Organisation eines Thesaurus für die Unterstützung der mehrsprachigen Suche in einer bibliographischen Datenbank im Bereich Planen und Bauen

Dimitri Busch, Fraunhofer-Informationszentrum Raum und Bau (IRB)

Zusammenfassung:
Das Problem der mehrsprachigen Suche gewinnt in der letzten Zeit immer mehr an Bedeutung, da viele nützliche Fachinformationen in der Welt in verschiedenen Sprachen publiziert werden. RSWB®plus ist eine bibliographische Datenbank zum Nachweis der Fachliteratur im Bereich Planen und Bauen, welche deutsch- und englischsprachige Metadaten-Einträge enthält. Bis vor kurzem war es problematisch Einträge zu finden, deren Sprache sich von der Anfragesprache unterschied. Zum Beispiel fand man auf deutschsprachige Anfragen nur deutschsprachige Einträge, obwohl die Datenbank auch potenziell nützliche englischsprachige Einträge enthielt. Um das Problem zu lösen, wurde nach einer Untersuchung bestehender Ansätze die RSWB®plus weiterentwickelt, um eine mehrsprachige (sprachübergreifende) Suche zu unterstützen, welche unter Einbeziehung eines zweisprachigen begriffsbasierten Thesaurus erfolgt. Der Thesaurus wurde aus bereits bestehenden Thesauri automatisch generiert. Die Einträge der Quell-Thesauri wurden in das SKOS-Format (Simple Knowledge Organisation System) umgewandelt, automatisch miteinander vereinigt und schließlich in einen Ziel-Thesaurus eingespielt, der ebenfalls in SKOS geführt wird. Für den Zugriff zum Ziel-Thesaurus werden Apache Jena und MS SQL Server verwendet. Bei der mehrsprachigen Suche werden Terme der Anfrage durch entsprechende Übersetzungen und Synonyme in Deutsch und Englisch erweitert. Die Erweiterung der Suchterme kann sowohl in der Laufzeit als auch halbautomatisch erfolgen. Das verbesserte Recherchesystem kann insbesondere deutschsprachigen Benutzerinnen und Benutzern helfen, relevante englischsprachige Einträge zu finden. Die Verwendung von SKOS erhöht die Interoperabilität der Thesauri und vereinfacht das Bilden des Ziel-Thesaurus und den Zugriff auf seine Einträge.

Summary:
In recent times, the problem of multi-lingual search is gaining more and more importance, because useful specialized information is published in different languages around the world. RSWB®plus is a bibliographic database, which includes German and English metadata entries in the field of construction and planning. Until recently it was difficult to find entries whose language differed from the query language. For example, German queries found only German entries, although the database also contained potentially useful English entries. After an investigation of existing approaches, the RSWB®plus database was improved to support cross-language information retrieval, which makes use of a bilingual concept-based thesaurus. The thesaurus was automatically generated from existing thesauri. The entries in the source thesauri were converted into SKOS format (Simple Knowledge Organization System), automatically merged and finally recorded in a target thesaurus, also in the SKOS format. Apache Jena and MS SQL Server are used to access the target thesaurus. For the multilingual retrieval, query terms are expanded by appropriate translations and synonyms in English and German. The expansion of the search terms can be carried out both semi-automatically and in the runtime. The improved retrieval system can especially help German users to find relevant English

entries. The use of the SKOS format increases interoperability of thesauri and simplifies the building of the target thesaurus and the access to its entries.

Zitierfähiger Link (DOI): http://dx.doi.org/10.5282/o-bib/2016H4S202-216
Schlagwörter: Mehrsprachige Suche; Sprachübergreifende Suche; Information Retrieval; Thesaurus; Ordnungssystem; Literaturdokumentation

1. Einführung

Das Problem der mehrsprachigen Suche gewinnt in der letzten Zeit immer mehr an Bedeutung, da viele nützliche Fachinformationen in der Welt in verschiedenen Sprachen publiziert werden. In diesem Artikel geht es um einen Thesaurus, der für die zweisprachige Suche in der Datenbank RSWB®plus verwendet wird. RSWB®plus ist eine bibliographische Datenbank zum Nachweis der Fachliteratur im Bereich Planen und Bauen, die im Fraunhofer-Informationszentrum Raum und Bau (IRB) produziert und online angeboten wird. RSWB®plus enthält deutschsprachige Metadaten-Einträge (Dokumentationseinheiten) der deutschen Baudatenbank RSWB® (Raumordnung, Städtebau Wohnungswesen, Bauwesen) und englischsprachige Metadaten-Einträge der internationalen Baudatenbank ICONDA®Bibliographic (the International Construction Database).[1] In Abbildung 1 und Abbildung 2 sind Beispiele der Einträge dargestellt.

Bei der Suche in RSWB®plus treten Probleme auf, die dadurch verursacht sind, dass die Datenbank Metadaten-Einträge in unterschiedlichen Sprachen enthält. Auf deutschsprachige Anfragen findet man nur deutschsprachige Einträge, obwohl die Datenbank auch potenziell nützliche englischsprachige Einträge enthalten kann. Auf englischsprachige Anfragen findet man nur dann deutschsprachige Einträge, wenn man nach Schlagwörtern sucht, da die Schlagwörter in den deutschsprachigen Einträgen auch ins Englische übersetzt sind.

Die Probleme können gelöst werden, indem die Anfragen in die jeweiligen Sprachen der Metadaten-Einträge übersetzt werden. Zum Beispiel kann man deutschsprachige Anfragen ins Englische übersetzen, um entsprechende englischsprachige Metadaten-Einträge zu finden. Die Übersetzung der Anfragen kann mit Hilfe eines Thesaurus erfolgen.

Im Folgenden werden bestehende Ansätze zur mehrsprachigen Suche und zur Organisation von Thesauri betrachtet. Nach der Analyse dieser Ansätze wird ein Verfahren zur Bildung eines zweisprachigen Thesaurus aus bereits bestehenden Thesauri und die Einbindung dieses Thesaurus in die Suche dargestellt.

1 Hier und im Folgenden handelt es sich um die Sprachen, in denen Metadaten in den jeweiligen Datenbanken geführt werden. Die Sprache eines Metadaten-Eintrags kann sich von der Sprache unterscheiden, in der die entsprechende dokumentarische Bezugseinheit, z.B. ein Buch oder ein Zeitschriftenartikel, publiziert wurde.

Originaltitel	Unverwechselbar. Fassade und Wärmedämmung
Autor	Müller, Kay-Uwe
Schlagwörter	Mehrfamilienhaus; Fassadengestaltung; Oberflächenstruktur; Passivhaus; Putzfassade; Farbkonzept; multiple dwelling; facade design; texture; passive house; plaster facade; color concept
Fachgebiet	10.060- Fassade; 14.170- Putzarbeit
Erscheinungsjahr	2015
Sprache	Deutsch
Publikationstyp	Zeitschriftenartikel
Quelle	Malerblatt (2015), Bd.86, Nr.3, S.58-60, ISSN: 1434-1360

Abb. 1: Deutscher Metadaten-Eintrag (RSWB®)

Original title	Advanced thermal insulation technologies in the built environment
Author	Livesey, Katie
Abstract	Reviews thermal insulation products, with a focus on advanced thermal insulation technologies such as aerogels, vacuum insulated panels, gas-filled panels and phase change materials.
Keywords	heat; insulation; efficiency; evaluation; insulating materials; materials; heat transmission; analysis
Publication year	2013
Language	English
Publication type	Journal Article
Source	BRE information paper (2013), no.4/13, p.1-16

Abb. 2: Englischer Metadaten-Eintrag (ICONDA®Bibliographic)

2. Bestehende Ansätze

2.1. Überlegungen zur mehrsprachigen Suche

Bei der mehrsprachigen Suche findet eine Suchanfrage in einer Sprache auch Dokumente in anderen Sprachen. Unter Dokumenten werden hier und im Folgenden Metadaten-Einträge sowie alle möglichen textuellen Einträge auf maschinellen Trägern verstanden. Im Folgenden werden bestehende Ansätze zur mehrsprachigen Suche genauer betrachtet.

2.1.1. Übersetzung der Dokumente vs. Übersetzung der Anfragen

Um das Wiederauffinden der Einträge in einer Sprache auf die Anfrage in einer anderen Sprache zu ermöglichen, gibt es zwei Hauptansätze:[2]

- Die Suchanfrage wird in die Sprache(n) der Dokumente übersetzt
- Dokumente werden in die Anfragesprache übersetzt

Um die o.g. Ansätze zu vergleichen, wurden bei der Firma IBM Retrieval-Experimente durchgeführt, bei welchen sowohl Dokumente als auch Anfragen mit gleichen Werkzeugen automatisch übersetzt wurden (Englisch-Französisch, Französisch-Englisch). Diese Experimente zeigten weder bei der Übersetzung der Dokumente noch bei der Übersetzung der Anfragen einen klaren Vorteil für die Retrieval-Qualität.[3]

Ein Vorteil der Dokumentenübersetzung kann darin bestehen, dass sie ermöglicht, die bei der Recherche gefundenen fremdsprachigen Dokumente zu lesen und zu verstehen. Der Vorteil ist jedoch nur dann erreichbar, wenn die Übersetzung bestimmten Qualitätsanforderungen entspricht. Um die Anforderungen zu erfüllen, sollte die Übersetzung maschinell nach einer sorgfältigen Anpassung des Übersetzungssystems an entsprechende Themenbereiche erfolgen.[4]

Ein wichtiger Vorteil der Anfragen-Übersetzung besteht in der Flexibilität bei der Anfrageformulierung. Die Benutzerinnen und Benutzer können die Sprachen angeben, in welche eine Anfrage zu übersetzen ist. Wenn sie die übersetzte Anfrage verstehen können, können sie bei Bedarf die Anfrage korrigieren, bevor sie sie für eine Recherche verwenden.[5]

Da die Übersetzung der Dokumente keine klaren Vorteile in Retrieval-Experimenten aufweist, wird derzeit meist die Übersetzung der Anfragen verwendet, weil dieser Ansatz mehr Flexibilität in die Anfrageformulierung bringt.[6] Auch für das Fraunhofer-Informationszentrum Raum und Bau (IRB) scheint die Übersetzung der Anfragen aus ähnlichen Gründen passender zu sein, auch weil die Beschaffung eines Systems zur Dokumentenübersetzung und dessen Anpassung an den Bereich Planen und Bauen kostenaufwändig sein können. Der Ansatz wird in Folgendem als sprachübergreifende Suche (cross-language information retrieval) bezeichnet.

2 Vgl. Wolfgang Stock, *Information Retrieval: Informationen suchen und finden*, Einführung in die Informationswissenschaft 1 (München: Oldenbourg Wissenschaftsverlag, 2007), 465 ff.

3 Vgl. J. Scott McCarley, „Should We Translate the Documents or the Queries in Cross-Language Information Retrieval?“ *In 37th Annual Meeting of the Association for Computational Linguistics: Proceedings of the Conference, 20-26 June 1999, University of Maryland, College Park, Maryland, USA*, hrsg. Association for Computational Linguistics (San Francisco: Morgan Kaufmann, 1999), 208-214.

4 Vgl. Paul Buitelaar, Klaus Netter und Feiyu Xu, „Integrating Different Strategies in Cross-Language Information Retrieval in the MIETTA Project,“ in *Language technology in multimedia information retrieval: Proceedings of the Fourteenth Twente Workshop on Language Technology, December 7-8, 1998, Enschede, The Netherlands*, hrsg. Djoerd Hiemstra, Franciska de Jong und Klaus Netter (Enschede: Univ. Twente, 1998), 9-17.

5 Vgl. Jian-Yun Nie, *Cross-Language Information Retrieval* (San Rafael: Morgan & Claypool, 2010), 19-20.

6 Vgl. ebd., 20.

2.1.2. Ansätze zur Übersetzung in der sprachübergreifenden Suche

Die Übersetzung der Suchanfragen bei der sprachübergreifenden Suche kann auf folgende Weise erfolgen:[7]

- Nutzung maschinenlesbarer mehrsprachiger Wörterbücher und Thesauri: Dieser Ansatz versucht, für jedes Wort einer Anfrage alle möglichen Übersetzungen in einem Wörterbuch zu finden. Die übersetzten Wörter bilden dann die Abfrage in der Zielsprache.
- Nutzung paralleler Korpora: Die parallelen Korpora sind Sammlungen von inhaltsgleichen Dokumenten, welche parallel in mehreren Sprachen abgespeichert sind. Ansätze, welche die parallelen Korpora nutzen, versuchen Übersetzungsrelationen zwischen zwei oder mehreren Sprachen herzustellen. Solche Relationen können entweder auf der Wortebene oder auf einer höheren Ebene, z.B. auf der Satzebene, hergestellt werden. Diese Übersetzungsrelationen können dann verwendet werden, um Abfragen oder Dokumente zu übersetzen.
- Nutzung eines „vollen" Systems für die maschinelle Übersetzung.

Dem Fraunhofer IRB liegen bereits mehrere Thesauri in den Bereichen Bauwesen und Raumordnung vor. Deswegen schien es sinnvoll, Suchanfragen mittels der Thesauri zu übersetzen. Im Folgenden werden Thesauri genauer betrachtet.

2.2. Allgemeineres zu Thesauri

Ein Thesaurus ist eine geordnete Zusammenstellung von Begriffen und ihren Benennungen, die zum Indexieren, Speichern und Wiederauffinden der Dokumente dient.[8]

Die Norm DIN 2342 versteht unter einem Begriff eine „Denkeinheit, die aus einer Menge von Gegenständen unter Ermittlung der diesen Gegenständen gemeinsamen Eigenschaften mittels Abstraktion gebildet wird."[9]

Begriffe werden sprachlich durch Benennungen repräsentiert. Eine Benennung ist eine aus einem Wort oder mehreren Wörtern bestehende Bezeichnung.[10] Zum Beispiel wird der Begriff COMPUTER durch die Benennungen „Computer" und „Rechner" repräsentiert. Bei Benennungen unterscheidet man zwischen Vorzugsbenennungen (Deskriptoren) und Alternativbenennungen (Nichtdeskriptoren). Die Vorzugsbenennungen sind Benennungen, welche zur Indexierung zugelassen werden.
Im Folgenden werden Benennungen auch als „Terme" bezeichnet.

7 Vgl. Carol Peters, Martin Braschler und Paul Clough, *Multilingual Information Retrieval: From Research to Practice* (Berlin Heidelberg: Springer, 2012), 60; vgl. dazu auch Nie, *Cross-language Information Retrieval*, 22.

8 Vgl. DIN-Normenausschuss Information und Dokumentation (NID) im DIN e.V., *DIN 1463-1, Erstellung und Weiterentwicklung von Thesauri: Einsprachige Thesauri* (Berlin: Beuth, 1987), 2.

9 DIN-Normenausschuss Terminologie (NAT) im DIN e.V., *DIN 2342-1, Begriffe der Terminologielehre: Grundbegriffe* (Berlin: Beuth, 1992), 1.

10 Vgl. ebd., 2.

2.3. Repräsentation der Thesauri

2.3.1. Modelle für die Repräsentation der Thesauri

Für die Repräsentation der Thesauri kann man zwei Modelle verwenden:[11]

- termbasiertes Modell
- begriffsbasiertes Modell (concept-based model)

Ein Thesaurus, welcher das termbasierte Modell unterstützt, besteht aus Benennungen (Termen) und Relationen zwischen den Benennungen. Ein begriffsbasierter Thesaurus besteht aus Begriffen und Relationen zwischen den Begriffen. Ein Begriff hat normalerweise einen eindeutigen Identifikator und eine oder mehrere Benennungen.

Traditionell wurden Thesauri nach dem termbasierten Modell aufgebaut. Alle Thesauri, welche im Fraunhofer IRB momentan vorliegen, sind termbasiert (siehe hierzu Kap. 3). In den letzten Jahren, insbesondere nach dem Erscheinen des Simple-Knowledge-Organisation-System-Formats (SKOS-Format), werden jedoch viele Thesauri nach dem begriffsbasierten Modell aufgebaut oder in dieses Modell umgewandelt. Genaueres über diese Entwicklung kann man in Kap. 2.3.3 erfahren.

2.3.2. Simple Knowledge Organisation System (SKOS)

In der Vergangenheit wurden Thesauri von verschiedenen Produzenten in unterschiedlichen Formaten dargestellt, was den Austausch der Thesauri und ihre Nutzung in unterschiedlichen Anwendungen erschwerte. Um den Austausch und die Nutzung der Thesauri zu erleichtern, ist es empfehlenswert, die Thesauri in standardisierten Formaten zu führen. Als Beispiel eines solchen Formats kann man das NISO-Format (ISO 25964) nennen.[12]

Das derzeit am häufigsten verwendete standardisierte Format ist SKOS. SKOS ist ein Standard vom W3C (World Wide Web Consortium) für die Repräsentation von Thesauri.[13] SKOS basiert wiederum auf den W3C-Standards für Semantic-Web RDF und OWL. Das fundamentale Element von SKOS ist der Begriff (Concept). Ein Begriff hat einen eindeutigen globalen Identifikator (URI), welcher den Begriff im Web eindeutig kennzeichnen kann, und eine oder mehrere Benennungen. Es gibt zwei Benennungstypen: Vorzugsbenennungen und Alternativbenennungen. Zwischen Begriffen können Relationen bestehen.

```
ex:computer rdf:type skos:Concept;
  skos:prefLabel „Computer“@de;
  skos:altLabel „EDV-Anlage“@de;
  skos:altLabel „Rechner“@de;
  skos:prefLabel “computer”@en;
  skos:broader ex:pc.
```

Abb. 3: Beispiel eines Begriffes in SKOS

Abbildung 3 zeigt ein Beispiel eines SKOS-Begriffes. Der Begriff hat einen Identifikator „ex:computer“, eine deutsche Vorzugsbenennung „Computer“, zwei deutsche Alternativbenennungen,

11 Vgl. Javier Lacasta, Javier Nogueras-Iso und Francisco Zarazags-Soria, *Terminological Ontologies: Design, Management and Practical Applications* (New York: Springer, 2010), 9-10.

12 „Format for Exchange of Thesaurus Data Conforming to ISO 25964-1,“ NISO, zuletzt geprüft am 02.09.2016, http://www.niso.org/schemas/iso25964/schema-intro.

13 Vgl. „SKOS Simple Knowledge Organization System Primer,“ W3C, zuletzt geprüft am 02.09.2016, https://www.w3.org/TR/skos-primer/.

„EDV-Anlage“ und „Rechner“, und eine englische Vorzugsbenennung „computer“. Der Begriff mit dem Identifikator „ex:computer“ ist ein Oberbegriff für den Begriff mit dem Identifikator „ex:pc“.

2.3.3. Verbreitung von SKOS

SKOS wurde am 18. August 2009 vom W3C als Empfehlung veröffentlicht.[14] Seit der Veröffentlichung wurden mehrere große Thesauri, wie z.B. AGROVOC und EUROVOC, sowie eine Vielzahl kleinerer Thesauri und anderer kontrollierter Vokabularien in SKOS publiziert.[15] Die weitere Verbreitung des SKOS-Formats wird u.a. von folgenden Faktoren verursacht:[16]

- Da SKOS vom W3C, einem internationalen Gremium zur Standardisierung der Techniken im World Wide Web, veröffentlich wird, werden Thesauri durch die Adoption von SKOS auf standardisierte Weise repräsentiert.
- Es ist relativ einfach, Thesauri, welche in anderen Formaten repräsentiert sind, in SKOS zu konvertieren. Bei einer solchen Konversion können Deskriptoren eines Quell-Thesaurus in SKOS-Begriffe und ihre Vorzugsbenennungen, Nichtdeskriptoren in Alternativbenennungen, und Relationen, wie Hierarchie und Assoziation, in entsprechende SKOS-Relationen, wie „broader“, „narrower“ und „related“, umgewandelt werden.
- Thesauri, welche in SKOS repräsentiert sind, können in das Semantic Web integriert und miteinander verlinkt werden.

2.4. Ansätze zum Aufbau von mehrsprachigen Thesauri

Ein mehrsprachiger Thesaurus ist ein Thesaurus, der für jeden Begriff äquivalente Benennungen in mehreren Sprachen enthält. Es gibt drei Ansätze zum Aufbau von mehrsprachigen Thesauri:[17]

1. Aufbau eines neuen Thesaurus von unten nach oben
   a. Man fängt mit einer Sprache an, und fügt eine andere Sprache oder Sprachen hinzu.
   b. Man fängt mit mehr als einer Sprache gleichzeitig an.
2. Die Kombination von bestehenden Thesauri
   a. Vereinigung von zwei oder mehreren bestehenden Thesauri in einen neuen Thesaurus.
   b. Verknüpfung bestehender Thesauri miteinander.
3. Die Übersetzung eines bestehenden Thesaurus in eine oder mehrere Sprachen.

Wie erwähnt, liegen im Fraunhofer IRB mehrere mehrsprachige Thesauri vor. Deswegen war die Kombination (Vereinigung oder Verknüpfung) von bestehenden Thesauri die naheliegende Methode.

Die Verknüpfung wird typischerweise durchgeführt, um einen einheitlichen Zugang zu verteilten heterogenen Informationsbeständen mehrerer Online-Anbieter zu ermöglichen. Jeder Online-Anbieter

14 „SKOS Current Status,“ W3C, zuletzt geprüft am 02.09.2016, https://www.w3.org/standards/techs/skos#w3c_all.
15 Vgl. „SKOS/Datasets,“ W3C, zuletzt geprüft am 02.09.2016, https://www.w3.org/2001/sw/wiki/SKOS/Datasets.
16 Vgl. Dean Allemang and James Hendler, *Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL*, 2. Auflage (Amsterdam u.a.: Morgan Kaufmann, 2011), 207-219.
17 IFLA, *Guidelines for Multilingual Thesauri* (The Hague: International Federation of Library Associations and Institutions, 2009), 2.

verwendet dabei für die Suche in eigenen Informationsbeständen eigene kontrollierte Vokabularien. Da es sich nicht um den Zugang zu verteilten Informationsbeständen, sondern um den Zugang zu unserer Datenbank RSWB®plus handelte, schien es sinnvoll, eine Vereinigung der bestehenden Thesauri in einem neuen Thesaurus durchzuführen und den neuen Thesaurus für die Suche in RSWB®plus zu verwenden.

2.5. Vereinigung der Thesauri

Unter der Vereinigung von Thesauri wird hier die Erstellung eines neuen Thesaurus auf Basis von Quell-Thesauri verstanden, wobei der neue Thesaurus die alten ersetzt.[18] Die Vereinigung bezieht sich sowohl auf Begriffe als auch auf Beziehungen der Quell-Thesauri. Je nach der Thematik der Quell-Thesauri unterscheidet man zwischen der Vereinigung von thematisch gleichen Thesauri (Merging) und der Vereinigung von thematisch ergänzenden Thesauri (Integration). Da die Thesauri, welche im Fraunhofer IRB vorliegen, ähnliche Themenbereiche enthalten, schien es sinnvoll, hauptsächlich Merging anzuwenden.

Da eine intellektuelle Vereinigung der Thesauri zeitaufwändig und teuer wäre, ist es sinnvoll, den Vorgang zu automatisieren. Für die Automatisierung der Vereinigung von Thesauri kann man kommerzielle Software, z.B. von Synaptica oder Wordmap, verwenden oder eigene Programme entwickeln.[19] Aus Kostengründen erscheint uns vorerst eine Eigenentwicklung besser geeignet. Man kann z.B. ein automatisches Verfahren implementieren, das von Lacasta, Nogueras-Iso und Zarazags-Soria ausführlich beschrieben wurde.[20] Nach diesem Verfahren erfolgt die Vereinigung auf folgende Weise: Zuerst werden die Quell-Thesauri in ein einheitliches Format, SKOS, umgewandelt. Dann werden Cluster, d.h. Gruppen von äquivalenten Begriffen, gebildet, wobei zwei Begriffe als äquivalent gelten, wenn sie mindestens eine gemeinsame Benennung enthalten. Aufgrund von Relationen zwischen den Quell-Begriffen werden Relationen zwischen den Clustern hergestellt. Schließlich werden die Cluster in Begriffe eines neuen Thesaurus umgewandelt, der auch in SKOS dargestellt wird.

Im Folgenden werden die Thesauri genauer betrachtet, welche vereinigt werden sollten.

3. Quell-Thesauri

3.1. FINDEX Bau

Das Facettenartige Indexierungssystem für das Bauwesen (FINDEX Bau) wurde vom Fraunhofer-Informationszentrum Raum und Bau (IRB) erstellt.[21] Der Thesaurus besteht aus einem systematischen und einem alphabetischen Teil und ist termbasiert. Der systematische Teil besteht aus vier Ebenen, wobei die erste Ebene 20 Hauptbenennungen enthält, z.B. „Bauart", „Baudurchführung",

18 Vgl. Wolfgang Stock und Mechtild Stock, *Wissensrepräsentation: Informationen auswerten und bereitstellen* (München: Oldenbourg Wissenschaftsverlag, 2008), 303 ff.

19 Heather Hedden, „Three M's: Mapping, Merging, and Multilingual Taxonomies" (Vortrag auf der Special Librarians Association Annual Conference, Chicago, Ill., 15.-18. Juli, 2012), Vortragsfolien, zuletzt geprüft am 02.09.2016, http://www.hedden-information.com/HeatherHedden-ThreeMs-SLA2012.pdf.

20 Vgl. Lacasta, Nogueras-Iso und Zarazags-Soria, *Terminological Ontologies*, 77 ff.

21 Fraunhofer-Informationszentrum Raum und Bau, Hrsg., *FINDEX Bau: Facettenartiges Indexierungssystem für das Bauwesen*, 2. Auflage (Stuttgart: IRB Verlag, 1985).

„Baunutzung“. Durch diese Hauptbenennungen wird der Thesaurus in separate Bereiche (Facetten) aufgeteilt. Im alphabetischen Teil wird zwischen Deskriptoren und Nichtdeskriptoren unterschieden. Zwischen Nichtdeskriptoren und Deskriptoren können folgende Äquivalenz-Beziehungen bestehen:

- BD – benutze Deskriptor; für den angeführten Nichtdeskriptor ist der nachfolgende Deskriptor zu verwenden, zum Beispiel: Wärmeisolierung BD Wärmedämmung.
- BF – benutzt für; der Deskriptor wird anstelle des angeführten Nichtdeskriptors verwendet, zum Beispiel: Wärmedämmung BF Wärmeisolierung.

Der Thesaurus enthält ca. 6500 Terme und ist zweisprachig: Deutsch und Englisch.

3.2. FINDEX Raum

Das Facettenartige Indexierungssystem für Raumordnung, Städtebau, Wohnungswesen (FINDEX Raum) wurde vom Fraunhofer-Informationszentrum Raum und Bau (IRB) erstellt und ist ähnlich wie FINDEX Bau strukturiert.[22] Der Thesaurus besteht aus einem systematischen und einem alphabetischen Teil und ist termbasiert. Der systematische Teil besteht aus vier Ebenen, wobei die erste Ebene 18 Hauptbenennungen enthält, z.B. „Raum und Siedlung“, „Stadt/Verwaltung“, „Technische Infrastruktur“. Im alphabetischen Teil wird zwischen Deskriptoren und Nichtdeskriptoren unterschieden. Zwischen Nichtdeskriptoren und Deskriptoren können, ähnlich wie in FINDEX Bau, Äquivalenz-Beziehungen BD (benutze Deskriptor) und BF (benutzt für) bestehen. Darüber hinaus können zwischen verwandten Deskriptoren Assoziationsbeziehungen SA (siehe auch) bestehen, zum Beispiel:

Flussbau SA Wasserwegebau.

Der Thesaurus enthält ca. 2300 Terme und ist zweisprachig: Deutsch und Englisch.

3.3. TCCS

Der Canadian Thesaurus of Construction Science and Technology (TCCS) wurde von der IF Research Group, University of Montreal, Kanada, erstellt.[23] Der Thesaurus befasst sich mit dem Bauwesen und ist termbasiert. Es wird zwischen Deskriptoren und Nichtdeskriptoren unterschieden. Zu den unterstützten Beziehungen gehören u.a. Äquivalenz (US - use term, UF - use term instead), Assoziation (AT - associated term, RT - related term), Hierarchie (BT - broader term, NT - narrower term), Ganzes/Teil (WT - whole term, PT - part term). Der Thesaurus enthält ca. 15000 Terme. Ursprünglich unterstützte TCCS zwei Sprachen, Englisch und Französisch. Im Rahmen eines Kooperationsprojektes, an welchem die University of Montreal und Fraunhofer IRB beteiligt waren, wurden TCCS-Terme auch auf Deutsch und Spanisch übersetzt.

22 Fraunhofer-Informationszentrum Raum und Bau, Hrsg., *FINDEX Raum: Facettenartiges Indexierungssystem für Raumordnung, Städtebau, Wohnungswesen (Stuttgart: IRB Verlag, 1985).*

23 „Canadian Thesaurus of Construction Science and Technology,“ NRC, zuletzt geprüft am 02.09.2016, http://irc-wae.irc.nrc.ca/thesaurus/welcome.html.

4. Erzeugung des Ziel-Thesaurus

Für die Vereinigung (Merging) der Quell-Thesauri wird ein Verfahren verwendet, das im Wesentlichen dem Ansatz von Lacasta, Nogueras-Iso und Zarazags-Soria ähnelt (vgl. Kap. 2.5). Bei der Erzeugung des Zielthesaurus werden folgende Schritte durchgeführt:

- Umwandlung von Quell-Thesauri in SKOS-Format
- Bilden von Clustern
- Umwandlung von Relationen zwischen Begriffen in Relationen zwischen Clustern
- Erzeugung von neuen Begriffen aus Clustern
- Ausgabe des neuen Thesaurus in SKOS

Das o.g. Verfahren wird im Folgenden anhand vereinfachter Beispiele verdeutlicht. Abbildung 4 zeigt die Umwandlung der Einträge der Quell-Thesauri in entsprechende SKOS-Begriffe. Links oben sind Einträge vom FINDEX BAU in deutscher und englischer Fassung dargestellt. Der Deskriptor „Belastungsversuch“ hat den Identifier „16.080.010.4“ und befindet sich in der Äquivalenzbeziehung BF („benutze für“) mit dem Term „Belastungsprobe“. Ein entsprechender englischer Deskriptor „loading test“ hat denselben Identifier. Der deutsche Eintrag und der englische Eintrag aus FINDEX BAU werden in einen SKOS-Begriff mit dem Identifier „ts:160800104“ umgewandelt, der rechts oben dargestellt wird. Der Begriff hat eine deutsche Vorzugsbenennung „Belastungsversuch“, eine englische Vorzugsbenennung „loading test“ und eine deutsche Alternativbenennung „Belastungsprobe“. Links unten kann man einen englischen Deskriptor „loading tests“ mit entsprechendem deutschen Term „Belastungstests“ sehen. Der TCCS-Eintrag wird in einen SKOS-Begriff mit dem Identifier „ts:CAN21889782235“ umgewandelt, der rechts unten dargestellt wird. Die Terme des ursprünglichen Eintrages wurden dabei in die Singularform gebracht. Der Begriff hat eine deutsche Vorzugsbenennung „Belastungstest“ und eine englische Vorzugsbenennung „loading test“. Da die beiden SKOS-Begriffe, welche aus FINDEX- und TCCS-Einträgen erzeugt wurden, eine gemeinsame Benennung „loading test“ haben, kann man die Begriffe in einem Cluster vereinigen.[24] Der Cluster ist in Abbildung 5 rechts oben dargestellt und enthält Verweise auf die beiden Quell-Begriffe. Der Cluster wird schließlich in einen Ergebnis-Begriff des Ziel-Thesaurus umgewandelt. Der Ergebnis-Begriff ist in Abbildung 5 rechts unten dargestellt. Er hat eine deutsche Vorzugsbenennung „Belastungsversuch“, eine englische Vorzugsbenennung „loading test“ und zwei deutsche Alternativbenennungen „Belastungsprobe“ und „Belastungstest“.

24 Bei dem Bilden von Clustern besteht ein Fehlerrisiko, wenn unterschiedliche Begriffe gleiche Benennungen (Homonyme) haben. Zum Beispiel wird der Term „Bank“ im FINDEX Bau als eine englische Vorzugsbenennung für *Ufer* und im FINDEX Raum als eine deutsche und eine englische Alternativbenennung für *Kreditunternehmen* verwendet. Um die Vereinigung solcher Begriffe in einem Cluster zu vermeiden, werden homonyme Benennungen anhand eines Wörterbuchs erkannt und bei der Clusterbildung nicht berücksichtigt.

Einträge in Quell-Thesauri	SKOS-Begriffe
FINDEX Bau Belastungsversuch 16.080.010.4 BF Belastungsprobe; loading test 16.080.010.4; **TCCS** loading tests DT Belastungstests …	**FINDEX Bau** ts:BAU16080010004 rdf:type skos:Concept; skos:prefLabel "Belastungsversuch"@de; skos:prefLabel "loading test"@en; skos:altLabel "Belastungsprobe"@de. **TCCS** ts:CAN21889782235 rdf:type skos:Concept; skos:prefLabel "Belastungstest"@de; skos:prefLabel "loading test"@en.

Abb. 4: Umwandlung der Quell-Thesaurus-Einträge in SKOS

Quell-Begriffe in SKOS	Cluster
ts:BAU16080010004 rdf:type skos:Concept; skos:prefLabel "Belastungsversuch"@de; skos:prefLabel "**loading test**"@en; skos:altLabel "Belastungsprobe"@de. ts:CAN21889782235 rdf:type skos:Concept; skos:prefLabel "Belastungstest"@de; skos:prefLabel "**loading test**"@en.	ts:BAU16080010004 ts:CAN21889782235 **Ergebnis-Begriff** ts:F21889782235 rdf:type skos:Concept; skos:prefLabel "Belastungsversuch"@de; skos:prefLabel "loading test"@en; skos:altLabel "Belastungsprobe"@de; skos:altLabel "Belastungstest"@de.

Abb. 5: Bilden eines Clusters und Erzeugung eines Ergebnis-Begriffes

5. Einbindung des Thesaurus in die Suche

Der Thesaurus wird in die Suche in der Datenbank RSWB®plus eingebunden, indem die Anfragen um die Terme des Thesaurus erweitert werden. Die Erweiterung der Anfrage kann sowohl automatisch als auch manuell erfolgen.

Bei der automatischen Erweiterung der Suchanfrage findet ein Computerprogramm alle SKOS-Begriffe, welche die Terme der Anfrage enthalten. Die Anfrage wird dann automatisch um alle Vorzugs- und Alternativbenennungen erweitert, die mit den Begriffen verbunden sind. Die automatische Einbindung der Suchanfrage wird anhand des Beispiels in Abbildung 6 verdeutlicht. Die primäre Anfrage enthält den Term „Wärmedämmung“. Für den Term wird automatisch ein entsprechender Begriff mit dem Identifier „ts:F1011250305“ gefunden. Die primäre Anfrage wird dann automatisch um eine englische Vorzugsbenennung „thermal insulation“, eine deutsche Alternativbezeichnung „Waermeisolierung“ und eine englische Alternativbezeichnung „heat insulation“ erweitert. Alle

Begriffe der erweiterten Anfrage werden durch den Booleschen Operator „or“ verknüpft. Die manuelle Erweiterung einer Anfrage erfolgt ähnlich der automatischen Erweiterung. Im Unterschied zur automatischen Erweiterung wird der gefundene Thesaurus-Begriff bzw. werden die Begriffe zuerst angezeigt. Die Auswahl der in die Suche einzubeziehenden Begriffe und der entsprechenden Benennungen erfolgt aktiv durch die Nutzerinnen und Nutzer. Für die o.g. Anfrage „Wärmedämmung“ wird z.B. der Thesaurus-Begriff so angezeigt, wie in Abbildung 7 dargestellt. Um bestimmte Benennungen in eine Anfrage zu übernehmen, müssen die Auswahlkästchen links von den Termen markiert werden. Um eine zweisprachige Suche mit der o.g. Anfrage durchzuführen, kann man z.B. das englische Schlagwort (Vorzugsbenennung) und das englische Synonym (Alternativbenennung) „heat insulation“ wählen und den Knopf „Übernehmen“ anklicken. Die Terme werden dann in die Anfrage übernommen, und mit dem primären Suchterm durch den Operator „or“ verknüpft.

Primäre Anfrage: Wärmedämmung

Der gefundene SKOS-Begriff:

ts:F10112503050 rdf:type skos:Concept;

skos:prefLabel "Waermedaemmung"@de;

skos:prefLabel "thermal insulation"@en;

skos:altLabel "Waermeisolierung"@de;

skos:altLabel "heat insulation"@en;

skos:broader ts:F21742064828.

Erweiterte Anfrage: Wärmedämmung or thermal insulation or Waermeisolierung or heat insulation

Abb. 6: Automatische Erweiterung einer Anfrage

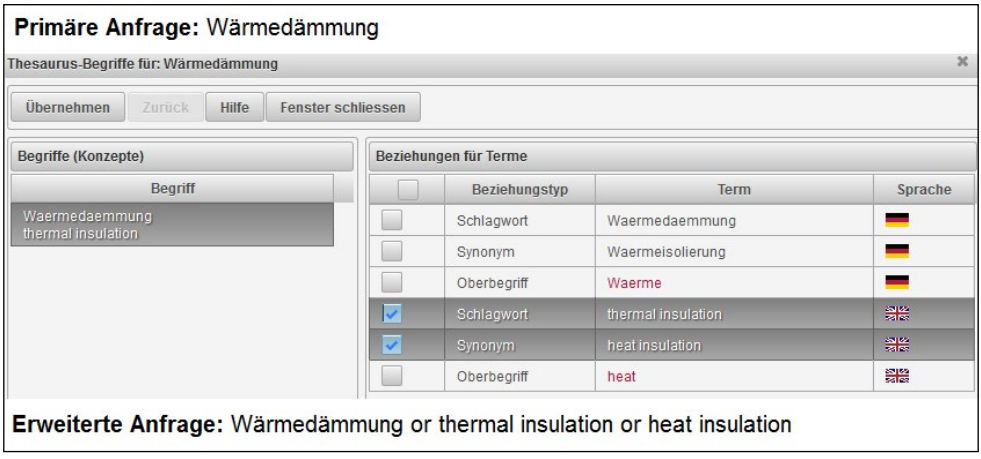

Abb. 7: Manuelle Erweiterung einer Anfrage

6. Software

Der neue Thesaurus wird als SKOS-Datei (RDF-TURTLE) gespeichert. Für die Arbeit mit dem Thesaurus wurden Java-Beans entwickelt, welche über Apache Jena auf den Thesaurus zugreifen. Apache Jena ist ein Open-Source-Framework für die Entwicklung von Semantic-Web- und Linked-Data-Anwendungen.[25]

Der Thesaurus wird momentan in die RSWB®plus -Datenbankanwendung eingebunden. RSWB®plus ist eine webbasierte Datenbankanwendung, welche das Datenbanksystem Microsoft SQL Server verwendet und unter Apache Tomcat und Microsoft Windows läuft. Wie erwähnt, wird bei der Thesaurus-Einbindung in Anwendungen nach Begriffen gesucht, welche bestimmte Terme enthalten. Um die Suche zu beschleunigen, enthält RSWB®plus eine spezielle Suchtabelle. Für jeden Begriff des Thesaurus wird in die Tabelle eine entsprechende Zeile eingetragen, welche den Identifikator des Begriffes sowie ein Textfeld enthält, dessen Wert aus Benennungen (Termen) des Begriffs besteht. Zum Beispiel werden für den Begriff aus Abbildung 6 sein Identifikator „ts:F10112503050“ und der Text „waermedaemmung; waermeisolierung; thermal insulation; heat insulation“ in die Suchtabelle eingetragen. Für die Suchtabelle wird ein Volltextindex angelegt, in dem eine schnelle Suche nach Texten (und entsprechenden Begriffen), welche bestimmte Terme enthalten, ermöglicht wird.

7. Fazit und Ausblick

Der in dem Artikel vorgestellte Ansatz zur mehrsprachigen Suche unter der Nutzung eines Thesaurus bietet folgende Vorteile:

- Deutschsprachige Benutzerinnen und Benutzer können auch englischsprachige Metadaten-Einträge finden. Damit wird die Effektivität der Recherchen erhöht.
- Die manuelle Führung und Strukturierung der Thesauri ist ein zeitaufwändiger und teurer Vorgang. Die Automatisierung der Erstellung eines neuen Thesaurus spart Zeit und Kosten.
- Durch die Umwandlung der Quell-Thesauri in ein einheitliches standardisiertes Format, SKOS, wird ihre Weiterverarbeitung (Abgleich und Vereinigung) erleichtert, da es bereits freie Software für die Arbeit mit SKOS/RDF gibt. Da der neue Thesaurus auch in SKOS geführt wird, kann man ihn nicht nur in RSWB®plus, sondern in andere Anwendungen einbinden. Durch die Standardisierung des Formats wird also Vergleichbarkeit (Kompatibilität) und die Zusammenarbeit (Interoperabilität) der Thesauri miteinander und mit anderen Anwendungen verbessert.[26]
- Da in dem Themengebiet immer wieder neue Begriffe erscheinen, wäre es sinnvoll, den Thesaurus künftig regelmäßig zu aktualisieren. Man kann z.B. Einträge von anderen freien Thesauri (falls vorhanden) in den Thesaurus übernehmen.

25 Vgl. „Apache Jena“, Apache, zuletzt geprüft am 02.09.2016, https://jena.apache.org/.
26 Vgl. Stock und Stock, *Wissensrepräsentation*, *293*.

Literaturverzeichnis

- Apache. „Apache Jena." Zuletzt geprüft am 02.09.2016. https://jena.apache.org.
- Allemang, Dean und James Hendler. *Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL*, 2. Auflage. Amsterdam u.a.: Morgan Kaufmann, 2011.
- Buitelaar, Paul, Klaus Netter und Feiyu Xu, „Integrating Different Strategies in Cross-Language Information Retrieval in the MIETTA Project." In *Language technology in multimedia information retrieval: Proceedings of the Fourteenth Twente Workshop on Language Technology, December 7-8, 1998, Enschede, The Netherlands*, herausgegeben von Djoerd Hiemstra, Franciska de Jong und Klaus Netter. Enschede: Univ. Twente, 1998, 9-17.
- *Fraunhofer-Informationszentrum Raum und Bau, Hrsg. FINDEX Bau: Facettenartiges Indexierungssystem für das Bauwesen*, 2. Auflage. Stuttgart: IRB Verlag, 1985.
- Fraunhofer-Informationszentrum Raum und Bau, Hrsg. *FINDEX Raum: Facettenartiges Indexierungssystem für Raumordnung, Städtebau, Wohnungswesen*, 1. Auflage. Stuttgart: IRB Verlag, 1985.
- DIN-Normenausschuss Information und Dokumentation (NID) im DIN e.V., *DIN 1463-1, Erstellung und Weiterentwicklung von Thesauri. Einsprachige Thesauri*. Berlin: Beuth, 1987.
- DIN-Normenausschuss Terminologie (NAT) im DIN e.V., *DIN 2342-1, Begriffe der Terminologielehre. Grundbegriffe*. Berlin: Beuth, 1992.
- Hedden, Heather. „Three M's: Mapping, Merging, and Multilingual Taxonomies". Vortrag auf der Special Librarians Association Annual Conference, Chicago, Ill., 15.-18. Juli, 2012. Vortragsfolien. Zuletzt geprüft am 02.09.2016. http://www.hedden-information.com/HeatherHedden-ThreeMs-SLA2012.pdf.
- IFLA. *Guidelines for Multilingual Thesauri*. The Hague: International Federation of Library Associations and Institutions, 2009.
- Lacasta, Javier, Javier Nogueras-Iso und Francisco Zarazags-Soria. *Terminological Ontologies: Design, Management and Practical Applications*. New York: Springer, 2010.
- McCarley, J. Scott. „Should We Translate the Documents or the Queries in Cross-Language Information Retrieval?" In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, College Park, 21. Juni 1999*, herausgegeben von Association for Computational Linguistics. 208-214. San Francisco: Morgan Kaufmann, 1999.
- Nie, Jian-Yun. *Cross-Language Information Retrieval*. San Rafael: Morgan & Claypool, 2010.

- NISO. „Format for Exchange of Thesaurus Data Conforming to ISO 25964-1.“ Zuletzt geprüft am 02.09.2016. http://www.niso.org/schemas/iso25964/schema-intro/.

- NRC. „Canadian Thesaurus of Construction Science and Technology.“ Zuletzt geprüft am 02.09.2016. http://irc-wae.irc.nrc.ca/thesaurus/welcome.html.

- Peters, Carol, Martin Braschler und Paul Clough. *Multilingual Information Retrieval: From Research to Practice*. Berlin Heidelberg: Springer, 2012.

- Stock, Wolfgang. *Information Retrieval: Informationen suchen und finden*. Einführung in die Informationswissenschaft 1. München: Oldenbourg Wissenschaftsverlag, 2007.

- Stock, Wolfgang und Mechtild Stock. *Wissensrepräsentation: Informationen auswerten und bereitstellen*. München: Oldenbourg Wissenschaftsverlag, 2008.

- W3C. „SKOS Current Status.“ Zuletzt geprüft am 02.09.2016. https://www.w3.org/standards/techs/skos#w3c_all.

- W3C. „SKOS/Datasets.“ Zuletzt geprüft am 02.09.2016. https://www.w3.org/2001/sw/wiki/SKOS/Datasets.

- W3C. „SKOS Simple Knowledge Organization System Primer.“ Zuletzt geprüft am 02.09.2016. https://www.w3.org/TR/skos-primer/.

