

Building a terminology network for search: the KoMoHe project

Philipp Mayr, Vivien Petras

GESIS Social Science Information Centre

Int'l Conf. on Dublin Core and Metadata Applications 2008

Berlin, 25. September 2008

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

Problem: Heterogeneous collections

- **Many databases:**
 - document types / formats
 - vocabularies
- **Controlled vocabularies:**
 - internal consistency (high)
 - intersystem compatibility (low) -> (semantic heterogeneity)
- **Goal:**
Seamless search across multiple heterogeneous collections/repositories based on semantically rich relations
- **Solution:**
translate → cross-walks → terminology mapping

KoMoHe Project (2004-2007)

KoMoHe (Competence Center Modeling and Treatment of Semantic Heterogeneity)

Goals:

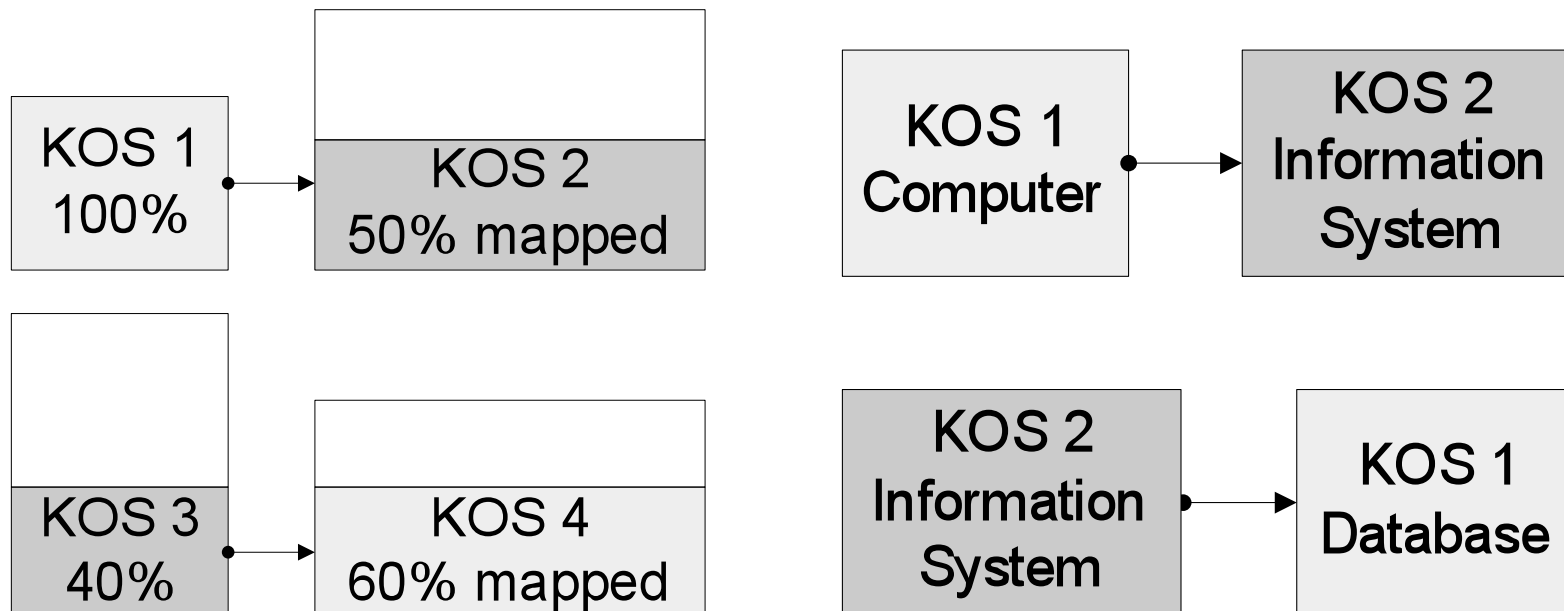
- Models for searching heterogeneous collections
- Development, organization & management of cross-walks between controlled vocabularies
- IR evaluation of the mappings (effectiveness of intellectual mapping)

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

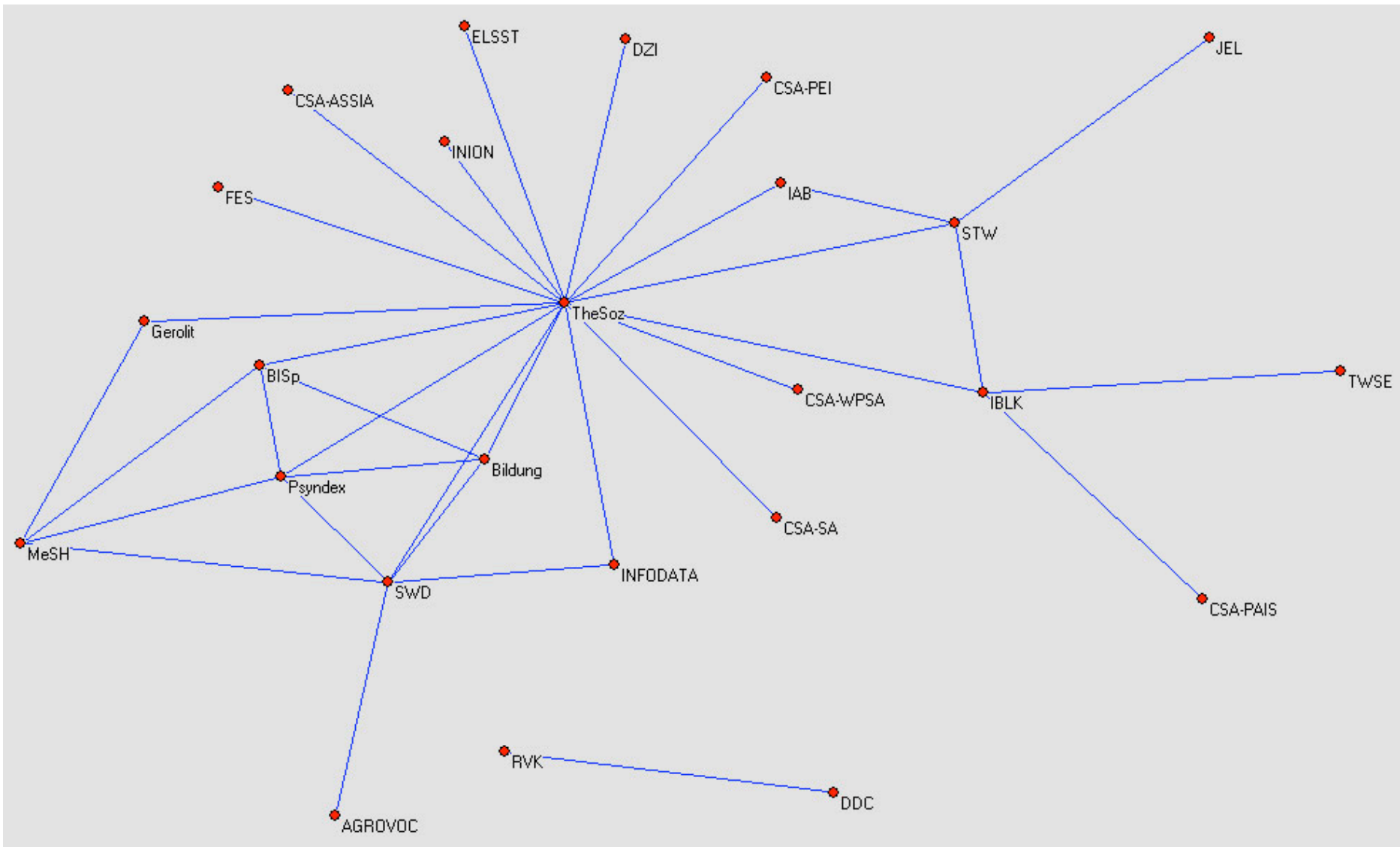
KOS 1	Relation	KOS 2
Library	=	Bibliothèque
Library	>	Special library
Thesaurus	<	KOS
Hacker	^	Computers + Security
Virus	0	

Cross-concordances

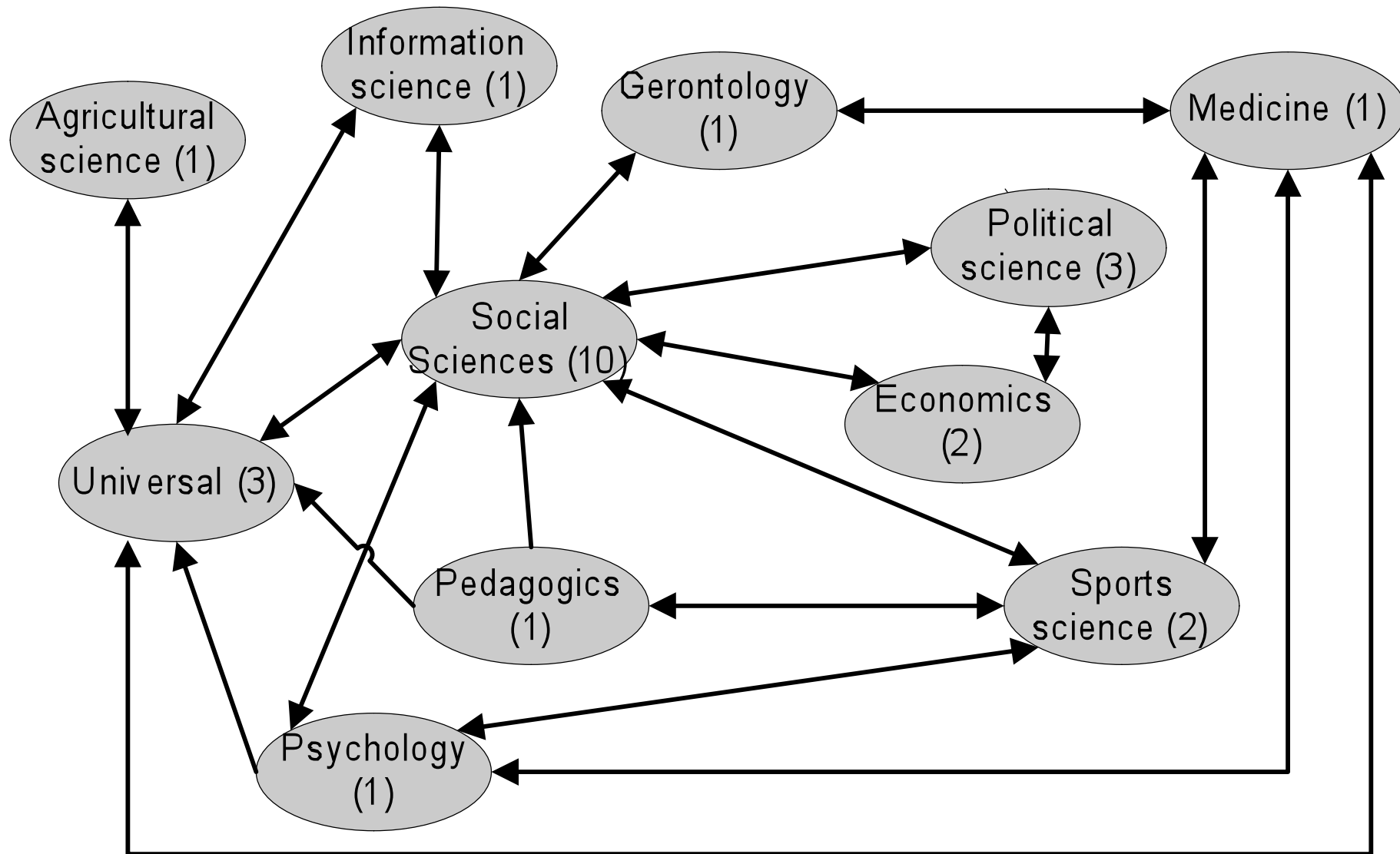
- 25 Vocabularies in 64 cross-concordances
 - Thesauri (16)
 - Descriptor lists (4)
 - Classifications (3)
 - Subject heading lists (2)

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

Net of Cross-concordances



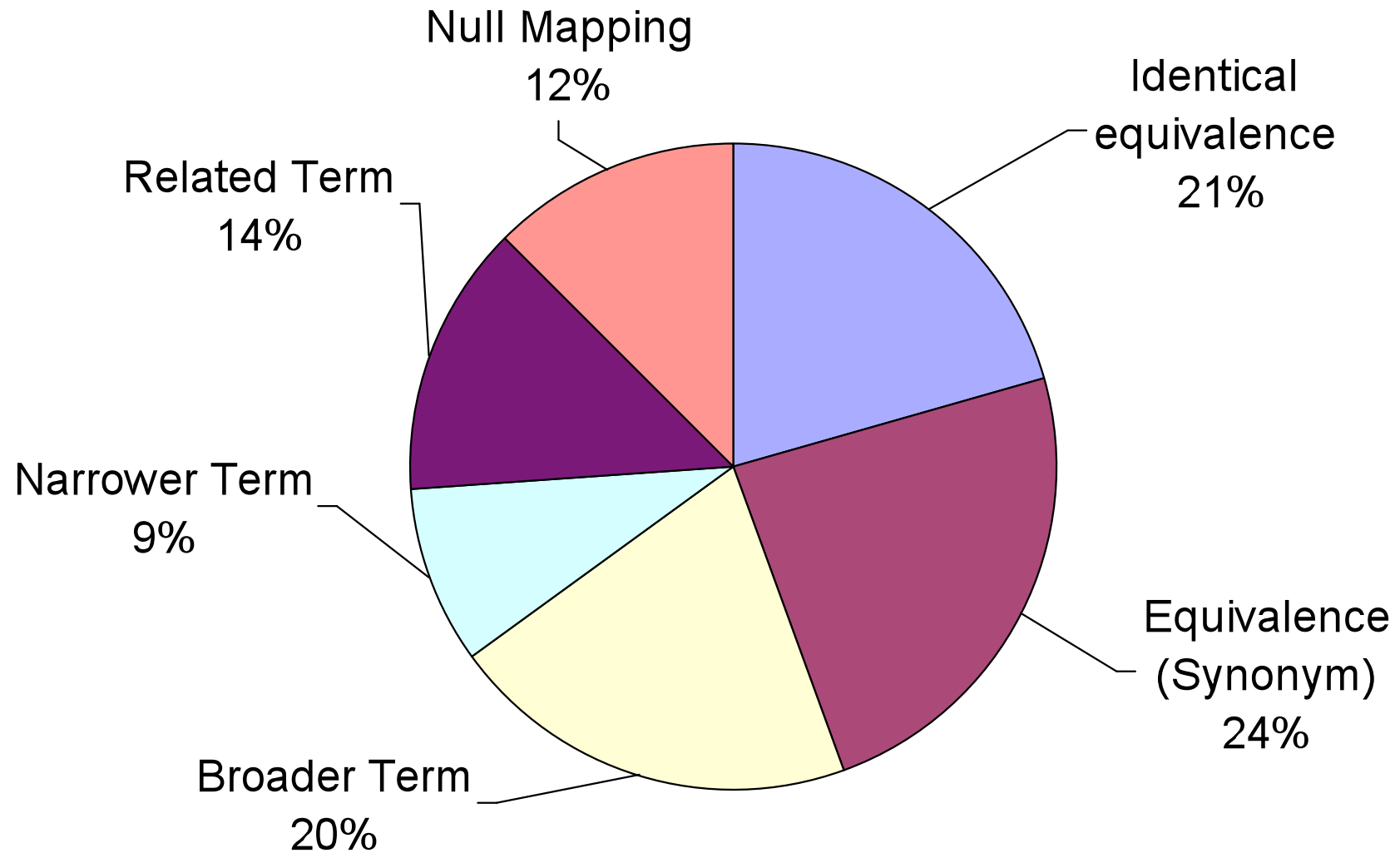
Disciplines



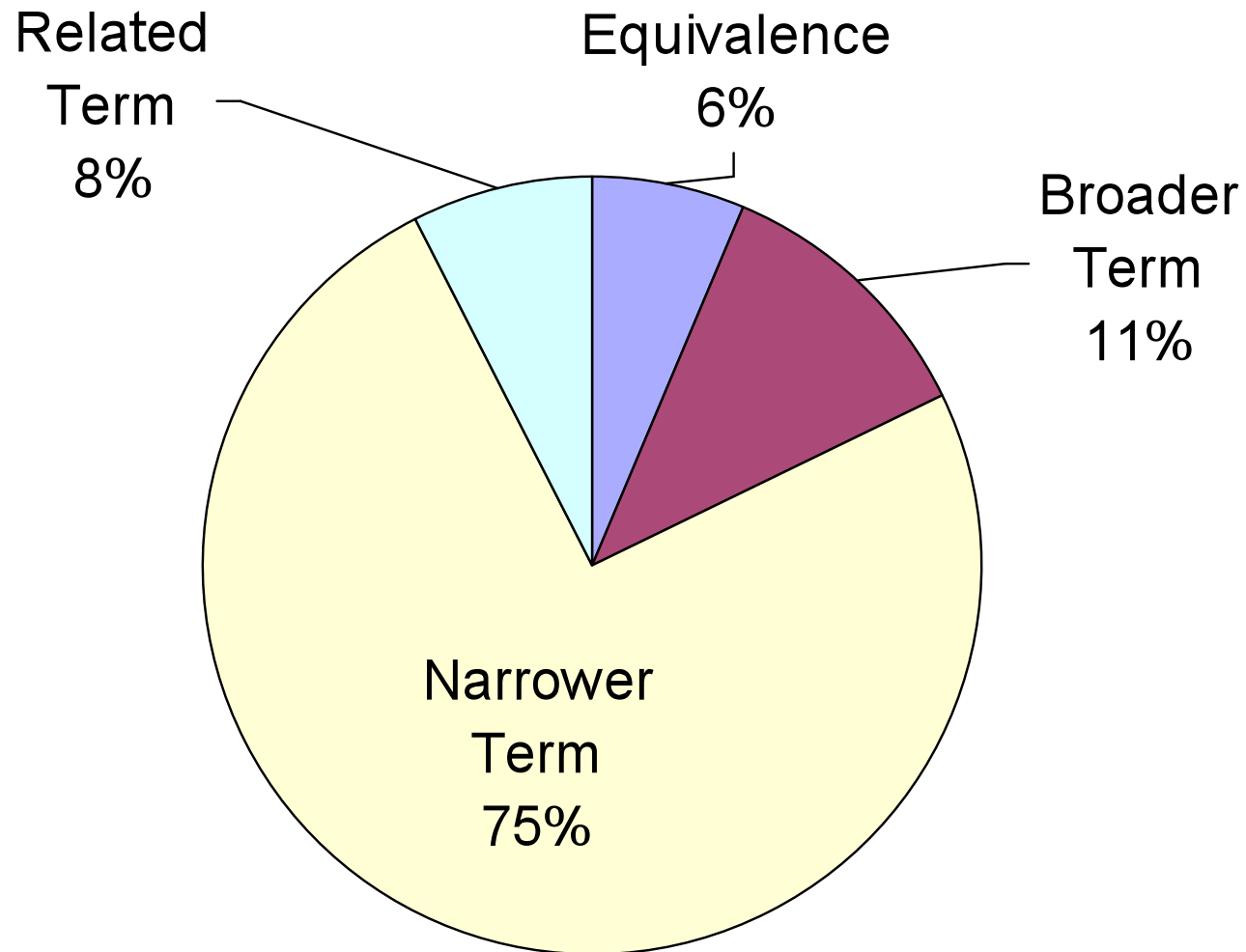
Differences

- Vocabulary type:
 - Thesaurus – Thesaurus
 - Classification – Thesaurus
 - (Classification – Classification)
 - (Thesaurus – Descriptor list)
- Change of discipline
- Change of language
- Size
- Combination / compounds

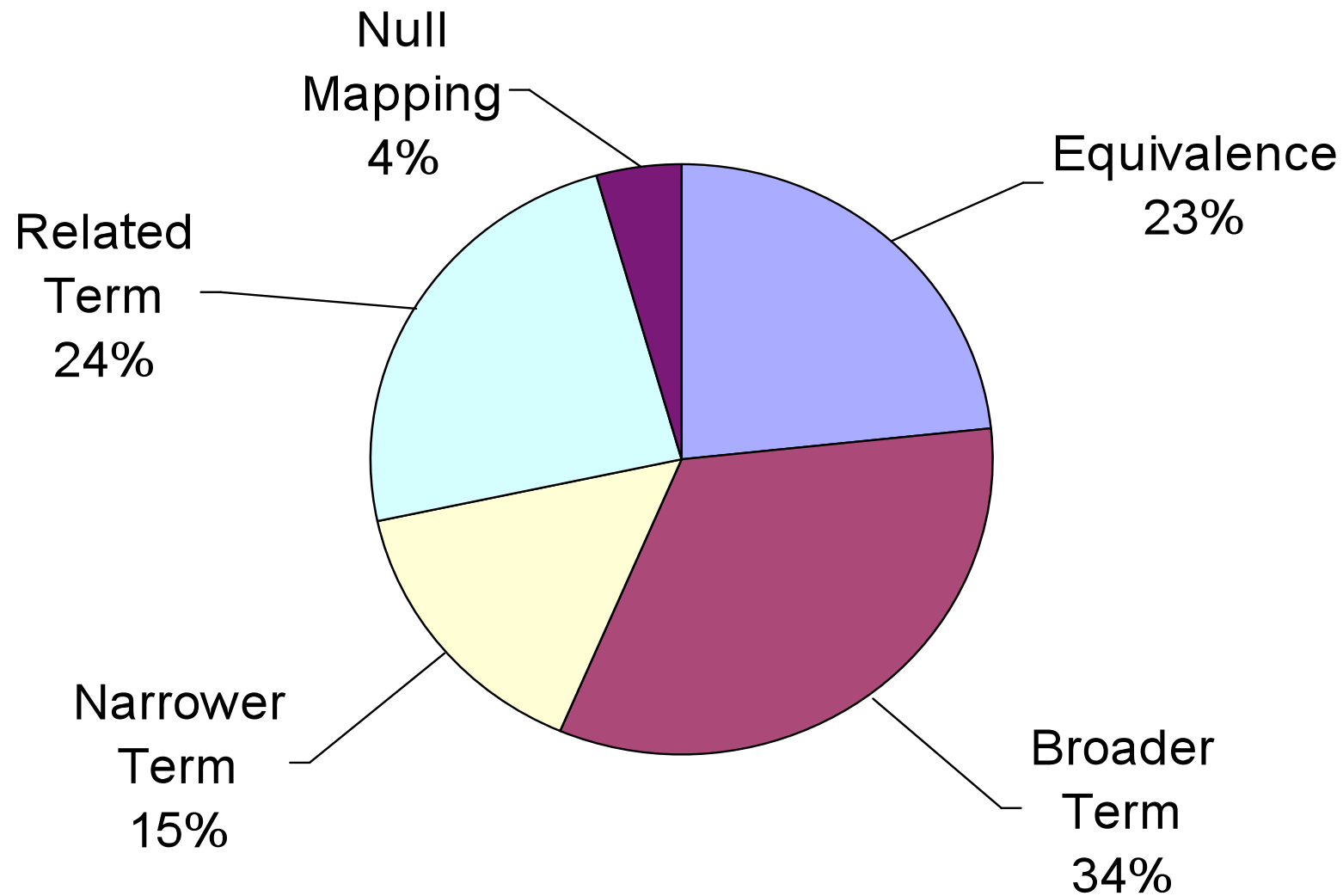
Thesauri – Thesauri



Classification – Thesaurus (JEL – STW)



Classification – Classification (RVK – DDC)



Information Retrieval Tests

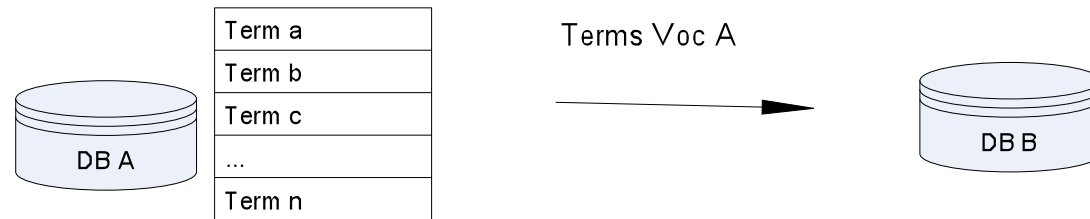
GOAL: Facilitate search across different databases

→ Navigate without semantic borders!

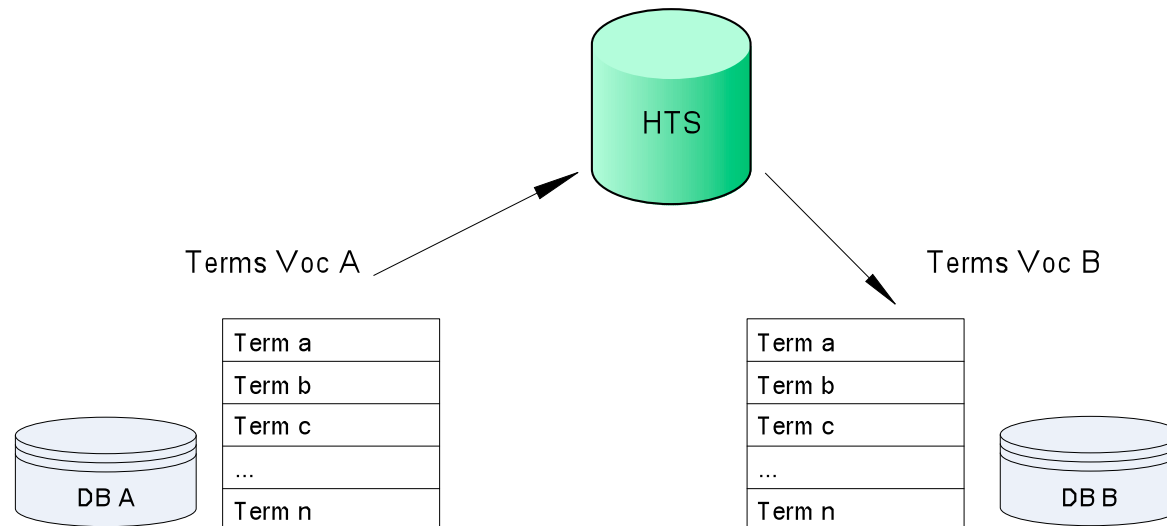
- Translate search terms into other terminologies
- Increase diversity of documents
- Improve search experience without effort for searcher

Information Retrieval Test CT-TT

Scenario CT



Scenario TT



HTS
(Heterogeneity Service) ~
Web service
providing the
mappings

Information Retrieval Tests

1. Do mappings improve subject search?

CT (start vocabulary) → TT → Destination database

Terms: Family relations → Family AND social relations →

2. Do mappings improve free-text search?

FT (start vocabulary) → FT + TT → Destination database

Terms: Family relations → Family relations OR (Family AND social relations) →

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 %)

	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%

Sowiport Thesaurus

sowiport.de Sozialwissenschaften auf den Punkt gebracht

Schnellstart Home Suche Produkte Themen Publikationen Service Kommunikation

Einfache Suche Erweiterte Suche **Thesaurus**

Deutsch English русский i

library Show: Translation Cross concordance

	Descriptor	Filter Translation	Filter Cross concordance
<input checked="" type="radio"/> alphabetic <input type="radio"/> systematic Hits: 2 of 10629 terms library library utilization	library Used for = <u>alliance</u> = <u>public library</u> Used in combination with = <u>utilization</u> for <u>library utilization</u> Notation = <u>4.2.07</u> Information Systems, Communications	Bibliothek библиотека	Thesaurus for the Social Sciences → ASSIA <input type="checkbox"/> Libraries → PEI <input type="checkbox"/> (null) → Soc.Abs. <input type="checkbox"/> Libraries → WPSA <input type="checkbox"/> Libraries → GeroLit <input type="checkbox"/> Bibliothekswissenschaft → DZI SoLit <input type="checkbox"/> Bibliothek → FES <input type="checkbox"/> Bibliothek <input type="checkbox"/> Arbeiterbibliothek → USB Köln <input type="checkbox"/> Bibliothek

Sowiport Search

Einfache Suche | **Erweiterte Suche** | Thesauri

Suche ▶ Trefferliste

Ihre Suche: Überall: [bibliothek und global]

Treffer: 27

Sortieren nach Relevanz (nach Titel)

markieren | Markierung entfernen

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 → CT translation
- More mappings which are not evaluated
- Sowiport: <http://www.sowiport.de>

Next steps

- Visualization of the network
- Combination with other value-added services (search term recommendation ~ mapping user terms and controlled terms)
- Conversion to SKOS
- Evaluation of indirect term transformation (term – switching term – end term)

Indirect term transformations

	A	B	C	D	E
94	CSA-SA	TheSoz	STW		
95			Startterm	Switching Term	Endterm
96			Abortion	Schwangerschaftsabbruch	Schwangerschaftsabbruch
97			Labor Force Participation	Erwerbsbeteiligung	Erwerbstätigkeit
98			Labor Theory of Value	Arbeitswerttheorie	Arbeitswertlehre
99			Management Styles	Führungsstil	Führungstheorie
100			Manufacturing Industries	produzierendes Gewerbe	Industrie
101			Mass Media	Massenmedien	Kommunikationsmedien
102			Measures (Instruments)	Messinstrument	Messgerät
103			Metropolitan Areas	Ballungsgebiet	Ballungsraum
104			Middle Class	Mittelschicht	Mittelstand
105			Needs	Bedürfnis	Bedürfnisse
106			Nuclear Weapons	Kernwaffe	Atomwaffe
107			Parks	Grünfläche	Städtische Grünfläche
108			Part Time Farming	Nebenerwerbsbetrieb	Nebenerwerbslandwirtschaft
109			Professional Associations	Berufsverband	Berufsverband
110			Religious Orders	Orden	Religiöser Orden
111			Social Consciousness	gesellschaftliches Bewusstsein	gesellschaftliches Bewusstsein

Social sciences – social sciences – economics

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

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>

KoMoHe Project

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

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