

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

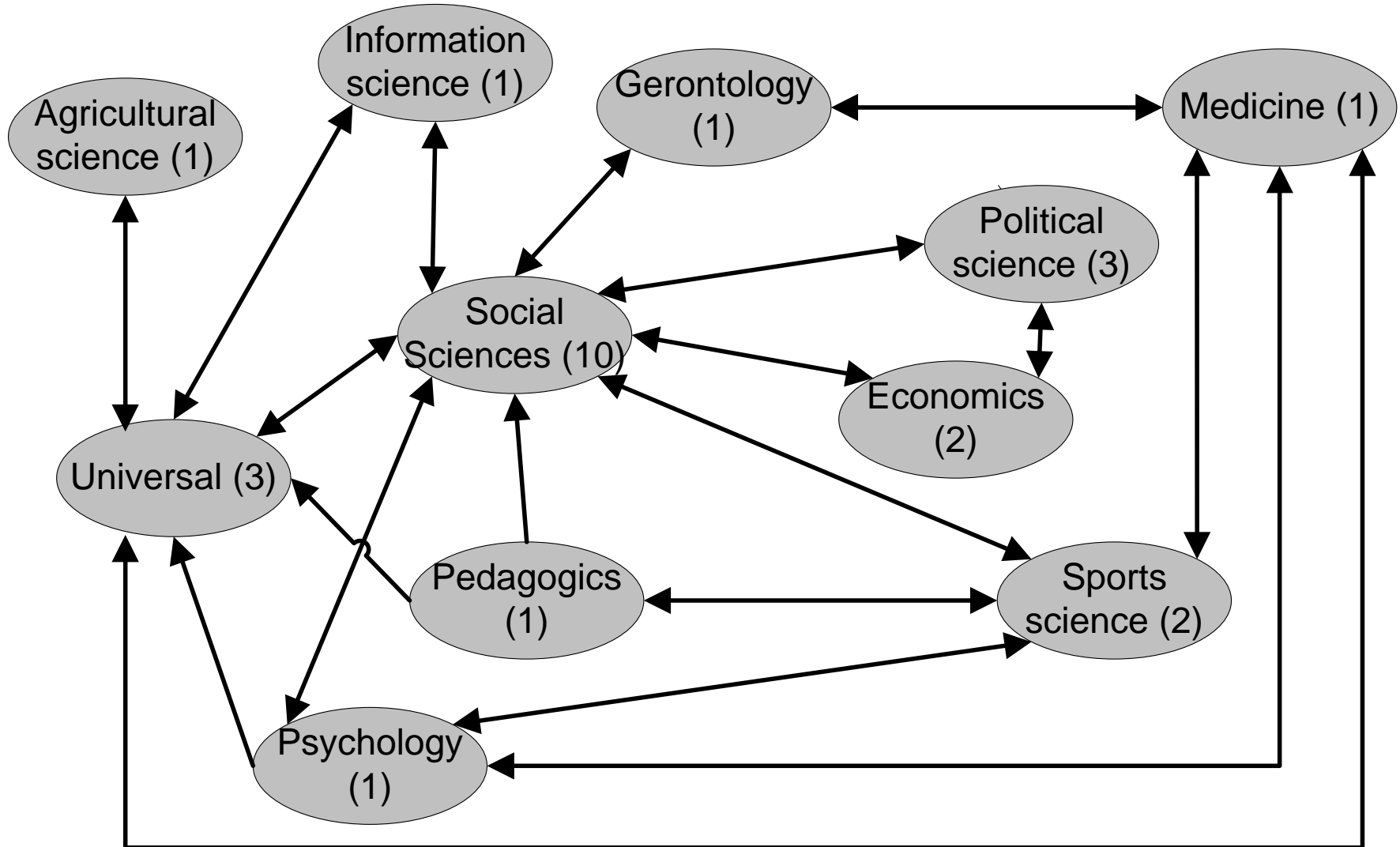
KOS 1	Relation	KOS 2
Library	=	Bibliothèque
Library	>	Special library
Thesaurus	<	KOS
Hacker	^	Computers + Security
Virus	0	

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

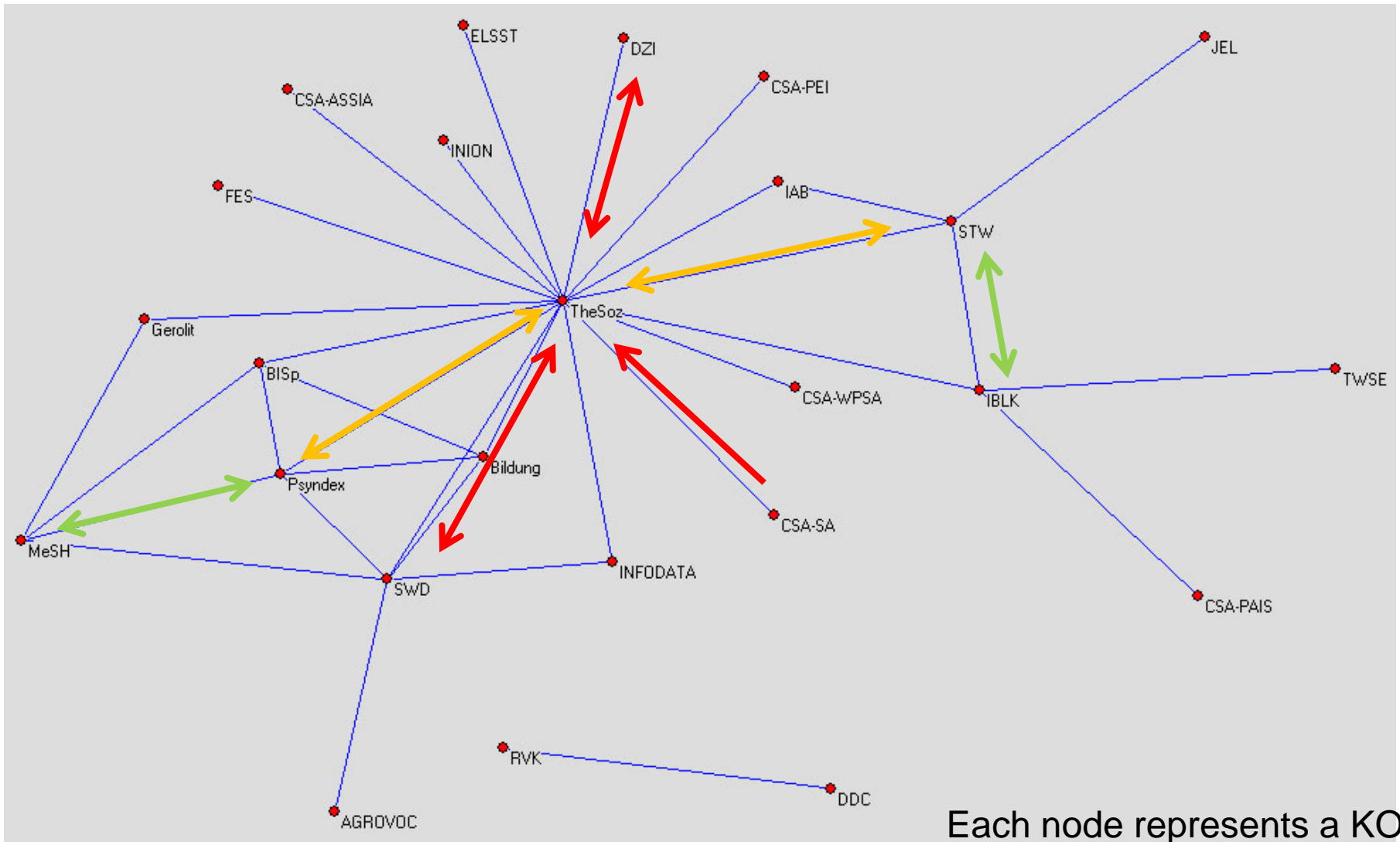
Cross-concordances

- 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



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?

Information Retrieval Test

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 → TT): 13 Cross-concordances
- Test 2 (FT → FT+TT): 8 Cross-concordances

Vocabulary	Discipline	Database	Documents in DB
TheSoz – Thesaurus Sozialwissenschaften (GESIS-IZ)	Social Sciences	SOLIS	345,086
DZI – Thesaurus des Deutschen Instituts für soziale Fragen	Social Sciences	SoLit	151,925
SWD – Schlagwortnormdatei	General (Social Sciences Excerpt)	USB Köln Sowi OPAC	72,729
CSA – Thesaurus of Sociological Indexing Terms (Cambridge Scientific Abstracts)	Social Sciences	CSA Sociological Abstracts	294,875
Psyndex - Psyndex Terms	Psychology	Psyndex (ZPID)	Ca. 200,000
STW – Standard Thesaurus Wirtschaft	Economics	Econis (ZBW Kiel)	Ca. 3,000,000
IBLK - Thesaurus Internationale Beziehungen und Länderkunde (Euro-Thesaurus)	Political Science	World Affairs Online WAO (SWP Berlin)	643,420
Mesh – Medical Subject Headings	Medicine	Medline (Dimdi)	Ca. 16,800,000

Table 4. Vocabularies and databases in the KoMoHe IR test

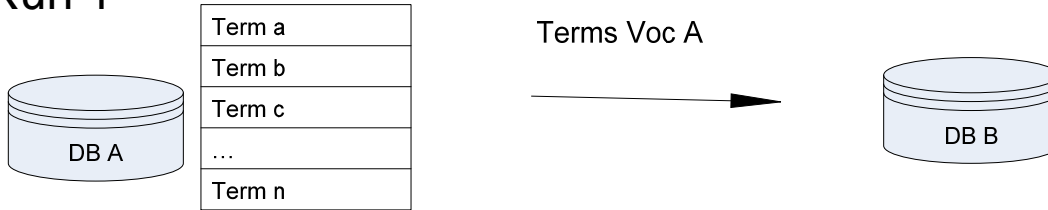
Steps

- 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

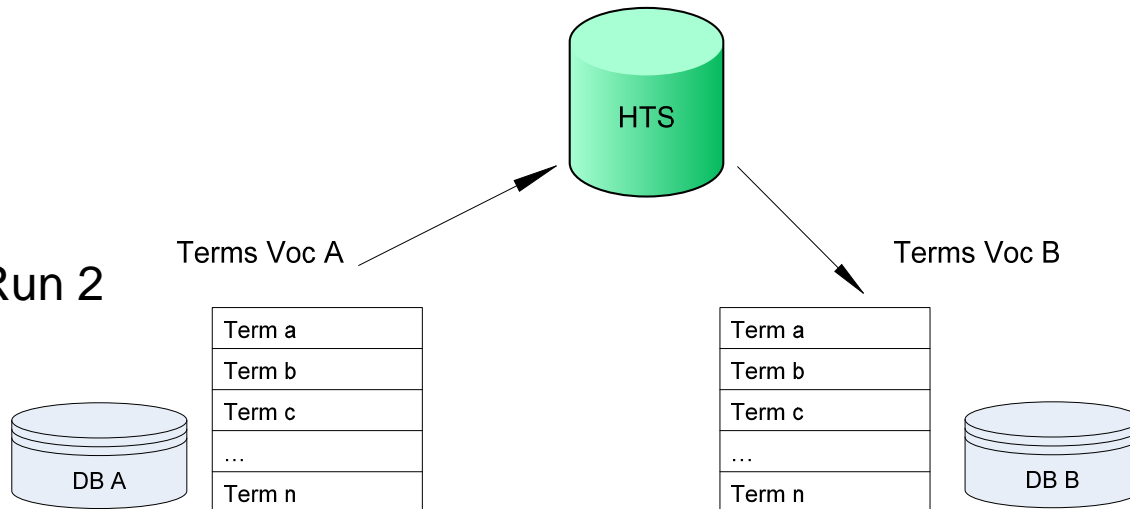
Scenario CT

Run 1



Scenario TT

Run 2



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

415: Arzneimittel für Kinder

Prev Next Relevant Non-relevant All Autostep

Topic Time Write Log Fontsize +

Relevant:10 NonRelevant:15 Not Assessed:0 Short

Entschädigung für Contergan	AUTHOR Derleder, Peter Winter, Gerd
Lernen und Gedächtnis bei Kindern : Eine Untersuchung über	PUBLICATION-YEAR 1976
Der Einfluß der Eltern auf den Drogenkonsum ihrer Kinder	DOCTYPE journalarticle
Ergebnisse medizin-soziologischer Untersuchungen zum Gesund	MEDIATYPE printed
Keine Angst vor der Angst : Risiko: Element unseres Lebens	ISBN 0341-3039
Pillen für den Klassenfrieden : zur medikamentösen Behandlu	HOST Das Argument Zeitschrift für Philosophie und Sozialwissenschaften, Sonderband 12
Als Allgemeinärztin in Nicaragua	ABSTRACT-DE Der vorliegende Aufsatz, beschäftigt sich unter 3 Gesichtspunkten mit den politischen, rechtlichen und medizinischen Problemen, die sich aus der Contergan-Affäre ergeben: der erste Teil über die Entschädigungsforderungen und ihre Behandlung dient einer theoretischen Verarbeitung, die einerseits die Ungewöhnlichkeit der juristischen Regelbildung aufgreift und andererseits das Lernpotential untersucht, welches die "Entschädigung" für Contergan politisch und rechtlich geborgen hat. Der zweite Teil beschäftigt sich mit den Funktionen, die dem zivilrechtlichen Schadenausgleich insgesamt verbleiben, und der dritte Teil mit der Frage, welche arzneimittelrechtlichen Lehren gezogen worden sind. Eine erweiterte Fassung dieses Aufsatzes, die sich an die juristische Fachöffentlichkeit richtet, ist in der Zeitschrift "Demokratie und Recht", Heft 3/1976, erschienen. (CK)
Arzneimittellisten in der Bundesrepublik Deutschland, der S	CLASSIFICATION-TEXT-DE Kriminalsoziologie, Rechtssoziologie, Kriminologie
Kindheit im Mittelalter : aus der Sicht eines heutigen Kind	Medizinsoziologie
Über Mittel und Maßnahmen der Mütter im Raum Königs Wusterh	CONTROLLED-TERM-DE Arzneimittel
Medikamentenkonsum und Alter : ein Blick in den 'Durchschni	Behinderung
Die Lebenssituation der Bürgerin höheren und hohen Lebensal	Bundesverfassungsgericht
Die Kosten jugendlicher Problembewältigung : Alkohol, Zigar	Entschädigung
Drogengebrauch und Prävention : von Problemen, die Kinder u	Gericht
Gesundheitszustand und Gesundheitsverhalten von Kindern : E	Gesetzgebung
Konsum von Schlaf- und Beruhigungsmitteln in der Schweiz :	Innovation
Von der Anstalts- zur Gemeindepsychiatrie : empirische Befu	Kind
www.wehwehchen.de	pharmazeutische Industrie
Gesundheitsversorgung in Entwicklungsländern : medizinische	Rechtslage
Bemerkungen des Bundesrechnungshofes 2003 zur Haushalts- un	Strafprozess
Der informierte Patient - Herausforderung für die Pharmakom	Urteil
Off-Label-Use bei der HIV-Behandlung	COUNTRY-CODE o.O. Bundesrepublik Deutschland
Vorstoß durch die Hintertür : die versuchte Einführung frem	DOCID iz-solis-90003590
Zu viel für manche kleine Seele : Trend zu immer mehr psych	LANGUAGE-CODE de
GEK-Arzneimittel-Report 2004 : Auswertungsergebnisse der GE	TITLE-DE Entschädigung für Contergan

Information Retrieval Tests - Results

- CT → TT (Improvements in %)

	Recall = Hitrate	Precision = Accuracy
Intradisciplinary	+39%	+34%
Interdisciplinary	+136%	+68%

- FT → FT+TT (Improvements in %)

	Recall = Hitrate	Precision = Accuracy
Intradisciplinary	+20%	-12%
Interdisciplinary	+24%	-24%

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 → 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)

Publications

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-
http://www.ifla.org/IV/ifla74/papers/129-Mayr_Petras-en.pdf

Mayr, Philipp; Mutschke, Peter; Petras, Vivien (2008): Reducing semantic complexity in distributed Digital Libraries: treatment of term vagueness and document re-ranking. In: Library Review. 57 (2008) 3. pp. 213-224.
<http://arxiv.org/abs/0712.2449>

Indirect term transformations

10	Thesoz - Gerolit -MESH			
11		Startterm	Switching Term	Endterm
12		Allergie	Allergische Erkrankungen	Hypersensitivity
13		älterer Arbeitnehmer	Ältere Erwerbstätige	Middle Aged + Employment
14		Alterssoziologie	Soziologische Gerontologie	Sociology + Geriatrics
15		Anomie	Abweichendes Verhalten	Behavioral Symptoms
16		Arbeitnehmer	Erwerbstätige	Employment
17		Arbeitspsychologie	Arbeitswissenschaft	Human Engineering
18		Ausbildungsstand	Bildungsstand	Educational Status
19		Autonomie	Selbständigkeit	Personal Autonomy
20		Behindertenwerkstätte	Beschützende Werkstätte	Sheltered Workshops
21		Beratungsgremium	Beirat	Advisory Committees
22		Berufsmobilität	Berufliche Mobilität	Career Mobility
23		Berufstätigkeit	Erwerbstätigkeit	Employment
24		Beschäftigungstherapeut	Ergotherapeuten	Occupational Therapy/MA
25		Bildungsniveau	Bildungsstand	Educational Status
26		Bildungsprogramm	Bildungsplanung	Education + Public Policy
27		biographische Methode	Biographische Analyse	Biography
28		Eigenarbeit	Schattenwirtschaft	Economics
29		Einsparung	Sparmaßnahmen	Cost Savings
30		Einwanderung	Migration	Emigration and Immigration
31		Emotionalität	Emotionen	Emotions

Social sciences – gerontology – medicine

Sowiport Search

Einfache Suche **Erweiterte Suche** Thesauri

Termtransformation

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

Suche ▶ Trefferliste

Ihre Suche: Überall: [**bibliothek** und **global**]

1 2 ...

Treffer: 27

Termtransformation

Der Term: global wurde transformiert in:
 Welt

Sortieren nach Relevanz (nach Titel)

markieren Markierung entfernen

- 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)**

KoMoHe Project

[http://www.gesis.org/en/research/
information_technology/komohe.htm](http://www.gesis.org/en/research/information_technology/komohe.htm)

E-mail: philipp.mayr@gesis.org