

# Zum Stand der Crosskonkordanzen in vascoda

*Philipp Mayr, Anne-Kathrin Walter*  
*GESIS / Informationszentrum Sozialwissenschaften, Bonn*

28. AGSB-Jahrestagung

28. März 2007

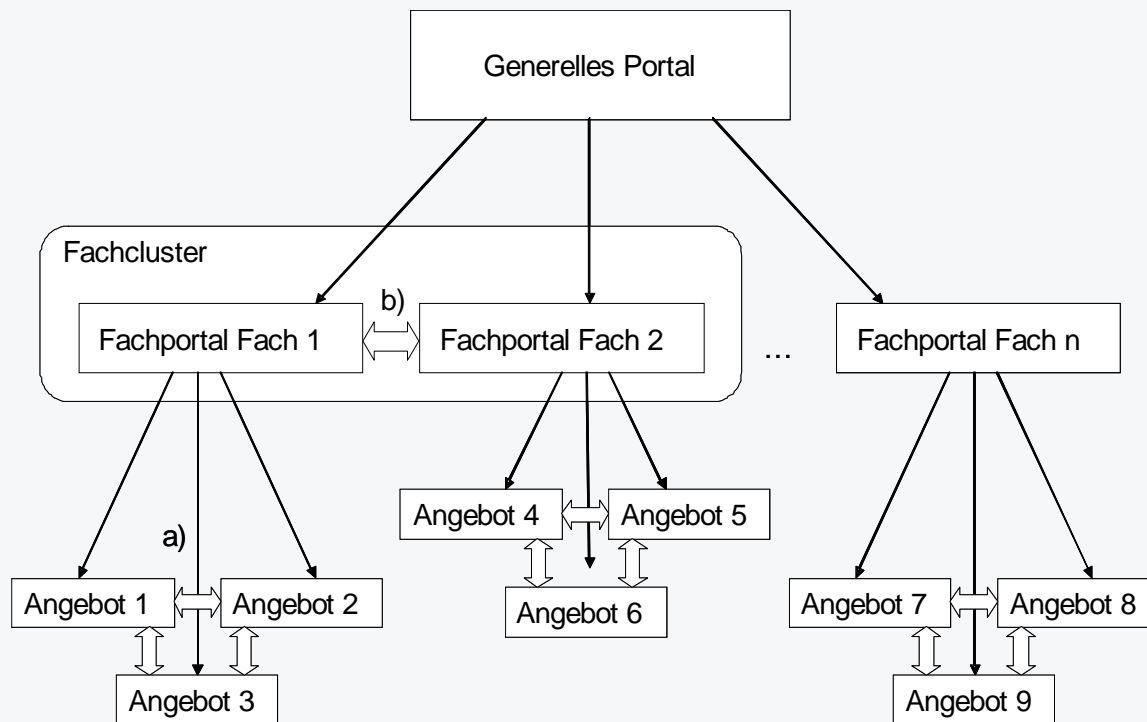


## Agenda

- Einführung
- Projekt Modellbildung/Heterogenitätsbehandlung
- Crosskonkordanzen
  - Verfahren
  - Übersicht
  - Einsatz
- Heterogenitätsservice
- Evaluation der Crosskonkordanzen
- Ausblick

# Leitlinien vascoda

- Integration: alle wichtigen Informationsanbieter und Datenkollektionen (Ausgangspunkt “deep web”)
- Einsatz kontrollierter Vokabulare
- Behandlung semantischer Heterogenität
- Fachcluster/Fachportale als Basis von vascoda
- Antwort auf Google/Google Scholar



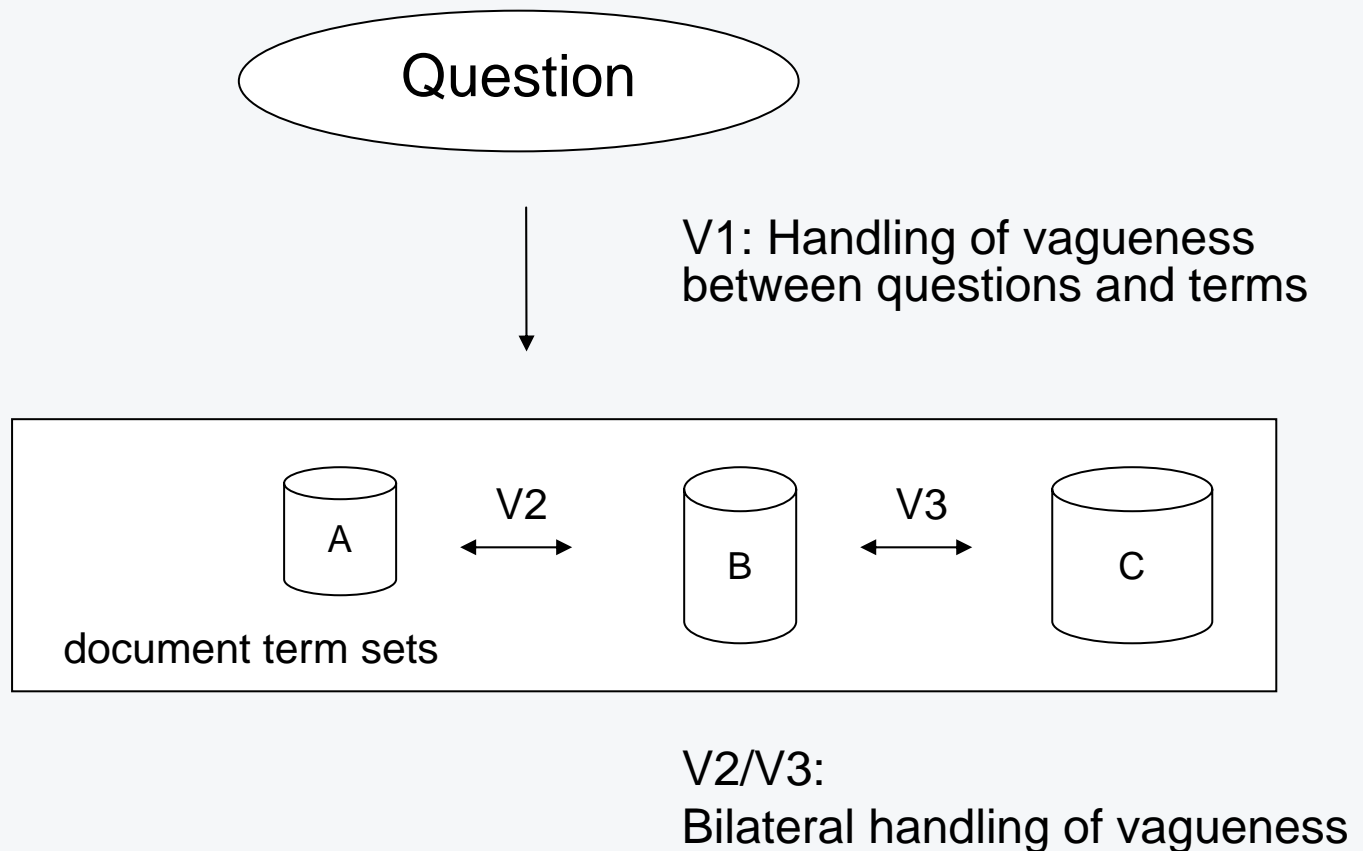
Kaskadierendes  
Modell einer  
Portalinfrastruktur  
(Mayr et al., 2005)



## Situation - Heterogenität in vascoda

- Fachportale sollen integrierte Recherche über mehrere Informationsangebote bieten (Metasuche – Standardierung der Metadaten)
  - Vielzahl unterschiedlicher Informationstypen
    - Internetquellen
    - Fachdatenbanken
    - OPACs
    - Volltexte
  - Unterschiedliche Erschließungssysteme der Informationsanbieter (Thesauri, Klassifikationen, andere kontrollierte Vokabulare)
- Semantische Heterogenität als unvermeidlich verbleibende Heterogenität

## Vagheitsbehandlung (Zweischritt-Modell)



- Behandlung von V1 (z.B. bei Petras, 2006)
- Behandlung von V2/V3 durch Module der semantischen Heterogenitätsbehandlung (insb. Crosskonkordanzen)



## 2) Projekt Modellbildung/Heterogenitätsbehandlung



## Projekte am IZ mit Crosskonkordanzen

### infoconnex:

Informationsverbund Bildung—Sozialwissenschaften—Psychologie  
Projektlaufzeit Juli 2001 bis Mai 2005 (gefördert vom BMBF)

### Kompetenznetzwerk Modellbildung und Heterogenitätsbehandlung:

Teilprojekt innerhalb des Kompetenznetzwerks „Neue Dienste,  
Standardisierung, Metadaten“ (gefördert vom BMBF)

- Modellbildung: übergreifende Modellbildung für komplexe Informationsinfrastrukturen, u. a. am Beispiel des Wissenschaftsportals „vascoda“ mit allen nachgeschalteten Ebenen
- Heterogenitätsbehandlung: als notwendige Ergänzung zur Standardisierung durch einheitliche Metadaten

Projektlaufzeit September 2004 bis August 2007

Verschiedene Ansätze, u.a.:

- Statistische Verfahren
- Intellektuelle Verfahren: **Crosskonkordanzen**

Crosskonkordanzen:

*Gerichtete, relevanzbewertete Relationen  
zwischen Termen zweier kontrollierter  
Vokabulare*

- Erweiterung des Suchraums, Erhöhung der Treffermenge
- Fokussierung auf ein Suchvokabular, kognitive Entlastung des Recherchierenden

# Was sind Crosskonkordanzen?

## Relationen

| Ausgangsterm      | Relation         | Relevanz | Zielterm                      |
|-------------------|------------------|----------|-------------------------------|
| Schifffahrtsrecht | =+o (Äquivalenz) | m        | Schiffahrt + Verkehrsrecht    |
| Schifffahrtsrecht | <o (Oberbegriff) | g        | Verkehrsrecht                 |
| Schifffahrtsrecht | ^o (verwandt)    | m        | Seerecht                      |
| Regionalstruktur  | > (Unterbegriff) | m        | Regionale Wirtschaftsstruktur |

Relevanzen bezogen auf die Treffermenge und Dokumentrelevanz

1:1- oder 1:n-Verknüpfungen

|                    |     |   |                               |
|--------------------|-----|---|-------------------------------|
| Biologieunterricht | <   | Unterricht                                    | DZI                           |
| Biologieunterricht | <   | Unterricht                                    | Standard Thesaurus Wirtschaft |
| Biologieunterricht | =   | Biologieunterricht                            | Schlagwortnormdatei           |
| Biologieunterricht | <+  | Biology + Teaching                            | CSA                           |
| Biologieunterricht | =+  | Naturwissenschaftlicher Unterricht + Biologie | Psyndex Terms                 |
| Biologieunterricht | =+  | Fachunterricht/Unterrichtsfach + Biologie     | IBLK                          |
| Biologieunterricht | =+o | Biologie + Schulfach                          | BISp-Liste                    |
| Biologieunterricht | <+o | Biologie + Unterrichtsstunde                  | BISp-Liste                    |
| Biologieunterricht | <+  | Biologie + Schule                             | DZA                           |
| Biologieunterricht | ^+  | Biologie + Unterricht                         | FES                           |

## Übersicht der kontrollierten Vokabulare (12/2006)

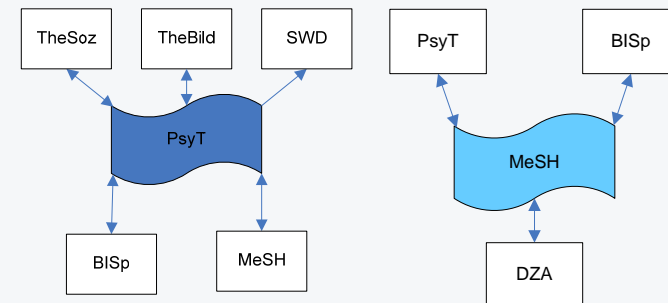
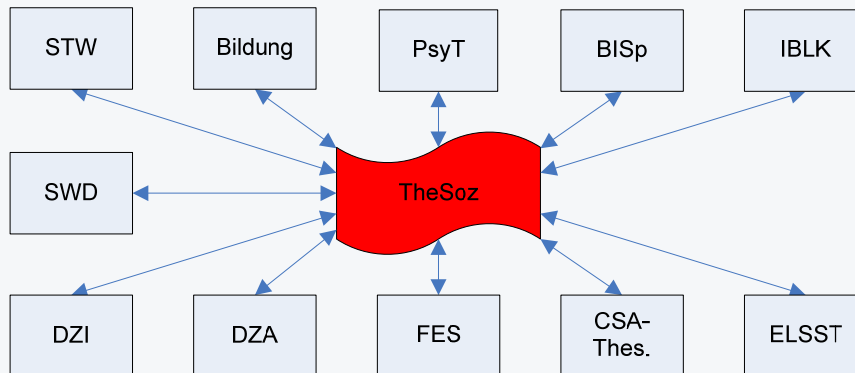
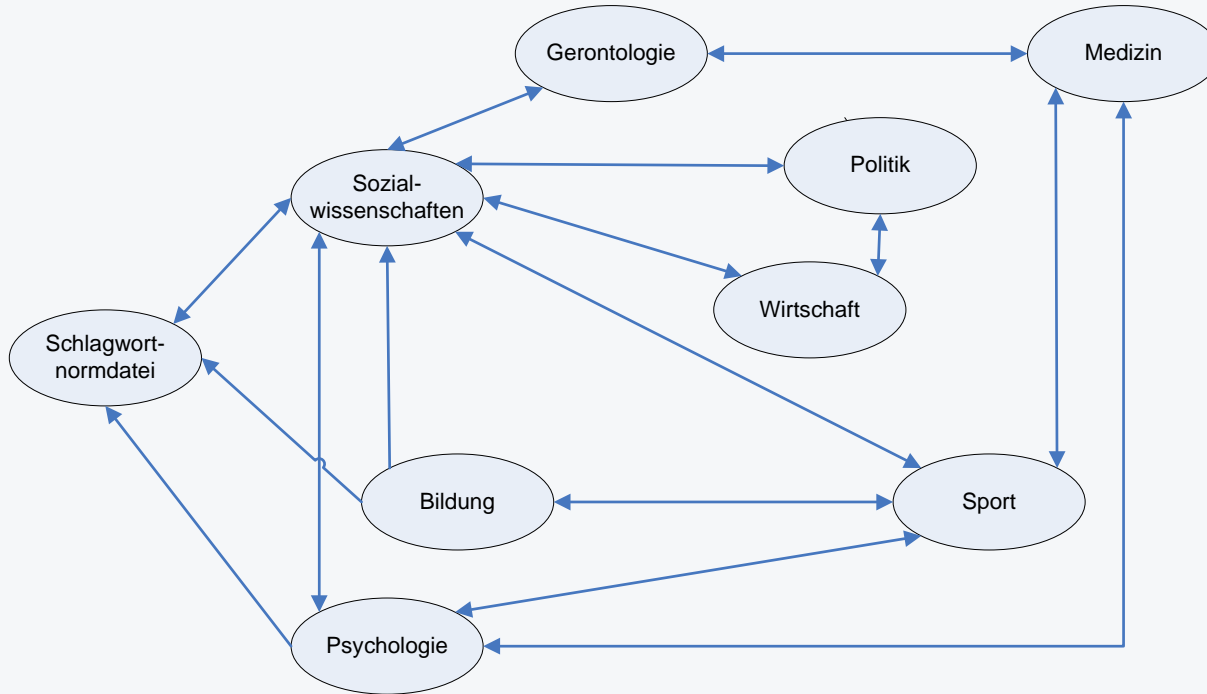
| <b>Kürzel</b> | <b>Name des Vokabulars</b>  | <b>Größe d. Vok. (ca.)</b> |
|---------------|---|----------------------------|
| Bildung       | Thesaurus Bildung   | 55.000                     |
| BISp          | Deskriptoren des Bundesinstituts für Sportwissenschaft                | 7.500                      |
| CSA-ASSIA     | CSA Thesaurus Applied Social Sciences Index and Abstracts             | 17.000                     |
| CSA-PAIS      | CSA Thesaurus PAIS International Subject Headings                     | 7.000                      |
| CSA-PEI       | CSA Thesaurus Physical Education Index                                | 1.800                      |
| CSA-SA        | Thesaurus of Sociological Indexing Terms                              | 4.000                      |
| CSA-WPSA      | CSA Thesaurus of Political Science Indexing Terms                     | 3.150                      |
| DZI           | Thesaurus des Deutschen Instituts für soziale Fragen                  | 2.000                      |
| ELSST         | European Language Social Science Thesaurus                            | 3.200                      |
| FES           | Deskriptoren der Friedrich-Ebert Stiftung                             | 4.000                      |
| GEROLIT       | Thesaurus des Deutschen Zentrums für Altersfragen                     | 2.000                      |
| IBLK          | Thesaurus Internationale Beziehungen und Länderkunde (Euro-Thesaurus) | 9.000                      |
| MeSH          | Medical Subject Headings  | 22.000                     |
| Psy           | Psyndex Terms   | 5.300                      |
| STW           | Standard Thesaurus Wirtschaft   | 5.600                      |
| SWD           | Schlagwortnormdatei   | 400.000                    |
| TheSoz        | Thesaurus Sozialwissenschaften (IZ)                                   | 7.500                      |
| TWSE          | Thesaurus für wirtschaftliche und soziale Entwicklung                 | 2.800                      |



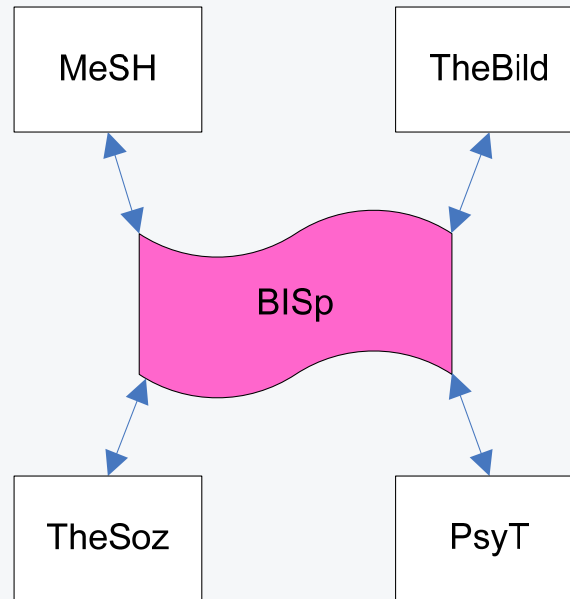
## Aktueller Stand Crosskonkordanzen (12/2006)

- 18 Vokabulare (KoMoHe und CARMEN)
- 8 Fachgebiete (Sozialwiss., Pädagogik, Psychologie, Wirtschaftswiss., Politikwiss., Sport, Medizin, Gerontologie)
- 21 bilaterale Konkordanzen (3 unilaterale)
- ca. 200,000 Relationen (Term-Term-Verbindungen)
- ca. 80,000 involvierte Deskriptoren

# Netz der Crosskonkordanzen (Fachgebiete)



# Netz der Crosskonkordanzen (Sport)



| Vok. | Vok.    | insg. | Äquiv. | OB   | UB   | VB   | Null | zg   | ZT   | Zielk. | Rel./AT |
|------|---------|-------|--------|------|------|------|------|------|------|--------|---------|
| BISp | TheSoz  | 7566  | 1978   | 1118 | 46   | 316  | 4108 | 1204 | 1744 | 2400   | 1,02    |
| BISp | Bildung | 7793  | 4417   | 1878 | 103  | 233  | 1162 | 2783 | 4098 | 4998   | 1,05    |
| BISp | Psyndex | 7624  | 1598   | 2890 | 181  | 471  | 2484 | 641  | 1728 | 2705   | 1,03    |
| BISp | MeSH    | 15083 | 2674   | 2151 | 7094 | 1006 | 2158 | 202  | 7925 | 8656   | 2,03    |

| =h   | =m  | =g | <h  | <m   | <g  | >h   | >m   | >g  | ^h  | ^m  | ^g  |
|------|-----|----|-----|------|-----|------|------|-----|-----|-----|-----|
| 1943 | 30  | 5  | 10  | 448  | 660 | 2    | 32   | 12  | 51  | 206 | 59  |
| 4320 | 88  | 9  | 10  | 972  | 896 | 0    | 91   | 12  | 57  | 118 | 58  |
| 1393 | 177 | 28 | 146 | 1978 | 766 | 5    | 140  | 36  | 96  | 309 | 65  |
| 2556 | 72  | 46 | 274 | 1311 | 566 | 4948 | 1890 | 256 | 434 | 450 | 122 |

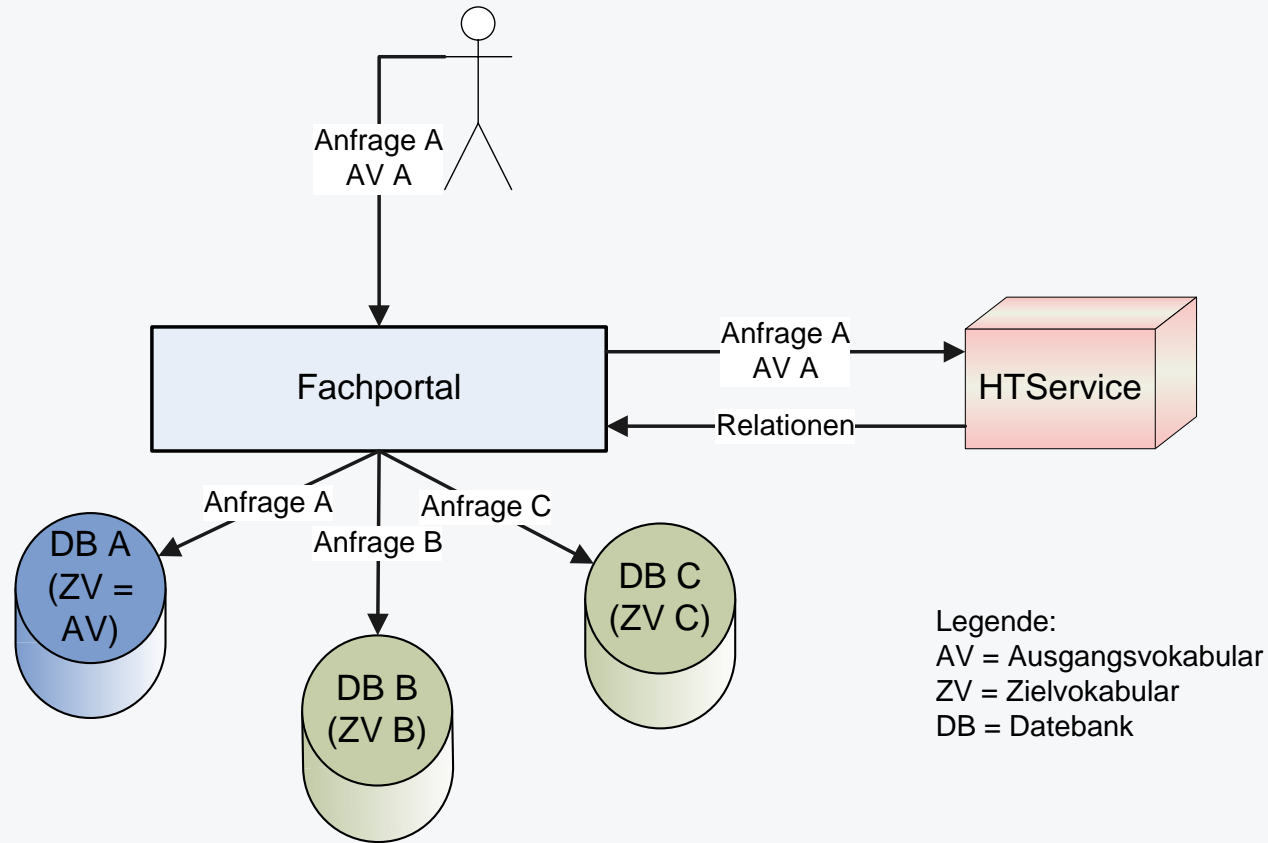
### Heterogenitätsservice:

- Web Service/SOAP
- Rückgabe von Termtransformationen für einen Anfrageterm
- Erste Testimplementierung
- Datenaustauschformat: XML

### Weitere Überlegung:

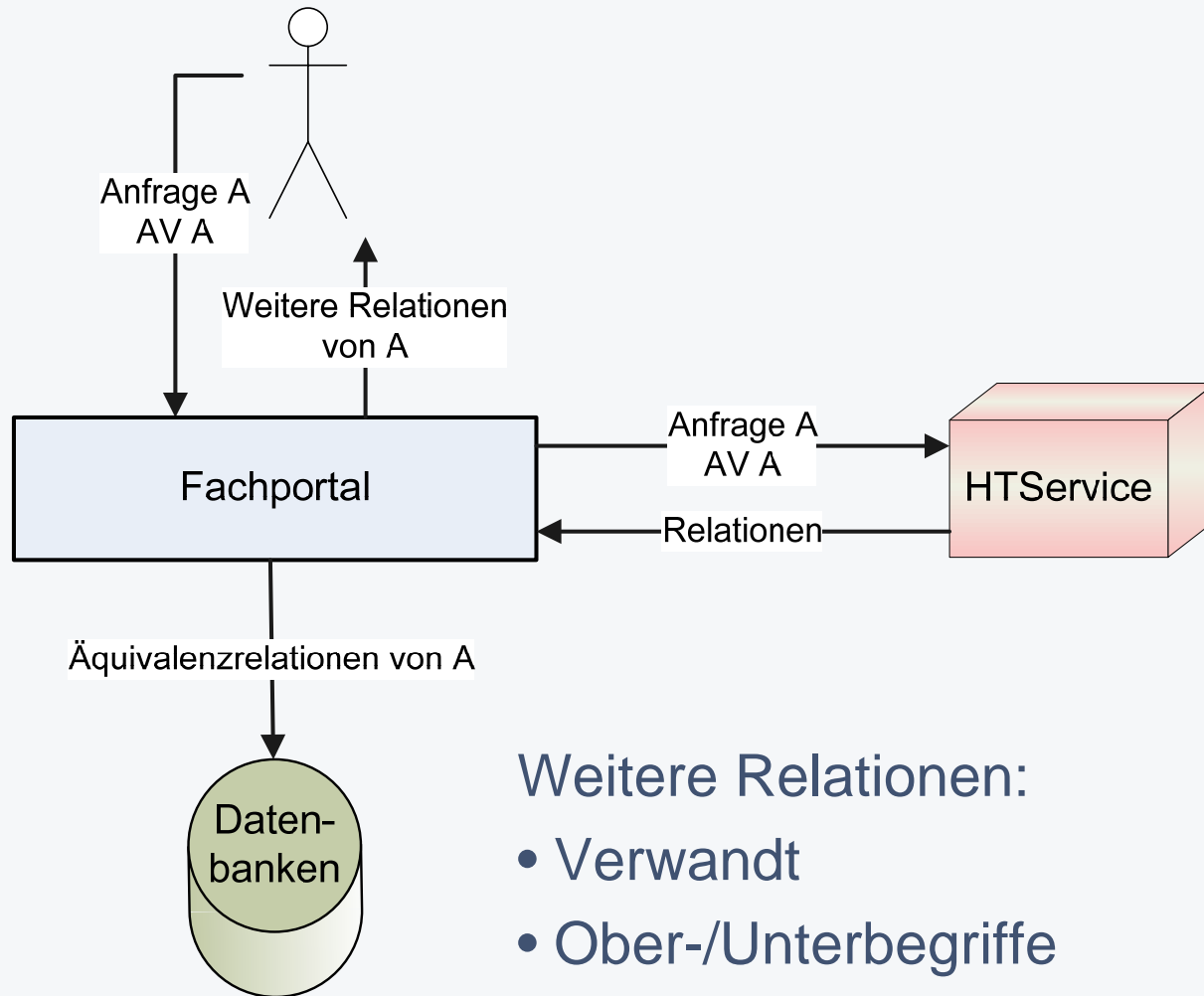
- Ggf. Verwendung von Standards für die technische Schnittstelle: SKOS (Simple Knowledge Organization System) für das Rückgabeformat

# Einsatzszenario 1: automatische Transformation der Anfrage



- Fachportal fragt HTService nach Termtransformationen
- Danach erfolgt Abfrage der Datenbanken

## Einsatzszenario 2 : Recherche-Unterstützung



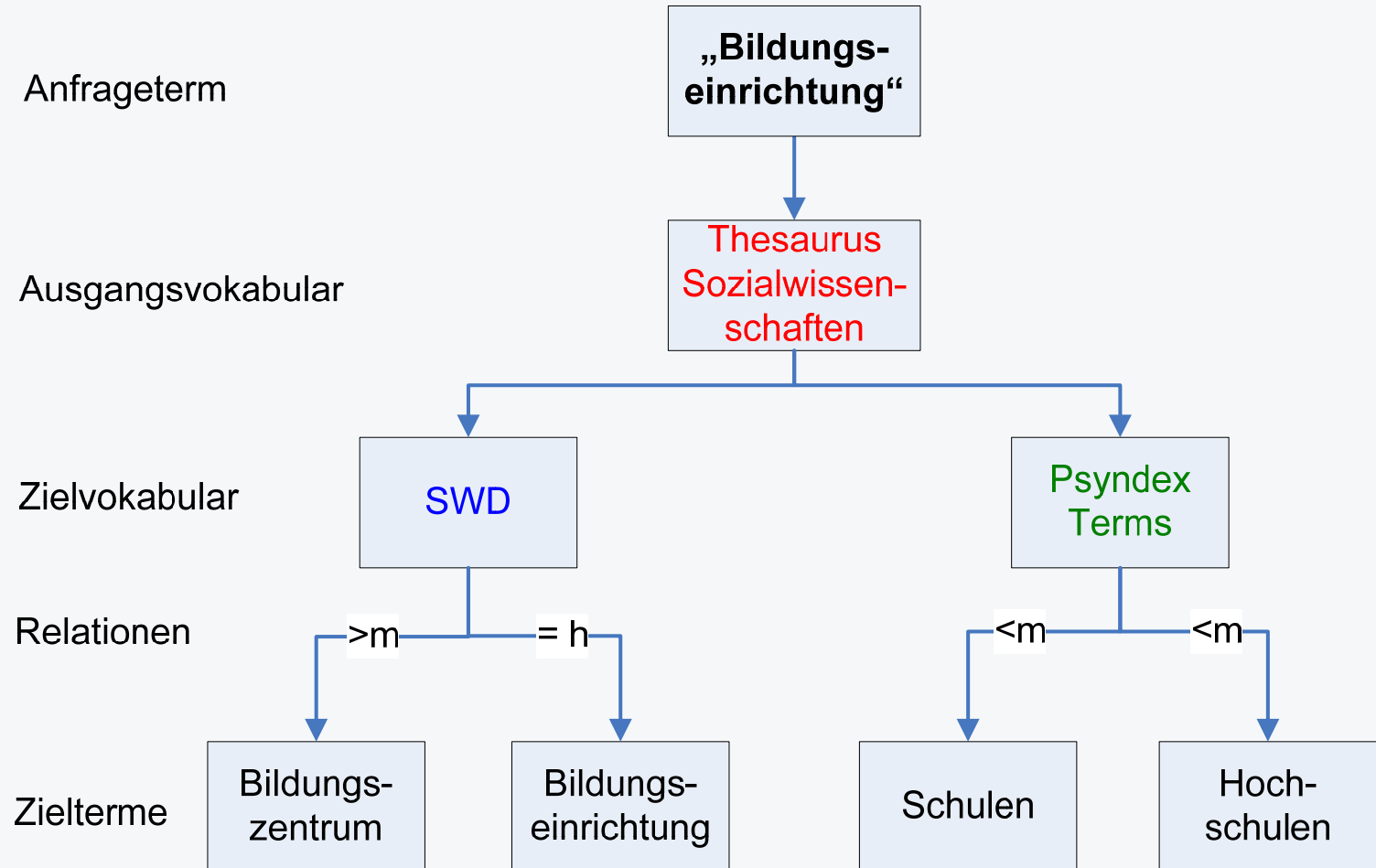
### Weitere Relationen:

- Verwandt
- Ober-/Unterbegriffe

→ Verfeinerung/Ausweitung der Recherche

# Beispiel: Antwort des Heterogenitätsservices

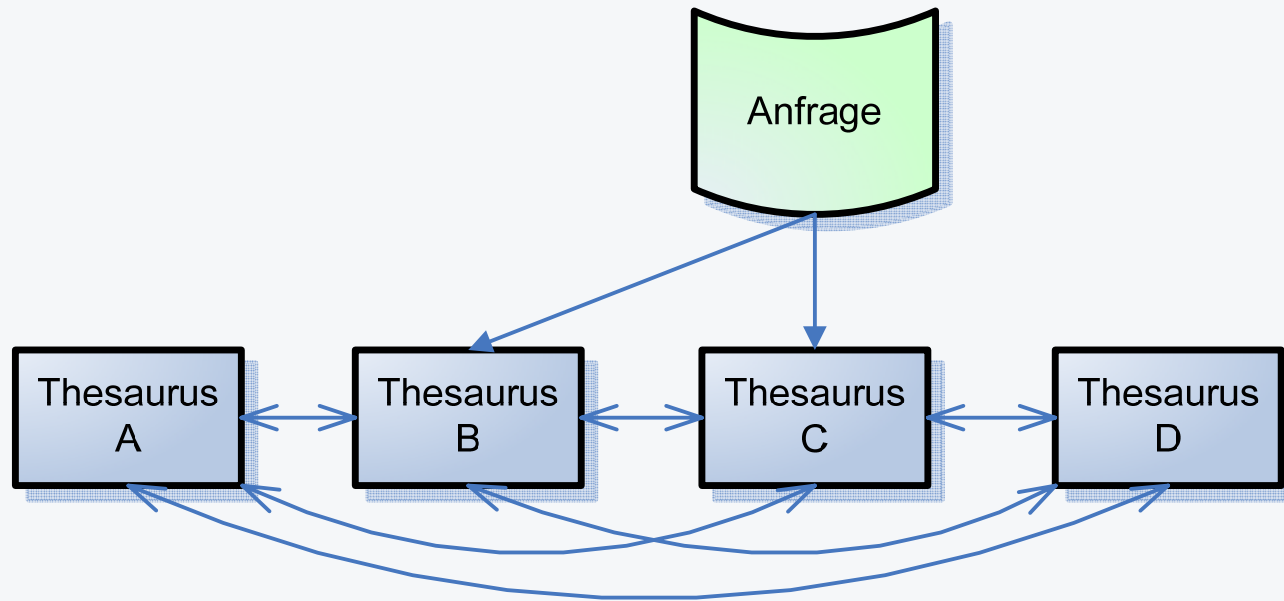
## Baumstruktur der Anfrage





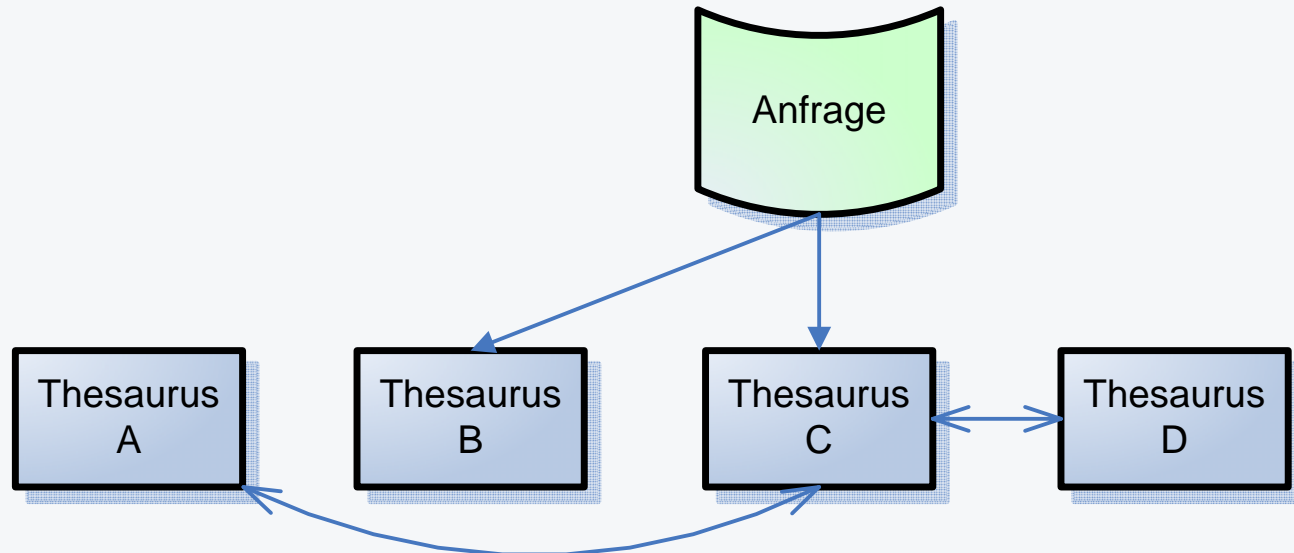
# Spezifika

# Spezifika beim Einsatz von Crosskonkordanzen



Praxis: keine vollständige Vermaschung der Vokabulare

## Spezifika beim Einsatz von Crosskonkordanzen II

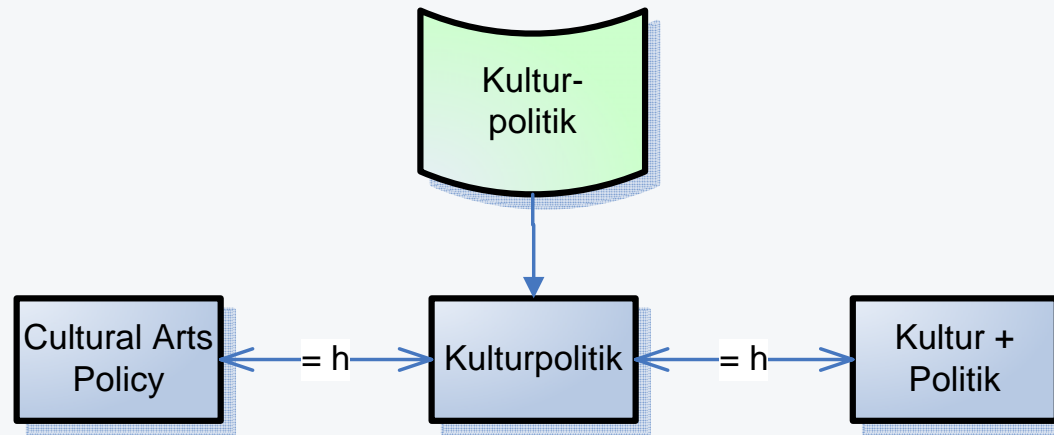


Anfrage kann in Thesaurus B und in Thesaurus C abgebildet werden

Welche Crosskonkordanzen werden angewendet?

Wahl eines Ausgangsthesaurus

## Einsatz von CK: Ausgangsthesaurus



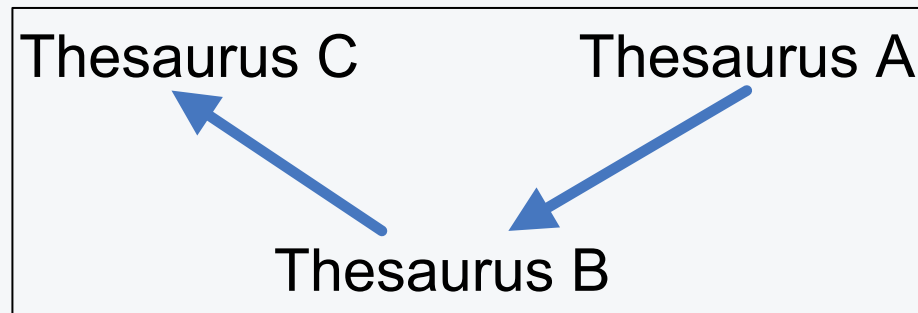
Möglichkeiten:

- Term wurde aus graphischer Oberfläche ausgewählt
- Einschränkung der Suche auf bestimmte Fächer
- Heuristiken, z.B. Trefferanzahl pro Term, Anzahl der Crosskonkordanzen des Vokabulars,...

Strategien zur Wahl des Ausgangsthesaurus sind noch zu testen!

## Erweiterter Einsatz der CK: indirekte Transformationen

Problem: Aufwand, um alle Vokabulare vollständig zu verknüpfen, ist zu hoch.



Besteht keine direkte Transformation:

Weg über ein (oder mehrere) „Switching Vocabulary“  
(weiteres Vokabular) → indirekte Transformation



# Evaluation



## Evaluation der Crosskonkordanzen

Bislang nur stichprobenartige Messungen in Vorgängerprojekten.

Fragen zur Evaluation:

- Zielgenauigkeit der Relationen
- Relevanz der durch die Crosskonkordanz zusätzlich gefundenen Treffer für das Suchbedürfnis des Nutzers?
- Auswirkungen der Fachgebiete der Thesauri auf die Zusammensetzung der Crosskonkordanz
- Auswirkungen der Struktur der Thesauri auf die Crosskonkordanz

Messungen:

- quantitativ (automatisch)
- qualitativ (intellektuelle Unterstützung)

### Ziel:

Feststellung von Mustern in der Crosskonkordanz,  
Zusammenhang mit

- Fachgebiet der beteiligten Thesauri
- Struktur der beteiligten Thesauri

### Verfahren: automatische Messungen u.a.:

- Aufteilung der Relationen auf Relationstypen
- Menge der getroffenen Deskriptoren im Zielthesaurus
- Deskriptoren pro Zielkonzept (bei Kombinationen)
- Auswertung der Thesauri



## Qualitative Evaluation

### Ziel:

Mehrwert für den Nutzer durch die zusätzlich gefundenen Dokumente

### Verfahren:

Recherche mit realen Nutzeranfragen

1. Natürlichsprachig in der Freitextsuche
2. Übersetzt in Deskriptoren in der Schlagwortsuche
3. Übersetzt in Deskriptoren in der Schlagwortsuche mit Einsatz der Crosskonkordanzen

Bewertung der Ergebnismengen bezüglich Relevanz der Treffer (analog TREC/CLEF)



## Qualitative Evaluation (Ablauf)

### Schritte:

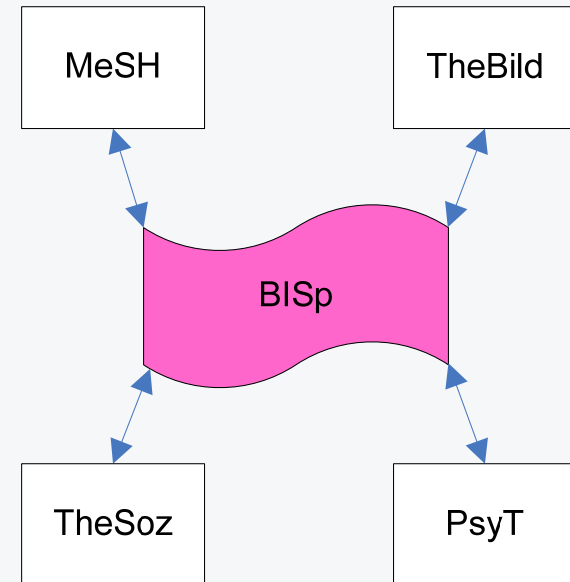
1. Lieferung realer Nutzeranfragen von den IZ- und Crosskonkordanz-Partnern (Operationalisiert)
2. Formulierung und Pretest der Suchanfragen zu den Evaluations-Szenarien
3. Suche mit den ausgewählten Suchanfragen (drei Anfragen je evaluierter Nutzeranfragen) in den entsprechenden Datenbanken und Download der Dokumente
4. Import der Dokumente in das Assessment-Tool und externe Relevanzbewertungen der Dokumente
5. Auswertung der Relevanzbewertungen

-> Ergebnisse August 2007

# Qualitative Evaluation (BISp)

Topics:

1. Neue methodische Ansätze in der Kinderleichtathletik
2. Krafttraining im Hochsprung
3. Doping im Radsport
4. Trainingsmethoden im Frauenfußball
5. Sicherheitsmaßnahmen zur Eindämmung gewaltsamer Fanausschreitungen im Fußball



# Termtransformationen (BISp) Topic 1

## Neue methodische Ansätze in der Kinderleichtathletik

### BISp

|                         |    |
|-------------------------|----|
| Leichtathletik          | <  |
| Kind                    | =  |
| Kindertraining          | =+ |
| Kinder- und Jugendsport | <+ |
| Methodik                | <  |
| Unterrichtsmethode      | =  |

### BISp

|                         |    |
|-------------------------|----|
| Leichtathletik          | =  |
| Kind                    | =  |
| Kindertraining          | =+ |
| Kinder- und Jugendsport | =+ |
| Methodik                | =  |
| Unterrichtsmethode      | =+ |

### BISp

|                    |     |
|--------------------|-----|
| Leichtathletik     | =   |
| Kind               | =   |
| Kindertraining     | ^+0 |
| Kindertraining     | ^+0 |
| Methodik           | =   |
| Unterrichtsmethode | <   |

### Psyndex

|   |
|---|
| Sport                                     |
| Kinder (Nachkommenschaft)                 |
| Kinder (Nachkommenschaft) + Sporttraining |
| Kinder (Nachkommenschaft) + Sporttraining |
| Lehrmethoden                              |
| Lehrmethoden                              |

### TheSoz

|                             |
|-----------------------------|
| Leichtathletik              |
| Kind                        |
| Kind + Training             |
| Kind + Jugendlicher + Sport |
| Methodik                    |
| Lehrmethode + Unterricht    |

### MeSH

|   |
|---|
| Track and Field                         |
| Child                                   |
| Child + Exercise                        |
| Child + Physical Education and Training |
| Methods                                 |
| Teaching                                |

# Termtransformationen (BISp) Topic 2

## Krafttraining im Hochsprung

### BISp

|                      |   |
|----------------------|---|
| Hochsprung           | < |
| Krafttraining        | = |
| Schnellkrafttraining | < |
| Maximalkrafttraining | < |
| Plyometrie           | < |

### Psyndex

|               |
|---------------|
| Springen      |
| Krafttraining |
| Krafttraining |
| Krafttraining |
| Krafttraining |

### BISp

|                      |   |
|----------------------|---|
| Hochsprung           | < |
| Krafttraining        | < |
| Schnellkrafttraining | < |

### TheSoz

|                |
|----------------|
| Leichtathletik |
| Training       |
| Training       |

### BISp

|                      |   |
|----------------------|---|
| Hochsprung           | = |
| Flop                 | = |
| Krafttraining        | = |
| Schnellkrafttraining | = |
| Maximalkrafttraining | < |
| Plyometrie           | < |

### Bildung

|                      |
|----------------------|
| Hochsprung           |
| Flop                 |
| Krafttraining        |
| Schnellkrafttraining |
| Krafttraining        |
| Sprungkrafttraining  |

### BISp

|                      |   |
|----------------------|---|
| Hochsprung           | < |
| Krafttraining        | < |
| Schnellkrafttraining | < |
| Maximalkrafttraining | < |
| Plyometrie           | < |

### MeSH

|                 |
|-----------------|
| Track and Field |
| Exercise        |
| Exercise        |
| Exercise        |
| Exercise        |

## Doping im Radsport

### BISp

|          |    |
|----------|----|
| Radsport | <  |
| Doping   | <+ |

### Psyndex

|  |
|--|
| Sport  |
| Drogen und Arzneimittel + Leistung (Fähigkeit) |

### BISp

|                  |    |
|------------------|----|
| Radsport         | =+ |
| Strassenradsport | <+ |
| Doping           | =  |

### TheSoz

|                 |
|-----------------|
| Sport + Fahrrad |
| Fahrrad + Sport |
| Droge + Sport   |

### BISp

|                  |   |
|------------------|---|
| Radsport         | = |
| Strassenradsport | < |
| Doping           | = |

### Bildung

|          |
|----------|
| Radsport |
| Radsport |
| Doping   |

### BISp

|                  |    |
|------------------|----|
| Radsport         | ^+ |
| Strassenradsport | <  |
| Doping           | =  |

### MeSH

|                    |
|--------------------|
| Bicycling + Sports |
| Sports             |
| Doping in Sports   |

- Weitere Crosskonkordanzen geplant
  - Technik
  - Agrovoc
  - Klassifikationen
- Einsatz statistischer Verfahren
  - MeSH-SWD
- Einsatz des Heterogenitätsservice in sowiport, vascoda, ...
- Heterogenitätsservice soll direkte und indirekte Term-Transformationen ermöglichen
- Anfrageabarbeitung an Benutzerschnittstelle (V1 Behandlung durch Search Term Recommender)
- Qualitative Evaluierung der Termtransformationen

# Vielen Dank für die Aufmerksamkeit!

Weiterführende Informationen zum Projekt unter  
<http://www.gesis.org/Forschung/Informationstechnologie/komohe.htm>

**Philipp Mayr**  
**Anne-Kathrin Walter**

Informationszentrum Sozialwissenschaften (IZ)  
Abt. Forschung und Entwicklung  
Lennéstr. 30  
53113 Bonn  
Tel. 0228 / 22 81 - 0  
email {mayr,walter}@bonn.iz-soz.de  
<http://www.gesis.org/IZ>