Normdatenpflege in Zeiten der Automatisierung: Erstellung und Evaluation automatisch aufgebauter Thesaurus-Crosskonkordanzen

Kempf, Andreas Oskar; Zapilko, Benjamin

Empfohlene Zitierung / Suggested Citation:
Kempf, A. O., & Zapilko, B. (2013). Normdatenpflege in Zeiten der Automatisierung: Erstellung und Evaluation automatisch aufgebauter Thesaurus-Crosskonkordanzen. *Information - Wissenschaft und Praxis*, *64*(4), 199-207. https://doi.org/10.1515/iwp-2013-0025

 DOI 10.1515/iwp-2013-0025 — Information. Wissenschaft & Praxis 2013; 64(4):199–207

Andreas Oskar Kempf und Benjamin Zapilko, Köln

Normdatenpflege in Zeiten der Automatisierung. Erstellung und Evaluation automatisch aufgebauter Thesaurus-Crosskonkordanzen

Thesaurus-Crosskonkordanzen bilden eine wichtige Voraussetzung für die integrierte Suche in einer verteilten Datenstruktur. Ihr Aufbau erfordert allerdings erhebliche personelle Ressourcen. Der vorliegende Beitrag liefert Evaluationsergebnisse des Library Track 2012 der Ontology Alignment Evaluation Initiative (OAEI), in dem Crosskonkordanzen zwischen dem Thesaurus Sozialwissenschaften (TheSoz) und dem Standard Thesaurus Wirtschaft (STW) erstmals automatisch erstellt wurden. Die Evaluation weist auf deutliche Unterschiede in den getesteten Matching-Tools hin und stellt die qualitativen Unterschiede einer automatisch im Vergleich zu einer intellektuell erstellten Crosskonkordanz heraus. Die Ergebnisse sprechen für einen Einsatz automatisch generierter Thesaurus-Crosskonkordanzen, um Domänenexperten eine maschinell erzeugte Vorselektion von möglichen Äquivalenzrelationen anzubieten.

Deskriptoren: Konkordanz; maschinell, Thesaurus, Ontology Matching, SKOS

Authority data maintenance in times of automation. Creation and evaluation of automatically established cross-concordances between controlled vocabularies

Crosswalks between thesauri are a fundamental prerequisite for an integrated search in distributed data structures. However, their manual creation is rather time-consuming. This article presents results of the Library Track of the Ontology Alignment Evaluation Initiative (OAEI) 2012, where for the first time crosswalks between the Thesaurus of the Social Sciences (TheSoz) and the Thesaurus for Economics (STW) have been detected automatically. The evaluation reveals clear differences in the performance of the participating matching tools. Moreover, qualitative differences between an automatical versus intellectual creation of crosswalks are exposed. As a result, it can be concluded that an automatic creation of crosswalks delivers a promising pre-selection of equivalence relationships for domain experts.

Keywords: thesaurus, cross-concordances, ontology matching, SKOS

La gestion de notices d'autorité à l'époque de l'automatisation. Création et évaluation de correspondances entre thesauri générés de façon automatique

Les correspondances entre différents thesauri forment une des préconditions fondamentales pour une recherche intégrée dans des structures de données distribuées. Cependant, leur création nécessite d'importantes ressources humaines. Cet article présente des résultats de la Library Track of the Ontology Alignment Evaluation Initiative (OAEI) 2012, où, pour la première fois, des correspondances entre le Thesaurus des Sciences Sociales (TheSoz) et le Thesaurus d´Économie (STW) ont été générées de manière automatisée. L'évaluation révèle de nettes différences entre les performances des outils utilisés. Par ailleurs elle démontre des différences qualitatives entre les modes de création automatiques et intellectuels. Les résultats de cette initiative montrent que la création automatique de correspondances peut fournir une très bonne présélection aux experts pour identifier les relations d'équivalence.

Mots-clés: correspondances entre thésauri, ontology matching, SKOS

1 Inhaltserschließung in Zeiten der Automatisierung

Die Automatisierung im Umgang mit Erschließungsinstrumenten betraf bisher vor allem die inhaltliche Erschließung von Datenbeständen selbst. So werden mit Zunahme digital verfügbarer Metadaten und Volltexte automatische Indexierungsverfahren als eine mögliche Antwort auf das Management unstrukturierter Daten diskutiert. Neben dem Projekt PETRUS der Deutschen Nationalbibliothek zur maschinellen Erschließung von Netzpublikationen

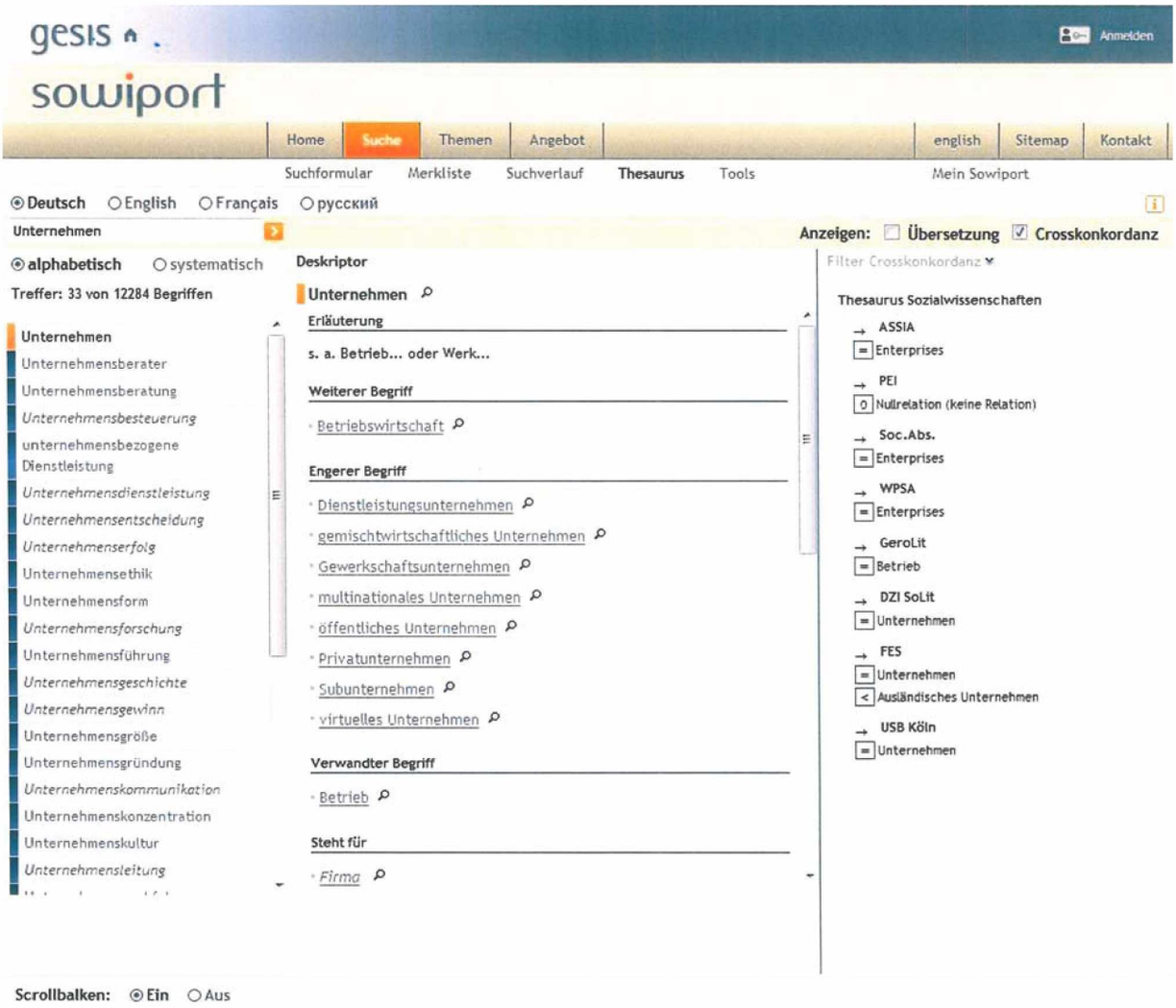

Abb. 1: Exemplarische Ansicht eines Crosskonkordanz-Datensatzes in der Online-Repräsentation des TheSoz in sowiport.

sammelten hierzu auch eine Reihe von Fachinformationszentren, wie DIPF (vgl. Wissel 2011), ZPID (vgl. Gerards 2011) und GESIS (vgl. Kempf 2012) Erfahrungen auf diesem Gebiet. Ein deutliches Desiderat bildet hingegen ein automatisierter Aufbau von Crosskonkordanzen zwischen unterschiedlichen kontrollierten Vokabularen.[1]

Thesaurus-Crosskonkordanzen bezeichnen den Aufbau von Relationen zwischen unterschiedlichen kontrollierten Vokabularen. Diese Beziehungen umfassen in der Regel neben Äquivalenz- sowie Unter- und Oberbegriffs- auch Ähnlichkeits- bzw. Verwandtschaftsrelationen. Dieserart ermöglichen Crosskonkordanzen die integrierte Suche in einer verteilten Datenstruktur. Ein Anwendungsbeispiel bildet das von GESIS betriebene sozialwissenschaftliche Fachportal sowiport. Neben den von GESIS selbst aufgebauten Datenbanken wie SOLIS (Sozialwissenschaftliches Literaturinformationssystem) und SOFIS (Sozialwissenschaftliches Forschungsinformationssystem) können darin Bestände von Zulieferern, wie etwa der Universitäts- und Stadtbibliothek Köln, erst dadurch über eine integrierte Suche zugänglich gemacht werden, dass der GESIS-Fachthesaurus Sozialwissenschaften (TheSoz) auf die Gemeinsame Normdatei (GND), wie sie von der USB Köln zur Erschließung verwendet wird, abgebildet wird.[2]

1 Bemühungen, automatisiert Crosskonkordanzen zwischen kontrollierten Vokabularen aufzubauen, lassen sich besonders in den Ontology Matching und Digital Library bzw. Information Retrieval Communities beobachten (vgl. exemplarisch Euzenat et al. 2007).

2 Zwischen der Deutschen Nationalbibliothek und GESIS existiert seit dem Jahr 2007 eine Kooperationsvereinbarung zur bilateralen Pflege und Weiterentwicklung der Crosskonkordanzen zwischen der GND und dem TheSoz für ausgesuchte Systematikstellen der GND.

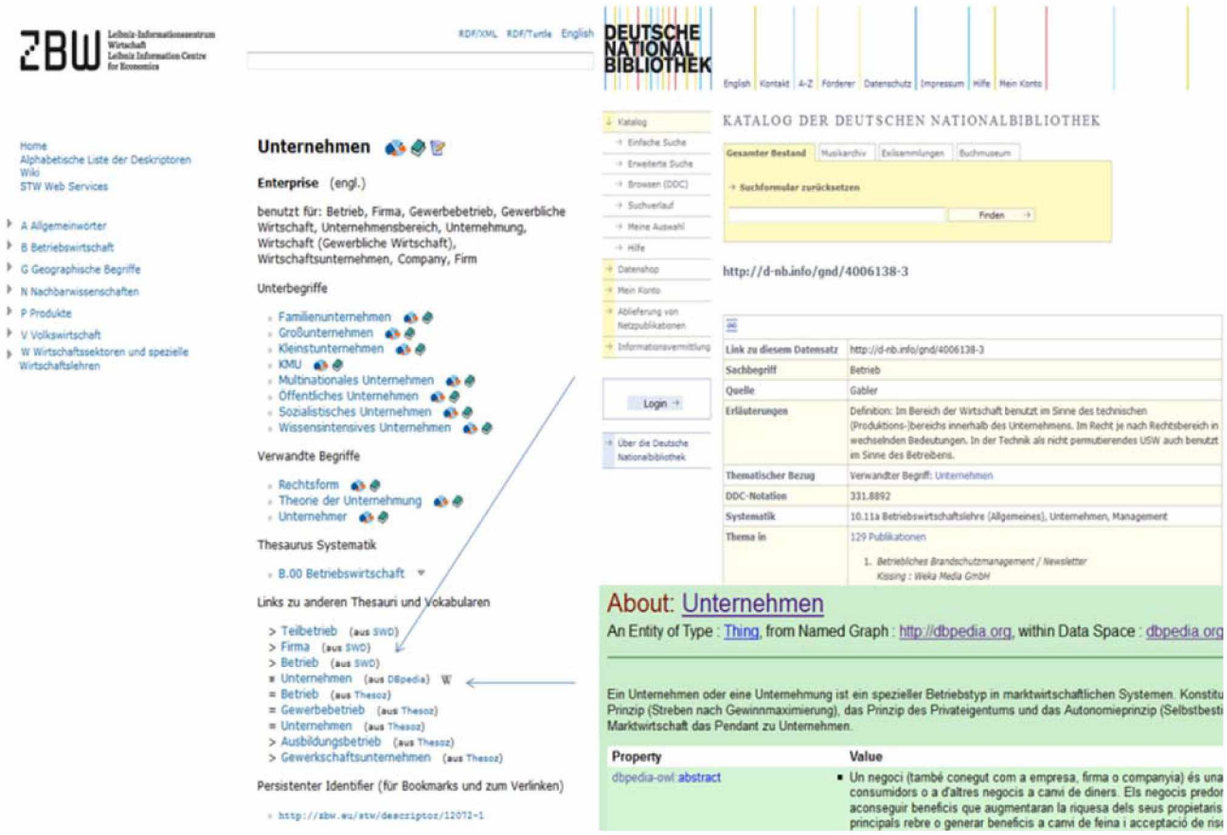

Abb. 2: Exemplarische Ansicht von Verlinkungsmöglichkeiten zwischen einem STW-Begriff und dem DNB-Katalog sowie DBpedia auf der Grundlage von Thesaurus-Crosskonkordanzen.

Daneben bilden Thesaurus-Crosskonkordanzen ein wichtiges Recherchetool für die Suche in verteilt vorhandenen Datenquellen, die immer häufiger nach dem Linked Open Data-Paradigma (LOD) im Web veröffentlicht werden.[3] So können kontrollierte Vokabulare, wenn sie etwa über das SKOS (Simple Knowledge Organization System) Datenmodell, dem maschinenlesbaren Standardformat für Thesauri und Klassifikationen des Semantic Web, im Web angeboten werden und über Crosskonkordanzen miteinander verlinkt sind, eine zentrale Brückenfunktion erfüllen, damit sich etwa Bibliotheksnutzer bei ihrer Suche zwischen unterschiedlichen im Web verfügbaren Datenquellen, wie z. B. Bibliothekskatalogen oder DBpedia, der die strukturierten Wikipedia-Informationen enthaltenden Datenansicht, bewegen können (vgl. Mayr et al. 2010). Auch in Deutschland hat die Veröffentlichung von Daten und Vokabularen nach Standards des Semantic Web Anklang gefunden. So wurden etwa der TheSoz (Zapilko et al. 2012) und der STW (Neubert 2009) im SKOS Format sowie die Gemeinsame Normdatei (GND) als Linked Open Data publiziert.

Ausgehend von dieser technischen Grundlage beschreibt der vorliegende Artikel das Vorgehen und die Ergebnisse eines automatischen Crosskonkordanz-Aufbaus zwischen den beiden inhaltlichen Erschließungsvokabularen TheSoz und STW. Nach einem Einblick in den bisherigen intellektuellen Aufbau von Thesaurus-Crosskonkordanzen wird der sog. Library Track der OAEI vorgestellt, in dessen Rahmen unterschiedliche sog. Matching-Tools zum Aufbau der Crosskonkordanzen eingesetzt wurden. Hieran folgen die Darstellung der Ergebnisse einer maschinellen Auswertung auf der Basis bereits bestehender Crosskonkordanzen zwischen beiden Thesauri sowie einer intellektuellen Evaluation, die ausschließlich die neu hinzugekommenen Crosskonkordanzen berücksichtigt. Im Anschluss an die Zusammenfassung der Ergebnisse wird ein Ausblick auf sich hieran anschließende Forschungsvorhaben geben.

3 Die Idee von Linked Open Data ist aus der Entwicklung des Semantic Web entstanden und beschreibt Methoden, um Daten und Metadaten jedweder Art mit Standards und Technologien des Semantic Web im World Wide Web zu veröffentlichen, zu teilen und miteinander zu verknüpfen (vgl. Berners-Lee 2006).

Tabelle 1: Übersicht über die im Rahmen des KoMoHe-Projekts aufgebauten Crosskonkordanzen zwischen TheSoz und STW.

Start-vokabular	End-vokabular	Relationen insgesamt	Äquivalenz-relationen	Ober-begriffsrel.	Unter-begriffsrel.	Verwandt-schaftsrel.	Null-relationen	Start-begriffe	End-begriffe	Begriffs-kombinationen
TheSoz	STW	7729	2977	1525	136	767	2324	7571	2563	136
STW	TheSoz	6648	4106	1500	115	427	500	5734	2928	1853

2 Vorgeschichte des Crosskonkordanz-Aufbaus bei GESIS

Thesaurus-Crosskonkordanzen werden bisher in erster Linie intellektuell erstellt. Ein prominentes Beispiel bildet das vom BMBF geförderte Projekt „Modellbildung und Heterogenitätsbehandlung (KoMoHe)“, in dessen Rahmen 25 kontrollierte Vokabulare über Crosskonkordanzen aufeinander abgebildet wurden.[4] Mit dem Ziel, ein umfassendes Netz an Crosskonkordanzen für Datenbanken im deutschsprachigen Raum zu erarbeiten und dieses um zentrale internationale kontrollierte Vokabulare zu erweitern, wurde ein wichtiges Instrument zur Heterogenitätsbehandlung in unterschiedlich erschlossenen Dokumentkollektionen geschaffen. So bildete das Projekt KoMoHe eine wichtige Voraussetzung für die integrierte Suche im gemeinsamen Wissenschaftsportal vascoda (vgl. Mayr/Petras 2008).[5]

Im Jahr 2004 wurden auch der STW und der TheSoz in diesem Projekt intellektuell aufeinander abgebildet. Gründe hierfür liegen zum einen in der deutlichen thematischen Überlappung zwischen beiden Thesauri. So bilden die Wirtschaftswissenschaften beispielsweise einen wichtigen Scope-Bestandteil der GESIS-Projektdatenbank SOFIS (Sozialwissenschaftliches Forschungsinformationssystem). Zum anderen enthält auch der STW explizit sozialwissenschaftliche Begriffe.

Um einen Eindruck von dem Aufwand der intellektuellen Crosskonkordanz-Erstellung zu bekommen, ist ein genauerer Blick auf die Datengrundlage hilfreich.

Der TheSoz ist das zentrale inhaltliche Erschließungsinstrument in den deutschsprachigen Sozialwissenschaften für Forschungsliteratur und -projektdaten. Über die GESIS-Datenbanken SOLIS (Sozialwissenschaftliches Literaturinformationssystem) und SOFIS (Sozialwissenschaftliches Forschungsinformationssystem) bzw. SOFISwiki ist er darüber hinaus das zentrale Retrieval-Instrument für die Recherche im sozialwissenschaftlichen Fachportal sowiport. So bestehen neben den unterschiedlichen Fachthesauri der CSA-Datenbanken (vgl. exemplarisch Fußnote 4) und der GND unter anderem Crosskonkordanzen zum Schlagwortvokabular der Friedrich-Ebert-Stiftung, die einen wichtigen Datenzulieferer in sowiport darstellt. Der TheSoz besteht aktuell (Stand 2012) aus über 8.000 Deskriptoren sowie über 5.000 Synonymverweisen, den sog. Nicht-Deskriptoren, und deckt damit entsprechend des weit gefassten sozialwissenschaftlichen Scopes eine große Bandbreite an Unterdisziplinen der Sozialwissenschaften ab, etwa die Politik- und die Erziehungswissenschaft.

Der STW ist hingegen das zentrale Erschließungsinstrument für die Wirtschaftswissenschaften. Insbesondere im Fachportal EconBiz und dem Bibliothekskatalog der ZBW Econis verwendet, besteht er aktuell aus etwa 6.000 Deskriptoren und 19.000 Nicht-Deskriptoren. Deutlich erkennbar ist das im Vergleich zum TheSoz sehr unterschiedliche Verhältnis zwischen Deskriptoren und Nicht-Deskriptoren.

Auf dieser Grundlage wurden im genannten KoMoHe-Projekt bilateral um die 7.000 Relationen zwischen den beiden Thesauri aufgebaut (vgl. Tab. 1). Für eine maschinell unterstützte Aktualisierung der Crosskonkordanzen zwischen beiden Thesauri, so wurde angenommen, ließe sich die große Anzahl bereits intellektuell aufgebauter Konkordanzen sehr gut nachnutzen (siehe hierzu weiter unten).

4 Bei diesem von GESIS in den Jahren 2004 bis 2007 durchgeführten Projekt handelte es sich um ein Teilprojekt des sog. „Kompetenznetzwerks Neue Dienste, Standardisierung, Metadaten“ (vgl. http://www.gesis.org/forschung/drittmittelprojekte/archiv/komohe/). Zu den aufeinander abgebildeten Thesauri zählten etwa der *Thesaurus of Sociological Indexing Terms* sowie der *Thesaurus of Political Science Indexing Terms*. Die erstellten Crosskonkordanzen zum TheSoz sind für die Einbindung der Cambridge Scientific Abstracts-Datenbanken (CSA) in sowiport von hoher Relevanz.

5 Zusammen mit den Vorgängerprojekten infoconnex und CARMEN (vgl. exemplarisch Schott/Schröder 2004) sind insgesamt 29 bilaterale sowie sechs unidirektionale Crosskonkordanzen entstanden. Daraus gingen über 500.000 Crosskonkordanzrelationen hervor (vgl. http://www.gesis.org/forschung/drittmittelprojekte/archiv/komohe/).

3 Der Library Track der Ontology Alignment Evaluation Initiative (OAEI) 2012

Im Rahmen der *Ontology Alignment Evaluation Initiative* (OAEI)[6], einer seit dem Jahr 2004 bestehenden internationalen Kampagne, deren Ziel es ist, Verfahren und Methoden zum Abgleich von Wissensorganisationssystemen, wie Ontologien, zu evaluieren, fand Ende des Jahres 2012 zum wiederholten Mal ein sog. Library Track statt. Wie zuvor in den Jahren 2007 bis 2009 ging es dabei darum, Korrespondenzen zwischen unterschiedlichen kontrollierten Vokabularen zu erkennen, um Thesauri perspektivisch noch stärker für die semantische Beschreibung von Daten im Web nutzen zu können.

Mit dem TheSoz und dem STW wurden Forschern und Entwicklern sog. Ontology Matching Tools in diesem jüngsten Library Track erneut Wissensorganisationssysteme aus dem Bibliotheks- bzw. Fachinformationswesen zur Verfügung gestellt. Konkretes Ziel des Library Tracks war es, möglichst viele korrekte Korrespondenzen zwischen beiden Thesauri automatisiert zu erkennen. 14 verschiedene Matching Tools bildeten schließlich erfolgreich beide Vokabulare mit unterschiedlichen Ergebnissen aufeinander ab.

Als wichtige Vorarbeit, um den Thesaurusabgleich zu realisieren, musste das SKOS-Format, in dem beide Thesauri vorliegen, zunächst in das Ontologie-Standardformat OWL (Web Ontology Language) transformiert werden. Den Hintergrund bildet der Umstand, dass Ontology-Matching Verfahren bisher ausschließlich auf dieser Formatgrundlage operieren. Dabei gilt es vor allem, Unterschiede in der Ausdrucksstärke zwischen SKOS und OWL zu beachten. Zwar verfügen Thesauri nicht über eine derartige Fülle an Begriffsbeziehungen, wie sie für Ontologien typisch sind. Vielmehr zeichnen sie sich in der Regel ausschließlich durch Synonym- und Ähnlichkeits- sowie Ober- und Unterbegriffsbeziehungen aus. Dennoch bietet SKOS aufgrund seines speziellen Fokus auf Thesauri umfangreiche und komplexe Beziehungen zwischen Termen an, wie sie im generischeren OWL-Format nicht darstellbar sind. So lässt sich in SKOS bspw. zwischen Ähnlichkeits- sowie Ober- und Unterbegriffsbeziehungen (skos:broader, skos:narrower, skos:related) explizit unterscheiden, in OWL können diese Beziehungen allerdings nicht derart präzise modelliert werden. Daneben kann in OWL im Gegensatz zum SKOS-Format die Unterscheidung zwischen Deskriptor und Nicht-Deskriptor nicht adäquat abgebildet werden. Während für die Darstellung der Synonymverweise in SKOS skos:altLabel gewählt wird, müssen diese in OWL ebenso wie der eigentliche Deskriptor als rdfs:label modelliert werden. Wie in Tabelle 2 zur Überführung des SKOS-Formats in OWL veranschaulicht, geht die Synonymbeziehung zwischen den Begriffen Unternehmen und Firma bei der Transformation nach OWL verloren.

Tabelle 2: Exemplarische Abbildung von SKOS in OWL.

SKOS	OWL
skos:Concept z.B. Unternehmen	**owl:Class** z.B. Unternehmen
skos:prefLabel Unternehmen **skos:altLabel** Firma	**rdfs:label** Unternehmen **rdfs:label** Firma
skos:scopeNote s.a. Betrieb... oder Werk... **skos:notation** 4.6.06	**rdfs:comment** s.a. Betrieb... oder Werk... **rdfs:comment** 4.6.06
A skos:narrower B Dienstleistungsunternehmen	**A rdfs:subClassOf B** Dienstleistungsunternehmen
A skos:broader B Betriebswirtschaft	**B rdfs:subClassOf A** Betriebswirtschaft
skos:related Betrieb	**rdfs:seeAlso** Betrieb

Ebenso gilt es darauf hinzuweisen, dass von den Matching Tools ausschließlich Äquivalenzbeziehungen zwischen Begriffen der beiden Thesauri aufgebaut wurden. Hierzu wurde vor allem das Umfeld der Begriffe, etwa Unter- und Oberbegriffsbeziehungen, sowie syntaktische Ähnlichkeiten berücksichtigt.[7] Daneben wurde von den Matching Tools, anders als bei einem intellektuellen Crosskonkordanz-Aufbau, keine Unterscheidung nach Richtung des Thesaurusabgleichs getroffen (TheSoz > STW vs. STW > TheSoz). Schließlich stellen viele Matching Tools außerdem nur 1:1-Beziehungen zwischen den Thesauri her, d. h., wenn ein System zu einem bestimmten Term eine mögliche Korrespondenz identifiziert hat, wird dieser Term für weitere mögliche Korrespondenzen nicht mehr berücksichtigt.[8]

6 Vgl. http://oaei.ontologymatching.org/

7 Generell existieren Ontology Matching Tools, die auch externe Quellen wie Wörterbücher oder Lexika berücksichtigen können (vgl. Euzenat et al. 2007). Unter den teilnehmenden Tools befand sich lediglich eines, das zu einzelnen Termen Wikipedia-Einträge identifizierte, um diese zum Abgleich zu nutzen.

8 Dieses Vorgehen liegt darin begründet, dass Klassen in Ontologien typischerweise sehr präzise und klar voneinander abgegrenzt definiert werden, weswegen beim Matching von Ontologien von eindeutigen Zuordnungen ausgegangen wird. Bei Thesauri ist dies durch die mitunter zahlreichen Synonymverweise und Ähnlichkeitsbeziehungen eher selten der Fall.

4 Bewertung der automatisch aufgebauten Thesaurus-Crosskonkordanzen

Die von den Matching Tools automatisch generierten Korrespondenzen zwischen beiden Thesauri wurden zum einen einer maschinellen und zum anderen einer intellektuellen Evaluation unterzogen.

Die maschinelle Bewertung erfolgte in Form eines sog. Reference Alignment. Hierzu wurden die von den Matching Tools anlässlich des Library Tracks erstellten Crosskonkordanzen zwischen TheSoz und STW mit den bereits im Rahmen des KoMoHe-Projekts erstellten Crosskonkordanzen verglichen. Mittels dieser Referenz konnten für die Matching Tools die zentralen Evaluationswerte Precision, Recall und F-Measure berechnet und die Ergebnisse der einzelnen Tools miteinander vergleichen werden.

Die Ergebnisse der verschiedenen Matching Tools fallen sehr heterogen aus (vgl. Tab. 3 sowie ausführlich Aguirre et al. 2012). Während die meisten Tools in der Qualität der Ergebnisse mit einem F-Measure zwischen 0.608 und 0.674 relativ dicht beieinander liegen, weisen nur wenige Tools Ergebnisse auf, die deutlich darunter liegen. Besonders auffällig sind die Unterschiede in Bezug auf die Zeitspannen (zwischen 21 und 144.070 Sekunden), die von den Tools benötigt wurden, um Crosskonkordanzen zu identifizieren, was auf die unterschiedlichen von den Tools verwendeten Algorithmen zurückzuführen ist. Insgesamt lässt sich feststellen, dass sich der Verlust der Ausdrucksstärke bei der Übertragung von spezifischen Thesaurusbeziehungen von SKOS nach OWL auch in den Ergebnissen der Matching Tools widerspiegelt. Dadurch, dass bspw. keine explizite Unterscheidung zwischen Synonymen besteht, entstehen sehr viele nicht korrekte Korrespondenzen, die auf sehr weit gefassten Synonymen basieren.

Als zweite Evaluation wurde eine intellektuelle Bewertung der gefundenen Korrespondenzen durchgeführt. Dabei wurden ausschließlich Beziehungen zwischen jenen Begriffen berücksichtigt, die bisher noch keine Relationen zu Begriffen des anderen Thesaurus aufwiesen. Im TheSoz betraf dies ca. 600 neue Deskriptoren seit dem Jahr 2004. Im STW lag die Zahl neu hinzugekommener Deskriptoren leicht darüber.

Die im Verlauf der Evaluation aufgetretenen Besonderheiten wurden nach Typen unterschieden. Bei den gelungenen Äquivalenzrelationen (vgl. Tab. 4, Zeilen 1 bis 5) zeigt sich, dass der Aufbau von Crosskonkordanzen sehr gut bei Begriffen gelingt, die auf der String-Ebene identisch sind (vgl. Levenshtein 1966 zu dem Begriff der Levenshtein-Distanz). Weisen beide Begriffe denselben Scope auf, lässt sich eine derartige Beziehung als korrekt aufgebaute Äquivalenzbeziehung werten (vgl. Tab. 4, Zeilen 1 und 2). Besonders gelungen erscheint aus diesem Grund der Aufbau von Äquivalenzrelationen zwischen Geographika und Ethnographika. In beiden Thesauri werden hierfür häufig dieselben Begriffe verwendet. Als besonders positiv sticht ein gelungener Aufbau von Äquivalenzrelationen dann heraus, wenn beide Begriffe auf String-Ebene nicht komplett identisch sind (vgl. Tab. 4, Zeilen 3 bis 5). In allen drei Beispielen sorgen die Synonymverweise bzw. Nicht-Deskriptoren dafür, dass eine gelungene Äquivalenzrelation aufgebaut wurde.

Tabelle 3: Ergebnisse der verschiedenen Ontology Matching Tools.

System	Precision	Recall	F-Measure	Zeit (sec.)	Anzahl CK	1:1
Tool 1	0.537	0.906	0.674	804	4712	
Tool 2	0.654	0.687	0.670	45	2938	
Tool 3	0.688	0.644	0.665	95	2620	
Tool 4	0.717	0.619	0.665	44	2413	ja
Tool 5	0.595	0.750	0.664	496	3522	
Tool 6	0.577	0.776	0.662	21	3756	
Tool 7	0.675	0.645	0.660	32773	2671	
Tool 8	0.465	0.925	0.619	14363	5559	
Tool 9	0.612	0.607	0.609	144070	2774	Ja
Tool 10	0.645	0.575	0.608	14494	2494	Ja
Tool 11	0.434	0.481	0.456	39869	3100	Ja
Tool 12	0.520	0.184	0.272	2171	989	ja
Tool 13	0.107	0.652	0.184	1096	17001	
Tool 14	0.321	0.072	0.117	37457	624	

Tabelle 4: Beispieldatensätze für die beschriebenen Fehlertypen.

	TheSoz	Hierarchieebene (Ober- und Unterbegriffe)	STW	Hierarchieebene (Ober- und Unterbegriffe)	Rel.-typ	Konf.-wert	Eval. (j/n)
1	**Abwasser**	OB: Ökologie, Umweltschutz, u.a.	**Abwasser**	OB: Wasser, Umweltbelastung, Abfall, u.a.	ER	1.0	j
2	**Afrikaner**	OB: Länder, Regionen, Nationalitäten	**Afrikaner**	UB: Senegalesen, Tunesier, u.a. OB: Völker und Ethnien	ER	1.0	j
3	**Abfallbeseitigung**	OB: Ökologie, Umweltschutz, u.a.	**Abfallentsorgung** (UF: Abfallbeseitigung, u.a.)	UB: Nukleare Entsorgung OB: Abfallwirtschaft und Recycling	ER	1.0	j
4	**Informationspflicht** (UF: Auskunftspflicht)	OB: Rechtsgrundlagen, Rechtsnormen	**Auskunftspflicht**	UB: Publizitätspflicht OB: Verwaltungsrecht	ER	1.0	j
5	**Exportwirtschaft**	OB: Wirtschaftszweige, -sektoren, -bereiche	**Außenhandels-wirtschaft** (UF: Exportwirtschaft)	UB: Außenhandelsvertretung, Exporthandel, u.a. OB: Außenhandelswirtschaft	ER	1.0	j
6	**Entwicklung**	UB: regionale Entwicklung, kulturelle Entwicklung, Schulentwicklung, u.a. OB: Allgemeinbegriffe	**Entwicklung**	UB: Gemeindeentwicklung, regionale Entwicklung, u.a. OB: Entwicklungsökonomik	ER	1.0	n
7	**Prävention** (SN: nicht im medizin. Sinne)	OB: Kriminalität, soziale Hilfen und Maßnahmen, soziale Leistungen, u.a.	**Prävention**	UB: Gesundheitsvorsorge OB: Allgemeinwörter	ER	1.0	n
8	**Akzeleration**	OB: Individuum, Persönlichkeit, Lebensalter, u.a.	**Akzelerator**	OB: Makroökonomik	ER	1.0	n
9	**Familieneinkommen**	OB: Einkommen, Wirtschaftsentwicklung, Wirtschaftsstatistik, natürliche Ressourcen	**Haushaltseinkommen** (UF: privates Einkommen, Familieneinkommen)	UB: Verfügbares Einkommen, ländliches Einkommen, u.a. OB: Private Haushalte, Vermögen, u.a.	ER	1.0	n
10	**Sekte**	*Keine hierarchische Einordnung vorhanden*	**Schaumwein** (UF: Sekt, Champagner)	OB: Wein	ER	1.0	n

Daneben lassen sich auch weniger gelungene Begriffsrelationen (vgl. Tab. 4, Zeilen 6 bis 10) nach Typen unterscheiden. So kann die unterschiedliche Reichweite von Begriffen, die auf der Zeichenebene identisch sind, ein besonderes Problem für einen automatischen Abgleich darstellen (vgl. Tab. 4, Zeile 6, der STW-Begriff „Entwicklung“ wäre in diesem Fall, ersichtlich an den aufgeführten Unterbegriffen, ein Unterbegriff zum TheSoz-Begriff.). Ebenso von besonderer Schwierigkeit ist der Fall, in dem die Begriffe auf der Zeichenebene erneut identisch sind, allerdings in

Tabelle 5: Übersicht über die Evaluationsergebnisse der angetretenen Ontology Matching Tools.

System	Äquivalenzrelationen insgesamt	Korrekte Äquivalenzrelationen	Nicht korrekte Äquivalenzrelationen
Tool 1	3500	215 (6,1 %)	3285
Tool 2	628	162 (25,8 %)	466
Tool 3	631	213 (33,8 %)	418
Tool 4	682	246 (36,1 %)	436
Tool 5	828	269 (32,5 %)	556
Tool 6	448	194 (43,3 %)	254
Tool 7	540	234 (43,3 %)	306
Tool 8	403	203 (50,4 %)	200
Tool 9	175	64 (36,6 %)	111
Tool 10	165	38 (23,0 %)	127
Tool 11	525	252 (48,0 %)	273
Tool 12	433	232 (53,8 %)	201
Tool 13	682	225 (33,0 %)	457
Tool 14	613	248 (40,5 %)	365

der Scope Note eines dieser Begriffe explizit eine Begriffsverwendung, wie sie im anderen Thesaurus vorgesehen ist, ausgeschlossen wird (vgl. Tab. 4, Zeile 7, in der Scope Note des TheSoz wird ausdrücklich darauf hingewiesen, dass der Begriff Prävention nicht im medizinischen Sinne gebraucht wird, hierfür wird der Begriff Prophylaxe verwendet). Dass von den Matching-Tools nicht die semantische Bedeutungsebene der Begriffe berücksichtigt wird, zeigt sich auch an einem weiteren Fehlertypus. So werden etwa inkorrekte Äquivalenzbeziehungen für String basiert sehr ähnliche Begriffe aufgebaut, auch wenn diese Begriffe sich in ihrer fachspezifischen Verwendung sehr stark voneinander unterscheiden (vgl. Tab. 4, Zeile 8).[9] Eine derart unterschiedliche Verwendung ist für die Softwaretools ebenfalls nicht erkennbar. Schließlich kann es auch zu inkorrekten Äquivalenzbeziehungen kommen, wenn Synonymverweise identische oder nahezu identische Begriffe aufweisen (vgl. Tab. 4, Zeilen 9 und 10, während Haushaltseinkommen zum TheSoz-Begriff Familieneinkommen einen Oberbegriff darstellt, weist Schaumwein im STW bis auf die Buchstabenfolge des Synonymverweises Sekt keinerlei Ähnlichkeit mit dem TheSoz-Begriff Sekte auf).

Die Gesamtergebnisse der Matching-Tools fallen für die intellektuelle Bewertung der neu aufgebauten Thesaurus-Crosskonkordanzen sehr unterschiedlich aus. Die Zahlen der als korrekt bewerteten neuen Äquivalenzrelationen bewegen sich zwischen knapp 40 bis 270 bzw. anteilig zwischen sechs und knapp 54 Prozent.

5 Fazit

Als Fazit der bisherigen Evaluation lässt sich festhalten, dass automatische Verfahren einerseits eine sehr schnelle Vorselektion beim Aufbau von Thesaurus-Crosskonkordanzen bieten. Durch den in erster Linie String basierten Abgleich werden korrekte Äquivalenzbeziehungen vor allem zwischen jenen Begriffen gefunden, die auf der Zeichenebene identisch sind. Einige Matching Tools lieferten hier bereits eine vielversprechende Menge an korrekten Ergebnissen mit einer Precision von bis zu 0,717. Andererseits ist eine Reihe von Auffälligkeiten beim automatischen Aufbau von Konkordanzen als sehr kritisch zu bewerten. Bis auf den Abgleich von Ethnographika und Geographika kann man davon ausgehen, dass der automatische Aufbau von Crosskonkordanzen unvollständig ist, was in besonderer Weise darin begründet liegt, dass einige Systeme nur 1:1-Relationen erkennen. So ist bei automatischen Verfahren keineswegs gesichert, dass sämtliche möglichen Äquivalenzrelationen ermittelt werden. Daneben bleiben die semantische Begriffsebene ebenso wie die thesaurusspezifische Reichweite von Begriffen unberücksichtigt. Des Weiteren wurde der Aufbau anderer Begriffsbeziehungen, wie Unter- und Oberbegriffsrelationen sowie Verwandtschaftsbeziehungen, ausgespart. Und schließlich sind die Verfahren richtungsblind – es wird kein Unterschied gemacht, ob ein Begriff vom TheSoz auf den STW abgebildet wird oder umgekehrt – sowie nicht in der Lage, zusammengesetzte Begriffe als Äquivalenzbegriffe aufzubauen. Es zeigt sich somit, dass beim derzeitigen Stand der Matching-Verfahren eine intellektuelle Einschätzung durch einen Domänenexperten unabdingbar ist.

Dennoch hat die Evaluation gezeigt, dass die erzielten Ergebnisse durchaus eine sinnvolle Vorauswahl an möglichen Crosskonkordanzen liefern können, was gerade bei umfangreichen Thesauri eine solide Basis für den manuellen Aufbau von Crosskonkordanzen bilden kann. Ziel weiterer Forschung wird es daher sein, eine Erhöhung des Konfidenzgrads der maschinellen Selektionsleistung zu erreichen. Hierzu sollen die Leistungen der unterschiedlichen Matching Tools auf der Ebene der einzelnen aufgebauten Crosskonkordanzen miteinander verglichen werden. So ist über die Ermittlung der Übereinstimmungen zwischen den unterschiedlichen Tools etwa die Bildung eines Schwellenwerts denkbar, der Aussagen über die Konfidenz der maschinell aufgebauten Thesaurus-Crosskonkordanzen zulässt.

9 Während Akzeleration in der Soziologie für die Entwicklungsbeschleunigung bei Jugendlichen steht, meint Akzelerator in den Wirtschaftswissenschaften eine bestimme Kennziffer, die ausdrückt, in welchem Maße eine bestimmte Veränderung der gesamtwirtschaftlichen Nachfrage zu einem bestimmten Investitionsvolumen führt.

Literatur

Aguirre, José Luis; Eckert, Kai; Euzenat, Jérôme; Ferrara, Alfio; Hage, Willem Robert van; Hollink, Laura; Meilicke, Christian; Nikolov, Andriy; Ritze, Dominique; Scharffe, François; Shvaiko, Pavel; Šváb-Zamazal, Ondřej; Trojahn, Cássia; Jiménez-Ruiz, Ernesto; Cuenca Grau, Bernardo; Zapilko, Benjamin (2012): Results of the ontology alignment evaluation initiative 2012. In: Shvaiko, Pavel; Euzenat, Jérôme; Kementsietsidis, Anastasios; Mao, Ming; Noy, Natasha; Stuckenschmidt, Heiner (Hrsg.): Proceedings of the 7th International Workshop on Ontology Matching (OM-2012) collocated with the 11th International Semantic Web Conference (ISWC-2012). CEUR Workshop Proceedings, Vol-946.

Berners-Lee, Tim (2006): Linked Data. http://www.w3.org/DesignIssues/LinkedData.html [16.12.2012].

Euzenat, Jérôme; Shvaiko, Pavel (2007): Ontology Matching, Berlin u. a.: Springer.

Gerards, Michael (2011): Semiautomatische Erschließung von Psychologie-Information. Präsentation im Rahmen des PETRUS-Workshops an der Deutschen Nationalbibliothek Frankfurt/Main, 21./22.03.2011 http://www.dnb.de/SharedDocs/Downloads/DE/DNB/wir/petrus/gerardsSemiautomatischeErschliessungZpid.pdf?__blob=publicationFile [07.06.2013].

Kempf, Andreas Oskar (2012): Automatische Indexierung in der sozialwissenschaftlichen Fachinformation. Eine Evaluationsstudie zur maschinellen Erschließung für die Datenbank SOLIS. Berliner Handreichungen zur Bibliotheks- und Informationswissenschaft, 329, Berlin: Institut für Bibliotheks- und Informationswissenschaft der Humboldt-Universität zu Berlin.

Levenshtein, Vladimir I. (1966): Binary codes capable of correcting deletions, insertions, and reversals. In: Soviet Physics Doklady, 10, 707–710.

Mayr, Philipp; Zapilko, Benjamin; Sure, York (2010): Ein Mehr-Thesauri-Szenario auf Basis von SKOS und Crosskonkordanzen. In: Ockenfeld, Marlies (Hrsg.): Recherche im Google-Zeitalter – vollständig und präzise?!: die Notwendigkeit von Informationskompetenz; Tagungsband/25. Oberhofer Kolloquium zur Praxis der Informationsvermittlung, Barleben/Magdeburg, Bd. 13, Frankfurt am Main: DGI, 163–172.

Mayr, Philipp; Petras, Vivien (2008): Cross-concordances – terminology mapping and its effectiveness for Information retrieval: Crosskonkordanzen – Terminologie Mapping und deren Effektivität für das Information Retrieval. In: World Library and Information Congress: 64th IFLA General Conference and Meeting, Québec http://archive.ifla.org/IV/ifla74/papers/129-Mayr_Petras-en.pdf [07.06.2013].

Neubert, Joachim (2009) Bringing the Thesaurus for Economics on to the Web of Linked Data. In: Proc. WWW Workshop on Linked Data on the Web, http://events.linkeddata.org/ldow2009/papers/ldow2009_paper7.pdf [07.06.2013].

Schott, Hannelore; Schroeder, Albert (2004) Crosskonkordanzen von Thesauri und Klassifikationen. In: Budin, G.;Ohly, Hans-Peter (Hrsg.):Wissensorganisation in kooperativen Lern- und Arbeitsumgebungen. Würzburg: Ergon-Verlag, 41–49.

Wissel, Verena (2011): Erfahrungsbericht und Schlussfolgerungen des DIPF zur Erprobung automatischer Erschließung. Präsentation im Rahmen des PETRUS-Workshops an der Deutschen Nationalbibliothek Frankfurt/Main, 21./22.03.2011 http://www.dnb.de/SharedDocs/Downloads/DE/DNB/wir/petrus/wisselDipfErfahrungsbericht.pdf?__blob=publicationFile [07.06.2013].

Zapilko, Benjamin; Schaible, Johann; Mayr, Philipp; Mathiak, Brigitte (2012): TheSoz: a SKOS representation of the thesaurus for the social sciences. In: Semantic Web: interoperabilty, usability, applicability. IOS Press.

http://www.gesis.org/forschung/drittmittelprojekte/archiv/komohe/ [18.12.2012].

http://oaei.ontologymatching.org/ [15.12.2012].

Dr. Andreas Oskar Kempf
GESIS – Leibniz-Institut für Sozialwissenschaften
Fachinformation Sozialwissenschaften
Unter Sachsenhausen 6–8
50667 Köln
andreas.kempf@gesis.org
www.gesis.org

Dr. Andreas Oskar Kempf ist wissenschaftlicher Mitarbeiter bei GESIS – Leibniz-Institut für Sozialwissenschaften. Verantwortlich für die Weiterentwicklung von Thesaurus und Klassifikation Sozialwissenschaften zählen zu seinen Forschungsgebieten die Stärkung ihrer Interoperabilität sowie die Einbindung der Erschließungsinstrumente in Anwendungen des Semantic Web.

Benjamin Zapilko
GESIS – Leibniz-Institut für Sozialwissenschaften
Wissenstechnologien für Sozialwissenschaften
Unter Sachsenhausen 6–8
50667 Köln
benjamin.zapilko@gesis.org
www.gesis.org

Benjamin Zapilko ist wissenschaftlicher Mitarbeiter bei GESIS – Leibniz-Institut für Sozialwissenschaften. Seine Forschungsinteressen liegen im Bereich Semantic Web und Linked Open Data mit dem Fokus auf der Integration und Verlinkung heterogener Datenquellen. In seiner Dissertation beschäftigt er sich mit Verfahren und Anwendungsszenarien, um Linked Open Data für die sozialwissenschaftliche Forschung nutzbar zu machen.