

# 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 webservice

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,...)

<u>uri</u>	<u>sumame</u>	<u>forename</u>	<u>variantNames</u>	<u>academicTitle</u>	<u>lifeData</u>	<u>affiliations</u>
<a href="http://g-nb.info/gnd/124825109">http://g-nb.info/gnd/124825109</a>	"Snower"	"Dennis J."	Snower, Dennis James Snower, D. J. Snower, Dennis Snower, D.	Prof. Ph.D.	1950-	<a href="http://g-nb.info/gnd/1007681-5">http://g-nb.info/gnd/1007681-5</a>

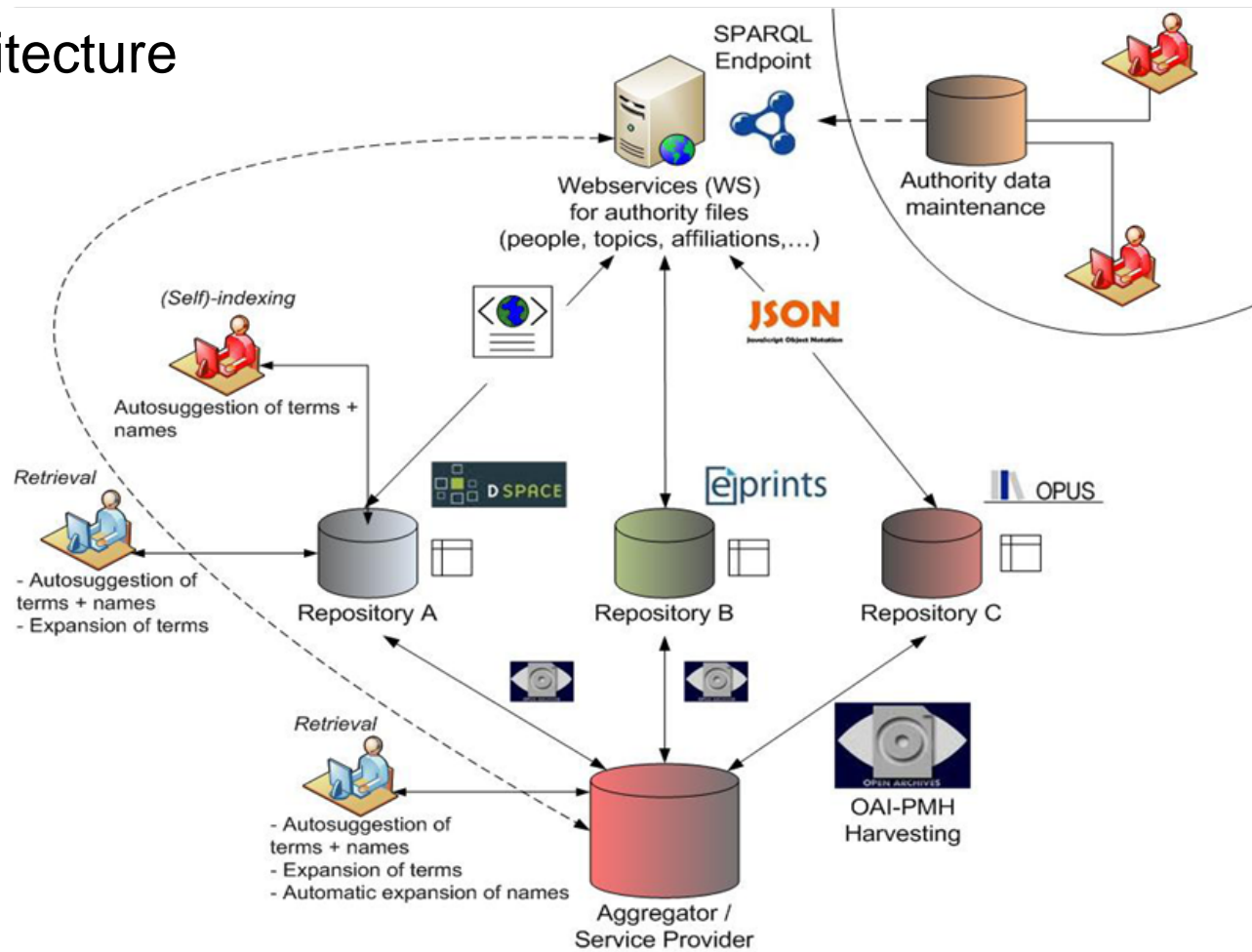
Repräsentation

<u>uri</u>	<u>lang_de</u>	<u>lang_en</u>	<u>lang_fr</u>	<u>broaderTerms</u>	<u>narrowerTerms</u>	<u>altLabel</u>	<u>closeMatch</u>
<a href="http://zbw.eu/stw/descriptor/19664-4">http://zbw.eu/stw/descriptor/19664-4</a>	"Finanzmar- tkrise"	"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 "Finanzkris- e"@de "Krise der Finanzmär- kte"@de	<a href="http://dbpedia.org/resource/Financial_crisis">http://dbpedia.org /resource/Financi- al_crisis</a>

Repräsentation

# Terminology webservice

## 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 webservice

How does it work?

The screenshot shows the EconStor search interface. The search term 'Steuersenkung' is entered, and the results are displayed in a table format. A red circle highlights the search results count 'Results 1-10 of 524.' and another red circle highlights 'Results 1-10 of 179.' in the left sidebar. A red dashed box highlights a specific search result entry.

**Search Results**

Search: All of EconStor  
for: (("Tax reduction") OR ("Steuersenkung"))

Steuersenkung Steuerbelastung Steuervereinfachung Steuerreform Steuerbemessung

Results 1-10 of 524.

Suchterme:  
• Tax reduction  
• Steuersenkung

Relevanz: 6  
Treffer (match) der Suchanfrage: Steuersenkung

Date	Title	Authors
1986	Bundesrepublik Deutschland : d. Aufschwung gewinnt wieder an Fahrt	Flemig, Günter / Langfeldt, Enno / Rosensch
2008	Lösungen für die Ölkrise	Kemfert, Claudia
2003	Hoffnungsvoller Richtungswechsel?	Straubhaar, Thomas
2004	Is the Household Demand for In-Home Services Sensitive to Tax Reductions? : The French Case	Flipo, Anne / Fougère, Denis / Olier, Lucile
2008	Qual der Wahl? Fiskalpolitik und Konjunktur	
2005	Steuern die Steuern Unternehmensentscheidungen?	Corneo, Giacomo
2006	Nicht-keynesianische Effekte der Fiskalpolitik: Eine Übersicht	Kösters, Wim / Schoewe, Inka / Zimme
2001	Konjunkturschlaglicht: Rezessionsfurcht in den USA	Bruck, Christiane
2001	Konjunkturschlaglicht: Steuerreform in den USA	Bruck, Christiane

# 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

```
16/Nov/2010:16:54:02 +0100] "GET /dspace/simple-search/query=innovationsverhalten+2003 HTTP/1.1" 200 40465 "http://econstor.eu/dspace/simp.
16/Nov/2010:16:54:03 +0100] "GET /dspace/simple-search?location=%2F&query=%28%28%22Environmental+degradation%22%29+OR+%28%22Umweltbelast
16/Nov/2010:16:54:04 +0100] "GET /dspace/simple-search?query=%28%28%22Environmental+degradation%22%29+OR+%28%22Umweltbelastung%22%29+OR-
16/Nov/2010:16:54:32 +0100] "GET /dspace/simple-search?location=%2F&query=innovationsverhalten+rammer HTTP/1.1" 302 - "http://econstor.eu/c
16/Nov/2010:16:54:32 +0100] "GET /dspace/simple-search?query=innovationsverhalten+rammer HTTP/1.1" 200 40497 "http://econstor.eu/dspace/sir
16/Nov/2010:16:55:45 +0100] "GET /dspace/simple-search?query=%28%28%22Environmental+degradation%22%29+OR+%28%22Umweltbelastung%22%29+OR-
16/Nov/2010:17:03:54 +0100] "GET /dspace/simple-search?query=%28%28jel%3AL20%29%29&start=50 HTTP/1.0" 200 39479 "-" "Mozilla/5.0 (compat:
16/Nov/2010:17:07:31 +0100] "GET /dspace/simple-search?query=coupon HTTP/1.1" 200 39258 "-" "Mozilla/5.0 (Windows; U; Windows NT 6.1; de;
16/Nov/2010:17:09:04 +0100] "GET /dspace/simple-search?location=%2F&query=coupon+psychologie HTTP/1.1" 302 - "http://www.econstor.eu/dsp
16/Nov/2010:17:09:04 +0100] "GET /dspace/simple-search?query=coupon+psychologie HTTP/1.1" 200 39495 "http://www.econstor.eu/dspace/simple
16/Nov/2010:17:10:45 +0100] "GET /dspace/simple-search?query=kapitalstock&submit=LOS HTTP/1.1" 200 39573 "http://www.econstor.eu/dspace/k
16/Nov/2010:17:11:06 +0100] "GET /dspace/simple-search?query=kapitalstock&start=10 HTTP/1.1" 200 39818 "http://www.econstor.eu/dspace/sir
16/Nov/2010:17:11:16 +0100] "GET /dspace/simple-search?query=kapitalstock&start=20 HTTP/1.1" 200 39882 "http://www.econstor.eu/dspace/sir
```

- 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

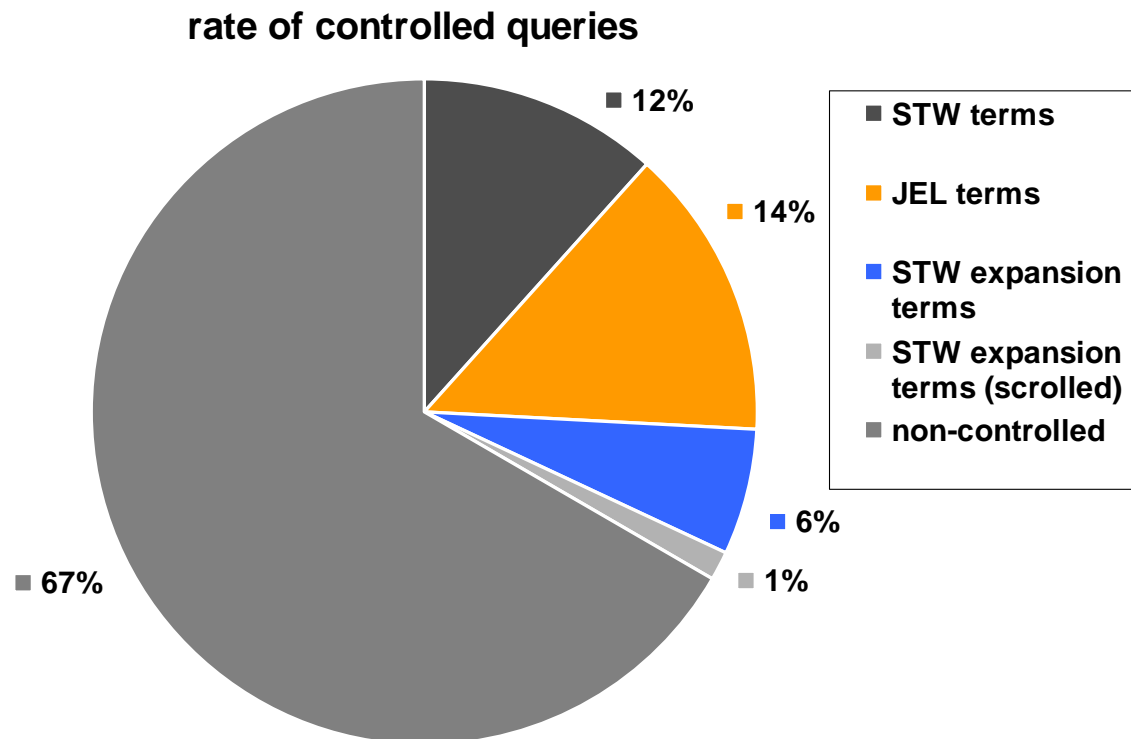
---

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)?

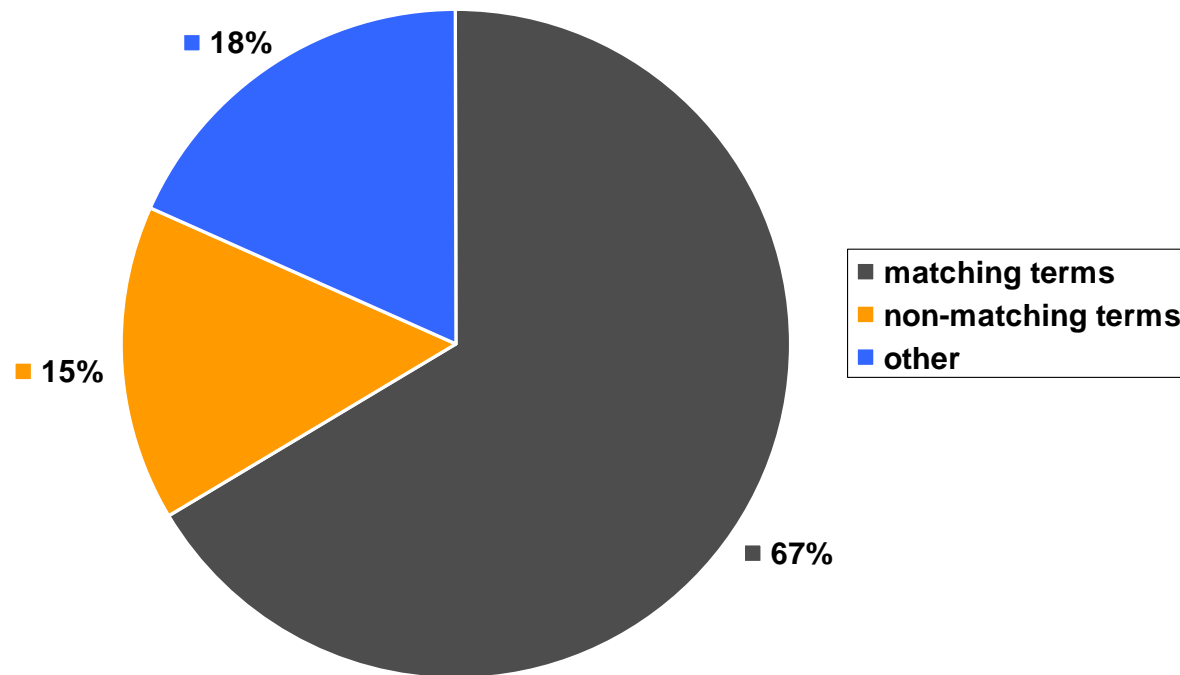


# Results

---

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

potential for controlled queries / internal search



\*approach:

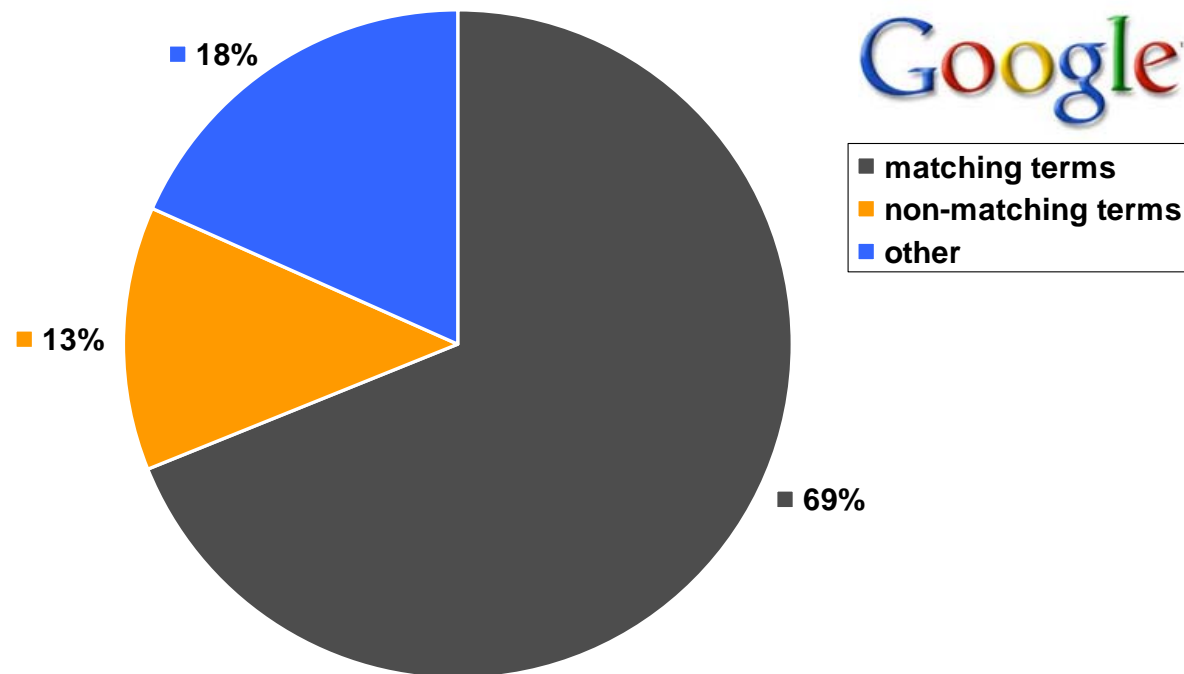
Running search terms against Lucene/SOLR-index of STW terms with stemming

# Results

---

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

potential for controlled queries / 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)

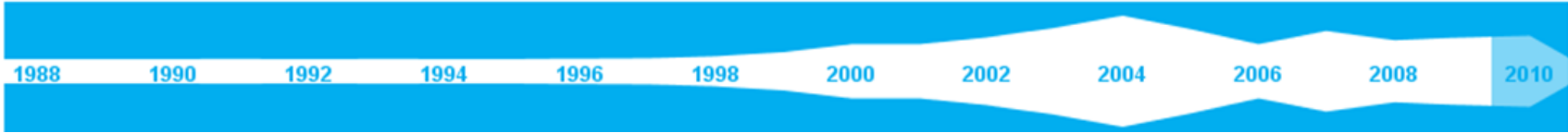
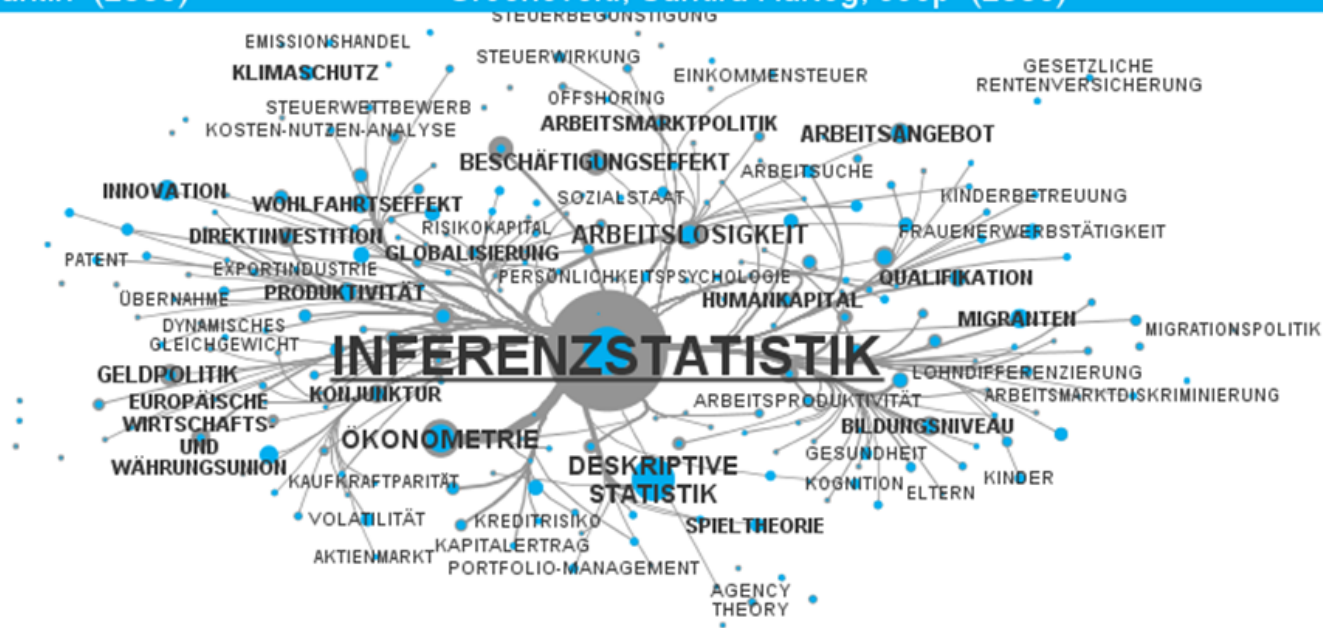
---

- 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*
-



...s 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](mailto:t.borst@zbw.eu)