Terminology for information retrieval: effectiveness of cross-concordances

Philipp Mayr & Vivien Petras
GESIS Social Science Information Centre, Bonn, Germany

Knowledge organization on the Web
ISKO-IWA meeting

Naples, Italy, 5 September 2008
German Social Science Infrastructure Services

www.gesis.org

- Digital Library
- Data archive
- Consulting
- Surveys & studies
- Society observation
German Social Science Infrastructure Services

Document types:
- Bibliographic
- Full texts
- Project data
- Institutions
- Web pages
- Statistical data
- Surveys
- People

Disciplines:
- Sociology
- Political Science
- Education
- Psychology
- Economics
- Business Administration
Heterogeneous collections

- Many databases:
  - document types / formats
  - vocabularies

- controlled vocabularies:
  - internal consistency \( \uparrow \)
  - intersystem compatibility \( \downarrow \) (semantic heterogeneity)

- Solution: translate \( \rightarrow \) cross-walks \( \rightarrow \) terminology mapping
KoMoHe Project

- September 2004 – 2007
- Goals:
  - Models for searching heterogeneous collections
  - Development, organization & management of cross-walks between controlled vocabularies
Terminology Mapping Initiatives

- **OCLC Terminology Services**
  - DDC, LCC, LCSH, Mesh

- **MACS (Multilingual Access to Subjects)**
  - LCSH – Rameau – SWD

- **CARMEN**
  - SWD, TheSoz, STW, …

- **Criss-Cross**
  - SWD – DDC
Cross-concordances

= manually created, directed relations between controlled terms of two knowledge organization systems (KOS)
**Relations**

- Equivalence
- Narrower Term
- Broader Term
- Related Term
- Null: no mapping

<table>
<thead>
<tr>
<th>KOS 1</th>
<th>Relation</th>
<th>KOS 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Library</td>
<td>=</td>
<td>Bibliothéque</td>
</tr>
<tr>
<td>Library</td>
<td>&gt;</td>
<td>Special library</td>
</tr>
<tr>
<td>Thesaurus</td>
<td>&lt;</td>
<td>KOS</td>
</tr>
<tr>
<td>Hacker</td>
<td>^</td>
<td>Computers + Security</td>
</tr>
<tr>
<td>Virus</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>
Cross-concordances

• 25 Vocabularies in 60 cross-concordances
  – Thesauri (16)
  – Descriptor lists (4)
  – Classifications (3)
  – Subject heading lists (2)

• 380,000 mapped terms
• 465,000 relations
• 205,000 equivalence relations
• 13 German, 8 English, 1 Russian, 3 multilingual
Disciplines

- Agricultural science (1)
- Information science (1)
- Gerontology (1)
- Medicine (1)
- Universal (3)
- Political science (3)
- Economics (2)
- Social Sciences (10)
- Pedagogics (1)
- Sports science (2)
- Psychology (1)
- Information science (1)
- Agricultural science (1)
- Universal (3)
- Social Sciences (10)
- Pedagogics (1)
- Psychology (1)
- Gerontology (1)
- Political science (3)
- Economics (2)
- Sports science (2)
Differences

• Vocabulary type:
  – Thesaurus – Thesaurus
  – Classification – Thesaurus
  – (Classification – Classification)
  – (Thesaurus – Descriptor list)

• Change of discipline

• Change of language

• Size

• Combination / compounds
Thesaurus – Thesaurus

- Identical equivalence: 21%
- Null Mapping: 12%
- Related Term: 14%
- Narrower Term: 9%
- Broader Term: 20%
- Equivalence (Synonym): 24%
Classification – Thesaurus (JEL – STW)

- Narrower Term: 75%
- Related Term: 8%
- Equivalence: 6%
- Broader Term: 11%
GOAL: Facilitate search across different databases

→ Navigate without semantic borders!

• Translate search terms into other terminologies
• Increase diversity of documents
• Improve search experience without effort for searcher
Information Retrieval Tests

1. Do mappings improve subject search?
   CT (start vocabulary) \(\rightarrow\) TT \(\rightarrow\) Destination database
   Family relations \(\rightarrow\) Family AND social relations \(\rightarrow\)

2. Do mappings improve free-text search?
   FT (start vocabulary) \(\rightarrow\) FT + TT \(\rightarrow\) Destination database
   Family relations \(\rightarrow\) Family relations OR (Family AND social relations) \(\rightarrow\)
Information Retrieval Tests

- Thesaurus mapping only
- Only equivalence relations
- Real queries (~6 per tested cross-concordance)
- Databases: 80,000 – 16 mio. documents
- Test 1 (CT → TT): 13 Cross-concordances
- Test 2 (FT → FT+TT): 8 Cross-concordances
### Information Retrieval Tests - Results

**CT → TT (Improvements in %)**

<table>
<thead>
<tr>
<th></th>
<th>Recall = Hitrate</th>
<th>Precision = Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intradisciplinary</td>
<td>+39%</td>
<td>+34%</td>
</tr>
<tr>
<td>Interdisciplinary</td>
<td>+136%</td>
<td>+68%</td>
</tr>
</tbody>
</table>

**FT → FT+TT (Improvements in %)**

<table>
<thead>
<tr>
<th></th>
<th>Recall = Hitrate</th>
<th>Precision = Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intradisciplinary</td>
<td>+20%</td>
<td>-12%</td>
</tr>
<tr>
<td>Interdisciplinary</td>
<td>+24%</td>
<td>-24%</td>
</tr>
</tbody>
</table>
Sowiport Portal: [www.sowiport.de](http://www.sowiport.de)
Sowiport Thesaurus

library

Library

Used for

- alliance
- public library

Used in combination with

- utilization for library utilization

Notation

- 4.2.07 Information Systems, Communications
Sowiport Search

Termtransformation
Der Term: bibliothek wurde transformiert in:
Bibliotheken oder
Bibliothekswesen oder
Elektronische Bibliothek oder
Libraries oder
PUBLIC LIBRARIES oder
Wissenschaftliche Bibliothek oder

Termtransformation
Der Term: global wurde transformiert in:
Welt

1 World Library and Information Congress: 'Libraries without borders: Navigating towards global understanding' ☞
Québec, Kanada; 10.08.2008 - 14.08.2008 URL: http://www.ifla.org/IV/ifla74/index.htm
Informationstyp: Veranstaltung
Datenbank: SocioGuide

2 IFLA Social Science Libraries Section Pre-Conference ☞
Toronto, Kanada; 06.08.2008 - 07.08.2008 URL: http://ilabs.inquiry.uiuc.edu/ilab/ssls
Informationstyp: Veranstaltung
Datenbank: SocioGuide

3 Refugees and asylum seekers in the Caribbean region: library service implications (Flüchtlinge und Asylsuchende in der Karibik: Implikationen für bibliothekarische Dienstleistungen) ☞
Autor: Brathwaite, Tamara
Erscheinungsjahr: 2007; Dokumenttyp: Buch
Datenbank: Sozialwissenschaftliches Literaturinformationssystem (GESIS)
Conclusion

• Cross-concordances improve subject search with controlled terms & free-text search

• Only 24% relations utilized (equivalence)

• Potential:
  – Other relations
  – Natural language query terms $\rightarrow$ CT translation

• More mappings which are not evaluated

• Sowiport: http://www.sowiport.de
Mayr, Philipp; Petras, Vivien (2008): Cross-concordances: terminology mapping and its effectiveness for information retrieval. In: 74th IFLA World Library and Information Congress. Québec, Canada-


„Databases without semantic borders“

KoMoHe Project

http://www.gesis.org/en/research/information_technology/komohe.htm

E-mail: philipp.mayr@gesis.org
vivien.petras@gesis.org