

## **Vom Thesaurus zum semantischen Netz: Potenziale von Data Mining in bibliographischen Datensätzen**

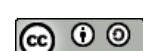
**Andreas Hartmann, S. Thänert**

### **Angaben zur Veröffentlichung / Publication details:**

Hartmann, Andreas, and S. Thänert. 2015. "Vom Thesaurus zum semantischen Netz: Potenziale von Data Mining in bibliographischen Datensätzen." *Digital Classics Online* 1 (1): 34–45. <https://doi.org/10.11588/dco.2015.1.18641>.

### **Nutzungsbedingungen / Terms of use:**

**CC BY-SA 4.0**



## Vom Thesaurus zum semantischen Netz. Potenziale von Data Mining in bibliographischen Datensätzen

Andreas Hartmann, Sabine Thänert

**Abstract:** The paper explores the potential of data mining in bibliographic records. Quantitative analysis of co-occurrences of keywords in bibliographic records makes it possible to create a semantic network. In contrast to the rigid hierarchy of traditional thesauri, this network maps polyvalent semantic relations between keywords across the distinct thematic branches of a thesaurus tree. Analysis of a comprehensive data pool made up from two major bibliographies (Gnomon Bibliographic Database and ZENON DAI) should allow for the semantic mapping of a core set of central topics (people, places, social structures, concepts) of classical scholarship.

### Einführung

Unser Beitrag beschäftigt sich mit den Möglichkeiten, bibliographische Datensätze einer automatisierten semantischen Auswertung zu unterziehen. Wir stellen dabei zum einen Ergebnisse des mittlerweile abgeschlossenen CLARIN-D-Kurationsprojektes „Bibliographische Datenbanken als visualisiertes Wissensnetz“ vor,<sup>1</sup> zum anderen wollen wir weitere Entwicklungsperspektiven skizzieren, die sich aus diesem Projekt ergeben haben.

### CLARIN-D-Kurationsprojekt „Bibliographische Datenbanken als visualisiertes Wissensnetz“

Ziel des CLARIN-D-Kurationsprojektes war der Aufbau einer altertumswissenschaftlichen Literaturplattform, über die der Zugriff auf die in einem gemeinsamen Index zusammengeführten Daten der *Gnomon Bibliographischen Datenbank*<sup>2</sup> und des ZENON DAI<sup>3</sup> möglich ist.<sup>4</sup> Zudem sollten neue Möglichkeiten der Auswertung und Visualisierung von Treffermengen und semantischen Kontexten erprobt sowie eine Anbindung an die Objektdatenbank *Arachne*<sup>5</sup> hergestellt werden. Die Umsetzung erfolgte durch den Lehrstuhl für Alte Geschichte der Universität Augsburg und das Deutsche Archäologische Institut.

1 URL: <http://de.clarin.eu/de/fachspezifische-arbeitsgruppen/f-ag-4-altertumswissenschaften/kurationsprojekt-2.html>.

2 URL: <http://www.gnomon-online.de>.

3 URL: <http://opac.dainst.org>.

4 Zugriff auf die in laufender Entwicklung befindliche Testplattform unter URL: <http://nighthorse06.dai-cloud.uni-koeln.de:8080/clarintest2/>.

5 URL: <http://arachne.uni-koeln.de>.

Aus Nutzerperspektive ist die stärkere Integration und Vernetzung der einzelnen bibliographischen Datenbanken und Objektdatenbanken zweifellos wünschenswert, um eine zeitintensive Parallelrecherche sowie den manuellen Abgleich von Dublettentreffern unnötig zu machen. Die Integration von bibliographischen Informationsangeboten und Datenbanken in einen gemeinsamen Index bietet dabei viel weitergehende Entwicklungsmöglichkeiten als eine Metasuche.<sup>6</sup> Die *Année Philologique* scheidet derzeit aufgrund ihrer Kostenpflichtigkeit für die Entwicklung einer offenen Fachinformationsplattform leider aus. Dasselbe gilt für *Dyabola*, was aber insofern weniger ins Gewicht fällt, da die entsprechenden Datenbestände zum ganz überwiegenden Teil über ZENON zugänglich sind.

Die *Gnomon Bibliographische Datenbank* entstand von vornherein als elektronische Literaturdatenbank und deckt, anders als die *Année Philologique* und ZENON, die Altertumswissenschaften in der Tradition der Zeitschrift Gnomon vorwiegend mit einem Erfassungsschwerpunkt in der Alten Geschichte (einschließlich aller Hilfswissenschaften) und Klassischen Philologie ab, wobei Alte Geschichte inklusiv verstanden wird als eine Geschichte der Kulturen des antiken Mittelmeerraumes, weshalb auch Arbeiten zum antiken Judentum und den sogenannten Randkulturen berücksichtigt werden. In die *Gnomon Bibliographische Datenbank* fließen Datenausspeicherungen der Bayerischen Staatsbibliothek, den Universitätsbibliotheken in Augsburg, Eichstätt und Tübingen sowie der *Joint Library of the Hellenic and Roman Societies* in London ein. ZENON zeichnet sich hingegen im Vergleich zur *Année Philologique* und *Gnomon Bibliographischer Datenbank* durch die Erfassung des gesamten archäologischen Fächerkanons ab, während Alte Geschichte und Klassische Philologie nicht systematisch abgedeckt werden. ZENON verzeichnet und erschließt den Bestand aller DAI-Bibliotheken, sowie des DEI Amman und der Winckelmann-Gesellschaft teilweise bis auf Aufsatzebene. Die Bibliothek der *British School at Athens* erfasst in enger Kooperation mit dem DAI anteilig griechisch-sprachige Zeitschriften formal und topographisch in ZENON und dem Gazetteer des DAI, iDAI.gazetteer.<sup>7</sup> Weiterhin werden elektronische Ressourcen nachgewiesen.

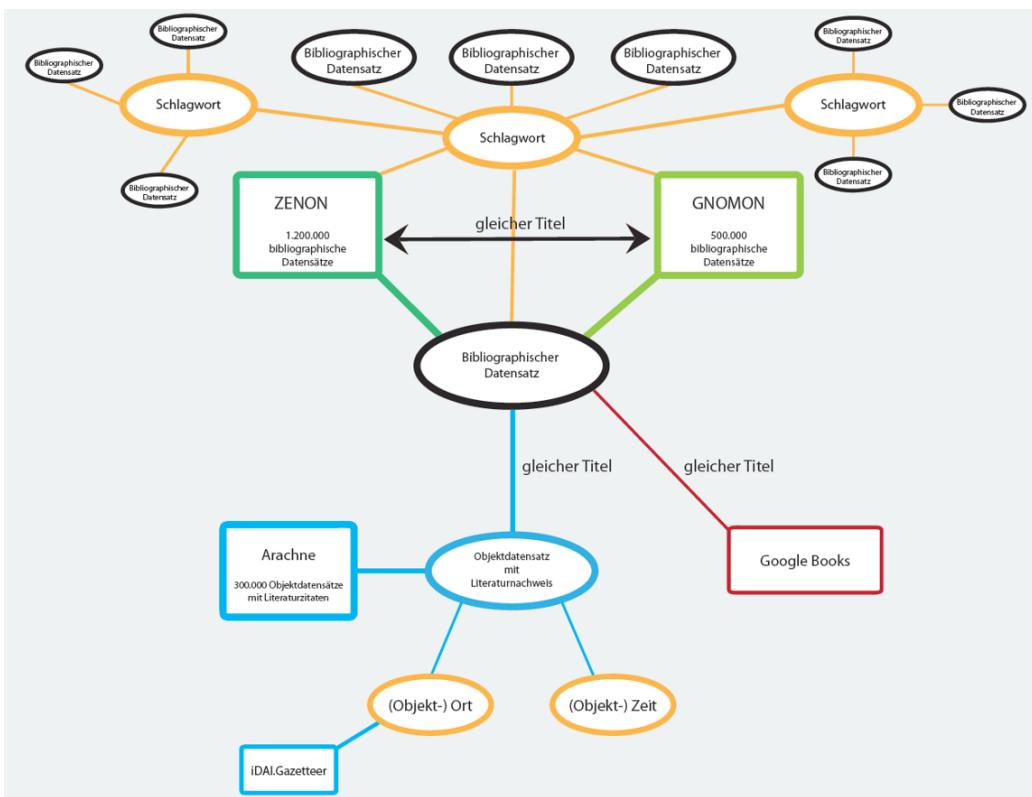
*Gnomon Bibliographische Datenbank* und ZENON zeichnen sich – im Gegensatz zur *Année Philologique*, die abseits der Gliederung nach antiken Autoren nur über eine grobe Systematik verfügt – durch die Verwendung differenzierter Thesauri aus. Diese Thesauri weisen Überschneidungen auf, sind aber den jeweiligen fachlichen Schwerpunkten entsprechend in weiten Teilen als komplementär zu betrachten. Bereits die bloße Zusammenführung der Daten in einer Plattform führt daher bei Titeln, die in beiden Quelldatenbanken vorhanden sind, oft zu einer signifikanten Verbesserung der Verschlagwortung. Gerade die Dichte der Verschlagwortung stellt den entscheidenden Mehrwert fachbibliographischer Informationsangebote dar.

Aus der Zusammenführung der sich thematisch ergänzenden Datenbestände von *Gnomon Bibliographischer Datenbank* und ZENON entstand ein umfassender altertumswissenschaftlicher Datenpool, der nicht nur für klassische bibliographische Recherchen herangezogen werden kann, sondern – und damit über die Metasuche *PropylaeumSearch* hinausgehend – auch in neuartiger Weise mit informatischen Methoden ausgewertet werden kann und sollte. Idealerweise erschließen neuartige Visualisierungen implizite thematische Zusammenhänge und ermöglichen neuartige Wissenszugänge.

---

<sup>6</sup> Eine solche wird seit geraumer Zeit im Rahmen der Virtuellen Fachbibliothek Propylaeum bereitgestellt:  
URL: <http://www.propylaeum.de/metaopac>.

<sup>7</sup> URL: <http://gazetteer.dainst.org>.



**Abb. 1: Vernetzung der Datenbestände von Gnomon Bibliographischer Datenbank und ZENON**

Abb. 1 verdeutlicht die geplanten Vernetzungen. Neben dem Mapping des Sacherschließungsvokabulars (Schlagwort-Thesauri) ging es im Projekt auch um die Datenanreicherung sowie Visualisierungen topologischer Art, d.h. um die Abbildung von Wissenskontexten. So sollten auch die in den Schlagwortbäumen der Datenbanken seit jeher enthaltenen, von den Benutzern aber nur selten explizit wahrgenommenen Wissenstopologien besser sicht- und nutzbar gemacht werden.

Das Projekt war in zwei Module untergliedert und diese wiederum in mehrere Arbeitsschritte. Das erste Modul umfasste Aufbereitung und Abgleich beider Datenbestände für die Integration in die gemeinsame Literaturplattform, die Integration der Sacherschließungsvokabulare und die Disambiguierung der Daten. So konnte durch den Abgleich der Literaturnachweise auf Textstringebene die gemeinsame Datenbasis für den Index erzeugt werden. Dabei wurde unter anderem die Bibliothek `diff_match_patch` verwendet,<sup>8</sup> um die durch unterschiedliche Aufnahmekonventionen und Tippfehler verursachten Probleme durch *fuzzy matching* zu umgehen.

Zusätzlich konnten durch eine umfassende Disambiguierung bestehender Daten eindeutige Identitäten nachgewiesen werden, unter anderem mit dem Ziel, Verlinkungen innerhalb der Datensätze zu ermöglichen. Hierbei wurden Ambivalenzen (Disambiguierung) und Dubletten (*coreference resolution*) im Bereich der antiken Personennamen sowie der Autoren- und Zeitschriftenindizes eliminiert. Diese Maßnahmen reagierten in erster Linie auf akute Matchingfehler. Sie können eine in Zukunft zu leistende durchgängige Umstellung auf Normdaten nicht ersetzen, verbessern aber dafür die Voraussetzungen substantiell. Die manuelle Disambiguierung führte somit zu einer über das Projekt weit hinauswirkenden Verbesserung der Datenqualität. Von dieser Basis ausgehend ließ sich der Datenabgleich zwischen der *Gnomon Bibliographischen Datenbank* und ZENON durch entsprechend angepasste Matching-Kriterien im Autorenfeld erfolgreich durchführen.

<sup>8</sup> URL: <https://code.google.com/p/google-diff-match-patch/>.

Technik und Material	TM		
Technik umfassend	TMUmf	Wissenschaft und Technik der Antike	2
Technik umfassend	TMUmf	Technologie der Antike	2
Ingenieurwesen	TMIng	Wissenschaft und Technik der Antike	4
Ingenieurwesen	TMIng	Technologie der Antike	4
Bauwesen	TMIngBau	Baupolitik & Bauprogramm	1
Bauwesen	TMIngBau	Bautechik	2
Bergbau, Steinbrüche und Rohstoffgewinnung	TMIngBerg	Bergbau	2
Bergbau, Steinbrüche und Rohstoffgewinnung	TMIngBerg	Steinbruch	2
Heiz- und Belüftungsanlagen	TMIngHeiz	Heizung	2
Mechanik	TMIngMech	Mechanik	1

Abb. 2: Ausschnitt der Mapping-Tabelle mit dem Sacherschließungsvokabular von Gnomon Bibliographischer Datenbank (re.) und ZENON (li.)

In einem zweiten Schritt wurde das Sacherschließungsvokabular der Thesaurusbäume von *Gnomon Bibliographischer Datenbank* und ZENON (Zweig „Klassische Archäologie“) erfolgreich aufeinander gemappt. Dabei wurden die Deskriptoren der beiden Thesauri gegenübergestellt (Abb. 2). Die Übereinstimmungen und Unterschiede jedes einzelnen Descriptors der *Gnomon Bibliographischen Datenbank* und des ZENON wurden mit einem Zahlencode zur Abbildung der semantischen Beziehung versehen. Dabei wurden folgende Fälle unterschieden: *Gnomon Bibliographische Datenbank*-Begriff im Vergleich zu ZENON-Begriff 100% entsprechend (1), exklusiver (2), inklusiver (4) oder nur je nach Kontext entsprechend (3). (5) zeigt das Fehlen jeder inhaltlichen Entsprechung an.

Im zweiten Modul erfolgte die technische Umsetzung der Konzeption, der Metadatenkonvertierung sowie der Bereitstellung. Für die OAI-Schnittstelle von ZENON wurde ein Harvester entwickelt. Alle relevanten Felder wurden in einer eigenen Datenbank abgespeichert. Das Harvesting war komplex, da die ZENON-Daten im vollständigen MARC21-XML mit einigen Erweiterungen ausgeliefert werden. Die Daten der *Gnomon Bibliographischen Datenbank* mussten aufbereitet, normalisiert und ebenfalls importiert werden. Nach der Normalisierung folgte das bereits erwähnte Matching der Daten (gemeinsamer Index). Zunächst erfolgte das Matching der Titeldaten. Die Prozedur konnte so erweitert werden, dass auch die Datensätze von Rezensionen den Datensätzen der jeweils besprochenen selbstständigen Publikationen zugeordnet werden. Zum Schluss wurde auf Grundlage der Mapping-Tabelle das Matching für die Schlagworte implementiert, damit auch über die Kategorien eine Beziehung hergestellt werden kann.

## Zenon

Titel:	Women in the Spartan revolutions of the third century B.C.
Verfasser:	Mossé, C.
Thesaurus:	griechisch,
Erschienen in:	Women's history and ancient history /
Gazetteer ID:	2043686
Arachne-Ort ID:	54

## Gnomon

Tit:	Women in the Spartan revolutions of the third century B.C. (Translated by Sarah B. Pomeroy)
Stit:	Women's history and ancient history. Sarah B. Pomeroy (Ed.). Chapel Hill/London: North Carolina UP
Ztit:	Sparta
Verfasser:	Mossé, Claude
Jahr:	1.991.
Ort:	:
Schlagwortkette:	Frauen -> Gesellschaft Frauen, spartanische -> Frauen, griechische -> Frauen -> Gesellschaft Sparta, hellenistische Zeit -> Hellenismus (323 - 31 v. Chr.) -> Griechische Geschichte -> Geschichte Hellenismus (323 - 31 v. Chr.) -> Griechische Geschichte -> Geschichte Frauen, griechische -> Frauen -> Gesellschaft

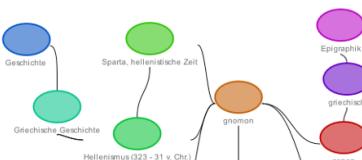


Abb. 3: Beispiel für einen gematchten Datensatz mit Kontext-Graph

Abb. 3 zeigt ein Beispiel für einen gematchten Datensatz. Der mit Hilfe der Bibliothek *Graph Dracula*<sup>9</sup> bereitgestellte Graph präsentiert mit den verknüpften Thesaurusbäumen von *Gnomon Bibliographischer Datenbank* und ZENON einen semantischen Kontext zum ausgewählten Datensatz und bildet gleichzeitig die unterschiedlichen Fachkontakte ab. Gegenwärtig werden für die Verknüpfung nur diejenigen Deskriptoren verwendet, die in beiden Datenbanken inhaltsgleich verwendet werden (Mapping-Kategorie 1). Die Einbeziehung der bereits erfassten asymmetrischen Beziehungen zwischen den Thesaurusbegriffen von *Gnomon Bibliographischer Datenbank* und ZENON in die Visualisierung des Graphen besitzt jedoch fraglos viel Potenzial und soll Ziel eines Folgeprojektes sein.

Aufgrund der unterschiedlichen Datensatzstruktur vor allem bei Sammelbänden entschieden wir uns für die Anzeige in einer gemeinsamen Einzeltrefferanzeige. Damit wurde auch das komplexe Problem der Zusammenführung von im Detail divergierenden Feldinhalten bzw. der Kriterienbildung für eine automatisierte Entscheidung für die Übernahme aus einer der beiden Ursprungsdatenbanken in jedem einzelnen Feld eines jeden Datensatzes vorerst umgangen.

Die auf der gemeinsamen Plattform zusammengeführten bibliographischen Datensätze sind eindeutigen und permanenten Linked-Data-URIs zugeordnet und somit für die Fachgemeinschaft nachnutzbar.

## Zenon

Titel:	La colonne Trajane et les forums impériaux.
Verfasser:	Galinier, Martin
Ort:	Rome :
Verlag:	École Française de Rome,
Thesaurus:	Trajanssäule, Bildsprache/Bildprogramme, Roma-Fori Imperiali, Roma-Forum Traiani, Dakerkriege,

### Verknüpfte Arachne Objekte

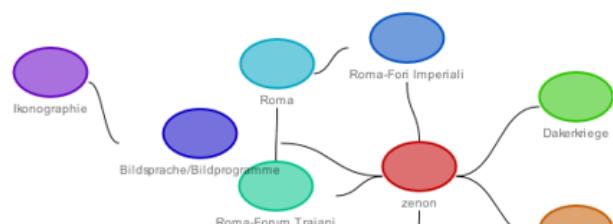
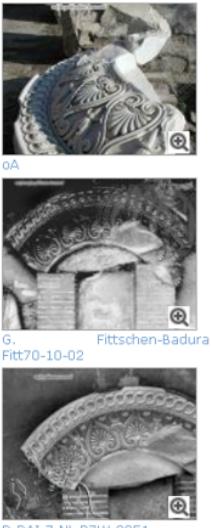


Abb. 4.1: Beispiel für eine Verknüpfung mit Arachne: ZENON-Datensatz

<sup>9</sup> URL: <http://www.graphdracula.net>.

47011: Fragment eines Clipeus von der Attika der Ost-Porticus  
Roma, Rom, Trajansforum

Vorhandene Bilder: 3



**Informationen zum Objekt - Allgemein**

**Aufbewahrung:**  
Roma, Antiker Ortsname: Roma, Italien, IT, Trajansforum. Inv.-Nr. 4037-4041  
→ [Ort im Gazetteer](#)

**Herkunft:**  
Italien, Trajansforum

**Gattung/Kulturepoche/Funktion:**  
Bauornamentik  
zu Monument gehörig: nein  
Kulturepoche: römisch

**Datierung:**  
1. Viertel 2. Jh. n. Chr.

**Erhaltung:**  
fragmentiert  
Bearbeitungen: nicht ergänzt

**Maße/Material:**  
Material: Marmor

**Literatur:**  
M. Galinier, La colonne Trajane et les forums impériaux (2007) 172 f. Taf. LXXIV a) Fig. 40;  
→ [Datensatz in ZENON DAI](#)

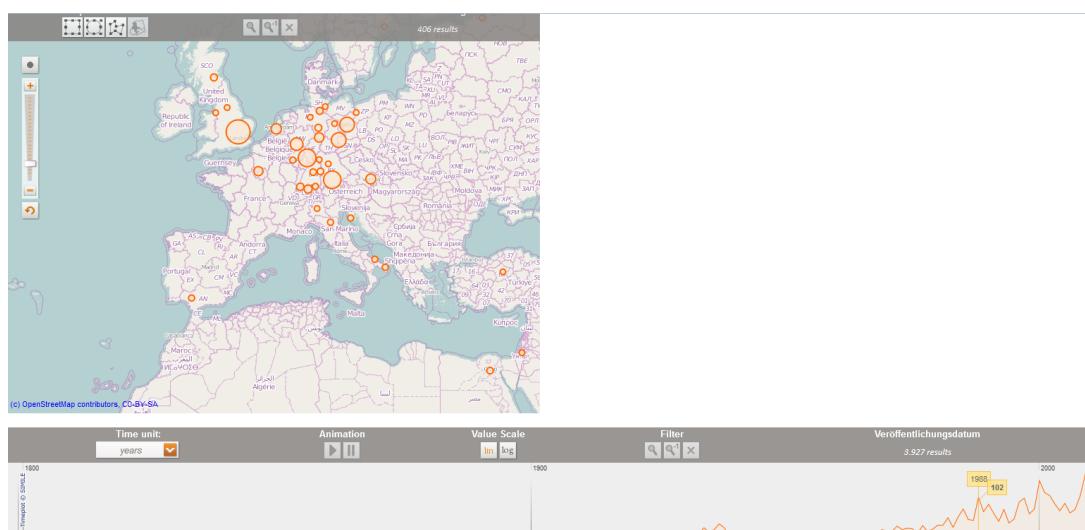
C.F. Leon, Die Bauornamentik des Trajansforums und ihre Stellung in der früh- und mittelkaiserzeitlichen Architekturdekorations Roms (Wien 1971) 80 ff.;  
→ [Datensatz in ZENON DAI](#)

Nachlass Konstantin Ronczewski (1874-1935);  
P. Zanker, Das Trajansforum in Rom, AA 1970, 511 Abb. 19;  
→ [Datensatz in ZENON DAI](#)

**Abb. 4.2: Beispiel für eine Verknüpfung mit Arachne: Arachne-Datensatz**

Der Abgleich der Literurnachweise mit der Objektdatenbank *Arachne* konnte für jene Objekte umgesetzt werden, deren Literaturangaben aussagekräftig waren und bereits auf ZENON verlinkt sind. Abb. 4 zeigt exemplarisch die Verknüpfung zwischen dem bibliographischen Datensatz und dem Literaturzitat in *Arachne*. Für jene Literaturangaben bei Objektdatensätzen, die verkürzt vorlagen, konnte vorerst keine Zuordnung erfolgen, da beispielsweise bei Aufsätzen der eigentliche Titel fehlt, sondern nur ein Verweis auf die Zeitschrift, in welcher der Beitrag enthalten ist, vorhanden ist. Ein einfacher Textstringabgleich scheidet daher als Instrument für ein umfassendes Matching aus.

Auf der übergeordneten Ebene der Treffermengen wurde für die Auswertung und Visualisierung von Zeit- und Ortsdaten das Framework *GeoTemCo* genutzt.<sup>10</sup>



**Abb. 5: Exemplarische Auswertung für den Suchbegriff „Augustus“**

10 URL: <http://www.informatik.uni-leipzig.de/geotemco>.

## Entwicklungsperspektiven

Ziel unserer zukünftigen Arbeiten soll nun die Berechnung und Darstellung eines Wissensnetzes werden, welches das in den bibliographischen Datensätzen abgebildete Fachvokabular in seinen semantischen Kontexten erschließt. Zwar liegen in Form elaborierter Thesaurus-Systeme bereits Wissenstopologien vor, die jedoch verschiedene Nachteile haben:

(1) Schlagwortthesauri sind prinzipiell hierarchisch-linear aufgebaut. Erste Ansätze dies zu ändern, verfolgt das DAI in verschiedenen Projekten. Grundsätzlich ist diese hierarchisch-lineare Struktur ein Erbe aus der Zeit der Zettelkataloge. Dieses Gliederungsprinzip hat allerdings den Nachteil, dass Zusammenhänge zwischen Begriffen, die in verschiedenen Zweigen angesiedelt sind, nicht unmittelbar transparent sind. Besonders deutlich wird das am Beispiel antiker Autoren: Dass „Arrian“ und Alexander d. Gr. in einem sachlichen Zusammenhang stehen, kann man in den Schlagwortbäumen nicht erkennen, weil Autoren und politische Persönlichkeiten verschiedenen Zweigen zugeordnet sind. Manuelle Querverweise können diesem Problem nur bedingt abhelfen, zumal ihre Setzung mit erheblichem Aufwand verbunden wäre.

(2) Die in den Thesauri beschriebenen semantischen Kontexte entsprechen einer normativen Setzung des jeweiligen Erstellers, sie sind keine Deskription, die aus den Datensätzen und damit mittelbar den Inhalten der indizierten Publikationen selbst abgeleitet ist. Sie spiegeln daher in erster Linie die Forschungstradition und bilden neue Tendenzen nur mit Verzögerung ab. Daraus folgt,

- a) dass man nicht immer Aussagen darüber machen kann, ob im Thesaurus benachbarte Begriffe eher eng oder lose miteinander verwandt sind (dies gilt *a fortiori* auch für nicht benachbarte Begriffe, vgl. [1]);
- b) dass die Schlagwortthesauri eine eher statische Wissensordnung darstellen, in die neue Konzeptionalisierungen der Forschung erst bewusst eingepflegt werden müssen.

Aus dieser Defizitanalyse ergeben sich folgende Handlungsziele:

(1) Über eine Visualisierung der verknüpften Thesaurusbäume hinaus müssen die Schlagwortketten in den einzelnen Datensätzen ausgewertet werden, in denen eine Vielzahl semantischer Verknüpfungen beschrieben sind, die gegenwärtig für den Benutzer eines bibliographischen Informationssystems nicht erfassbar sind. In diese informationstechnische Auswertung sollten auch jene „Nicht-Deskriptoren“ einfließen, mit denen das DAI bereits sein Vokabular anreichert und die in der *Gnomon Bibliographischen Datenbank* bei einzelnen Datensätzen in einem freien Textfeld erfasst werden.

Die sich daraus ergebenden Bezüge sind prinzipiell nicht hierarchisch, sie sind deskriptiv und dynamisch. Um bei dem oben angeführten Beispiel zu bleiben: Eine Auswertung der Schlagwortketten würde zeigen, dass die Schlagworte „Arrian“ und „Alexander d. Gr.“ sehr häufig miteinander kombiniert sind, das heißt, es zeichnen sich Kookkurrenzen bzw. Kollokationen ab (Abb. 6). Darüber hinaus lassen sich aufgrund der Quantifizierung der jeweils untersuchten Schlagwortkopplungen auch Gewichtungen der hergestellten Relationen ermitteln. „Arrian“ und „Alexander d. Gr.“ würden demnach als näher verwandt erscheinen als „Arrian“ und „Schwarzmeergebiet“. Dieses Ergebnis könnte sich jedoch (theoretisch) bei einer Verschiebung der Forschungsinteressen auch dynamisch ändern.

	GBD	ZENON	Gesamt
Alexander d. Gr.	60	9	69
Hellenismus	6		6
Quintus Curtius	3	2	5
Xenophon	3	2	5
Diodorus Siculus	1	4	5
Epiktet	4		4
Thukydides	2	2	4
Herodot	2	2	4
Plutarch	2	2	4
Strabon		4	4
Kavallerie		4	4
Strategie/Kampfesweise		4	4
Appian	3		3
Plutarchi Alexander	3		3
Alexanderzug	3		3
Griechische Geschichtsschreibung	3		3
Lukian	3		3
Alanen	2	1	3
Schwarzmeergebiet	2	1	3
Militärwesen		3	3
Schutzwaffen/Rüstungen		3	3
Meteorologie	2		2
Luciani Quomodo historia	2		2
Sprachwissenschaft		2	2
Strategie		2	2

Abb. 6: Exemplarische manuelle Auswertung von Kookkurrenzen des Deskriptors „Arrian“ mit anderen Deskriptoren

(2) Die vorgeschlagene Auswertung der Schlagwortketten generiert Querverbindungen zwischen den verschiedenen Zweigen der Thesaurusbäume, sodass sich zusammen mit dem bereits durchgeführten Matching der verschiedenen Thesaurusvokabulare ein dichtes Netz von semantischen Bezügen ergibt. Diese offensichtliche Visualisierung als Graph durch Punkte und Verbindungslinien ist nicht ideal, und auch die hierarchische Baumdarstellung ist keine geeignete Form zur Visualisierung der errechneten Bedeutungskontexte. Es bietet sich für die Darstellung aller Beziehungen zu einem einzigen Knoten (z. B. „Arrian“) beispielsweise eine mit Farben und Schriftgrößen codierte Tag-Cloud an. Das gezeigte Beispiel (Abb. 7) soll nur eine ungefähre Vorstellung vermitteln. Die Visualisierung beruht auf einer einfachen Auszählung – nicht berücksichtigt sind dabei die im Thesaurus-Mapping ermittelten asymmetrischen Beziehungen zwischen einander zugeordneten weiteren und engeren Begriffen. Die in Abb. 6 aus Gründen der Übersichtlichkeit nicht gezeigten Schlagworte mit nur einem Treffer würden in der Summe zu einer deutlichen Aufwertung des Begriffes Militärwesen führen. Auf der Ebene der Visualisierung selbst ergibt sich das Problem, dass die üblichen Cloud-Generatoren – in diesem Fall *Voyant Cirrus*<sup>11</sup> – nicht in der Lage sind, die Tags entsprechend vorgegebener Sachgruppen zu ordnen. Somit können sie jedoch aus der vorhandenen Wissens-topologie der Thesauri keinen Nutzen ziehen.

11 URL: <http://voyeurtools.org/tool/Cirrus>.



**Abb. 7:** Visualisierung der Kookkurrenzen aus Abb. 5 als Tag-Cloud, erzeugt mit Voyant Cirrus

(3) Das Wissensnetz wird umso dichter und aussagekräftiger, je breiter die Datenbasis ist. Bisher existiert ein gemeinsamer Index von *Gnomon Bibliographischer Datenbank* und ZENON. Schon die experimentelle Auswertungstabelle (Abb. 6) zeigt, wie sich die Deskribierungen beider Datenbanken ergänzen und in der Summe zu einem ausgewogeneren Bild führen. Darüber hinaus kann eine weitere Verdichtung des Wissensnetzes durch eine systematische Auswertung aller in der Objektdatenbank *Arachne* in unstrukturierter Form hinterlegten Literaturhinweise erfolgen (Abb. 8). Man kann davon ausgehen, dass Literaturtitel, die denselben Objekt zugeordnet sind, miteinander verwandt sind und damit mittelbar auch Aussagen über die Relation der jeweils in *Gnomon Bibliographischer Datenbank* und ZENON zugeordneten Schlagwörter abgeleitet werden können. Um dieses Ziel zu erreichen, muss jedoch ein völlig neuer Matchingalgorithmus entwickelt werden, der zwischen verschiedenen Publikationstypen differenziert und nicht primär auf einem Textstringabgleich im Titelfeld basiert.

(4) Die Crosskonkordanzen der beiden Thesauri sollen über SKOS (*Simple Knowledge Organisation System*) modelliert und damit semantisch interoperabel gemacht werden.<sup>12</sup> Diese Zusammenführung kann beispielsweise in einem Triplestore erfolgen.

Literatur

- Helbig4 II, 74, Taf. 13, 2 ♂

M. Moretti, Ara Pacis Augustae (1948), 32, Taf. 4 f. ♂

G. M. Koeppel, Bjb 187, 1987,, 196, Anm. 35, Abb. 62, 5

G. M. Koeppel, Bjb 188, 1988,, 61, Anm. 496 ♂

E. Simon, Ara Pacis Augustae, Monumenta Artis Antiquae 1 (1967), 51 f. ♂

S. Settis, Die Ara Pacis, in: M. R. Hofter (Hrsg.), Kaiser Augustus und die verlorene Republik, Ausstellungskat. Berlin (1988)

Th. Kraus, Die Ranken der Ara Pacis (1953)

N. Hannestad, JRA 13, 2000,

M. Torelli, Typology and Structure of Roman Historical Relief (1982) ♂

Abb. 8: Beispiel für Literaturangaben in Arachne

<sup>12</sup> URL: <http://www.w3.org/2004/02/skos>. Vgl. Mayr u.a. 2010.

Abschließend wäre die Frage nach dem wissenschaftlichen Nutzen eines solchen Wissensnetzes zu stellen. Auf den ersten Blick liegen Berührungs punkte mit Morettis Konzept des *distant reading* auf der Hand,<sup>13</sup> insofern vorliegende Metadaten zu einer großen Zahl an Publikationen ausgewertet und die Ergebnisse in graphischer Form visualisiert werden. Freilich ist zu beachten, dass Moretti sein Konzept in Abgrenzung zu einer bestimmten stark kanonbezogenen Praxis der Literaturwissenschaft entwickelte. Ein exegetisches *close reading* in diesem Sinne ist in der Geschichtswissenschaft seit jeher nicht üblich und Moretti hat sich denn auch nicht von ungefähr auf einen Historiker, nämlich Fernand Braudel, als eines seiner Vorbilder bezogen.<sup>14</sup>

Neue Perspektiven könnte das Konzept des *distant reading* in Verbindung mit dem Wissensnetz demnach vor allem in zwei Bereichen eröffnen: Zum einen für eine Globalgeschichte, die ähnlich der von Moretti anvisierten Weltliteraturgeschichte aus pragmatischen Gründen nicht mehr auf einem *close reading* aller relevanten Quellen aufbauen kann,<sup>15</sup> zum anderen für die Wissenschaftsgeschichte. Denkbar wäre beispielsweise, die Rekonfigurationen des Wissensnetzes für bestimmte chronologisch oder regional definierte Teilmengen des Datenpools zu untersuchen.

Freilich wäre es falsch, *distant* und *close reading* in einen strikten Gegensatz zu bringen: Die im Wissensnetz visualisierten Zusammenhänge basieren letztlich auf Publikationen, die ihrerseits das Ergebnis eines *close reading* sind. Auch die primär ausgewertete Verschlagwortung geht (noch?) nicht auf quantifizierendes *distant reading* zurück, sondern auf eine direkte Inspektion der jeweiligen Forschungstexte.

Man könnte nun einwenden, dass das skizzierte Wissensnetz auch nur die Verknüpfungen anzeigt, die ohnehin jeder erwartet hätte, der im Feld einigermaßen orientiert ist. Ein Mehrwert bei der Einbindung in Literaturrecherchesysteme ergibt sich daher zunächst für studentische und interdisziplinäre Nutzer. Der entscheidende Wert des Wissensnetzes liegt jedoch darin, dass es das in den bibliographischen Datensätzen implizit vorliegende und elementarisierte Wissen der *scientific community* über inhaltliche Zusammenhänge expliziert und formalisiert. Damit lässt sich dieses Wissen über semantische Bezüge zwischen den Begriffen des Fachvokabulars für weitere computergestützte Auswertungen nutzbar machen.

Das größte Potenzial könnte das Wissensnetz daher in einem *distant reading* spezieller Art entfalten, nämlich der Auswertung elektronischer Volltexte. Die im Wissensnetz abgebildeten semantischen Kontexte dürften es erlauben, in unstrukturierten Daten aufgrund von Konkurrenzen und Kollokationen Sachkontakte zu identifizieren, die in der Folge die eindeutige Identifizierung genannter Personen, Orte und Konzepte zulassen.

Da die Schlagwortthesauri von *Gnomon Bibliographischer Datenbank* und ZENON bereits mehrsprachig vorliegen, ist auch das Wissensnetz prinzipiell multilingual und zur Anwendung auf Texte in allen etablierten Wissenschaftssprachen geeignet.

In einem ersten Schritt könnte man ein solches Vorgehen auf die Titeldaten der bibliographischen Datensätze selbst anwenden: Dort, wo Daten ohne bzw. mit unzureichender Verschlagwortung vorliegen, könnte der Abgleich der Titelstichworte mit dem Wissensnetz eine automatisierte Basisverschlagwortung ermöglichen. Langfristig könnten sich aus der Kombination von bibliographischen Datensätzen, daraus abgeleiteten Wissensnetzen und deren Anwendung auf Volltexte ganz neue Formen der Erschließung wissenschaftlicher Forschungsliteratur entwickeln und dadurch ein großer Beitrag für die Informationsrückgewinnung gelingen.

---

13 Moretti 2005.

14 Moretti 2000, S. 56–57; Moretti 2003 *passim*.

15 Moretti 2000.

## Literatur

Mayr u.a. (2010): P. Mayr / B. Zapilko / Y. Sure, Ein Mehr-Thesauri-Szenario auf Basis von SKOS und Crosskonkordanzen, in: 25. Oberhofer Kolloquium, Magdeburg/Barleben, 22. April 2010.  
(URL: <http://eprints.rclis.org/14500/>, abgerufen am 4.2.2015)

Moretti (2000): F. Moretti, Conjectures on world literature, New Left Review 1 (2000), S. 54–68.

Moretti (2003): F. Moretti, Graphs, maps, trees. Abstract models for literary history”, New Left Review 24 (2003), S. 67–93.

Moretti (2005): F. Moretti, Graphs, maps, trees: abstract models for literary history, London / New York 2005.

## Weitere Ressourcen (zuletzt abgerufen am 04.05.2015):

Arachne, zentrale Objektdatenbank des Deutschen Archäologischen Instituts, URL:  
<http://arachne.uni-koeln.de>

CLARIN-D-Kurationsprojekt „Bibliographische Datenbanken als visualisiertes Wissensnetz“, URL:  
<http://de.clarin.eu/de/fachspezifische-arbeitsgruppen/f-ag-4-altertumswissenschaften/kurationsprojekt-2.html>  
<http://nighthorse06.dai-cloud.uni-koeln.de:8080/clarintest2/>

Comparative Visualization of Geospatial-Temporal Data, URL:  
<http://www.informatik.uni-leipzig.de/geotemco>

Diff, Match und Patch Library für Plain Text, URL:  
<https://code.google.com/p/google-diff-match-patch>

Gazetteer des Deutschen Archäologischen Instituts, URL:  
<http://gazetteer.dainst.org>

Gnomon Bibliographische Datenbank, URL:  
<http://www.gnomon-online.de>

JavaScript Graph Library, URL:  
<http://www.graphdracula.net>

Propylaeum, Virtuelle Fachbibliothek Altertumswissenschaften, URL:  
<http://www.propylaeum.de/metaopac>

Voyant Cirrus, Wordcloud Tool, URL:  
<http://voyeurtools.org/tool/Cirrus>

W3C, URL:  
<http://www.w3.org/2004/02/skos>

ZENON, Zentraler Online-Katalog der Bibliotheken des Deutschen Archäologischen Instituts, URL:  
<http://opac.dainst.org>

## Autorenkontakt<sup>16</sup>

### **Dr. phil. Andreas Hartmann**

Akademischer Oberrat am Lehrstuhl für Alte Geschichte  
Universität Augsburg  
Mail: [andreas.hartmann@phil.uni-augsburg.de](mailto:andreas.hartmann@phil.uni-augsburg.de)

URL: <https://www.philhist.uni-augsburg.de/lehrstuehle/geschichte/alte/MitarbeiterInnen/ahartmann/>

### **Dipl.-Bibl. Sabine Thänert M.A.**

Bibliotheksamtsrätin am Deutschen Archäologischen Institut (Berlin)  
Mail: [sabine.thaenert@dainst.de](mailto:sabine.thaenert@dainst.de)

URL: <http://www.dainst.org/dai/mitarbeiter>

---

<sup>16</sup> Die Rechte für Inhalt, Texte, Graphiken und Abbildungen liegen, wenn nicht anders vermerkt, bei den Autoren.