zerlegung und die damit einhergehende Abstraktion, die bei der Übertragung ins Digitale eine Verzerrung der Wirklichkeit mit sich bringen. Im kulturhistorischen Bereich kommt eine weitere Schwierigkeit hinzu, die es in letzter Konsequenz schlichtweg unmöglich macht, im digitalen Raum ein adäquat proportioniertes Abbild der geschichtlichen Wirklichkeit herzustellen – und sei es auch nur eines kleinen Ausschnitts davon. Es ist das Grundproblem jeder historischen Forschung: In den allerseltensten Fällen sind Angaben so vollständig und flächendeckend vorhanden, dass es möglich erscheint, ein ausgewogenes, ›statistisch‹ korrektes Abbild der historischen Lage zu zeichnen. Die Lückenhaftigkeit sowohl der Überlieferung als auch in der historischen Forschung sorgt dafür, dass in aller Regel nur ausschnitthafte, mehr oder minder exemplarische Daten vorliegen – zum Beispiel für einzelne Personen in einer Gruppe, oder für wenige Jahre in einem länger dauernden Vorgang. Natursprachliche Texte können durch eine Vielzahl von Formulierungsweisen die wissenschaftliche Darstellung so nuancieren, dass einerseits die Sachverhalte im Einzelnen zutreffend gewichtet werden und andererseits

[13] Wikipedia: GraphQL.

ein reflektiertes, angemessenes Gesamtbild entsteht. Datenbanken können dies nicht; sie können gegebenenfalls die Quellen getreu abbilden, eine Wirklichkeit zusammenfassend rekonstruieren oder gar synthetisieren jedoch nicht.

Ein pragmatischer Ausweg aus dem Dilemma kann daher sein, sich bei Entscheidungen in Bezug auf Abstraktion und Granulation von den vorliegenden Quellen leiten zu lassen. Einteilungen von Zeit und Raum werden zum Beispiel bei Melderegister-Einträgen, die nur einmal im Jahr erhoben werden, durch die Quelle vorgegeben. Das gleiche gilt für die Benennung von Bauabschnitten, die aus Bauabrechnungen übernommen werden. Alle Probleme, die sich aus der Lesung und Deutung der Quellen ergeben, finden sich dann in der Datenbank wieder. Die Unsicherheit resultiert in diesem Fall nicht aus der Abstraktion, sondern aus dem Verzicht auf historische Interpretation. Konsequenterweise wäre eine solche Datenbank dann kaum mehr als ein ›kulturhistorisches Informationssystem‹ zu bezeichnen, sondern hätte den Status einer strukturierten Quellenedition.

Ein Informationssystem wie ZUCCARO, das nicht ausschließlich Originalquellen reproduziert, sondern mit impliziten, aus der Sekundärliteratur oder sogar allgemein bekannten Fakten arbeitet, hat also aus systemischen Gründen keine Chance, den *modelling gap*, also den intellektuellen Abstand zwischen Wirklichkeit und Modell, zu überbrücken. Ein Informationssystem hat lediglich Hinweischarakter. Es ist ein Findmittel, das im besten Falle rasch und übersichtlich zum Forschungsstand hinführt, aber keine virtuelle Historie. Dem Missverständnis, Datenbankinhalte könnten die historische Wirklichkeit auch nur näherungsweise darstellen beziehungsweise ›repräsentieren‹, muss stets aufs Neue entgegengetreten werden: Der Wegweiser zeigt den Weg, aber er geht ihn nicht. Ziel eines digitalen Informationssystems ist es, auf bekannte Fakten, Zusammenhänge und Forschungsmeinungen aufmerksam zu machen und dieses Material so transparent und sinnfällig wie möglich aufzubereiten und anzubieten. Der historische Wahrheitsgehalt ist – wie bei jedem wissenschaftlichen Katalog – vom Benutzer selber anhand der angegebenen Quellen zu verifizieren und einzuschätzen. Der adäquaten Wiedergabe von Unschärfen, Wahrscheinlichkeiten und Qualitäten historischer Ereignisse und Sachverhalte sind insofern von vornherein pragmatische Grenzen gesetzt.

2. Genau, wahrscheinlich, eher nicht...

2.1 Genau: Präzision und Trennschärfe

Zunächst ist festzuhalten: In einem graphbasierten Datenmodell sind jeweils zwei Knoten – die Entitäten repräsentieren – durch eine Kante – also eine Beziehung – miteinander verbunden. Unsicherheit kann ausschließlich in der Beziehung ausgedrückt werden. Eine vage oder ungewisse Entität ist zwar vielleicht denkbar (zum Beispiel der mythische Ort Thule oder der unmerklich wirkende »Zeitgeist«), es hat jedoch wenig Sinn, diese in wissenschaftlichen Aussagen zu verwenden – es sei denn, wenn man das gedankliche Konzept meint und nicht den Ort oder Akteur. Auch bei fest eingeführten Fachbegriffen gibt es Randunschärfen, zum

Beispiel: Fällt eine Mixtur aus Leinöl, Eigelb und Pigment bereits unter den Begriff Ölfarbe? oder: Kann man diesen Stein als Würfelkapitell bezeichnen, oder ist er nur ein roh behauener Kämpferblock? oder: Wann beginnt, wann endet die Stilepoche des Barock? Trotzdem sind diese Begriffe wohldefiniert. Unsicher oder umstritten ist vielmehr, ob man eine Entität – also einen Forschungsgegenstand – mit vollem Recht einem bestimmten begrifflichen Konzept zuordnen kann. Auch hier liegt also die Ungewissheit in der Relation, nicht in der allzu vagen Definition der Entität.

Ein Grundproblem bei der Modellierung von Unschärfen und Unsicherheiten besteht darin, dass das übliche Datenmodell eines Graphen, das sich von der mathematischen Graphentheorie herleitet, überhaupt keine unterschiedlich gewichteten Kanten (in unserem Fall Wissensrelationen) vorsieht. Um eine solche Gewichtung vorzunehmen zu können, muss man einen *property graph* verwenden, bei dem die Relationen mit zusätzlichen Angaben angereichert, also *reifiziert* werden können. In einem oder mehreren dieser Felder können Werte gespeichert werden, die den Grad der Unsicherheit beziehungsweise Vagheit abbilden.

Der einfachste Fall ist hierbei die *Präzision*: Die Relation enthält eine Angabe über den Grad der Genauigkeit, mit dem sie zutrifft. Hierbei steht in der Regel nicht die Relation selber in Frage, sie wird als gegeben betrachtet; vielmehr ist eines ihrer Attribute oder ein Parameter mehr oder weniger zutreffend. Das kann zum Beispiel die Datierung eines Ereignisses, die Zuordnung einer Entität zu einer Kategorie oder auch die Lokalisierung eines Ortes betreffen.

Ein Datenmodell kann hierzu festlegen, wie etwa zeitliche Ungenauigkeit oder inhaltliche Unklarheit kodiert werden sollen. Ein ungewisses Datum kann in einen gewissen Zeitraum fallen. Man kann also einen *terminus post quem* und einen *terminus ante quem* angeben, ohne genauer festzulegen, welches Datum als das wahrscheinlichste gelten kann; man kann aber auch nur ein Datum angeben und dazu die mögliche Streuung festhalten.[14]

Im geisteswissenschaftlichen Kontext tritt häufig der Fall auf, dass es für ein bestimmtes Ereignis unterschiedliche, einander ausschließende Datierungsvorschläge gibt. In diesem Fall erscheint es sinnvoll, mehrere Ereignis-Beziehungen anzulegen und mit unterschiedlichen Zeitspannen zu versehen. Die jeweilige Begründung kann dann mit der entsprechenden Beziehung verknüpft werden. Es ist auch möglich, zusätzlich eine allgemeinere Ereignis-Beziehung anzulegen, die den gesamten vorgeschlagenen Zeitraum umfasst. Wählt man die Lösung mit mehreren, alternativ datierten Ereignissen, so schließt sich sofort die Frage an, wie diese dargestellt werden, in welcher Reihenfolge oder in welcher Auswahl. Je genauer man es mit der Präzision nimmt, umso problematischer werden die Visualisierung und die intuitive Verständlichkeit der Daten.

An dieser Stelle wird deutlich, dass es keine allgemeingültige Lösung für derartige Präzisionsprobleme geben kann, die jeden Einzelfall abdeckt. Die Modellierung von Datierungen kann sich nicht allein danach richten, wie die objektive Wirklichkeit ausgesehen

[14] Vgl. dazu die Definition der Entität E52 Time-Span im CIDOC-CRM.

hat, da diese in vielen Fällen unbekannt oder nicht verfügbar ist. Vielmehr wird man sich oft auf vermutetes, aber verschiedenartig begründetes Wissen stützen. Die Frage, wie ein ungenau bekannter Sachverhalt in einer Datenbank technisch umgesetzt und damit inhaltlich dargestellt werden soll, kann also nicht allein anhand dessen entschieden werden, wie es sich in der historischen Wirklichkeit verhielt. Entscheidend ist vielmehr die Überlegung, wie der Sachverhalt dem Benutzer der Datenbank im Suchergebnis dargestellt werden soll. Modellierungsfragen können sich also nicht primär an den Charakteristika der Inhalte orientieren, sondern müssen die Bedürfnisse der Benutzer beziehungsweise den Verwendungszweck der Datenbank zu Grunde legen.

2.2 Wahrscheinlich: Plausibilität als Gewichtungskriterium

Aus dem geschilderten Beispiel, bei dem mehrere Datierungsvorschläge miteinander konkurrieren, ergibt sich ein zweiter, prinzipiell anders gelagerter Fall von Ungewissheit in einer kulturhistorischen Datenbank. Hier geht es nicht mehr allein darum, die mangelnde Präzision einer Aussage festzuhalten, sondern ihre *Plausibilität* im Vergleich zu anderen, widersprechenden Aussagen, oder ihre Wahrscheinlichkeit im Vergleich zu ihrem Gegenteil abzubilden.

In den wenigsten Fällen ist historisches Wissen mit absoluter Sicherheit belegbar. Es ist also durchaus verständlich und legitim, einen Grad an Wahrscheinlichkeit angeben zu wollen, ob eine im Datenbestand angelegte Beziehung zwischen zwei Entitäten tatsächlich historisch bestanden hat. Hier kann man in der Tat von ›verschieden stark‹ ausgeprägten oder gewichteten Graph-Kanten sprechen.

Es leuchtet unmittelbar ein, dass dieser Fall zwei Probleme aufwirft, die im Fall der mangelnden Genauigkeit nicht auftreten. Zum einen betrifft die fehlende Sicherheit hier die Existenz der Beziehung als solcher. Wenn wir die genaue Datierung eines Ereignisses nicht kennen, so ist dadurch noch nicht in Frage gestellt, dass es das Ereignis gab. Wenn aber unklar ist, ob das Ereignis in dem angegeben Zeitraum überhaupt stattgefunden hat, dann kann der Fall eintreten, dass in der Datenbank ein Sachverhalt gespeichert wird, der möglicherweise der historischen Wahrheit widerspricht.

Das alleine wäre noch kein grundsätzliches Problem, denn auch in der analogen Geisteswissenschaft werden regelmäßig hypothetisch mögliche Sachverhalte vorgeschlagen, sofern ihnen eine gewisse Wahrscheinlichkeit zu eigen ist. Ausschlaggebend ist hier erneut die mit der Aussage verknüpfte, durch Argumentation untermauerte wissenschaftliche Begründung, die dem Adressaten die Möglichkeit gibt, das Wissensfaktum und seine Plausibilität selber zu beurteilen.

Das zweite Problem betrifft den Begriff der Wahrscheinlichkeit, der mit dieser Art von Unsicherheit verbunden ist. Gemeint ist hier nicht die objektivistische, mathematisch berechenbare Wahrscheinlichkeit,[15] die üblicherweise aufgrund von empirisch beobachteten Vorgängen auf das mögliche Eintreten zukünftiger Ereignisse schließt. Auf unseren Fall ist der subjektive Wahrscheinlichkeitsbegriff[16] anzuwenden, bei dem die persönliche Einschätzung als Maß für die Sicherheit eines Sachverhalts dient. Da es sich hier um gemeinschaftliches Wissen handelt, ist es nicht die Einschätzung des Einzelnen, die zählt, sondern die der wissenschaftlichen Community. Daher ist der Begriff ›Plausibilität‹ zu bevorzugen, um den Aspekt der Zustimmung zu betonen.

Zwar kann Plausibilität als Zahlenwert, und zwar als Quotenverhältnis[17] dargestellt werden, das Problem liegt jedoch darin, dass die Zahlenwerte sich aus dem Verhältnis der vorhandenen Alternativen zueinander ergeben. Plausibilität ist ein relativer Wert, er kann nicht absolut angegeben werden: Eine Vermutung verliert an Wahrscheinlichkeit, wenn eine plausiblere hinzukommt. Wenn mehrere alternative Sachverhalte mit ihrer jeweiligen Plausibilität festgehalten werden sollen, dann steht jeder Zahlenwert in einem festgelegten Verhältnis zu jedem anderen und zur Gesamtsumme. Dies hat zur Folge, dass man nicht einen Plausibilitätswert verändern, hinzufügen oder löschen kann, ohne dass alle anderen angepasst werden müssten. Andernfalls würde sich das Gesamtverhältnis verschieben, und die Datenbank geriete in einen inkonsistenten Zustand – es sei denn, in solchen Fällen würde ein automatischer Korrekturmechanismus greifen.

Aber auch, wenn es diesen gäbe, wäre es schwierig, das Verhältnis von Plausibilitäten verschiedener Sachverhalte zueinander mit einem Algorithmus zu ermitteln. Wonach soll geurteilt werden? Nach einem ähnlich vagen *credibility factor* der ForscherInnen, welche die jeweilige Aussage getätigt haben? Nach der Zahl der angegebenen Quellen? Nach einer Abstimmung durch die wissenschaftlichen Benutzer? Die Absurdität und der fragwürdige wissenschaftliche Nutzen einer solchen Berechnung liegen auf der Hand.

Kulturhistorische Aussagen bestehen außerdem oft aus einer Vielzahl von abhängigen Sachverhalten, die sich kaum separat beurteilen lassen. Wenn ein Kunstwerk einer Person zugeschrieben wird, impliziert dies Überlegungen zur Datierung, zur Malweise, zum Entstehungskontext, zur Schaffensphase und so weiter. In der Praxis wird man nicht die einzelnen Parameter separat hinsichtlich ihrer Stichhaltigkeit bestimmen und daraus die wahrscheinlichste mutmaßliche ›Realität‹ berechnen, sondern mehrere Relationen bilden, die unterschiedliche wissenschaftliche Meinungen darstellen. Ein Beispiel ist der ZUCCARO-Datensatz zu einer Zeichnung in der Berliner Kunstbibliothek (Abbildung 2). Zu dem Blatt, das wohl in Rom entstanden ist und vermutlich einen Entwurf für eine Villa zeigt, gibt es voneinander abweichende wissenschaftliche Meinungen zur Bestimmung und zur Autorschaft. Diese sind in den zugehörigen Relationen in Plausibilitätswerten, Kommentaren und zusätzlichen bibliographischen und archivalischen Belegen ausgedrückt.[18]

[15] Wikipedia: Objektivistischer Wahrscheinlichkeitsbegriff.
[16] Wikipedia: Subjektiver Wahrscheinlichkeitsbegriff.
[17] Wikipedia: Chancenverhältnis.
[18] Vgl. Kieven / Schelbert 2014.

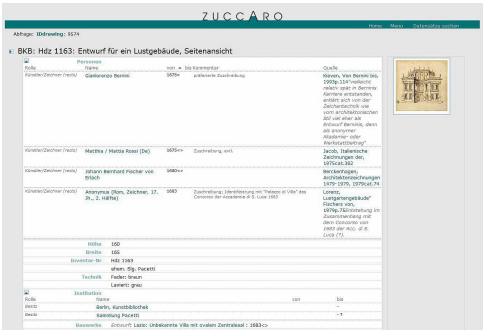


Abb. 2: Unbekannter Künstler (Gianlorenzo Bernini?): Entwurf für ein Lustgebäude, vermutlich eine Villa in Rom. Datensatz in ZUCCARO. CC-BY-NC-SA 4.0.

Erneut stellt sich also die Frage, wie der Unsicherheitsfaktor in einer Datenbank eingebaut werden kann, und erneut kann die abgebildete Wirklichkeit nicht als Kriterium dienen. Wieder müssen wir festhalten: Es kommt darauf an, wozu wir den Aspekt der Plausibilität eigentlich benötigen, was der Datenbank-Benutzer daraus ersehen kann beziehungsweise welchen Zweck wir mit der Datenbank anstreben.

2.3 Eher nicht: Negative Aussagen als Basis kulturwissenschaftlichen Wissens

Kulturhistorisches Wissen, insbesondere die dazu notwendige Argumentation, basiert nicht selten auf dem Ausschlussprinzip. Häufig wissen wir lediglich positiv, dass eine bestimmte historische Beziehung mit Sicherheit niemals bestanden hat. Mit treffendem Witz hat der Karikaturist Freimut Woessner den Vorgang beispielhaft in einem Cartoon dargestellt, der 1991 anlässlich der Ausstellung »Rembrandt – Der Meister und seine Werkstatt« im Berliner Stadtmagazin »zitty« erschienen ist (Abbildung 3). Die in der Karikatur erzählte Geschichte bezieht sich auf das Faktum, dass die Mitglieder des Rembrandt Research Project 1986 das Gemälde mit dem Titel *Der Mann mit dem Goldhelm*, bis dahin eine Zimelie der Berliner Museen, aus dem eigenhändigen Oeuvre des Meisters ausgeschieden hatten. Diese neue wissenschaftliche Erkenntnis erregte großes Aufsehen und wurde mit Erstaunen, gelegentlich auch mit Empörung aufgenommen. Die Berliner Ausstellung von 1991 stellte die Ergebnisse des Projekts einem größeren Publikum vor Augen.

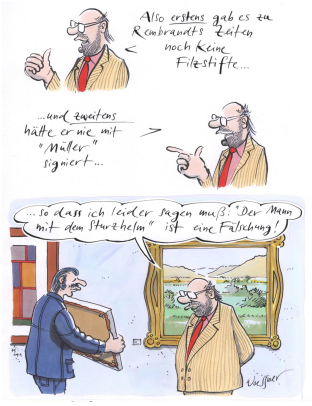


Abb. 3: Freimut Woessner: Der Mann mit dem Sturzhelm, 1991, Zeichnung, Archiv des Künstlers. Mit freundlicher Genehmigung des Künstlers.

In Woessners Bildergeschichte wird der kunsthistorische Argumentationsprozess auf die Schippe genommen. Angesichts eines ihm zu Prüfung vorgelegten Gemäldes bringt der Experte zwei Einwände gegen eine Zuschreibung an Rembrandt, um dann schlussfolgernd darzulegen, dass es sich um eine Fälschung handeln müsse. Die Komik besteht darin, dass abgesehen davon, dass die vorgeführte Filzstiftzeichnung anscheinend einen Mann im Helm zeigt, offenkundig überhaupt kein Bezug zu dem niederländischen Meister des 17. Jahrhunderts besteht und insofern die Bezeichnung als ›Fälschung‹ völlig absurd ist.

Kulturhistorisches Wissen besteht – und das ist charakteristisch für unsere Disziplinen – nicht selten in solchen negativen Statements. Der Experte kann keine positive Zuschreibung vornehmen, sondern lediglich aussagen, dass zwischen der Filzstiftzeichnung und Rembrandt keine Beziehung besteht. Wenn wir diesen Sachverhalt in unser Datenmodell übertragen wollen, so tritt der paradoxe Fall ein, dass wir eine Beziehung zwischen den zwei Entitäten ›Rembrandt‹ und ›Filzstiftzeichnung‹ anlegen müssen (also eine Kante zwischen zwei Knoten), um damit auszusagen, dass gerade *keine* Beziehung besteht.

Man könnte sich hier aus der Verlegenheit helfen, indem man kulturwissenschaftlich festgestellte Nicht-Beziehungen definiert als normale Beziehungen mit einer Plausibilität von Null. In unserem Fall gibt daneben keine weitere Beziehung zwischen Werk und Künstler, denn der Urheber ist ja unbekannt. Die Gleichbehandlung von Beziehungen und Nicht-Beziehungen führt also zu einer weiteren Paradoxie, nämlich dass sich beide systematisch nicht mehr unterscheiden lassen. Praktisch bedeutet das, dass in einer Auflistung der Werke Rembrandts auch Werke erscheinen, die überhaupt nichts mit dem Meister zu tun haben – und je mehr solche negativen Statements wir in die Datenbank aufnehmen, umso deutlicher treten diese in Erscheinung. Aus einem Wissensgraphen würde ein Graph des Nichtwissens. »Die Welt ist alles, was nicht *nicht der Fall* ist«: Eine Datenbank, die für sämtliche Verbindungen, die mit Gewissheit auszuschließen sind, Beziehungsdatensätze anlegt, würde sich selber ad absurdum führen.

Aus dem Dilemma gibt es keinen prinzipiellen Ausweg, denn man kann derartige affirmativ-negative Aussagen auch nicht dadurch darstellen, dass man sie einfach weglässt. Es besteht nämlich grundsätzlich keine Unterscheidungsmöglichkeit zwischen einer Angabe, die prinzipiell unbekannt ist, und einer Angabe, die lediglich nicht in die Datenbank eingetragen wurde. Wenn zu einem Kunstwerk kein Künstler angegeben ist, heißt dies nicht, dass es keinen Urheber gab, es heißt auch nicht, dass der Urheber unbekannt ist. Es bedeutet lediglich ›keine Information eingetragen‹. Es wäre absurd, die Datenbank überall dort, wo Angaben tatsächlich nicht bekannt oder verfügbar sind, mit Aussagen über eben diesen Sachverhalt zu füllen. Trotzdem ist es in der Kulturgeschichte durchaus üblich, auch Unbekanntes positiv festzuhalten, etwa bei anonymen, aber künstlerisch profilierten Meistern Notnamen zu vergeben, oder mit allgemeinen Bezeichnungen wie ›Niederländischer Maler des 15. Jahrhunderts‹ zu arbeiten. Insbesondere im zweiten Fall wird deutlich, dass zu einer näheren Bestimmung der Person durchaus weitere Kriterien vorhanden sind. Man könnte statt der Künstlerperson eine abstrakte Personengruppe anlegen, deren Merkmale sind, dass sie ungefähr im 15. Jahrhundert bestanden hat und mit der kulturellen Region Niederlande in Beziehung stand. Dadurch wird vermieden, dass man eine Unzahl anonymer Künstler anlegen muss, die jeweils einzeln näher zu spezifizieren sind. Hier wird erneut deutlich, dass eine Datenbank kein Abbild der geschichtlichen Realität sein kann: Eine solche Personengruppe hat historisch nicht existiert. Sie ist lediglich ein Platzhalter, der zur Organisation des Wissens dient.

Zu einem ähnlichen Schluss wird man kommen, wenn man das Problem der negativen Aussagen dadurch löst, dass man es auf solche Fälle beschränkt, wo die Forschung auf vergleichbare Weise im Ausschlussverfahren gearbeitet und sich vorläufig auf negative Aussagen zurückgezogen hat. Wo es einen Anlass gab, eine Urheberschaft Rembrandts zu vermuten, da kann man auch positiv festhalten, wenn sich herausstellt, dass sich diese Beziehung nicht bestätigt hat. Diesen Fall zeigt auch die Karikatur, wenn sie als Beleg zwei weitere Negativaussagen anführt. Ob es allerdings sinnvoll ist, Selbstverständlichkeiten digital festzuhalten wie ›Rembrandt hat nicht mit Müller signiert‹, steht dahin. Nicht jedes Argument einer kunsthistorischen Erörterung muss in einem Informationssystem Aufnahme finden. Andererseits sind Argumentationsketten wie ›Rembrandt hat keine Filzstifte verwendet, weil es sie zu seiner Zeit noch nicht gab‹ durchaus nicht trivial. In bestimmten Fällen kann es durchaus sinnvoll sein, sie in einem Informationssystem zu dokumentieren. Sie in Beziehungen zu zerlegen und im Rahmen eines *property graph* zu modellieren, macht die Sache allerdings beträchtlich komplizierter.

Zwei Dinge werden daran deutlich. Erstens: Negative Aussagen sind nicht dazu geeignet, die historische Wirklichkeit zu repräsentieren. Dennoch haben sie in einem kulturhistorischen Informationssystem ihre Berechtigung. Zweitens: Am Beispiel der negativen Aussagen wird deutlich, dass eine Datenbank nicht die reale Vergangenheit, sondern nur das durch Forschung erworbene Wissen darüber widerspiegeln und verwalten kann. Dieses wird immer lückenhaft sein. Die Unsicherheit ist eine Funktion der systembedingten Unvollständigkeit. Auch eine Forschungsdatenbank kann uns nichts wirklich Neues lehren, sie kann nicht selber Forschung treiben und automatisch Wissenslücken füllen. Sie lässt uns lediglich bereits Bekanntes rascher

wiederfinden als zuvor. Darüber hinaus ermöglicht sie uns, durch intelligente Strukturierung anhand eines Datenmodells das scheinbar altbekannte Material nahezu mühelos neu anzuordnen und aus anderen Blickwinkeln zu betrachten.

3. Letztlich alles eine Frage der Darstellung ...

3.1 Das Problem des ranking in der Wissensrepräsentation

Der Fall der negativen Beziehungen verweist auf ein weiteres, grundlegendes Problem kulturwissenschaftlicher Informationssysteme. Ausschlaggebend für die Vermittlung des gespeicherten Wissens ist nämlich nicht allein das Datenmodell, durch welches die Informationen in kleinste Einheiten zerlegt werden, sondern in mindestens gleichem Umfang die Frage nach der Anordnung der Relationen bei der Ergebnisausgabe im Rahmen einer Datenbankabfrage. Nach welchen Kriterien werden unscharfe beziehungsweise unsichere Beziehungen sortiert? Vom *ranking* der Suchergebnisse nach wissenschaftlicher Relevanz hängt die Verständlichkeit des Materials und Vertrauenswürdigkeit der im System gespeicherten Informationen in hohem Maße ab. Lässt sich die unterschiedliche inhaltliche Gewichtung im Hinblick auf Genauigkeit oder Plausibilität so allgemeingültig quantifizieren, dass damit ein Sortier-Algorithmus umgehen kann? Welche Rolle spielen dabei der Abfragekontext, das Erkenntnisinteresse und der subjektive Blickwinkel des Benutzers? Ist es überhaupt zulässig, das Ranking intelligenten Algorithmen zu überlassen, oder ist es Bestandteil der wissenschaftlichen Aussage und muss daher separat modelliert werden? Diese Fragen sollen hier nur anhand einiger praktischer Beispiele aufgeworfen, aber nicht abschließend beantwortet werden.

Die Kehrseite des Umgangs mit Ungenauigkeit und Plausibilität, nämlich das Problem der Anordnung und Visualisierung, wurde im Rahmen der internen Evaluation von *ZUCCARO* immer wieder deutlich. Nicht selten gab es beim wissenschaftlichen Publikum Reaktionen wie ›da ist ja alles durcheinander‹, ›und was ist nun wichtig?‹. Einen charakteristischen Fall bildeten die Einträge zu der heutzutage nicht mehr existierenden *Villa del Pigneto Sacchetti* von Pietro da Cortona.[19] Um diese aufzurufen, wurde über das Abfrageformular nach einem Bauwerk gesucht, dessen Bezeichnung die Zeichenkette ›Villa Sacchetti‹ enthält.

Zwar fand das System alle zu der Villa gehörenden Daten, das Ergebnis befriedigte die Erwartungen dennoch keineswegs: In der Ausgabetabelle wurde die Trefferliste alphabetisch angeordnet. Dadurch erschienen sämtliche Teilgebäude der Villa, das Casino, die Grotte, das Nymphäum zuoberst. Der Haupteintrag der Villa erschien erst ganz unten, weil bei ihm die Bezeichnung mit dem Buchstaben ›V‹ anfängt.

[19] Vgl. hierzu auch Kieven 2011.

Selbst die semantische Hierarchisierung der Bestandteile untereinander durch Beziehungen wie ›ist enthalten in‹ oder ›bildet Einheit mit‹ reichte nicht aus, um für den Benutzer Klarheit zu schaffen. Ursache dafür ist letztlich erneut die Diskrepanz zwischen dem Zwang zur Festlegung und der daraus resultierenden Trennschärfe des Digitalen im Angesicht der fließenden Gegebenheiten in der Realität. Der Ausgabe-Algorithmus war nicht darauf vorbereitet, in der Trefferliste die hierarchische Struktur zu erkennen und diese in der Anordnung zu berücksichtigen. Inzwischen wurde die Ausgabelogik für diesen Fall geändert (Abbildung 4), aber natürlich müssten noch viele andere Ausnahmen berücksichtigt werden. Der *property graph* allein reicht nicht aus, einen vernünftigen Überblick zu gewährleisten, damit der Zusammenhang des Wissens nicht durch die Fülle an Detailinformationen verloren geht.

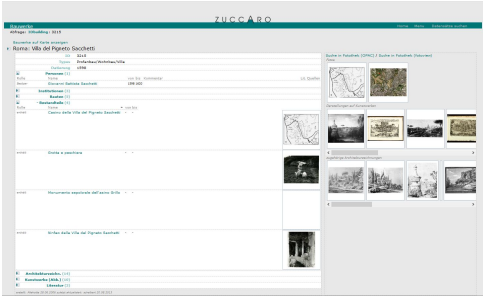


Abb. 4: Rom, Villa del Pigneto Sacchetti. Datensatz in ZUCCARO. CC-BY-NC-SA 4.0.

Es wird deutlich, dass ohne die zusätzliche Berücksichtigung semantischer Zusammenhänge die visuelle Repräsentation des Wissens in einem Informationssystem so unverständlich ausfallen kann, dass der Benutzer eher verwirrt als informiert wird. Dies trifft besonders dann zu, wenn er nicht weiß, mit welchen Ordnungskriterien der Ausgabe-Algorithmus arbeitet. Die Anordnung der Daten nach ihrer Relevanz für den Benutzer erfordert eine tiefere semantische Analyse sowohl der Abfragesituation als auch der im System enthaltenen Ergebnismenge – und dies, obwohl mit der Suche nach einem Bauwerk bereits ein inhaltlicher Kontext vorgegeben war.

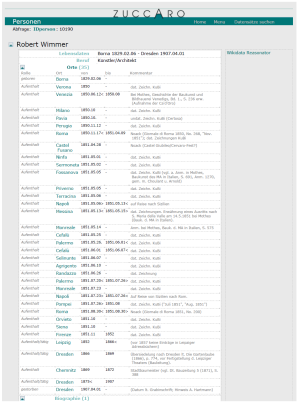


Abb. 5: Aufenthalte und Reisen des sächsischen Architekten Robert Wimmer. Datensatz in ZUCCARO. CC-BY-NC-SA 4.0.

Die Anordnung von Suchergebnissen nach Relevanz, Bedeutung oder Umfang eines Sachverhalts wird außerdem erschwert durch die oben beschriebenen Unverhältnismäßigkeiten, die aus der digitalen Segmentierung des Materials resultieren. Lebensstätten, Reisestationen, Studien- und Wirkungsorte eines Künstlers lassen sich in einem Graphenmodell hervorragend und – im Gegensatz zum Karteiformular – in beliebiger Tiefe abbilden. In ZUCCARO wird jeder Aufenthalt als eine zeitlich und modal definierte Beziehung zu einer Lokalität ausgedrückt. Als Beispiel diene hier die – soweit bislang bekannte – Biographie des sächsischen Architekten Robert Wimmer (1829–1907), dessen in der Kunstbibliothek Berlin aufbewahrte Skizzen einer Italienreise kürzlich ausgewertet werden konnten (Abbildung 5). In der Standardansicht erscheint jeder dieser Aufenthalte gleichwertig, ob es sich um einen mehrjährigen Studienaufenthalt handelt oder nur um eine kurze, aber dokumentierte Reisestation. Immerhin enthält der Datensatz die Zeitdauer, so dass die Relevanz mit einem geeigneten Visualisierungsinstrument, etwa dem Geo-Browser von DARIAH-DE,[20] anschaulich besser gewürdigt werden kann (Abbildung 6).

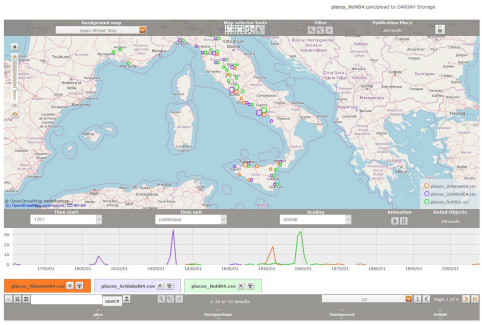


Abb. 6: Visualisierung eines Teils der Reisestationen des Architekten Robert Wimmer im Geo-Browser von DARIAH-DE. CC-BY-NC-SA 4.0.

[20] Beschreibung des Geo-Browsers und der Benutzeroberfläche.

3.2 Wissensrepräsentation als fragestellungsabhängige Visualisierung

Im Umgang mit Unsicherheiten und Unschärfen spielt die Anordnung des Wissens eine vergleichbar wichtige Rolle. Eine Standardlösung gibt es auch hier nicht. Dazu genügt es, sich die Probleme bei der Anordnung von sowohl präzise als auch unscharf datierten Datensätzen vor Augen zu führen. Hat man einen Datensatz, wo als Datierung der Zeitraum 1408–1415 eingetragen ist, und einen anderen, der präzise auf 1411 datiert ist – welcher wird zuerst ausgegeben? Und was geschieht mit einem weiteren Datensatz, der den Zeitraum 1405–1420 umfasst?

Bei der Anordnung des Wissens spielt der jeweilige Fragekontext eine wichtige Rolle. Nicht in jedem Fall kann das gleiche Sortierkriterium angewandt werden. Wenn man einen Künstler betrachtet, dann möchte man seine Werke vermutlich in der Reihenfolge ihres Entstehens aufgeführt haben; vielleicht möchte man außerdem zunächst nur die sicher zugeschriebenen Werke sehen und die unsicheren später auf Wunsch einblenden. Betrachtet man die Werke aus der Sicht einer Sammlung, dann möchte man sie vielleicht lieber nach Ankaufsdatum oder nach dem Saal, in dem sie hängen, geordnet sehen. Was geschieht in diesen Fällen mit Datensätzen, die zu wenige oder ungenaue Informationen enthalten? Hierzu sind vermutlich für jeden Einzelfall eigene Überlegungen anzustellen. Festzuhalten bleibt, dass *Relevanz* in jedem Betrachtungskontext anders definiert ist. Ob ein befriedigender generischer Algorithmus für das *ranking* überhaupt existiert, muss offenbleiben.

Es ist darüber hinaus grundsätzlich fraglich, ob ein Ergebnis, das durch *opaque algorithms*, also undurchsichtige Berechnungen, zustande kommt, überhaupt als Repräsentation von Wissen aufgefasst werden kann. Zur Wissenschaft gehört die Überprüfbarkeit, und die ist in diesem Fall nicht gegeben. Demzufolge wäre es nur konsequent, das jeweilige Ranking, also die Gewichtung der Datensätze bei der Ausgabe, bei der Datenredaktion explizit festzulegen. Dem stehen jedoch große Schwierigkeiten entgegen. Die eine liegt darin, dass bei einem *property graph* jeder Knoten in den Mittelpunkt der Betrachtung rücken kann. Daher müssten explizite Angaben zum Ranking der zugehörigen Beziehungen im Knoten selber gespeichert und dort gegebenenfalls auch modifiziert werden. Es ist fraglich, ob sich der damit verbundene *overhead* nicht negativ auf die Komplexität des Systems und seine Performanz auswirken würde. Die andere Schwierigkeit entsteht dadurch, dass nicht immer ein einzelner Knoten betrachtet wird, sondern oft eine ganze Auswahl, wie in dem genannten Beispiel der Villa Sacchetti. In diesem Fall gibt es keinen Speicherort für Ranking-Angaben im Datenbestand des Wissensgraphen, sondern das System selbst müsste darüber Buch führen. Wie dies im Datenmodell aussehen könnte, wäre gesondert zu überlegen.

Die Visualisierung von Unsicherheiten und Unschärfen ist ein eigenes Gebiet, das von der Forschung auch in anderen Fächern längst thematisiert worden ist.[21] In ZUCCARO haben wir zum Beispiel damit experimentiert, bei der Zuschreibung einer Architekturzeichnung an einen historischen Zeichner den jeweiligen Grad der Gewissheit mit unterschiedlichen Farben zu kodieren (Abbildung 7).

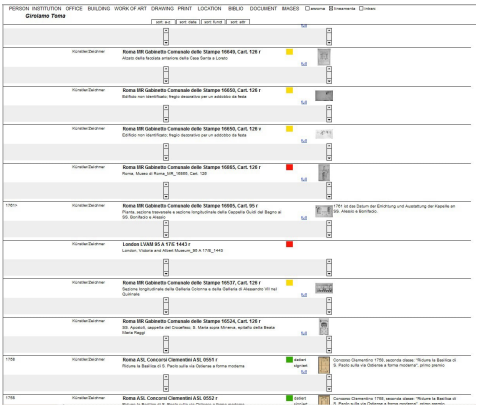


Abb. 7: Kodierung der Zuschreibungswahrscheinlichkeit mit Farben in ZUCCARO (Desktopansicht 2007). CC-BY-NC-SA 4.0.

Besonders augenfällig wird das Thema im Bereich der digitalen räumlichen Modellierung. Ein CAD-Modell kann einen Mauerzug prinzipiell nur in einer ganz konkreten Ausdehnung und Position wiedergeben. Verglichen mit den tatsächlichen historischen Kenntnissen sind die im CAD-System gespeicherten Maßangaben demzufolge oft viel zu präzise und spiegeln dadurch eine Genauigkeit des Wissens vor, die keine fundierte Grundlage hat. Die virtuelle Präzision ist die digitale Kehrseite der *fuzziness* des Wissens. Daher haben sich auf diesem Gebiet bereits verschiedene Strategien im Umgang mit Unsicherheit und Unschärfe herausgebildet.[22]

Unsicherheit, Unvollständigkeit und Granularität der Daten werfen weitere Fragen auf. Zwar legt Datenhaltung in einem *property graph* die Möglichkeit nahe, darauf Methoden des automatisierten *reasoning* anzuwenden, also mit Hilfe von Netzwerkalgorithmen implizites Wissen in explizites umzuwandeln und dadurch den Datenbestand zu konsolidieren. Wie weit solche Verfahren durch die genannten Probleme behindert würden, wäre zu prüfen.

Ein automatisiertes Ranking beispielsweise, das sich auf die Anzahl der verknüpften Elemente stützt, oder auf die Häufigkeit von Anfragen, oder Benutzerbewertungen, dürfte stets fehlerbehaftet bleiben. Ebenso problematisch erscheint es, aus den vorhandenen Inhalten automatisiert Schlussfolgerungen zu ziehen oder statistische Auswertungen vorzunehmen. Zwar können derartige Verfahrensweisen den Charakter umfangreicher Simulationen

[21] Vgl. Bonneau et al. 2014; eher praktisch Yau 2018.
[22] Standards für die Kenntlichmachung der hypothetischen Bestandteile eines Modells wurden durch die London Charter am 19.09.2017 definiert. Vgl. speziell hierzu auch – aus eher gestalterischer als modeltheoretischer Sicht: Lengyel / Toulouse 2011 und Lengyel / Toulouse 2015.

annehmen, um nicht ausreichende Informationen zu interpolieren und damit wieder in den Datenbestand zurückzuwirken, doch die Zulässigkeit ist fraglich. Derartige Versuche und Überlegungen werden im Projekt *Venice Time Machine* und im Projektverbund *Time Machine* angestellt (Abbildung 8).[23]

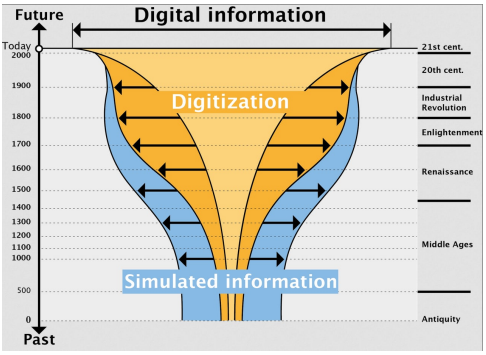


Abb. 8: Frédéric Kaplan: Überbrückung von fehlenden Quellen durch Simulation. Kaplan 2013, fig. 3.

3.3 Fazit

Unsere Erfahrungen bei der zwölfjährigen Arbeit mit dem Informationssystem ZUCCARO sind bescheidener. Als Quintessenz der hier angestellten Überlegungen lässt sich die Erkenntnis formulieren, dass ein kulturhistorisches Informationssystem der Forschung in erster Linie der inhaltlichen Erschließung des historischen Materials dienen sollte. Die Ausarbeitung des Datenmodells und die Algorithmen des Datenbanksystems sollten stärker darauf Rücksicht nehmen, wie man die Inhalte dem Benutzer möglichst übersichtlich und durchschaubar präsentieren kann. Das Datenmodell extrem elaboriert und detailliert zu strukturieren, damit es der historischen Wirklichkeit möglichst nahe kommt, ist demgegenüber zweitrangig und in manchen Fällen sogar kontraproduktiv. Eine möglichst getreue Simulation einer geschichtlichen Wirklichkeit anzustreben, sollte nicht das Ziel eines *knowledge graph* sein. Unsicherheiten im Datenbestand zu dokumentieren und für den Benutzer kenntlich zu machen, ist sinnvoll – sie automatisiert auszuwerten jedoch nicht.

[23] Vgl. Kaplan 2013 und Kaplan 2017.

Bibliographische Angaben

Georges-Pierre Bonneau / Hans-Christian Hege / Chris R. Johnson / Manuel M. Oliveira / Kristin Potter / Penny Rheingans / Thomas Schultz: Overview and State-of-the-Art of Uncertainty Visualization. In: Scientific visualization: uncertainty, multifield, biomedical, and scalable visualization. Hg. von Charles D. Hansen / Min Chen / Christopher R. Johnson / Arie E. Kaufman / Hans HageLondon u.a. 2014, S. 3-27. (= Mathematics and Visualization, 37) [Nachweis im GBV]

Thomas Frisendal: Property Graphs: The Swiss Army Knife of Data Modeling. In: dataversity.net. Big Data Blogs. Blogbeitrag vom 22.09.2017. [online]

Edith Hamilton: Spokesmen for God. The Great Teachers of the Old Testament. New York 1936. [Nachweis im GBV]

Frédéric Kaplan: Lancement de la »Venice Time Machine«. In: fkaplan.wordpress.com. Frederic Kaplan. Blogbeitrag vom 14.03.2013. [online]

Frédéric Kaplan / Isabella di Lenardo: Big Data of the Past. In: Frontiers in Digital Humanities (2017). Artikel vom 29.05.2017. DOI : 10.3389/fdigh.2017.00012

Elisabeth Kieven / Georg Schelbert: Architekturzeichnungen, Architektur und digitale Repräsentationen. Das Projekt LINEAMENTA. In: kunsttexte.de 4 (2014). DOI: 10.18452/6832

Elisabeth Kieven: Research Infrastructures for Historic Artefacts: Knowledge Networks. In: Research Infrastructures in the Digital Humanities. Hg. von der European Science Foundation. (ESF Science Policy Briefing: 42, Straßburg, 09.2011) Straßburg 2011, S. 13-15. [online]

Dominik Lengyel / Catherine Toulouse: Darstellung von unscharfem Wissen in der Rekonstruktion historischer Bauten. In: Von Handaufmaß bis High Tech III. 3D in der historischen Bauforschung. Hg. von Katja Heine. Darmstadt u.a. 2011, S. 182-186. [Nachweis im GBV]

Dominik Lengyel / Catherine Toulouse: Die Bedeutung architektonischer Gestaltung in der Vermittlung von Unschärfe. In: gams.uni-graz.at. Präsentation vom 25.02.2015. (DHd 2015, Graz, 23.-27.02.2015) Graz 2015. [online]

Alan M. MacEachren: Visualizing Uncertain Information. DOI: 10.14714/CP13.1000 In: Cartographic Perspectives 13 (1992), S. 10-19. [online] [Nachweis im GBV]

Martin Raspe / Georg Schelbert: ZUCCARO. Ein Informationssystem für die historischen Wissenschaften. In: IT Information Technology 51 (2009), H.4, S. 207-215. DOI: 10.11588/artdok.00005812 [Nachweis im GBV]

Nathan Yau: Visualizing the Uncertainty in Data. In: flowingdata.com. Guides. Beitrag vom 08.01.2018. [online]

Abbildungslegenden und -nachweise

Abb. 1: Schematisches Datenmodell des Informationssystems ZUCCARO (Stand 2013). CC-BY-NC-SA 4.0.

Abb. 2: Unbekannter Künstler (Gianlorenzo Bernini?): Entwurf für ein Lustgebäude, vermutlich eine Villa in Rom. Datensatz in ZUCCARO. CC-BY-NC-SA 4.0.

Abb. 3: Freimut Woessner: Der Mann mit dem Sturzhelm, 1991, Zeichnung, Archiv des Künstlers. Mit freundlicher Genehmigung des Künstlers.

Abb. 4: Rom, Villa del Pigneto Sacchetti. Datensatz in ZUCCARO. CC-BY-NC-SA 4.0.

Abb. 5: Aufenthalte und Reisen des sächsischen Architekten Robert Wimmer. Datensatz in ZUCCARO. CC-BY-NC-SA 4.0.

Abb. 6: Visualisierung eines Teils der Reisestationen des Architekten Robert Wimmer im Geo-Browser von DARIAH-DE. CC-BY-NC-SA 4.0.

Abb. 7: Kodierung der Zuschreibungswahrscheinlichkeit mit Farben in ZUCCARO (Desktopansicht 2007). CC-BY-NC-SA 4.0.

Abb. 8: Frédéric Kaplan: Überbrückung von fehlenden Quellen durch Simulation. Kaplan 2013, fig. 3.

ZfdG

Zeitschrift für digitale Geisteswissenschaften

Beitrag aus:
Sonderband 4 der ZfdG: Die Modellierung des Zweifels – Schlüsselideen und -konzepte zur graphbasierten Modellierung von Unsicherheiten. Hg. von Andreas Kuczera, Thorsten Wübbena und Thomas Kollatz. 2019. DOI: 10.17175/sb004

Titel:
Graphbasierte Modellierung von Faktenprovenienz als Grundlage für die Dokumentation von Zweifel und die Auflösung von Widersprüchen

Autor/in:
Thomas Efer

Kontakt:
efer@informatik.uni-leipzig.de

Institution:
Universität Leipzig, Institut für Informatik

GND:
1125649186

ORCID:
0000-0002-8376-3884

DOI des Artikels:
10.17175/sb004_011

Nachweis im OPAC der Herzog August Bibliothek:
1037074947

Erstveröffentlichung:
14.08.2019

Lizenz:

Sofern nicht anders angegeben (cc) BY-SA

Medienlizenzen:
Medienrechte liegen bei den Autoren

Letzte Überprüfung aller Verweise:
14.08.2019

GND-Verschlagwortung:
Geschichtswissenschaft | Graphdatenbank | Konzeptionelle Modellierung | Wissensrepräsentation |

Zitierweise:
Thomas Efer: Graphbasierte Modellierung von Faktenprovenienz als Grundlage für die Dokumentation von Zweifel und die Auflösung von Widersprüchen. In: Die Modellierung des Zweifels – Schlüsselideen und -konzepte zur graphbasierten Modellierung von Unsicherheiten. Hg. von Andreas Kuczera, Thorsten Wübbena und Thomas Kollatz. 2019 (= Sonderband der Zeitschrift für digitale Geisteswissenschaften, 4). PDF Format ohne Paginierung. Als text/html abrufbar unter DOI: 10.17175/sb004_011.

Thomas Efer

Graphbasierte Modellierung von Faktenprovenienz als Grundlage für die Dokumentation von Zweifel und die Auflösung von Widersprüchen

Abstracts

Ziel dieses Beitrags ist es, die Wichtigkeit einer nachvollziehbaren Herkunft von Aussagen in Wissensbasen der Digitalen Geisteswissenschaften herauszustellen. Neben der Vorstellung genereller Aspekte der Aussagenmodellierung auf abstrakter und beispielgeleiteter Ebene wird das Konzept einer Faktenprovenienz entwickelt und in Aussagemodelle integriert. Auf Basis von Provenienzketten wird demonstriert, wie eine im System erfasste Herkunftsdokumentation von Einzelaussagen zur Behandlung von Widersprüchen und der Reduzierung von Unsicherheit genutzt werden kann.

This contribution aims to demonstrate the importance of traceable provenance information within knowledge bases in the Digital Humanities. Besides presenting rather general aspects of how to model statements in an abstract and in an exemplary manner, the concept of fact provenance is introduced and integrated with statement expression models. Using so-called provenance chains, it is shown how provenance information that is captured within an information system can be utilized to handle contradictions and reduce the overall uncertainty of the knowledge base.

1. Motivation und Einführung

Die Geistes- und Sozialwissenschaften befinden sich gegenwärtig unter dem Eindruck und innerhalb der Dynamik einer sich rapide digitalisierenden Gesellschaft vor zahlreichen neuen Herausforderungen.

Wird von *den* Geistes- und Sozialwissenschaften gesprochen, schwingt die gewagte Grundannahme einer Kohärenz der darunter subsumierten wissenschaftlichen Akteure und Aktivitäten mit. Diese lässt sich bei genauerer Betrachtung höchstens für kleine Teilbereiche einzelner Fachrichtungen oder gewisse interdisziplinäre Querschnittsthemen rechtfertigen. Während eine solche bewusst gleichmachende Abstraktion sehr hilfreich für die Initiierung interdisziplinärer Arbeiten ist und darüber hinaus gewinnbringend für die Institutionalisierung und Lobbyarbeit genutzt werden kann, so kann sie problematisch werden: Oft zeigt sich dann erst spät in der Praxis, wie groß und zum Teil unüberwindbar die Differenzen in inhaltlicher und methodischer Dimension tatsächlich sind. Die Forschungsziele der einzelnen Disziplinen und konkreten Einzelforschungsarbeiten verteilen sich (entsprechend dem akademischen Selbstverständnis der Fächer und Forscher*innen) meist sehr breit auf der Skala zwischen dem Erreichen eines konkreten Erkenntnisgewinns und der davon angekoppelten Interpretation bestimmter Sachverhalte in neuen, aktuellen Kontexten. Entsprechend zeigen sich auch deutliche Unterschiede im Umgang mit Quellenmaterial, sekundärer Fachliteratur, vorherrschenden Fächertraditionen, disziplinären Strömungen und einzelnen fachlichen Aussagen im Forschungsprozess – insbesondere im Umgang mit allen Arten von »Daten«.

Im Rahmen der Digital Humanities wird die Unterstützung jeglicher geistes- und sozialwissenschaftlicher Forschungstätigkeit durch »generische« digitale Werkzeuge für diese wichtigen datenzentrierten Forschungstätigkeiten angestrebt. Dafür müssen geeignete Abstraktionen und Verallgemeinerungen gefunden werden. Bei der Überführung bisheriger Forschungstätigkeiten in die digitale Welt werden somit auch digitale Modelle für Daten benötigt, die den Grundbedürfnissen der Disziplinen gerecht werden. Da sich unbestritten ein Großteil der Forschungsarbeiten im Einzugsbereich der DH mit dem Sichten, Sammeln, Erschließen, Bewerten und Verknüpfen von Befunden aus Quellenmaterial beschäftigt (vgl. dazu die Taxonomy of Digital Research Activities in the Humanities.[1]), liegt es nahe, in diesem Bereich die gemeinsamen Anforderungen zu ergründen und dafür entsprechende digitale Unterstützung bereitzustellen.

Während die meisten Forschungsdatenbanken ausschließlich die Endprodukte dieser Arbeitsschritte beinhalten – welches bereits eine große Unterstützung darauf aufbauender weiterführender Forschung sein kann –, soll im Rahmen dieses Beitrags eine Herangehensweise vorgestellt werden, mit deren Hilfe bei einer Weiter- und Nachnutzung tiefer in die Entstehungszusammenhänge dieses Faktenwissens hineingesehen werden kann.

Mit Fakten sind in diesem Fall vom Menschen »gemachte«, als plausibel angesehene Aussagen gemeint – für die Zwecke digitaler Datenbanken speziell solche Aussagen, die sich formalisiert abbilden lassen. Der Faktenbegriff soll hier noch nicht mit absoluten Kategorien wie einem »tatsächlichen Wahrheitsgehalt« gleichgesetzt werden. Auch bei allen als wahr angenommenen Gegebenheiten handelt es sich (speziell im Kontext der Wissenschaft) schließlich immer nur um »Tatsachenbehauptungen«, die analytisch und interpretativ im Forschungsprozess jederzeit angezweifelt, ausgeklammert, zurückgewiesen oder aber übernommen und kombiniert, und damit in weitergreifende Aussagen überführt werden können.

Um die Fakten qualifiziert einschätzen zu können, ist in erster Linie die Kenntnis ihrer genauen Herkunft von Interesse. Die »Entstehungsumstände« des Faktenwissens und die »Überlieferungshistorie« sind dabei die wesentlichen Bestandteile der in diesem Beitrag Faktenprovenienz genannten Herkunftsinformationen. Nicht immer sind diese Komponenten in ausreichender Form bekannt oder belegt. Faktenprovenienz ist inhaltlich verwandt, aber bei weitem nicht deckungsgleich zur sogenannten Datenprovenienz (wie etwa bei Simmhan et al. beschrieben[2]). Datenprovenienz (auch Data- Lineage genannt) beschreibt Anforderungen und Verfahren, um die Transformationen von digital vorliegenden Quelldaten hin zu den in einem System vorgehaltenen oder aus ihm exportierten Enddaten maschinenlesbar zu dokumentieren. Als Vorteile werden verbesserte oder vereinfachte Qualitätssicherung, Attribution von Rechteinhabern und Reproduzierbarkeit angesehen. Ähnliche Ziele verfolgt auch die Berücksichtigung von Faktenprovenienz in Forschungsdatenbanken. Ihr Fokus liegt jedoch nicht allein auf bereits digital vorliegenden Rohdaten, sondern auf der kompletten Historie der enthaltenen Fakten. Nur durch diese erweiterte Sichtweise kann die Arbeit mit Fakten im Forschungsprozess ganzheitlich unterstützt werden.

[1] Borek et al. 2016
[2] Simmhan et al. 2005, S. 31–36.

Bei den digital in einer Datenbank zu erfassenden Aussagen sind im Kontext von fachfragen-orientierten DH-Methoden in erster Linie Aussagen zu Gegebenheiten der Fachdomäne von Interesse und weniger technische oder organisatorische Metainformationen. Im Umgang mit Forschungsdaten der Fachdomänen stellen sich für Datenbanksysteme dabei sehr grundsätzliche Fragen: Welche Daten sollen erfasst werden? Welche Arten von Aussagen sollen im System abbildbar sein? Wie können erfasste Einträge zueinander in Beziehung gesetzt werden (bei der Eingabe und bei der Abfrage)? Wie können externe Daten übernommen und eigene Daten exportiert werden? Für diese Fragen bieten generische Werkzeuge, Repositoriums- und Datenbanksysteme in der Regel akzeptable bis sehr gute Antworten. Die Provenienz der so in den Datenbanken kodierten Aussagen kann jedoch im Allgemeinen nicht abgebildet werden! Dieser Fehlstelle widmet sich der vorliegende Beitrag.

Für die gemeinsame Speicherung und Abfrage von Fakten und ihrer Provenienz werden Systeme benötigt, mit denen eine sehr flexible Datenmodellierung möglich ist. Ohne Abwertung möglicher Alternativtechnologien soll im Folgenden eine Festlegung auf eine bestimmte Gruppe von dafür geeigneten Systemen getroffen werden.

2. Technologischer Rahmen

Dieser Beitrag bezieht sich auf die Anwendung von Graphdatenbanksystemen, speziell solchen, die das Property-Graph-Datenmodell umsetzen.[3] In diesem Datenmodell stehen für die Repräsentation der zu erfassenden Daten die folgenden einfachen Bausteine zur Verfügung, durch deren Kombination sich komplexere Sachverhalte abbilden lassen:

Die so genannten Knoten (nodes vertices) können als Repräsentanten für Aussagegegenstände angesehen werden. Sie lassen sich zählen, auflisten und über intern vergebene IDs einzeln adressieren.

Daneben stehen mit den so genannten Kanten (edges) Konstrukte zur Verfügung, mit denen genau zwei Knoten miteinander verbunden werden können. Kanten unterscheiden dabei zwischen Start- und Zielknoten, so dass die Verbindung »gerichtet« ist. Zwischen beliebigen Knotenpaaren können beliebig viele Kanten in beliebiger Richtung existieren. Den jeweiligen Zielknoten werden die Kanten dabei als »eingehende« Kanten, dem Startknoten als »ausgehende« Kanten zugeordnet. Kanten besitzen genau ein sogenanntes Label. Dieses ist eine kategoriale Größe, die verwendet wird, um verschiedene (semantisch oder technisch zu unterscheidende) Arten von Beziehungen zwischen den mit einer Kante verbundenen Knoten auszudrücken. Auch Kanten besitzen interne IDs und lassen sich zählen und auflisten. Darüber hinaus ist es möglich, sie nach ihrem Label zu filtern. Das Label hat dabei üblicherweise eine textuelle Repräsentation, wie ›IST_VATER_VON‹ oder ›FOLLOWS‹. Damit ist es möglich und üblich, die Kante (als technische Verbindung von Knoten) auch als Abbild einer zu modellierenden Beziehung zwischen zwei Aussagegegenständen anzusehen.

[3] Vgl. dazu Robinson et al. 2013.

Schließlich existieren mit den sogenannten Properties noch Konstrukte, mit denen sich »Eigenschaften« von Knoten und Kanten notieren lassen. Diese Eigenschaften sind mit maschinenlesbaren Werten befüllt. Erst dadurch ergibt sich der Datenbankcharakter des Systems. Sie werden in Form von Schlüssel-Wert-Paaren (key-value pairs) gespeichert. Diese bestehen aus einem Schlüssel, also einer kategorialen Größe, die den in der Property notierten Eigenschaftstyp bestimmt, und einem Wert, welcher in der Regel in einem primitiven Datentyp, wie etwa Ganzzahl oder Zeichenkette, vorliegt. Die Property-Schlüssel besitzen (genau wie die Kantenlabels) eine textuelle Repräsentation. Alle Schlüssel-Wert-Paare sind jeweils genau einem Knoten oder genau einer Kante zugewiesen. Losgelöst von diesen Konstrukten können sie nicht existieren.

Damit ist die Palette der verwendbaren Modellierungskonstrukte auch schon vollständig. Im nächsten Unterabschnitt des Artikels wird noch genauer auf die damit umsetzbare Datenmodellierung eingegangen. In der Praxis existieren zahlreiche Nuancen dieses grundlegenden Datenmodells. In der populären und systemübergreifend genutzten Programmierschnittstelle für Graphdatenbanken in Java namens Tinkerpop wird erst ab Version 3 erlaubt, einem Knoten oder einer Kante mehrere Properties mit gleichem Schlüssel beizufügen. Damit einher geht auch die Möglichkeit, Properties für Properties zu definieren (welche intern von den Systemen jedoch oft nur mittels der oben beschriebenen Basiskonstrukte »virtuell« umgesetzt wird). Ebenso gibt es unterschiedliche Ansichten darüber, ob im Property-Graph-Modell auch Knoten über ein Label verfügen sollten, also einen Typen haben können (oder müssen), wie z. B. im populärsten System, Neo4j, üblich. Da im Rahmen dieses Beitrags nicht allzu tiefgehend auf direkter technischer Ebene mit den Modellierungskonstrukten gearbeitet wird, sollen diese und weitere Feinheiten an dieser Stelle jedoch nicht weiter eruiert werden.

Graphdatenbanken weisen abseits der Spezialisierung auf dieses Datenmodell viele Gemeinsamkeiten mit klassischen, relationalen Systemen auf. Auch sie bewegen sich im Segment der Echtzeitabfragen im so genannten Online Transaction Processing (OLTP), unterstützen strukturierte Abfragesprachen und oft auch Transaktionen, also eine Persistierung der Änderungen am Datenbestand nach einem Alles-oder-nichts-Prinzip im logischen Einbenutzerbetrieb. Im OLTP-Betrieb ist die Geschwindigkeit der Beantwortung einer Abfrage von großer Bedeutung. Abfragen in Graphdatenbanksystemen liefern meist Mengen von Knoten zurück, welche entweder direkt über die Werte ihrer Properties ausgewählt werden oder aber indirekt durch das Überspringen von Kanten von einem bereits ausgewählten Knoten aus erreicht werden können. Die Nutzung der Kanten zur Navigation innerhalb des durch sie aufgespannten Knotennetzwerks wird Traversierung genannt. Diese Traversierung kann in Graphdatenbanken effizient über sehr viele Zwischenstationen geschehen, wodurch sich meist ein erheblicher Geschwindigkeitsvorteil gegenüber relationalen Datenbanken und deren Tabellenverknüpfung über Joins ergibt.[4]

[4] Vgl. Rodriguez / Neubauer 2011, S. 29–46.

Diesen Geschwindigkeitsvorteil können die Systeme allerdings nur geltend machen, wenn zur Beantwortung der Anfrage ein kleiner, »lokaler« Ausschnitt der Datenbankeinträge (Knoten und Kanten) »besucht« wird. In der Praxis zeigt sich, dass sich nicht wenige Probleme in den geisteswissenschaftlichen Disziplinen durch begrenzte Umkreissuchen um »interessante« Einträge herum ausdrücken lassen. Oft ist die Analyse aller mit einem Objekt verknüpften anderen Objekte interessanter und zielführender, als die aller nichtverknüpften.[5]

Für »globale« Auswertungen, welche den kompletten Datenbestand oder große Teile davon traversieren, beispielsweise um statistische Kennzahlen zu ermitteln, können keine schnellen Antwortzeiten erwartet werden. Ganz im Gegenteil kann es dazu kommen, dass statt einer Antwort sogar ein durch die Überschreitung vorhandener Systemressourcen (meist des Arbeitsspeichers) hervorgerufener Fehler vermeldet wird. Hierfür werden künftig verstärkt Lösungen im Umfeld der Graphentechnologie benötigt, welche neben OLTP auch ein Online Analytical Processing unterstützen und die Systemressourcen unter Aufgabe der Echtzeit-Abfragbarkeit effektiver für Analysezwecke nutzen können. Idealerweise verwenden solche Systeme dasselbe Datenmodell (und dieselben Abfrageschnittstellen und -sprachen) wie Graphdatenbanken und sind ggf. sogar fest mit ihnen verbunden oder in sie integriert. Daher wird im Folgenden nicht weiter auf diese Unterscheidung eingegangen, auch wenn sie für die konkrete Umsetzung in der Praxis sehr relevant ist.

Neben dieser technischen Sichtweise soll die Graphdatenbank hier hauptsächlich als ein Wissensspeicher fungieren. Ähnlich einer Wissensbasis (Knowledge Base) werden darin einzelne Aussagen erfasst, kontextualisiert und für eine strukturierte Abfrage vorgehalten. Während in semantischen Netzwerken vorwiegend Semantic-Web-Technologien zur Anwendung kommen, wird im Folgenden auf maschinenlesbare Semantik auf Schema-Ebene und auf die Möglichkeiten, ein automatisches logisches Schließen (Reasoning) durchzuführen, verzichtet. Dies geschieht mit Hinblick auf Geschwindigkeitsaspekte für die Speicherung, Indizierung und Abfrage großer Datenmengen und um im Speichersystem technische Schemainformationen und Stamm- bzw. Instanzdaten nicht zu vermischen. Eine Überführung von Property Graphs in eine Semantic-Web-Repräsentation ist jedoch jederzeit problemlos möglich, wie beispielsweise die Arbeit von Hartig aus dem Jahr 2014 zeigt.[6] Fehlende technische Möglichkeiten für das generische Reasoning lassen sich in der wissenschaftlichen Anwendungsdomäne zudem leicht verschmerzen, da die so erzielten abgeleiteten Aussagen meist deutlich unspezifischer sind als solche, die durch zielgerichtete Datenbankabfragen mit Kenntnis der Fachdomäne erzielt werden können. Ein solcher Abfragemodus erlaubt einen flexibleren Umgang mit Forschungshypothesen, welche über die nötigen Kontexte (und Konfidenzen), die für eine Folgerbarkeit auf Faktenebene entscheidend sind. Parallel dazu wird ein semantisches Modell des Wissensspeichers in der Regel anwendungsspezifisch durch Festlegung von Geschäftslogik und Interaktionsmöglichkeiten in der Recherchesoftware abgedeckt. Basis aller Folgerung und Interpretation der Daten muss im Kontext von Wissensbasen allgemein und im Kontext der digitalen Geisteswissenschaften im Besonderen

[5] Vgl. Efer 2017.
[6] Hartig 2014.

die so genannte Open World Assumption sein, wie sie etwa von Moore und Pham beschrieben[7] und genauer analysiert wird. Sie sagt aus, dass nicht enthaltene Aussagen nicht zwingend unwahr sind. Ihr Wahrheitsgehalt ist somit ›unbekannt‹.

Mit den hier umrissenen modelltechnischen Möglichkeiten und Einschränkungen lässt sich nun ein System zur Abbildung von Aussagen entwickeln, welches die benötigten Ankerpunkte und Konstrukte für die Annotation von Faktenprovenienz bereitstellt.

3. Möglichkeiten der Modellierung von Aussagen

Die folgenden Beispiele orientieren sich am zu diesem Beitrag gehörigen, gleichnamigen Vortrag auf der Tagung Graphentechnologien 2018 – Die Modellierung des Zweifels in Mainz. Das Ziel hinter der Wahl der eher »informellen« Inhalte ist es, einen vergleichsweise abstrakten und dennoch mit nachvollziehbaren Gegebenheiten verknüpften Zugang zur Aussagenmodellierung zu erreichen – zunächst losgelöst von der wissenschaftlichen Praxis.

Als Ausgangspunkt der Modellierung soll hier zunächst die natürlichsprachlich vorliegende Aussage ›Peter isst eine Banane‹ betrachtet werden. Diese wirft freilich für sich genommen deutlich mehr Fragen auf als sie beantwortet: Welcher Peter ist gemeint? Kann dieser genauer charakterisiert oder gar eindeutig identifiziert werden? Ist es möglich, die mit unbestimmtem Artikel bedachte Banane irgendwie von anderen Bananen zu unterscheiden (oder ist ihr einziges Alleinstellungsmerkmal, von Peter gegessen zu werden)? Was ist der räumlich-zeitliche Kontext der Aussage?

Selbst ohne Kenntnis dieser zusätzlichen Details ist es möglich, die Aussage als einzelnen Fakt in der Datenbank abzubilden. Dazu wird ein Knoten als Repräsentant für ›Peter‹ und ein weiterer Knoten für die in der Aussage referenzierte ›Banane‹ erzeugt. Zwischen beiden wird eine als ›isst‹ typisierte Kante ausgehend von ›Peter‹ eingefügt. Damit ist in der Datenbank immerhin hinterlegt, dass ein Vorgang des ›Essens‹ (Kantentyp ›isst‹) dokumentiert ist.

Diesem Vorgang können nun alle denkbaren Kontextinformationen angefügt werden, beispielsweise der Zeitpunkt, welcher in Form einer Kantenproperty erfasst werden kann. An dem Knoten, der für Peter steht, lässt sich in jedem Fall sein Name als Property erfassen. Eventuelle darüber hinaus bekannte Fakten, wie sein Alter oder die Farbe der Banane ließen sich ebenso als Knotenproperties erfassen. Ebenso kann das implizit oder explizit gegebene Hintergrundwissen, dass Peter ›eine Person‹ und die Banane ›ein Gegenstand‹ ist, als Typenzuweisung für die Knoten abgebildet werden. Dies kann entweder durch die Verwendung von Knotenlabels oder die Erstellung von Repräsentanten für den jeweiligen Typ und die Verwendung einer geeignet typisierten Kante (etwa ›is_a‹ oder ›typeof‹ genannt) realisiert werden.

[7] Moore / Van Pham 2015.

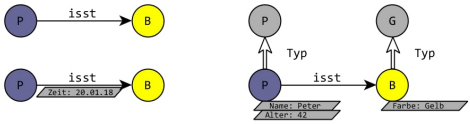


Abb. 1: Modellierung einfacher Aussagen im Graphen. [Efer 2019.]

Abbildung 1 zeigt visuelle Repräsentationen aller bisher verwendeten Modellierungsvarianten für die unterschiedlich stark kontextualisierte Aussage in einer (hoffentlich) intuitiven, jedoch nicht formalisierten oder standardisierten graphischen Notation.

Einfache Aussagen nach diesem Muster lassen sich stets gleich erfassen. Ist klar, dass es sich bei der handelnden Person in einer zweiten Aussage um denselben Peter wie zuvor handelt, so sollte für diesen kein neuer Repräsentant eingefügt, sondern der vorhandene Knoten weiter genutzt werden. Ob innerhalb der Datenbank zur plausiblen Abbildung einer weiteren Aussage dieselbe Banane ein weiteres Mal gegessen werden kann (von Peter oder einer anderen Person), muss dann bereits sehr individuell entschieden werden (indem etwa ›isst‹ auch die Bedeutung ›isst Teile von‹ umfasst). Solche semantischen Probleme sollen hier jedoch zunächst ausgeklammert werden. Stattdessen gilt es, einen weiteren, sehr häufigen und dabei durchaus problematischen Typ von Aussage zu untersuchen:

Lautet die abzubildende Aussage nun nämlich: ›Jürgen sagt, dass Peter eine Banane isst‹, kann der bisherige Ansatz einer ausschließlich direkten Nutzung von Graphkonstrukten nicht mehr verwendet werden. Für Peter, die Banane und die ›isst‹-Kante zwischen ihnen kann noch der selbe Modellierungsansatz wie oben gewählt werden. Auch Jürgen kann einen Knoten als Repräsentanten erhalten. Doch wohin zeigt eine von diesem ausgehende ›sagt-Kante‹? Diese sollte auf die komplette Aussage von oben verweisen. Technisch kann sie jedoch nicht auf ›zwei Knoten und eine Kante‹ verweisen, ebenso wenig auf ›die Kante, an der beide Knoten anliegen‹. Sie kann nur zwei Knoten miteinander verbinden. Es wird also ein Knoten als Repräsentant der obigen »geschachtelten« Aussage benötigt. Die Erstellung eines solchen Knotens wird im Umfeld des Semantic Web Reifizierung genannt. In Graphdatenbanksystemen ist dafür kein eigenes Modellierungskonstrukt vorgesehen. Das Prinzip lässt sich dennoch anwenden, indem die im Semantic Web übliche Verwendung von Tripeln aus Subjekt, Prädikat und Objekt als Abstraktion verwendet wird.

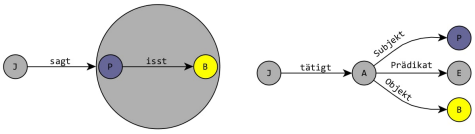


Abb. 2: Modellierung von Aussagen über Aussagen mittels Reifizierung. [Efer 2019.]

Abbildung 2 zeigt, wie ein separater Knoten, welcher für die Aussage (A) steht, mit dem Repräsentanten von Peter, dem der Banane, sowie einem separaten Knoten, der für ›den Vorgang des Essens‹ (E) steht, über entsprechende Kanten (Subjekt, Objekt und Prädikat)

verknüpft ist. Nun kann das »Tätigen« der so formell beschriebenen Aussage durch das Verbinden des Repräsentanten für Jürgen mit diesem Aussageknoten abgebildet werden. Eine solche Modellierungsweise ist sehr populär. Da neben dem Tripel Subjekt-Prädikat-Objekt (SPO) mit dem Repräsentanten für die Aussage selbst nun ein vierter Baustein für jede elementare Aussage existiert, werden Systeme, die sich auf die effiziente Speicherung (und bedingt auch Abfrage) von Daten in einem solchen Datenmodell spezialisiert haben, auch Quadruple Stores genannt. Das Property-Graph-Modell kann diesen Ansatz sehr effizient und elegant unterstützen. Das Vokabular wird hierbei anstatt im Graphenschema (mit dem Kantentyp ›isst‹) nun als Teil der Daten vorgehalten (mit einem eigenen Knoten als Repräsentant von ›Essen‹).

Nicht alle Aussagen sind sinnvoll mit einem einfachen SPO-Muster erfassbar. Wird etwa der Satz ›Peter kauft im Supermarkt eine Banane.‹ betrachtet, so wird leicht begreiflich, dass ein Verknüpfen von Peter mit der Banane mittels ›kauft_im_Supermarkt‹-Kante nicht sehr weitsichtig ist. Das Vorgehen impliziert, dass künftig auch ›kauft_auf_dem_Wochenmarkt‹- oder ›kauft_im_Internet‹-Kanten benötigt werden, welche nach dem Property-Graph-Modell noch nicht einmal komfortabel in eine Hierarchie von Kantenlabels einsortiert werden können.

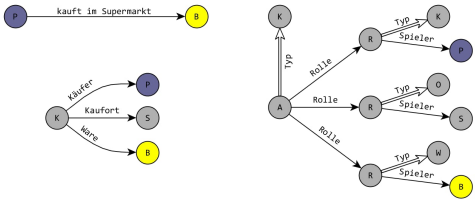


Abb. 3: Modellierung von mehrgliedrigen Assoziationen. [Efer 2019.]

Abbildung 3 zeigt zusätzlich zur so gebildeten Repräsentationsform zwei weitere Ansätze: Zum einen ist es möglich, für die einzelnen Aussage-Teile Knoten zu erzeugen (Peter, der Supermarkt, die Banane) und sie einem »abstrakten« Knoten, der für die »Kaufhandlung« (K) steht. zuzuordnen. Hierbei können nun beliebig viele Aussagegegenstände in Relation gesetzt werden (nicht mehr nur zwei). Dieses Vorgehen ähnelt dabei der bekannten Reifizierung, mit dem Unterschied, dass nun spezielle Kantentypen, ›Käufer‹, ›Kaufort‹ und ›Ware‹, benötigt werden, um die Repräsentanten qualifiziert mit der Kaufhandlung zu verbinden.

Auch solche unterstützenden Kantentypen werden mit der Aufnahme weiterer Aussagen in die Datenbank perspektivisch in immer neuen Ausprägungsformen vorkommen. Sie werden dringend benötigt, um die Rolle der an der Aussage beteiligten Knoten zu definieren. Ohne eine solche Unterscheidung könnte die modellierte Aussage ebenso gut als ›Ein Supermarkt kauft in der Banane Peter.‹ interpretiert werden. Sollen diese Rollendefinitionen nicht Teil des Graphenschemas werden, so kann als weitere Abstraktion das Assoziationsmodell der semantischen Technologie Topic Maps (ISO 13250) verwendet werden. Abbildung 3 zeigt auf der rechten Seite die dafür angelegte Struktur. In dieser existiert ein generischer abstrakter Knoten (A) für die Assoziation (also die Gesamtaussage) sowie einzelne generische abstrakte Knoten für alle Rollen (R). Diesen abstrakten Knoten ist ein Typ zugeordnet, hier

›Kaufhandlung‹ als Assoziationstyp sowie ›Käufer‹, ›Kaufort‹ und ›Ware‹ als Rollentypen. Die Rollen-Knoten sind darüber hinaus mit ihren Rollenspielern , also den konkreten Aussagegegenständen verbunden. So ist das komplette Domänenvokabular innerhalb der Instanzdaten abgebildet. Die Zahl der Kantentypen bleibt auch bei Erweiterung um neue Aussagentypen konstant. Dieses Prinzip der abstrakten Stellvertreterknoten wird als Indirektion im nächsten Unterabschnitt noch näher vorgestellt.

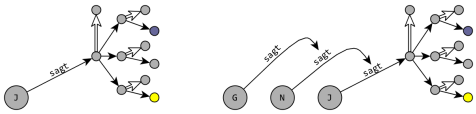


Abb. 4: Problematik rekursiver Reifizierung mittels Assoziationsmodellen. [Efer 2019.]

Abbildung 4 zeigt, dass auch referenzierte Aussagen (im Stil von ›Jürgen sagt, dass…‹) mit dem Assoziationsmodell problemlos aufgegriffen werden können. Allerdings ist es nach wie vor nicht möglich, eine solche »indirekte Rede» selbst zu referenzieren. Beginnt ein Satz also mit ›Laut Günther hat Nora erzählt, Jürgen würde behaupten, dass Peter…‹ so würde für jede Ebene der Meta-Aussage wieder eine eigene Assoziation gebildet werden müssen, wodurch das eigentliche Domänenwissen (über Kauf- und Ernährungsgewohnheiten von Menschen) in der Datenbank sehr stark von technischen Modellierungskonstrukten überschattet würde.

Der nächste Unterabschnitt baut nun auf den eingangs besprochenen Überlegungen zur Notwendigkeit der Modellierung von Faktenprovenienz und den bis hierhin vorgestellten Modellierungsansätzen auf, um ein einfaches System zu entwickeln, in dem sich lange Folgen von Überlieferung einzelner Fakten effizient abbilden und für die Forschung nutzbar machen lassen.

4. Provenienzketten und Indirektion

Für die weitere Veranschaulichung soll von hier an (ebenfalls analog zum Vortrag) ein neues, eventuell forschungsnäheres Beispiel eingeführt werden. Es soll der übliche Fall der Erfassung historischer Aussagen aus digitalen Versionen älterer Quellen betrachtet werden (ohne hierbei allzu realistische Fach- und Forschungsfragen zu beachten).

Bei der Durchsicht einer ins Englische übersetzten Internetversion von Herodots Historien könnte beispielsweise die folgende Textstelle den Autor dieses Artikels zur Aufnahme eines neuen Fakts in eine Forschungsdatenbank zur Beziehung von Völkern und Stämmen im antiken Mittelmeerraum animieren:

> »[…] they further had a huge vase made in bronze, […] which they sent to Croesus as a return for his presents to them. The vase, however, never reached Sardis. […] The Lacedaemonian

> story is that when it reached Samos, on its way towards Sardis, the Samians having knowledge of it, put to sea in their ships of war and made it their prize.«[8]

Es könnten sehr einfach Knoten für eine Vase, die Spartaner (Lakedaimonier), die Samier und Krösus (bzw. allgemein die Lydier in Sardis) erstellt werden. Dazu gesellen sich eine Schenkungs-Assoziation und eine Raub-Assoziation. Für letztere ist es besonders interessant, die Faktenprovenienz zu kennen. Denn es ist bei weitem nicht sicher, dass der so dokumentierte Fakt der historischen Wahrheit entspricht – und latenter Zweifel daran ist bereits in der Textquelle enthalten.

Für eine qualifizierte Angabe der Faktenprovenienz sollten an dieser Stelle alle bekannten Quellen und Überlieferungsschritte identifiziert, abgegrenzt und geordnet erfasst werden, von den ältesten Belegen bis hin zur letzten Instanz vor der Eingabe in die Datenbank.

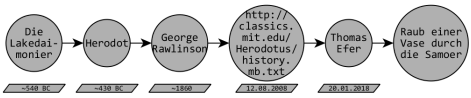


Abb. 5: Beispielhafte Provenienzkette. [Efer 2019.]

Abbildung 5 zeigt, wie sich daraus eine Kette von Knoten in der Datenbank ergibt, welche über einen geeigneten Kantentyp miteinander verbunden sind, bis zur Aussage (bzw. dem dafürstehenden Assoziationsknoten). Die einzelnen Zeitpunkte der Überlieferung (falls bekannt), können direkt an dem Überlieferungsknoten notiert werden. Anders als bei der indirekten Wiedergabe (›Laut Günther hat Nora erzählt, …‹) steht hier das von der Originalquelle am weitesten entfernte Element am nächsten an der Aussage. Denn nur über dieses Element »erfährt« das System von der Existenz der anderen Kettenglieder. Das entspricht der typischen Angabe von Provenienzen bei Objekten, wo auch die »letzten« Besitzer*innen und Aufenthaltsorte zuerst genannt werden.

Diese Überlieferungskette gibt dem eigentlichen Fakt wertvollen Kontext: Er sollte nur dann in der Forschung direkte Beachtung finden, wenn davon auszugehen ist, dass er sich nicht aufgrund von Fehlinterpretation des Eingebenden in die Datenbank, inkonsistenter Textwiedergabe in der Internetquelle, falscher Übersetzung durch George Rawlinson, Fiktion des griechischen Historikers oder falscher Bezichtigung durch die Spartaner entstanden ist. Die Provenienzkette macht diese logischen und quellenkritisch relevanten Abhängigkeiten für die Forschung erstmals explizit und ihre (zum Teil auch mangelhafte) Berücksichtigung im Forschungsprozess für externe Betrachter*innen endlich transparent nachvollziehbar.

Aus Modellierungssicht sollte das bisher sehr einfache Konstrukt der Provenienzkette noch etwas verfeinert werden. Zur Demonstration der Notwendigkeit kann eine zufällige weitere Textstelle aus derselben Onlinequelle herangezogen werden:

[8] Herodotus / Rawlinson 1994-2009.

> »[…] Archias, a man named Archias like his grandsire […] told me that his father was called Samius, because his grandfather Archias died in Samos so gloriously, [that he] was buried with public honours by the Samian people.«[9]

Hieraus kann neben prosopographischen Abhängigkeiten auch leicht der Fakt gewonnen werden, dass Archias (der Großvater) ein Ehrenbegräbnis erhalten hat. Dieser kann mit den üblichen Mitteln als Assoziation in der Datenbank erfasst werden. Die Faktenprovenienz könnte nun in einer eigenen, von der obigen unabhängigen Kette erfasst werden. Doch dann gäbe es beispielsweise für George Rawlinson zwei Repräsentanten in der Datenbank. Dies ist in Wissensbasen generell sehr unerwünscht, kann Inkonsistenzen und Mehrarbeit bei der Pflege der Daten hervorrufen und erschwert eine gemeinsame Betrachtung der Überlieferungslage aller Fakten. Stattdessen sollten bedeutungstragende Knoten, welche für Entitäten stehen, stets nachgenutzt werden.

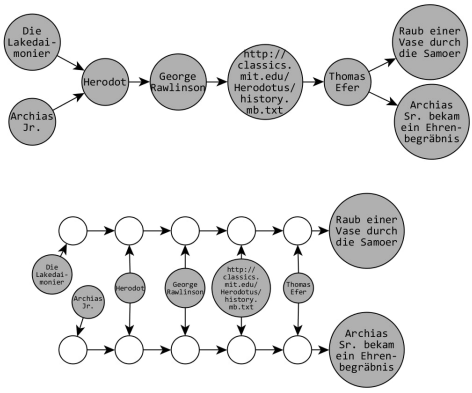


Abb. 6: Trennung von überlappenden Provenienzketten mittels Indirektion. [Efer 2019.]

Abbildung 6 zeigt oben zunächst, wie eine kombinierte Provenienzkette unter Knotennachnutzung aussehen könnte. Es zeigt sich, dass beide Aussagen auf den gleichen Eintragenden, die gleiche Onlinequelle, etc. zurückgehen. Was zunächst wie eine elegante Lösung erscheint, birgt nun allerdings bei der Abfrage das Problem, dass die Zuordenbarkeit der Ursprungsquellen Herodots zu beiden Aussagen nicht mehr gegeben ist. Über eine Graphenabfrage kann nicht mehr rekonstruiert werden, ob die Spartaner oder Archias (der Enkel) vom Raub der Vase berichtet haben.

Deshalb ist es notwendig, für die gemeinschaftliche Abbildung mehrerer Provenienzketten, welche sich in den beteiligten Entitäten überlappen, eine Stufe der Indirektion in der Modellierung zu verwenden. Es werden abstrakte, »leere« Knoten als Platzhalter für die Position in der Kette erzeugt. Diese sind dann mit den eigentlichen Entitäten verknüpft

[9] Herodotus / Rawlinson 1994-2009.

– grob vergleichbar mit den oben vorgestellten Rollen-Knoten und Rollenspielern im Assoziationsmodell von Topic Maps.[10] Abbildung 6 zeigt, wie mit diesem Ansatz beide Ketten sauber getrennt und dennoch auf der Ebene der Entitäten unifiziert sind.

Provenienzketten sind ein erster Vorschlag zur Dokumentation der Faktenherkunft. Eventuell lassen sich für bestimmte Anwendungsgebiete andere, geeignetere Repräsentationsformen finden. Worin genau besteht der Vorteil von Ketten? Was sagt die Reihenfolge aus und lässt sie sich immer genau definieren? Diese Fragen weisen auf viele praktische Probleme hin, die die Erfassung und Verwaltung der Provenienzinformation mit sich bringen kann. Alternativ zur Kette ließen sich einzelne Überlieferungsstationen auch als Menge wiedergeben. Mit angegebenen und gegebenenfalls approximierten Zeitinformationen ließe sich eine (oder mehrerer hypothetische) Ketten rekonstruieren, unter Berücksichtigung von Unsicherheiten bei der Datierung. Allerdings erschwert ein solcher Ansatz die Abfrage im Graphenmodell. Ein weiterer damit verknüpfter und noch ungeklärter Punkt betrifft die Fragen: Woher stammen die Provenienzangaben und muss eine Provenienz der Provenienz abgebildet werden (oder ist diese identisch oder zumindest sehr stark verschränkt mit der Faktenprovenienz)?

Im Lichte dieser Unklarheiten und da die vorgeschlagene Umsetzung bisher nur auf einfachster technischer Ebene mit den Basiskonstrukten des Property-Graph-Modells operiert, welche sich vielseitig verstehen und verwenden lassen, ist der Wunsch nach einem gemeinsam zu nutzenden Rahmenwerk für eine formale semantische Interpretation von Provenienzketten (oder -mengen) mehr als nachvollziehbar. Bisherige Arbeiten in verwandten Bereichen sind etwa die PROV-Ontologie des W3C. Diese umfasst ein umfangreiches Vokabular und enthält viele gute Überlegungen auf der Abstraktionsebene von Entity, Activity und Agent. Zudem wartet es mit einem eigenen (vergleichsweise komplizierten) RDF-basierten Datenmodell auf, wie im Standard von Moreau und Missier beschrieben.[11] Damit mutet es allerdings ähnlich schwerfällig an, wie allgemeine semantische Basis-Referenzmodelle, etwa das CIDOC-CRM, wie von Ore, Doerr und anderen definiert.[12] Letzteres ließe sich konzeptionell sehr gut als Grundlage für die Definition von Events im Sinne von Überlieferungsereignissen nutzen. Insbesondere für die semantisch sinnvolle Charakterisierung von Aktionen als Vorgängen der Überlieferung oder aktiven Bewahrung von Fakten besteht noch weiterer Forschungsbedarf, der sich auch nur im engen Dialog mit den Fachdisziplinen der Geistes- und Sozialwissenschaften auflösen lässt. Es stellt sich nun die Frage, ob Provenienzketten bereits in ihrer bisherigen Form nutzbringend für die Forschung sein können.

In der arbeitsteiligen Forschung wurde die Frage nach der Herkunft von Faktenwissen bisher oft implizit durch aufwendige Editionsprozesse von der oder dem Forschenden ferngehalten. Glaubwürdigkeit und Plausibilität von Fakten wurden unter Einhaltung guter wissenschaftlicher Praxis vorab (nach Ermessen der Editor*innen und geknüpft an ihre Reputation) geprüft. Die Herkunft der enthaltenen (und ausgesparten) Fakten sind für die Editor*innen nachverfolgbar und werden meist allenfalls ausschnittsweise in Form eines kritischen Apparats oder durch

[10] ISO/IEC 13250 International Organization for Standardization 2003, Stage 90.92.
[11] Moreau / Missier 2013.
[12] Ore et al. 2018.

Begleittexte kommuniziert. Digital erfasste Provenienzketten ersetzen diese Praxis nicht. Sie erlauben jedoch die Introspektion und falls nötig auch ein qualifiziertes Abweichen von der durch sie gefestigten »Lehrmeinung«. Fakten, die erfasst werden und deren Herkunft dokumentiert wird, sind dabei bisher explizit noch nicht an Wahrheitsgehalte geknüpft. Denn oftmals ist es in der Forschung auch von großem Interesse, woher »falsche« Informationen stammen, ebenso wem sie wann vorlagen und wessen Urteile sie eventuell beeinflussten. Diese wissenschaftsgeschichtlich spannenden Fragen setzen sich bis in heutige Theoriegebilde zahlreicher Disziplinen fort. Ihre Offenlegung und damit auch die kritische Hinterfragung des aktuellen Forschungsstandes ist eine wesentliche Aufgabe moderner (digital unterstützter) Forschungstätigkeit.

Bevor die Nutzung von Provenienzketten für den Umgang mit Zweifel vorgestellt wird, soll an dieser Stelle noch ein kurzer Exkurs zu den möglichen Kettengliedern der Überlieferung, Modi der Übernahme und Extraktion von Fakten sowie ihre digitale Repräsentation im Graphen eingeschoben werden.

5. An Faktenprovenienz beteiligte Entitäten

Entitäten sind konkrete oder abstrakte, belebte oder unbelebte »Dinge«, die in der Regel benannt werden können. Besitzen sie keinen Namen, so können sie doch insoweit durch Nummerierung oder über ihre Relation zu anderen Entitäten charakterisiert werden, dass man über sie (und nur sie) mit anderen kommunizieren kann. Entitäten benötigen in einer digitalen Datenbank eindeutige Identifikationsmerkmale, so genannte Identifier, damit auf sie verwiesen werden kann. Projektintern kann dies über einfache Datenbank-IDs geschehen, beim Datenaustausch über Projekte hinweg bieten sich global eindeutige Identifier an, wie sie beispielsweise im Semantic Web über Webadressen (hinter denen üblicherweise Informationsressourcen hinterlegt sind) realisiert wird. Über die Prinzipien von Linked Open Data (LOD) können Anbieter und Konsumenten offener Datensammlungen verteilt dieselben Entitäten referenzieren. Über diese technologischen und organisatorischen Möglichkeiten ist zudem ein Übertrag zwischen verschiedenen Identifier-Systemen möglich. Zahlreiche Gremien und Autoritäten der kulturellen und staatlichen Domäne werden für Entitäten von übergeordneter Bedeutung Normdatensätze erzeugt. Heutzutage werden diese durchweg in maschinenlesbarer Form angeboten.

Da Benennungen an sich keine gute Identifier sind (gleiches kann unterschiedlich benannt werden und unterschiedliches gleich) muss für eine saubere Nutzung von Normdaten und für die aussagefähige Modellierung der Domänendaten eine Disambiguierung und eine Zusammenführung von Entitäten und den mit ihnen verbundenen Datensätzen stattfinden. Für die Faktenprovenienz sind verschiedene Typen von Entitäten relevant. Diese können (und werden es in vielen Fällen auch) Teil der erfassten Domänendaten sein.

Personen sind im historischen Kontext nicht selten schwer zu greifen und zuweilen nur schwer eindeutig zu identifizieren. Die Informationen, die sich über sie aus den Quellen gewinnen lassen, sind nicht immer ausreichend, um eine Verknüpfung zu Normdaten zu ermöglichen.

Oft ist die Forschung auch so spezifisch, dass noch gar keine ausreichenden Normdaten für die besondere Domäne im genau passenden geo-temporalen Kontext existieren. Dazu kommt, dass zuweilen fiktionale, mythische Charaktere attribuiert werden oder idealisierte Variationen von realen historischen Personen beschrieben werden. Das Herausarbeiten solcher Identität(en) ist (auch ohne explizite Modellierung von Faktenprovenienz) eine wichtige Forschungstätigkeit und ein Schlüssel zum Verständnis des Materials.

Falls eine Überlieferung oder originale Bekundung eines Fakts in Form von Dokumenten stattgefunden hat, gilt es, eine geeignete Granularitätsstufe für die Referenzierung zu finden. Zwischen dem abstrakten Werk eines Autors oder einer Autorin und einem konkreten physischen Buch in einer spezifischen Auflage können mehrere konzeptionelle Ebenen liegen, wie sie etwa in den Functional Requirements for Bibliographic Records (FRBR) unterschieden werden.[13] Auch Paratexte auf Textträgern können als Quellen dienen und sollten damit als eigene referenzierbare Entitäten vorliegen (welche im Graphen freilich mit ihren übergeordneten Einheiten verbunden werden können). Für unumstrittene, bekannte und edierte Werke bieten sich eher Referenzierungsschemata für kanonische Texte an, für Handschriften kann über die Verwendung von Identifiern für den physischen Träger des Textes nachgedacht werden, da dieser direkt mit ihm verbunden ist und auch im Forschungskontext oft als Einheit verstanden wird.

Bei Dokumenten aus dem Web gilt zu beachten, dass diese selbstverständlich über ihre Webadresse eindeutig identifiziert werden können. Jedoch: Im Semantic Web werden alle Arten von Ressourcen durch Webadressen referenziert! Wird eine Person beispielsweise über die URL ihrer Instituts-Webseite identifiziert, so ist bei einer Aussage ›laut der Ressource mit dieser URL‹ nicht mehr zu erkennen, ob sie ›laut der Person‹ oder ›laut der Webseite‹ getroffen wurde. Im Referenzmodell von Topic Maps wird daher zwischen Identifier und Locator einer Ressource unterschieden. Dies ist eventuell auch für die Entitäten in Provenienzketten sinnvoll.

Neben den prinzipiell klar umreißbaren Entitätentypen ›Person‹ und ›Dokument‹ existieren auch diffusere Glieder in Überlieferungsketten. Beispiele können Gruppen (wie ›die Illuminaten‹), Ethnien, Sammelidentitäten (etwa ›der Senat‹) oder Institutionen sein. Diese können zu unterschiedlichen Zeitpunkten (was abgeschwächt auch für andere Entitätentypen gilt) ganz unterschiedlichen Charakter besitzen und sich in ihrer Zusammensetzung, Ausrichtung, Wirkmächtigkeit und nicht zuletzt Glaubwürdigkeit (allgemein und in Bezug auf spezielle Themenbereiche) sehr unterscheiden. Auch gilt es hier, bei der Referenzierung ein angemessenes Granularitätsniveau zu wählen: Ist nun die Pressesprecherin persönlich, die Presseabteilung oder die Institution selbst in die Kette zu übernehmen? Warum nicht alle drei in dieser Reihenfolge? Eng damit verbunden ist die Frage einer Rollenidentität (also eine bestimmte Person ›als Vater einer anderen Person‹ oder ›als König‹ oder ›als Zeuge in einem Prozess‹). Möglicherweise sollten für solche Ausdifferenzierungen einer Entität stets spezielle Repräsentanten erstellt und mit der Hauptentität verknüpft werden.

[13] International Federation of Library Associations and Institutions 2009.

Schließlich kann es auch sinnvoll sein, Werkzeuge zur Faktenextraktion, etwa physikalische Messungen zur Datierung oder Programme bzw. von ihnen verwendete Algorithmen an passender Stelle in die Provenienzkette aufzunehmen. Verfahren, die Fakten generieren, sind als initiales Element einzusetzen, während Verfahren, die Informationen verarbeiten (und nicht nur einer unveränderten Weitergabe dienen), entsprechend später in die Kette eingefügt werden sollten. Jede Form der automatischen oder manuellen Veränderung, Edierung, Übersetzung oder Interpretation sollte nachvollziehbar sein.

Wenn durch Identifizierung, Disambiguierung und Referenzierung eine vollständige Kette aus einzelnen beteiligten Entitäten gebildet ist, kann diese im nächsten Schritt für Umgang mit unklarer Faktenlage verwendet werden.

6. Unsicherheit, Zweifel und Widersprüche

Bis hierhin wurde ein System entwickelt, mit dem sich die Überlieferungsgegebenheiten für Fakten auf der Basis von beteiligten Entitäten abbilden lassen. Ob die dabei beschriebenen Aussagen (über Aussagen, über Aussagen… usw.) tatsächlich getätigt wurden, oder nicht, lässt sich im Nachhinein nicht ermitteln. Die Open-World-Assumption suggeriert zudem, dass weitere unterstützende oder auch gegenteilige Aussagen existieren können.

Die Faktenlage muss daher auf ihre Plausibilität hin und ihre Konsistenz untersucht werden. Für eine höhere Sicherheit über den Wahrheitsgehalt sorgt das Sammeln von Belegen. Zur guten Praxis der Wissenschaft gehört es jedoch genauso, als wahr angenommene Fakten (insbesondere im historischen Kontext) immer wieder anzuzweifeln. Nichts sollte von Forscher*innen diskurslos akzeptiert werden. Viele der Aktivitäten in den Geisteswissenschaften zielen auf genau diese kritische (Re-)Kontextualisierung bekannter Fakten ab. Statt absoluter Sicherheit muss mit unterschiedlichen Graden von Unsicherheit umgegangen werden. Wie kann diese Unsicherheit abgebildet werden?

Unsicherheit auf Eigenschaftsebene, die sich aus Vagheit von Wertezuweisungen oder einer bloßen Angabe von Schranken für Werte (anstatt der Werte selbst) ergibt, soll hier im Weiteren nicht behandelt werden. Quellen für diese Unsicherheiten können divers sein, wie etwa Toleranzen und die Unschärfe von Messungen, Subjektivität, Erinnerungslücken oder mögliche Divergenzen, die sich aus Übersetzung und Datenüberführung zwischen konzeptionell unterschiedlichen Maßeinheiten oder diskretisierten Einteilungen einer Größe ergeben.

Unsicherheit über die Tatsächlichkeit von Begebenheiten, also der generelle Zweifel am Wahrheitsgehalt eines Fakts in der Wissensbasis ist dagegen stets gleichbedeutend mit dem Zweifel an Zeugen, Quellen oder einzelnen Überlieferungsschritten.

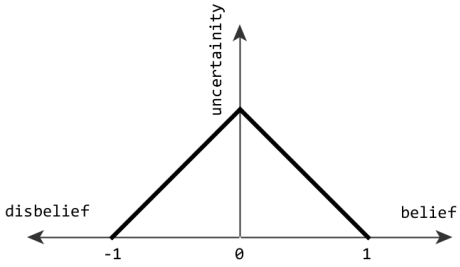


Abb. 7: Unsicherheit im Spannungsfeld zwischen ›belief‹ und ›disbelief‹, nach Hartig 2009. [Efer 2019.]

Abbildung 7 zeigt eine Übersicht, in der sich Unsicherheit als Zweifel an konkreten Aussagen interpretieren lässt. Der Bereich zwischen ›belief‹ und ›disbelief‹ gegenüber einem Fakt (hervorgerufen z. B. durch mehrere widersprüchliche Aussagen oder unbeweisbarem Misstrauen gegenüber einzelnen Quellen) weist die höchste uncertainty auf.

Die Spezifikation von Unsicherheit innerhalb der Provenienzkette berührt wieder das Themenfeld der »Provenienz der Provenienz«. Sie kann im Modell prinzipiell auf verschiedene Arten und Weisen ausgedrückt werden, kann dabei oft nur den Charakter eines »Kommentars« für menschliche Bearbeiter*innen haben, denn als maschinell zu behandelnde Kategorie gelten. Davon ausgenommen sind zwei gut abzubildende Fälle: Zum einen können offensichtliche Fehler und Ungenauigkeiten bei der Übernahme von Fakten direkt an die passende Stelle notiert werden, indem eine entsprechende Property an die korrespondierende Kante innerhalb der Kette hinzugefügt wird. Zweitens kann die generelle Unglaubwürdigkeit einzelner Entitäten ebenfalls direkt an diesen notiert werden. Dies bietet sich etwa an bei gefälschten Dokumente, identifizierten Hochstapler*innen., aber auch integren Forscher*innen. die jedoch ihre Fakten erzeugende Interpretation von Quellen auf Basis eines mittlerweile überholten Forschungsstandes vorgenommen haben.

7. Mögliche Auflösungsmechanismen

Die Einschätzung einer Entität als (möglicherweise) unglaubwürdig hat nun im Modell den Effekt, dass das betroffene Kettenglied bei der Auflistung der Faktenprovenienz nicht einfach ausgelassen oder übersprungen werden darf, da alle vorherigen Kettenglieder nur durch die Annahme seiner Integrität im Graphen »erreichbar« sind.

Angenommen, der Autor dieses Artikels würde behaupten, dass bereits Herodot sagte: ›Peter isst eine Banane‹. In diesem abstrusen Fall würde es sich lohnen, auch alle weiteren Fakten erneut zu überprüfen, in deren Provenienzkette er prominent vorkommt.

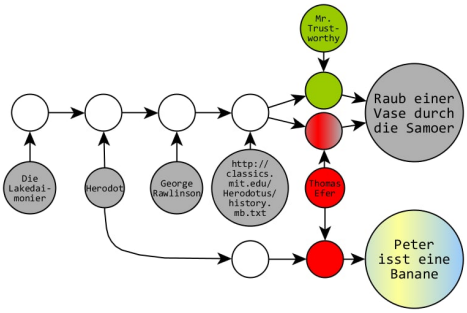


Abb. 8: Invalidierung unglaubwürdiger Kettenglieder und Ausbesserung durch Nachrecherche. [Efer 2019.]

Abbildung 8 zeigt in Rot die unglaubwürdige Überlieferungsbekundung und die dadurch unglaubwürdig gemachte Entität. Wenn dieser Entität nun nicht mehr getraut wird, fallen auch glaubhafte Belege für den Fakt des Raubes der Vase durch die Samoer weg. Ein Mitglied der fachwissenschaftlichen Community müsste nun selbst eine Recherche durchführen (z. B. durch Aufruf der Online- Ressource oder Konsultation der Printausgabe oder einer unübersetzten griechischen Edition) und durch diese Rechercheleistung die Fehlstelle »auffüllen«. Diese vertrauenswürdige Person kann nun als Entität (mit eigenem Stellvertreterknoten) in die Provenienzkette eingefügt werden. Aus der Kette wird gewissermaßen ein Strang (wie bei aller Form der mehrfachen »Absicherung« der Provenienz, z. B. durch mehrere Überlieferungen in Manuskriptform oder Ähnliches).

Nicht immer lassen sich Fakten als falsch widerlegen oder mit überwältigender Evidenz unterfüttern. Die Existenz unauflösbarer Widersprüche kann durchaus plausibel sein und ist grundsätzlich für die Nutzung einer Wissensbasis im Rechercheprozess nicht schädlich. Die explizite Dokumentation solch strittiger und unklarer Punkte erzeugt Transparenz und erleichtert zudem die Ursachenforschung und im Zuge dessen ggf. sogar das Aufspüren neuen Quellenmaterials im Umfeld der bisher konsultierten Überlieferungszeugen.

Die Notation der Unglaubwürdigkeit einzelner Quellen und Interpretationen kann in einem solchen Modell unabhängig von der Inklusion einzelner Fakten in die Wissensbasis erfolgen. Die kritische Beschäftigung mit Faktenprovenienz kann somit in einer zielgerichteten Stand-Off-Annotation der einzelnen Aussagen und ihrer Herkunft münden. Permanente und verlustbehaftete Filterung, Löschung, Korrektur und Edition an Ort und Stelle sind so nicht mehr nötig und versperren künftigen Forscher*innen nicht mehr die Sicht auf die (komplexe) Quellenlage. Gleichzeitig kann die »Übernahme« fremder Einschätzungen zur Fakten-Glaubwürdigkeit effizient und transparent innerhalb eines gemeinsamen Referenzsystems erfolgen.

Zum wichtigen Werkzeug zur Unterstützung von Recherchearbeit könnte eine (semi-)automatische Identifikation »schwacher« Provenienzketten werden. Dafür könnten beispielsweise die Zeiträume zwischen den einzelnen Kettengliedern beleuchtet werden: Mehrere hundert Jahre Abstand sind unplausibel für eine direkte Überlieferung zwischen

einzelnen Personen. Hier könnte eine genauere Recherche weitere Zwischenstationen zu Tage fördern. Weiterhin könnten (für Zeiten ohne briefliche oder fernmündliche Kommunikationsmöglichkeiten) große räumliche Distanzen zwischen den Wirkungsorten von überliefernden Personen als Schwachstelle identifiziert werden. Diese machen eine direkte Faktenüberlieferung unwahrscheinlicher, außer es existieren Belege für ein persönliches Treffen. Lange Zeiträume sind bei physischen Trägern der Fakten zunächst nicht unüblich, wie z. B. bei handgeschriebenen Autographen und lange aufbewahrten Archivalien. Wenn allerdings Kopien, Teil-Abschriften und Kompilationen beteiligt sind, gilt es dort, mit den üblichen Herangehensweise der Quellenkritik (entsprechend des genauen Entitätentyps) vorzugehen.

Eine solch differenzierte Auswertung benötigt letztlich einen geeigneten ontologischen Zugang. Dabei ist noch unklar, ob dies generisch, wie bei den bereits genannten Vokabularen, geschehen kann, oder auf die spezifische Fachkultur und Arbeitsweise einzelner Fachgebiete sowie deren jeweilige Quellenlage zugeschnitten sein muss, insbesondere bezüglich Art, Zustand, Umfang, Alter und Heterogenität der Quellensammlungen. Spezielle Anforderungen ergeben sich dort aus den Berührungspunkten der Forschung mit dem Sektor des kulturellen Erbes. Für die Provenienzdokumentation von Fakten sind auch gegenständliche Provenienzen von größtem Interesse und es ist leicht ersichtlich, dass hierbei eine Fülle von kuratorischen Gepflogenheiten, materiellen Besonderheiten und organisatorischen Standards zu berücksichtigen ist: Transkripte von Wachszylindern, Fachartikel zur Grabungsdokumentation archäologischer Quellen, digitale Korpussammlungen und Forschungsdatenbanken, Sammlungen zur Oral History – all diese Überlieferungswege müssen adäquat abbildbar sein.

8. Kritik, Desiderate und Ausblick

Die hier vorgestellten Ansätze können (so die Hoffnung des Autors) zu einem gewissen Grad helfen, Forschungsergebnisse und ihre Herleitung übersichtlicher und für andere anschlussfähiger zu erfassen. Dem gewählten Konzept wohnt (wie jedem Modell) eine bewusste Simplifizierung inne. Damit greift es für spezielle Anwendungsfälle sicher in vielerlei Hinsicht noch kurz, zumal viele Aspekte insbesondere der Abbildung von Identität und normierten Ankerpunkten für Entitäten bislang sehr unterspezifiziert sind. Das System wie hier beschrieben ist in der Praxis bisher nicht systematisch erprobt worden., zumindest eine Nachjustierung einzelner Teile ist im Zuge dessen zu erwarten.

Wie im Abstract erwähnt, liegt dem Beitrag der Wunsch zugrunde, eine Diskussion über die Inklusion von Provenienzinformationen in Forschungsdatenbanken und Recherchessysteme zu initiieren. Auch wenn diese Idee auf fachwissenschaftlichen Zuspruch stoßen sollte, bleibt zu untersuchen, wie bestehende Forschungsmethodik und eine explizite, digitale und transparente Modellierung von Faktenprovenienz zusammenpassen und welche Implikationen sich für Forschungsprozesse ergeben. Insbesondere die Identifikation von Grenz- und Sonderfällen des Modells kann nur sinnvoll aus praxisnahen Erwägungen und durch realitätsnahe Testfälle geschehen.

Jede im Forschungsalltag zusätzlich zu erfassende Information ist mit erhöhten Arbeitsaufwänden verbunden. Hinzu kommen weitere Aufwände bei der kontinuierlichen Pflege und Aktualisierung der Daten. Solange sich im Arbeitsalltag kein dauerhafter, sichtbarer Nutzen aus der Erfassung von Provenienzinformationen gesammelter Fakten ergibt, ist der Wunsch verständlich, die dafür aufzuwendende Arbeitszeit lieber für die »eigentliche« Forschung zu verwenden. Hier können sich ähnliche Reibungspunkte ergeben, wie sie für die Dokumentation von Programmquelltext existieren. Welche Dokumentationsaufwände sind mindestens nötig? Welche angemessen? Welchen Grad der öffentlichen oder internen Nachnutzung der Fakten und ihrer Provenienzinformationen wird es voraussichtlich geben? Insbesondere bei öffentlicher Verfügbarmachung ist zudem der subjektive Eindruck der eintragenden Nutzer*innen nicht zu unterschätzen, eventuell wertvolle Recherchearbeit »für Andere« zu leisten, ohne dabei direkte akademische Reputationsvorteile zu erhalten. Als Gegenargument könnte angemerkt werden, dass die Provenienzketten eindeutig den jeweiligen Bearbeiter, die jeweilige Bearbeiterin als letztes Glied enthalten können und damit die kleinen individuellen Beiträge zum kollektiven Wissen sogar besser als bisher herausgestellt werden können. Solange dies jedoch keinen anerkannten »Wert« in der Fachwelt darstellt ist, stellt dies wohl nur einen schwachen Trost dar.

Aus der eher kollektiven Sicht heraus lässt sich feststellen, dass sich Provenienzketten für die Qualitätskontrolle einer Wissensbasis nutzen lassen und darüber hinaus einen transparenteren Forschungsprozess befördern können. Neben den vielen positiven Aspekten für die informiertere und fundiertere Ableitung von Aussagen bestehen dabei auch Gefahren in der Praxis: Durch die Nutzungsmuster der Fakten wird ggf. sichtbar, wer welche (noch lebenden) Forscherkolleg*innen als »unglaubwürdig« einstuft. Dies birgt großes Konfliktpotential, zumindest bis durch die weite Adaption solcher Forschungsmethodik eine Versachlichung der Debatten und eine kollaborative Konfliktlösung eintritt, falls dies realistisch ist.

Ein weiterer Kritikpunkt an den vorgestellten Lösungsansätzen für die Modellierung von Faktenprovenienz kann die Vermischung von primären Domänendaten und (sekundären) Provenienzdaten in der Wissensbasis sein. Eine solche komplexitätserhöhende Anreicherung von diversen Informationen ist nicht immer gewünscht und nicht immer praktisch für die Erstellung von Präsentations- und Fachanwendungen. Eine logische Trennung aller erfassten Daten durch speziell zugewiesene Properties, einer knotenweisen Verknüpfung zu Data Collections oder die Nutzung logischer Subgraphen (falls vom Datenbanksystem unterstützt) löst dieses Problem, bedeutet allerdings zusätzlicher Aufwand bei Konzeption und Umsetzung des Systems, sowie bei der Formulierung und Abarbeitung von Anfragen. Hier existiert definitiv ein Bedarf für weitere Überlegungen und praktische Untersuchungen.

Als weiteres Desiderat kann die Konzeption (standardisierter) Interaktionsweisen und Nutzeroberflächen für die Visualisierung und schnelle Introspektion, aber auch für die unterstützte Dateneingabe von Provenienzketten angesehen werden. Daneben ist zu untersuchen, wie sich die Workflows mit weiteren externen Softwareprodukten verknüpfen lassen. Für die Nachrecherche von Fakten wäre es beispielsweise wünschenswert, die entsprechenden zusätzlichen parallelen Kettenglieder von dem oder der Forschenden zur

konsultierten Quelle in der Provenienzkette automatisch hinzuzufügen, wenn die damit verknüpfte Rechercheaufgabe in einem Ticketsystem als erfolgreich abgearbeitet markiert wird.

Die hier vorgestellten Konzepte werden derzeit im Langzeitvorhaben Bibliotheca Arabica der Sächsischen Akademie der Wissenschaften zu Leipzig auf ihre Praxistauglichkeit hin untersucht und entsprechend weiterentwickelt. Das Projekt beschäftigt sich mit der reichhaltigen Produktion, Rezeption und Transmission von Literatur im bisher als post-klassisch bezeichneten Zeitalter der Manuskriptkultur in der arabischsprachigen Welt. In diesem Rahmen ist die Erstellung einer graphbasierten bio-bibliographischen Datenbank, in welche unter anderem die Daten von über 100 gedruckten Handschriftenkatalogen eingespeist werden sollen, vorgesehen. Durch sie wird eine digitale Arbeitsweise unterstützt, welche nicht nur neuartige Arbeitsweisen bei der Analyse der Manuskriptdaten ermöglicht, sondern insbesondere auch die Bereitstellung von Provenienzinformationen für Fakten zur Kontextualisierung der abgebildeten kodikologischen, bio- und bibliographischen und aller weiteren übernommenen Aussagen erfordert. Aus diesem Projekt werden also neue Impulse für die graphbasierte Modellierung von Faktenprovenienz hervorgehen.

Allgemein wird angestrebt, für die hier beschriebenen Ideen eine generische webbasierte Datenbanklösung als Mischung aus dokumentierter Referenzimplementierung und einfach nutzbarer Endanwendersoftware zur Verfügung zu stellen. Diese sollte generische Oberflächen mit entsprechenden individuellen Anpassungsmöglichkeiten, etwa im Stil der beliebten MyCoRe-Repositoriumssoftware bieten. Ein realistischer Zeitplan für die Umsetzung dieses Vorhabens ist vorerst jedoch noch nicht abzusehen.

Bibliographische Angaben

Luise Borek / Quinn Dombrowski / Jody Perkins / Christof Schöch: TaDiRAH: a Case Study in Pragmatic Classification. In: Digital Humanities Quarterly 10 (2016), H. 1. [online]

Thomas Efer: Graphdatenbanken für die textorientierten e-Humanities. Leipzig, 2017. URN: urn:nbn:de:bsz:15-qucosa-219122 [Nachweis im GBV]

Olaf Hartig: Querying Trust in RDF Data with tSPARQL. In: The Semantic Web: Research and Applications. Hg. von Lora Aroyo et al. (ESWC: 6, Heraklion, 31.05.-04.06.2009) Berlin u.a. 2009, S. 5-20. DOI: 10.1007/978-3-642-02121-3_5 [Nachweis im GBV]

Olaf Hartig: Reconciliation of RDF* and Property Graphs. In: arXiv.org. Technical Report vom 11.09.2014. [online]

Herodotus: The History of Herodotus. Translated by George Rawlinson. In: The Internet Classics Archive. 1994-2009. [online]

Functional Requirements for Bibliographic Records. Hg. von IFLA Study Group on the Functional Requirements for Bibliographic Records. In: ifla.org. Version von 02.2009. [online]

Information technology – SGML applications – Topic maps / ISO. Hg. Von International Organization for Standardization. ISO/IEC 13250 Stage 90.92. Second Edition vom 23.10.2003. [online]

Philip Moore / Hai Van Pham: On Context and the Open World Assumption. In: 2015 IEEE 29th International Conference on Advanced Information Networking and Applications Workshops. Hg. von Leonard Barolli. (IEEE AINA: 29, Gwangiu, 24.-27.03.2015) Piscataway, NJ. 2015. [Nachweis im GBV]

Luc Moreau / Paolo Missier: PROV-DM: The PROV Data Model. In: w3.org. W3C Recommendation vom 30.04.2013. [online]

Definition of the CIDOC Conceptual Reference Model. Hg. von Christian Emil Ore / Martin Doerr / Patrick LeBœuf / Stephen Stead. In: cidoc-crm.org. Version 6.2.3. von 05.2018. [online]

Ian Robinson / Jim Webber / Emil Eifrem: Graph Databases. Beijing u.a. 2013. [Nachweis im GBV]

Marko A. Rodriguez / Peter Neubauer: The Graph Traversal Pattern. In: Graph Data Management: Techniques and Applications. Hg. von Sherif Sakr / Eric Pardede. Hershey, PA 2012, S. 29–46. [Nachweis im GBV]

Graph Data Management: Techniques and Applications. Hg. von Sherif Sakr / Eric Pardede. Hershey, PA 2012, S. 29–46. Siehe auch [Nachweis im GBV]

Yogesh L. Simmhan / Beth Plale / Dennis Gannon: A survey of Data Provenance in e-Science. In: ACM SIGMOD Record 34 (2005), H. 3, S. 31–36. [Nachweis im GBV]

Abbildungslegenden und -nachweise

Abb. 1: Modellierung einfacher Aussagen im Graphen. [Efer 2019.]

Abb. 2: Modellierung von Aussagen über Aussagen mittels Reifizierung. [Efer 2019.]

Abb. 3: Modellierung von mehrgliedrigen Assoziationen. [Efer 2019.]

Abb. 4: Problematik rekursiver Reifizierung mittels Assoziationsmodellen. [Efer 2019.]

Abb. 5: Beispielhafte Provenienzkette. [Efer 2019.]

Abb. 6: Trennung von überlappenden Provenienzketten mittels Indirektion. [Efer 2019.]

Abb. 7: Unsicherheit im Spannungsfeld zwischen ›belief‹ und ›disbelief‹, nach Hartig 2009. [Efer 2019.]

Abb. 8: Invalidierung unglaubwürdiger Kettenglieder und Ausbesserung durch Nachrecherche. [Efer 2019.]