Intra- and interdisciplinary cross-concordances for information retrieval

Philipp Mayr
GESIS – Leibniz Institute for the Social Sciences, Bonn, Germany

8th NKOS Workshop at the 13th ECDL Conference
Corfu, Greece, 01. October 2009
KoMoHe Project (2004-2007)

KoMoHe (Competence Center Modeling and Treatment of Semantic Heterogeneity)

Goals:

- Models for searching heterogeneous collections
- Development, organization & management of cross-walks between controlled vocabularies
- IR evaluation of the mappings (effectiveness of intellectual mapping)
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>

manually created, directed relations between controlled terms of two knowledge organization systems (KOS)
Cross-concordances

• 25 Vocabularies in 64 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)
- Social Sciences (10)
- Psychology (1)
- Pedagogics (1)
- Universial (3)
- Gerontology (1)
- Economics (2)
- Medicine (1)
- Political science (3)
- Sports science (2)
Net of Cross-concordances

Each node represents a KOS
Objectives

• Translate search terms into other terminologies
• Increase diversity of documents from different databases
• Improve search experience without effort for searcher
• Test the effect for IR in different disciplines (social science and others)
Main questions

• What is examined?
  – the quality of the mappings
  – or the quality of the associated search

• Can we enable distributed search with the subject access tools over several information systems?
  – In one discipline
  – Between at least two disciplines

• Is the impact of terminology mapping on recall and precision measurable?

• The mappings are helpful to whom?
Question: How effective are the mappings in an actual search? Does the application of term mappings improve search over a non-transformed subject (i.e. controlled vocabulary) search?
Information Retrieval Tests

- Thesauri mappings only
- Only equivalence relations
- Real queries (~6 per tested cross-concordance)
- Databases: 80,000 – 16 mio. documents
- Test 1 (CT $\rightarrow$ TT): 13 Cross-concordances
- Test 2 (FT $\rightarrow$ FT+TT): 8 Cross-concordances
<table>
<thead>
<tr>
<th>Vocabulary</th>
<th>Discipline</th>
<th>Database</th>
<th>Documents in DB</th>
</tr>
</thead>
<tbody>
<tr>
<td>TheSoz – Thesaurus Sozialwissenschaften (GESIS-IZ)</td>
<td>Social Sciences</td>
<td>SOLIS</td>
<td>345,086</td>
</tr>
<tr>
<td>DZI – Thesaurus des Deutschen Instituts für soziale Fragen</td>
<td>Social Sciences</td>
<td>SoLit</td>
<td>151,925</td>
</tr>
<tr>
<td>SWD – Schlagwortnormdatei</td>
<td>General (Social Sciences Excerpt)</td>
<td>USB Köln Sowi OPAC</td>
<td>72,729</td>
</tr>
<tr>
<td>CSA – Thesaurus of Sociological Indexing Terms (Cambridge Scientific Abstracts)</td>
<td>Social Sciences</td>
<td>CSA Sociological Abstracts</td>
<td>294,875</td>
</tr>
<tr>
<td>Psyindex - Psynex Terms</td>
<td>Psychology</td>
<td>Psynex (ZPID)</td>
<td>Ca. 200,000</td>
</tr>
<tr>
<td>STW – Standard Thesaurus Wirtschaft</td>
<td>Economics</td>
<td>Econis (ZBW Kiel)</td>
<td>Ca. 3,000,000</td>
</tr>
<tr>
<td>IBLK - Thesaurus Internationale Beziehungen und Länderkunde (Euro-Thesaurus)</td>
<td>Political Science</td>
<td>World Affairs Online WAO (SWP Berlin)</td>
<td>643,420</td>
</tr>
<tr>
<td>Mesh – Medical Subject Headings</td>
<td>Medicine</td>
<td>Medline (Dimdi)</td>
<td>Ca. 16,800,000</td>
</tr>
</tbody>
</table>

Table 4. Vocabularies and databases in the KoMoHe IR test
Steps

- Requesting recent research topics from our partners (social science and others)
- Intellectually translating the topics into controlled term searches in a KOS A
- Automatically translating the controlled terms via HTS into the controlled terms of a KOS B
- Retrieving documents from two runs
  1. Controlled term (CT) search (KOS A) in database B
  2. Translated term (TT) search (KOS B) in database B
Information Retrieval Test CT-TT

**Scenario CT**

- **Run 1**
  - DB A
  - Terms Voc A: Term a, Term b, Term c, ..., Term n
  - DB B

**Scenario TT**

- **Run 2**
  - DB A
  - Terms Voc A
  - DB B

HTS (Heterogeneity Service) ~ Web service providing the mappings
Information Retrieval Tests

Test 1
Intradisciplinary:
Social sc. – Social sc.
TheSoz – DZI
DZI – TheSoz
TheSoz – SWD
SWD – TheSoz
CSA – TheSoz

• 5 concordances
• 3 databases
• 35 topics

Test 2
Interdisciplinary:
Social sc. – Psychology
Social sc. – Economics

TheSoz – Psyndex
Psyndex – TheSoz
TheSoz – STW
STW – TheSoz

• 4 concordances
• 3 databases
• 19 topics

Test 3
Interdisciplinary:
Int. Relations – Economics
Medical sc. – Psychology

IBLK – STW
STW – IBLK
Mesh – Psyndex
Psyndex – Mesh

• 4 concordances
• 4 databases
• 28 topics
Methodology

• Downloading the documents for both runs (CT, TT), cutt-off: 1,000 docs
• Pooling both runs (CT, TT) for each topic
• Importing the documents into a assessment tool
• Relevance assessment of the documents by experts
• Analysis of the assessment data
  – Retrieved: average number of retrieved documents (across all search types)
  – Relevant: average number of relevant retrieved documents (across all search types)
  – Rel_ret: average number of relevant retrieved documents for a particular search type
  – Recall: proportion of relevant retrieved documents out of all relevant documents (averaged across all queries of one search type)
  – Precision: proportion of relevant retrieved documents out of all retrieved documents (averaged across all queries of one search type)
Assessment of the documents: by experts
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>

Detailed results can be found in Mayr & Petras, 2008
Discussion

• Overlap and more identical terms in intradisciplinary mappings
  – Mapping in one discipline is simpler: just one expert
  – Lesser effect on search
  – Automatic mapping may be more useful in intradisciplinary sets: mainly syntactic matching

• **Language plays a major role**
  – we had just one bilingual mapping in the test

• Restrictions of the study: no real users or interactions, only thesauri, KOS in German
Summary

Why are cross-concordances in one discipline less effective for IR?

• Amount of identical terms are significantly higher in one discipline (one language)
• No effective transformation possible for IR, if you have identical terms

Mapping projects should more often perform IR tests to measure the effect of their mappings.
Conclusion

• Cross-concordances improve subject search with controlled terms & free-text search: larger measurable effects on interdisciplinary mappings

• Only 24% relations utilized (equivalence)

• Potential:
  – Other relations
  – STR $\rightarrow$ CT translation

• More mappings which are not evaluated

• Mappings are used e.g. in portals like sowiport, vascoda, ireon, … and other projects
Next steps

• Visualization of the terminology network
• Combined evaluation with other value-added services (search term recommendation)
• Conversion to SKOS
• Evaluation of other disciplines
• Evaluation of indirect term transformation (term – switching term – end term)

## Indirect term transformations

<table>
<thead>
<tr>
<th>Thesoz - Gerlit - MESH</th>
<th>Switching Term</th>
<th>Endterm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allergie</td>
<td>Allergische Erkrankungen</td>
<td>Hypersensitivity</td>
</tr>
<tr>
<td>älterer Arbeitnehmer</td>
<td>Ältere Erwerbstätige</td>
<td>Middle Aged + Employment</td>
</tr>
<tr>
<td>Alterssoziologie</td>
<td>Soziologische Gerontologie</td>
<td>Sociology + Geriatrics</td>
</tr>
<tr>
<td>Anomie</td>
<td>Abweichendes Verhalten</td>
<td>Behavioral Symptoms</td>
</tr>
<tr>
<td>Arbeitnehmer</td>
<td>Erwerbstätige</td>
<td>Employment</td>
</tr>
<tr>
<td>Arbeitspsychologie</td>
<td>Arbeitswissenschaft</td>
<td>Human Engineering</td>
</tr>
<tr>
<td>Ausbildungsstand</td>
<td>Bildungsstand</td>
<td>Educational Status</td>
</tr>
<tr>
<td>Autonomie</td>
<td>Selbständigkeit</td>
<td>Personal Autonomy</td>
</tr>
<tr>
<td>Behindertenwerkstätte</td>
<td>Beschützende Werkstätte</td>
<td>Sheltered Workshops</td>
</tr>
<tr>
<td>Beratungsgremium</td>
<td>Beratung</td>
<td>Advisory Committees</td>
</tr>
<tr>
<td>Berufsmobilität</td>
<td>Berufliche Mobilität</td>
<td>Career Mobility</td>
</tr>
<tr>
<td>Berufstätigkeit</td>
<td>Erwerbstätigkeit</td>
<td>Employment</td>
</tr>
<tr>
<td>Beschäftigungstherapeut</td>
<td>Ergotherapeuten</td>
<td>Occupational Therapy/MA</td>
</tr>
<tr>
<td>Bildungsniveau</td>
<td>Bildungsstand</td>
<td>Educational Status</td>
</tr>
<tr>
<td>Bildungsprogramm</td>
<td>Bildungspianung</td>
<td>Education + Public Policy</td>
</tr>
<tr>
<td>biographische Methode</td>
<td>Biographische Analyse</td>
<td>Biography</td>
</tr>
<tr>
<td>Eigenarbeit</td>
<td>Schattenwirtschaft</td>
<td>Economics</td>
</tr>
<tr>
<td>Einsparung</td>
<td>Sparmaßnahmen</td>
<td>Cost Savings</td>
</tr>
<tr>
<td>Einwanderung</td>
<td>Migration</td>
<td>Emigration and Immigration</td>
</tr>
<tr>
<td>Emotionalität</td>
<td>Emotionen</td>
<td>Emotions</td>
</tr>
</tbody>
</table>
KoMoHe Project

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

E-mail: philipp.mayr@gesis.org