Cross-Language Information Retrieval und automatische Sacherschließung in Suchmaschinen am Beispiel der „Bielefeld Academic Search Engine“ (BASE)

- Dirk Pieper, UB Bielefeld -

Inhalt:

- BASE: Einführung
- Cross-Language Information Retrieval (CLIR)
- Der EUROVOC-Thesaurus
- **Einbindung des EUROVOC-Thesaurus in BASE**
- Integration der Schlagwort-Normdatei in BASE
- Literatur und weitere Informationen (Auswahl)

BASE-Einführung

- BASE = Bielefeld Academic Search Engine
- www.base-search.net
- Wissenschaftliche Suchmaschine: OAI-Service-Provider plus ausgewählte wissenschaftliche Web-Quellen
- Inhalt: über 10,4 Mio. Dokumente aus mehr als 750 Quellen (Stand: Ende Mai 2008)

BASE-Einführung

- BASE-Features Suche: Suchhistorie, Suchverfeinerung, Sortierung, Google-Scholar-Integration, Suche über verschiedene Wortformen, multilinguale Suche, SWD
- Weitere BASE-Features: Transparenz, differenzierte Anzeige von bibliographischen Daten (sofern vorhanden), Volltextindexierung mit Verbindung zu den entsprechenden Metadaten (quellenabhängig), HTTP- und SOAP-Schnittstelle zum BASE-Index

BASE-Einführung

- Einige Meilensteine:

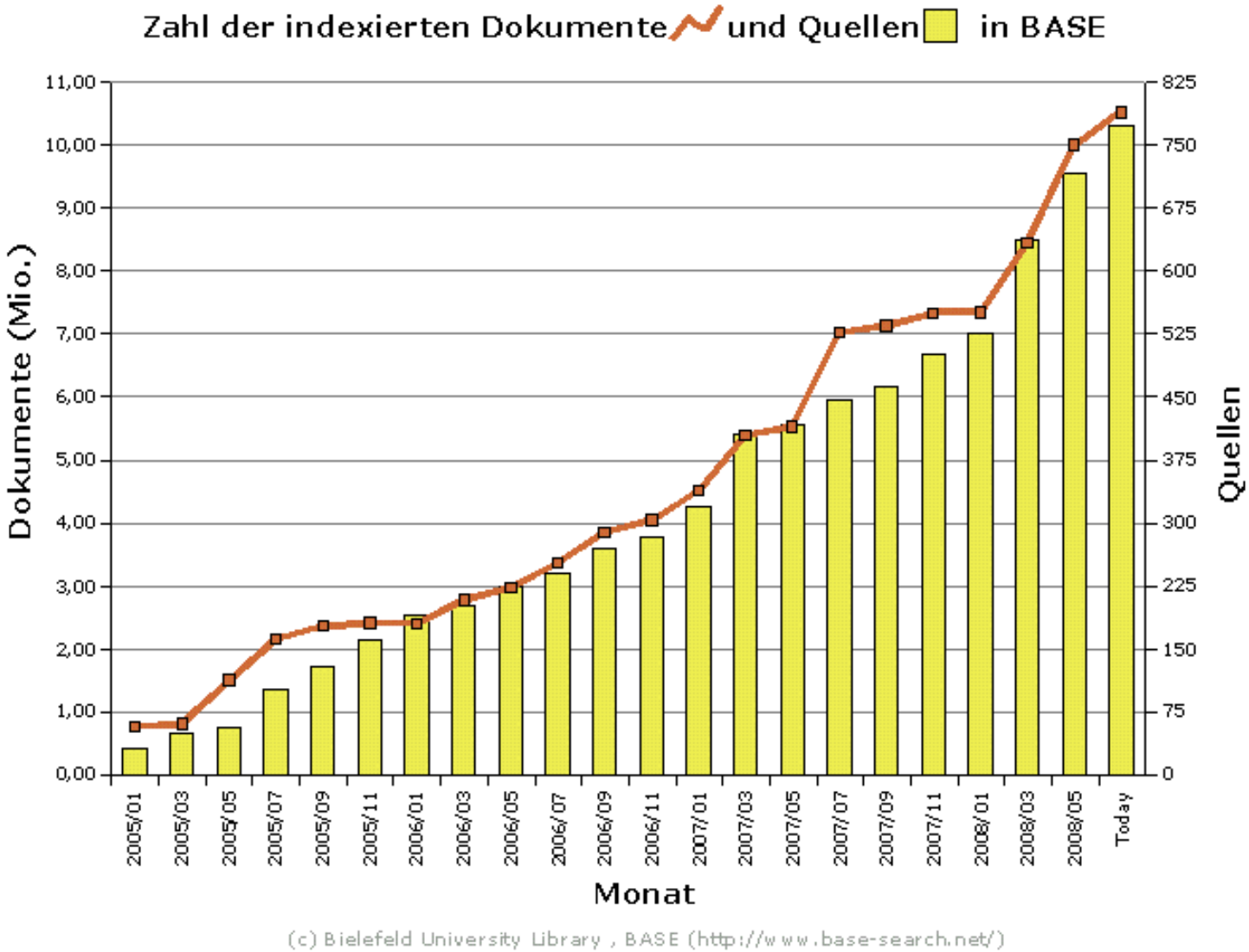
✓ Start 2004

✓ 2005: Suchhistorie, Sortierung, Uni-Suche

✓ 2006: Google-Scholar-Integration, Server-Farm, Beginn der Teilnahme am EU-Projekt DRIVER

✓ 2007: Multilinguale Suche (EUROVOC-Thesaurus)

✓ 2008: Index größer 10 Millionen Dokumente


Zahl der indexierten Dokumente und Quellen in BASE
Dokumente (Mio.)
Quellen
Monat
11,00
10,00
9,00
8,00
7,00
6,00
5,00
4,00
3,00
2,00
1,00
0,00
825
750
675
600
525
450
375
300
225
150
75
0
2005/01
2005/03
2005/05
2005/07
2005/09
2005/11
2006/01
2006/03
2006/05
2006/07
2006/09
2006/11
2007/01
2007/03
2007/05
2007/07
2007/09
2007/11
2008/01
2008/03
2008/05
Today
(c) Bielefeld University Library , BASE (http://www.base-search.net/)

Cross-Language Information Retrieval

- CLIR: Suchanfrage nach einem Dokument D in einer Sprache L findet auch Dokumente D' in der Sprache L'
- Problem: Repräsentation von Zeichen (z.B. Lexemen) und deren Semantik von L in L'
- Ein Lösungsansatz: Verwendung eines multilingualen Thesaurus

Der EUROVOC-Thesaurus

- Amt für amtliche Veröffentlichungen der Europäischen Gemeinschaften
- Beinhaltet in V4.2 rd. 6.500 Basisbegriffe in 21 Sprachen plus Synonyme sofern vorhanden (rd. 239.000 Einträge aus 21 Sprachen)
- http://europa.eu/eurovoc/sg/sga_doc/eurovoc_dif!SERVEUR/menu!prod!MENU?langue=DE

Einbindung des EUROVOC-Thesaurus in BASE

- Jeder Basisbegriff hat eine eindeutige ID über alle Sprachen
- Zu jedem Basisbegriff existieren used-for-terms (UF), Verknüpfung über die entsprechende ID
- Ausgangspunkt deutschsprachige Basisbegriffe: Mapping auf andere Sprachversionen 1 -> S_n
- Zweiter Schritt: 1 -> S_n + UF_n

Einbindung des EUROVOC-Thesaurus in BASE

- Beispiel Basisbegriff ID 1158 deutschsprachig

<RECORD>

<DESCRIPTEUR_ID>1158</DESCRIPTEUR_ID>

<LIBELLE>Abfallwirtschaft</LIBELLE>

</RECORD>

Einbindung des EUROVOC-Thesaurus in BASE

- Beispiel Basisbegriff ID 1158 -> S_n (Ausschnitt)

<term lang="de" id="de1158">

<label>Abfallwirtschaft</label>

<spellVar>управление на отпадък</spellVar>

<spellVar>nakládání s odpadem</spellVar>

<spellVar>forvaltning af affald</spellVar>

<spellVar>διαχείριση των αποβλήτων</spellVar>

<spellVar>waste management</spellVar>

<spellVar>gestión de residuos</spellVar>

<spellVar>jätehuolto</spellVar>

<spellVar>ravnanje z odpadki</spellVar>

<spellVar>atliekų tvarkyba</spellVar>

<spellVar>avfallshantering</spellVar>

... </term>

Einbindung des EUROVOC-Thesaurus in BASE

- Beispiel Basisbegriff ID 1158 -> UF

<RECORD>

<DESCRIPTEUR_ID>1158</DESCRIPTEUR_ID>

<UF><UF_EL>Abfallbehandlung</UF_EL>

<UF_EL>Abfallmanagement</UF_EL>

<UF_EL>öffentliche Deponie</UF_EL></UF>

</RECORD>

Einbindung des EUROVOC-Thesaurus in BASE

- ID 1158 -> S_n + UF_n (Ausschnitt)

<term lang="de" id="1158">

<label>Abfallwirtschaft</label>

<spellVar>управление на отпадък</spellVar> <spellVar>hulladékgazdálkodás</spellVar> <spellVar>Abfallbehandlung</spellVar>

<spellVar>nakládání s odpadem</spellVar> <spellVar>gestione dei rifiuti</spellVar> <spellVar>Abfallmanagement</spellVar>

<spellVar>forvaltning af affald</spellVar> <spellVar>atliekų tvarkyba</spellVar> <spellVar>öffentliche Deponie</spellVar>

<spellVar>διαχείριση των αποβλήτων</spellVar> <spellVar>atkritumu apsaimniekošana</spellVar> <spellVar>skládka odpadu</spellVar>

<spellVar>waste management</spellVar> <spellVar>beheer van afvalstoffen</spellVar> <spellVar>ravnanje z odpadki</spellVar>

<spellVar>gestión de residuos</spellVar> <spellVar>gospodarka odpadami</spellVar> <spellVar>avfallshantering</spellVar>

<spellVar>jätehuolto</spellVar> <spellVar>gestão de resíduos</spellVar> <spellVar>affaldsbehandling</spellVar>

<spellVar>gestion des déchets</spellVar> <spellVar>managementul deşeurilor</spellVar> ... </term>

Einbindung des EUROVOC-Thesaurus in BASE

- Ergebnis: desc_desc_ufall.xml
- Schritt 3: Erzeugung diverser Wörterbücher mit FAST-Dictionary-Tool (Lemmatisierung, Komposita-Zerlegung, **Permutationen**, Phrasierung, spell check, ...) S_n + UF_n -> S_{np} +UF_{np}
- Erweiterung der FAST-Query-Pipelinestufen
- Einbindung des Thesaurus über optionale Erweiterung der Queries
- Vorteile: keine Indexerweiterung nötig, neue Thesaurus-Versionen sind leicht aktualisierbar

Einbindung des EUROVOC-Thesaurus in BASE

- Ergebnis Suchanfrage ohne Query-Erweiterung (Ausschnitt)

<SNIP>

<QUERYTRANSFORMS><QUERYTRANSFORM NAME="Original query" ACTION="NOP" QUERY="abfallwirtschaft" CUSTOM=""
MESSAGE="Original query" MESSAGEID="1"/>

<QUERYTRANSFORM NAME="FastQT_DefaultIndex" ACTION="Suggested new query" QUERY="lemcontent:abfallwirtschaft"
CUSTOM="" MESSAGE="Default index suggested for textual terms in the query" MESSAGEID="2"/></QUERYTRANSFORMS>

<RESULTSET FIRSTHIT="1" LASTHIT="10" HITS="10" TOTALHITS="293" MAXRANK="15937" TIME="0.0461">

<HIT NO="1" RANK="5469" SITEID="0" MOREHITS="0">

</SNIP>

Einbindung des EUROVOC-Thesaurus in BASE

- Ergebnis Suchanfrage mit Query-Erweiterung mit einfacher Spracherweiterung (Ausschnitt)

<SNIP>

QUERYTRANSFORMS><QUERYTRANSFORM NAME="Original query" ACTION="NOP" QUERY="abfallwirtschaft" CUSTOM="" MESSAGE="Original query" MESSAGEID="1"/>

<QUERYTRANSFORM NAME="FastQT_Synonym" ACTION="Modified the query" QUERY="(abfallwirtschaft)" CUSTOM="" MESSAGE="Inserted parentheses to indicate words for which synonyms are available." MESSAGEID="32"/>

<QUERYTRANSFORM NAME="FastQT_Synonym" ACTION="Modified the query" QUERY="(abfallwirtschaft "atkritumu apsaimniekošana" "atliekų tvarkyba" avfallshantering "beheer van afvalstoffen" "forvaltning af affald" "gestion des déchets" "gestione dei rifiuti" "gestión de residuos" "gestão de resíduos" "gospodarenje otpadom" "gospodarka odpadami" hulladékgazdálkodás jätehuolto "managementul deşeurilor" "nakládání s odpadem" "odpadové hospodárstvo" "ravnanje z odpadki" "waste management" "διαχείριση των αποβλήτων" "управление на отпадък")" CUSTOM="" MESSAGE="Semantic query transformation performed on parts of the query." MESSAGEID="16"/></QUERYTRANSFORMS>

<RESULTSET FIRSTHIT="1" LASTHIT="10" HITS="10" TOTALHITS="2735" MAXRANK="48799" TIME="0.0179">

</SNIP>

Einbindung des EUROVOC-Thesaurus in BASE

- Ergebnis Suchanfrage mit Query-Erweiterung mit Sprach-erweiterung und used-for-terms (Ausschnitt)

<SNIP><QUERYTRANSFORMS><QUERYTRANSFORM NAME="Original query" ACTION="NOP" QUERY="abfallwirtschaft" CUSTOM="" MESSAGE="Original query" MESSAGEID="1"/><QUERYTRANSFORM NAME="FastQT_Synonym" ACTION="Modified the query" QUERY="(abfallwirtschaft)" CUSTOM="" MESSAGE="Inserted parentheses to indicate words for which synonyms are available." MESSAGEID="32"/>

<QUERYTRANSFORM NAME="FastQT_Synonym" ACTION="Modified the query" QUERY="(abfallbehandlung abfallmanagement abfallwirtschaft affaldsbehandling afvalbeheer "atkritumu apsaimniekošana" "atkritumu apstrāde" "atkritumu izgāztuve" "atkritumu poligons" "atkritumu pārvaldība" "atliekų tvarkyba" "atliekų tvarkymas" avfallshantering "behandeling van afvalstoffen" "behandling av avfall" "beheer van afvalstoffen" "discarica pubblica" "décharge publique" "forvaltning af affald" "gestion des déchets" "gestione dei rifiuti" "gestión de desechos" "gestión de residuos" "gestão de resíduos" "gospodarenje otpadom" "gospodarka odpadami" "hospodaření s odpadem" hulladékgazdálkodás hulladékkezelés jätehuolto jätteenkäsittely "landfill site" lixeira "managementul deşeurilor" "nakládání s odpadem" "odpadové hospodárstvo" "odpadové hospodářství" "offentlig deponi" "offentlig losseplads" "openbare stortplaats" "ravnanje z odpadki" "rubbish dump" "skládka odpadu" szeméttelep sąvartynas "traitement des déchets" "tratamento de resíduos" "tratamiento de desechos" "tratamiento de residuos" "trattamento dei rifiuti" "waste management" "waste treatment" "yleinen kaatopaikka" "öffentliche deponie" "úprava odpadu" "šiukšlių sąvartynas" "δημόσιος χώρος διάθεσης" "διαχείριση των αποβλήτων" "κατεργασία αποβλήτων" "управление на отпадък")" CUSTOM="" MESSAGE="Semantic query transformation performed on parts of the query." MESSAGEID="16"/></QUERYTRANSFORMS>

<RESULTSET FIRSTHIT="1" LASTHIT="10" HITS="10" TOTALHITS="3615" MAXRANK="48799" TIME="0.0536"></SNIP>

Integration der Schlagwort-Normdatei in BASE

- SWD wird in BASE nur für OAI-Metadaten mit Sprachkennung “de” verwendet
- Begrenzung auf Sachschlagwörter (rd. 160.000 Datensätze)
- Unterschiede zum EUROVOC-Thesaurus: Benutzung sprachabhängiger FAST-Dictionary-Tools, Index-Erweiterung statt Query-Erweiterung

Integration der Schlagwort-Normdatei in BASE

2. Technisch-organisatorischer Wandel und europäischer Binnenmarkt

» Treffer in neuem Browser-Fenster öffnen

Titel: Technisch-organisatorischer Wandel und europäischer Binnenmarkt
Autor: Spannhake, B.
Schlagwörter: Projektträger des BMBF für Arbeit, Umwelt und Gesundheit ; Bauarbeit, Bauhandwerk, Überalterung, Belastung, Gesundheitsschutz, Prävention ; erwerbsarbeit ; gesundheitsstatus ; gesundheitszustand ; fitness ; krankheit ; last ; beanspruchung ; lastannahme ; verhütung ; vorbeugung ; prophylaxe ; vorsorge ; verhinderung
Inhalt: Spannhake, B. ; Technisch-organisatorischer Wandel und **europäischer Binnenmarkt** ; Projektträger des BMBF für Arbeit, Umwelt und Gesundheit ; Bauarbeit, Bauhandwerk, Überalterung, Belastung, Gesundheitsschutz...
Veröffentlicht: 1992

http://elib.dlr.de/39280/ (0.8k) [HTML]
Datenlieferant Deutsches Zentrum für Luft und Raumfahrt: elib - DLR electronic library
» Diesen Titel in Google Scholar suchen

Literatur und weitere Informationen (Auswahl)

- Cross-Language Evaluation Forum (CLEF). http://www.clef-campaign.org/
- EU-Projekt „Cross-language Access to Catalogues And On-line libraries“ (CACAO). http://www.cacaoproject.eu/
- Kremp, M. (2007): Chinesische Web-Surfer bleiben unter sich. In: Spiegel Online vom 29. Oktober 2007. http://www.spiegel.de/netzwelt/web/0,1518,514120,00.html
- Lewandowski, D. (2008): Problems with the use of Web search engines to find results in foreign languages. In: Online Information Review 32 (to appear). Preprint: http://www.durchdenken.de/lewandowski/doc/OIR2008_Preprint.pdf
- Oard, D.W. (1997): Serving Users in Many Languages. Cross-Language Information Retrieval for Digital Libraries. In: D-Lib Magazine, December 1997. http://www.dlib.org/dlib/december97/oard/12oard.html
- Pieper, D./Summann, F. (2006): Bielefeld Academic Search Engine (BASE): An end-user oriented institutional repository search service. In: Library Hi Tech, Bd. 24, Nr. 4, S. 614 – 619. http://www.emeraldinsight.com/10.1108/07378830610715473

Literatur und weitere Informationen (Auswahl)

- Pieper, D./Wolf, S. (2007): BASE - Eine Suchmaschine für OAI-Quellen und wissenschaftliche Webseiten. In: Information, Wissenschaft & Praxis, Bd. 58, Nr. 3, S. 179 – 182.
- Plank, B. (2006): Multilingual Access to Library Catalogues. Students Symposium, LCT Colloquia Free University of Bolzano- Bozen, October 26, 2006. http://www.inf.unibz.it/mcs/lct/slides/symposium06/plank.pdf
- Potthast, M./Stein, B./Anderka, M. (2008): A Wikipedia-Based Multilingual Retrieval Model. In: Macdonald, C./Ounis, I./ Plachouras, V./Ruthven, I./White, R.W. (Ed.): 30th European Conference on IR Research, ECIR 2008, Glasgow, Vol. 4956 of Lecture Notes in Computer Science, S. 522 – 530. http://www.uni-weimar.de/medien/webis/publications/downloads/papers/stein_2008b.pdf
- Youssef, M.A. (2001): Cross Language Information Retrieval, University of Maryland, Department of Computer Science, April 2001. http://www.otal.umd.edu/uupractice/clir/

Vielen Dank für Ihre Aufmerksamkeit!