Searching in a Multi-Thesauri-Scenario –
experiences with SKOS and terminology mappings

Philipp Mayr
GESIS Leibniz Institute for the Social Sciences &
University of Applied Sciences in Darmstadt

Conference on Interoperability and Semantics in Knowledge Organization
Köln, 7/20/2010

Topics of the talk

Projects using controlled terms and terminology mappings for improving search

- KoMoHe (mapping terminology)

- IRM project (recommending descriptors, providing contexts)

- Searching in a multi-thesauri-scenario (connecting services)

Topic of the conference

Importance of controlled terms/concepts on the Web?

- Today

- Tomorrow?

Search term recommenders

Situation

Heterogeneity of controlled vocabularies

- **Internal consistence** → high

- **Compatibility** between controlled vocabularies → low

- Searching with a set of vocabularies is fairly unexplored

KoMoHe Project (2004-2007)

KoMoHe (Competence Center Modeling and Treatment of Semantic Heterogeneity)

- <u>Translate</u> search terms into other terminologies
- <u>Increase</u> diversity of documents from different databases
- <u>Improve</u> search experience without effort for searcher
- <u>Test</u> the effect for IR in different disciplines (social science and others)

Net of Cross-concordances

Each node represents a KOS
Evaluated mappings in colored arrows
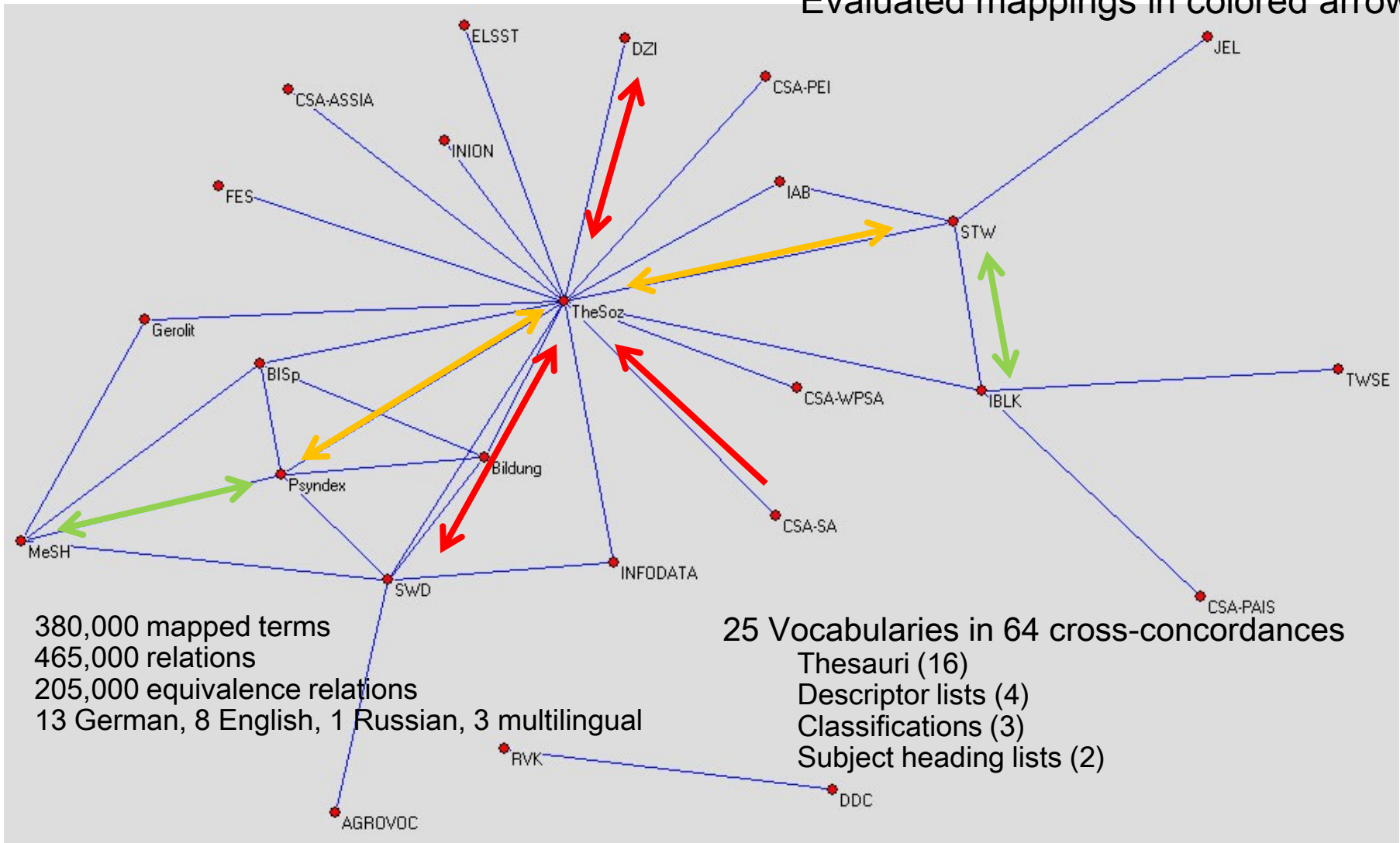
380,000 mapped terms
465,000 relations
205,000 equivalence relations
13 German, 8 English, 1 Russian, 3 multilingual

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

Main questions

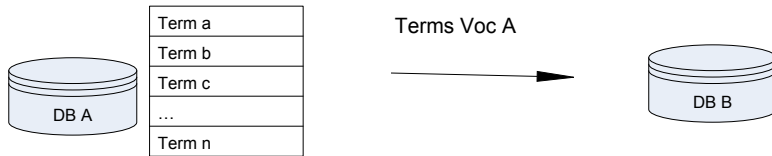- ## What is examined?
  - the quality of the mappings
  - or the <u>quality of the associated search</u>
- ## Can we enable distributed search with the subject access tools over several information systems?
  - In <u>one discipline</u>
  - Between at least <u>two disciplines</u>
- ## Is the impact of 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?

Scenario CT

Run 1

Terms Voc A

DB A → DB B

| Term a |
| Term b |
| Term c |
| … |
| Term n |

Scenario TT

HTS

Run 2

Terms Voc A   Terms Voc B

| Term a |
| Term b |
| Term c |
| … |
| Term n |

| Term a |
| Term b |
| Term c |
| … |
| Term n |

DB A   DB B
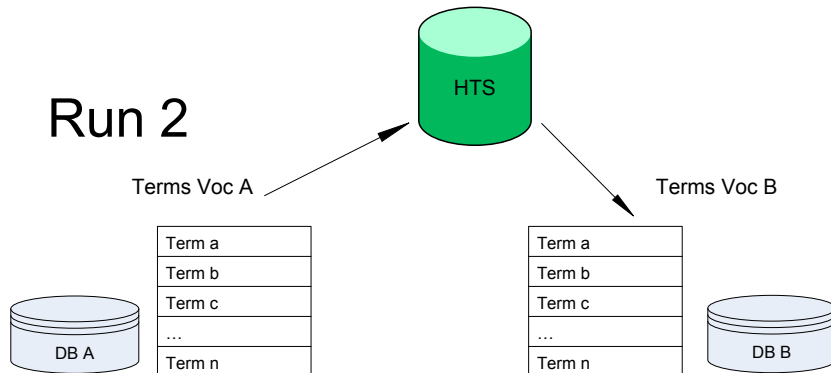
CT → TT (Improvements in %)

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

Mayr/Petras, 2008

HTS (Heterogeneity Service) ~ Web service providing the mappings

8

Experiences

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

- <u>Language plays a major role</u>
  - we had just one bilingual mapping in the test

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

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

Einfache Suche   Erweiterte Suche   Thesa...

Suche ▸ Trefferliste

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

◀ 1 2 ... ▶   **Treffer: 27**   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)**

Mappings for query expansion
Current implementation
No interaction

http://www.gesis.org/sowiport/

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

- Amount of <u>identical terms</u> 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.

IRM project

Motivation:

- Imprecise search results:

  - Vagueness of search terms

  - High ratio of irrelevant hits due to large results sets

- Improvement of search results/experience:

  - Search term recommendation (bridge between user and controlled voc.)

  - Non-textual ranking techniques (providing context)

http://www.gesis.org/beta/prototypen/irm

Search Term Recommender (STR)
Recommendation of highly associated controlled terms

Query: jugendsoziologie   Search

Only show metadata sets which include an abstract (slower) ☐

Suggest search terms

Controlled vocabulary
Thesaurus Sozialwissenschaften ▼
Automatic query expansion ☐

Rerank the result list

Rerank method
Bradfordizing ▼

Interactive query enhancement

term cloud   search term suggestions   central journals

Total hits: 2699

1. Neidhardt, Friedhelm (1973): *Beitraege zur Kritik der Jugendsoziologie und Jugendforschung* in **Deutsche Jugend Zeitschrift für Jugendfragen und Jugendarbeit Bd. 21 H. 4** (0012-0332) toggle details
2. Eckert, Roland (1973): *Grenzuebergaenge jugendsoziologischer Forschung* in **Deutsche Jugend Zeitschrift für Jugendfragen und Jugendarbeit H. 4** (0012-0332) toggle details
3. Baus, Magdalene (1965): *Deutsche Maedchen in England* in **Deutsche Jugend Zeitschrift für Jugendfragen und Jugendarbeit Bd. 2** (0012-0332) toggle details
4. Kreutz, Henrik (1971): *Theorie der Jugend und soziale Schichtung* in **Deutsche Jugend Zeitschrift für Jugendfragen und Jugendarbeit** (0012-0332) toggle details

Eltern-Kind-Beziehung   Handlungstheorie   Heim
Informationsgehalt   Jugend   Jugendbewegung
Jugendforschung   <small>Jugendlicher</small>
Jugendreligion   Jugendsoziologie   Klassenantagonismus
Körperlichkeit   Lebensbedingungen   Peer Group
Schelsky, H.   Stadtsoziologie   familiale Sozialisation

Statistical relations:
Mappings between user terms
and controlled terms

13

Service: Reranking (Bradfordizing)
Sorting central papers in core journals/publishers on top

Query: jugendsoziologie and collection:iz*   [Search]
Only show metadata sets which include an abstract (slower) ☐

Journal context

Suggest search terms

Controlled vocabulary
Thesaurus Sozialwissenschaften ▾
Automatic query expansion ☐

Rerank the result list

Rerank method
Bradfordizing ▾

Interactive query enhancement

term cloud   search term suggestions   central journals

Total hits: 2539

1. Bendit, Rene; Gaiser, Wolfgang (1992): *Einführung in das Thema: 'Europa und sein Nachwuchs'* in **DISKURS Studien zu Kindheit, Jugend, Familie und Gesellschaft Jg. 2 H. 2** (0937-9614) toggle details
2. Mauger, Gerard (1992): *Jugend in Frankreich : Reproduktion sozialer Ungleichheit im gesellschaftlichen Wandel* in **DISKURS Studien zu Kindheit, Jugend, Familie und Gesellschaft Jg. 2 H. 2** (0937-9614) toggle details
3. (1992): *Die Horizontalisierung der Lebenswelt italienischer Jugendlicher* in **DISKURS Studien zu Kindheit, Jugend, Familie und Gesellschaft Jg. 2 H. 2** (0937-9614) toggle details
4. Bendit, Rene; Gaiser, Wolfgang; Nissen, Ursula (1992): *Aufwachsen in der Bundesrepublik Deutschland : Risiken und Chancen im modernen Sozialstaat* in **DISKURS Studien zu Kindheit, Jugend, Familie und Gesellschaft Jg. 2 H. 2** (0937-9614) toggle details

- DISKURS (0937-9614) [150]
- Aus Politik und Zeitgeschichte (0479-611X) [150]
- Deutsche Jugend (0012-0332) [146]
- Neue Praxis (0342-9857) [117]
- Zeitschrift für Soziologie der Erziehung und Sozialisation (1436-1957) [89]
- Televizion (0943-4755) [86]
- Kölner Zeitschrift für Soziologie und Sozialpsychologie (0023-2653) [74]
- Neue Sammlung (0028-3355) [69]
- Zeitschrift für Sozialisationsforschung und Erziehungssoziologie (0720-4361) [68]
- Medien und Erziehung (0723-399X) [52]

14

Service: Reranking (Author centrality)

Leibniz Gemeinschaft

Ranking by Author Centrality: sorting central author papers on top

Query: jugendsoziologie and collection:iz*   Search

Only show metadata sets which include an abstract (slower) ☐

Author context

Suggest search terms

Controlled vocabulary

Thesaurus Sozialwissenschaften ▾

Automatic query expansion ☐

Rerank the result list

Rerank method

Author Centrality ▾

Interactive query enhancement

term cloud   search term suggestions   central journals   central persons

Total hits: 12113

Co-authorship network

1. Heitmeyer, Wilhelm; C...
   Bedeutung für die Ver...
   JugendforschungIndiv...
   subjektive Verarbeitu...
   (3-7799-0414-4) toggle...

2. Heitmeyer, Wilhelm; C...
   gesellschaftliche Proz...
   Konsequenzen in **Juge...**

3. Heitmeyer, Wilhelm; M...
   Chancen und Belastun...
   Orientierungssicherhe...
   gesellschaftliche Proz...
   Konsequenzen (3-779...
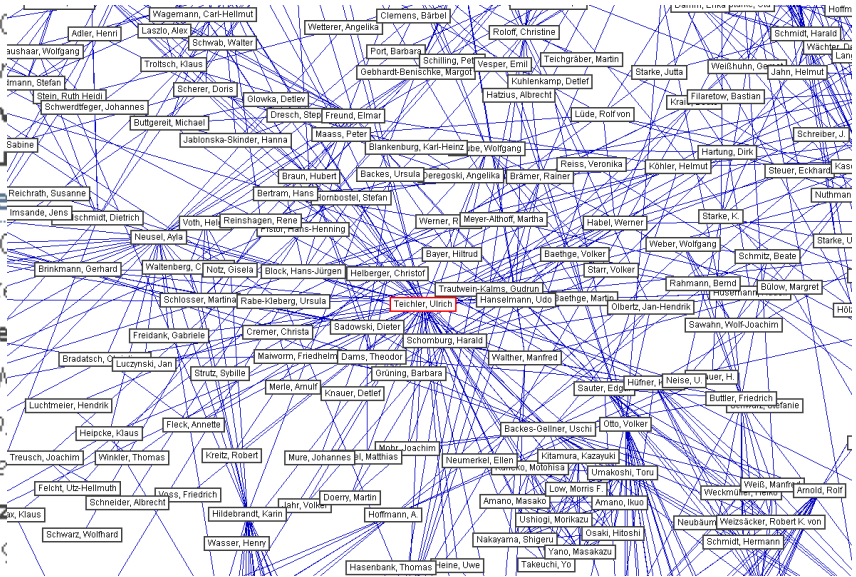
- Heitmeyer, Wilhelm [0.02113342]
- Schubarth, Wilfried [0.02075069]
- Krüger, Heinz-Hermann [0.01957099]
- DuBois-Reymond, Manuela [0.018002631]
- Boehnke, Klaus [0.017718809]
- Hübner-Funk, Sibylle [0.01680161]
- Hurrelmann, Klaus [0.014602789]
- Zinnecker, Jürgen [0.012975149]
- Melzer, Wolfgang [0.012351238]
- Lindner, Bernhard [0.011778912]
- Becker, Helmut [0.010350749]
- Neumann-Braun, Klaus [0.009717477]
- Eigenbrodt, Jörg [0.009552841]

Evaluation
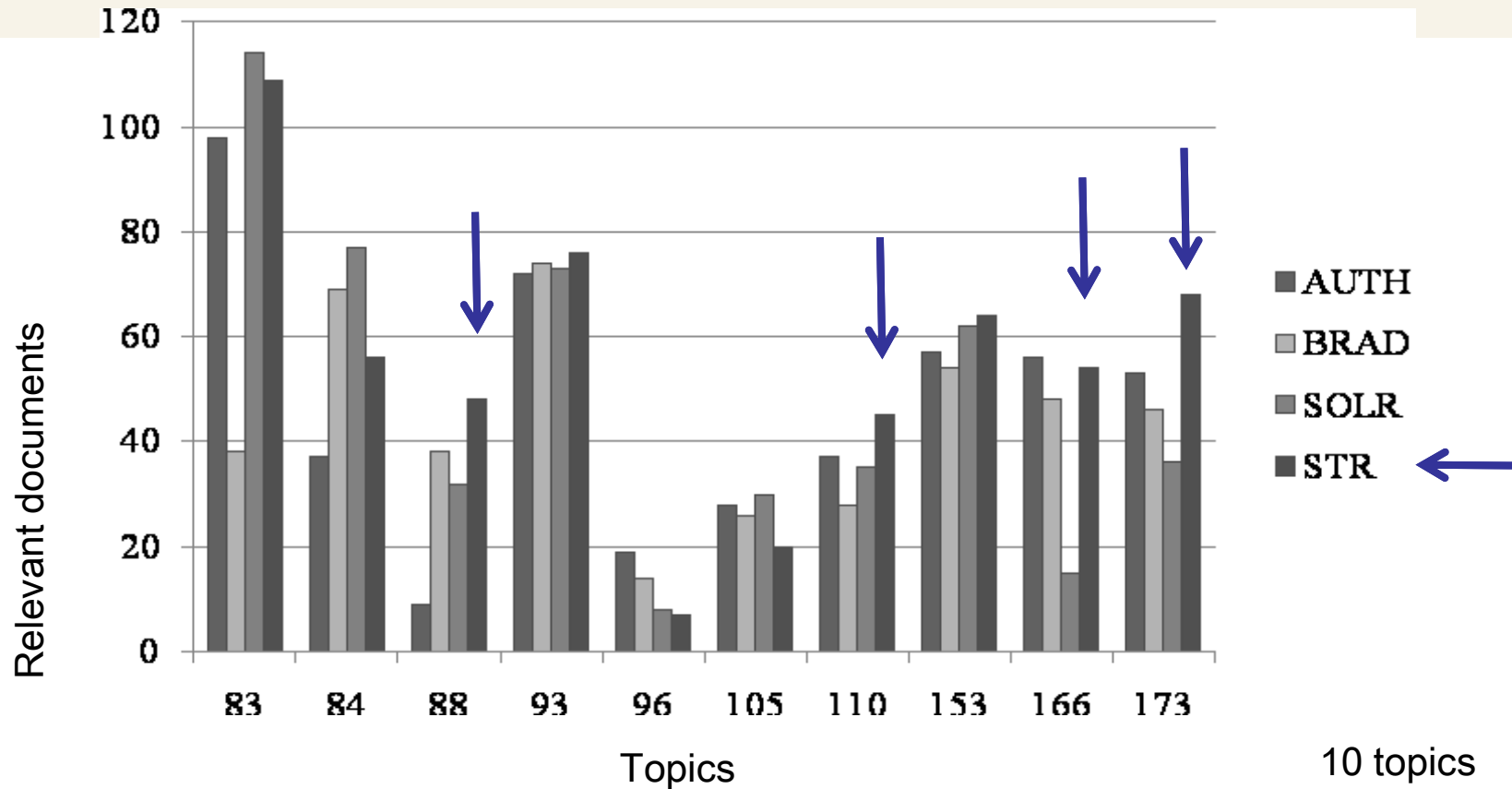
Number of documents judged relevant by the assessors per topic

- STR by adding the 4 descriptors with the highest confidence
- STR was best service (topics 88, 110, 166 and 173 with an improvement of 20% at least compared to the baseline)

10 topics
1 database
top 10 docs
4 services
73 students

Mayr et al., submitted

Evaluation

SOLR

STR

14

5 5

0

6 3

3

Mayr et al., submitted

BRAD

AUTH
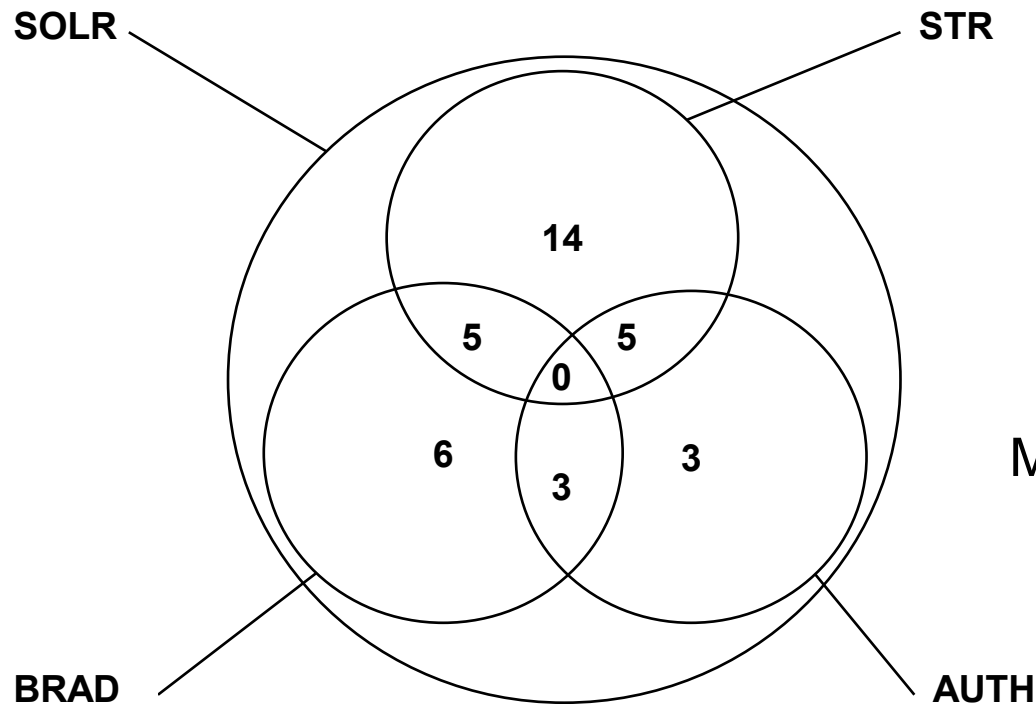
36 intersections of suggested top n=10 documents over all topics
and services (total of 400 documents)
• Result sets are nearly disjoint
• Services provide quite different views onto the document space

Next steps

Searching in a multi-thesauri-scenario (connecting services)

Interactive retrieval with terminology services support

SKOS

One open standard for representing controlled vocabularies on the <u>Web</u>

Usage will be eased for:

- Web users

- <u>**Applications of the Semantic Web (SPARQL)**</u>

Objective: Sharing and interlinking with other standardized data sources → And see what happens!

Leibniz
Gemeinschaft

Multi-Thesauri-Scenario

SKOS Mapping Properties

- „skos:exactMatch"

- „skos:closeMatch"

- „skos:broadMatch"

- „skos:narrowMatch"

- „skos:relatedMatch"

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

Modeling of terminology mappings with SKOS
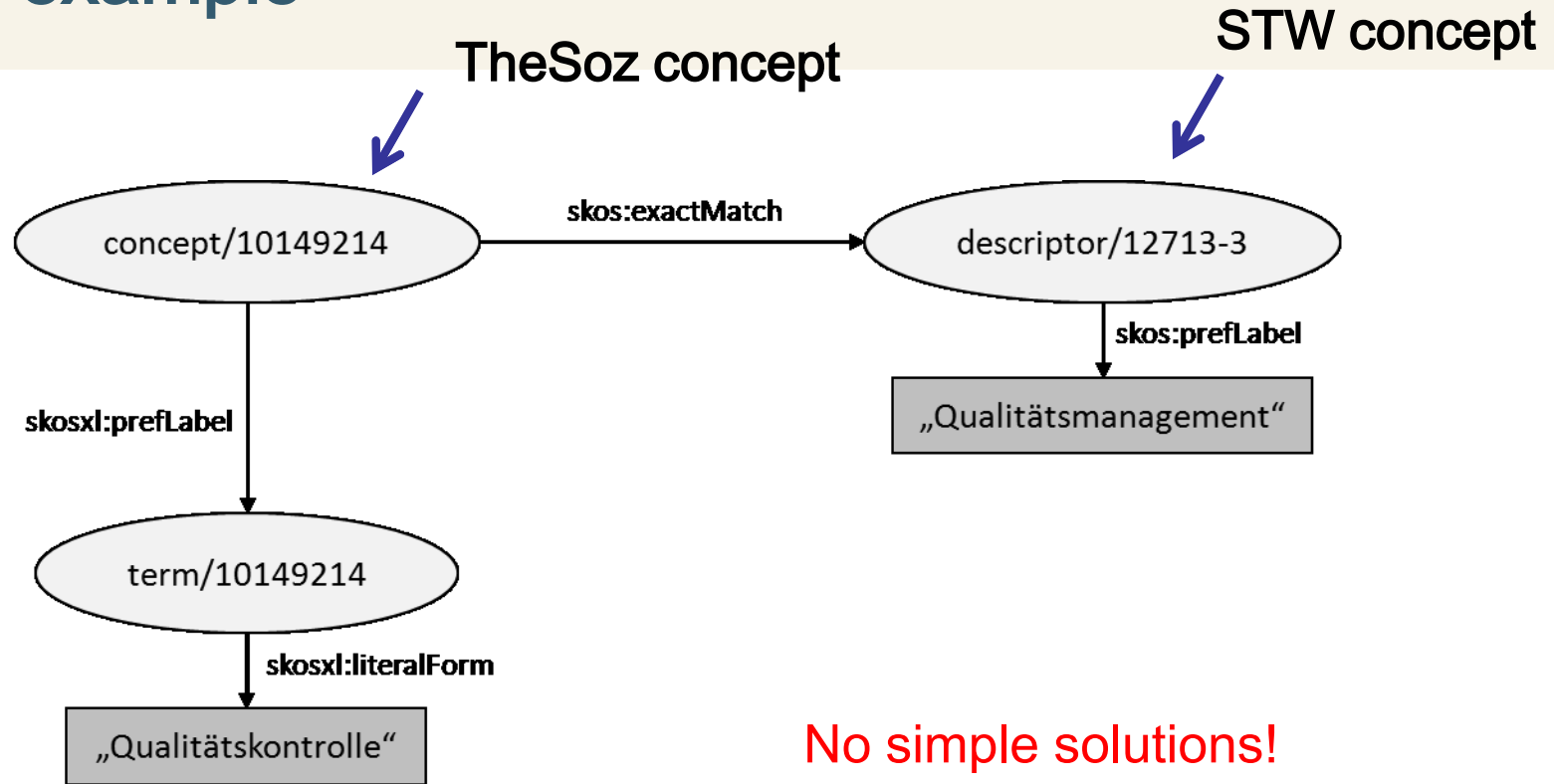
Mapping between **concepts**

vs.

Mapping between **terms**

SKOS approach: Mapping <u>between URIs of involved concepts</u>

Problem: Most mapping have been build between terms

SKOS example

STW concept

TheSoz concept

No simple solutions!

Example: Mapping between TheSoz and STW in SKOS

Mapping between URIs of involved concepts

LOD as a first step to publish data

http://lod.gesis.org

Thesaurus Sozialwissenschaften (TheSoz) at GESIS Linked Data Prototype

http://lod.gesis.org/thesoz/

Property	Value
cc:attributionName	■ GESIS - Leibniz Institute for the Social Sciences (en) ■ GESIS - Leibniz-Institut für Sozialwissenschaften (de)
cc:attributionURL	■ <http://www.gesis.org>
dcterms:creator	■ GESIS - Leibniz Institute for the Social Sciences (en) ■ GESIS - Leibniz-Institut für Sozialwissenschaften (de)
dcterms:hasVersion	■ 0.5
is skos:inScheme	

at GESIS Linked Data Prototype

http://lod.gesis.org/thesoz/concept/10034400

Property	Value
skosxl:altLabel	■ <http://lod.gesis.org/thesoz/term/10034398>
skos:broader	■ <http://lod.gesis.org/thesoz/concept/10034393>
skos:inScheme	■ <http://lod.gesis.org/thesoz/>
is skos:narrower of	■ <http://lod.gesis.org/thesoz/concept/10034393>
skosxl:prefLabel	■ <http://lod.gesis.org/thesoz/term/10034400>
is skos:related of	■ <http://lod.gesis.org/thesoz/concept/10036106>
skos:related	■ <http://lod.gesis.org/thesoz/concept/10036106>
rdf:type	■ ext:Descriptor

This page shows information obtained from the SPARQL endpoint at http://lod.gesis.org/thesoz/sparql.
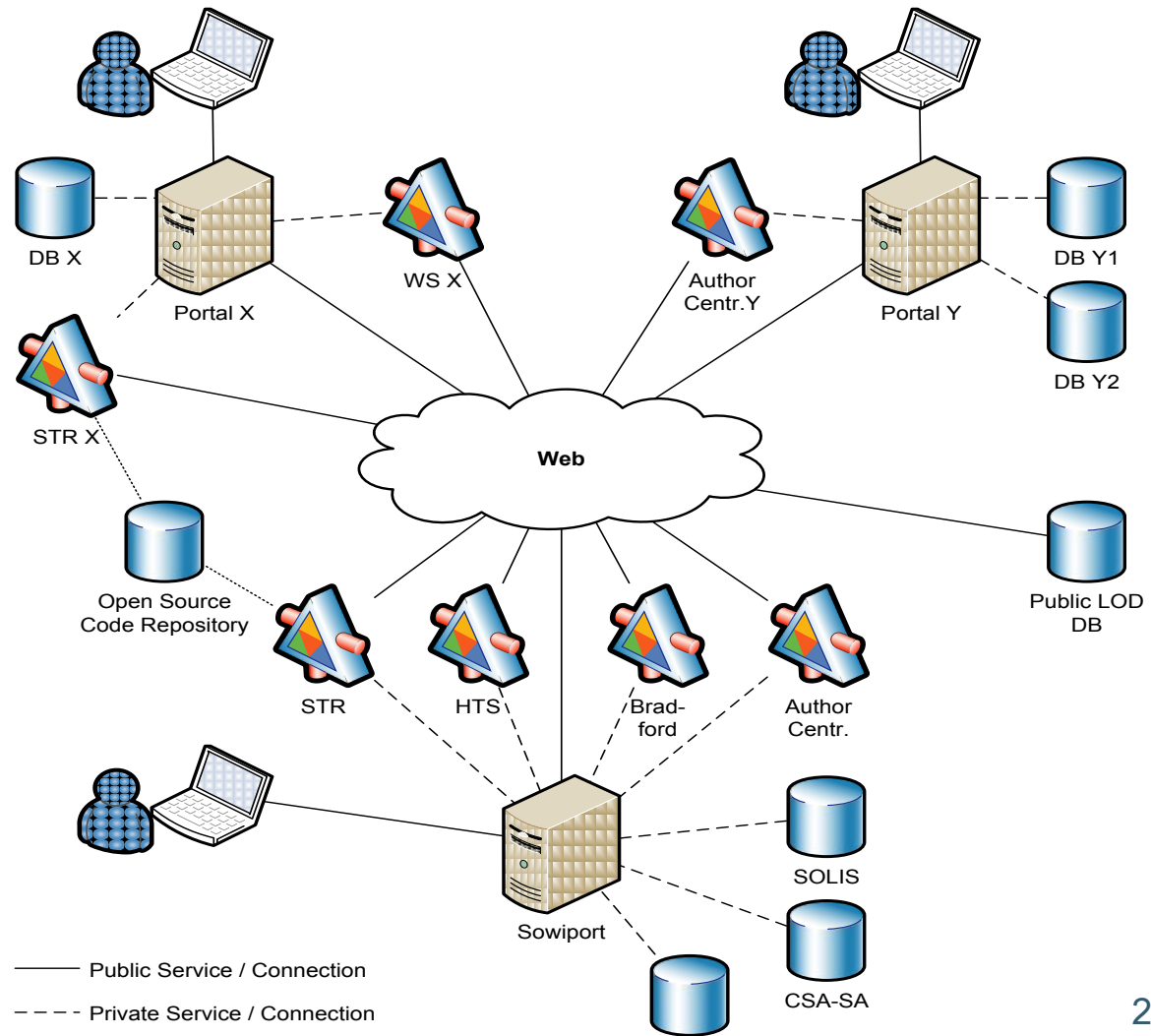As N3 | As RDF/XML

The GESIS Linked Data Prototype uses the Pubby Linked Data Frontend. More information on Linked Open Data at GESIS can be found here.

Perspectives

Next steps:
- further user evaluation of the services
- Combination of services
- Reusability of the services

Value-added services as public web services

DB X
Portal X
WS X
Author Centr.Y
Portal Y
DB Y1
DB Y2
STR X
Web
Open Source Code Repository
Public LOD DB
STR
HTS
Brad-ford
Author Centr.
Sowiport
SOLIS
CSA-SA

——— Public Service / Connection
- - - - Private Service / Connection

Short conclusions

1. Mappings matter! See retrieval results …

   - Interdisciplinary applications?

2. Contexts matter! See retrieval results …

   - Providing context to stimulate and support users

   - Support interactive systems for exploration

Next steps

1. Provide data as linked open data

2. Provide linked open services

3. Experiment with users

4. Experiment with applications

Visual Terminology Browser

Suchbeg jugend
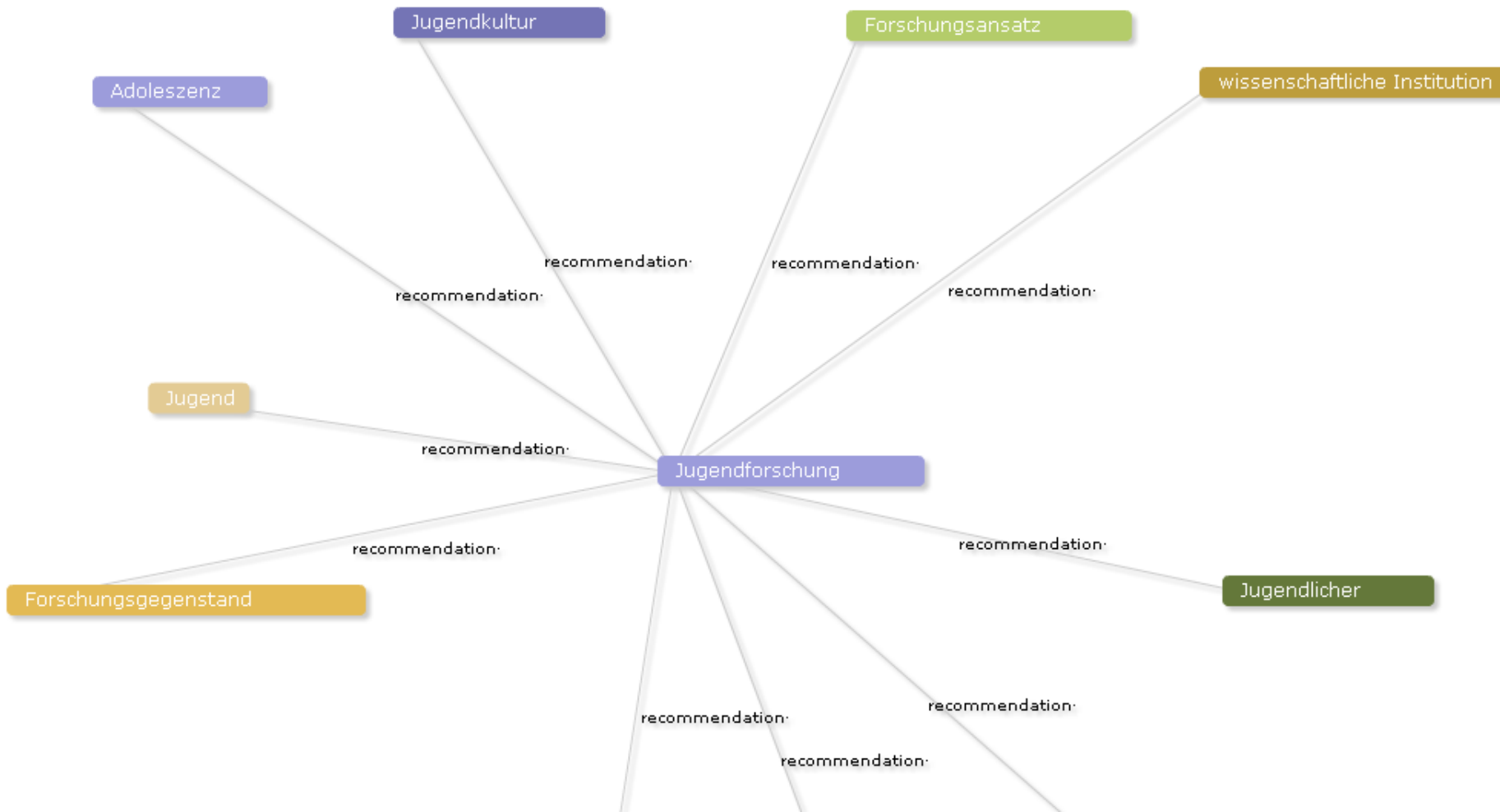
References

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; Schaer, Philipp; Sure, York (submitted): Science Models as Value-Added Services for Scholarly Information Systems.

Dr. Philipp Mayr
GESIS
Lennéstr. 30
53113 Bonn

mailto:philipp.mayr@gesis.org

http://www.gesis.org/index.php?id=2479