

Hochschule Hannover

Fakultät III - Medien, Information und Design

Abteilung Information und Kommunikation

Sommersemester 2016

Nicht-standardisierte Erweiterungen von SKOS-Thesauri und ihre Auswirkungen auf die Kompatibilität

vorgelegt von:

Daniel Freytag

Matrikelnr.: 1243201

Erstprüfer: Prof. Dr. Christian Wartena

Zweitprüferin: Dr. Ina Blümel

Einbeck, den 04.05.2016

Abstract:

Vorliegende Arbeit beschäftigt sich mit den Auswirkungen von selbst-definierten Extensions auf die Kompatibilität von SKOS-Thesauri untereinander. Zu diesem Zweck werden als Grundlage zunächst die Funktionsweisen von RDF, SKOS, SKOS-XL und Dublin Core Metadaten erläutert und die verwendete Syntax geklärt. Es folgt eine Beschreibung des Aufbaus von konventionellen Thesauri inkl. der für sie geltenden Normen. Danach wird der Vorgang der Konvertierung eines konventionellen Thesaurus in SKOS dargestellt. Um dann die selbst-definierten Erweiterungen und ihre Folgen betrachten zu können, werden fünf SKOS-Thesauri beispielhaft beschrieben. Dazu gehören allgemeine Informationen, ihre Struktur, die verwendeten Erweiterungen und ein Schaubild, das die Struktur als Übersicht darstellt. Anhand dieser Thesauri wird dann beschrieben wie Mappings zwischen den Thesauri erstellt werden und welche Herausforderungen dabei bestehen. Ein Fazit schließt die Arbeit ab.

Inhaltsverzeichnis

Abkürzungsverzeichnis	I
Tabellenverzeichnis	II
Abbildungsverzeichnis	III
1 Einleitung	1
2 SKOS	2
2.1 RDF	3
2.2 Syntax	4
2.2.1 Turtle	4
2.2.2 RDF/XML	5
2.3 Funktionsweise von SKOS	6
2.4 SKOS-XL	10
2.5 Dublin Core	14
3 Thesauri	16
3.1 Funktionsweise	16
3.2 Normen	18
4 Kodierung von SKOS Thesauri am Beispiel des Thesaurus Sozialwissenschaften	19
4.1 Analyse des Thesaurus	19
4.2 Mapping von Thesaurus- und SKOS-Klassen	21
5 Beschreibung der Beispielthesauri	24
5.1 Agrovoc	25
5.2 Eurovoc	29
5.3 Standard Thesaurus Wirtschaft	33
5.4 Thesaurus Sozialwissenschaften	37
5.5 UNESCO Thesaurus	41

6 Mapping von Thesauri	45
6.1 Methoden des Thesaur-Mappings	47
6.2 Herausforderungen beim Thesaurus-Mapping	48
6.2.1 Modellierung von Konzepten	49
6.2.2 Modellierung von Klassifikationen	50
6.2.3 Verwendung von Compound Equivalence	51
6.2.4 Unterschiede in der Erschließungstiefe	54
6.2.5 Unterschiede in der semantischen Struktur	55
7 Fazit	58

Literaturverzeichnis

Abkürzungsverzeichnis

KOS	Knowledge Organization System
SKOS XL	SKOS - Extension for Labels
RDF	Ressource Description Framework
OWL	Web Ontology Language
PPN	Pica Produktionsnummer
XML	EXtensible Markup Language
URI	Uniform Ressource Identifier
W3C	World Wide Web Consortium
SWAD	Semantic Web Advanced Development in Europe
OCLC	Online Computer Library Centre
NCSA	National Center for Supercomputing Applications
DCMI	Dublin Core Metadata Initiative
ISO	International Organization for Standardization
FAO	Food and Agriculture Organization
GVK	Gemeinsamer Verbundkatalog
CC	Creative Commons Rights Expression Language
GND	Gemeinsame Norm Datei

Tabellenverzeichnis

3.1	Thesaurus-Kürzel nach ISO 25964-1	17
4.1	Übersicht über selbst-definierte Erweiterungen des TheSOZ	22
6.1	Mapping Typen nach ISO-25964-2	45
6.2	Mapping Types nach ISO 25964-2 inkl. optionaler Relationen	46
6.3	Mapping Ergebnisse Agrovoc	48
6.4	Aufbau der Thesaurus Konzepte in Beispielthesauri	49
6.5	Mappings zwischen Agrovoc, STW und Eurovoc	56

Abbildungsverzeichnis

2.1	Ein RDF-Graph mit Literalen zur Beschreibung von Datenwerten	3
5.1	Struktur des Agrovoc Thesaurus	28
5.2	Struktur des Eurovoc Thesaurus	32
5.3	Struktur des Standard Thesaurus Wirtschaft (STW)	36
5.4	Struktur des Thesaurus Sozialwissenschaften	40
5.5	Struktur des UNESCO Thesaurus	44

1 Einleitung

Diese Arbeit beschäftigt sich mit der Kodierung von Thesauri im Simple Knowledge Organization System (SKOS) und beleuchtet die dabei entstehenden Schwierigkeiten. Besonderes Augenmerk liegt dabei auf den unterschiedlichen SKOS-Modellierungen der Thesauri, den dabei entstandenen selbst-definierten Erweiterungen von SKOS und den Auswirkungen all dieser Unterschiede auf die Kompatibilität der Thesauri untereinander, bzw. das Mapping der Thesauri aufeinander. Zu diesem Zweck wird im ersten Teil der Arbeit zunächst beschrieben, was SKOS ist und welche Regeln für seine Anwendung gelten. Dazu gehört auch die verwendete Syntax, die Erweiterung SKOS-XL, sowie Beschreibungen von RDF als Unterbau von SKOS und dem Dublin Core Metadaten-Standard, der in SKOS-Vokabularen häufige Anwendung findet. In nächstem Abschnitt der Arbeit wird dann beschrieben, was ein Thesaurus ist und welche Normen für seine Erstellung galten und gelten. Nachdem diese wichtigen Grundlagen beschrieben wurden beschäftigt sich Kap. 4 mit der Kodierung eines Thesaurus in SKOS. Dabei dient die Konvertierung des Thesaurus Sozialwissenschaften als Beispiel.

Um dann im weiteren Verlauf der Arbeit die beim Thesaurus-Mapping verwendeten Methoden und die resultierenden Herausforderungen besser beschreiben zu können folgt eine detaillierte Beschreibung von 5 Beispielthesauri. Diese sind der Agrovoc Thesaurus der Vereinten Nationen, der Eurovoc von der Europäischen Union, der Standard Thesaurus Wirtschaft des Leibniz-Informationszentrums für Wirtschaft, der Thesaurus Sozialwissenschaften vom Leibniz-Institut für Sozialwissenschaften (GESIS) und die SKOS-Version des UNESCO-Thesaurus. Ausgewählt wurden diese Thesauri anhand von Größe, Relevanz und ihrer für diese Arbeit interessanten Verwendung des SKOS-Standards. Zu der Beschreibung der Thesauri gehören neben allgemeinen Informationen eine genaue Beschreibung der Struktur des Thesaurus, sowie ein Schaubild, anhand dessen sich diese Struktur übersichtlich nachvollziehen lässt. Der folgende Teil der Arbeit beschäftigt sich dann damit, wie Thesauri miteinander verlinkt bzw. aufeinander gemappt werden und welche Herausforderungen dabei auftreten. Abgeschlossen wird die Arbeit mit einem Fazit, das die aktuelle Situation von SKOS-Thesauri und beschreibt und die Ergebnisse der Arbeit zusammenfasst.

Zur Darstellung der Arbeit sei angemerkt, dass alle Beispiele für SKOS aus Gründen der Lesbarkeit in Turtle Syntax aufgeführt sind. Alle SKOS-Klassen und Prädikate sind auf folgende Art gekennzeichnet:

```
skos:concept .
```

2 SKOS

SKOS ist ein vom World Wide Web Consortium (W3C) veröffentlichter Standard, mit dem verschiedene kontrollierte Vokabulare, wie Taxonomien, Klassifikationen oder Thesauri für die Verwendung in Semantic Web Anwendungen kodiert werden können. Zum Zeitpunkt dieser Arbeit gibt es laut einer Untersuchung von Manaf et al¹ etwa 478 SKOS-Vokabulare, von denen 54 Thesauri sind, mit denen sich diese Arbeit beschäftigt. Die Entwicklung von SKOS begann im Rahmen des Semantic Web Advanced Development in Europe (SWAD) Projekts, eines EU-Projekts zur Weiterentwicklung und Verbreitung der Semantic Web Aktivitäten des W3C, zwischen 2002 und 2004 und wurde dann zwischen 2004 und 2009 vervollständigt. Die ursprünglichen Dokumente *SKOS Core Guide* und *SKOS Core Vocabulary Specification* aus dem Jahr 2005 sind dabei inzwischen vom *SKOS Primer*² und *SKOS Reference*³ abgelöst worden. Diese Dokumente stellen jetzt die grundlegenden Dokumente dar, in denen SKOS beschrieben wird. Beide Dokumente, sowie weitere Informationen, finden sich auf der SKOS Website des W3C⁴. Im Folgenden Kapitel wird zunächst, hauptsächlich anhand dieser Dokumente, die Funktionsweise von SKOS insoweit erläutert, wie es für das Verständnis der späteren Kapitel notwendig ist. Dafür wird zunächst das Resource Description Framework (RDF), als Grundlage für SKOS, kurz erläutert und die meist für SKOS verwendete Syntaxen RDF/XML und Turtle vorgestellt. Danach folgt eine Beschreibung der grundlegenden Funktionsweise von SKOS und der SKOS-Erweiterung SKOS - Extension for Labels (SKOS XL)⁵. SKOS-XL findet in der Praxis häufige Anwendung und ist wichtig für die weiteren Kapitel der Arbeit. Abgeschlossen wird das Kapitel mit einem kurzen Abschnitt über den Dublin Core Metadaten Standard, der in vielen SKOS-Thesauri für das Hinzufügen von Metadaten verwendet wird.

¹ ABDUL MANAF ET AL. (2012, S.279)

² ISAAC & SUMMERS (2009)

³ MILES & BECHHOFER (2009b)

⁴ <https://www.w3.org/2004/02/skos/>; zuletzt geprüft am 09.03.2016

⁵ MILES & BECHHOFER (2009a)

2.1 RDF

Das RDF ist ein vom W3C entwickelter Standard zur Beschreibung von Ressourcen und ihren Relationen im Semantic Web in maschinenlesbarer Form. Entwickelt wurde es zwischen 1999 und 2004. Version 1.1 wurde 2014 veröffentlicht. RDF stellt ein Standard Modell für den Datenaustausch im Semantic Web dar⁶ und bildet auch die Grundlage für SKOS. RDF ist für Situationen gedacht in denen Daten im Netz von Applikationen verarbeitet und unter Applikationen ausgetauscht werden sollen ohne dass die Bedeutung und die Beziehungen der Daten verloren gehen⁷.

Daten werden in RDF in Graphen beschrieben, die aus Ressourcen gebildet werden. Um Ressourcen zu beschreiben werden diese in RDF zunächst mit Uniform Resource Identifiers (URIs) eindeutig identifiziert und dann mit Aussagen beschrieben. Die Aussagen bestehen aus Subjekt, Prädikat und Objekt und bilden damit sogenannte Tripel. In folgender Abbildung nach Hitzler⁸ wird ein RDF-Graph dargestellt. Die Kreise des Graphen sind Ressourcen mit den ihnen gegebenen URIs. Diese können sowohl Subjekt als auch Objekt eines Tripels sein. Die Kanten stellen die Prädikate dar. Es handelt sich hierbei ebenfalls um Ressourcen mit einer URI. Die Kästen sind Literale. Diese bilden immer das Objekt eines Tripels und beschreiben das Subjekt mit einem Wert. Der abgebildete Graph beschreibt drei Tripel:

1. Die Ressource `http://example.org/SemanticWeb` hat den Titel *Semantic Web - Grundlagen*
2. Die Ressource `http://example.org/SemanticWeb` wurde verlegt bei der Ressource `http://www.springer.com/Verlag`
3. Die Ressource `http://www.springer.com/Verlag` hat den Namen *Springer-Verlag*

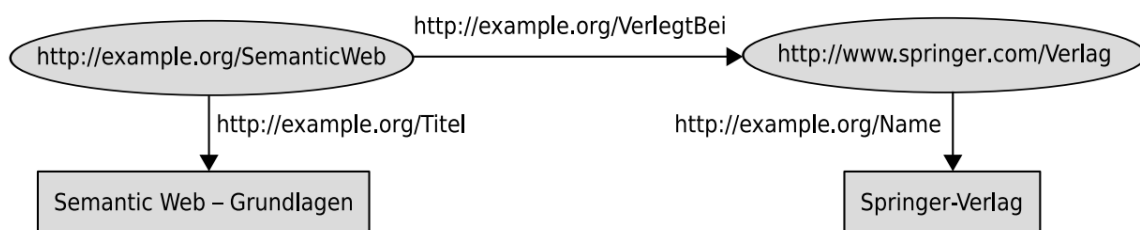


Abbildung 2.1: Ein RDF-Graph mit Literalen zur Beschreibung von Datenwerten

Um diese Tripel in maschinenlesbarer Form darzustellen werden verschiedene Syntaxen benutzt, die im

⁶ W3C SEMANTIC WEB WIKI (2014)

⁷ SCHREIBER & RAIMOND (2014)

⁸ HITZLER (2008, S.39)

nächsten Abschnitt erläutert werden. RDF und seine Art der Beschreibung von Ressourcen bilden auch die Grundlage für SKOS und sind für nahezu alle Semantic Web Anwendungen wichtig. Eine detaillierte Beschreibung der Regeln und Funktionsweise von RDF soll im Rahmen dieser Arbeit nicht stattfinden. Weitere Informationen finden sich im *RDF 1.1 Primer*⁹.

Während RDF ein wichtiger Grundstein für das Semantic Web ist, wären die meisten Anwendungen und Vokabulare wie SKOS nicht denkbar ohne die Verwendung von weiteren auf RDF aufbauenden W3C-Standards. Besonders wichtig sind dabei RDFS und die Web Ontology Language (OWL). Eine ausreichende Beschreibung dieser Standards würde den Rahmen dieser Arbeit übersteigen. Stattdessen sei auf die Dokumente des W3C verwiesen, die diese Standards beschreiben¹⁰. Für weitere Informationen empfehlen sich auch die Werke von Allemang¹¹, Hitzler¹² und Pellegrini¹³. Elemente von RDFS und OWL die in SKOS besonders wichtig und für das Verständnis unerlässlich sind, werden an entsprechender Stelle direkt erläutert.

2.2 Syntax

Als Syntax für RDF und SKOS wird im Allgemeinen entweder RDF/XML oder Turtle verwendet. Die meisten Online verfügbaren SKOS-Thesauri bieten dabei die Möglichkeit die vorhandenen Konzepte in beiden Varianten darzustellen bzw. zu exportieren. Teilweise besteht zusätzlich die Möglichkeit anderer Darstellungen, wie beispielsweise JSON¹⁴. Im Folgenden sollen Turtle und RDF/XML jeweils kurz beschrieben werden.

2.2.1 Turtle

Turtle ist eine Syntax für RDF und verwandte Standards, die N-Tripel aus der N3-Notation von Tim Berners-Lee¹⁵ mit der Verwendung von *Qualified Names* (QNames) verbindet, um eine möglichst lesbare und für den Anwender einfach zu editierende Darstellung der Tripel zu erreichen¹⁶. QNames sind dabei

⁹ SCHREIBER & RAIMOND (2014)

¹⁰ HITZLER (2012), BRICKLEY & GUHA (2014)

¹¹ ALLEMANG & HENDLER (2011)

¹² HITZLER (2008)

¹³ PELLEGRINI & BLUMAUER (2006)

¹⁴ Die JavaScript Object Notation, kurz JSON, ist ein kompaktes Datenformat in für Mensch und Maschine einfach lesbarer Textform zum Zweck des Datenaustauschs zwischen Anwendungen. Quelle: https://de.wikipedia.org/wiki/JavaScript_Object_Notation; zuletzt geprüft am 02.05.2016

¹⁵ BERNERS-LEE & CONNOLLY (2011)

¹⁶ ALLEMANG & HENDLER (2011, S. 45)

Kürzel die URIs abkürzen und damit einfach verwendbar machen. Dafür werden am Anfang einer in Turtle vorliegenden Ansammlung von Tripeln Kürzel definiert, die dann für alle Tripel Anwendung finden. Im Beispiel werden die Kürzel für SKOS und RDF, sowie ein Namensraum für Beispiele definiert:

```
@prefix skos: <http://www.w3.org/2004/02/skos/core#>.
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix ex: <http://www.example.com/> .
```

Die Darstellung der eigentlichen Tripel findet dann mit den Kürzeln statt:

```
ex:animals      rdf:type      skos:concept
                 skos:prefLabel "animals"@en;
                 skos:altLabel  "creatures"@en;
                 skos:hiddenLabel "animals"@en.
```

`ex:animals` ist durch die vordefinierten Kürzel gleichbedeutend mit `http://www.example.com/animals` usw. Wie am Beispiel zu erkennen kann das Subjekt des Tripels bei wiederholtem Vorkommen entfallen. Die Tripel mit gleichem Subjekt werden dabei mit Semikolon abgetrennt. Am Ende eines Blocks von Tripeln folgt ein Punkt. Turtle bietet zur weiteren Verkürzung auch die Möglichkeit `rdf:type` durch `'a'` zu ersetzen.

```
ex:animals      a      skos:concept
```

ist also gleichbedeutend mit:

```
ex:animals      rdf:type      skos:concept
```

Eine detaillierte Beschreibung des Standards liefert das W3C¹⁷.

2.2.2 RDF/XML

RDF/XML ist eine Variante von EXtensible Markup Language (XML) die das W3C aus Gründen der Geläufigkeit der Verwendung von XML im Internet, als Darstellungsform für RDF und damit auch SKOS empfiehlt¹⁸.

Im Beispiel werden die gleichen Tripel wie oben in RDF/XML dargestellt:

```
<?xml version="1.0" encoding="utf-8" ?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:skos="http://www.w3.org/2004/02/skos/core#">
```

¹⁷ PRUD'HOMMEAUX & CAROTHERS (2014)

¹⁸ ALLEMANG & HENDLER (2011, S.46)

```

<skos:concept rdf:about="http://www.example.com/animals">
  <skos:prefLabel xml:lang="en">animals</skos:prefLabel>
  <skos:altLabel xml:lang="en">creatures</skos:altLabel>
  <skos:hiddenLabel xml:lang="en">animals</skos:hiddenLabel>
</skos:concept>

</rdf:RDF>

```

Auch in dieser Syntax werden zunächst Kürzel definiert. Danach werden die Tripel in XML dargestellt. Diese Darstellungsweise ist in der Praxis häufig, da XML ohnehin weitverbreitete Anwendung im Internet findet. Der Nachteil dieser Darstellung ist die schlechte Lesbarkeit für den Anwender. Auch eignet sich diese Syntax auf Grund ihrer Länge nicht gut für die Darstellung in Texten. Eine detaillierte Beschreibung von RDF/XML findet sich beim W3C¹⁹

2.3 Funktionsweise von SKOS

SKOS als Datenmodell basiert auf RDF und liegt in der Praxis in XML oder Turtle-Syntax vor. Es beschreibt die Ressourcen und ihre Relationen im zu kodierenden Vokabular anhand verschiedener Klassen und Prädikate. Im Folgenden wird in Grundzügen die Funktionsweise von SKOS vorgestellt. Dies geschieht zur Sicherung der notwendigen Grundkenntnisse für die Ausführungen in späteren Kapiteln dieser Arbeit. Die Darstellung hier ist dabei keinesfalls als erschöpfend zu betrachten. Detaillierte Beschreibungen finden sich in den SKOS-Dokumenten des W3C, insbesondere des *SKOS-Primers*²⁰ und dem *SKOS-Reference*-Dokument²¹, aus denen auch die Beispiele in diesem Kapitel stammen.

Konzepte

Die Grundeinheit von SKOS stellt das Konzept dar. Konzepte repräsentieren die Idee hinter einem Begriff. Um Ressourcen als Konzepte darzustellen wird die Ressource zunächst mit einer einzigartigen URI versehen, um sie zweifelsfrei identifizieren zu können und dann mit RDF als `skos:concept` festgelegt. Dafür wird `rdf:type` verwendet, dass verwendet wird um eine Ressource einem bestimmten Typ zuzuweisen. In Turtle-Syntax, mit vorheriger Deklaration von QNames zur Verkürzung²², sieht dies folgendermaßen aus:

¹⁹ GANDON & SCHREIBER (2014)

²⁰ ISAAC & SUMMERS (2009)

²¹ MILES & BECHHOFER (2009b)

²² Bemerkung: Die Deklaration der Kürzel entfällt in den folgenden Beispielen

```
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix ex: <http://www.example.com/> .
```

```
ex:animals    rdf:type      skos:concept
```

Label

Wenn das Konzept festgelegt ist, wird es im zweiten Schritt gelabelt, d.h. das Konzept bekommt ein Literal in Form eines Strings zugewiesen. Es gibt drei Arten von Labels: `skos:prefLabel`, `skos:altLabel` und `skos:hiddenLabel`.

```
ex:animals    rdf:type      skos:concept
              skos:prefLabel "animals"@en;
              skos:altLabel  "creatures"@en;
              skos:hiddenLabel "animals"@en.
```

`skos:prefLabel` beschreibt dabei die Vorzugsbenennung des Konzepts und `skos:altLabel` eine alternative Benennung. `skos:hiddenLabel` dient dazu nicht verwendete Bezeichnungen oder häufig vorkommende falsche Schreibweisen abzudecken, damit z.B. im Fall einer Suche mit einem häufig auftretendem Schreibfehler das Konzept trotzdem gefunden werden kann. Wie im Beispiel sichtbar, werden die Strings der Labels mit Sprachkürzeln versehen. Dies bietet die Möglichkeit multilinguale Versionen der zu kodierenden Vokabulare darzustellen. Während `skos:prefLabel` pro Sprache nur einmal vorkommen darf, können die anderen Labels beliebig oft vergeben werden. Versehen mit mit URI und Labels ist das Konzept nun für die weitere Verwendung bereit.

Relationen

Um Knowledge Organization System (KOS) korrekt abbilden zu können, muss SKOS die Möglichkeit bieten, semantische Relationen darzustellen. Dies geschieht in SKOS durch verschiedene Klassen von Relationen, die die SKOS-Konzepte miteinander verbinden. Wichtig dabei zu beachten ist, dass Relationen in Standard-SKOS nur zwischen Konzepten bestehen können und nicht zwischen einzelnen Labels oder Relationen. SKOS bietet Möglichkeiten für die Darstellung von hierarchischen und assoziativen Relationen. Hierarchische Relationen werden durch die Klassen `skos:broader` und `skos:narrower` beschrieben, die über- bzw. untergeordnete Konzepte beschreiben.

```
ex:animals    rdf:type      skos:concept;
              skos:prefLabel "animals"@en;
```

```

                skos:narrower      ex:mammals.

ex:mammals      rdf:type          skos:concept;
                skos:prefLabel    "mammals"@en;
                skos:broader      ex:animals.

```

Assoziative Relationen, die ein nicht-hierarchisches Verwandtschaftsverhältnis beschreiben, werden mit `skos:related` beschrieben:

```

ex:bird         rdf:type          skos:concept;
                skos:prefLabel    "bird"@en;
                skos:related      ex:ornithology.

ex:ornithology  rdf:type          skos:concept;
                skos:prefLabel    "ornithology"@en;
                skos:related      ex:birds.

```

Metadaten

Zusätzlich zu der Darstellung von Relationen ist es in den meisten Anwendungsfällen auch gewünscht die Konzepte des KOS mit Metadaten anzureichern, um einen hohen Informationsgehalt zu erreichen und die Weiterverwertung der Daten zu vereinfachen. SKOS bietet standardmäßig eine Klasse `skos:note`, die es mit verschiedenen Spezialisierungen ermöglicht unterschiedliche Arten von Metadaten zu verzeichnen. Dazu gehören beispielsweise `skos:definition`, die die Definition eines Konzepts enthält, `skos:scopeNote`, die den Bereich angibt in dem das jeweilige Konzept Anwendung findet oder `skos:historyNote`, die die geschichtliche Entwicklung, bzw. veränderte Bedeutungen, eines Konzepts enthält.

Zusätzlich zu diesen Klassen, die allesamt inhaltliche Metadaten des Konzepts enthalten, gibt es auch Möglichkeiten Metadaten über das KOS an sich, bzw. seine Bearbeitung, zu verzeichnen. Für diese Zwecke gibt es `skos:editorialNote` und `skos:changeNote`. Diese Klassen sind für Administratoren oder Mitarbeiter des KOS gedacht, um Informationen über Wartung, Veränderung und Ähnliches aufnehmen zu können.

Alle Metadaten Klassen von SKOS werden mit Strings bezeichnet und bieten wie die Labels die Möglichkeit Notes in verschiedenen Sprachen zu verzeichnen.

```

ex:bird         rdf:type          skos:concept;
                skos:prefLabel    "bird"@en;
                skos:definition    "A warm-blooded egg-laying vertebrate animal..."@en;
                skos:changeNote    "last updated 01-01-2011; by Max Mustermann".

```

Concept Schemes

SKOS-Konzepte, versehen mit Labels, Relationen und Metadaten werden schließlich in einem `skos:conceptScheme` zusammengefasst. `skos:conceptScheme` beschreibt die Gesamtheit eines KOS, wie beispielsweise eines kompletten Thesaurus. Diese Klasse bietet die Möglichkeit das KOS an sich mit Metadaten zu versehen. Im Beispiel ist der *Animal Thesaurus* mit Dublin Core Metadaten (siehe 2.5) versehen worden:

```
ex:animalThesaurus    rdf:type          skos:conceptScheme;
                      dc:title           "Animal Thesaurus";
                      dc:creator        "Max Mustermann".
```

Um ein `skos:conceptScheme` mit den darin enthaltenen Konzepten zu verbinden gibt es das Prädikat `skos:inScheme`, das Konzepte einem `skos:conceptScheme` zuordnet und `skos:hasTopConcept`, das dieselbe Relation in umgekehrter Richtung beschreibt und einen Einstieg in die Hauptkonzepte eines KOS darstellt:

```
ex:bird               rdf:type          skos:concept;
                      skos:prefLabel    "bird"@en;
                      skos:inScheme     ex:animalThesaurus.

ex:animalThesaurus   rdf:type          skos:conceptScheme;
                      skos:hasTopConcept ex:bird.
```

Mappings zwischen Concept Schemes

Da jedes Konzept innerhalb eines KOS mit einer URI versehen und damit eindeutig identifizierbar ist, gibt es die Möglichkeit Relationen zwischen Konzepten verschiedener KOS darzustellen. Dafür gibt es in SKOS die Prädikate `skos:exactMatch`, das eine genaue Entsprechung in einem anderem KOS kennzeichnet und `skos:closeMatch`, das eine beinahe Entsprechung kennzeichnet. Zusätzlich gibt es Relationen äquivalent zu den hierarchischen Relationen innerhalb des KOS: `skos:broadMatch`, `skos:narrowMatch` und `skos:relatedMatch`.

```
@prefix ex2: <http://www.AnotherExample.com/> .
```

```
ex:bird               rdf:type          skos:concept;
                      skos:prefLabel    "bird"@en;
                      skos:inScheme     ex:animalThesaurus;
                      skos:exactMatch   ex2:vogel.

ex2:vogel             rdf:type          skos:concept;
                      skos:prefLabel    "Vogel"@de;
                      skos:inScheme     ex:TierThesaurus;
                      skos:exactMatch   ex:bird.
```

Im Beispiel gibt es zwei Thesauri für Tiere, einen auf Deutsch und einen auf Englisch. Beide enthalten das Konzept Vogel. Die Konzepte haben, da sie das gleiche beschreiben, eine `skos:exactMatch` Relation. Bemerkenswert ist hier, dass es durch die Verwendung von Sprachkürzeln möglich ist, sowohl das deutsche als auch das englische Konzept innerhalb eines einzigen Thesaurus darzustellen. Die Konstruktion im Beispiel ist also möglich, aber nicht unbedingt sinnvoll. Während SKOS im Einzelnen noch weitere Möglichkeiten bietet, soll die Darstellung an dieser Stelle ausreichen, um die späteren Kapitel verständlich zu machen. In diesem Kapitel ist auch bewusst nicht davon ausgegangen worden, dass es sich beim Beispiel-KOS um einen Thesaurus handelt. Wie SKOS für Thesauri verwendet wird, wird in Kapitel 4 und 6 näher beschrieben.

2.4 SKOS-XL

SKOS XL ist eine optionale Erweiterung für SKOS. Sie wird im Appendix B des SKOS Reference Documents²³ und dem SKOS-XL Namespace Dokument²⁴ beschrieben. SKOS-XL erweitert SKOS um neue Möglichkeiten, lexikale Entitäten zu beschreiben und zu verlinken. An dieser Stelle soll die Funktionsweise von SKOS-XL anhand von zwei Anwendungsfällen beschrieben werden. Dabei handelt es sich um das Anhängen von Metadaten an Labels und die Definition eigener Extensions für Relationen zwischen Labels.

Grundsätzlich erweitert SKOS-XL das SKOS Schema um 6 neue Klassen, bzw. Prädikate:

- `skosxl:label`
- `skosxl:literalForm`
- `skosxl:prefLabel`
- `skosxl:altLabel`
- `skosxl:hiddenLabel`
- `skosxl:labelRelation`

Die ersten 5 Klassen erlauben eine größere Flexibilität in der Vergabe von Metadaten in SKOS-Vokabularen. Beispielsweise nehme man folgende SKOS-Konstruktion:

²³ MILES & BECHHOFFER (2009b)

²⁴ MILES & BECHHOFFER (2009a)

```

@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix ex: <http://www.example.com/> .

```

```

ex:animalThesaurus    rdf:type          skos:conceptScheme;
                     skos:hasTopConcept  ex:birds;
                     skos:hasTopConcept  ex:mammals.

ex:mammals            rdf:type          skos:concept;
                     skos:prefLabel    "Mammals"@en;
                     skos:narrower     ex:cats.

ex:cats              rdf:type          skos:concept;
                     skos:prefLabel    "Cats"@en;
                     skos:altLabel     "Felines"@en;
                     skos:broader      ex:mammals.

```

Um dieser Konstruktion Metadaten hinzuzufügen, besteht die Möglichkeit die verschiedenen vorhandenen Instanzen von `skos:note` zu verwenden. Beispielsweise könnte man `skos:definition` verwenden um den Konzepten eine Definition hinzuzufügen:

```

ex:cats              rdf:type          skos:concept;
                     skos:prefLabel    "Cats"@en;
                     skos:altLabel     "Felines"@en;
                     skos:broader      ex:mammals;
                     skos:definition   "Some Definition of Cats..."@en.

```

Was aber hat man für Möglichkeiten Metadaten an einzelne Labels zu hängen? Will man z.B. an das `skos:altLabel` des Beispiels anhängen wer die alternative Benennung hinzugefügt hat oder wann sie hinzugefügt wurde, ist das mit Standard SKOS nur schwierig umzusetzen. Da SKOS in RDF-Tripeln funktioniert, lassen sich die Metadaten nicht einfach an das Literal "Felines" hängen, da ein Literal nicht das Subjekt eines RDF-Tripels sein kann. Ebenfalls würde es eine Regelverletzung darstellen `skos:altLabel` als Subjekt eines Tripels zu verwenden, da die Range, also der mögliche Wert der aufgenommen werden kann, von `skos:altLabel` ein Literal ist²⁵ und das Literal wiederum nur Objekt eines Tripels sein kann. Man könnte `skos:editorialNote` an das `skos:skos:concept` hängen und die gewünschten Informationen dort in Form eines Strings unterbringen, hätte damit aber keine ideale Lösung. Viele Informationen über verschiedene Labels, die alle direkt am Konzept hängen werden schnell unübersichtlich. Eine Möglichkeit wäre die Verwendung von sogenannten Blank Nodes. Ein Blank Node ist eine Ressource der keine URI gegeben wurde, die aber als Subjekt eines Tripels dienen kann. Eine Konstruktion mit

²⁵ MILES & BECHHOFFER (2009b, Kap. 5.3)

Blank Nodes könnte folgendermaßen aussehen:

```
ex:cats          rdf:type          skos:concept;
                 skos:prefLabel    "Cats"@en;
                 skos:altLabel      BN1;
                 skos:broader        ex:mammals;
                 skos:definition     "Some Definition of Cats..."@en.

BN1              rdf:value          "Felines"@en;
                 skos:editorialNote "written by Max Mustermann";
                 skos:historyNote    "last changed, 22-05-2015".
```

Die Verwendung von Blank Nodes umgeht das Problem des fehlenden Subjekts. Der Nachteil dieser Variante ist, dass Blank Nodes selbst keine URI haben und damit nicht zweifelsfrei identifiziert werden können.

Besser ist die Verwendung von SKOS-XL. Es bietet die Möglichkeit folgender Konstruktion:

```
@prefix skosxl: <http://www.w3.org/2008/05/skos-xl#>.
@prefix dc: <http://purl.org/dc/elements/1.1/>.
```

```
ex:cats          rdf:type          skos:concept;
                 skosxl:prefLabel  ex:Label1;
                 skosxl:altLabel    ex:label2;
                 skos:broader        ex:mammals;
                 skos:definition     "Some Definition of Cats..."@en.

ex:label2        rdf:type          skosxl:label;
                 skosxl:literalForm "Felines"@en;
                 dc:creator          "A. Catman";
                 dc:created          "01-01-1999".
```

Hier wird werden die Standard SKOS-Elemente `skos:prefLabel` & `skos:altLabel` durch Instanzen ihrer SKOS-XL Varianten ersetzt. Diese haben als Range kein Literal sondern die Klasse `skosxl:label`, deren Instanzen wiederum eine Ressource mit eigener URI darstellen (hier z.B. `ex:label2`). Diese Ressource enthält das Prädikat `skosxl:literalForm`, das das Literal enthält. Jetzt besteht die Möglichkeit an die neue Ressource `ex:label2` beliebige Metadaten anzuhängen. Im Beispiel werden *DC Element Set* Metadaten verwendet.

SKOS-XL bietet so die Möglichkeit Metadaten flexibel an verschiedene Bereiche eines Konzeptes anzuhängen. Zwei Instanzen der Klasse `skosxl:label` sind dabei nicht zwangsweise die gleiche Ressource,

auch wenn sie die gleiche `skosxl:literalForm` haben²⁶.

Eine Zweite Möglichkeit die SKOS-XL bietet ist die Definition eigener Relationen zwischen Instanzen von `skosxl:label` mit Hilfe von `skosxl:labelRelation`. Dieses Prädikat macht eine Konstruktion wie die folgende möglich:

```
ex:pc                rdf:type          skos:concept;
                    skosxl:prefLabel  ex:Label1;
                    skosxl:prefLabel  ex:Label2;

ex:label1            rdf:type          skosxl:label;
                    skosxl:literalForm "Personal Computer"@en.

ex:label2            rdf:type          skosxl:label;
                    skosxl:literalForm "PC"@en.

ex:label1            skosxl:labelRelation  ex:label2.
```

In diesem Beispiel hat das Konzept `ex:pc` zwei unterschiedliche `skosxl:prefLabel`, die zu zwei Instanzen von `skosxl:label` führen. Bei einer handelt es sich um die ausgeschriebene Bezeichnung "Personal Computer", bei der anderen um die geläufigere Abkürzung "PC". `skosxl:labelRelation` legt nun fest, dass zwischen beiden Labels eine Relation besteht ohne das festgelegt wird welcher Art diese Relation ist. Um dies festzulegen besteht die Möglichkeit `skosxl:labelRelation` mit einer eigenen Definition zu versehen, die vor der Verwendung mit `rdfs:subPropertyOf` deklariert wird:

```
ex:acronym           rdfs:subPropertyOf  skosxl:labelRelation .

ex:label2            ex:acronym          ex:label1.
```

Hier wurde `ex:acronym` als eine `skosxl:labelRelation` festgelegt um auszudrücken, dass es sich bei *PC* um eine Abkürzung von *Personal Computer* handelt. Sinnvoll wäre es dazu `ex:acronym` mit Metadaten, die die Bedeutung der Klasse beschreiben zu versehen:

²⁶ MILES & BECHHOFFER (2009b, Kap. B.2)

ex:acronym	rdfs:subPropertyOf skos:definition	skosxl:labelRelation; "Declares one Label as an acronym of the other" .
ex:label2	ex:acronym	ex:label1.

Diese Funktion von SKOS-XL ermöglicht also die flexible Definition eigener Relationen. In der Praxis ist die Benutzung von SKOS-XL weit verbreitet. Jeder der in Kapitel 6 beschriebenen Thesauri verwendet SKOS-XL und das Verständnis von SKOS-XL ist essentiell für die Betrachtung selbst-definierter Extensions in dieser Arbeit.

2.5 Dublin Core

Dublin Core (DC) ist ein Metadaten Standard der 1995 aus einem Workshop vom Online Computer Library Centre (OCLC) und dem National Center for Supercomputing Applications (NCSA) in Dublin, Ohio entstanden ist. Ziel des Standards ist es einen einfachen Weg zu schaffen strukturierte Metadaten zu Web Ressourcen hinzuzufügen²⁷. Betreut wird DC von der Dublin Core Metadata Initiative (DCMI)²⁸. Seit seiner Einführung hat sich der DC Standard in zwei Sets entwickelt. Das *DC Metadata Element Set* umfasst 15 Metadaten Elemente²⁹, die später entwickelten *DC Metadata Terms* umfassen ein stark erweitertes Set von 55 Elementen für die Beschreibung von sowohl digitalen als auch physischen Medien³⁰. Bei den Elementen des *Element Sets* handelt es sich um ein einfaches Set von Metadaten, wie man sie auch in beispielsweise bibliographischen Daten eines Bibliothekskatalogs finden würde³¹. Das stark erweiterte *DC Metadata Terms* Set enthält neben den 15 Elementen des *Element Sets* weitere 40 Elemente, die einen weiten Bereich von möglichen Metadaten abdecken³².

Dublin Core Metadaten haben inzwischen weite Verbreitung gefunden und sind im ISO Standard 15836-2009³³ standardisiert.

Für die Verwendung von DC Metadaten Sets in SKOS wird zunächst ein entsprechender QName definiert:

²⁷ DE KEYSER (2012, S.147)

²⁸ DUBLIN CORE METADATA INITIATIVE (2016a)

²⁹ URL: <http://dublincore.org/documents/dces/>; zuletzt geprüft am 11.03.2016

³⁰ WOOD (2014, S. 259)

³¹ WOOD (2014, S. 259)

³² URL: <http://dublincore.org/documents/dcmi-terms/>; zuletzt geprüft am 11.03.2016

³³ ISO INTERNATIONALE ORGANISATION FÜR NORMUNG (2009)

```
@prefix dc: <http://purl.org/dc/elements/1.1/> .
```

```
@prefix dct: <http://purl.org/dc/terms/> .
```

Man beachte die unterschiedlichen URIs für die Verwendung des *DC Metadata Element Set* und des *DC Metadata Terms*. Die eigentlichen Metadaten lassen sich in SKOS an geeignete Subjekte anhängen:

```
@prefix ex: <http://www.example.com/>
```

```
ex:bird          rdf:type          skos:concept ;
                  skos:prefLabel    "birds"@en ;
                  dc:creator         "A. Birdman"
                  dc:created         "01-01-1999"
```

Eine genaue Beschreibung der Verwendung von Dublin Core Metadaten und eine Auflistung der verfügbaren Elemente beider Sets findet sich im Wiki der DCMI³⁴.

³⁴ DUBLIN CORE METADATA INITIATIVE (2016b)

3 Thesauri

Diese Arbeit beschäftigt sich mit in SKOS-kodierten Thesauri. Um die nötigen Hintergrundinformationen zu vervollständigen, folgt an dieser Stelle nach der Beschreibung von SKOS und seiner Funktionsweise, ein Kapitel über den Thesaurus. Ein Thesaurus kann auf folgende Art definiert werden:

„Ein Thesaurus im Bereich der Information und Dokumentation ist eine geordnete Zusammenstellung von Begriffen und ihren (vorwiegend natürlichsprachigen) Bezeichnungen, die in einem Dokumentationsgebiet zum Indexieren, Speichern und Wiederauffinden dient.“¹

Ein Thesaurus sammelt also Begrifflichkeiten mit ihren Bezeichnungen. Diese Begriffe werden in einem Thesaurus in Relation zueinander gesetzt. Der Thesaurus legt fest welche Begriffe einander über- oder untergeordnet sind, welche Begriffe verwandt sind und welche Benennungen für einzelne Begriffe zu bevorzugen oder zu vernachlässigen sind. Im Folgenden soll die Funktionsweise eines Thesaurus kurz dargestellt werden. Eine ausführliche Beschreibung soll an dieser Stelle entfallen, da sie den Rahmen dieser Arbeit übersteigen würde. Die Beschreibung geht nur so weit, wie es für das Verständnis von SKOS-Thesauri nötig ist. In diesem Kapitel sollen auch die Normen beschrieben werden, die für Thesauri galten und gelten.

3.1 Funktionsweise

Ein Thesaurus geht zur Beschreibung seiner Begriffe entweder vom Term oder vom Konzept als Grundeinheit aus. Dabei meint Term an dieser Stelle die Bezeichnung des Begriffs und Konzept geht einen Schritt weiter und meint die *Idee* hinter dem Begriff. Während moderne Thesauri, nach den neuesten Normen und auch Thesauri in SKOS vom Konzept ausgehen, gehen die meisten klassischen Thesauri vom Term aus. Terme werden in Deskriptoren und Nicht-Deskriptoren unterschieden. Deskriptor meint dabei eine Bezeichnung, die für den jeweiligen Begriff zu bevorzugen ist und für beispielsweise die Verschlagwortung verwendet werden kann. Nicht-Deskriptoren sind demnach Bezeichnungen die nicht zu verwenden sind. Es handelt sich dabei meist um Synonyme, umgangssprachliche Versionen oder ähnliche Varianten eines Deskriptors. Der Thesaurus sammelt alle Deskriptoren und Nicht-Deskriptoren des behandelten Gebiete und stellt sie in Relation zueinander. Dabei werden hierarchische Relationen,

¹ KUHLEN ET AL. (2004, S. 209)

Tabelle 3.1: Thesaurus-Kürzel nach ISO 25964-1

Kürzel	Relation	
TT	Top Term	Term, der in hierarchischer Ordnung an erster Stelle steht
BT	Broader Term	Ein übergeordneter Term
NT	Narrower Term	Ein untergeordneter Term
RT	Related Term	Ein verwandter Begriff
BTG	Broader Term generic	Ein generischer übergeordneter Term
NTG	Narrower Term generic	Ein generischer untergeordneter Term
BTP	Broader Term partitive	Ein partitativer übergeordneter Begriff
NTP	Narrower Term partitive	Ein partitiver untergeordneter Begriff
USE	Use	Bezeichnet den Deskriptor
UF	Use For	Bezeichnet Synonyme, bzw. nicht-Deskriptoren
UFC	Used for Combination	Bezeichnet ein kombiniertes Synonym

wie über- und untergeordnete Begriffe, assoziative Relationen, wie verwandte Begriffe und Äquivalenz-Relationen, wie Synonyme, beschrieben. Die Relationen in einem konventionellen Thesaurus werden üblicherweise mit Kürzeln bezeichnet. Hier die Kürzel nach ISO 25964-1 in ihrer englischsprachigen Variante: Ein Beispiel-Deskriptor:

Poultry

UF Domesticated Birds

NT chickens, ducks, geese

BT domesticated animals

RT eggs, birds

Dieser Eintrag sagt aus, dass es sich bei der zu verwendenden Benennung um *Poultry* handelt. Das Synonym *Domesticated Birds*, ist ein Nicht-Deskriptor und steht deshalb in **UF** Relation zu *Poultry*. *Chickens, Ducks und Geese* sind spezifischere Bezeichnungen (also untergeordnete Begriffe) und *Domesticated Animals* ist die übergeordnete, weniger spezifische Bezeichnung. *Eggs und Birds* sind Begriffe, die in einem assoziativen Verwandtschaftsverhältnis zu *Poultry* stehen. Der Beispiel-Deskriptor enthält nicht alle in 3.1 aufgezählten Relationen. Nicht jeder Thesaurus verwendet jede mögliche Relation, die Auswahl findet je nach Verwendungszweck des Thesaurus und der gewünschten Komplexität der beschriebenen Relationen statt.

3.2 Normen

Für die Erstellung und Verwendung von Thesauri gibt es verschiedene ISO-Normen. Die erste Norm für Thesauri war ISO 2788², die einsprachige Thesauri für das Information retrieval beschrieb. Die Norm wurde 1974 von der International Organization for Standardization (ISO) erstveröffentlicht und 1986 überarbeitet. 1985 erschien zusätzlich die ISO 5964³, die multilinguale Thesauri behandelt. Beide Normen gehen vom Term als Grundeinheit eines Thesaurus aus. Beide Normen wurden 2011 von der ISO 25964 abgelöst, die in zwei Teilen erschien. ISO 25964-1⁴ beschreibt die Erstellung, Darstellung und Weiterentwicklung mono- und multilingualer Thesauri. Zusätzlich enthält die Norm ein Datenmodell und ein XML-Schema, die Hilfestellung bei der Handhabung von Thesaurus-Daten geben sollen, vor Allem wenn es um den Austausch zwischen verschiedenen Thesauri geht. Der zweite Teil der Norm, ISO 25964-2⁵, beschäftigt sich mit der Vernetzung von Thesauri untereinander und mit anderen KOSs. Dazu gehören auch Empfehlungen für das Mapping von Termen in verschiedenen Systemen. ISO 25964 geht, wie SKOS, statt vom Term vom Konzept als Grundeinheit eines Thesaurus aus.

Im nächsten Abschnitt werden SKOS und Thesauri, nachdem sie einzeln beschrieben wurden, zusammengefügt indem der Vorgang des kodierens eines Thesaurus in SKOS beschrieben wird.

² ISO INTERNATIONALE ORGANISATION FÜR NORMUNG (1986)

³ ISO INTERNATIONALE ORGANISATION FÜR NORMUNG (1985)

⁴ ISO INTERNATIONALE ORGANISATION FÜR NORMUNG (2011)

⁵ ISO INTERNATIONALE ORGANISATION FÜR NORMUNG (2013)

4 Kodierung von SKOS Thesauri am Beispiel des Thesaurus Sozialwissenschaften

Soll ein konventioneller Thesaurus in SKOS dargestellt, oder kodiert werden, müssen die vorhanden Konzepte, Terme und Relationen des Thesaurus in ihre SKOS-Entsprechungen konvertiert werden. Dieser Konvertierungsvorgang wird in diesem Kapitel, anhand des Beispiels Thesaurus Sozialwissenschaften (TheSOZ), näher beschrieben. Die Verwendung des Thesaurus Sozialwissenschaften als Beispiele ergibt sich aus der Tatsache, dass die Konvertierung sowohl bei van Assem¹, als auch vom Leibniz Institut für Sozialwissenschaften (GESIS) selbst dokumentiert ist². Bei der Beschreibung des Vorgangs kommen auch die Probleme und Herausforderungen zur Sprache die bei der Konvertierung auftreten können.

Grundsätzlich besteht der Vorgang der Konvertierung eines Thesaurus in SKOS nach van Assem et al aus drei grundsätzlichen Schritten³:

1. Analyse des Thesaurus
2. Mapping der Teile des Thesaurus zu ihren SKOS-Klassen
3. Technische Konvertierung

An dieser Stelle soll der Konvertierungsvorgang, anhand der ersten beiden Schritte beschrieben werden. Der dritte Schritt, die technische Konvertierung soll an dieser Stelle nicht weiter behandelt werden, da sie für das Thema der Arbeit keine große Relevanz hat.

4.1 Analyse des Thesaurus

Als erster Schritt des Konvertierungsvorgangs steht die Analyse des zu kodierenden Thesaurus. Dabei werden zunächst Struktur und Aufbau des Thesaurus betrachtet. Dazu gehört, welche Relationen im Thesaurus verwendet werden und ob die Terme des Thesaurus außer den üblichen Relationen in Klassifikationen oder Kategorien organisiert sind⁴. Betrachtet wird hier auch ob der Thesaurus gängigen

¹ VAN ASSEM ET AL. (2006, S. 5)

² ZAPILKO (2013); ZAPILKO & SURE (2009)

³ VAN ASSEM ET AL. (2006, S. 5)

⁴ MAYR (2010, S. 3)

Thesaurus-Normen entspricht oder nicht-standardisierte Relationen enthält die in der Norm nicht vorgesehen sind. Auch ob der Thesaurus schon von Konzepten als Grundeinheit ausgeht, wie SKOS und die aktuelle Thesaurus-Norm ISO 25964, oder noch von Termen, wie in den abgelösten Normen ISO 2788 und 5964, muss festgestellt werden, da dies später große Auswirkungen auf das Mapping der Thesaurus-Klassen hat.

Der Thesaurus Sozialwissenschaften (TheSOZ) hat nach dieser Analyse ca. 12.000 Terme⁵, davon 8.000 Deskriptoren und 4.000 nicht-Deskriptoren aus allen Bereichen der Sozialwissenschaften und aus einigen verwandten Gebieten⁶. Diese liegen in Deutsch, Englisch und Französisch vor. Zusätzlich sind die Begriffe des Thesaurus noch in einer eigenen Klassifikation organisiert und jeweils mit einer entsprechenden Notation versehen.

In den meisten Relationen entspricht der Thesaurus den Normen ISO 2788 und ISO 5964. Er enthält allerdings auch Relationen die nicht dem Standard entsprechen. Dazu gehört beispielsweise der sogenannte *Alternative nicht-Deskriptor (AD)*. Der AD ist eine eigene Relation des TheSOZ, bei der ein Nicht-Deskriptor verschiedene gleichwertige *USE*- oder *USE Combination*-Relationen zu verschiedenen Deskriptoren hat⁷. Welcher Deskriptor jeweils verwendet wird richtet sich nach dem Kontext. Ein Beispiel nach Mayr:

„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.“⁸

Spezielle Relationen wie diese sind bei der Analyse des Thesaurus zu beachten, da im nächsten Schritt eine Möglichkeit gefunden werden sollte um diese Relationen in SKOS abzubilden um einen Informationsverlust gegenüber dem konventionellen Thesaurus zu vermeiden.

Wenn durch die Analyse festgestellt wurde, welche Eigenheiten der Thesaurus hat, die zu Problemen bei der Kodierung führen könnten, werden anschließend unter Berücksichtigung dieser Eigenheiten die Klassen des Thesaurus in ihre SKOS-Entsprechungen konvertiert.

⁵ Stand 2013

⁶ ZAPILKO (2013, S. 1ff)

⁷ MAYR (2010, S. 4)

⁸ MAYR (2010, S. 4)

4.2 Mapping von Thesaurus- und SKOS-Klassen

Im zweiten Schritt der Konvertierung werden die Klassen des Thesaurus mit ihren SKOS-Entsprechungen gemappt. Mapping meint in diesem Fall, dass in Beziehung setzen zweier unterschiedlicher Systeme, bzw. das verbinden ihrer Entsprechungen. In den meisten Fällen sind diese Entsprechung eindeutig. Die Beziehung *Broader Term (BT)* des Thesaurus ist beispielsweise äquivalent zu *skos:broader*, die Klasse *Related Term (RT)* zu *skos:related*, usw⁹.

Ein Beispielbegriff aus dem Agrovoc¹⁰, einmal in einer konventionellen Version, einmal in SKOS:

Konventionell:

Poultry¹¹

UF domesticated birds

NT chickens, ducks, geese

BT domesticated animals

RT eggs, birds

SKOS-Version:

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
```

```
@prefix ex: <http://www.example.com/>.
```

```
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
```

```
ex:123 rdf:type          skos:Concept ;
      skos:prefLabel    "poultry" @en ;
      skos:altLabel     "domesticated birds" @en ;
      skos:narrower     "chickens" @en ;
      skos:narrower     "ducks" @en ;
      skos:narrower     "geese" @en ;
      skos:broader      "domestic animals" @en ;
      skos:related      "eggs" @en ;
      skos:related      "birds" @en .
```

Im Beispiel ist sichtbar, dass aus dem Term `Poultry` das `skos:concept ex:123` mit dem Literal `Poultry` geworden ist. Die URI `ex:123` für die Ressource ist dabei in diesem Beispiel zufällig gewählt. Wie die

⁹ MAYR (2010, S. 4)

¹⁰ Anmerkung: Der verwendete Begriff stammt aus dem Agrovoc, das Beispiel verwendet aber nicht die SKOS-Konstruktion des Agrovoc-Thesaurus

¹¹ Anm.:Relationskürzel nach ISO 25964; Beispiel aus Agrovoc Thesaurus

URI einer Ressource gebildet wird, ist unterschiedlich und wird in jedem KOS neu festgelegt. Verwendet wird beispielsweise eine einfache Durchnummerierung der Ressourcen oder eine andere Benennung nach einem bestimmten Schema. Die Relationen des Terms sind durch ihre SKOS-Entsprechungen ersetzt worden, wobei die Terme aus dem konventionellen Thesaurus zu Literalen der entsprechenden SKOS-Prädikate geworden sind.

Ein Problemfall, der an dieser Stelle demonstrierbar ist, ist das SKOS von Konzepten als Grundeinheit ausgeht, der konventionelle Thesaurus allerdings von Termen. Es kann also sein, dass im konventionellen Thesaurus sowohl der Deskriptor *Poultry*, als auch der nicht-Deskriptor *Domesticated Birds* als eigenständige Terme auftauchen können. Beide Terme können dabei ihre eigenen Relationen haben. In der SKOS-Variante taucht der nicht-Deskriptor *Domesticated Birds* nicht mehr als eigenständiger Term auf, sondern geht im `skos:concept` des Deskriptors als `skos:altLabel` auf. Er kann selbst keine Relation mehr haben, da er nicht das Subjekt eines Tripels sein kann. Der Unterschied zwischen Termen und Konzepten verursacht also Probleme einige Relationen des Thesaurus korrekt abzubilden¹².

Zu diesen schwer abzubildenden Relationen gehört auch der in Kap. 5.1 beschriebene *alternative nicht-Deskriptor* des TheSOZ. Um diese Probleme zu lösen wird SKOS-XL (siehe Kap. 2.5) verwendet. Im TheSoz wurden unter Verwendung von SKOS-XL eigene Extensions definiert um die Relationen des AD und andere problematische Fälle ohne Informationsverlust abbilden zu können¹³.

Extension	Description
<code>thesoz:Descriptor</code>	Descriptors of the TheSoz, which are defined as subclasses of "skos:Concept".
<code>thesoz:Classification</code>	Notation of the classification hierarchy of the TheSoz, which is defined as a subclass of "skos:Concept".
<code>thesoz:EquivalenceRelationship</code>	An equivalence relationship between two terms, where the terms are assigned via "thesoz:use" and "thesoz:usedFor" properties. This is a subclass of "skosxl:Label".
<code>thesoz:CompoundEquivalence</code>	A compound equivalence between terms. For constructing "use combination" and "used for combination" relations between terms. The non-preferred term is assigned by the "thesoz:compoundNonPreferredTerm" property, the preferred terms by the "thesoz:preferredTermComponent" property. This is a subclass of "skosxl:Label".
<code>thesoz:use</code>	Use relation, which is defined as a subproperty of "skosxl:labelRelation".
<code>thesoz:usedFor</code>	Used for relation, which is defined as a subproperty of "skosxl:labelRelation".
<code>thesoz:preferredTermComponent</code>	A preferred term as a component for a "use combination" and "used for combination" relation. This property is defined as a subproperty of "skosxl:labelRelation".
<code>thesoz:compoundNonPreferredTerm</code>	The non-preferred term as a component for a "use combination" and "used for combination" relation. This property is defined as a subproperty of "skosxl:labelRelation".
<code>thesoz:isPartOfEquivalenceRelationship</code>	Relation from a term to the class "thesoz:EquivalenceRelationship".
<code>thesoz:isPartOfCompoundEquivalence</code>	Relation from a term to the class "thesoz:CompoundEquivalence".
<code>thesoz:hasTranslation</code>	Relation between different languages of a term, which is defined as a subproperty of "skosxl:labelRelation".
<code>thesoz:isTranslationOf</code>	Inverse property of "thesoz:hasTranslation".

Tabelle 4.1: Übersicht über selbst-definierte Erweiterungen des TheSOZ

¹² MAYR (2010, S. 4)

¹³ ZAPILKO (2013, S. 3)

Zunächst wurden alle Terme des Thesaurus innerhalb der `skos:concept`s mit der SKOS-XL Klasse `skosxl:Label` modelliert und danach die Klasse `skosxl:labelRelation` verwendet um Relationen zwischen ihnen darzustellen. Dafür wurden Instanzen von `skosxl:labelRelation` als verschiedene Relationen deklariert und mit eigenen Kürzeln versehen. Eine Darstellung aller dabei entstandenen Extensions findet sich nach Zapilko¹⁴ in Tabelle 4.1.. Es lässt sich unschwer erkennen, dass eine große Anzahl von eigenen Extensions nötig war um die Relationen des Thesaurus korrekt abzubilden. Weitere Informationen zu der Modellierung des TheSOZ finden sich in Kap. 5.4. Wie sich die Extensions auf die Kompatibilität des Thesaurus zu anderen SKOS-Thesauri auswirkt ist Thema in Kap. 6.

Im nächsten Abschnitt folgt eine detaillierte Beschreibung von 5 Beispielthesauri, anhand derer in Kap. 6 die Auswirkungen verschiedener SKOS-Modellierungen betrachtet werden.

¹⁴ ZAPILKO (2013, S. 3)

5 Beschreibung der Beispielthesauri

Im Folgenden werden die fünf Thesauri vorgestellt, die beispielhaft auf ihre unterschiedliche Modellierung, ihre selbst-definierten Extensions und die Auswirkungen dieser Eigenheiten auf die Kompatibilität mit anderen SKOS-Thesauri hin untersucht werden sollen. Dazu werden zunächst einige allgemeine Informationen über den Thesaurus angegeben, wie die Menge der vorhandenen Deskriptoren, die verfügbaren Sprachen, die behandelten Fachgebiete oder in welchen Formaten die Konzepte des Thesaurus angezeigt bzw. heruntergeladen werden können. Danach wird die Modellierung des Thesaurus beschrieben. Dazu gehören Struktur, verwendete Standards, Umgang mit Metadaten und die für den Thesaurus definierten Klassen. Zusätzlich enthält jede Thesaurus Beschreibung ein Schaubild, welches die Struktur des Thesaurus in grafischer Form erläutert. Der Aufbau der Schaubilder wird hier nun kurz beschrieben.

Beschreibung der Schaubilder

Die Schaubilder stellen den Aufbau des Thesaurus und die Verbindungen zwischen seinen einzelnen Teilen dar. Dabei stellt jeder Kasten entweder eine Ressource oder eine Klasse bzw. Prädikat dar. Diese Klassen und Prädikate stammen nicht nur aus SKOS, sondern auch aus den anderen im Thesaurus verwendeten Vokabularen. Die verwendeten Vokabulare werden im Schaubild mit ihren Präfixen aufgelistet. Zusätzlich sind die Vokabulare farblich voneinander abgesetzt: selbst-definierte Vokabulare werden in Orange, Dublin Core Elemente blau und alle Standard-Vokabulare in schwarz dargestellt. Ressourcen werden in den Schaubildern fett gedruckt, die Klassen und Prädikate in normaler Schrift und Literale, wie Strings, kursiv. Die Verbindungen zwischen den einzelnen Teilen des Thesaurus werden in Form von Linien zwischen ihnen dargestellt.

5.1 Agrovoc

Der Agrovoc ist ein Thesaurus der Food and Agriculture Organization (FAO) der Vereinten Nationen, der die Bereiche Landwirtschaft, Forstwirtschaft, Ernährung, Umwelt und artverwandte Bereiche abdeckt. Der Thesaurus wurde zu Beginn der 80er Jahre erstveröffentlicht, zunächst in Englisch, Spanisch und Französisch, und sollte als Hilfe bei der Indexierung von Veröffentlichungen in den abgedeckten Bereichen dienen. Im Jahr 2000 wurde die Print-Version des Thesaurus durch eine Online-Variante ersetzt und seit 2009 steht der Thesaurus in SKOS zur Verfügung.

URL: <http://aims.fao.org/standards/agrovoc>

Deskriptoren / Nicht-Deskriptoren: >32.000 Konzepte

Sprachen: 23 Sprachen

Features: Suche, Hierarchisches Browsen, Download in RDF oder als LOD-Set, Vorschlag von neuen Termen

Fachgebiete: Landwirtschaft, Forstwirtschaft, Fischerei, Nahrung und Ernährung, Umwelt, u.a.

Formate: RDF/XML, Turtle, JSON-LD

Mappings: ASFA, Biotechnology Glossary (FAO), Chinese Agriculture Thesaurus (CAT), DBpedia, Dewey Decimal Classification (DDC), EUROVOC, GEMET, GeoNames, Geopolical Ontology, Library of Congress Subject Headings (LCSH), NAL Thesaurus, RAMEAU, STW - Thesaurus for Economics, TheSoz - Thesaurus for the Social Sciences, SWD (Schlagwortnormdatei), EARTH

Klassifikation: AgrovocCode als Notation einer mehrstufigen Klassifikation. Nicht in SKOS abgebildet

Verwendete Standards: SKOS, SKOS-XL, DC

Der Agrovoc Thesaurus ist ein `skos:conceptScheme`, mit der URI `http://aims.fao.org/aos/agrovoc` (Kürzel: `agrovoc`). Den Einstieg in den Thesaurus bieten eine Reihe von Top Concepts, also Konzepten die keine übergeordneten Konzepte haben. Diese sind mit `skos:hasTopConcept` direkt an den Thesaurus gehängt. Alle Konzepte sind nach dem selben Muster benannt und als `skos:concept` modelliert. Die Terme der Konzepte werden mit `skos:prefLabel` und `skos:altLabel` modelliert und enthalten

einen String + Sprachcode für jede Sprachvariante. Dazu hat jedes Konzept ein `skosxl:prefLabel` und ein `skosxl:altLabel`. Beide führen für jede Sprache zu einer Ressource, die nach dem Muster `agrovoc:xl\sprache\Nummer` bezeichnet ist. Jedes dieser Labels enthält `skos:notation` und ein Literal vom Datentyp Agrovoc Code. Der Agrovoc Code entspricht bei Instanzen von `skosxl:prefLabel` der Nummer des Konzepts (entspricht der Nummer in der Konzept URI). Im Beispiel ist gut sichtbar, dass das `skosxl:prefLabel` `agrovoc:xl_en_1299491039255` als Notation die Nummer des Konzepts `agrovoc:c_8313` hat.

```
agrovoc:c_8313
    a                skos:Concept ;
    skos:altLabel    "Carabao"@de,
    skosxl:prefLabel agrovoc:xl_en_1299491037411,
    skosxl:altLabel  agrovoc:xl_en_1299491039255,
    skos:prefLabel   "Wasserbüffel"@de
    skos:broader     agrovoc:c_25410 ;
    skos:inScheme    <http://aims.fao.org/aos/agrovoc> ;
```

```
agrovoc:xl_en_1299491039255
    skos:notation    "6614"^^agrovoc:AgrovocCode ;
    a                skosxl:Label .
```

```
agrovoc:xl_en_1299491037411
    skos:notation    "8313"^^agrovoc:AgrovocCode ;
    a                skosxl:Label .
```

Hierarchische Relationen zwischen den Konzepten werden mit den Standard Relationen `skos:broader` und `skos:narrower` modelliert. Nicht-hierarchische Relationen werden entweder mit `skos:related` oder mit der für den Agrovoc erstellten *Agrontology*¹ mit dem Kürzel `ns0:` dargestellt. Die *Agrontology* umfasst mehr als 200 speziell für den Agrovoc definierte Relationen. Alle diese Relationen hier aufzuführen würde den Rahmen dieses Textes sprengen, darum folgen an dieser Stelle nur einige Beispiele. Die Relationen der *Agrontology* teilen sich in Object, Data und Annotation Properties, sowie Named Individuals auf. Sie decken eine Vielzahl von sehr spezialisierten Relationen ab, die für die vom Thesaurus abgedeckten Fachbereiche relevant sind. Beispielsweise finden sich bei den Object Properties Relationen für die wissenschaftlichen Namen von Pflanzen und Tieren, (`ns0:hasScientificName`). Beispiele für weitere Prädikate sind `ns0:CasautiveRelationship`, das eine Anzahl von Unterklassen bzw. SubProperties für unterschiedliche kausale Zusammenhänge enthält, das Prädikat `ns0:HasSynonym`, mit Varianten

¹ FAO (2012)

für verschiedene Benennungen eines Terms oder die Klasse `ns0:QuantitativeRelationship`, die Klassen und Prädikate für den quantitativen Vergleich verschiedener Konzepte enthält. Im Bereich der Data Properties finden sich beispielsweise Relationen für Singular oder Plural Formen eines Terms oder verschiedene Schreibweisen. Annotation Properties enthält Klassen und Prädikate die für die Annotation eines Konzepts verwendet werden können. Die Klassen und Prädikate in diesem Bereich sind jeweils auch als Object oder Data Property definiert. In der Agrontology sind sehr viele Klassen definiert und sie alle zu beschreiben würde hier zu weit führen.

Als Beispiel hier eine Verwendung des Prädikats `ns0:produces`, das ein Konzept als ein Produkt eines anderen auszeichnet:

```
agrovoc:c_8313
    ns0:produces      agrovoc:c_16076, agrovoc:c_23996 ;
```

Metadaten werden auf zwei Arten angehängt. DC-Klassen werden verwendet um den Zeitpunkt der Erstellung und den Zeitpunkt der letzten Veränderung eines Konzepts zu verzeichnen. Die Prädikate `skos:scopeNote` und `skos:Definition` werden ebenfalls verwendet. Die Definition führt dabei nicht direkt zu einem String, sondern zu einer Ressource die nach dem Muster `agrovoc:xDEFNummer` bezeichnet ist. Erst an diese Ressource wird mit `rdf:value` der String gehängt. Mappings auf Ressourcen in anderen Concept Schemes werden mit den Standard SKOS-Elementen modelliert. Dabei werden die Konzepte, je nach Relationen, mit den Prädikaten `skos:exactMatch`, `skos:closeMatch`, `skos:broadMatch`, `skos:narrowMatch` oder `skos:relatedMatch`, mit der URI der Konzepte in den anderen KOS verbunden. Im Beispiel eine `skos:closeMatch`-Relation zwischen einem Agrovoc-Konzept und einem Konzept aus dem ASFA-Thesaurus².

```
<http://aims.fao.org/aos/asfa/c_8123>      skos:closeMatch      agrovoc:c_49874 .
```

² <http://www4.fao.org/asfa/asfa.htm>

5.2 Eurovoc

Der Eurovoc ist ein multidisziplinärer Thesaurus der Europäischen Union. Die erste Ausgabe des Eurovoc erschien 1984 in zwei gedruckten Bänden, einem alphabetischen und einem thematischen, und in sieben Sprachen. Die erste Online-Version des Thesaurus erschien im Jahr 2000.

URL: <http://eurovoc.europa.eu/>

Deskriptoren / Nicht-Deskriptoren: 6.645 Deskriptoren, zwischen 150 und 13.139 nicht-Deskriptoren je nach Sprache

Sprachen: 23 Amtssprachen der EU + Albanisch und Serbisch

Features: Suche, erweiterte Suche, Browsen nach Thema.

Download des Thesaurus nach Bereichen (pdf), als alphabetische Liste nach Bereich, mehrsprachige Liste nach Bereich, alphabetischer Index

Fachgebiete: Politisches Leben, Internationale Beziehungen, Europäische Gemeinschaften, Recht, Wirtschaftsleben, Wirtschafts- und Handelsverkehr, Finanzwesen, Soziale Fragen, Bildung und Kommunikation, Wissenschaften, Unternehmen und Wettbewerb, Beschäftigung und Arbeit, Verkehr, Umwelt, Land- und Forstwirtschaft, Ernährung, Produktion, Technologie und Forschung, Energie, Industrie, Geografie, Internationale Organisation

Formate: RDF/XML, Turtle

Mappings: Agrovoc, GEMET, ECLAS, European Commissions Libraries Catalogue

Klassifikation: 2-stufige hierarchische Klassifikation mit 21 Bereichen (2-stellige Zahl) und 127 Mikrothesauri (4-stellige Zahl)

Verwendete Standards: SKOS, SKOS-XL, Dublin Core, OWL, XML, RDF, FOAF

Im Eurovoc werden viele eigene Klassen und Prädikate definiert. Für diese Klassen wird der Namespace <http://eurovoc.europa.eu/schema#> mit dem Kürzel `eu:` verwendet. Die oberste Klasse bildet `eu:Thesaurus`, eine `rdfs:subClassOf` von `skos:ConceptScheme`. `eu:Eurovoc` ist eine Instanz dieser Klasse. An `eu:Eurovoc` wird mit `eu:language` eine Liste von Sprachen angehängt, die die Sprachen repräsentieren in denen ein Konzept eine Vorzugsbenennung hat. Jede Instanz von `eu:language`

ist eine mit `rdfs:label` bezeichnete Sprache. Wie oben erwähnt hat der Eurovoc eine zweistufige Klassifikation mit 21 Bereichen (Domains) und 127 Mikrothesauri. Um diese Klassifikation abzubilden wurden `eu:Domain` und `eu:Microthesaurus` definiert. Beide Klassen sind Unterklassen von `skos:conceptScheme`. Jede Domain und jeder Mikrothesaurus werden mit einem `skos:prefLabel` für jede Sprache benannt. Zusätzlich zum Namen wird im Literal auch immer die Notation der Klassifikation mit aufgenommen. Die Klasse `dc:identifier` wird zusätzlich verwendet um die Notation noch einmal einzeln aufzuführen. In der Notation werden 2-stellige Zahlen für Domains und 4-stellige Zahlen für Mikrothesauri verwendet. Die Mikrothesauri enthalten zusätzlich `skos:hasTopConcept`. Hier werden die Konzepte innerhalb des Mikrothesaurus aufgenommen die keine übergeordneten Begriffe haben. Jeder Mikrothesaurus ist jeweils in einer `eu:Domain`, um Polyhierarchie zu vermeiden.

```
<[...]/100180>      a          eu:Microthesaurus ;
                   dc:identifier    "1216" ;
                   eu:domain        <http://eurovoc.europa.eu/100145> ;
                   skos:prefLabel   "1216 Strafrecht"@de,[...];
                   skos:hasTopConcept <http://eurovoc.europa.eu/3513> ,[...].
```

Die einzelnen Konzepte des EuroVoc werden mit der Klasse `eu:ThesaurusConcept`, einer Subclass von `skos:concept`, dargestellt. Unter Verwendung von `skos:inScheme` ist jedes dieser Konzepte Mitglied von `eu:EuroVoc` und dem jeweiligen `eu:Microthesaurus`. Für Relationen zwischen Konzepten werden `skos:broader`, `skos:broaderTransitive` und `skos:related` verwendet. Mit seinen Termen wird das Konzept mit `xl:prefLabel` oder `xl:altLabel` verbunden³.

```
<[...]/1251>      a          eu:ThesaurusConcept ;
                   xl:prefLabel     <http://eurovoc.europa.eu/167690> ;
                   skos:broader      <http://eurovoc.europa.eu/3943> ;
                   skos:broaderTransitive <http://eurovoc.europa.eu/1432> ;
```

```
<[...]/167690>   a          eu:PreferredTerm ;
                   xl:literalForm   "Tötung"@de ;
                   skos:inScheme     <http://eurovoc.europa.eu/100180> ;
                   xl:altLabel       [:::].
```

Die Terme sind Instanzen von `eu:ThesaurusTerm`, deklariert als Unterklasse von `xl:Label`, und sind entweder Vorzugs- oder Nicht-Vorzugsbenennungen. Vorzugsbenennungen werden als `eu:PreferredTerm` und nicht-Vorzugsbenennungen mit `eu:SimpleNonPreferredTerm` oder `eu:CompoundNonPreferredTerm` modelliert. Benannt werden sie mit `xl:literalForm`. Die Relationen zwischen den Termen verwenden ebenfalls nicht die normalen SKOS-Klassen sondern `eu:EquivalenceRelationship` oder `eu:CompoundEquivalence`.

³ Anmerkung: Der Eurovoc verwendet als Kürzel für SKOS-XL statt des üblichen `skosxl:` die Variante `xl:`

Letztere Klasse teilt sich dabei nochmal in `eu:compoundNonPreferredTerm` und `eu:preferredTermComponent` . Die Äquivalenz-Relationen sind jeweils auf Term-Level modelliert, d.h. es werden keine Relationen zwischen Konzepten, also `eu:ThesaurusConcept` modelliert. Es wurden unter Verwendung von `xl:labelRelation` weitere Klassen und Prädikate modelliert. Beispielsweise werden Abkürzungen von Thesaurus-Termen mit der ausgeschriebenen Version durch die Klassen `eu:acronym` , bzw. `eu:fullName` verbunden. Schließlich wird noch `eu:translationOf` verwendet um Übersetzung eines Terms anzugeben. Für die Dokumentation der Thesaurus-Begriffe werden die Standard Varianten von `skos:note` verwendet: `skos:scopeNote` , `skos:historyNote` und `skos:definition` . Diese Klassen werden nicht in jedem Konzept verwendet, sondern nur wenn entsprechende Informationen vorhanden sind.

Mappings auf andere Thesauri werden mit den Standard SKOS-Mapping-Relationen dargestellt und hängen am Konzept. Dies findet auf die gleiche Art und Weise statt wie im Agrovoc.

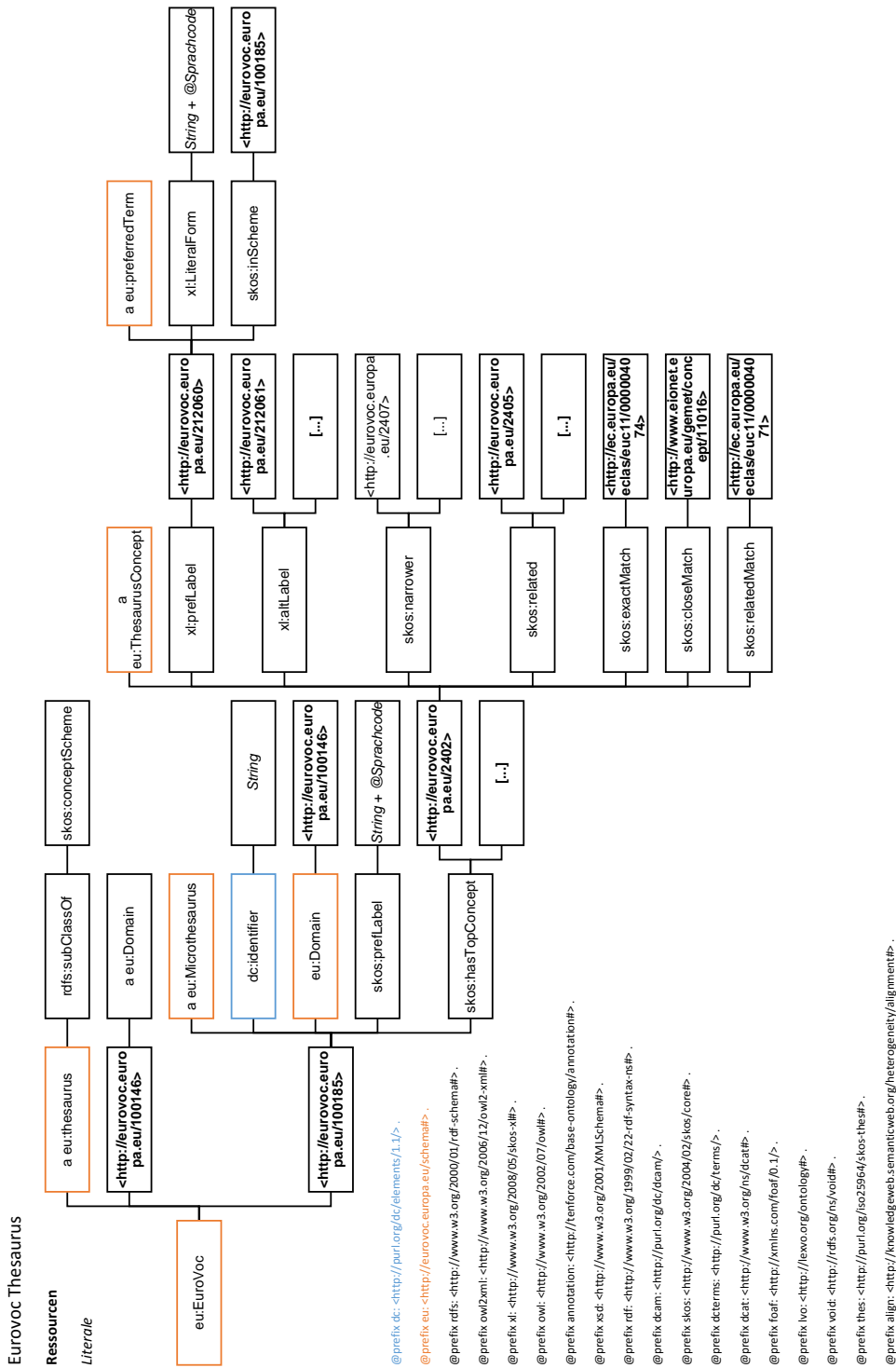


Abbildung 5.2: Struktur des Eurovoc Thesaurus

5.3 Standard Thesaurus Wirtschaft

Der Standard Thesaurus Wirtschaft (STW) ist ein Thesaurus des Leibniz-Informationszentrum Wirtschaft (ZBW) und soll die Suche im EconBiz Fachportal für Wirtschaft⁴ erleichtern.

URL: <http://zbw.eu/stw/version/latest/about>

Deskriptoren / Nicht-Deskriptoren: 6.000 / 20.000

Sprachen: Deutsch und Englisch

Features: Thematisches Browsen in den Konzepten, Suche, Download in RDF/XML und Turtle, Mailingliste mit aktuellen Updates, Download von STW und Mappings

Fachgebiete: Volkswirtschaft (V), Betriebswirtschaft (B), Wirtschaftssektoren und spezielle Wirtschaftslehren (W), Produkte(P), Nachbarwissenschaften (N), Geographische Begriffe (G), Allgemeinwörter (A)

Formate: RDF/XML, Turtle

Mappings: Gemeinsame Normdatei (GND), DBPedia, Thesaurus Sozialwissenschaften, Agrovoc, WKD Arbeitsrechtsthesaurus

Klassifikation: 7 Subthesauri (Anfangsbuchstabe der Domain), 2 weitere Gliederungsstufen

Verwendete Standards: SKOS, DC Terms & Elements, FOAF, GBV-Ontologie, Creative Commons Rights Expression Language

Der STW unterteilt sich zunächst in 7 Subthesauri. Bereich A beinhaltet fachlich unspezifische Allgemeinwörter. Die Bereiche P und W lehnen sich an die Güter- und Branchensystematik des Statistischen Bundesamtes an. Der Geografische Teil G, richtet sich nach Kontinenten und Die Bereiche B und V richten sich nach den fachlich üblichen Einteilungskriterien. Bereich N beinhaltet sinnvolle Begriffe aus Nachbarwissenschaften. Der Thesaurus als ganzes ist als `skos:conceptScheme` definiert und hat ein `skos:prefLabel` und die 7 Subthesauri in `skos:hasTopConcept`. Die Subthesauri haben eine URI nach dem Schema `<thsys/Notation>`. Zusätzlich sind Metadaten direkt an den Thesaurus gehängt. DC Elements und Terms Klassen beschreiben Veröffentlichungsdatum, Titel und Creator. Klassen der

⁴ <http://www.econbiz.de/>; zuletzt geprüft am 05.04.2016

Creative Commons Rights Expression Language (cc:) beschreiben die Copyright Regelungen des Thesaurus.

```
<../stw>
  cc:attributionName "ZBW - Leibniz Information Centre for Economics"@en,
  cc:attributionURL <http://zbw.eu> ;
  cc:license <http://opendatacommons.org/licenses/odbl/1-0/> ;
  cc:morePermissions "http://www.zbw.eu/e_imprint.htm"@en;
  dc:creator "Deutsche Zentralbibliothek für Wirtschaftswissenschaften [...]";
  dc:publisher "ZBW - Leibniz Information Centre for Economics"@en;
  dcterms:issued "2015-06-15"^^xsd:date ;
  dcterms:title "STW Thesaurus for Economics"@en;
  a skos:ConceptScheme ;
  rdfs:isDefinedBy stw:about ;
  owl:versionInfo "9.0"@de, "9.0"@en ;
  skos:hasTopConcept <thsys/a>, <thsys/b>, <thsys/g>[...];
  skos:prefLabel "STW Thesaurus for Economics"@en, [...].
```

Die Subthesauri sind als `skos:concept` definiert und mit `skos:prefLabel` benannt. Den Subthesauri untergeordnet sind die spezifischeren Kategorien der STW-Klassifikation⁵. Diese werden ebenfalls wie Konzepte behandelt und stehen unter Verwendung von `skos:narrower` in hierarchischer Beziehung zu den Subthesauri. Die Notation der STW-Klassifikation wird dabei mit `skos:notation` angegeben.

```
<thsys/v>
  a skos:Concept, zbwest:Thsys ;
  rdfs:isDefinedBy <thsys/v/about> ;
  rdfs:label "Economics"@en, "Volkswirtschaft"@de ;
  skos:inScheme <../stw> ;
  skos:narrower <thsys/70021>, <thsys/70039>, [...];
  skos:notation "V"^^xsd:string ;
  skos:prefLabel "V Economics"@en, "V Volkswirtschaft"@de .
```

Die Unterkategorien der Klassifikation sind ähnlich modelliert. Sie enthalten entweder weitere Unterkategorien mit `<thsys/Notation>` oder Deskriptoren mit `<descriptor/Nummer>`. In beiden Fällen wird `skos:narrower` verwandt. Die Deskriptoren schließlich sind ebenfalls als `skos:concept` definiert. Zusätzlich sind sie ein `zbwest:Descriptor`. Die Relationen des Deskriptors sind hauptsächlich mit den Standard SKOS-Klassen modelliert. Dabei werden sowohl hierarchische als auch assoziative Relationen verwendet, die entweder zu weiteren STW-Deskriptoren, zu Systemstellen der Klassifikation oder zu Mappings auf andere Thesauri führen. Die Deskriptoren haben unter Verwendung der GBV-Ontologie⁶ eine Pica Produktionsnummer (PPN) für den Gemeinsamer Verbundkatalog (GVK), wobei diese im GVK

⁵ http://zbw.eu/wikis/wikisaurus/uploads/Main/STW_Klassifikation.doc

⁶ VOB (2014)

selbst nicht auffindbar scheinen. Für die Literatursuche über das EconBiz werden im `foaf:page` und `zbw:indexedItem` die Suchbegriffe (als Ressourcen) aufgenommen mit denen der Deskriptor in der Datenbank gesucht wird.

```
<descriptor/10253-0>
  gbv:gvkppn      "091381304"^^xsd:string ;
  a               skos:Concept, zbwext:Descriptor ;
  rdfs:isDefinedBy <descriptor/10253-0/about> ;
  skos:altLabel   "Macroeconometric model"@en,[...];
  skos:broader    <descriptor/15373-0>, <thsys/70234> [...];
  skos:inScheme   <../stw> ;
  skos:narrowMatch <http://d-nb.info/gnd/4120788-9>,[...];
  skos:prefLabel  "Macroeconometrics"@en,[...] ;
  skos:related    <descriptor/29938-1> ;
  skos:relatedMatch <http://d-nb.info/gnd/4074486-3> ;
  foaf:page       <http://www.econbiz.de/subject/Macroeconometrics/>,[...];
  zbwext:indexedItem <http://www.econbiz.de/subject/Macroeconometrics/>,[...].
```

Mappings auf andere Thesauri finden, wie im Agrovoc oder Eurovoc, mit den Standard SKOS-Mapping-Relationen statt.

5.4 Thesaurus Sozialwissenschaften

Der Thesaurus Sozialwissenschaften (TheSOZ) ist ein Thesaurus des GESIS, der als Tool für die Recherche in den Datenbanken SOFIS und SOLIS dient.

URL: <http://sowiport.gesis.org/Thesaurus> (Interaktive Version im Fachportal Sowiport)

Deskriptoren / Nicht-Deskriptoren: 8.000 / 4.000

Sprachen: Deutsch, Englisch, Französisch

Features: Suche, Systematisch Browsen

Fachgebiete: Sozialwissenschaft und verwandte Sachgebiete

Formate: RDF/XML, N3

Mappings: Standard Thesaurus Wirtschaft, Agrovoc

Klassifikation: 3-stufige Systematik mit 5 Hauptgruppen (0-5)

Verwendete Standards: SKOS, SKOS-XL, DC, Creative Commons Rights Expression Language (CC), Provenance Vocabulary Core Ontology Specification

Der TheSOZ Thesaurus ist mit SKOS und SKOS-XL modelliert. Zusätzlich sind einige eigene Extensions definiert. Der Thesaurus als Ganzes ist als `skos:conceptScheme` definiert, mit `skos:prefLabel` benannt und mit DC Metadaten versehen. Ebenfalls vorhanden sind die schon beim STW verwendeten CC Daten, die die Copyright Situation klären. Zusätzlich wird die *Provenance Vocabulary Core Ontology Specification* (*prv*:)⁷ verwendet, die zur Darstellung der Provenienz von Daten im Web dient. Hiermit werden Metadaten über Autoren, Erstellungsdaten und Ähnliches aufgenommen.

⁷ HARTIG & ZHAO (2012)

```
<../stw>    prv:createdBy    <http://lod.gesis.org/thesoz/Creation> .
```

```
<http://lod.gesis.org/thesoz/Creation>    a    prv:DataCreation ;
    prv:completedAt    "2014-02-25" ;
    prv:performedBy    <http://www.gesis.org/> .
```

Die selben Informationen sind zumindest für den Thesaurus als Ganzes, auch in DC vorhanden. Die Systematik des Thesaurus ist in einer eigenen Klasse `thesoz:Classification` modelliert, die eine `rdfs:subClassOf` von `skos:concept` ist. Die Instanzen der Klasse sind die einzelnen Systemstellen der Klassifikation. Die Unterpunkte sind jeweils ihre eigene Ressource mit einer `skos:notation` und einem `skos:prefLabel`. Systemstellen sind mit ihren übergeordneten Stellen mit `skos:broader` und mit den untergeordneten Stellen mit `skos:narrower` verknüpft. Die Konzepte des Thesaurus sind Instanzen von `thesoz:descriptor` mit Instanzen von `skosxl:prefLabel` für jede Sprache. `thesoz:descriptor` ist eine sub-Class von `skos:concept`. Die einzelnen Sprachversionen sind `skosxl:Label` mit einer `skosxl:literalForm`. Die nicht-deutschen Varianten sind zusätzlich jeweils eine `thesoz:isTranslationOf` des Konzepts. Das Konzept hat eine `skos:notation` und ist zusätzlich als `skos:narrower` der entsprechenden Systemstelle definiert.

```
<[...]/concept/10034303>    a    thesoz:Descriptor , prv:DataItem ;
    skos:inScheme    <http://lod.gesis.org/thesoz/> ;
    prv:containedBy    <http://lod.gesis.org/thesoz/> ;
    prv:createdBy    <http://lod.gesis.org/thesoz/Creation> ;
    skos:related    <[...]/concept/10034304>,[...] ;
    rdfs:label    "Abbrecher"@de ;
    skosxl:prefLabel    <[...]/thesoz/term/10034303>[...];
    skosxl:altLabel    <[...]/thesoz/term/10034307> .
```

```
<[...]/term/10034303>    skosxl:literalForm    "Abbrecher"@de ;
    skos:inScheme    <http://lod.gesis.org/thesoz/> ;
    prv:containedBy    <http://lod.gesis.org/thesoz/> ;
    prv:createdBy    <http://lod.gesis.org/thesoz/Creation> ;
    a    skosxl:Label , prv:DataItem .
```

```
<[...]/concept/10034303>    skos:notation    "3.2.00" ;
    skos:broader    <[...]/classification/3.2.00> .
```

Zusätzlich zu dieser Modellierung gibt es noch weitere für TheSOZ definierte Erweiterungen, die verschiedene Relationen abbilden (siehe Tabelle 4.1). Die Definition dieser Relationen findet nach folgendem Muster statt:

```
thesoz:EquivalenceRelationship rdfs:label      "Equivalence Relationship"@en ;
    skos:definition "An equivalence relationship between two terms"@en ;
    rdfs:isDefinedBy <http://lod.gesis.org/thesoz/ext/thesoz_ext.rdf> ;
    rdfs:subClassOf skosxl:Label .
```

Die Klasse bekommt ein Label zugewiesen und mit `skos:definition` kurz beschrieben. Jede Klasse, bzw. jedes Prädikat, ist dabei als `rdfs:subClassOf` oder `rdfs:subPropertyOf` einer bestehenden Klasse definiert. Häufige Verwendung findet dabei das Prädikat `skosxl:labelRelation` (siehe Kap. 2.5).

Thesaurus Sozialwissenschaften (TheSOZ)

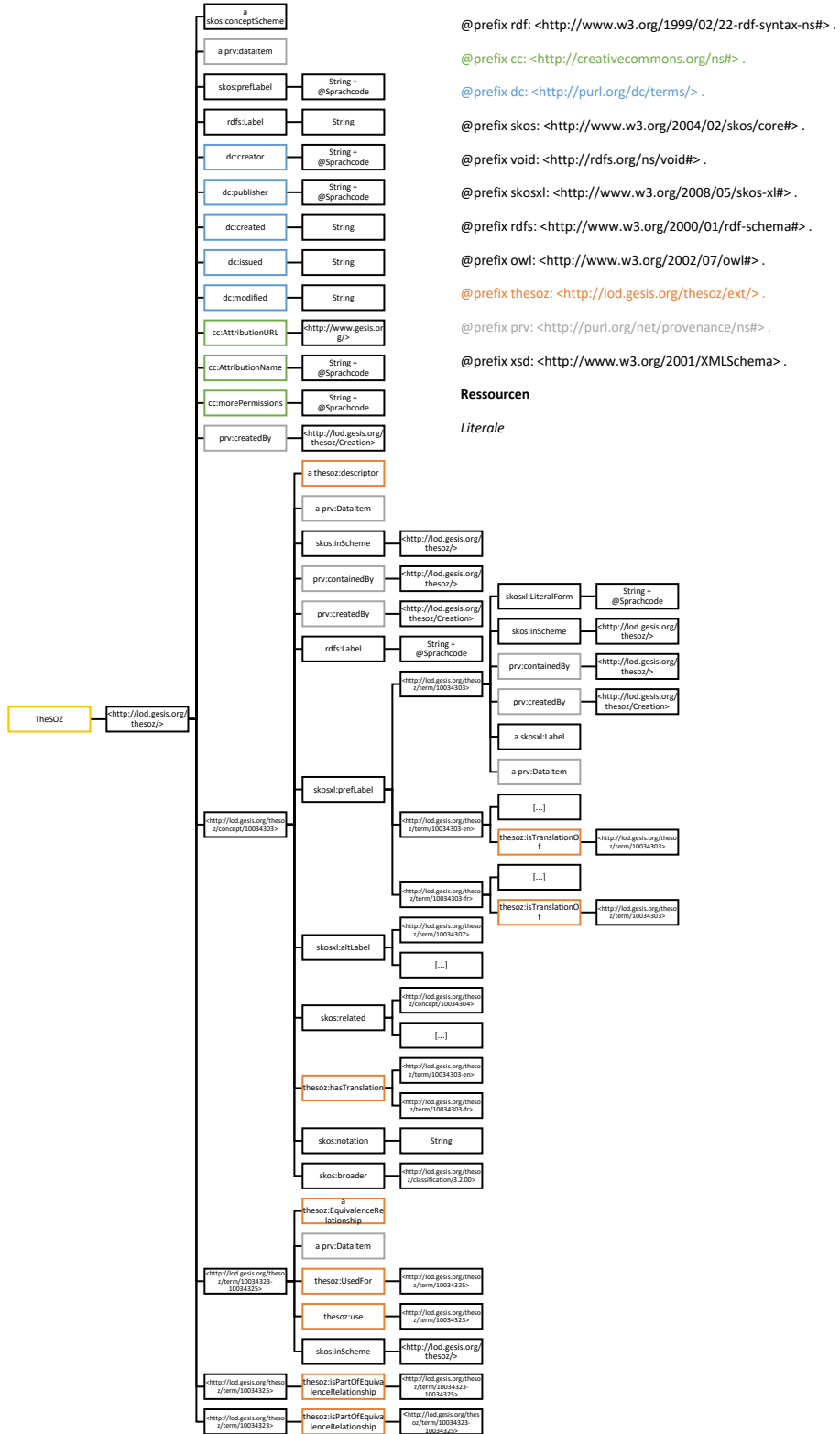


Abbildung 5.4: Struktur des Thesaurus Sozialwissenschaften

5.5 UNESCO Thesaurus

Der UNESCO Thesaurus ist ein in SKOS verfügbarer Thesaurus der von der UNESCO als Hilfsmittel für die inhaltliche Erschließung entwickelt wurde.

URL: <http://skos.um.es/unescothes/>

Deskriptoren / Nicht-Deskriptoren: 4.408 Konzepte mit 17.980 prefLabels und 13.874 altLabels

Sprachen: Englisch, Französisch, Spanisch, Russisch

Features: Suche, Hierarchisches Browsen, Download in RDF/XML und Turtle, SPARQL Endpoint

Fachgebiete: Bildung, Wissenschaft, Kultur, Sozialwissenschaft, Information und Kommunikation, Politik, Recht, Wirtschaft, Länder

Formate: RDF/XML, N3, Turtle, JSON, JSON-LD

Mappings: keine

Klassifikation: 7 Hauptgruppen (1-7), 88 Mikrothesauri (x.05-85)

Verwendete Standards: SKOS, DC, iso-Thes

Grundlegend ist der UNESCO Thesaurus hauptsächlich mit Standard SKOS-Klassen modelliert, ohne Verwendung von SKOS-XL. Er verwendet dazu eine Gruppe von selbst definierten Extensions die im *UNESKOS Vocabulary*⁸ festgelegt sind und einige Klassen aus dem Korrespondenz-Modell zwischen SKOS und ISO 25964 von Isaac und De Smedt⁹ mit dem Namespace `iso-thes: http://purl.org/iso25964/skos-thes#`. Diese dienen hauptsächlich zur Abbildung der Klassifikation des Thesaurus. Der Namespace des Uneskos-Vokabulars ist `http://purl.org/umu/uneskos#`, mit dem Kürzel `uneskos:`.

Strukturell baut sich der UNESCO Thesaurus folgendermaßen auf: Die 7 Hauptgruppen des Thesaurus werden als Ressourcen vom Typ `iso-thes:ConceptGroup` definiert, mit `skos:prefLabel` benannt und enthalten ihre Notation mit `skos:Notation`. Sie enthalten das Prädikat `iso-thes:microThesaurusOf` um sie als Mitglieder des `skos:conceptScheme` s auszuzeichnen und das Prädikat `iso-thes:subGroup`, mit

⁸ PASTOR-SÁNCHEZ (2015)

⁹ ISAAC & DE SMEDT (2013)

dem die in der Hauptgruppe enthaltenen Mikrothesauri aufgeführt sind.

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
```

```
@prefix iso-thes: <http://purl.org/iso25964/skos-thes#> .
```

```
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
```

```
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
```

```
<http://skos.um.es/unescothes/COL003> rdf:type iso-thes:ConceptGroup ;
    skos:prefLabel "Culture"@en ;
    skos:notation "3" ;
    rdfs:label "Culture"@en ;
    iso-thes:microThesaurusOf <http://skos.um.es/unescothes/CS000> ;
    iso-thes:subGroup <http://skos.um.es/unescothes/COL365>[...].
```

Die Mikrothesauri sind wiederum als `iso-thes:ConceptGroup` definiert und haben ebenfalls ein `skos:prefLabel` und eine Notation. Die Klasse `uneskos:hasMainConcept` listet dann die Konzepte innerhalb des Mikrothesaurus auf, die keine übergeordneten Begriffe haben. Diese und alle anderen enthaltenen Konzepte sind dazu als `skos:member` aufgelistet. Als rückwärtige Verbindung zur Hauptgruppe wird die Klasse `iso-thes:superGroup` verwendet.

```
<http://skos.um.es/unescothes/COL330> rdf:type iso-thes:ConceptGroup ;
    skos:prefLabel "linguistics"@en ;
    skos:notation "3.30" ;
    rdfs:label "linguistics"@en ;
    skos:member <http://skos.um.es/unescothes/C04384>[...];
    iso-thes:superGroup <http://skos.um.es/unescothes/COL003> ;
    uneskos:hasMainConcept <http://skos.um.es/unescothes/C04382>[...].
```

Die einzelnen Konzepte werden unterschiedlich modelliert, je nach dem ob es sich um Hauptkonzepte des jeweiligen Mikrothesaurus handelt. Beide Varianten enthalten `skos:prefLabel` und ggf. `skos:altLabel` für die Benennung des Konzepts und verwenden die Standard SKOS-Prädikate um assoziative und hierarchische Verwandtschaften zu anderen Konzepten darzustellen. `Skos:scopeNote` wird verwendet um den Kontext des Konzepts festzulegen. Außerdem wird `uneskos:memberOf` verwendet um die Zugehörigkeit zum Mikrothesaurus festzulegen. Konzepte die Hauptkonzepte eines Mikrothesaurus sind, enthalten zusätzlich `uneskos:mainConceptOf` um sie als Hauptkonzept des Mikrothesaurus zu kennzeichnen. Alle Konzepte enthalten `skos:inScheme`, um die Mitgliedschaft zum Thesaurus als Ganzes zu modellieren und `skos:topConceptOf` wenn sie keine übergeordneten Konzepte haben.

```
<http://skos.um.es/unescothes/C03738> rdf:type skos:concept ;
    skos:prefLabel "Sociolinguistics"@en ;
    skos:inScheme <http://skos.um.es/unescothes/CS000> ;
    skos:topConceptOf <http://skos.um.es/unescothes/CS000> ;
    skos:narrower <http://skos.um.es/unescothes/C01429> ;
```

```
skos:related      <http://skos.um.es/unescothes/C03740>[...];
skos:scopeNote   "Fonction sociale et culturelle du langage."@fr[...];
uneskos:memberOf <http://skos.um.es/unescothes/COL330> ;
uneskos:mainConceptOf <http://skos.um.es/unescothes/COL330> .
```

Nachdem an dieser Stelle die Beschreibungen der Beispiellesen abgeschlossen sind, folgt im nächsten Kapitel eine Beschreibung des Mappings von Thesauruskonzepten und eine Analyse der dabei auftretenden Herausforderungen.

6 Mapping von Thesauri

Um das volle Potenzial von Thesauri im Semantic Web auszunutzen ist es wünschenswert verschiedene Thesauri und andere kontrollierte Vokabulare miteinander zu verlinken, bzw aufeinander zu mappen. Dies ermöglicht es mit einer Suche verschiedenste Vokabulare abzufragen und damit den Suchraum und die Qualität der Ergebnisse deutlich zu erhöhen¹. Da beispielsweise die meisten Thesauri, wie aus dem vorherigen Kapitel ersichtlich, nur für einen beschränkten Fachbereich entwickelt wurden und auch nur diesen umfassend abbilden, ist es für eine übergreifende und semantisch angereicherte Suche sinnvoll die verschiedenen Thesauri miteinander zu verbinden². Diese Verbindung findet in Form von Mappings, bzw. Crosskonkordanzen statt. Crosskonkordanzen sind nach Krause³ eine *intellektuell erstellte Verbindungen zwischen Termen zweier Thesauri oder Klassifikationen*. Wie diese Mappings aussehen und welche Möglichkeiten und Methoden es gibt sie zu erstellen ist Thema dieses Kapitels.

Die derzeit gültige ISO-Norm für Thesauri, ISO-25964, legt in ihrem zweiten Teil Richtlinien für die Interoperabilität von Thesauri und anderen Vokabularen fest und empfiehlt Mapping-Relationen für Thesauri. Clarke definiert ein Mapping als:

„(...)a relationship between a concept in one vocabulary and one or more concepts in another“⁴.

Tabelle 7.1, nach Clarke⁵, zeigt die Empfehlungen für drei Basis-Mappingtypen:

Mapping Type	Tag	Example
Equivalence	EQ	Laptop EQ Notebook
Hierarchical	NM	Roads NM Streets
	BM	Streets BM Roads
Associative	RM	Journals RM Magazines

Tabelle 6.1: Mapping Typen nach ISO-25964-2

Die Mapping-Typen entsprechen den internen Relationen von Thesauri. **NM** und **BM** sind äquivalent zu **NT** und **BT** und **RM** entspricht **RT**, jeweils mit dem Unterschied, dass es sich bei den Mapping-Relationen um Relationen *zwischen* Thesauri handelt. Das **EQ** Mapping entspricht in etwa einer **USE/UF** Beziehung, mit der wichtigen Unterscheidung, dass es sich um eine Äquivalenz zwischen

¹ CLARKE (2012, S.2)

² MAYR (2010, S. 1)

³ KRAUSE (2004, S.81)

⁴ CLARKE (2012, S.2)

⁵ CLARKE (2012, S.2)

Konzepten und nicht Termen handelt⁶.

Zusätzlich zu diesen grundlegenden Mappings enthält ISO-25964-2 noch optionale Mappings, die die Relationen genauer spezifizieren sollen. In der folgenden Tabelle, ebenfalls nach Clarke⁷, werden diese aufgelistet:

Mapping Type	Optional differentiation	Examples
Equivalence	Exact	Laptop EQ Notebook
	Inexact	mad cow disease =EQ bovine spongiform encephalopathy
	Compound	lawns ~EQ turfs
	Intersecting	women executives EQ women + executives
	Cumulative	inland waterways EQ rivers canals
Hierarchical		roads NM streets; streets BM roads
	Generic	rats BMG rodents; rodents NMG rats
	Instantial	Paris BMI capital cities; capital cities NMI Paris
Associative	Partitive	fingers BPI hands; hands NPI fingers
		journals RM magazines

Tabelle 6.2: Mapping Types nach ISO 25964-2 inkl. optionaler Relationen

In SKOS finden sich für die meisten dieser Mappings direkte Entsprechungen. SKOS enthält dafür das Prädikat `skos:mappingRelation`, das wiederum verschiedene untergeordnete Prädikate zur Darstellung von verschiedenen Mappings enthält⁸. Die hierarchischen Mappings `NM` und `BM` lassen sich in SKOS mit `skos:narrowMatch` und `skos:broadMatch` und ein assoziatives Mapping mit `skos:relatedMatch` darstellen. Für ein Äquivalenz-Mapping bietet SKOS `skos:exactMatch` für `EQ` und `skos:closeMatch` für `~EQ`. SKOS bietet keine genauen Entsprechungen für die übrigen optionalen Mappings in Tabelle 7.2, was das abbilden solcher Mappings in Thesauri erschwert.

Mappings mit SKOS sind stets Konzept basiert, d.h. es werden zwei Konzepte verbunden. Dabei werden die URIs der beteiligten Konzepte verwendet. Im Beispiel ein Mapping aus dem Agrovoc:

```
<http://aims.fao.org/aos/asfa/c_8123> skos:closeMatch agrovoc:c_49874 .
```

Das dargestellte Tripel verbinden ein Konzept aus dem ASFA Thesaurus mit der URI `http://aims.fao.org/aos/asfa/c_8123` mit einem Konzept mit der URI `http://aims.fao.org/aos/agrovoc/c_49874` aus dem Agrovoc.

Zum Zeitpunkt dieser Arbeit sind bereits eine Vielzahl von Mappings zwischen den untersuchten Beispielthesauri erstellt worden. Diese Mappings sind entweder voll automatisch mit Hilfe von String-

⁶ CLARKE (2012, S.3)

⁷ CLARKE (2012, S.4)

⁸ MILES & BECHHOFFER (2009b)

matching Verfahren, halbautomatisch, d.h. mit intellektueller Überprüfung der String-Matching Ergebnisse oder vollständig manuell erstellt worden. Wie genau ein Mapping mit einem halbautomatischen Verfahren erstellt wird, ist Thema des nächsten Abschnitts.

6.1 Methoden des Thesauri-Mappings

Um Mappings zwischen Thesauri zu erstellen gibt es verschiedene Methoden. Grundsätzlich lassen sich diese in automatische, semi-automatische und manuelle bzw. intellektuelle unterscheiden. In jedem Fall werden bei der Erstellung von Mappings zwischen zwei Thesauri Konzepte in beiden gesucht, die eine der im vorigen Abschnitt vorgestellten Mapping-Relationen zueinander haben. Bei einem manuellen Mapping würde dies bedeuten, dass entsprechende Fachleute sich beide Thesauri ansehen und die Konzepte miteinander vergleichen. Diese Methode ist aufgrund der Menge der Konzepte in den meisten Thesauri ausgesprochen arbeitsintensiv. Automatische Verfahren vergleichen die Konzepte der Thesauri mit String-matching Verfahren. Während diese Methode schnell durchführbar ist, ist sie nicht annähernd so genau wie die manuelle Methode. In der Praxis werden dementsprechend beide Verfahren kombiniert.

Morshed et al⁹ stellen beispielsweise in einem Paper ihre Methode dar, mit der sie Mappings zwischen der SKOS-Version des Agrovoc und sechs anderen in SKOS vorhandenen Thesauri erstellt haben. Sie verwenden eine Kombination aus automatischem String-matching und einer manueller Überprüfung der Daten. Dazu werden zunächst alle Tripel von Agrovoc und dem zu mappenden Thesaurus in einem Tripel-Store gespeichert, einer Datenbank zur Speicherung von RDF-Tripeln. Um danach die `skos:exactMatch` -Mappings zu finden werden die Instanzen von `skos:prefLabel` aus beiden Thesauri mit Hilfe verschiedener String-matching Verfahren verglichen. Da die betrachteten Thesauri verschiedene Sprachversionen anbieten, wird jeweils in der Sprache verglichen die beide Thesauri gemeinsam haben. Dies ist in den meisten Fällen englisch. Aus dem Vergleich werden Ähnlichkeitswerte zwischen den Strings der `skos:prefLabel` ermittelt. Um die Genauigkeit zu erhöhen werden verschiedene String-matching Verfahren verwendet und aus den resultierenden Werten das arithmetische Mittel errechnet. Jede Ähnlichkeit, die unter einem festgelegten Grenzwert liegt, wird entfernt, alle anderen gespeichert.

Anschließend werden die gefundenen Mappings von einem Experten auf den von den Thesauri betrachteten

⁹ MORSHED ET AL. (2011)

Gebieten überprüft. Um sicherzustellen, dass ein Mapping zwischen zwei `skos:prefLabel`s korrekt ist, wurden vier Kriterien überprüft:

1. Die Bedeutung der `skos:prefLabel` wird anhand der vorhandenen `skos:altLabel` nochmals verglichen
2. Die `skos:prefLabel` werden, falls vorhanden, in anderen gemeinsamen Sprachen verglichen
3. Es wird verglichen, an welcher Stelle der Hierarchie des Thesaurus beide Konzepte zu finden sind und inwieweit dies übereinstimmt. Auf diese Weise lassen sich Hinweise auf die genaue Bedeutung des Konzepts finden
4. Definition und Scope Notes werden, falls vorhanden, verglichen

Tabelle 7.3 aus Morshed et al¹⁰, zeigt die Ergebnisse des Vergleichs. Es ist deutlich sichtbar, dass die meisten der vom String-matching Verfahren erstellen Mappings korrekt waren und nur ein geringer Teil inkorrekt.

Aligned thesauri	N. of candidate exact matches	Manual evaluation		Precision
		N. of correct matches	N. of incorrect matches	
AGROVOC-EUROVOC	1,321	1,298	23	98.26
AGROVOC-GEMET	1,240	1,190	50	95.97
AGROVOC-LCSH	1,166	1,095	71	93.90
AGROVOC-NALT	13,609	13,393	216	98.41
AGROVOC-STW	1,165	1,142	23	98.02
AGROVOC-RAMEAU	728	687	41	94.37
TOTAL	19,229	18,805	424	0.98

Tabelle 6.3: Mapping Ergebnisse Agrovoc

Die inkorrekt gemappten Paare dieses Beispiels, als auch anderer Mapping-Situationen, sind auf unterschiedliche Gründe zurückzuführen, von denen einige auf inhaltliche Besonderheiten und Unterschiede der gemappten Thesauri und andere auf die unterschiedliche Verwendung von SKOS zurückzuführen sind. Welche Gründe dies sein können, ist Thema des nächsten Abschnitts.

6.2 Herausforderungen beim Thesaurus-Mapping

Das Mappen verschiedener SKOS-Thesauri aufeinander bringt eine Anzahl Herausforderungen mit sich. Dieses Herausforderungen entstehen teilweise durch unterschiedliche Modellierung in SKOS, teilweise durch Besonderheiten und Unterschiede der Thesauri. Beide Arten von Herausforderungen sollen hier

¹⁰ MORSHED ET AL. (2011, S.7)

Thema sein. An dieser Stelle sollen auch die in Kapitel 5 beschriebenen Thesauri wiederaufgenommen und die Auswirkungen ihrer Besonderheiten auf ihre Kompatibilität betrachtet werden.

6.2.1 Modellierung von Konzepten

Zunächst sollen die SKOS bedingten Herausforderungen Thema sein. Eine davon, ist die sehr unterschiedliche Struktur der verschiedenen SKOS Thesauri. Will man beispielsweise ein Mapping zwischen zwei SKOS-Thesauri erstellen und wendet dabei ein Verfahren ähnlich dem in vorigen Kapitel beschriebenen an, d.h. man vergleicht die `skos:prefLabel` der `skos:concept`s mit verschiedenen String-matching Methoden, fällt schnell auf, dass nicht jeder Thesaurus diese Konzepte oder Label auf die gleiche Weise modelliert. Sieht man sich die in Kapitel 5 beschriebenen Thesauri an, stellt man fest, dass nur 2, nämlich der UNESCO Thesaurus und mit Einschränkungen der STW, klassisch `skos:concept` und `skos:prefLabel` verwenden. Die anderen Thesauri gehen andere Wege. Der Eurovoc Thesaurus beispielsweise, ersetzt `skos:concept` mit der selbst-definierten Klasse `eu:ThesaurusConcept` und verwendet zusätzlich SKOS-XL zur Modellierung der Labels. D.h. man müsste, wollte man den Eurovoc Thesaurus mit einem anderen, klassisch modellierten, Thesaurus mappen, wissen, dass man nicht `skos:concept` und `skos:prefLabel` auf einander mappt, sondern auf Eurovoc Seite `eu:ThesaurusConcept`, `skosxl:prefLabel` und darin die `skosxl:literalForm` verwenden muss. Tabelle 6.4 zeigt die unterschiedlichen Positionen der zu vergleichenden Strings im Überblick.

Thesaurus	Konzept	Label
STW	<code>skos:concept</code> <code>zbwext:descriptor</code>	-> <code>skos:prefLabel</code>
Eurovoc	<code>eu:ThesaurusConcept</code>	-> <code>skosxl:prefLabel</code> -> <code>skosxl:LiteralForm</code>
Agrovoc	<code>skos:concept</code>	-> <code>skos:prefLabel</code> -> <code>skosxl:LiteralForm</code> <code>skosxl:prefLabel</code>
TheSOZ	<code>thesoz:descriptor</code>	-> <code>skosxl:prefLabel</code> -> <code>skosxl:LiteralForm</code>
UNESCO	<code>skos:concept</code>	-> <code>skos:prefLabel</code>

Tabelle 6.4: Aufbau der Thesaurus Konzepte in Beispielthesauri

Während diese Probleme der unterschiedlichen Modellierung bei einem intellektuellen Mapping leicht zu überwinden sind, behindern sie ein automatisches Mapping zukünftiger Thesauri erheblich. Es ist immer nötig vor jedem Mapping die Thesaurusstruktur manuell zu analysieren und das automatische Mapping entsprechend anzupassen¹¹.

Ein anderer Fall der eng mit diesen Modellierungsunterschieden zusammenhängt, ist die irreguläre

¹¹ MAYR (2010, S.8)

Verwendung von `skos:concept` im STW. Der STW beinhaltet neben seinen Konzepten, eine Klassifikation mit 7 Subthesauri, die wiederum 2 weitere Gliederungsstufen haben. Um diese Klassifikation abzubilden verwendet der STW `skos:concept`, indem die Systemstellen der Klassifikation jeweils als `skos:concept` modelliert werden. Benutzt wird diese Form der Modellierung um die Systemstellen der Klassifikation mit `skos:broader` und `skos:narrower` verschachteln zu können¹². Beispielsweise findet sich im STW der Subthesaurus Volkswirtschaft mit der URI `[...]<thsys/v>`. Der Subthesaurus stellt kein Konzept an sich dar, ist aber als `skos:concept` modelliert um die weiteren Stufen der Klassifikation mit `skos:narrower` an den Subthesaurus hängen zu können. Der Subthesaurus ist mit einem `skos:prefLabel` gelabelt, das in seinem String zusätzlich zu Volkswirtschaft noch die Notation des Subthesaurus (V) enthält. Die erste untergeordnete Klassifikationsstufe unter V, heißt ebenfalls Volkswirtschaft und ist wie der Subthesaurus als `skos:concept` modelliert und mit `skos:prefLabel` benannt. Der String des Labels ist wieder Volkswirtschaft mit der Notation, in diesem Fall "V.00 Volkswirtschaft"@de. Um diese Klassifikationsstufen von den Deskriptoren des Thesaurus zu trennen, sind sie neben `skos:concept` zusätzlich als `zbwext:Thsys` deklariert. Die Deskriptoren des Thesaurus sind neben `skos:concept` auch als `zbwext:descriptor` deklariert. Beide Klassen sind als `rdfs:subClassOf` von `skos:concept` modelliert. Während diese Modellierung keine Probleme verursacht solange es nur um Thesaurus interne Relationen geht, kann diese Verwendung von `skos:concept` Probleme verursachen sobald versucht wird ein automatisches Mapping mit anderen Thesauri zu erstellen. Ein String-matching Programm, das `skos:concept` s miteinander vergleicht könnte nicht unterscheiden welche Instanzen tatsächliche Konzepte meinen und welche Systemstellen der Klassifikation darstellen. `zbwext:Thsys` und `zbwext:descriptor` unterscheiden zwar im Thesaurus diese beiden Varianten voneinander, dies müsste im String-matching allerdings vorher auch entsprechend konfiguriert werden.

6.2.2 Modellierung von Klassifikationen

Diese Modellierung der Klassifikation im STW ist nur ein Beispiel für ein Problem, das in allen in Kap. 5 betrachteten Thesauri, außer dem Agrovoc, besteht. Die Thesauri enthalten jeweils ihre eigene Klassifikation und konstruieren zur Darstellung dieser Klassifikation eigene Lösungen, da die Möglichkeiten die SKOS zur Darstellung von Klassifikationen bietet ihnen nicht ausreichen. Die Verwendung von `skos:collection` ist nicht praktikabel, da `skos:collection` nicht mit den in SKOS vorhandenen Relationen geschachtelt werden kann^{13 14}. Der Eurovoc enthält eine 2-stufige Klassifikation, mit 21

¹² NEUBERT (2009, S.3)

¹³ NEUBERT (2009, S.3)

¹⁴ ISAAC & SUMMERS (2009, Kap. 4.1)

Bereichen (Domains) und 127 Mikrothesauri. Um diese Klassifikation abzubilden werden die Klassen `eu:Domain` und `eu:microthesaurus` definiert, die jeweils `rdfs:subClassOf skos:conceptScheme` sind. Da `skos:conceptScheme` aber keine Möglichkeiten bietet verschiedene Instanzen mit `skos:broader` und `skos:narrower` in Relation zu stellen, wird `skos:hasTopConcept` bei den Mikrothesauri verwendet um die einzelnen Konzepte anzuhängen. Der Thesaurus Sozialwissenschaften verwendet eine dreistufige Systematik mit 5 Hauptgruppen. Um diese abzubilden wurde `thesoz:classification` als `rdfs:subClassOf skos:concept` definiert. Diese wird wie im STW dann mit `skos:narrower` und `skos:broader` geschachtelt.

Die Notwendigkeit in diesem Ausmaß eigene Extensions definieren zu müssen um eine Klassifikation abzubilden, zeigt deutlich, dass SKOS in dieser Hinsicht noch großen Verbesserungsbedarf hat. Der UNESCO Thesaurus verwendet ein relativ neues Vokabular von De Smedt und Isaac¹⁵, welches zum Ziel hat, die Interoperabilität zwischen SKOS und der aktuellen ISO-Norm 25964 zu verbessern. Neben einer Vielzahl anderer Verbesserungen und Erweiterungen, enthält das `iso-thes` Vokabular Möglichkeiten zur Darstellung von Klassifikationen. Es wird die Klasse `iso-thes:conceptGroup` eingeführt, eine `rdfs:subClassOf skos:Collection`, die die Möglichkeit bietet verschiedene Gruppen von Konzepten mit den Prädikaten `iso-thes:subGroup` und `iso-thes:superGroup` zu schachteln. Zusätzlich gibt es das Prädikat `iso-thes:microthesaurusOf` mit dem sich Mikrothesauri innerhalb eines `skos:conceptScheme` darstellen lassen. Während diese Möglichkeiten schon eine starke Verbesserung gegenüber Standard-SKOS darstellen, kommt auch der UNESCO-Thesaurus nicht darum herum weitere eigene Klassen zu definieren um seine Mikrothesaurus Struktur korrekt abzubilden. Im UESKOS-Vocabulary¹⁶ werden weitere Klassen und Prädikate definiert, bei denen es sich um inverse Versionen vorhandener Relationen handelt, die die Möglichkeit bieten sollen diese in beide Richtungen darzustellen.

In jedem Fall erschweren die diversen Ansätze der Darstellung von Klassifikationen die Kompatibilität der Thesauri untereinander, da die zunehmend unterschiedlichen Strukturen automatische Mappings sehr schwierig und damit eine intellektuelle Analyse unumgänglich machen.

6.2.3 Verwendung von Compound Equivalence

Eine andere Situation die die Kompatibilität von Thesauri untereinander erschwert, ist die Verwendung von Compound Equivalence. Compound Equivalence (USE+), meint eine Relation in der ein

¹⁵ ISAAC & DE SMEDT (2013)

¹⁶ PASTOR-SÁNCHEZ (2015)

zusammengesetzter Begriff nicht als Deskriptor im Thesaurus vorkommt, sondern als Kombination von Deskriptoren ausgedrückt werden muss¹⁷. Im Beispiel ist der Begriff *Coal Mining* nicht als Deskriptor im Thesaurus aufgenommen, sondern wird als Kombination von *Coal* und *Mining* ausgedrückt:

Coal Mining

USE+ coal

USE+ mining

Wie eine solche Relation das Mapping zwischen Thesauri behindern kann, lässt sich gut anhand eines Beispiels aus dem Thesaurus Sozialwissenschaften erläutern. Der TheSOZ verwendet an vielen Stellen die selbst definierte Relation `thesoz:CompoundEquivalence`, einer `rdfs:subClassOf` von `skosxl:Label`. Beispielsweise ist der Nicht-Deskriptor *Luftverschmutzung* folgendermaßen im Thesaurus aufgenommen:

```
<[...]/thesoz/term/10051428>  skosxl:literalForm  "Luftverschmutzung"@de ;
      a                                           skosxl:Label , prv:DataItem ;
      skos:inScheme                             <http://lod.gesis.org/thesoz/> ;
      [...].
```

```
<[...]/term/10051428-10045125-10045124>      a      thesoz:CompoundEquivalence ,
      prv:DataItem ;
      skos:inScheme                             <http://lod.gesis.org/thesoz/> ;
      [...].
      thesoz:compoundNonPreferredTerm          <[...]/thesoz/term/10051428> ;
      thesoz:preferredTermComponent           <[...]/thesoz/term/10045125> ,
      <[...]/thesoz/term/10045124> .
```

Zunächst ist *Luftverschmutzung* lediglich als Term und nicht als Konzept im Thesaurus aufgeführt. Die `thesoz:CompoundEquivalence` bekommt eine neue URI und enthält `thesoz:compoundNonPreferredTerm`, um den zusammengesetzten Nicht-Deskriptor *Luftverschmutzung* aufzunehmen, und `thesoz:preferredTermComponent`, die die Terme enthält aus denen sich der Begriff zusammensetzt, in diesem Fall *Schadstoff*(10045125) und *Luft*(10045124).

Will man nun beispielsweise ein Mapping mit dem Eurovoc Konzept *Luftverunreinigung* (im Eurovoc der Deskriptor zu *Luftverschmutzung*) oder dem Agrovoc Konzept *Air Pollution* erstellen ergeben sich Probleme. Zunächst gibt es im TheSOZ kein Konzept *Luftverschmutzung* auf das sich mappen ließe, sondern lediglich den Term *Luftverschmutzung*. Es kann kein Mapping zu den Konzepten von *Luft* UND *Schadstoff* erstellt werden, da die SKOS Mapping-Relationen keine 1:n Relation erlauben und

¹⁷ WILL (2012, S.4)

außerdem symmetrisch sind, d.h. ein `skos:exactMatch` mapping von *Luftverunreinigung* auf *Luft* UND *Schadstoff* würde bedeuten das beide für sich genommen das gleiche wären wie *Luftverunreinigung*. In der grundlegenden Form unterstützt SKOS also keine Compound-Mappings¹⁸. Umgehen lässt sich das Problem wenn beide Thesauri SKOS-XL verwenden. Da der Term *Luftverschmutzung* im TheSOZ ein `skosxl:label` ist, könnte man eine Variante von `skosxl:labelRelation` verwenden. Dies würde allerdings voraussetzen, dass eine entsprechende Variante `skosxl:LabelRelation` in beiden Thesauri definiert ist, da `skosxl:LabelRelation` nur für sich keine Aussage außer einer nicht näher spezifizierten Relation zwischen den Labels macht.

Es gibt verschiedene Lösungsansätze Compound Equivalence innerhalb eines Thesaurus abzubilden. Während TheSOZ die selbst definierte Relation `thesoz:CompoundEquivalence` nutzt, beschreiben De Smedt und Isaac in ihrem `iso-thes`-Vokabular, dass wie bereits erwähnt, auch Lösungen für Klassifikationen bietet, andere Klassen, mit denen sich Compound Equivalence darstellen ließe¹⁹. Keine dieser Varianten bietet im Moment jedoch die Möglichkeit Mappings zwischen Thesauri zu erstellen. An dieser Stelle ist noch eine Lösung zu finden, um alle in ISO-25964-2 beschriebenen Mapping-Relationen auch in SKOS abbilden zu können. Während die unterschiedliche Verwendung von SKOS in den Thesauri offenkundig Probleme verursacht, ergeben sich auch Herausforderungen, die durch inhaltliche Unterschiede der Thesauri auftreten. Keil beschreibt vier Gründe für Vagheit zwischen den Vokabularen²⁰. Zunächst werden in unterschiedlichen Thesauri, teilweise unterschiedliche Terme für Begriffe gewählt. Ein gutes Beispiel dafür ist das bereits erwähnte Konzept *Luftverschmutzung*. In STW, TheSOZ und Agrovoc ist der deutsche Deskriptor *Luftverschmutzung* gewählt worden. Der UNESCO-Thesaurus hat keinen deutschen Deskriptor für das Konzept. Auch in anderen Vokabularen, wie der Gemeinsame Norm Datei (GND) der Deutschen Nationalbibliothek oder der DBpedia wird als Vorzugsbenennung *Luftverschmutzung* verwendet. Der Term scheint also der gebräuchlichste für das Konzept zu sein. Dennoch verwendet der Eurovoc-Thesaurus den Begriff *Luftverunreinigung* als Vorzugsbenennung und *Luftverschmutzung* lediglich als alternativen Term. Das setzt sich auch in der englischen Version des Thesaurus fort. Hier verwendet der Eurovoc *atmospheric pollution*, alle anderen geprüften Thesauri aber den Term *air pollution*. Unterschiede wie diese erschweren das Mapping zwischen den Thesauri, in dem beispielsweise ein String-Matching zwischen Eurovoc und den anderen Thesauri hier kein exaktes Mapping produzieren würde, da unterschiedliche `skos:prefLabel` vorhanden sind.

¹⁸ KEIL (2012, S.53)

¹⁹ ISAAC & DE SMEDT (2013)

²⁰ KEIL (2012, S.47)

6.2.4 Unterschiede in der Erschließungstiefe

Ein anderes inhaltliches Problem ist die unterschiedliche Erschließungstiefe der Thesauri. Je nach fachlichem Schwerpunkt des Thesaurus kann es vorkommen, dass Begriffe unterschiedlich spezifisch aufgenommen werden und in der Thesaurus Klassifikation an unterschiedlichen Stellen zu finden sind. Als Beispiel ist hier das Konzept *Apfel* geeignet. *Apfel* kommt in Agrovoc, STW und Eurovoc vor. Der TheSOZ hat das Konzept nicht aufgenommen, da es nicht in seinen behandelten Fachbereich passt. Dies gilt auch für den UNESCO-Thesaurus. Agrovoc, STW und Eurovoc behandeln Äpfel alle unterschiedlich. Der Agrovoc, als landwirtschaftlicher Thesaurus hier am genauesten, enthält den Begriff Apfel in der Begriffskette Produkt->Pflanzliches Produkt->Obst->Kernobst->Äpfel. Der STW enthält Apfel auch als Konzept. Hier findet sich der Begriff allerdings unter Obst->Apfel. Der Eurovoc enthält Apfel nicht als eigenens Konzept, sondern nur als Term mit einer USE Beziehung zu Kernobst. Hier ist die Begriffskette also Obst->Kernobst. Zusätzlich enthält der Eurovoc auch *Pflanzliches Erzeugnis*, das man mindestens als bedeutungsähnlich zu *Pflanzliches Produkt* im Agrovoc betrachten könnte. *Pflanzliches Erzeugnis* ist im Eurovoc allerdings nicht als Konzept sondern als `eu:microthesaurus`, einer `rdfs:subClassOf` von `skos:conceptScheme` aufgenommen und dementsprechend für ein Mapping ungeeignet.

Eine genauere Betrachtung der Mappings zwischen den Konzepten dieser Thesauri fördert zu Tage, dass vor Allem von Seiten des Eurovoc teilweise fehlerhafte Mappings vorhanden sind. Tabelle 6.5 stellt alle Mappings dar, die zwischen den Thesauri für die betrachteten Konzepte vorhanden sind. Zunächst gibt es eine Reihe von `skos:exactMatch` zwischen Agrovoc und STW. Diese verbinden die in beiden Thesauri vorhandenen Konzepte und überspringen die Teile der Hierarchie des Agrovoc die im STW nicht vorkommen. Es werden keine weiteren SKOS Mapping-Relationen verwendet, es wäre jedoch möglich z.B. ein `skos:broadMatch` zwischen STW:Obst und Agro:Apfel und umgekehrt oder ähnliche Mappings zu erstellen. Die vorhandenen Mappings tauchen in beiden Thesauri auf und sind korrekt. Problematischer sind die Mappings zwischen Agrovoc und Eurovoc.

Als erstes fällt auf, dass der Agrovoc neben dem Begriff Obst(Fruits) auch den Begriff Frucht(Fruit) enthält. Ersterer meint das Produkt, letzterer ist unter dem übergeordneten Konzept *plant reproductive organs* eingeordnet und meint die Frucht in botanischem Sinne (beschrieben in der `skos:scopeNote` des Konzepts). Beide sind als `skos:exactMatch` auf das Konzept Obst im Eurovoc gemappt. Diese Mappings tauchen auch im Eurovoc auf und produzieren mehrere Probleme. Zunächst ist es semantisch problematisch, dass hier zwei so unterschiedliche Konzepte aufeinander gemappt werden. Während das Agrovoc Konzept Obst, gut auf das Eurovoc Konzept Obst passt, ist dies bei Agro:Frucht und eu:Obst

nicht der Fall. Obwohl der Eurovoc keine Scope Note oder Definition enthält, die die Bedeutung des Konzepts genauer spezifizieren würden, lässt sich anhand der untergeordneten Begriffe gut erkennen, dass hier zwei sehr unterschiedliche Konzepte aufeinander gemappt wurden. Obst im Eurovoc hat als untergeordnete Begriffe verschiedene spezifischere Obstsorten, wie Kernobst oder Zitrusfrucht, und spezifischere Arten von Frucht, wie Beerenfrucht und Schalenfrucht. Das Agrovoc Konzept Frucht hingegen hat als Unterbegriffe die biologischen Komponenten einer Frucht wie Schale und Samen.

Abgesehen von diesen semantischen Problemen ergibt sich noch ein SKOS-relevantes Problem. Das Prädikat `skos:exactMatch` ist transitiv. Das bedeutet, dass ein `skos:exactMatch` Mapping von <A> auf und von auf <C> bedeutet, dass auch <A> `skos:exactMatch` <C> der Fall ist. Wenn diese Regel beachtet wird, bedeutet das im vorliegenden Fall, dass wenn `agro:Obst` und `eu:Obst` und `agro:Frucht` und `eu:Obst` beide exakte Matches sind, das auch bedeutet, dass `agro:Obst` und `agro:Frucht` ein `skos:exactMatch` sind. Dies ist natürlich offensichtlich falsch, da beide Konzepte sehr unterschiedliche Bedeutungen haben, wie aus ihren Positionen in der Klassifikation und den untergeordneten Termen erkennbar ist.

Auf Seiten des Eurovoc finden sich zusätzlich zu diesem Mapping noch korrekte `skos:broadMatch` -Mappings von `eu:Kernobst` und den einzelnen Kernobst-Sorten im Agrovoc. Ebenfalls vorhanden ist ein falsches `skos:exactMatch` -Mapping von `eu:Kernobst` und `Agro:Feigen`, das so offensichtlich falsch ist, dass es sich um einen simplen Fehler handeln muss. In jedem Fall ist erkennbar, dass die unterschiedlichen Erschließungstiefen der Thesauri und die dadurch entstandenen unterschiedlichen Konzepte ein Hindernis beim Mapping der Thesauri darstellen. An diesem Beispiel ist ebenfalls gut zu erkennen, dass verschiedene Sprachversionen von Konzepten sprachspezifische Probleme verursachen können. Die Konzepte *Obst* und *Frucht* im Agrovoc haben in der englischen Version eine so ähnliche Schreibweise (*Fruits* und *Fruit*), dass ein automatisches Mapping mit String-Matching auf jeden Fall problematische Ergebnisse liefern würde.

6.2.5 Unterschiede in der semantischen Struktur

Der Dritte Problembereich der bei Keil genannt wird, ist hier ebenfalls gut zu sehen. Keil beschreibt, dass unterschiedliche semantische Strukturen in den beteiligten Thesauri Probleme bereiten können, indem sich unterschiedliche Begriffsinhalte finden, je nach dem welches Fachgebiet der Thesaurus behandelt²¹.

²¹ KEIL (2012, S.47)

Konzept	Mapping	Zielthesaurus
Agrovoc		
Produkt	skos:exactMatch	STW:Produkt
Pflanzliches Produkt		
Obst (Fruits)	skos:exactMatch	STW:Obst
Kernobst		
Apfel	skos:exactMatch	STW:Apfel
Frucht (Fruit)	skos:exactMatch	EuV:Obst
STW		
Produkt	skos:exactMatch	Agro:Produkt
Obst (Fruit)	skos:exactMatch	Agro:Obst(Fruits)
Apfel	skos:exactMatch	Agro:Apfel
Eurovoc		
Pflanzliches Produkt	nur als Mikrothesaurus	
Obst	skos:exactMatch	Agro:Obst(Fruits)
	skos:exactMatch	Agro:Frucht(Fruit)
Kernobst	skos:exactMatch	Agro:Kernobst
	skos:exactMatch	Agro:Feigen
	skos:broadMatch	Agro:Apfel
	skos:broadMatch	Agro:Birnen
	skos:broadMatch	Agro:Quitten

Tabelle 6.5: Mappings zwischen Agrovoc, STW und Eurovoc

Je nach Sichtweise kann beispielsweise der Term Universität das Gebäude der Universität oder die Institution Universität meinen. Im beschriebenen Beispiel ist deutlich zu sehen, dass der Agrovoc mit dem Term *fruit* etwas anderes meint als der Eurovoc, dessen englische Variante von Obst ebenfalls *fruit* ist. Während der Agrovoc mit seinem landwirtschaftlichen und biologischen Fachschwerpunkt auf sehr spezifische biologische und botanische Teile einer Frucht bezieht, geht die Frucht im Eurovoc nicht nur im Konzept Obst auf, sondern führt auch zu sehr unterschiedlichen Unterbegriffen.

Der letzte Bereich den Keil nennt²² sind die unterschiedlich verwendeten semantischen Relationen, die je nach dem verwendeten Ordnungssystem der Begriffe zustande kommen. Dies lässt sich am sehr unterschiedlichen Umgang mit dem Konzept Apfel in den verschiedenen Thesauri gut nachvollziehen. Die `skos:broader` und `skos:narrower`-Relationen die von und zu dem Konzept Apfel führen sind in allen betrachteten Thesauri unterschiedlich. Dies wird deutlich, wenn man sich ansieht wohin `skos:broader` von den unterschiedlichen Varianten von *Obst* führt. Im STW führt `skos:broader` von Obst zu einer Systemstelle der Klassifikation mit dem Namen *P.01.02 Obst und Gemüse*, die zwar als `skos:concept`

²² KEIL (2012, S.47)

deklariert ist, nicht aber als `zbwext:descriptor` wie die anderen Konzepte des Thesaurus (siehe Anfang dieses Kapitels). Im Eurvoc gibt es kein `skos:broader` das zu einem Überbegriff von Obst führt. Stattdessen ist das Konzept *Obst* mit `skos:inScheme` an den `eu:microthesaurus` mit dem `skos:prefLabel 6006 Pflanzliches Erzeugnis` gehängt. Im Agrovoc führt `skos:broader` zum übergeordneten Konzept *Planzliches Produkt* von dem aus weitere übergeordnete Konzepte führen.

Alle in diesem Kapitel beschriebenen Herausforderungen erschweren das Mapping verschiedener SKOS-Thesauri aufeinander und behindern eine gemeinsame Nutzung der Thesauri im Sinne des Semantic Web erheblich. Es ist deutlich erkennbar, dass ein automatisches Mapping-Verfahren, das in jedem Fall angewendet werden kann nicht möglich und ein manuelles bzw. intellektuelles Eingreifen immer nötig ist. Während alle beschriebenen Thesauri bereits Mappings zueinander und auch zu anderen SKOS-Thesauri im Netz haben, ist das volle Potential von Mappings bei Weitem noch nicht ausgeschöpft.

7 Fazit

Zielsetzung der vorliegenden Arbeit war es, die Auswirkungen von selbst-definierten Extensions in SKOS-Thesauri auf ihre Auswirkungen auf die Kompatibilität zu prüfen. Nach dieser Prüfung können verschiedene Schlussfolgerungen gezogen werden. Zunächst muss zweifellos festgestellt werden, dass SKOS, auch unter der Hinzunahme von SKOS-XL, in seiner aktuellen Form für die Darstellung der meisten Thesauri keine ausreichenden Möglichkeiten bietet. Dies ist offenkundig wenn man betrachtet, dass für alle der fünf untersuchten Beispielthesauri eine Vielzahl von Extensions erstellen werden mussten, um die Eigenheiten und Anforderungen ihrer Modellierungen befriedigend abbilden zu können. Dies hat zur Folge, dass die untersuchten Thesauri strukturell große Unterschiede aufweisen.

Ein in vier der fünf untersuchten Thesauri aufgetretenes Problem ist dabei die Darstellung der den Thesauri eigenen Klassifikationssysteme. SKOS bietet in seiner Standard-Form keine ausreichenden Möglichkeiten Klassifikationen abzubilden. Die Klasse `skos:collection`, die für eine Sammlung von Konzepten gedacht ist, ist dadurch dass sie nicht schachtelbar ist für Klassifikationen ungeeignet und `skos:notation` bietet zwar die Möglichkeit die Notation einer Klassifikation aufzunehmen, hilft aber nicht bei der Darstellung der eigentlichen Klassifikationshierarchie. Um dieses Problem zu umgehen haben die untersuchten Thesauri unterschiedliche Wege eingeschlagen: Der STW und der Thesaurus Sozialwissenschaften verwenden `skos:concept` oder dazu selbst-definierte Unterklassen um die hierarchischen Relationen `skos:broader` und `skos:narrower` zur Schachtelung der Klassifikation zu verwenden. Der Eurovoc verwendet selbst-definierte Versionen von `skos:conceptScheme` um seine Struktur aus Domains und Mikrothesauri darzustellen und der UNESCO-Thesaurus verwendet eine Kombination aus eigenen Extensions und des `iso-thes:`-Vokabulars von De Smedt und Isaac¹.

Zusätzlich war es in allen Thesauri nötig andere Relationen abzubilden, als jene die in SKOS zur Verfügung stehen. Ein gutes Beispiel dafür ist die sehr umfangreiche Agrontology des Agrovoc Thesaurus mit einer Großen Zahl an Klassen und Prädikaten zur Darstellung sehr spezifischer Relationen im behandelten Fachgebiet. Viele Relationen sind auch für den Thesaurus-Sozialwissenschaften definiert worden (siehe Tabelle 4.1), wie etwa der dem TheSOZ eigene Alternative Descriptor. Eine Relation die SKOS nicht bietet, in Thesauri aber häufige Anwendung findet, ist die Compound Equivalence, also eine Äquivalenz zwischen einem Term und einer Kombination von anderen Termen. Sowohl der Eurovoc

¹ ISAAC & DE SMEDT (2013)

als auch TheSoz und der UNESCO-Thesaurus, mussten für diese Relation eigene Extensions definieren. Allgemeine Lösungen für diese problematischen Fälle haben sich noch nicht gezeigt. Einen Lösungsansatz bieten De Smedt und Isaac in ihrem `iso-thes`-Vokabular. Dieses wurde mit dem Ziel erstellt, eine größere Annäherung von SKOS und den Empfehlungen der aktuellen ISO-25964 zu erreichen und bietet durch die Verwendung schachtelbarer Gruppen von Konzepten die Möglichkeit Klassifikationen abzubilden. Es bietet ebenfalls Möglichkeiten für die Darstellung von Compound Equivalence. Im Bearbeitungszeitraum dieser Arbeit verwendet von den untersuchten Thesauri nur der UNESCO-Thesaurus dieses Vokabular.

Ein anderes Problem, das mit Standard-SKOS momentan noch nicht lösbar ist, ist die Darstellung aller in ISO-25964-2 empfohlenen Mapping-Properties. Die in SKOS vorhandenen Möglichkeiten für das Mapping verschiedener KOS aufeinander decken nicht alle Empfehlungen ab. Beispielsweise bietet SKOS keine Möglichkeit eine Compound Equivalence zu mappen. Auch die spezifischeren Varianten von hierarchischen Mappings, Generic, Instantial und Partitive², sind mit SKOS nicht abgedeckt. An dieser Stelle wird es nötig sein weitere Mapping-Relationen in SKOS zu definieren.

Was allgemein bei den Mappings zwischen den untersuchten Thesauri feststellbar ist, ist dass zwar schon viele Mappings vorhanden sind, diese aber quantitativ und qualitativ noch deutlich ausbaufähig sind. Es fehlen noch Mappings zwischen großen Thesauri, wie beispielsweise dem STW und dem Eurovoc und viele der erstellten Mappings sind entweder fehlerhaft, wie das in Kap. 6.2 beschriebene Mapping zwischen Eurovoc und Agrovoc, oder die Verlinkung der Weboberflächen funktioniert nicht. Grundsätzlich ist es nötig für das Mapping verschiedener Thesauri, zumindest teilweise, intellektuell einzugreifen. Ein vollautomatisches und kontinuierliches Mapping ist durch die großen strukturellen Unterschiede und die anderen in Kapitel 6.2 beschriebenen Herausforderungen unmöglich. Zusätzlich folgt aus der Verwendung von String-matching Verfahren für das Mapping, dass die meisten erstellten Relationen `skos:exactMatch` oder `skos:closeMatch` Relationen sind und die hierarchischen und assoziativen Mappings die SKOS bietet sehr viel weniger Verwendung finden.

Abschließen lässt sich sagen, dass SKOS in seiner aktuellen Form trotz seiner umfangreichen Verbreitung, noch nicht die ideale Lösung für KOS im Semantic Web darstellt. Die Notwendigkeit der Erstellung von eigenen Extensions steht dem Wunsch der Interoperabilität der Vokabulare untereinander genauso im Weg, wie die sehr unterschiedliche Modellierung der einzelnen Thesauri.

² CLARKE (2012)

Eidesstattliche Erklärung

Hiermit erkläre ich an Eides statt, dass ich die eingereichte Bachelor-/Masterarbeit selbstständig und ohne fremde Hilfe verfasst, andere als die von mir angegebenen Quellen und Hilfsmittel nicht benutzt und die den benutzten Werken wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Einbeck, den 04.05.2016

(Unterschrift)

Literaturverzeichnis

- ABDUL MANAF, NOR AZLINAYATI., BECHHOFFER, SEAN. & STEVENS, ROBERT. (2012): *The Semantic Web: Research and Applications: 9th Extended Semantic Web Conference, ESWC 2012, Heraklion, Crete, Greece, May 27-31, 2012. Proceedings*, Kap. The Current State of SKOS Vocabularies on the Web, 270–284. Springer, Berlin [u.a.]. URL http://dx.doi.org/10.1007/978-3-642-30284-8_25, zuletzt geprüft am 20.02.2016.
- ALLEMANG, DEAN. & HENDLER, JIM. (2011): *Semantic Web for the working ontologist. Effective modeling in RDFS and OWL*. 2. Aufl. Aufl. Elsevier, Amsterdam [u.a.].
- BERNERS-LEE, TIM. & CONNOLLY, DAN. (2011): *Notation3 (N3): A readable RDF syntax*. URL <https://www.w3.org/TeamSubmission/n3/>; zuletzt geprüft am: 10.03.2016.
- BRICKLEY, DAN. & GUHA, RV. (2014): *RDF Schema 1.1*. URL <https://www.w3.org/TR/rdf-schema/>; zuletzt geprüft am: 20.02.2016.
- CLARKE, STELLA DEXTRE. (2012): *ISO 25964: a standard in support of KOS interoperability*. URL http://www.iskook.org/sites/default/files/ISKOUK-2011_StellaDextreClarke.pdf; zuletzt geprüft am: 20.02.2016.
- DE KEYSER, PIERRE. (2012): *Indexing - From thesauri to the semantic web*. Chandos, Oxford [u.a.].
- DUBLIN CORE METADATA INITIATIVE (2016a): *Dublin Core Metadata*. URL <http://dublincore.org/>; zuletzt geprüft am: 20.02.2016.
- DUBLIN CORE METADATA INITIATIVE (2016b): *Dublin Core User Guide*. URL http://wiki.dublincore.org/index.php/User_Guide; zuletzt geprüft am: 20.02.2016.
- FAO (2012): *Agrontology*. URL <http://aims.fao.org/sites/default/files/uploads/file/aos/agrontology/index.htm#objectproperties>; zuletzt geprüft am: 31.03.2016.
- GANDON, FABIEN. & SCHREIBER, GUUS. (2014): *RDF 1.1 XML Syntax*. URL <https://www.w3.org/TR/rdf-syntax-grammar/>; zuletzt geprüft am: 10.03.2016.
- HARTIG, OLAF. & ZHAO, JUN. (2012): *Provenance Vocabulary Core Ontology Specification*. URL <http://trdf.sourceforge.net/provenance/ns.html#>; zuletzt geprüft am: 31.03.2016.
- HITZLER, PASCAL. (2008): *Semantic Web : Grundlagen*. 1. Aufl. Aufl. Springer, Berlin [u.a.].

- HITZLER, PASCAL [U.A]. (2012): *OWL 2 Web Ontology Language Primer (Second Edition)*. URL <https://www.w3.org/TR/owl2-primer/>; zuletzt geprüft am: 20.02.2016.
- ISAAC, ANTOINE. & DE SMEDT, JOHAN [HG.]. (2013): *Correspondece between ISO 25964 and SKOS/SKOS-XL Models*. URL http://www.niso.org/apps/group_public/download.php/12351/Correspondence%20ISO25964-SKOSXL-MADS-2013-12-11.pdf; zuletzt geprüft am: 20.02.2016.
- ISAAC, ANTOINE. & SUMMERS, ED. (2009): *SKOS Simple Knowledge Organization System Primer*. URL <http://www.w3.org/TR/skos-primer/>; zuletzt geprüft am: 20.02.2016.
- ISO INTERNATIONALE ORGANISATION FÜR NORMUNG (1985): *ISO 5964 - Richtlinien für die Schaffung und Weiterentwicklung von mehrsprachigen Thesauren*.
- ISO INTERNATIONALE ORGANISATION FÜR NORMUNG (1986): *ISO 2788 - Richtlinien für die Erstellung und Entwicklung einsprachiger Thesauri*.
- ISO INTERNATIONALE ORGANISATION FÜR NORMUNG (2009): *ISO 15836 - Information and documentation - The Dublin Core metadata element set*.
- ISO INTERNATIONALE ORGANISATION FÜR NORMUNG (2011): *ISO 25964-1 - Thesauri und die Interoperabilität mit anderen Vokabularen - Teil 1: Thesauri zur Informationssuche*.
- ISO INTERNATIONALE ORGANISATION FÜR NORMUNG (2013): *ISO 25964-1 - Thesauri und die Interoperabilität mit anderen Vokabularen - Teil 2: Interoperabilität mit anderen Vokabularen*.
- KEIL, STEFAN. (2012): *Terminologie Mapping: Grundlagen und aktuelle Normungsvorhaben*. In: *Information - Wissenschaft und Praxis*, **63**, 1: 45–55.
- KRAUSE, JÜRGEN. (2004): *Konkretes zur These, die Standardisierung von der Heterogenität her zu denken*. In: *Zeitschrift für Bibliothekswesen und Bibliographie*, **51**, 2: 76–89.
- KUHLEN, RAINER., SEEGER, THOMAS. & STRAUCH, DIETMAR. (2004): *Grundlagen der praktischen Information und Dokumentation: Band 1: Handbuch zur Einführung in die Informationswissenschaft und -praxis*. 5. Aufl. Aufl. De Gruyter, Berlin.
- MAYR, PHILIPP. (2010): *Ein Mehr-Thesauri-Szenario auf Basis von SKOS und Crosskonkordanzen*. In: *25. Oberhofer Kolloquium*. DGI, Magdeburg [u.a]. URL <http://www.ib.hu-berlin.de/~mayr/arbeiten/Oberhof2010.pdf>, zuletzt geprüft am 20.02.2016.

- MILES, ALIFTAIR. & BECHHOFFER, SEAN. (2009a): *SKOS Simple Knowledge Organization System eXtension for Labels (SKOS-XL) Namespace Document*. URL <http://www.w3.org/TR/skos-reference/skos-xl.html>; zuletzt geprüft am: 20.02.2016.
- MILES, ALIFTAIR. & BECHHOFFER, SEAN. (2009b): *SKOS Simple Knowledge Organization System Reference*. URL <https://www.w3.org/TR/skos-reference/>; zuletzt geprüft am: 20.02.2016.
- MORSHED, AHSAN., CARACCILO, CATERINA., JOHANNSEN, GUDRUN. & KEIZER, JOHANNES. (2011): *Thesaurus Alignment for Linked Data Publishing*. In: *International Conference on Dublin Core and Metadata Applications 2011*. The Hague. URL <http://www.fao.org/docrep/015/an895e/an895e00.pdf>, zuletzt geprüft am 21.04.2016.
- NEUBERT, JOACHIM. (2009): *Bringing the Thesaurus for Economics to the Web of Linked Data*. URL http://ceur-ws.org/Vol-538/ldow2009_paper7.pdf, zuletzt geprüft am 21.04.2016.
- PAFOR-SÁNCHEZ, JUAN-ANTONIO. (2015): *UNESKOS Vocabulary*. URL <http://skos.um.es/TR/uneskos/>; zuletzt geprüft am: 31.03.2016.
- PELLEGRINI, TAFSILO. & BLUMAUER, ANDREAS [HG.]. (2006): *Semantic Web : Wege zur vernetzten Wissensgesellschaft*. Springer, Berlin [u.a.].
- PRUD'HOMMEAUX, ERIC. & CAROTHERS, GAVIN. (2014): *RDF 1.1 Turtle*. URL <https://www.w3.org/TR/turtle/>; zuletzt geprüft am: 20.02.2016.
- SCHREIBER, GUUS. & RAIMOND, YVES. (2014): *RDF 1.1 Primer*. URL <https://www.w3.org/TR/2014/NOTE-rdf11-primer-20140624/>; zuletzt geprüft am: 20.02.2016.
- VAN AFSEM, MARK., MALAISE, VERONIQUE., MILES, ALIFTAIR. & SCHREIBER, GUUS. (2006): *The Semantic Web: Research and Applications: 3rd European Semantic Web Conference, ESWC 2006 Budva, Montenegro, June 11-14, 2006 Proceedings*, Kap. A Method to Convert Thesauri to SKOS, 95–109. Springer. URL http://dx.doi.org/10.1007/11762256_1, zuletzt geprüft am 20.02.2016.
- VOß, JAKOB. (2014): *GBV-Ontology*. URL <http://gbv.github.io/gbvontology/gbvontology.html>; zuletzt geprüft am: 20.02.2016.
- W3C SEMANTIC WEB WIKI (2014): *Resource Description Framework (RDF)*. URL <https://www.w3.org/RDF/>; zuletzt geprüft am: 10.03.2016.

WILL, LEONARD. (2012): *The ISO 25964 data model for the structure of an information retrieval thesaurus*. In: Bulletin of the American Society for Information Science and Technology, **38**, 4: 48–51. URL <http://dx.doi.org/10.1002/bult.2012.1720380413>.

WOOD, DAVID. (2014): *Linked Data : structured data on the Web*. Manning, Shelter Island.

ZAPILKO, BENJAMIN. & SURE, YORK. (2009): *Converting the TheSoz to SKOS*. URL http://www.gesis.org/fileadmin/upload/forschung/publikationen/gesis_reihen/gesis_methodenberichte/2009/TechnicalReport_09_07.pdf; zuletzt geprüft am: 20.02.2016.

ZAPILKO, BENJAMIN [U.A.]. (2013): *TheSoz - A SKOS Representation of the Thesaurus for the Social Sciences*. In: Semantic web : interoperability, usability, applicability, **4**, 3: 257–263. URL http://www.semantic-web-journal.net/sites/default/files/swj279_2.pdf, zuletzt geprüft am 20.02.2016.