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

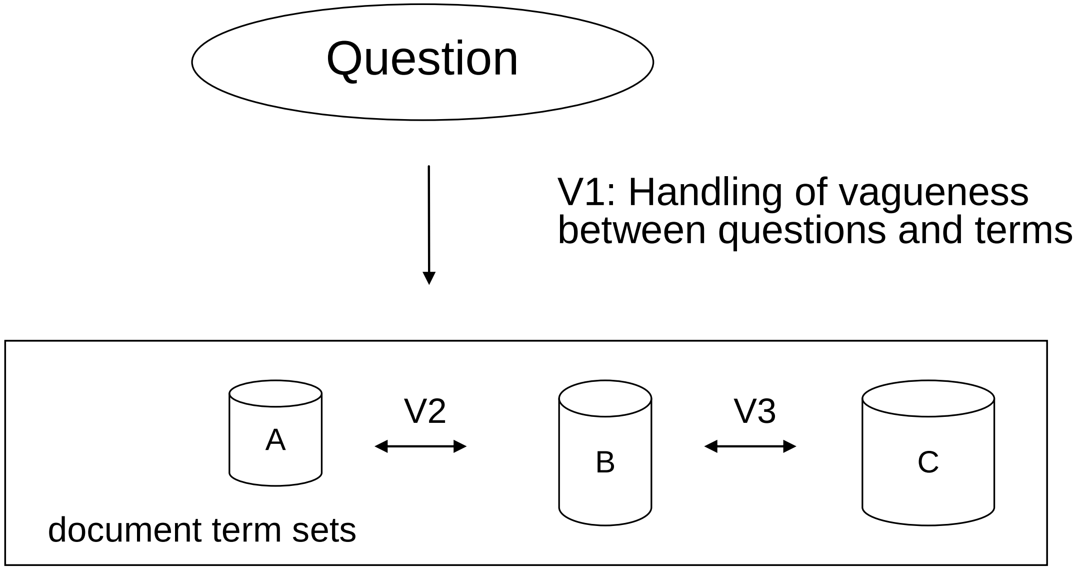
Theoretical background

- Vagueness between terms
  - Language ambiguity
  - Meaning of terms
- Semantic heterogeneity in document collections
- Problems while indexing documents
  - Consistency
  - Precision
  - Topicality

Background

2 step methodology

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



V1: Handling of vagueness between questions and terms

Cross-concordances are used for V2 and V3

V2/V3:
Bilateral handling of vagueness

Project - background

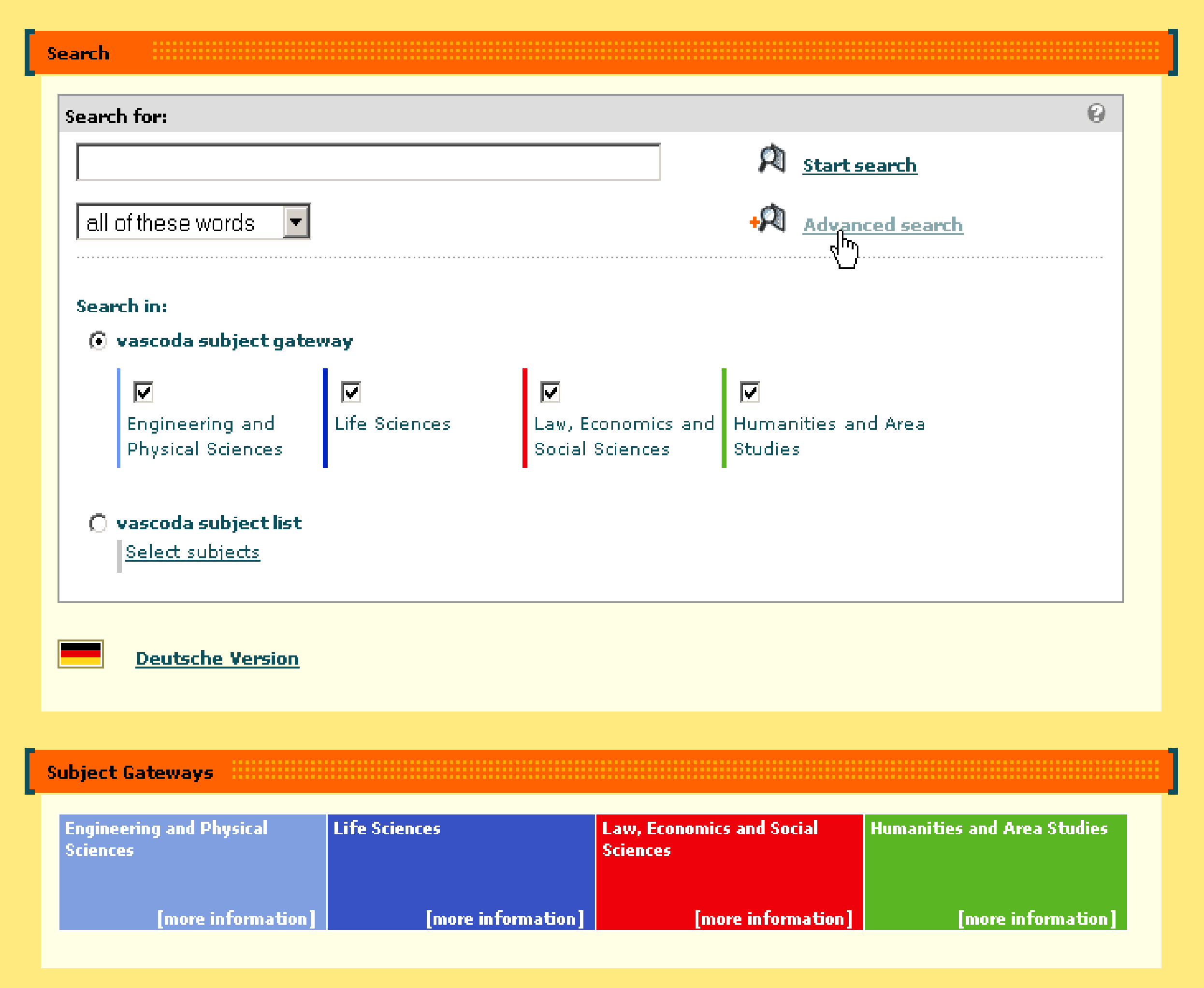
vascoda approach: an interdisciplinary portal (DL) for scientific information

- 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



Search

Search for:

Start search

all of these words ▾

Advanced search

Search in:

● vascoda subject gateway

☑ Engineering and Physical Sciences

☑ Life Sciences

☑ Law, Economics and Social Sciences

☑ Humanities and Area Studies

○ vascoda subject list

Select subjects

Deutsche Version

Subject Gateways

Engineering and Physical Sciences [more information]

Life Sciences [more information]

Law, Economics and Social Sciences [more information]

Humanities and Area Studies [more information]

Project

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

Project

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

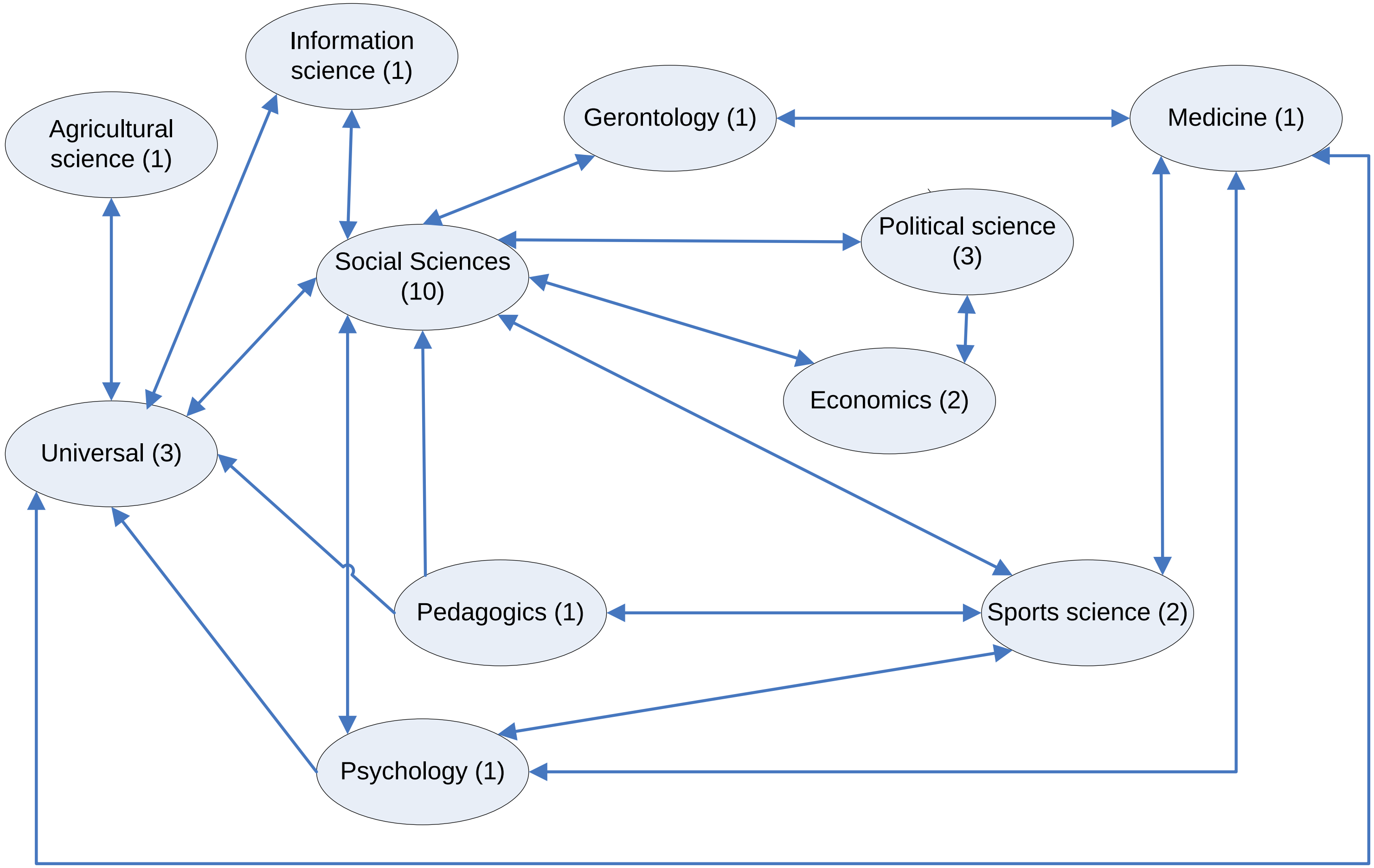
Controlled vocabularies

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



Controlled vocabularies – overview 1

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

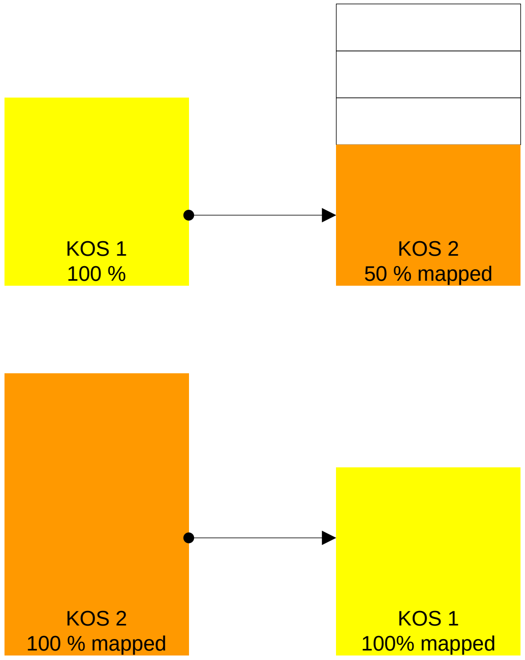
Controlled vocabularies – overview 2

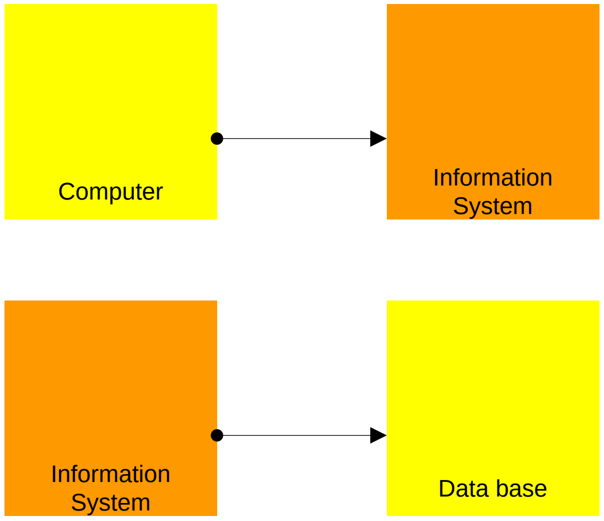
16	RVK	Regensburger Verbundklassifikation	universal	Classification	part
17	SWD	Schlagwortnormdatei	universal	Subject Headings	part
18	STW	Standard Thesaurus Wirtschaft	economics	Thesaurus	complete
19	Bildung	Thesaurus Bildung	pedagogics	Thesaurus	part
20	DZI	Thesaurus of the Deutschen Instituts für soziale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

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)

term KOS 1	relation	term(s) KOS n
hacker	=	Hacking
hacker	^+	Computers + Crime
hacker	^+	Internet + Security
ISDN	0	
ISDN	<	Telecommunications
documentation system	>	Abstracting services

Cross-concordances - overview

	Voc. name	Voc. name	type	status	year
1	TheSoz	STW	bilateral	imported	2004
2	TheSoz	BiSp	bilateral	imported	2004
3	Psyndex	BISp	bilateral	imported	2004
4	BISp	Bildung		imported	2004
5	TheSoz	DZI	bilateral	imported	2005
6	TheSoz	FES	bilateral	imported	2005
7	TheSoz	IBLK	bilateral	imported	2005
8	TheSoz	Gerolit	bilateral	imported	2005
9	MeSH	BISp	bilateral	imported	2005
10	STW	IBLK	bilateral	imported	2005
11	TheSoz	CSA-WPSA	bilateral	imported	2006
12	TheSoz	CSA-ASSIA	bilateral	imported	2006
13	TheSoz	ELSST	bilateral	imported	2006
14	TheSoz	CSA-PEI	bilateral	imported	2006
15	MeSH	Psyndex	bilateral	imported	2006
16	MeSH	Gerolit	bilateral	imported	2006
17	IBLK	CSA-PAIS	bilateral	imported	2006
18	IBLK	TWSE	bilateral	imported	2006
19	INION	TheSoz			2007
20	INFODATA	SWD	bilateral	imported	2007
21	INFODATA	TheSoz	bilateral	imported	2007
22	IAB	TheSoz	bilateral	imported	2007
23	IAB	STW	bilateral	imported	2007
24	SWD	MeSH	bilateral		2007
25	SWD	AGROVOC	bilateral		2007
26	JEL	STW		ready	2007
27	RVK	DDC		ready	2007

7 further mappings from the previous projects

infoconnex and CARMEN

Data base

Vocabularies: 25

Mappings: 28 bilateral, 6 unilateral

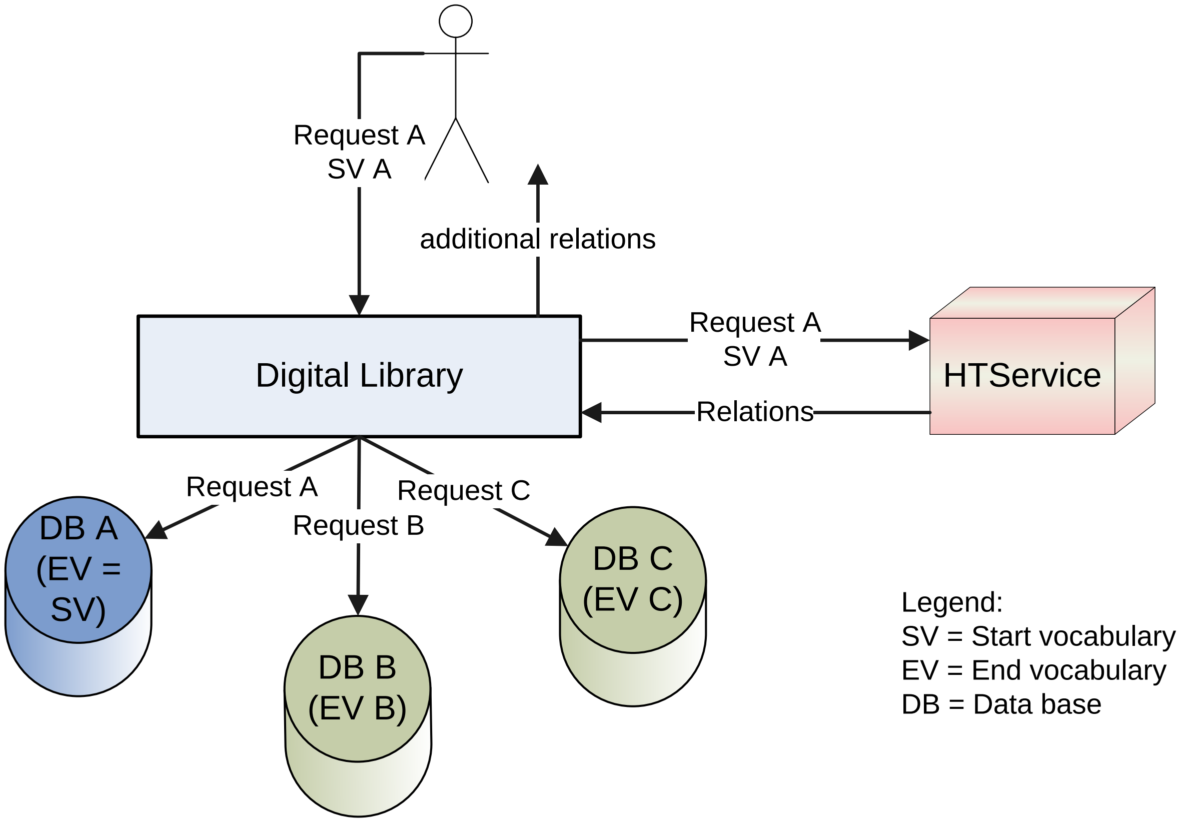
Size: round 396,000 relations to date

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)



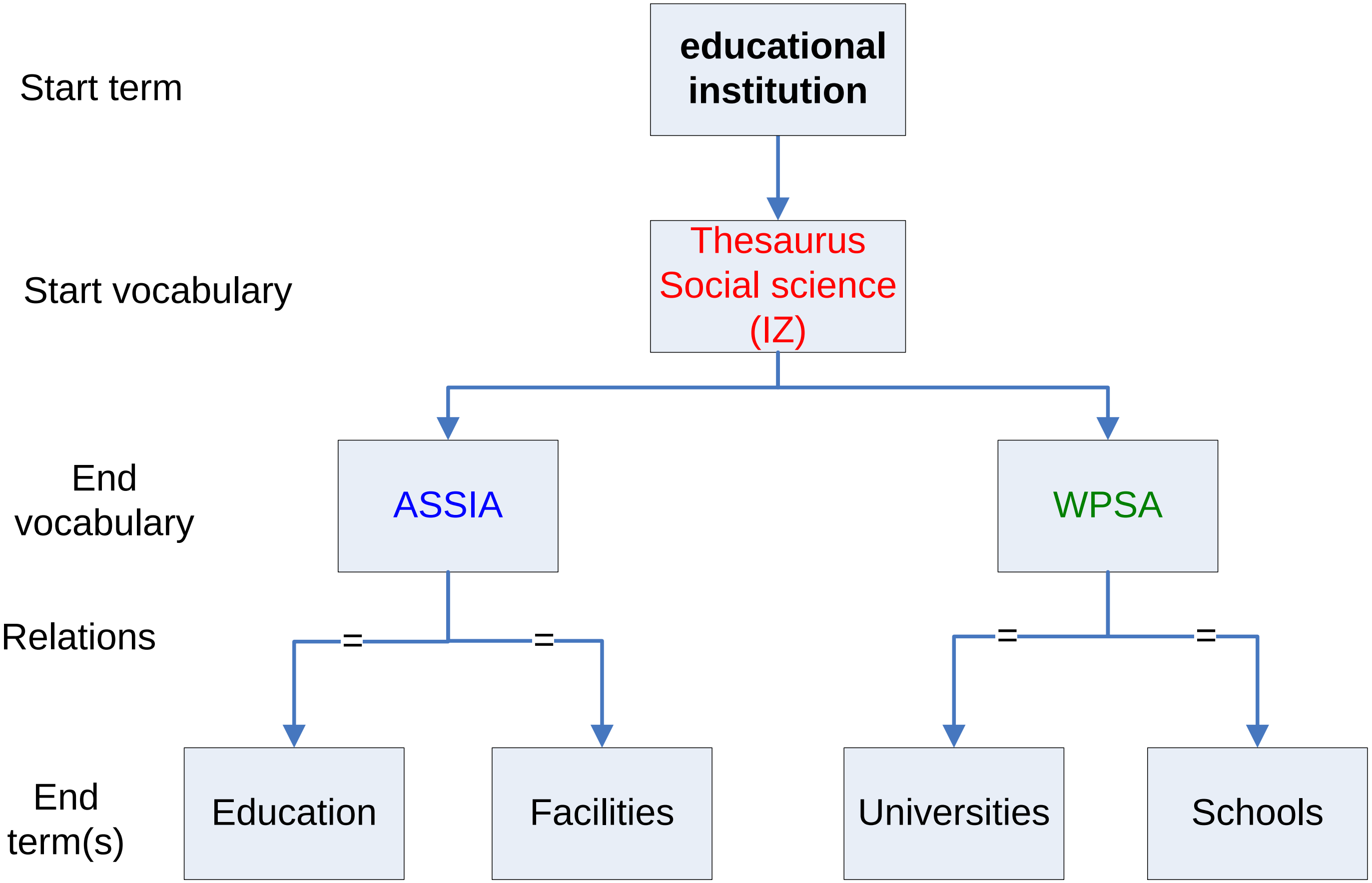
Legend:
SV = Start vocabulary
EV = End vocabulary
DB = Data base

2 scenarios

- Just transform into equivalence relations

- Present additional relations to users

Heterogeneity Service



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

Evaluation – preliminary results

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

SWD-TheSoz		CT	TT	FT
5 topics	Recall	**0,5817**	**0,5817**	**0,684**
	Precision	**0,3663**	**0,3663**	**0,3642**

TheSoz-DZI		CT	TT	FT
10 topics	Recall	**0.5907**	**0.7602**	**0.5327**
	Precision	**0.3173**	**0.3914**	**0.6760**

STW-TheSoz		CT	TT	FT
6 topics	Recall	**0.2807**	**0.5859**	**0.5944**
	Precision	**0.2351**	**0.3634**	**0.3199**

- TT improves over CT, but not necessarily over FT
- FT generates more doc (FT search controlled terms too)

Summary & Outlook

All related cross-concordances will be used in sowiport

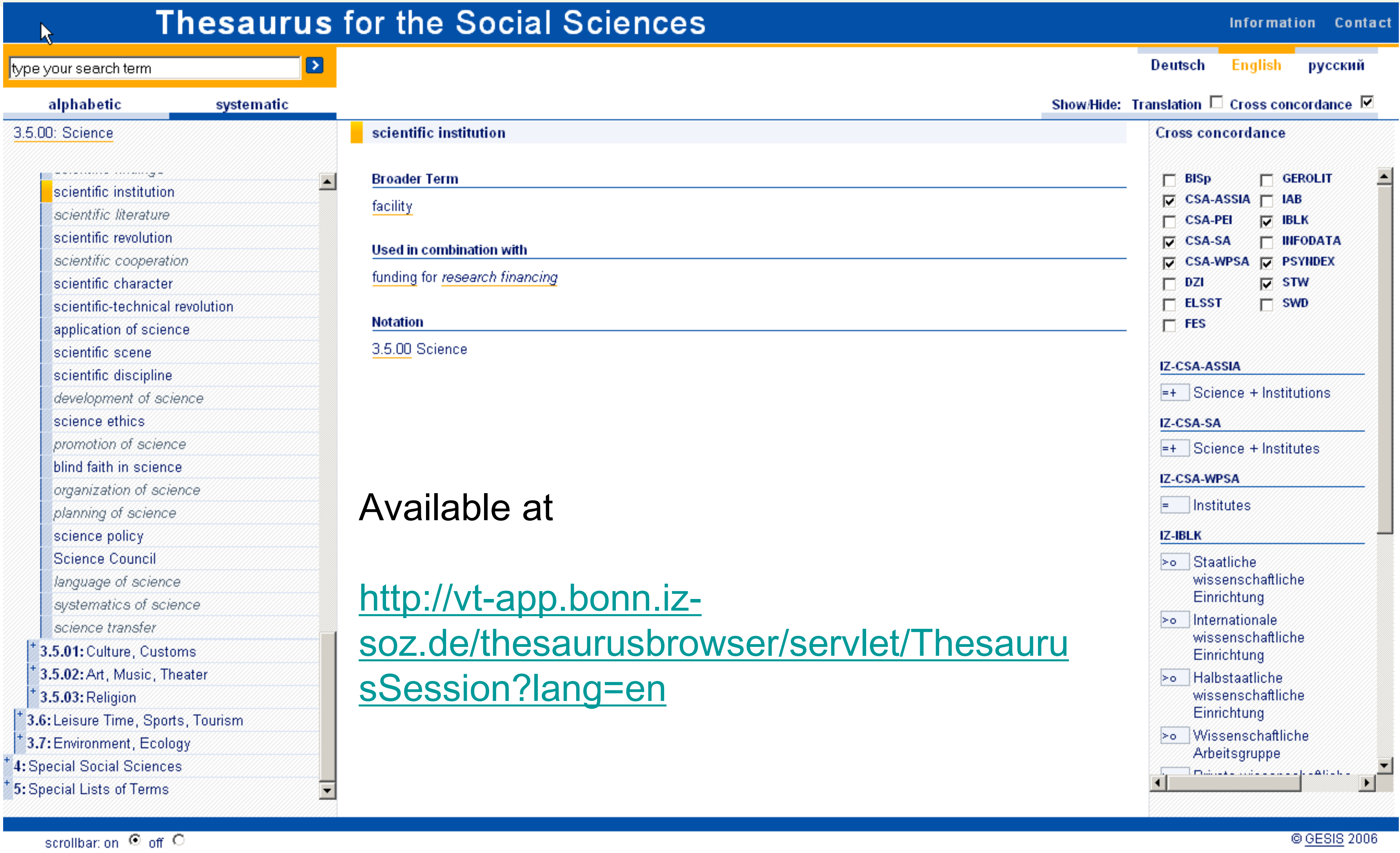
Results of the quantitative and retrieval evaluation will be finished next month

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



Thesaurus for the Social Sciences

Information Contact

type your search term

Deutsch English русский

alphabetic systematic

Show/Hide: Translation □ Cross concordance ☑

3.5.00: Science

- scientific institution
- *scientific literature*
- scientific revolution
- *scientific cooperation*
- scientific character
- scientific-technical revolution
- application of science
- scientific scene
- scientific discipline
- *development of science*
- science ethics
- *promotion of science*
- blind faith in science
- *organization of science*
- *planning of science*
- science policy
- Science Council
- *language of science*
- *systematics of science*
- *science transfer*
- + **3.5.01:** Culture, Customs
- + **3.5.02:** Art, Music, Theater
- + **3.5.03:** Religion
- + **3.6:** Leisure Time, Sports, Tourism
- + **3.7:** Environment, Ecology
- + **4:** Special Social Sciences
- + **5:** Special Lists of Terms

scientific institution

Broader Term

facility

Used in combination with

funding for *research financing*

Notation

3.5.00 Science

Cross concordance

□ BISp
☑ CSA-ASSIA
□ CSA-PEI
☑ CSA-SA
☑ CSA-WPSA
□ DZI
□ ELSST
□ FES
□ GEROLIT
□ IAB
☑ IBLK
□ INFODATA
☑ PSYNDEX
☑ STW
□ SWD

IZ-CSA-ASSIA

=+ Science + Institutions

IZ-CSA-SA

=+ Science + Institutes

IZ-CSA-WPSA

= Institutes

IZ-IBLK

>o Staatliche wissenschaftliche Einrichtung

>o Internationale wissenschaftliche Einrichtung

>o Halbstaatliche wissenschaftliche Einrichtung

>o Wissenschaftliche Arbeitsgruppe

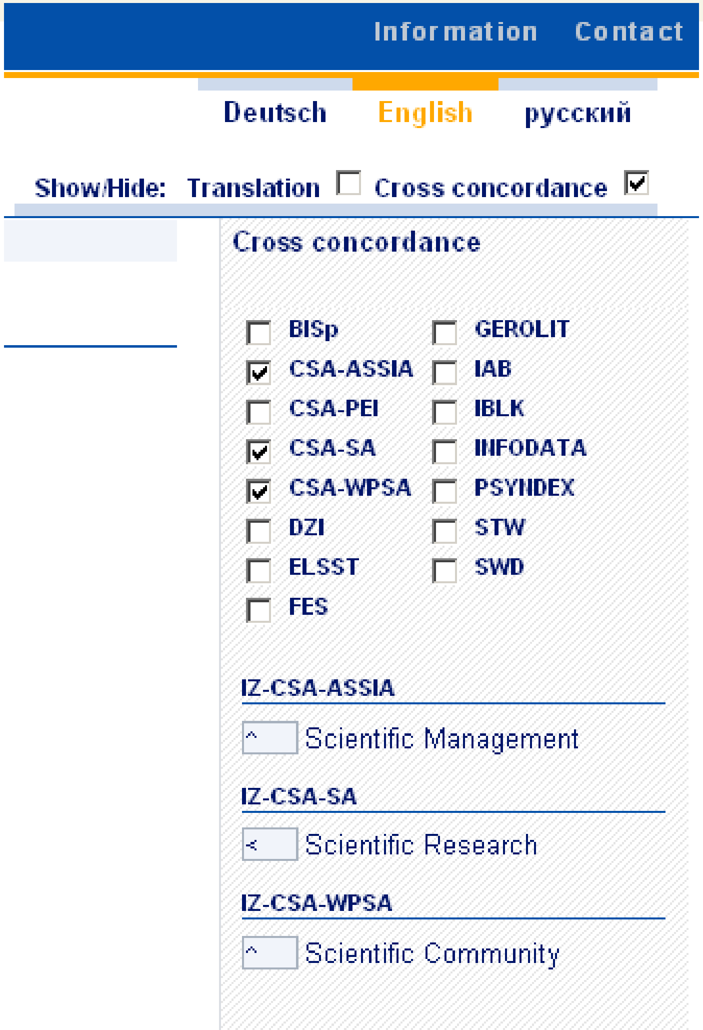
scrollbar: on ● off ○

© GESIS 2006

Available at

http://vt-app.bonn.iz-soz.de/thesaurusbrowser/servlet/ThesaurusSession?lang=en

Online-Thesaurus



Information Contact

Deutsch English русский

Show/Hide: Translation □ Cross concordance ☑

Cross concordance

- □ BISp
- ☑ CSA-ASSIA
- □ CSA-PEI
- ☑ CSA-SA
- ☑ CSA-WPSA
- □ DZI
- □ ELSST
- □ FES
- □ GEROLIT
- □ IAB
- □ IBLK
- □ INFODATA
- □ PSYNDEX
- □ STW
- □ SWD

IZ-CSA-ASSIA

^ Scientific Management

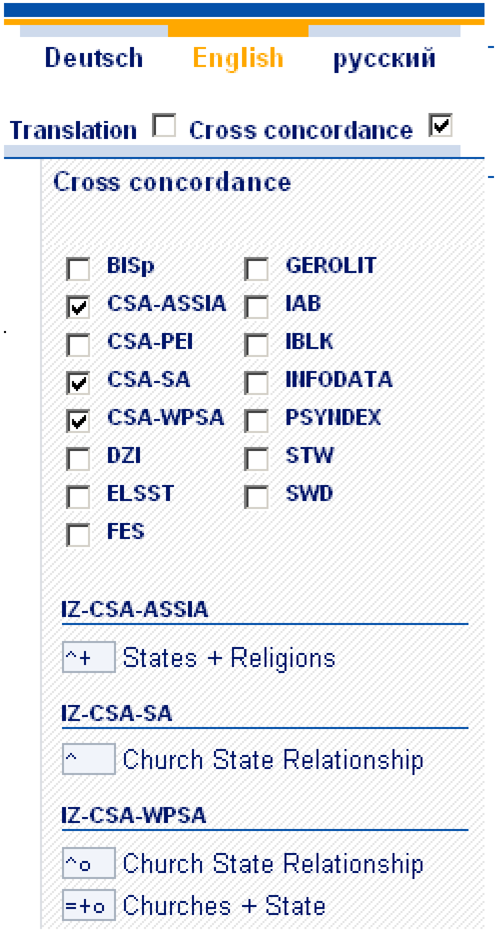
IZ-CSA-SA

< Scientific Research

IZ-CSA-WPSA

^ Scientific Community

1) Scientific scene



Deutsch English русский

Translation □ Cross concordance ☑

Cross concordance

- □ BISp
- ☑ CSA-ASSIA
- □ CSA-PEI
- ☑ CSA-SA
- ☑ CSA-WPSA
- □ DZI
- □ ELSST
- □ FES
- □ GEROLIT
- □ IAB
- □ IBLK
- □ INFODATA
- □ PSYNDEX
- □ STW
- □ SWD

IZ-CSA-ASSIA

^+ States + Religions

IZ-CSA-SA

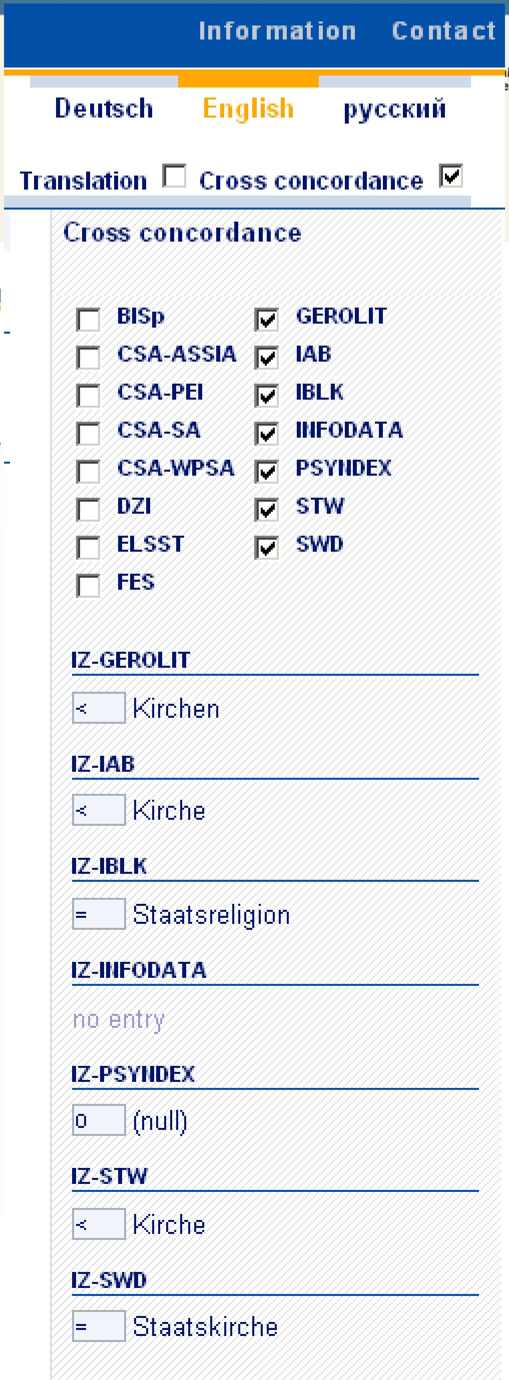
^ Church State Relationship

IZ-CSA-WPSA

^o Church State Relationship

=+o Churches + State

2) State church



Information Contact

Deutsch English русский

Translation □ Cross concordance ☑

Cross concordance

- □ BISp
- □ CSA-ASSIA
- □ CSA-PEI
- □ CSA-SA
- □ CSA-WPSA
- □ DZI
- □ ELSST
- □ FES
- ☑ GEROLIT
- ☑ IAB
- ☑ IBLK
- ☑ INFODATA
- ☑ PSYNDEX
- ☑ STW
- ☑ SWD

IZ-GEROLIT

< Kirchen

IZ-IAB

< Kirche

IZ-IBLK

= Staatsreligion

IZ-INFODATA

no entry

IZ-PSYNDEX

o (null)

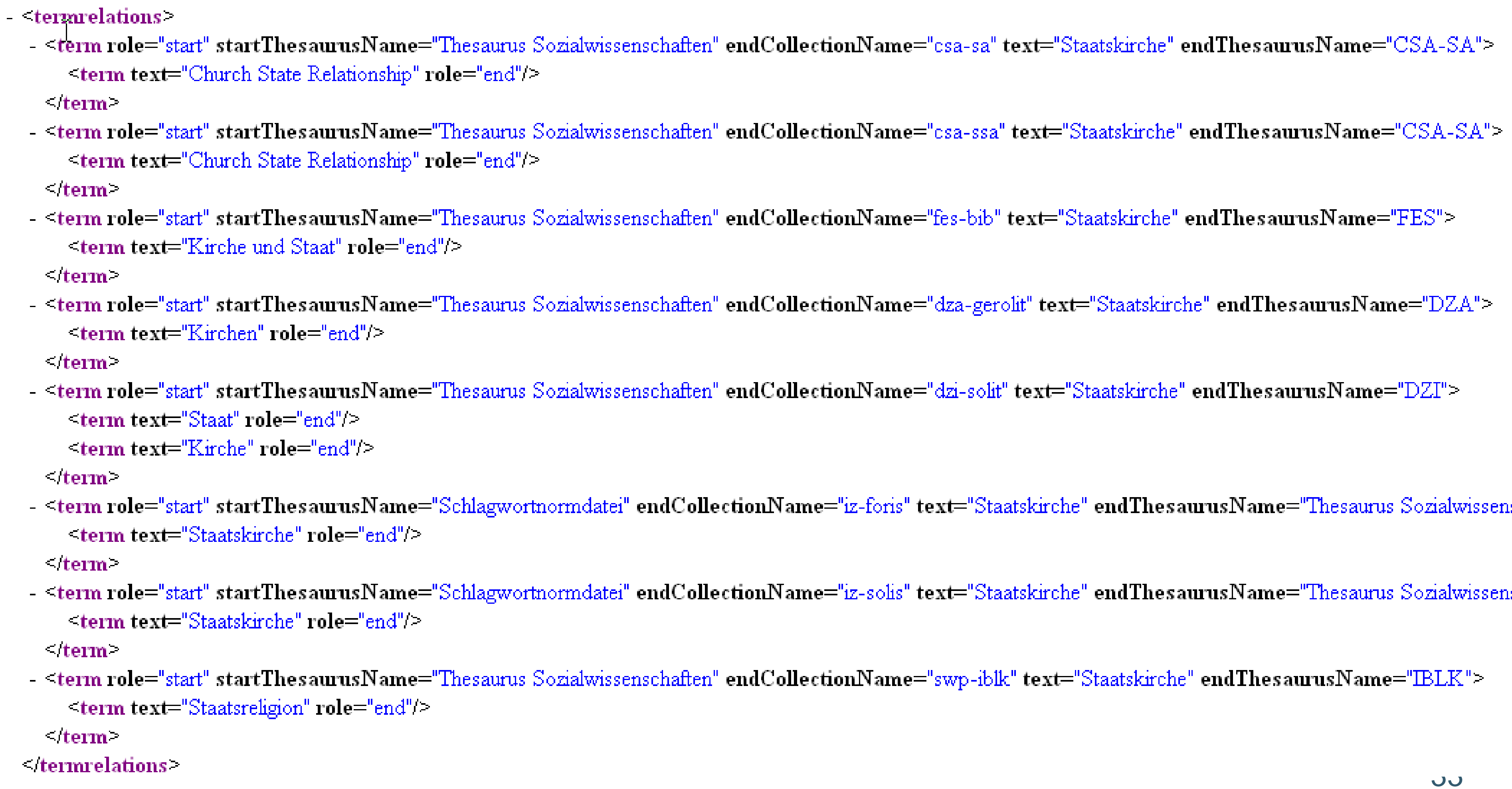
IZ-STW

< Kirche

IZ-SWD

= Staatskirche

Heterogeneity Service



```
- <termrelations>
  - <term role="start" startThesaurusName="Thesaurus Sozialwissenschaften" endCollectionName="csa-sa" text="Staatskirche" endThesaurusName="CSA-SA">
      <term text="Church State Relationship" role="end"/>
    </term>
  - <term role="start" startThesaurusName="Thesaurus Sozialwissenschaften" endCollectionName="csa-ssa" text="Staatskirche" endThesaurusName="CSA-SA">
      <term text="Church State Relationship" role="end"/>
    </term>
  - <term role="start" startThesaurusName="Thesaurus Sozialwissenschaften" endCollectionName="fes-bib" text="Staatskirche" endThesaurusName="FES">
      <term text="Kirche und Staat" role="end"/>
    </term>
  - <term role="start" startThesaurusName="Thesaurus Sozialwissenschaften" endCollectionName="dza-gerolit" text="Staatskirche" endThesaurusName="DZA">
      <term text="Kirchen" role="end"/>
    </term>
  - <term role="start" startThesaurusName="Thesaurus Sozialwissenschaften" endCollectionName="dzi-solit" text="Staatskirche" endThesaurusName="DZI">
      <term text="Staat" role="end"/>
      <term text="Kirche" role="end"/>
    </term>
  - <term role="start" startThesaurusName="Schlagwortnormdatei" endCollectionName="iz-foris" text="Staatskirche" endThesaurusName="Thesaurus Sozialwissen
      <term text="Staatskirche" role="end"/>
    </term>
  - <term role="start" startThesaurusName="Schlagwortnormdatei" endCollectionName="iz-solis" text="Staatskirche" endThesaurusName="Thesaurus Sozialwissen
      <term text="Staatskirche" role="end"/>
    </term>
  - <term role="start" startThesaurusName="Thesaurus Sozialwissenschaften" endCollectionName="swp-iblk" text="Staatskirche" endThesaurusName="IBLK">
      <term text="Staatsreligion" role="end"/>
    </term>
  </termrelations>
```

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