Usage and impact of controlled vocabularies in a subject repository for indexing and retrieval

Dr. Timo Borst

LIBER 2011
Barcelona
29.6.-2.7.2011

Overview

1. Terminology webservices as a means for supporting retrieval in the realm of library applications

2. Logfile analysis as an approach for analysing users‘ search behaviour

3. Results

4. Conclusions and suggestions for improving search interfaces

Terminology webservices

General idea: “Provide a framework for integrating authority data, which is both normative and flexible enough to tolerate local idiosyncrasies on a string level.”

Approach: Concept modelling based on Semantic Web / SKOS standards (for concepts, persons, institutions,…)

uri	surname	forename	variantNames	academicTitle	lifeData	affiliations
http://d-nb.info/gnd/124825109	"Snower"	"Dennis J."	Snower, Dennis James Snower, D. J. Snower, Dennis Snower, D.	Prof. Ph.D.	1950-	http://d-nb.info/gnd/1007681-5

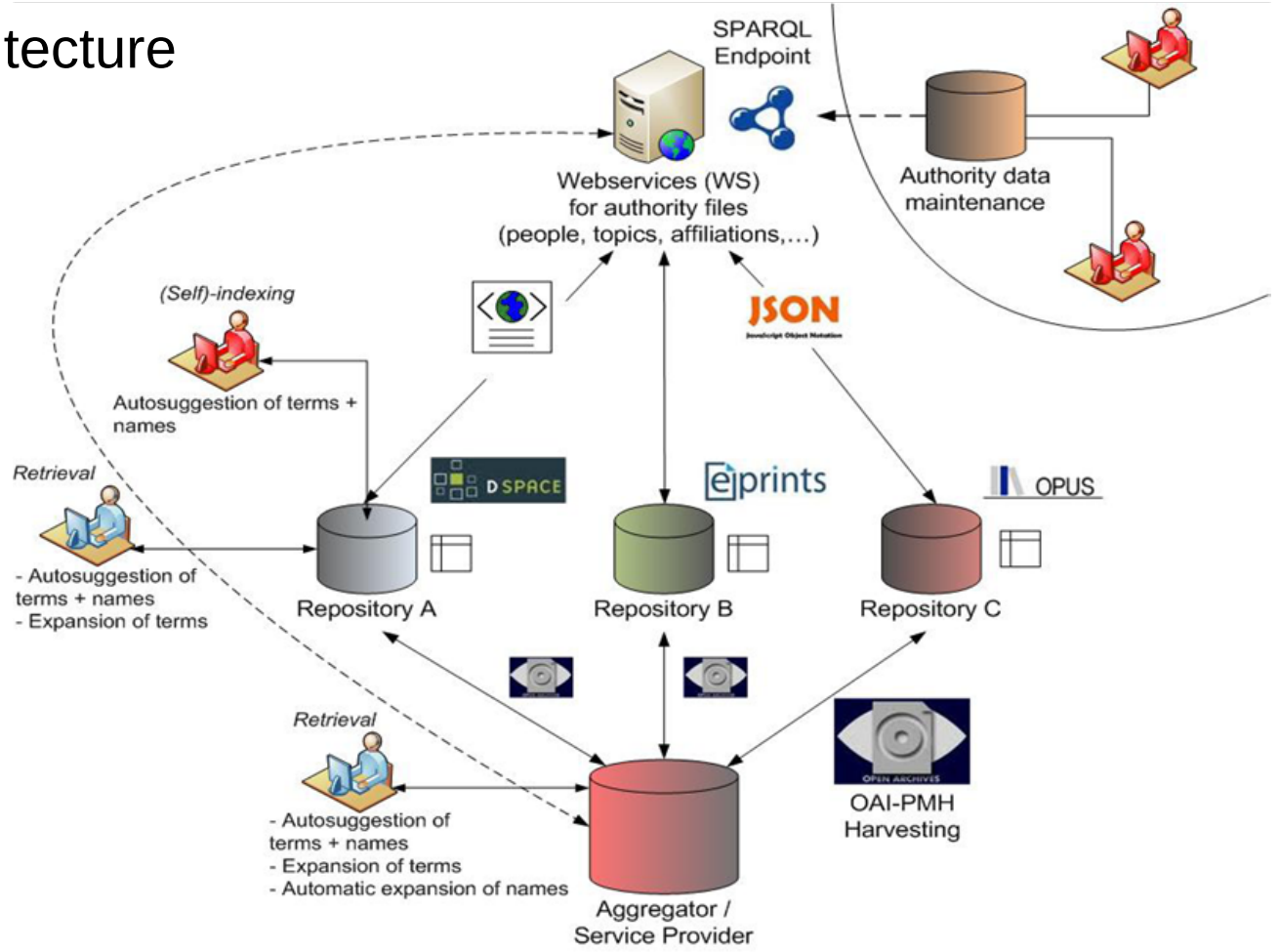
Repräsentation

uri	lang_de	lang_en	lang_fr	broaderTerms	narrowerTerms	altLabel	closeMatch
http://zbw.eu/stw/descriptor/19664-4	"Finanzmarktkrise"	"Financial crisis"	"crise financière"	<stw/descriptor/10343-6> <stw/thsys/70187> <stw/thsys/71089>	<stw/descriptor/13688-6> <stw/descriptor/18730-1>	"Financial instability"@en "Finanzkrise"@de "Krise der FinanzmÄrkte"@de	http://dbpedia.org/resource/Financial_crisis

Repräsentation

Terminology webservices

Architecture



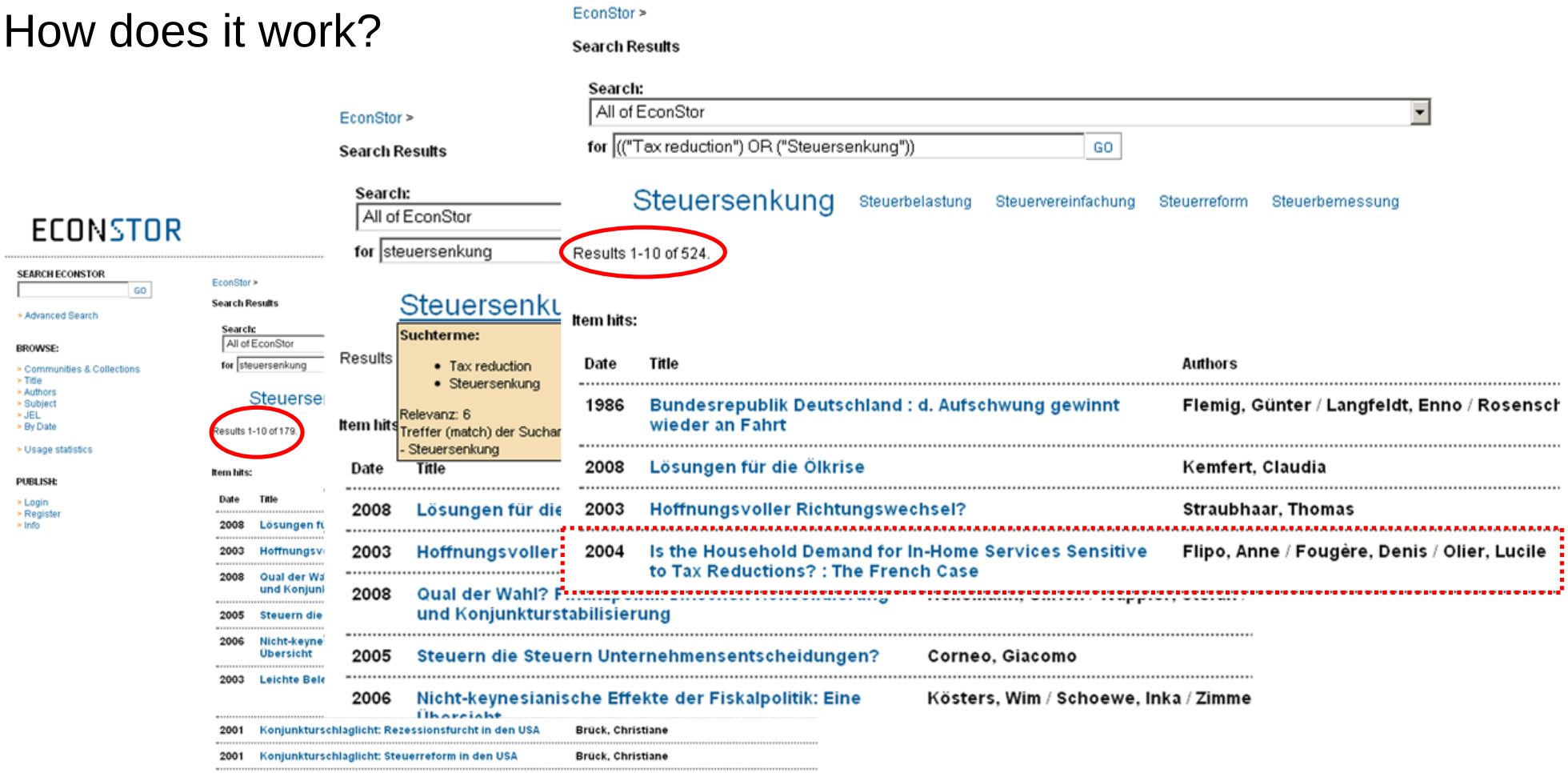
Terminology webservices

Terminology?

- „STW Thesaurus for Economics“, http://zbw.eu/stw/versions/latest/about.en.html
- More than 6,000 standardized subject headings and 18,000 entry terms
- Contains concepts from Economics and Business Research, but also from law, sociology and politics
- Part of the Semantic Web and the LOD cloud
- Integrated into our own retrieval applications, downloaded from many institutions

Terminology webservices

How does it work?



Logfile analysis

- Many approaches to analysis of user behaviour (logfile analysis, real-time tracking, usability studies, questionnaires…)
- To us, logfiles serve as a basis for analysing string patterns in queries, hence search behaviour on a linguistic level
- Basic idea: each user request is logged in a standardized way, e.g. by a web server

- Query strings are automatically processed and analysed e.g. through scripts (PERL), regexp or UNIX Shell commands (grep, sed, awk,…)

Logfile analysis

Pros

+ Automatic generation and persistence of logfiles
+ Can be processed at any time by different tools
+ Filtering of robots, crawlers etc. possible

Cons

- Access through proxies and browser caches -> no user identification and counting possible
- Sometimes restricted to data privacy rules (e.g., no IP tracking allowed)
- No real-time processing

Logfile analysis

To be investigated:

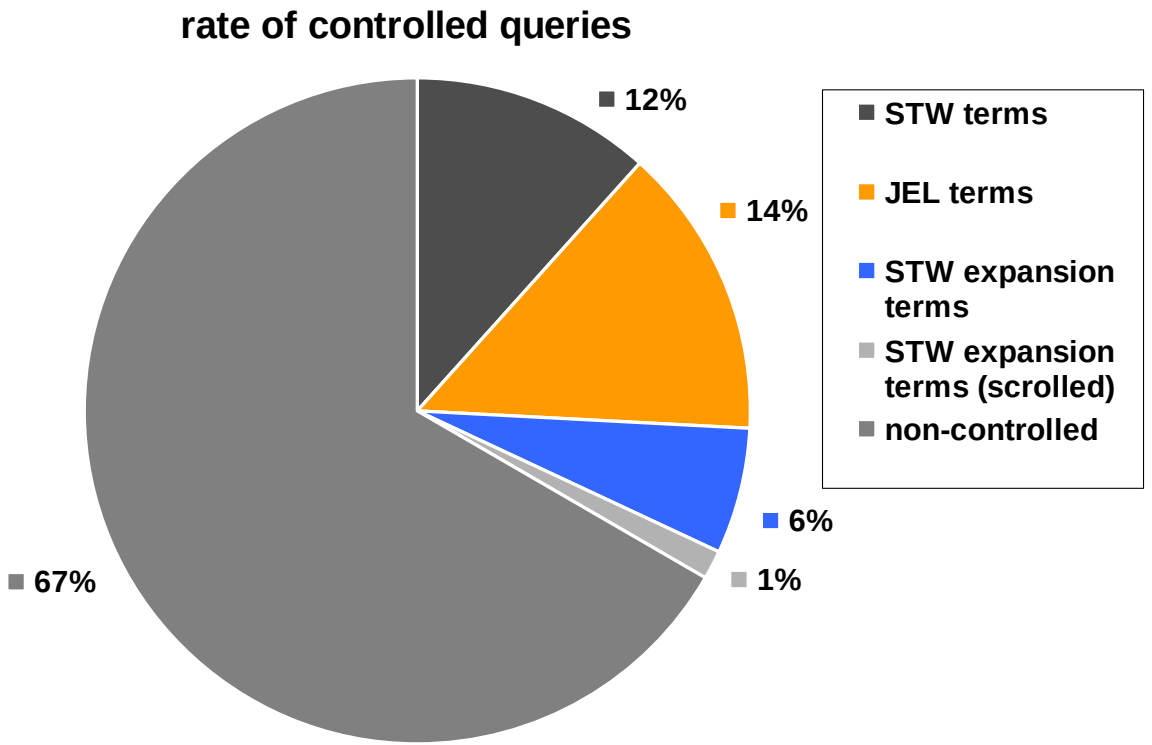
1. What is the *current rate* of search queries with controlled vocabulary?

2. What is the *potential mapping* of uncontrolled search terms to controlled vocabulary?

3. How does the use of controlled vocabulary affect *document views*?

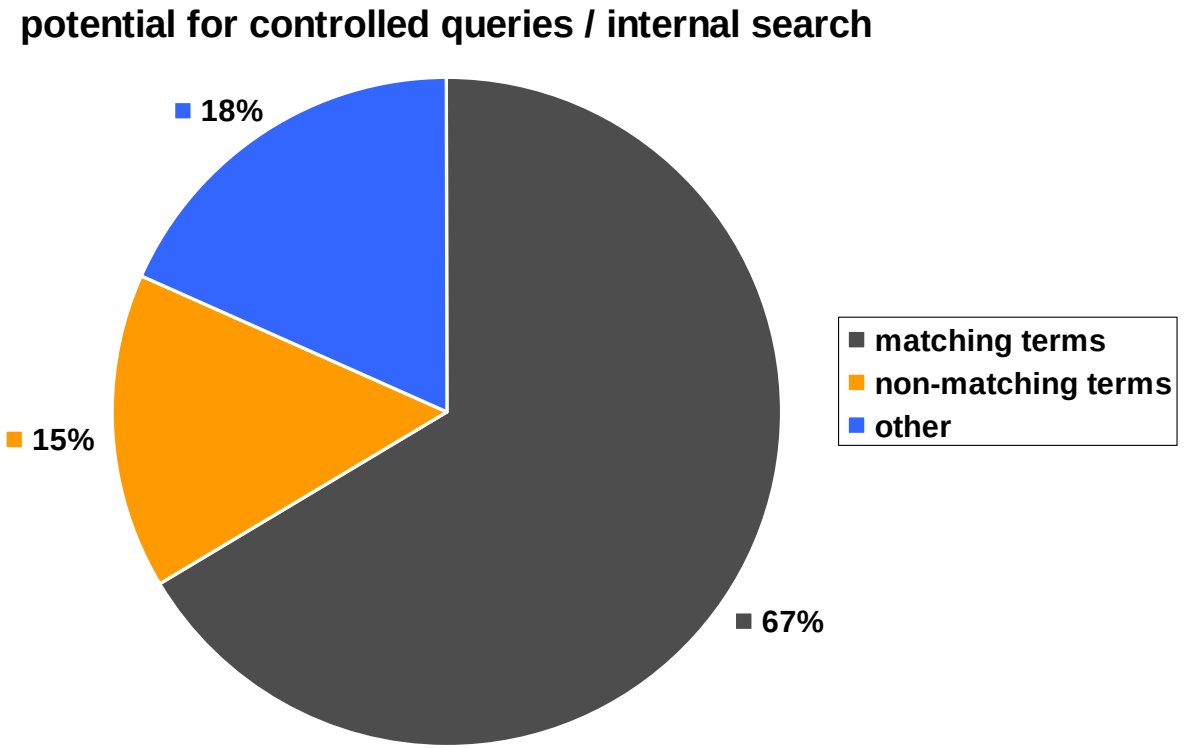
Results

What is the *current rate* of search queries with controlled vocabulary (JEL, STW terms by autosuggest and search term expansion with/without scrolling)?



Results

What is the *potential mapping* of uncontrolled search terms to controlled vocabulary (internal search)?*

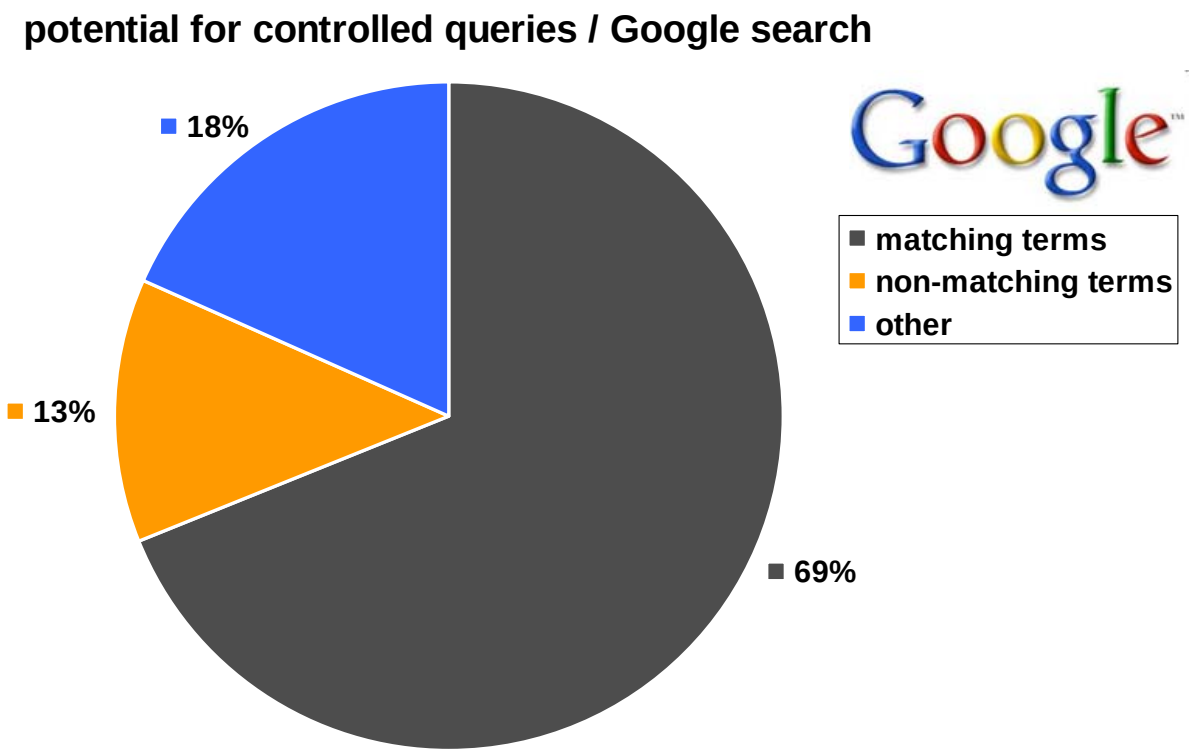


*approach:

Running search terms against Lucene/SOLR-index of STW terms with stemming

Results

How does the use of controlled vocabulary affect document views (Google search)?*



*approach:

Running search terms against Lucene/SOLR-index of STW terms with stemming

Conclusions and suggestions for improving search interfaces (I)

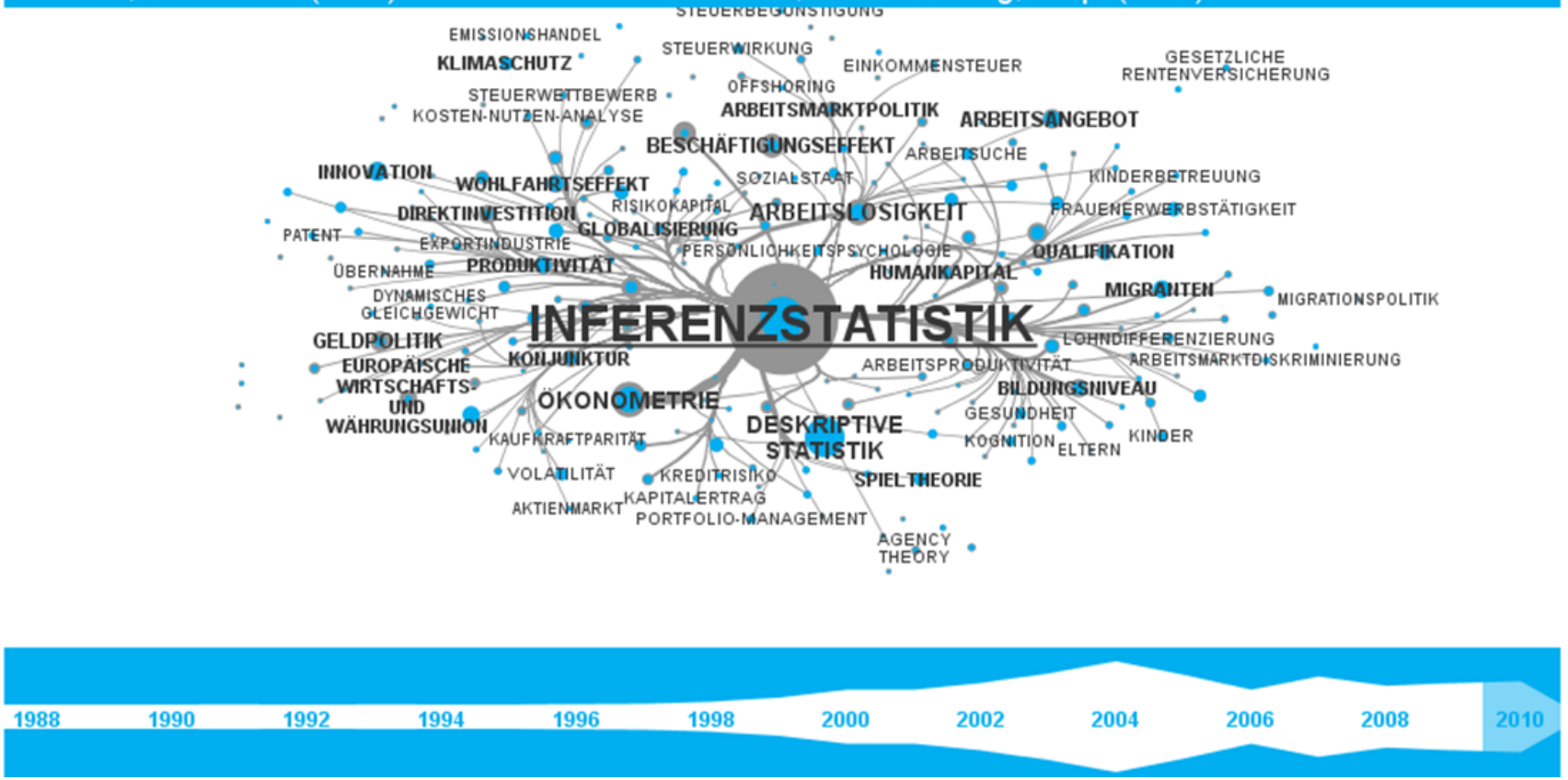
- Significant use of and potential for controlled vocabulary – if the vocabulary is big enough and constantly maintained
- Significant rate of uncontrolled terms belonging to other-categories like „names“ and „document titles“ – how to support this better?
  - *Different searches for names according to different roles (e.g. search for (co-)authors, in citations, author information etc.)*
  - *Suggesting names by authority files*
- Result sets resulting from search term expansion are scrolled quite often – how to avoid this?
  - *Adding filters*
  - *Sorting by column*
  - *Cascading search*

Conclusions and suggestions for improving search interfaces (II)

- Mapping of uncontrolled terms to vocabulary still may be further improved by linguistic techniques – main goal: convergence between „information system‘s language“ and user language
- Uncontrolled internal search (in our repository) and Google search formally do not differ much - what does that mean?
  - Statement: Adaptation to Google text based search is not appropriate for domain specific scientific search. Instead, we do need
    - *suggest services based on authority data for terms, names, institutions etc. to better anticipate domain specific queries*
    - *visible real-time information about other users‘ search behaviour (community building)*
    - *visualization and navigation of domain specific topics*

rs in regional business cycles
Kholodilin, Konstantin (2009)

Overeducation, Wages and Promotions within the
Groeneveld, Sandra Hartog, Joop (2003)



Thank you!

Questions?

Dr. Timo Borst

t.borst@zbw.eu