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

16. März 2016

Dimitri Busch

Fraunhofer Informationszentrum
Raum und Bau IRB, Stuttgart

6. Bibliothekskongress,

Leipzig, 14.-17. März 2016

Fraunhofer
IRB

Einführung

- Es geht um einen zweisprachigen Thesaurus, der u.a. für die Suche in der Datenbank RSWBPlus verwendet wird.
- RSWBPlus ist eine bibliographische Datenbank zum Nachweis der Fachliteratur im Bereich Planen und Bauen.
- RSWBPlus enthält deutschsprachige Einträge von der deutschen Baudatenbank RSWB und englischsprachige Einträge von der internationalen Baudatenbank ICONDA Bibliographic.

Deutscher Eintrag (RSWB)

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

Englischer Eintrag (ICONDA)

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.
Keyword	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

Problem

- Auf deutschsprachige Anfrage findet man nur deutschsprachige Einträge, obwohl die Datenbank auch potenziell nützliche englischsprachige Einträge enthalten kann. Auf englischsprachige Anfrage findet man nur dann deutschsprachige Einträge, wenn man nach Schlagwörtern sucht.
- Lösung – Mehrsprachige Suche: Suchanfrage in einer Sprache findet auch Einträge in anderen Sprachen.

Ansätze zur mehrsprachigen Suche

- Die Suchanfrage wird in die Sprache(n) der Einträge übersetzt (sprachübergreifende Suche, cross-language Information Retrieval)
- Einträge werden in die Anfragesprache übersetzt

Zu uns passt besser der erste Ansatz (sprachübergreifende Suche), da es zu zeitaufwändig und teuer würde, englische Einträge ins Deutsche zu übersetzen

Ansätze zur sprachübergreifenden Suche

(Peters et al., 2012, S. 59) , (Stock, 2007, S. 465)

- Übersetzung der Suchanfrage mittels maschinenlesbarer Wörterbücher, Thesauri usw.
- Übersetzung der Suchanfrage nach statistischen Verfahren unter Nutzung paralleler Korpora
- Die Nutzung eines „vollen“ Systems für maschinelle Übersetzung

Zu uns passt besser der erste Ansatz auf Basis eines Thesaurus. Der Thesaurus wird aus bereits bestehenden Thesauri in den Bereichen Bauwesen und Raumordnung erzeugt, welche dem Fraunhofer IRB vorliegen.

Quell-Thesauri

- **FINDEX BAU**; Autor: Fraunhofer IRB; Thema: Bauwesen; Beziehungen: Äquivalenz (BD, BF) und Hierarchie; facettenartig; zweisprachig: Deutsch und Englisch.
- **FINDEX RAUM**; Autor: Fraunhofer IRB; Themen: Raumordnung, Städtebau, Wohnungswesen; Beziehungen: Äquivalenz (BD, BF), Assoziation (SA) und Hierarchie; facettenartig; zweisprachig: Deutsch und Englisch.
- Canadian Thesaurus of Construction Science and Technology (**TCCS**); Autor: IF Research Group, University of Montreal; Thema: Bauwesen; Beziehungen: Äquivalenz (US, UF), Assoziation (AT, RT) , Hierarchie (BT,NT), Ganzes/Teil (WT, PT); mehrsprachig: Englisch, Französisch, Deutsch und Spanisch.

Format für die Repräsentation der Thesauri

- Alle Quell-Thesauri sind termbasiert, d.h. sie enthalten Terme und Beziehungen zwischen den Termen.
- Die Quell-Thesauri sind in unterschiedlichen Formaten dargestellt, was die gemeinsame Verarbeitung erschwert.
- Um die Verarbeitung der Thesauri zu erleichtern, werden sie in ein gemeinsames standardisiertes Format, SKOS (Simple Knowledge Organisation System) umgewandelt.
- Das SKOS-Format ist konzeptbasiert. Konzepte sind abstrakte Dinge, welche durch Terme bezeichnet werden. SKOS basiert auf RDF (Resource Description Framework).

Erzeugung des Ziel-Thesaurus

- Umwandlung von Quell-Thesauri in SKOS
- Bilden von Clustern. Ein Cluster ist eine Gruppe von äquivalenten Konzepten. 2 Konzepte gelten als äquivalent, wenn sie mindestens eine gemeinsame Bezeichnung haben
- Umwandlung von Beziehungen zwischen Konzepten in Beziehungen zwischen Clustern
- Erzeugung von neuen Konzepten aus Clustern
- Ausgabe des neuen Thesaurus in SKOS

Der Ansatz ähnelt sich dem Ansatz von Lacasta et al. (2010).

Umwandlung der Quell-Thesaurus-Einträge in SKOS: Beispiel

Einträge in Quell-Thesauri

FINDEX Bau
Belastungsversuch 16.080.010.4
BF Belastungsprobe;
loading test 16.080.010.4;

TCCS
loading test
DT Belastungstest

SKOS-Konzepte

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.

Bilden eines Clusters und Erzeugung eines Ergebniskonzepts : Beispiel

Quell-Konzepte in SKOS

```
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.
```

Cluster

```
ts:BAU16080010004
ts:CAN21889782235
```

Ergebnis-Konzept

```
ts:F21889782235 rdf:type skos:Concept;
   skos:prefLabel "Belastungsversuch"@de;
   skos:prefLabel "loading test"@en;
   skos:altLabel "Belastungsprobe"@de;
   skos:altLabel "Belastungstest"@de.
```

Automatische Einbindung des Thesaurus in die Suche

- Eingabe von Suchbegriffen (Termen).
- SKOS-Konzepte werden gefunden, welche die Suchbegriffe enthalten.
- Die Anfrage wird um alle bevorzugten und alternativen Bezeichnungen erweitert.

Automatische Einbindung des Thesaurus in die Suche: Beispiel

- Primäre Anfrage: Wärmedämmung
- SKOS-Konzept:

  ts:F10112503050 rdf:type skos:Concept;

  skos:prefLabel "Waermedaemmung"@de;

  skos:prefLabel "thermal insulation"@en;

  skos:altLabel "Waermeisolierung"@de;

  skos:altLabel "heat insulation"@en.
- Erweiterte Anfrage:

  Wärmedämmung

  or thermal insulation

  or Waermeisolierung

  or heat insulation

Einbindung des Thesaurus in die Suche: manuell

Eingangsanfrage: Wärmedämmung

Ausgangsanfrage: Wärmedämmung or thermal insulation or heat insulation

Software

- Java /JDBC
- Jena
- Microsoft SQL SERVER
- Apache Tomcat
- Microsoft Windows

Fazit und Ausblick

- Deutschsprachige Benutzer können auch englische Dokumente finden
- Iteroperabilität der Thesauri durch die Darstellung im standardisierten Format (vgl. ISO 25964-2, Abs. 3.38)
- SKOS (RDF) vereinfach das Bilden des Ziel-Thesaurus und den Zugriff zu seinen Einträgen, da es bereits freie Software für die Arbeit mit RDF gibt, z.B. Jena, Sesame.
- Der neue Thesaurus ist nicht an RSWBPlus gebunden und kann zukünftig auch in anderen Anwendungen verwendet werden.

Literatur

- Coprian, W.; Kaiser, K. (1985): FINDEX Bau. Stuttgart: IRB Verlag
- Fraunhofer IRB (1985): FINDEX. Facet-Oriented Indexing System for Architecture and Construction Engineering. Stuttgart: IRB Verlag
- ISO 25964-2: Information and documentation – Thesauri and interoperability with other vocabularies – Part 2: Interoperability with other vocabularies
- Koengeter, B. (1985): FINDEX Raum. Stuttgart: IRB Verlag
- Lacasta, J.; Nogueras-Iso, J.; Zarazags-Soria, F. (2010): Terminological Ontologies. New York: Springer
- Peters, C.; Braschler, M; Clough, P. (2012): Multilingual Information Retrieval. Heidelberg: Springer
- Stock, W. (2007): Information Retrieval. München: Oldenbourg
- TCCS, Canadian Thesaurus of Construction Science and Technology. Ottawa: Government of Canada, Industry, Trade and Commerce, 1978

Vielen Dank für Ihre Aufmerksamkeit!