Results from a German terminology mapping effort: intra- and interdisciplinary cross-concordances between controlled vocabularies

Philipp Mayr, Vivien Petras, Anne-Kathrin Walter
GESIS Social Science Information Centre,
Bonn, Germany

Budapest, September 21, 2007
Outline

• Introduction & background
• Project KoMoHe
• Controlled vocabularies & cross-concordances
• Database and HTS
• Evaluation effort
• Summary & Outlook
• Demo (Online-Thesaurus)
Introduction

Theoretical background

• Vagueness between terms
  • Language ambiguity
  • Meaning of terms
• Semantic heterogeneity in document collections
• Problems while indexing documents
  • Consistency
  • Precision
  • Topicality
2 step methodology

- V1: between user terms and document terms
- V2: between document terms in different collections

Cross-concordances are used for V2 and V3
vascoda approach: an interdisciplinary portal (DL) for scientific information
- Transfers queries to specialized portals
- Covers information services from more than 40 partners

Consequences:
- Very complex structures (dozens of collections, schemata, interfaces, indexing languages, …)
- Necessity for semantic integration of relevant information services
Title: Kompetenzzentrum Modellbildung und Heterogenitätsbehandlung (Competence Center Modeling and Treatment of Semantic Heterogeneity)

Financing: Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung, BMBF)
Subproject of "Kompetenznetzwerk Neue Dienste, Standardisierung, Metadaten" (Competence Network: New Services, Standardization, Metadata)

Persons involved: Jürgen Krause, Philipp Mayr, Vivien Petras, Max Stempfhuber, Anne-Kathrin Walter

Project Duration: September 2004 through August 2007
Task: creation, organization and management of cross-concordances

Modeling and implementation of modules to treat semantic heterogeneity for vascoda collections

Largest terminology mapping effort in Germany

First major effort to evaluate the results of using cross-concordance for distributed retrieval
Controlled vocabularies

Various types of KOS: thesauri, classification systems, subject heading lists, descriptor lists

Cross-concordances for vascoda (respective sowiport)

- Mainly KOS centred around the social sciences
- Other disciplines are covered

25 KOS altogether
Types of KOS: Thesauri (16), Descriptor lists (4), Classifications (3), Subject headings (2)

Sizes of KOS: between 1,000 and 17,000 mapped terms; some KOS are mapped partly because of their size

Subjects of KOS: social science and related, political science, economics, medicine – subject specific parts of universal vocabularies
Controlled vocabularies - disciplines

Diagram showing relationships and counts for various disciplines, including:
- Agricultural science (1)
- Information science (1)
- Gerontology (1)
- Medicine (1)
- Social Sciences (10)
- Political science (3)
- Economics (2)
- Pedagogics (1)
- Sports science (2)
- Psychology (1)

Diagram layout indicates connections and overlaps among disciplines.
## Controlled vocabularies – overview 1

<table>
<thead>
<tr>
<th>#</th>
<th>Vocabular</th>
<th>Name</th>
<th>Subject</th>
<th>Type</th>
<th>Mapped</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>AGROVOC</td>
<td>AGROVOC Thesaurus</td>
<td>agricultural science</td>
<td>Thesaurus</td>
<td>part</td>
</tr>
<tr>
<td>2</td>
<td>CSA-ASSIA</td>
<td>CSA Thesaurus Applied Social Sciences Index and Abstracts</td>
<td>social sciences</td>
<td>Thesaurus</td>
<td>complete</td>
</tr>
<tr>
<td>3</td>
<td>CSA-WPSA</td>
<td>CSA Thesaurus of Political Science Indexing Terms</td>
<td>social sciences</td>
<td>Thesaurus</td>
<td>complete</td>
</tr>
<tr>
<td>4</td>
<td>CSA-PAIS</td>
<td>CSA Thesaurus PAIS International Subject Headings</td>
<td>political science</td>
<td>Thesaurus</td>
<td>complete</td>
</tr>
<tr>
<td>5</td>
<td>CSA-PEI</td>
<td>CSA Thesaurus Physical Education Index</td>
<td></td>
<td>Thesaurus</td>
<td>complete</td>
</tr>
<tr>
<td>6</td>
<td>FES</td>
<td>Descriptors of the Friedrich-Ebert Stiftung</td>
<td>social sciences</td>
<td>Descriptor list</td>
<td>complete</td>
</tr>
<tr>
<td>7</td>
<td>BiSp</td>
<td>Descriptors of the Bundesinstitut für Sportwissenschaft</td>
<td>sports science</td>
<td>Descriptor list</td>
<td>complete</td>
</tr>
<tr>
<td>8</td>
<td>INION</td>
<td>Descriptors of the Institute of Scientific Information on Social Sciences of the Russian Academy of Sciences</td>
<td>social sciences</td>
<td>Descriptor list</td>
<td>part</td>
</tr>
<tr>
<td>9</td>
<td>IAB</td>
<td>Descriptors of the Institut für Arbeitsmarkt- und Berufsforschung</td>
<td>social sciences</td>
<td>Descriptor list</td>
<td>complete</td>
</tr>
<tr>
<td>10</td>
<td>DDC</td>
<td>Dewey Decimal Classification</td>
<td>universal</td>
<td>Classification</td>
<td>part</td>
</tr>
<tr>
<td>11</td>
<td>ELSST</td>
<td>European Language Social Science Thesaurus</td>
<td>social sciences</td>
<td>Thesaurus</td>
<td>complete</td>
</tr>
<tr>
<td>12</td>
<td>INFODATA</td>
<td>INFODATA Thesaurus</td>
<td>information science</td>
<td>Thesaurus</td>
<td>complete</td>
</tr>
<tr>
<td>13</td>
<td>JEL</td>
<td>Journal of Economic Literature Classification System</td>
<td>economics</td>
<td>Classification</td>
<td>complete</td>
</tr>
<tr>
<td>14</td>
<td>MeSH</td>
<td>Medical Subject Headings</td>
<td>medicine</td>
<td>Subject Headings</td>
<td>part</td>
</tr>
<tr>
<td>15</td>
<td>Psyndex</td>
<td>Psyndex Terms</td>
<td>psychology</td>
<td>Thesaurus</td>
<td>complete</td>
</tr>
</tbody>
</table>
## Controlled vocabularies – overview 2

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>RVK</td>
<td>Regensburger Verbundklassifikation</td>
<td>universal</td>
<td>Classification part</td>
</tr>
<tr>
<td>17</td>
<td>SWD</td>
<td>Schlagwortnormdatei</td>
<td>universal</td>
<td>Subject Headings part</td>
</tr>
<tr>
<td>18</td>
<td>STW</td>
<td>Standard Thesaurus Wirtschaft</td>
<td>economics</td>
<td>Thesaurus complete</td>
</tr>
<tr>
<td>19</td>
<td>Bildung</td>
<td>Thesaurus Bildung</td>
<td>pedagogics</td>
<td>Thesaurus part</td>
</tr>
<tr>
<td>20</td>
<td>DZI</td>
<td>Thesaurus of the Deutschen Instituts für soziale Fragen</td>
<td>social sciences</td>
<td>Thesaurus complete</td>
</tr>
<tr>
<td>21</td>
<td>GEROLIT</td>
<td>Thesaurus of the Deutschen Zentrums für Altersfragen</td>
<td>social sciences</td>
<td>Thesaurus complete</td>
</tr>
<tr>
<td>22</td>
<td>TWSE</td>
<td>Thesaurus für wirtschaftliche und sozialen Entwicklung</td>
<td>political science</td>
<td>Thesaurus complete</td>
</tr>
<tr>
<td>23</td>
<td>IBLK</td>
<td>Thesaurus Internationale Beziehungen und Länderkunde (Euro-Thesaurus)</td>
<td>political science</td>
<td>Thesaurus complete</td>
</tr>
<tr>
<td>24</td>
<td>CSA-SA</td>
<td>Thesaurus of Sociological Indexing Terms</td>
<td>social sciences</td>
<td>Thesaurus complete</td>
</tr>
<tr>
<td>25</td>
<td>TheSoz</td>
<td>Thesaurus Sozialwissenschaften</td>
<td>social sciences</td>
<td>Thesaurus complete</td>
</tr>
</tbody>
</table>
Cross-concordances

Definition: Directed, relevance evaluated/estimated relations between controlled terms of two KOS
Most KOS were bilaterally mapped, but not always symmetrically or completely.
Cross-concordances - steps

- Estimation of the costs for an inter-thesaurus mapping
  - Analysis of the vocabularies
  - Sizes of the vocabularies
  - Topical overlap
- Selection of the cross-concordance contributors and partners
  - Mostly indexers & terminology workers
  - Institutions holding the rights of a vocabulary
- Project coordination and quality assurance
  - Review of parts of the relations (semantics)
  - Recall measures & syntax check
- Import into the cross-concordance database
- Integration in the terminology service (heterogeneity web service)
Cross-concordances

Mapping is done intellectually by: researchers, terminology experts, domain experts, postgraduates

Practical rules and guidelines:
1. Use intra thesaurus relations (e.g. ND->D)
2. Test the recall and precision of combinations
3. Relevances of the relations are normally depended on the relation type
4. Use 1:1 relations first
5. Map word groups consistently
Cross-concordances

Workflow

1. Understand the meaning of a start descriptor (use start thesaurus relations and database)
2. Search term in end thesaurus
   - Search word stem
   - Search equivalence, synonyms
   - Stop if you find an equivalence, otherwise build a combination or an other relation type
3. Map the term in the cross-concordance file
4. Add a relevance for the relation
Cross-concordances - examples

Equivalence (=) means identity, synonym, quasi-synonym

Hierarchy (< >)
  • Broader terms (<) from a narrower to a broad
  • Narrower terms (>) from a broad to a narrower

Association (^) for related terms
Null (0) no mapping possible

- Additional relevance for Relations
  (high, medium, low)

<table>
<thead>
<tr>
<th>term KOS 1</th>
<th>relation</th>
<th>term(s) KOS n</th>
</tr>
</thead>
<tbody>
<tr>
<td>hacker</td>
<td>=</td>
<td>Hacking</td>
</tr>
<tr>
<td>hacker</td>
<td>^+</td>
<td>Computers + Crime</td>
</tr>
<tr>
<td>hacker</td>
<td>^+</td>
<td>Internet + Security</td>
</tr>
<tr>
<td>ISDN</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>ISDN</td>
<td>&lt;</td>
<td>Telecommunications</td>
</tr>
<tr>
<td>documentation system</td>
<td>&gt;</td>
<td>Abstracting services</td>
</tr>
</tbody>
</table>
Cross-concordances - overview

<table>
<thead>
<tr>
<th>Voc. name</th>
<th>Voc. name</th>
<th>type</th>
<th>status</th>
<th>year</th>
</tr>
</thead>
<tbody>
<tr>
<td>TheSoz</td>
<td>STW</td>
<td>bilateral</td>
<td>imported</td>
<td>2004</td>
</tr>
<tr>
<td>TheSoz</td>
<td>BiSp</td>
<td>bilateral</td>
<td>imported</td>
<td>2004</td>
</tr>
<tr>
<td>Psyndex</td>
<td>BISp</td>
<td>bilateral</td>
<td>imported</td>
<td>2004</td>
</tr>
<tr>
<td>BiSp</td>
<td>Bildung</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TheSoz</td>
<td>DZI</td>
<td>bilateral</td>
<td>imported</td>
<td>2005</td>
</tr>
<tr>
<td>TheSoz</td>
<td>FES</td>
<td>bilateral</td>
<td>imported</td>
<td>2005</td>
</tr>
<tr>
<td>TheSoz</td>
<td>IBLK</td>
<td>bilateral</td>
<td>imported</td>
<td>2005</td>
</tr>
<tr>
<td>TheSoz</td>
<td>Gerolit</td>
<td>bilateral</td>
<td>imported</td>
<td>2005</td>
</tr>
<tr>
<td>MeSH</td>
<td>BISp</td>
<td>bilateral</td>
<td>imported</td>
<td>2005</td>
</tr>
<tr>
<td>STW</td>
<td>IBLK</td>
<td>bilateral</td>
<td>imported</td>
<td>2005</td>
</tr>
<tr>
<td>TheSoz</td>
<td>CSA-WPSA</td>
<td>bilateral</td>
<td>imported</td>
<td>2006</td>
</tr>
<tr>
<td>TheSoz</td>
<td>CSA-ASSIA</td>
<td>bilateral</td>
<td>imported</td>
<td>2006</td>
</tr>
<tr>
<td>TheSoz</td>
<td>ELSST</td>
<td>bilateral</td>
<td>imported</td>
<td>2006</td>
</tr>
<tr>
<td>TheSoz</td>
<td>CSA-PEI</td>
<td>bilateral</td>
<td>imported</td>
<td>2006</td>
</tr>
<tr>
<td>MeSH</td>
<td>Psyndex</td>
<td>bilateral</td>
<td>imported</td>
<td>2006</td>
</tr>
<tr>
<td>MeSH</td>
<td>Gerolit</td>
<td>bilateral</td>
<td>imported</td>
<td>2006</td>
</tr>
<tr>
<td>IBLK</td>
<td>CSA-PAIS</td>
<td>bilateral</td>
<td>imported</td>
<td>2006</td>
</tr>
<tr>
<td>IBLK</td>
<td>TWSE</td>
<td>bilateral</td>
<td>imported</td>
<td>2006</td>
</tr>
<tr>
<td>INION</td>
<td>TheSoz</td>
<td></td>
<td></td>
<td>2007</td>
</tr>
<tr>
<td>INFODATA</td>
<td>SWD</td>
<td>bilateral</td>
<td>imported</td>
<td>2007</td>
</tr>
<tr>
<td>INFODATA</td>
<td>TheSoz</td>
<td>bilateral</td>
<td>imported</td>
<td>2007</td>
</tr>
<tr>
<td>IAB</td>
<td>TheSoz</td>
<td>bilateral</td>
<td>imported</td>
<td>2007</td>
</tr>
<tr>
<td>IAB</td>
<td>STW</td>
<td>bilateral</td>
<td>imported</td>
<td>2007</td>
</tr>
<tr>
<td>SWD</td>
<td>MeSH</td>
<td>bilateral</td>
<td></td>
<td>2007</td>
</tr>
<tr>
<td>SWD</td>
<td>AGROVOC</td>
<td>bilateral</td>
<td></td>
<td>2007</td>
</tr>
<tr>
<td>JEL</td>
<td>STW</td>
<td></td>
<td>ready</td>
<td>2007</td>
</tr>
<tr>
<td>RVK</td>
<td>DDC</td>
<td></td>
<td>ready</td>
<td>2007</td>
</tr>
</tbody>
</table>

7 further mappings from the previous projects infoconnex and CARMEN
Vocabularies: 25
Mappings: 28 bilateral, 6 unilateral
Size: round 396,000 relations to date
Concepts: round 124,000 (incl. combinations)

Cross-concordance relations:
  • Equivalence: 165,000 (42%)
  • Broader: 84,000 (21%)
  • Narrower: 36,000 (9%)
  • Association: 56,000 (14%)
  • Null: 56,000 (14%)
Heterogeneity Service (HTS)

2 scenarios
- Just transform into equivalence relations
- Present additional relations to users

Legend:
SV = Start vocabulary
EV = End vocabulary
DB = Database
Heterogeneity Service

Start term

Start vocabulary

End vocabulary

Relations

End term(s)

---

educational institution

Thesaurus Social science (IZ)

ASSIA

WPSA

Education

Facilities

Universities

Schools
Evaluation

To date only very small evaluations in previous projects

Do cross-concordances improve search?

How?

Objective: to test and measure the effectiveness of cross-concordance in an real distributed environment

Questions:

• Exactness of the relations
• Relevance of the additional documents
• Intra- vs. Interdisciplinary cross-concordances

Measuring: quantitative analysis and retrieval test
Evaluation - Quantitative analysis

Objective: find trends in the cross-concordances
  • depended on the subject and structure of the vocabularies

Measures:
• Distribution of relations
• Ratio of mapped term in the end vocabulary
• Ratio of identities (term a is exact the same as term b)
• Relations for an end term or concept
Evaluation – preliminary results

In the same discipline generally more equivalence relations (TheSoz, DZI, SWD)

• Exact match in the same discipline is high
• Exact match in the same language is high (German)

In interdisciplinary cross-concordances generally more associations and Null relations (TheSoz, Psyndex, STW, IBLK, MeSH)

But differences in creating the cross-concordances (human factor) are visible
Evaluation – Retrieval test

Objective: value-added for the user (additional documents)

Task: Evaluating real user topics (operationalized in controlled terms)

1. Free text query (FT)
2. Descriptor query in the controlled term field (CT)
3. Translated descriptors via cross-concordance (only EQ-relations) (TT)

Relevance assessment of the retrieved documents
Evaluation – Retrieval test

Steps:
1. Real user topics by partners (in operationalized form)
2. Formulation of the queries and pretest of the test
3. Searching the databases (3 queries for a topic) and download of the documents (max. 1,000 doc)
4. Import of the documents in assessment tool and assessment of the documents
5. Analysis of the assessments
Evaluation – Retrieval test

Collections:
Test 1 - Social sciences: SOLIS, CSA Sociological Abstracts, SoLit, OPAC University Library Cologne
Test 2 - Social sciences interdisciplinary: SOLIS, Econis, Psyndex
Test 3 - Interdisciplinary: Medline, Psyndex, Econis, World Affairs online

Topics: between 5-10 for a mapping
Documents: max. 1,000 documents for a topic, documents are not ranked
Recall is the percentage of retrieved relevant documents out of all relevant documents. Precision is the percentage of relevant documents out of the retrieved doc.
Evaluation – preliminary results

<table>
<thead>
<tr>
<th></th>
<th>CT</th>
<th>TT</th>
<th>FT</th>
</tr>
</thead>
<tbody>
<tr>
<td>SWD-TheSoz</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 topics</td>
<td>Recall</td>
<td>0.5817</td>
<td>0.5817</td>
</tr>
<tr>
<td></td>
<td>Precision</td>
<td>0.3663</td>
<td>0.3663</td>
</tr>
<tr>
<td>TheSoz-DZI</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10 topics</td>
<td>Recall</td>
<td>0.5907</td>
<td>0.7602</td>
</tr>
<tr>
<td></td>
<td>Precision</td>
<td>0.3173</td>
<td>0.3914</td>
</tr>
<tr>
<td>STW-TheSoz</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 topics</td>
<td>Recall</td>
<td>0.2807</td>
<td>0.5859</td>
</tr>
<tr>
<td></td>
<td>Precision</td>
<td>0.2351</td>
<td>0.3634</td>
</tr>
</tbody>
</table>

- TT improves over CT, but not necessarily over FT
- FT generates more doc (FT search controlled terms too)
All related cross-concordances will be used in sowiport.

Results of the quantitative and retrieval evaluation will be finished next month.

Other relation types and their utilization in search.

Indirect term transformations (experiments).

Merging V1 treatment (V1 is the vagueness between user terms and descriptors) and cross-concordances.
Online-Thesaurus

Available at

1) Scientific scene

2) State church
Heterogeneity Service
Project „Competence Center Modeling and Treatment of Semantic Heterogeneity”:
http://www.gesis.org/en/research/information_technology/komohe.htm

Email:
philipp.mayr@gesis.org
vivien.petras@gesis.org