

Autorin: Lena-Luise Stahn
Betreuer: Prof. Reinhard Förtsch (DAI Berlin)
Prof. Dr. Ulf Leser (HU Berlin)

I. Einführung/ Motivation

Einen Vorteil der Verwendung traditioneller Erschließungswerkzeuge im Semantic Web Kontext stellt einerseits die Möglichkeit dar, der fortschreitenden Zersplitterung des Erschließungsvokabulars entgegenzuwirken. Andererseits wird so die Verbindung heterogener Informationsquellen erleichtert. Die Umsetzung eines 'Thesaurus' in maschinenlesbares, auf Linked Data ausgerichtetes Format ermöglicht die Erweiterung der Recherche auch mithilfe anderer, auf diese Weise verbundener Vokabulare. Diese so erreichte Interoperabilität der Terminologien bietet leichteren Zugriff auf verteilte Informationsquellen, wodurch eine Vervollständigung der Recherche-Ergebnisse erreicht werden soll.

Von einer Umsetzung der spezifischen Terminologien in das dazu notwendige Linked Data Format sind viele Fachbereiche jedoch noch weit entfernt. Die Thesauri des Deutschen Archäologischen Instituts (DAI) [DAI 2014] bilden hier ein gutes Beispiel. Seit der Gründung des Instituts im 19. Jahrhundert werden im Zuge der Erstellung des Realkatalogs, der in den heutigen Archäologischen Bibliographien aufgegangen ist, (zunächst als „wohl älteste ausführliche archäologische Systematik faßbar“ [Blanck 1979 S. 19]) umfangreiche Thesauri entwickelt, die durch ihre weitreichende Verwendung zu den wichtigsten Terminologien für diesen Bereich geworden sind und weiterhin umfassend gepflegt werden. Traditionell allerdings geprägt durch die auf die klassischen Gebiete Griechenland und Italien ausgerichtete Forschung, können sie den inzwischen wesentlich breiter orientierten wissenschaftlichen Tätigkeiten am Institut und in der Fach-Gemeinschaft nicht mehr umfassend und vollständig gerecht werden. Die Entwicklung paralleler, topographisch und chronologisch andersartig ausgerichteter „Schattenvokabulare“, zu nennen sind hier beispielsweise der iDAI.gazetteer als kontrolliertes Vokabular topographischer Ausrichtung oder die im Zuge der Gründung der Außenstelle Peking von deren Leiterin entwickelte Terminologie für die ostasiatische Archäologie, bilden daher eine notwendige Ergänzung zu den traditionellen Thesauri, um die gesamte inzwischen am DAI betriebene Forschung abzubilden. Ihre Eigenständigkeit erschwert jedoch die Recherche und kann nur durch mehrere separate Suchansätze umständliche umgangen werden. Diese parallele Existenz von Terminologien innerhalb eines einzigen Fachbereiches macht eine Verbesserung

der Interoperabilität der Terminologien dringend erforderlich. Durch die Umsetzung der DAI-Daten und -vokabulare in ein Linked Data-Format soll eine solche Verbesserung erreicht werden, um durch einen vereinheitlichten Datenraum möglichst vollständige Retrievalergebnisse gewährleisten zu können.

2. Ziel

Ziel dieser Arbeit ist es, zunächst eine Vorgehensweise zu entwickeln, die archäologische Erschließungsvokabulare des Deutschen Archäologischen Instituts in ein Semantic Web-Format zu bringen. Diese Migration soll die Grundlage bilden, Lösungsansätze dafür zu finden, inwieweit eine Interoperabilität zwischen diesen Vokabularen mit einem derzeit verfügbaren Tool automatisiert hergestellt werden kann. Aufgrund des eingeschränkten Umfangs dieser Arbeit wird das Vorgehen für ein Mapping zunächst beispielhaft durchgeführt.

Die unterschiedliche semantische Tiefe und Granularität jedes einzelnen Vokabulars soll dabei erhalten bleiben, die Interoperabilität lediglich durch Verknüpfungen der Daten hergestellt werden. Es wird dabei zu prüfen sein, inwieweit sich besonders das Datenformat SKOS (Simple Knowledge Organization System) [SKOS-Homepage 2014], welches das W3C 2009 als Standard-Datenmodell und -Vokabular für die Darstellung von Klassifikationssystemen und Thesauri ausgezeichnet hat [W3C 2013] [Baker et. al. 2013], auf die archäologischen Terminologien anwenden lässt. Dabei sollen besonders die SKOS Mapping Properties im Mittelpunkt stehen. Als Alternative, eventuell bedingt durch die Auswahl des Mapping-Tools, bietet sich OWL (Web Ontology Language) [OWL-Homepage 2014] an. Migration und Mapping sollen anhand ausgewählter, im Folgenden aufgeführter, Vokabulare durchgeführt werden:

- „römischer“ Thesaurus
- iDAI.gazetteer
- Thesaurus der Römisch-Germanischen Kommission (RGK)
- Terminologie des Langzeitprojektes „Archäologie der Oasenstadt Tayma: Kontinuität und Wandel der Lebensformen im ariden Nordwesten der Arabischen Halbinsel vom Neolithikum bis zur Islamisierung“)
- Fachvokabular für die ostasiatische Archäologie (geführt an der Außenstelle Peking)

Als langfristiges Ziel wird angestrebt, die ausführliche Dokumentation der Umsetzung in SKOS sowie des Mappings als Leitfaden für eine spätere vollständige Umsetzung auszuführen.

3. Herangehensweise

Für die Vorgehensweise dienen mehrere bereits erfolgte Projekte als Orientierung, wobei vor allem die Dokumentation der im Rahmen des DARIAH-DE-Projektes am DAI bereits durchgeführten Umsetzung des römischen 'Thesaurus' [Beer et. al. 2014] als Leitfaden genannt werden muss, des weiteren die Umsetzung des 'Thesaurus' Sozialwissenschaften (TheSOZ) des GESIS Leibniz-Institut für Sozialwissenschaften [Kempf et. al. 2014] [Zapilko & Sure 2009] [Mayr et. al. 2008] sowie das in [Mcgregor 2007] beschriebene Projekt der Strathclyde Universität, in dem zur Umsetzung von Terminologie-Interoperabilität ebenfalls SKOS angewendet wird, zu nennen sind. Eine weitere Anwendung von SKOS ist in [Keil 2012] zu finden. Der hier unternommene Versuch, eine Standardisierung des Terminologie Mappings einzuführen, unter Einbeziehung von SKOS, soll als „Best Practice Paper“ zur Orientierung für die durchzuführende Arbeit dienen. In den genannten Arbeiten hat sich allgemein eine drei Schritte umfassende Vorgehensweise zur Migration herausgebildet [Assem et. al. 2006] [Dellschaft & Hachenberg 2011] [Zapilko et. al. 2013], der in dieser Arbeit gefolgt wird.

3.1. Analyse der umzusetzenden Vokabulare

Zunächst erfolgt eine ausführliche Dokumentation der Lage der Terminologien am Institut, gefolgt von einer Analyse der oben genannten Beispielvokabulare. Hier stehen Fragen zu den hierarchischen und assoziativen Relationen im Vordergrund. Es muss geklärt werden, welche Strukturen bestehen, beispielsweise ob Relationen auch zwischen Nicht-Deskriptoren existieren. Dabei kann eine den ISO-Normen entsprechende Umsetzung des Vokabulars die Transformation in SKOS u. U. erleichtern, was jedoch nicht immer der Fall ist, wie mehrere Use Cases zeigen¹.

3.2. Zuordnung der Terme und Relationen zu SKOS classes und properties

Die im ersten Schritt durchgeführte Analyse ist erforderlich, um einschätzen zu können, inwieweit eine einfache Umsetzung in SKOS, in diesem Schritt zunächst durch Mapping der ermittelten Terme auf die

¹ <http://www.w3.org/TR/skos-ucr/> (Zugriff: 01.07.2014).

SKOS-Semantik, erfolgen kann und ob dazu Erweiterungen notwendig sind, um den Anforderungen der Terminologien zu entsprechen. Als großes Problem gilt hier beispielsweise die Konzeptbasiertheit von SKOS („Concepts“ werden mit jeweils mehreren Labeln angereichert, die als *prefLabel* und *altLabel* die Deskriptoren und Nicht-Deskriptoren als „term“ auszeichnen [SKOS Reference 2009] [Mcgregor 2007]), die Beziehungen zwischen Nicht-Deskriptoren (in SKOS dargestellt als „term“) nicht zulassen. Zu klären ist, inwieweit dementsprechende Erweiterungen vorgenommen werden müssen (u. a. im Projekt von GESIS wurde dazu SKOSxl [SKOS Reference Appendix SKOS-XL 2009] verwendet, beispielsweise um die dem TheSOZ eigenen Relationen „UF“ bzw. „USE“ umsetzen zu können [Mayr et. al. 2010 6f]). Eine eigene Erweiterung zu erstellen soll dabei jedoch möglichst unterlassen werden, um eventuelle spätere Unvereinbarkeiten mit externen Vokabularen zu vermeiden. U.U. muss eine Alternative als Beschreibungsvokabular gefunden werden.

Für das anschließende Mapping soll geklärt werden, welche SKOS Mapping Properties (*skos:exactMatch*, *skos:closeMatch*, *skos:broadMatch*, *skos:narrowMatch* sowie *skos:relatedMatch*) verwendet werden können.

3.3. Automatisierte Transformation und Mapping mittels eines Konvertierungs- und Mapping-Tools

Für diese Schritte ist zu klären, in welchem Format die Daten bisher vorgehalten worden sind und wie die zukünftigen Anwendungen darauf zugreifen sollen. Dies alles trägt zu den Entscheidungen bei, welche Tools und Programmiersprachen zur Umsetzung angewendet werden (etwa XSL) und in welchem Format die Daten gespeichert werden sollen (Triplestore). Auf Grundlage der transformierten Daten soll in einem zweiten Schritt das Mapping der einzelnen Vokabulare erfolgen, wobei ein Tool zum Einsatz kommt, welches die Cross-Referenzen der Vokabulare anhand von Syntax- und Kontext-Vergleichen automatisiert erstellt. Bei der Auswahl des Tools half eine Empfehlung aus der Fach-Community², da hier bereits positive Erfahrungen mit der Anwendung gemacht wurden. Die Entstehung des Tools „amalgame“³ erfolgte zudem im Rahmen der Entwicklung der Europeana⁴, der virtuellen Bibliothek zur Verbreitung des kulturellen Erbes Europas, sodass von einem verwandten Kontext der zu bearbeitenden Vokabulare ausgegangen werden kann. Die Eignung von amalgame für archäologisches Vokabular, außerhalb des eigentlichen Entwicklungskontextes, soll im Verlauf der Arbeit untersucht werden.

² B. Zapilko vom GESIS Leibniz-Institut für Sozialwissenschaften nannte auf Nachfrage hin drei Tools, die im Rahmen der OAEI derzeit angewendet werden, wobei amalgame aufgrund des verwandten Kontextes als am geeignetsten für diese Arbeit scheint.

³ Amalgame wurde als Erweiterung des auf Prolog basierenden frameworks für Semantic Web Anwendungen ClioPatria, entwickelt an der Freien Universität Amsterdam, erstellt, um das Mapping von SKOS-Vokabularen zu unterstützen. <http://www.w3.org/2001/sw/wiki/ClioPatria> (Zugriff: 26.01.15).

⁴ <http://www.europeana.eu/> (Zugriff: 26.01.15).

3.4. Evaluation

Eine abschließende Evaluation des erstellten Mappings kann im Rahmen dieser Arbeit nur beispielhaft erfolgen. Dies liegt in erster Linie am begrenzten zeitlichen Rahmen, der durch eine intellektuelle Überprüfung jeder einzelnen erstellten Cross-Konkordanz weit überschritten werden würde. Ein zweiter Punkt betrifft das für statistische Vergleiche notwendige Referenz-Mapping (auch reference alignment), d. h. ein intellektuell erstelltes vollständiges Mapping der beteiligten Vokabulare, welches nicht vorliegt, sodass hier nur von Schätzwerten ausgegangen werden kann.

Für die Vorgehensweise wird deshalb festgelegt:

3.4.1. Es wird ein prozentualer Anteil an Mappings zufällig herausgezogen und auf syntaktische und semantische Korrektheit überprüft (wie hoch dieser Anteil ist, wird abhängig von der Gesamtmenge der gefundenen Mappings festgelegt);

3.4.2. Bei Fragen nach dem Bedeutungserhalt bzw. nach zu erwartenden Verschiebungen auf der semantischen Ebene kann ein Fachwissenschaftler als Unterstützung/Ergänzung herangezogen werden (hier besteht bereits Kontakt zur Sacherschließungsabteilung des DAI, wo der römische Thesaurus erstellt und gepflegt wird),

3.4.3. Dann erfolgt die statistische Darstellung der Datenerhebung in Tabellenform, zusätzlich ausgewertet nach verschiedenen Vorgehensweisen des Mappings (Vergleich auf String-Ebene, an Scope notes, am Kontext), abhängig von den Möglichkeiten, die das Tool bietet.

3.4.4. In erster Linie kann so die Precision für die oben ausgewählten Mappings, beruhend auf Schätzwerten⁵, gemessen werden. Inwieweit dies auch für den Recall möglich ist, soll sich im Verlauf der Untersuchung und mit Unterstützung des oben erwähnten Sacherschließungsexperten geklärt werden.

3.4.5. Aus diesen punktuellen Daten können dann als Endergebnis Schätzungen zu Precision und evtl. Recall für das gesamte automatisiert erstellte Mapping abgeschätzt werden.

4. Ausblick

Der Ausblick wird versuchen, eine Antwort auf die Frage zu geben, ob ein automatisiertes Vorgehen

⁵ Evaluierungen solcher Mappings basieren stets auf statistischen Werten für Precision und Recall, die durch Vergleiche mit einem manuell erstellten Referenz-Alignment gewonnen werden. Ein solches liegt jedoch für die DAI-Vokabulare nicht vor.

bei der Erstellung von Mappings für archäologisches Vokabular mithilfe eines auf dem derzeitigen Entwicklungsstand befindlichen Tools machbar und sinnvoll ist. Da ein solches Mapping weit umfangreicherer Vorarbeiten in Form von Recherchen und Entscheidungen und anschließender Evaluierung, im Idealfall durch Vergleiche mehrerer angewendeter Tools und ihrer Ergebnisse, bedarf, ist dies jedoch zunächst als Wegweiser und weniger als konkret anwendbares Material gedacht. Auch Ansätze des teilautomatisierten „reference alignment“, des intellektuell unterstützten Mappings, werden einbezogen.

Literatur

- [Assem et. al. 2006] Van Assem, M., Malaisé, V., Miles, A., & Schreiber, G. (2006). A method to convert thesauri to SKOS (pp. 95-109). Springer Berlin Heidelberg.
- [Baker et. al. 2013] Baker, T., Bechhofer, S., Isaac, A., Miles, A., Schreiber, G., & Summers, E. (2013). Key choices in the design of Simple Knowledge Organization System (SKOS). *Web Semantics: Science, Services and Agents on the World Wide Web*, 20, 35-49.
- [Beer et. al. 2014] Beer, N., Herold, K., Kolbmann, W., Kollatz, Th., Romanello, M., Rose, S., & Walkowski, N.-O. (2014). Interdisciplinary Interoperability. DARIAH-DE working papers 3.
- [Blanck 1979] Blanck, H. (1979). Die Bibliothek des Deutschen Archäologischen Instituts in Rom (p. 19), Deutsches Archäologisches Institut, Geschichte und Dokumente ; Bd. 7.
- [DAI 2014] Homepage des DAI. URL: <http://www.dainst.org/de/> (Zugriff: 17.06.2014).
- [Dellschaft & Hachenberg 2011] Dellschaft, K. & Hachenberg, C. (2011) Repräsentation von Wissensorganisationssystemen (KOS) im Semantic Web: Ein Best Practice Guide 2011.
- [De Roo et. al. 2013] Sun, H., De Roo, J., Mels, G., Depraetere, K., Twagirumukiza, M., De Vloed, B., & Colaert, D. (2013). Assessing and Improving SKOS Mapping Quality. In SWAT4LS.
- [Doerr 2001] Doerr, M. (2006). Semantic problems of thesaurus mapping. *Journal of Digital information*, 1(8).
- [Doerr 2004] Doerr, M. (2004). Semantic interoperability: theoretical considerations. ICSFORTH. URL: http://www.ics.forth.gr/ftp/techreports/2004/2004.TR345_Semantic_Interoperability_Theoretical_Considerations.pdf (Zugriff: 17.06.2014).
- [Gantert 2011] Gantert, K. (2008). Bibliothekarisches Grundwissen, 177-184.
- [Knorz 2011] Knorz, G. (2011) B5 Informationsaufbreitung II: Indexierung, Grundlagen der praktischen Information und Dokumentation: Handbuch zur Einführung in die Informationswissenschaft und –praxis, 179-188.
- [Gradmann 2013] Gradmann, S. (2013). B 7 Semantic Web und Linked Open Data, Grundlagen der praktischen Information und Dokumentation: Handbuch zur Einführung in die Informationswissenschaft und -praxis, 219-228.
- [Keil 2012] Keil, S. (2012). Terminologie Mapping: Grundlagen und aktuelle Normungsvorhaben. *Information-Wissenschaft & Praxis*, 63(1).

- [Kempf et. al. 2014] Kempf, A. O., Ritze, D., Eckert, K., & Zapilko, B. (2014). New Ways of Mapping Knowledge Organization Systems. Using a Semi-Automatic Matching-Procedure for Building Up Vocabulary Crosswalks. Knowledge Organization, 41(1), 66-75.
- [Mayr et. al. 2008] Mayr, P., & Petras, V. (2008). Crosskonkordanzen: Terminologie Mapping und deren Effektivität für das Information Retrieval. Online verfügbar unter http://archive.ifla.org/IV/ifla74/papers/129-Mayr_Petras-trans-de.pdf, zuletzt geprüft am, 14, 2011.
- [Mayr et. al. 2010] Mayr, P., Zapilko, B., & Sure, Y. (2010). Ein Mehr-Thesauri-Szenario auf Basis von SKOS und Crosskonkordanzen. Recherche im Google-Zeitalter-vollständig und präzise, 163-172.
- [Mcgregor 2007] Macgregor, G., Joseph, A., & Nicholson, D. (2007). A SKOS Core approach to implementing an M2M terminology mapping server. In International Conference on Semantic Web and Digital Libraries (ICSD-2007).
- [OAEI Homepage 2013] Ontology Alignment Evaluation Initiative Homepage. URL: <http://oaei.ontologymatching.org/> (Zugriff: 28.08.2014).
- [OWL-Homepage 2014] OWL-Homepage des W3C. URL: http://www.w3.org/standards/techs/owl#w3c_all (Zugriff: 28.08.2014).
- [Peters 2013] Peters, I. (2013). B 8 Benutzerzentrierte Erschließungsverfahren. Grundlagen der praktischen Information und Dokumentation: Handbuch zur Einführung in die Informationswissenschaft und -praxis, 229-237.
- [SKOS-Homepage 2014] SKOS-Homepage des W3C. URL: <http://www.w3.org/2004/02/skos/> (Zugriff: 17.06.2014).
- [SKOS Reference Appendix SKOS-XL 2009] SKOS Simple Knowledge Organization System Reference. W3C Recommendation 18 August 2009. Appendix B. SKOS eXtension for Labels (SKOS-XL), URL: <http://www.w3.org/TR/skos-reference/#xl> (Zugriff: 24.06.14).
- [SKOS Reference 2009] SKOS Simple Knowledge Organization System Namespace Document - HTML Variant 18 August 2009 Recommendation Edition, URL: <http://www.w3.org/2009/08/skos-reference/skos.html> (Zugriff: 25.06.14).
- [Umlauf 2014] Umlauf, Informationsorganisation 1. Inhaltserschließung in Bibliotheken. Vorlesungsskript. (Berliner Handreichungen zur Bibliothekswissenschaft. 82) URL: <http://www.ib.hu-berlin.de/~kumlau/handreichungen/h82/> (Zugriff: 24.06.2014).
- [W3C 2013] W3C Press Release, From Chaos, Order: W3C Standard Helps Organize Knowledge. SKOS Connects Diverse Knowledge Organization Systems to Linked Data, 2013, URL: <http://www.w3.org/2009/07/skos-pr> (Zugriff: 24.6.14).

- [Zapilko & Sure 2009] Zapilko, B., & Sure, Y. (2009). Converting the TheSoz to SKOS. GESIS Report.
- [Zapilko et al. 2013] Zapilko, B., Schaible, J., Mayr, P., & Mathiak, B. (2013). TheSoz: A SKOS representation of the thesaurus for the social sciences. *Semantic Web*, 4(3), 257-263.