Ein Mehr-Thesauri-Szenario auf Basis von SKOS und Crosskonkordanzen

Philipp Mayr, Benjamin Zapilko, York Sure

Abstract
Im August 2009 wurde SKOS „Simple Knowledge Organization System“ als neuer Standard für web-basierte kontrollierte Vokabulare durch das W3C veröffentlicht[1]. SKOS dient als Datenmodell, um kontrollierte Vokabulare über das Web anzubieten sowie technisch und semantisch interoperabel zu machen. Perspektivisch kann die heterogene Landschaft der Erschließungsvokabulare über SKOS vereinheitlicht und vor allem die Inhalte der klassischen Datenbanken (Bereich Fachinformation) für Anwendungen des Semantic Web, beispielsweise als Linked Open Data[2] (LOD), zugänglich und stärker miteinander vernetzt werden. Vokabulare im SKOS-Format können dabei eine relevante Funktion einnehmen, indem sie als standardisiertes Brückenvokabular dienen und semantische Verlinkung zwischen erschlossenen, veröffentlichten Daten herstellen.
Die folgende Fallstudie skizziert ein Szenario mit drei thematisch verwandten Thesauri, die ins SKOS-Format übertragen und inhaltlich über Crosskonkordanzen aus dem Projekt KoMoHe verbunden werden. Die Mapping Properties von SKOS bieten dazu standardisierte Relationen, die denen der Crosskonkordanzen entsprechen. Die beteiligten Thesauri der Fallstudie sind a) TheSoz (Thesaurus Sozialwissenschaften, GESIS), b) STW (Standard-Thesaurus Wirtschaft, ZBW) und c) IBLK-Thesaurus (SWP).

1 Thesauri und Crosskonkordanzen

Thesauri sind wichtige Retrieval-Instrumente zur Suche in größeren Dokumentbeständen, i. d. R. Literaturnachweissystemen. Thesauri werden häufig exklusiv für eine Kollektion entworfen und decken damit hauptsächlich ein Wissensgebiet möglichst umfassend ab. Inzwischen werden zunehmend andere Dokumenttypen, z. B. Forschungsdaten, in die Suche mit einbezogen. Teilweise sind diese Dokumenttypen auch über Thesauri oder andere kontrollierte Vokabulare erschlossen (Beispiel cessda-Portal[3] und ELSST-Thesaurus). Um eine kollektionsübergreifende, semantisch angereicherte Suche zu realisieren, kann das Verbinden (Mapping) von involvierten Thesauri als ein Lösungsansatz angesehen werden (vgl. Krause, 2008). Im Projekt KoMoHe[4] (Mayr & Petras, 2008) wurde damit begonnen, prototypisch für den Anwendungsfall vascoda Thesauri miteinander zu verbinden und diese Verbindungen (sog. Crosskonkordanzen) für das Information Retrieval zu nutzen. Diese Crosskonkordanzen liegen momentan in einer relationalen Datenbank bei der GESIS vor und sind über einen Webservice zugänglich.
Crosskonkordanzen werden im Projekt definiert als intellektuell und manuell erstellte Verbindungen, die Äquivalenz, Hierarchie und Verwandtschaft zwischen Termen zweier kontrollierter Vokabulare über Relationen bestimmen. Typischerweise werden die Vokabulare bilateral verbunden, d. h. eine Konkordanz verbindet Terme eines Vokabulars A zu einem Vokabular B und eine weitere Konkordanz verbindet Terme von Vokabular B zurück zu A. Bilaterale Relationen sind dabei nicht notwendigerweise symmetrisch. Beispielsweise wird der Term ‚Computer’ aus Vokabular A auf den Term ‚Information System’ in Vokabular B abgebildet, aber der gleiche Term ‚Information System’ in Vokabular B

[1] http://www.w3.org/TR/skos-reference/
[2] Informationen über Linked Open Data finden sich unter: http://linkeddata.org
[3] http://www.cessda.org/accessing/search/

[4] http://www.gesis.org/forschung-lehre/programme-projekte/informationswissenschaften/projektuebersicht/komohe/

kann mit einem anderen Term, z. B. ‚Data base', in Vokabular A relationiert werden. Die Crosskonkordanzen im Projekt KoMoHe involvieren die gesamten oder umfangreiche Teile der Vokabulare (siehe Übersicht in Abb. 1). Bevor das Mapping der Terme beginnt, werden die Vokabulare bezüglich thematischen und syntaktischen Überlappungen untersucht. Die Crosskonkordanzen werden von Wissenschaftlern oder Terminologie-Experten innerhalb einer Domäne erstellt. Es ist essentiell für ein erfolgreiches Mapping, dass die Bedeutung und Semantik der Terme und der internen Relationen der beteiligten Vokabulare vollständig verstanden werden. Insbesondere das Verstehen und semantisch korrekte Relationieren der Terme unterscheidet die intellektuellen Verfahren von den rein automatischen Ansätzen. Der Mapping-Prozess basiert auf einem Set von praktischen Regeln und Richtlinien. Während des Mappings der Terme werden alle Intra-Thesaurusrelationen (Scope Notes eingeschlossen) verwendet. Recall und Precision der erstellten Relationen sollen in den entsprechenden Datenbanken geprüft werden. Diese Prüfung ist insbesondere für Kombinationen von Termen (1:n-Relationen) wichtig. 1:1-Termrelationen sollen beim Mapping bevorzugt werden. Zum Schluss wird die semantische Korrektheit der Crosskonkordanzen von Experten kontrolliert, zudem werden Stichproben empirisch auf Dokument-Recall und -Precision geprüft.

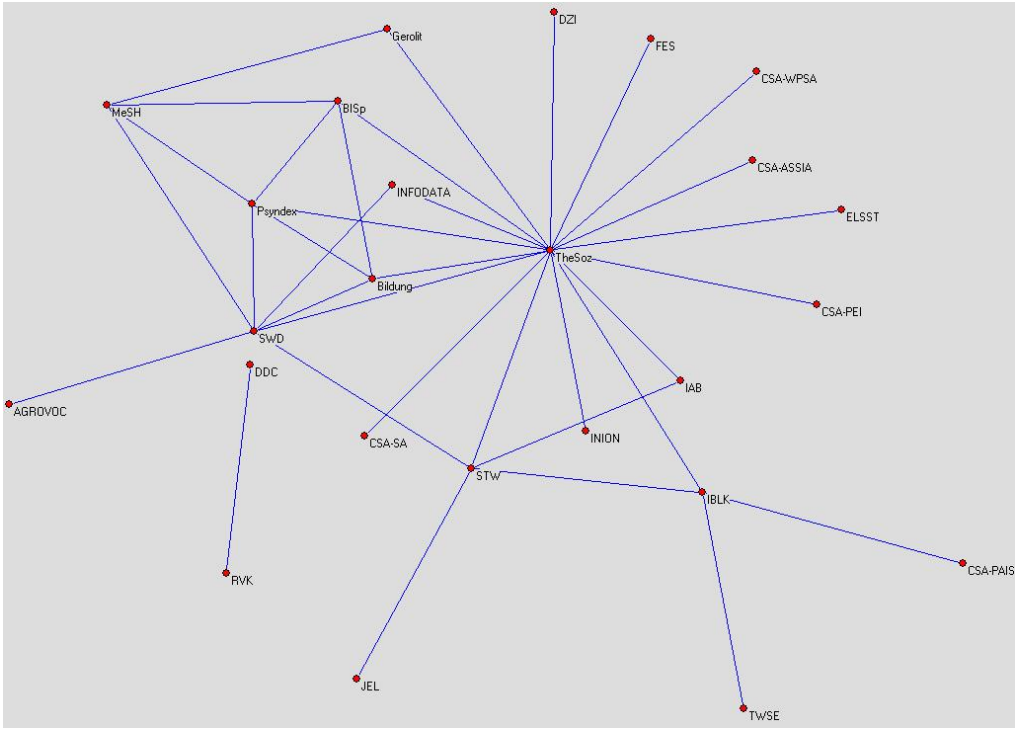

Abb. 1: Übersicht der vernetzten kontrollierten Vokabulare im Projekt KoMoHe (Auflösung der Kürzel siehe KoMoHe-Website)

2 SKOS (Simple Knowledge Organization System)

Mit SKOS (Simple Knowledge Organization System) wurde im August 2009 vom W3C ein Datenmodell als Standard deklariert, das die Veröffentlichung von Erschließungsvokabularen (Thesauri, Klassifikationen, Taxonomien etc.) in einem maschinenlesbaren Standardformat für das semantische Web möglich macht und dadurch ein Teilen und Verlinken mit anderen standardisierten Datenquellen ermöglicht.

> "The SKOS data model provides a standard, low-cost migration path for porting existing knowledge organization systems to the Semantic Web" (W3C, 2009).

Basierend auf dem grundlegenden Semantic Web-Standardformat RDF[5] (Resource Description Framework) gewährleistet SKOS eine hohe Interoperabilität in Verbindung mit anderen Anwendungen und Formaten. Durch die Klassen und Properties von SKOS können hierarchische und assoziative Relationen von Vokabularen, aber auch dokumentarische und lexikale Eigenschaften modelliert werden. Die „SKOS eXtension for Labels“ (SKOS-XL) stellt dabei eine optionale Erweiterung dar, die eine Identifizierung, Beschreibung und Verlinkung von einzelnen lexikalen Entitäten unterstützt, wie es in vielen Vokabularen benötigt wird. Verlinkungen zwischen verschiedenen SKOS-Vokabularen (vergleichbar mit Crosskonkordanzen zwischen traditionellen Erschließungsvokabularen) können über die so genannten SKOS Mapping Properties hergestellt werden.

3 Transformation von Thesauri nach SKOS

Der Transformationsprozess eines Thesaurus ins SKOS-Format lässt sich meist in drei Schritte gliedern. Assem et al. (2006) haben vor diesem Hintergrund eine strukturierte Methode eingeführt, die aus folgenden Schritten besteht: (1) Analyse der Struktur des Thesaurus, der enthaltenen Terme und der Relationen zwischen den Termen, (2) Mapping der Elemente und Relationen des Thesaurus auf äquivalente SKOS-Elemente und -Relationen und (3) technische Konvertierung des im zweiten Schritt festgelegten Mappings.

Selbst wenn ein Thesaurus den gängigen ISO-Normen[6] entspricht, zeigt sich, dass eine Übertragung nach SKOS in vielen Fällen nicht so trivial ist wie angenommen, wie viele Case Studies[7] belegen. Zwar ist SKOS, da es auf RDF basiert, erweiterbar um die Definition eigener Relationen, jedoch bleibt dabei stets zu beachten, dass es dadurch zu Inkompatibilitäten mit anderen Vokabularen und Daten kommen kann, denen die selbst definierten Relationen nicht bekannt sind oder die diese nicht bzw. nicht vollständig weiter verarbeiten können. Dieses Szenario kann schnell eintreten, wenn eigens definierte Klassen oder Properties nicht ausreichend durch gängige Formate (z. B. RDF-Schema[8]) deklariert und beschrieben und somit in einen Kontext gestellt werden können.

Die rege Aufnahme und weitere Entwicklung von SKOS lassen aber durchaus darauf hoffen, dass auch für komplexere, nicht unbedingt den ISO-Standards entsprechende Relationen Modellierungslösungen entstehen werden. Einen Schritt in diese Richtung von anderer Seite könnte die Arbeit an der „ISO-Norm 25964: Thesauri and interoperability with other vocabularies“[9] bilden.

3.1 Thesaurus-Analyse

Grundlage für jede Thesaurus-Transformation bildet die Analyse der zugrunde liegenden Strukturen des Vokabulars. Dabei sollten nicht nur assoziative und hierarchische Relationen berücksichtigt werden, sondern auch die generelle Struktur bzw. der Aufbau des Thesaurus. Für eine Repräsentation eines Thesaurus in SKOS ist höchst relevant, ob bzw. inwieweit er den gängigen ISO-Normen entspricht, die Existenz von und Beziehung zwischen Deskriptoren und Nicht-Deskriptoren sowie die Existenz von Kategorien oder Klassifikationshierarchien.

a) TheSoz (Thesaurus Sozialwissenschaften, GESIS)

Der Thesaurus Sozialwissenschaften enthält etwa 11600 Terme, von denen ca. 7750 Deskriptoren (autorisierte Schlagwörter) und ca. 3850 Nicht-Deskriptoren sind. Es lassen sich die gängigen assoziativen und hierarchischen Relationen zwischen Termen identifizieren (Oberbegriff, Unterbegriff, verwandter Begriff etc.), darunter jedoch auch komplexere Relationen wie „Use Combination“ und „Used For Combination“. Eine Sonderstellung nehmen im TheSoz die etwa 200 so genannten „Alter-

[5] http://www.w3.org/RDF/

[6] ISO-Normen für Thesauri: ISO 2788 (Guidelines for the establishment and development of monolingual thesauri) und ISO 5964 (Guidelines for the establishment and development of multilingual thesauri)

[7] Use Cases zu SKOS finden sich u. a. unter: http://www.w3.org/TR/skos-ucr

[8] http://www.w3.org/TR/rdf-schema/

[9] Weitere Informationen unter: http://www.niso.org/workrooms/iso25964

nativen Nicht-Deskriptoren“ (AD) ein, die Ambiguitäten in den Termrelationen ausdrücken, d. h. ein alternativer Nicht-Deskriptor verfügt z. B. über mehrere, gleichwertige Relationen vom Typ „Use“ oder „Use Combination“ zu verschiedenen Deskriptoren.

> Beispiel: Der alternative Nicht-Deskriptor „Erhebung“ enthält die Relationen „Use Datengewinnung“, „Use Revolution“ sowie „Use Widerstand“, die alle gleichwertig zu behandeln sind, und behandelt dadurch die Mehrdeutigkeit des Begriffes „Erhebung“.

Zusätzlich existiert im TheSoz eine systematische Klassifikationshierarchie. Jeder Term des Thesaurus ist dabei einem oder mehreren Klassifikationstermen zugeordnet.

b) STW (Standard-Thesaurus Wirtschaft, ZBW)
Der Standard-Thesaurus Wirtschaft enthält etwa 5800 Deskriptoren und 17000 Nicht-Deskriptoren zu allen ökonomischen Themen sowie aus benachbarten Disziplinen wie Politik, Recht und Soziologie (vgl. Neubert, 2009). Die Deskriptoren sind darüber hinaus in einer Taxonomie mit etwa 500 Einträgen verortet.

c) IBLK-Thesaurus (Europäischer Thesaurus Internationale Beziehungen und Länderkunde, SWP)
Der Europäische Thesaurus Internationale Beziehungen und Länderkunde[10] enthält etwa 8200 Deskriptoren, darunter ca. 600 Eigennamen von bedeutenden internationalen Institutionen und internationalen Abkommen. Der IBLK-Thesaurus enthält schwerpunktmäßig Deskriptoren zu internationalen und regionalwissenschaftlichen Themen. Die Deskriptoren sind darüber hinaus in 24 Themenfeldern eingeordnet.

3.2 Mapping der Thesaurus-Strukturen auf SKOS-Klassen und -Properties

Für viele Elemente eines Thesaurus lassen sich schnell die adäquaten SKOS-Elemente identifizieren („skos:narrower“, „skos:broader“ etc.). Eine grundlegende Besonderheit von SKOS, die allerdings vor jedem Mapping beachtet werden sollte, ist, dass Vokabulare in SKOS konzept-basierend dargestellt werden, d. h. es gibt Konzepte („skos:Concept“) mit jeweils mehreren Labels, wobei pro Konzept ein präferiertes Label („skos:prefLabel“) existiert und daneben weitere alternative oder versteckte Labels („skos:altLabel“ und „skos:hiddenLabel“). Die meisten Thesauri sind jedoch eher term-basierend aufgebaut und verfügen über Relationen zwischen einzelnen Termen, wobei auch präferierte und nicht-präferierte Terme beteiligt sein können. Dieser grundsätzliche Unterschied stellt oft eine große Hürde dar, wenn Vokabulare nach SKOS transformiert werden sollen, die über eine große Anzahl an Relationen zwischen präferierten und nicht-präferierten Termen für dasselbe Begriffskonzept verfügen, da dort aus dem Blickwinkel von SKOS meist viele Terme in einem Konzept zusammengefasst werden können. Die folgende Abbildung (Abb. 2) skizziert die konzept-basierte Repräsentation in SKOS.

[10] http://www.fiv-iblk.de/ip/thesaurus.htm

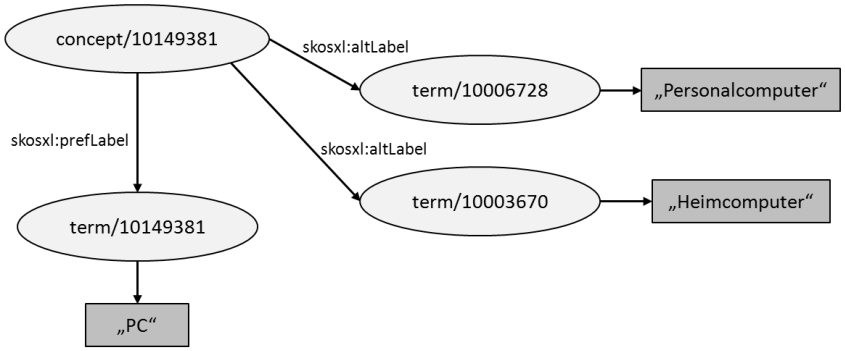

Abb.2: Konzept-basierte Repräsentation von Termen in SKOS

a) TheSoz (Thesaurus Sozialwissenschaften, GESIS)
Aufgrund der in 3.1 beschriebenen Besonderheiten des TheSoz und der stark ausgeprägten Term-Zentrierung stellte sich ein Mapping auf das SKOS-Format aufwändiger dar als zunächst angenommen (Zapilko & Sure, 2009). Um der konzept-basierten Struktur von SKOS zu entsprechen, ohne die relevanten Beziehungen zwischen präferierten und nicht-präferierten Termen zu verlieren, wurde auf die Klassen und Relationen von SKOS-XL zurückgegriffen, die explizit für die Repräsentation lexikaler Sachverhalte entwickelt wurden. Die Nutzung von SKOS-XL ermöglicht es, zum einen Termrelationen innerhalb eines SKOS-Konzeptes zu realisieren und zum anderen hierarchische Relationen zwischen Konzepten zu modellieren. In der folgenden Abbildung (Abb. 3) wird dieser Aspekt in abstrahierter Darstellung verdeutlicht.

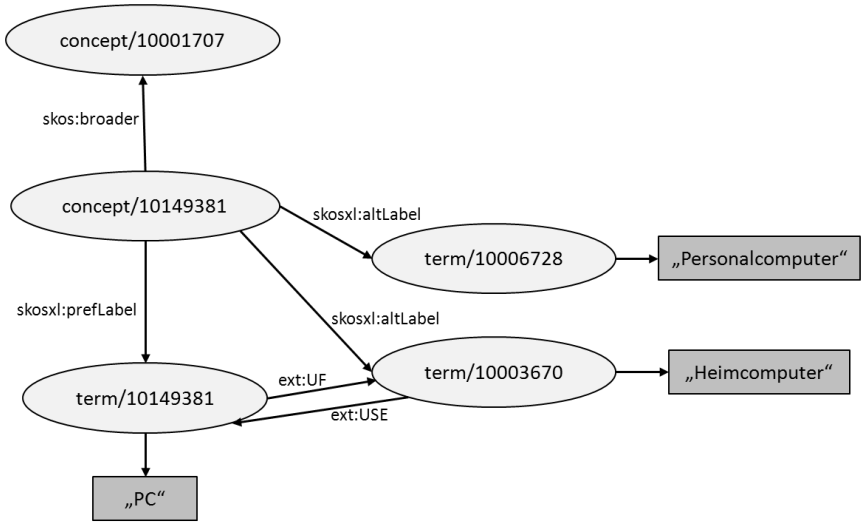

Abb. 3: Zusammenspiel von Konzepten und Termen in der SKOS-Version des TheSoz

Die so genannten „Label Relations“ von SKOS-XL ermöglichen die Definition eigener Relationen zwischen lexikalen Labels, wie in diesem Fall einzelne Terme bezeichnet werden. Daher ist durch ihre Verwendung eine solide Modellierung von Äquivalenzrelationen wie „Use“ und „Used For“ möglich, aber auch „Use Combination“ sowie „Used For Combination“ können auf diese Weise dargestellt werden. Durch die Zuordnung mehrerer „Label Relations“ zu einem Term wird gleichzeitig die Problematik der alternativen Nicht-Deskriptoren des TheSoz behandelt (siehe Abb. 4).

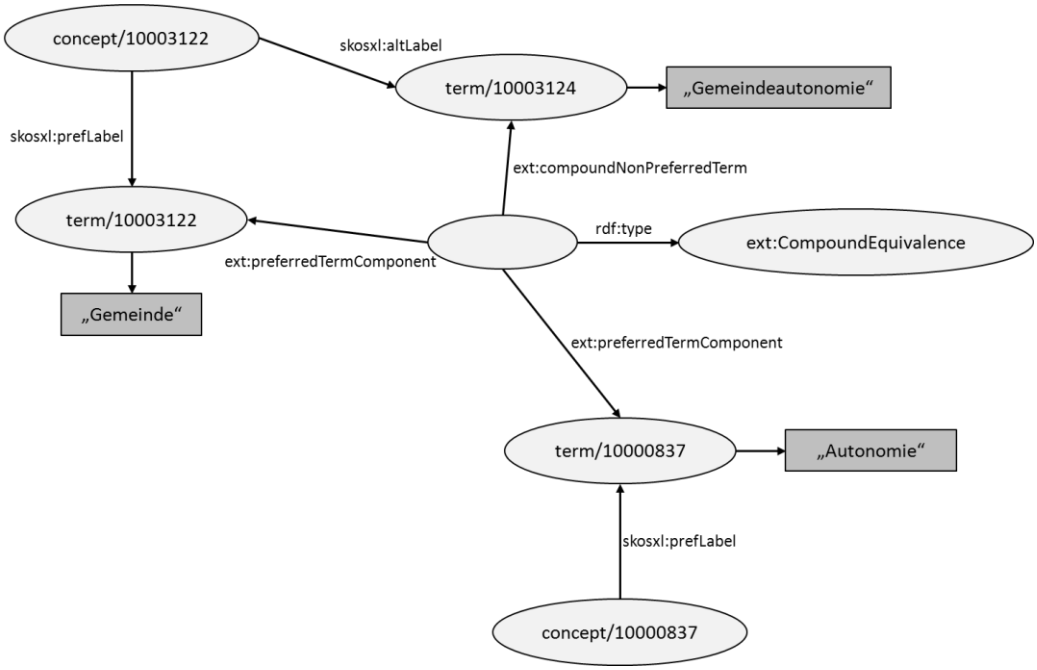

Abb. 4: Relation „Used For Combination“

Die Verwendung von SKOS-XL für diese komplexen Thesaurus-Strukturen orientiert sich im Anwendungsfalle des TheSoz an den vorgestellten SKOS-Erweiterungen für den EUROVOC-Thesaurus (Smedt, 2009). Bei eigens definierten Erweiterungen ist stets zu beachten, dass diese ausführlich durch die Beschreibung mit Standard-Klassen und -Properties beschrieben werden, damit sie von anderen Daten und Anwendungen möglichst korrekt verarbeitet werden können. Die folgende Tabelle (siehe Tab. 1) zeigt eine Aufstellung über das Mapping des TheSoz ins SKOS-Format.

Thesaurus Element	Beschreibung / Funktion	SKOS Klasse / Property
DD	Deskriptor	skosxl:prefLabel
ND	Nicht-Deskriptor	skosxl:altLabel
AD	Alternativer Nicht-Deskriptor	skosxl:altLabel
NT	Unterbegriff (Narrower Term)	skos:narrower
BT	Oberbegriff (Broader Term)	skos:broader
RT	Verwandter Term (Related Term)	skos:related
USE	Benutze Y statt X (Use)	ext:USE als Property der Klasse ext:equivalenceRelationship
UF	Gegenteil von USE: Y benutzt statt X (Used For)	ext:UF als Property der Klasse ext:equivalenceRelationship
USK	Benutze Kombination X UND Y (Use Combination)	ext:preferredTermComponent als Property der Klasse ext:CompoundEquivalence
UFK	Gegenteil von USK: Benutzt für X mit Y (Used For Combination)	ext:compoundNonPreferredTerm und ext:preferredTermComponent als Properties der Klasse ext:CompoundEquivalence
translation	Engl. Übersetzung des Terms	ext:hasTranslation
scope	Scope Notes	skos:scopeNote
notationcode	Numerischer Code der systematischen Klassifikation, der der Term zuzuordnen ist	skos:notation

Tab. 1: Mapping des TheSoz nach SKOS

b) STW (Standard-Thesaurus Wirtschaft, ZBW)
Der STW konnte relativ geradlinig auf adäquate SKOS-Elemente gemappt werden. Lediglich zwei Bereiche erforderten die Definition von Erweiterungen. Um zwischen Deskriptoren und Einträgen der Taxonomie unterscheiden zu können (beispielsweise bei Anfragen), wurden zugehörig zum „skos:Concept“ die zwei Subklassen „zbwext:Descriptor“ und „zbwext:Thsys“ eingeführt. Außerdem wurde eine Unterklasse für die „skos:note“ definiert. Diese „zbwext:useInsteadNote“ enthält spezielle Hinweise über „Use Instead“-Relationen zwischen Termen, die nicht verloren gehen sollen. Detaillierte Ausführungen über die Transformation des STW nach SKOS finden sich in (Neubert, 2009).

c) IBLK-Thesaurus (SWP)
Der IBLK-Thesaurus wurde bisher noch nicht nach SKOS übertragen, ein Mapping der Thesaurus-Elemente und -Strukturen steht noch aus. Der IBLK-Thesaurus wurde im Projekt KoMoHe sowohl zum TheSoz als auch zum STW über Crosskonkordanzen verbunden. Aus diesem Grund bietet sich der Thesaurus für das Mehr-Thesauri-Szenario in dieser Fallstudie an.

3.3 Technische Konvertierung nach SKOS
Basierend auf den in 3.2 erstellten Mappings können die Thesauri mit Hilfe einer Konvertierungsroutine nach SKOS transformiert werden. Zusätzlich zu den Mappings erhält in diesem Schritt jedes Element des Thesaurus (also jedes Konzept und jeder Term) eine eigene eindeutige URI, über die das Element identifiziert, referenziert und somit auch verlinkt werden kann. Der technische Transformationsprozess läuft in der Regel automatisiert ab. Da der Thesaurus Sozialwissenschaften bereits als XML-Datei vorlag, wurde er mittels XSL-Transformation nach SKOS RDF/XML konvertiert. Durch die Umwandlung der ausgeprägten Term-Zentralität des TheSoz in eine konzept-basierte Repräsentation entstand zusätzlicher manueller Aufwand.

4 Modellierung bestehender Crosskonkordanzen mit den SKOS Mapping Properties
Liegen wie in unserem Beispiel zwei Thesauri im SKOS-Format vor, so können auch bereits existierende Crosskonkordanzen zwischen ihnen in SKOS repräsentiert werden. Unter Anwendung der SKOS Mapping Properties „skos:exactMatch“, „skos:closeMatch“, „skos:broadMatch“, „skos:narrowMatch“ sowie „skos:relatedMatch“ lassen sich grundsätzlich den Crosskonkordanzen entsprechende Relationen zwischen Konzepten zweier Vokabulare modellieren.

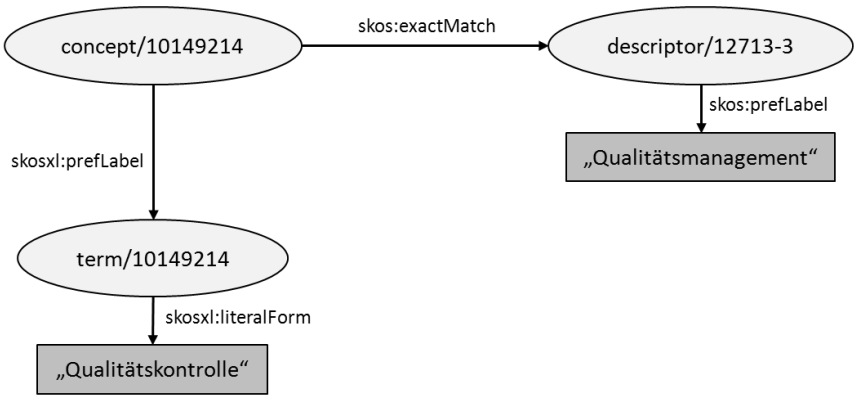

Abb. 5: Versuch eines Mappings zwischen TheSoz und STW in SKOS

Ein Mapping erfolgt, wie das obige Beispiel zeigt (siehe Abb. 5), nur zwischen den URIs der beteiligten Konzepte. Über die URIs wird jeweils die Referenz zu dem beteiligten Konzept mit all seinen enthaltenen Termen und Relationen hergestellt. Bei bilateralen Crosskonkordanzen ist zudem zu beachten, dass die Mappings zwischen Konzepten in beide Richtungen nicht zwingend identisch sein müssen, sondern sich voneinander unterscheiden können. Daher sollten die entsprechenden Mappings in

SKOS auch bilateral dargestellt werden, damit keine relevanten Beziehungen verloren gehen oder verfälscht dargestellt werden.
Es bleibt weiterhin ein relevanter Aspekt zu beachten und zu untersuchen, nämlich ob Mappings generell eher zwischen Konzepten oder auf Termebene erstellt werden. Die Grundlage der Crosskonkordanzen, traditionelle Thesauri, ist term-basierend, SKOS sieht aber eine konzept-basierende Repräsentation vor. Die teilweise stark variierende Struktur der Thesauri (z. B. Repräsentation von Konzepten und Termen) erschwert diese Problematik. Eine automatisierte Übertragung der Crosskonkordanzen nach SKOS als Ganzes wird dadurch insofern erschwert, dass SKOS-Mappings für jedes Thesauri-Paar ggf. aufs Neue modelliert werden müssen.
Darüber hinaus existieren auch bei den Crosskonkordanzen Sonderfälle, deren Auftreten sich durch disziplinäre Unterschiede und Überlappungen in den beteiligten Thesauri nicht vermeiden lässt und deren Darstellung mit den SKOS Mapping Properties sich derzeit als schwierig erweist. Ein Konzept eines Thesaurus kann beispielsweise einer Kombination aus Konzepten des anderen Thesaurus entsprechen, beispielsweise entspricht das Konzept „Produktionsstatistik“ im STW der Kombination der TheSoz-Konzepte „Produktion“ und „Statistik“. Für diesen Sachverhalt müsste eine Relation definiert werden, die sich äquivalent zu einer eigens definierten „Use Combination“-Relation (siehe 3.2) verhält, allerdings auf Ebene der Mapping Properties angewandt werden kann. Eine Lösung dieser Problematik steht derzeit noch aus.

5 Realisierung eines Mehr-Thesauri-Szenarios als LOD-Anwendung

Um die Interoperabilität der SKOS-Thesauri untereinander sowie mit anderen frei verfügbaren Daten zu ermöglichen, müssen sie zunächst über eine Schnittstelle nach außen hin ansprechbar gemacht werden. Ein etabliertes Verfahren hierfür ist, Datenquellen über ein SPARQL[11]-Protokoll, auch als SPARQL-Endpoint bezeichnet, verfügbar zu machen. SPARQL (SPARQL Protocol and RDF Query Language) ist eine graph-basierte Anfragesprache, mit der RDF-Daten angefragt werden können. Da die Thesauri verschiedener Anbieter in der Regel verteilt über verschiedene Schnittstellen angeboten werden, ist das SPARQL-Protokoll zumindest eine standardisierte Möglichkeit der Abfrage. Um die Inhalte mehrerer Thesauri, die über verschiedene SPARQL-Endpoints angeboten werden, gleichzeitig verlinkt repräsentieren zu können, kann das Tool Pubby[12] eingesetzt werden, das an der FU Berlin entwickelt wurde und auch im Rahmen des DBpedia[13]-Projekts eingesetzt wird. Pubby erweitert SPARQL-Endpoints um ein Linked Data Interface und ermöglicht durch die Dereferenzierung von URIs, die in den SKOS-Thesauri hinterlegten URIs im Browser aufzulösen und eine HTML-Darstellung zu generieren, die verfügbare Inhalte zu der im Browser angefragten Ressource (im Fall von Thesauri wären dies Terme und Konzepte) anzeigt. Auch wenn die zugrundeliegenden Thesauri über verschiedene SPARQL-Endpoints angesprochen werden, kann auf diesem Weg eine einheitliche HTML-Darstellung generiert werden.

[11] http://www.w3.org/TR/rdf-sparql-query/
[12] http://www4.wiwiss.fu-berlin.de/pubby/
[13] http://dbpedia.org/

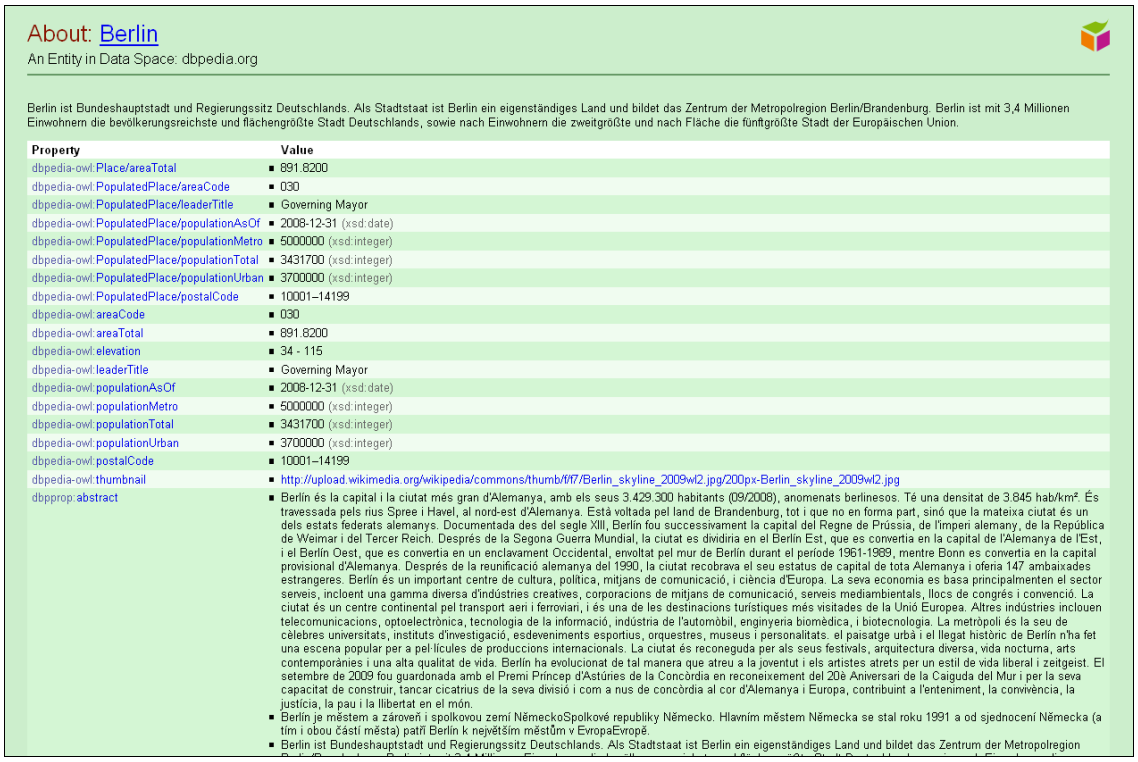

Abb. 6: Pubby als HTML-Darstellung für DBpedia

In einem ersten Schritt können so Terme und Konzepte der beteiligten Thesauri mit ihren zugehörigen Relationen dargestellt werden. Einen wesentlichen Bestandteil des Mehr-Thesauri-Szenarios stellen allerdings die zuvor transformierten Crosskonkordanzen dar, denn sie verbinden die Thesauri letztendlich erst miteinander. Zu jedem in Pubby angeforderten Term oder Konzept können nun zusätzlich bestehende Crosskonkordanzen in der HTML-Darstellung mit angezeigt werden. Die miteinander verlinkten Thesauri dieses Szenarios können somit wiederum als Brücke zwischen weiteren im Web verfügbaren Datenquellen dienen, beispielsweise in Form von Schlagwörtern.

6 Zusammenfassung und Ausblick

Die Fallstudie zeigt, dass perspektivisch die heterogene Landschaft der Erschließungsvokabulare über SKOS vereinheitlicht und vor allem die Inhalte der klassischen Datenbanken (Bereich Fachinformation) für Anwendungen des Semantic Web, beispielsweise als Linked Open Data (LOD), zugänglich und stärker miteinander vernetzt werden können. Vokabulare im SKOS-Format können dabei eine wichtige Funktion einnehmen, indem sie als standardisiertes Brückenvokabular dienen und semantische Verlinkung zwischen erschlossenen, veröffentlichten Daten herstellen.

Die Fallstudie zeigt außerdem, dass bestehende Thesauri nicht rein automatisch nach SKOS überführt werden können, sondern dass spezifische Anpassungen an den Mappings vorgenommen werden müssen, wenn die Spezifika der Thesauri erhalten bleiben sollen. Die Realisierung eines Mehr-Thesauri-Szenarios auf Basis von Crosskonkordanzen und SKOS, die hier nur konzeptionell erfolgt ist, zeigt deutliche Potentiale, insbesondere für die verteilte Suche in heterogenen Suchumgebungen. Problematisch ist, dass Instrumente wie Crosskonkordanzen, die zeitlich vor der Standardisierung von SKOS entstanden sind, und in dem Fall ohne persistente Identifikatoren (z. B. URI) realisiert wurden, in SKOS-Szenarien konzeptionell umgestaltet werden müssen.

Als nächster konkreter Entwicklungsschritt, der aus dieser Fallstudie folgt, ist zunächst die technische Realisierung des vorgestellten Mehr-Thesauri-Szenarios in Form eines Demonstrators auf Basis von SKOS-Thesauri und -Crosskonkordanzen zu nennen. Darüber hinaus bedarf es weiterer Untersuchungen, inwieweit sich ein generelles Verfahren zur Überführung der Crosskonkordanzen nach SKOS entwickeln lässt, das einen möglichst minimalen Aufwand der Nachbearbeitung beinhaltet. Vor dem

Hintergrund, dass die Crosskonkordanzen auf einer „traditionellen“ Repräsentation der Thesauri basieren, diese allerdings in SKOS wiederum sehr unterschiedlich repräsentiert werden, bringt dieser Aspekt weiteres Forschungspotenzial mit sich. Schlussendlich sind auf diesem Szenario aufsetzende Suchanwendungen zu entwickeln, die mehrere SKOS-Thesauri involvieren, um Chancen und Potenziale vertieft analysieren zu können.

Literatur

Assem, Mark van; Malaisé, Véronique; Miles, Alistair; Schreiber, Guus (2006): A Method to Convert Thesauri to SKOS. In: European Semantic Web Conference. URL: http://www.cs.vu.nl/~mark/papers/Assem06b.pdf

Krause, Jürgen (2008): Semantic heterogeneity: comparing new semantic web approaches with those of digital libraries. In: Library Review 57, Nr. 3, S. 235-248

Mayr, Philipp; Petras, Vivien (2008a): Building a terminology network for search: the KoMoHe project. S. 177-182. In: Greenberg, Jane; Klas, Wolfgang (Hrsg.): Metadata for semantic and social applications: Proceedings of the 8. International Conference on Dublin Core and Metadata Applications. Berlin: Uni.-Verl. Göttingen. URL: http://edoc.hu-berlin.de/docviews/abstract.php?lang=ger&id=29148

Neubert, Joachim (2009): Bringing the "Thesaurus for Economics" on to the Web of Linked Data. In: Linked Data on the Web (LDOW2009). URL: http://events.linkeddata.org/ldow2009/papers/ldow2009_paper7.pdf

Smedt, Johan De (2009): SKOS Extensions for the EUROVOC Thesaurus. In: 3rd Annual European Semantic Technology Conference.

W3C (2009): SKOS Simple Knowledge Organization System Reference: W3C Recommendation 18 August 2009. W3C. URL: http://www.w3.org/TR/skos-reference/

Zapilko, Benjamin; Sure, York (2009): Converting the TheSoz to SKOS. GESIS Technical Report 2009/07. GESIS - Leibniz Institute for the Social Sciences. ISSN: 1868-905. (2009) URL: http://www.gesis.org/fileadmin/upload/forschung/publikationen/gesis_reihen/gesis_methodenberichte/2009/TechnicalReport_09_07.pdf