gesis
Leibniz Institute for the Social Sciences

Creation of custom KOS-based recommendation systems

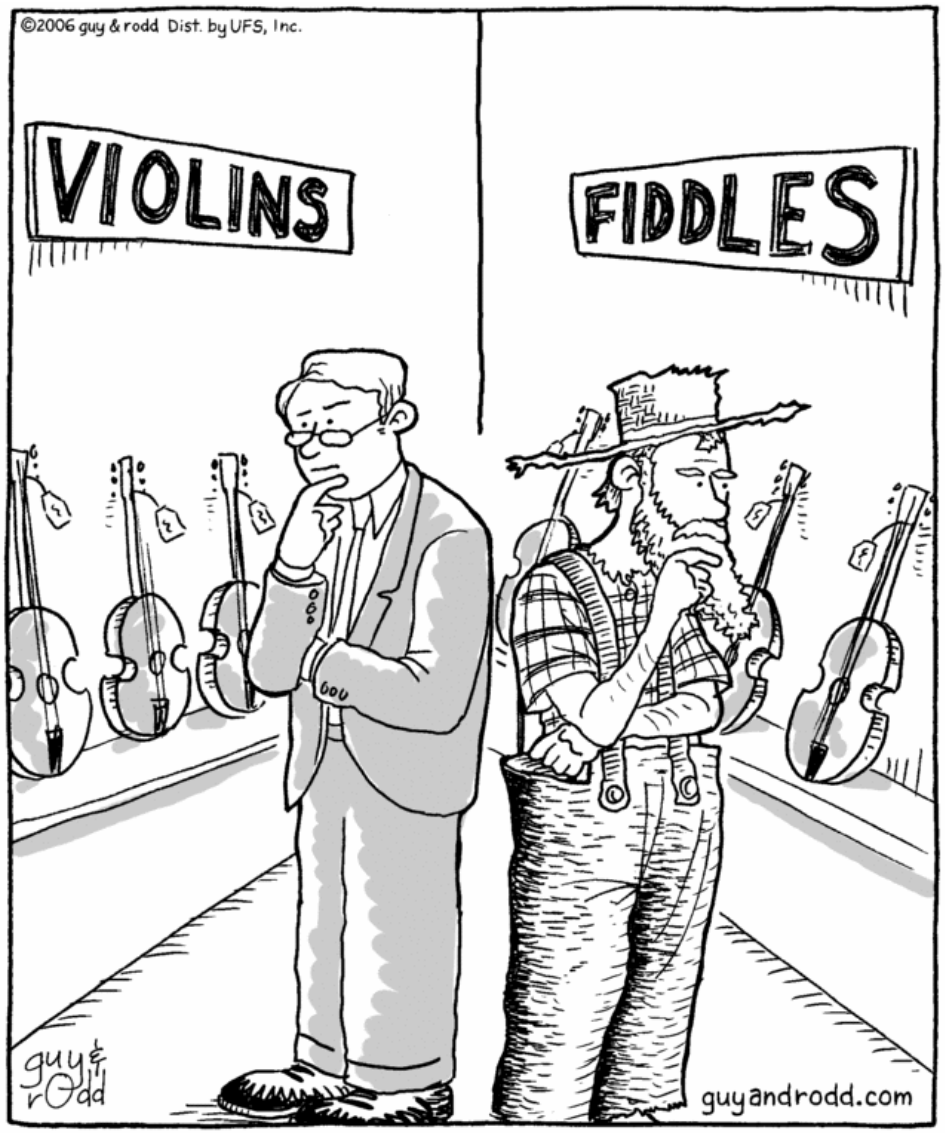
Thomas Lüke, Wilko van Hoek,
Philipp Schaer, Philipp Mayr

NKOS-Workshop @ TPDL 2012
Paphos, Cyprus, 2012-09-27

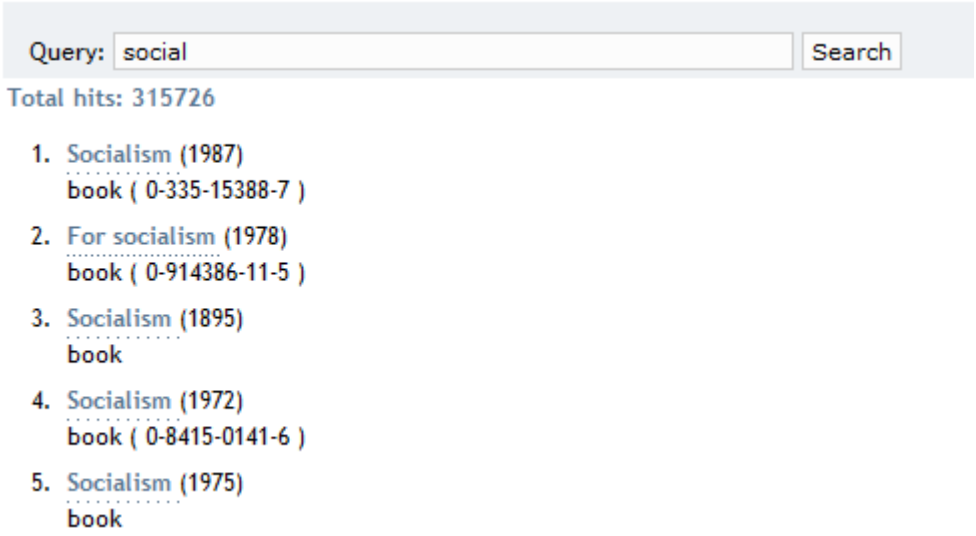
Thomas.Lueke@gesis.org

Overview

1. Motivation: Finding the matching terms in IR
2. Use Cases for recommendation systems
3. Creating custom recommenders
   - Workflow
   - Interface
4. Demonstration
5. Conclusion


©2006 guy & rodd Dist. by UFS, Inc.
VIOLINS
FIDDLES
guyandrodd.com

Motivation

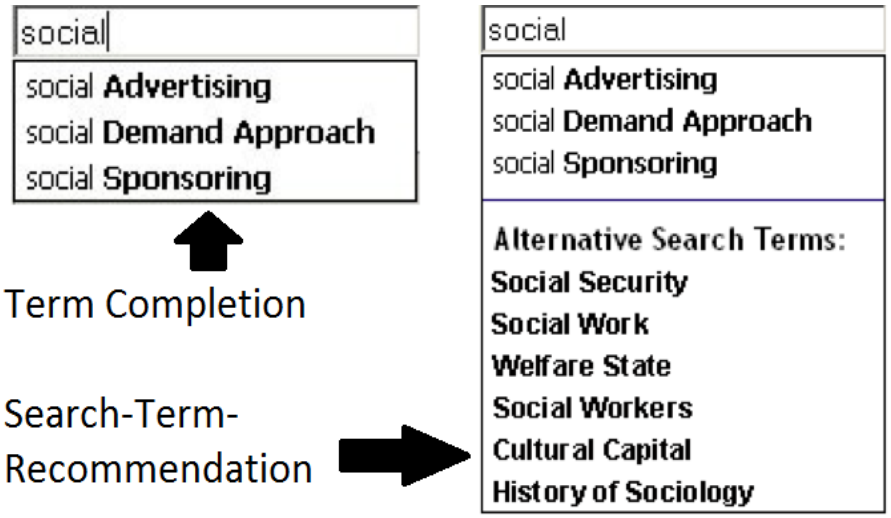


- Databases are vastly growing
  - empty result sets are rare
  - too unspecific results are a problem
- Users need to refine their search

Motivation

see Hienert et al., 2011

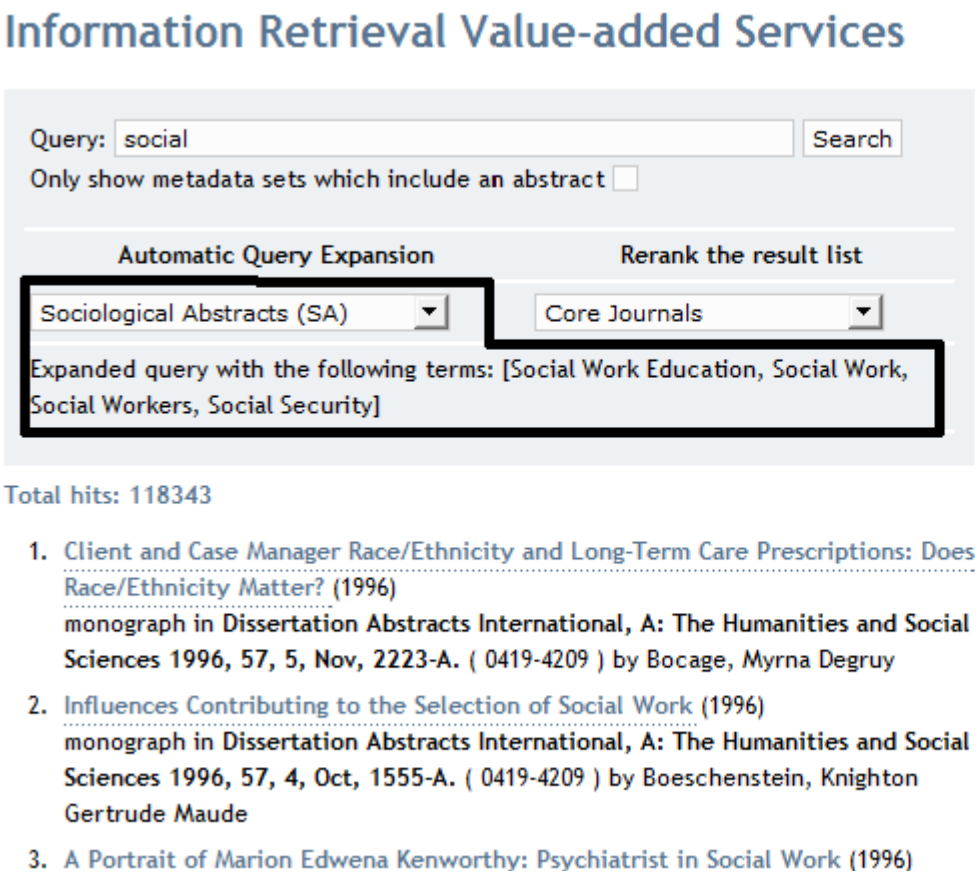
Use-Case 1: Manual Query Expansion



Standard Search Term Recommender (STR)

- Maps any query term onto controlled Thesaurus-concepts
- Trained with many different databases and vocabularies (SOLIS, CSA-SA, SPOLIT, FIS Bildung, …)
- Real Life usage: Portal Sowiport (cf. TPDL 2011: Hienert et al.)

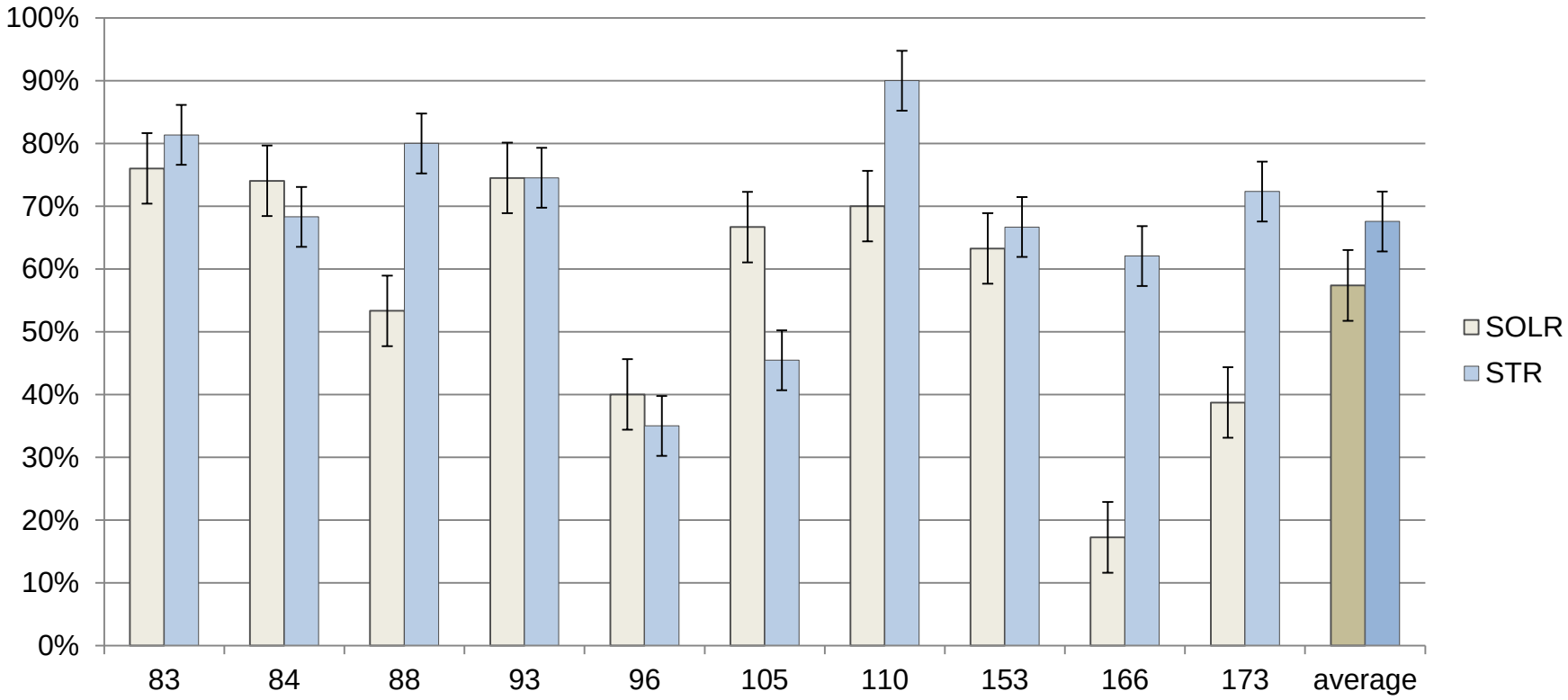
Use-Case 2: Automatic Query Expansion



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

See NKOS presentation Mayr et al., 2010

Use-Case Evaluation



Result: On Average the usage of an STR can improve the search process

See (Mutschke et. al: Science models as value-added services for scholarly information systems. Scientometrics. 89, 349–364 (2011).

Creating custom recommenders

- Recommender Service in IRM I was based on commercial software
- Goals in IRM II:
  - Replacing old technology with new self-written version
  - Making technology available to others by being open-source
  - Provide Web-Interfaces to use recommenders services
  - Allow the creation of custom recommenders on our servers

- Why Custom STRs?
  - The more specific the dataset, the more specific the recommendations
  - Customized for your specific information need
    → see our Poster/Paper
    (Improving Retrieval Results with Discipline-specific Query Expansion, TPDL 2012, Lüke et. Al, http://arxiv.org/abs/1206.2126)

OAI-PMH Dublin Core Data

Free Terms in Title →

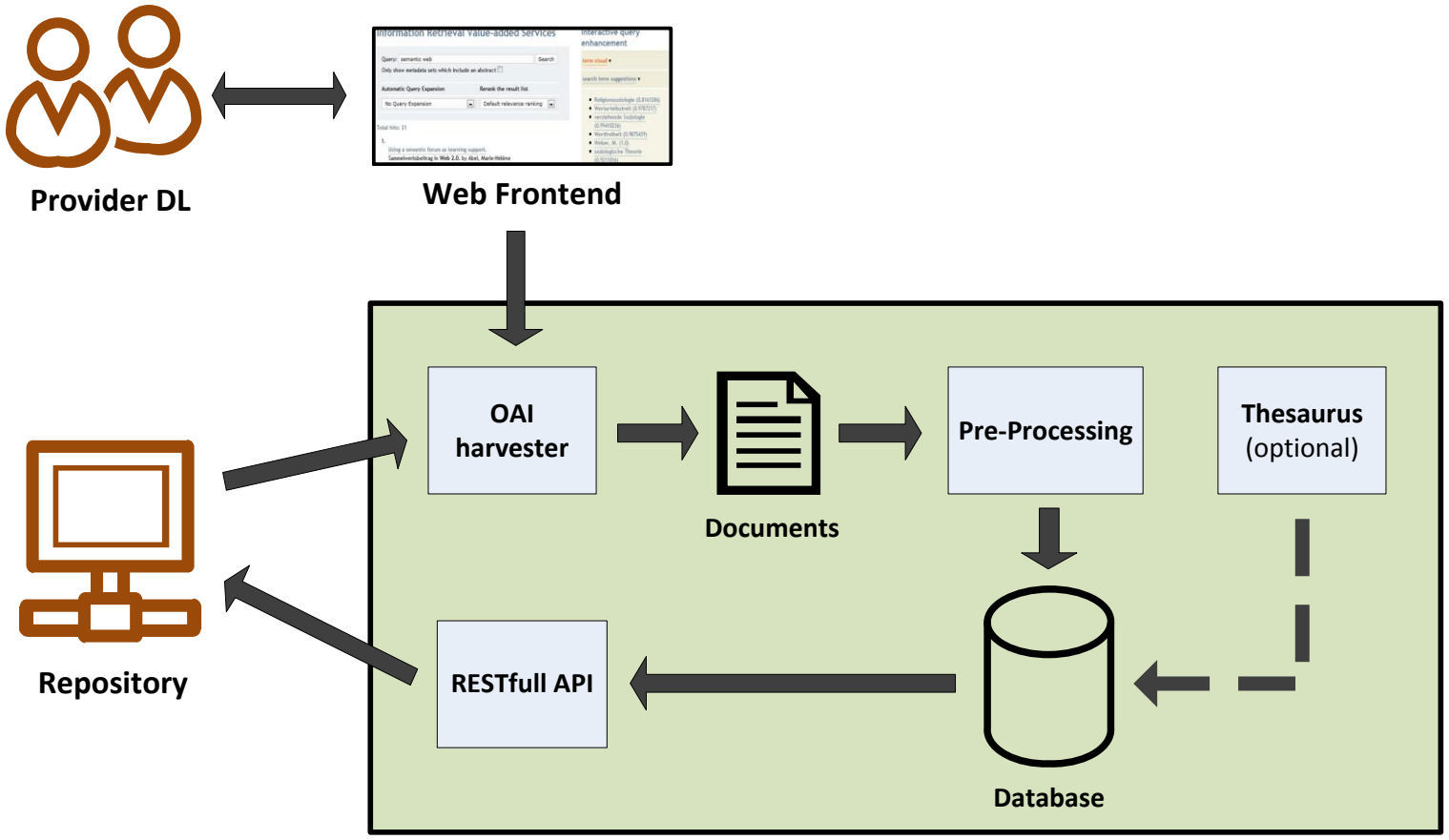
Controlled Vocabulary →

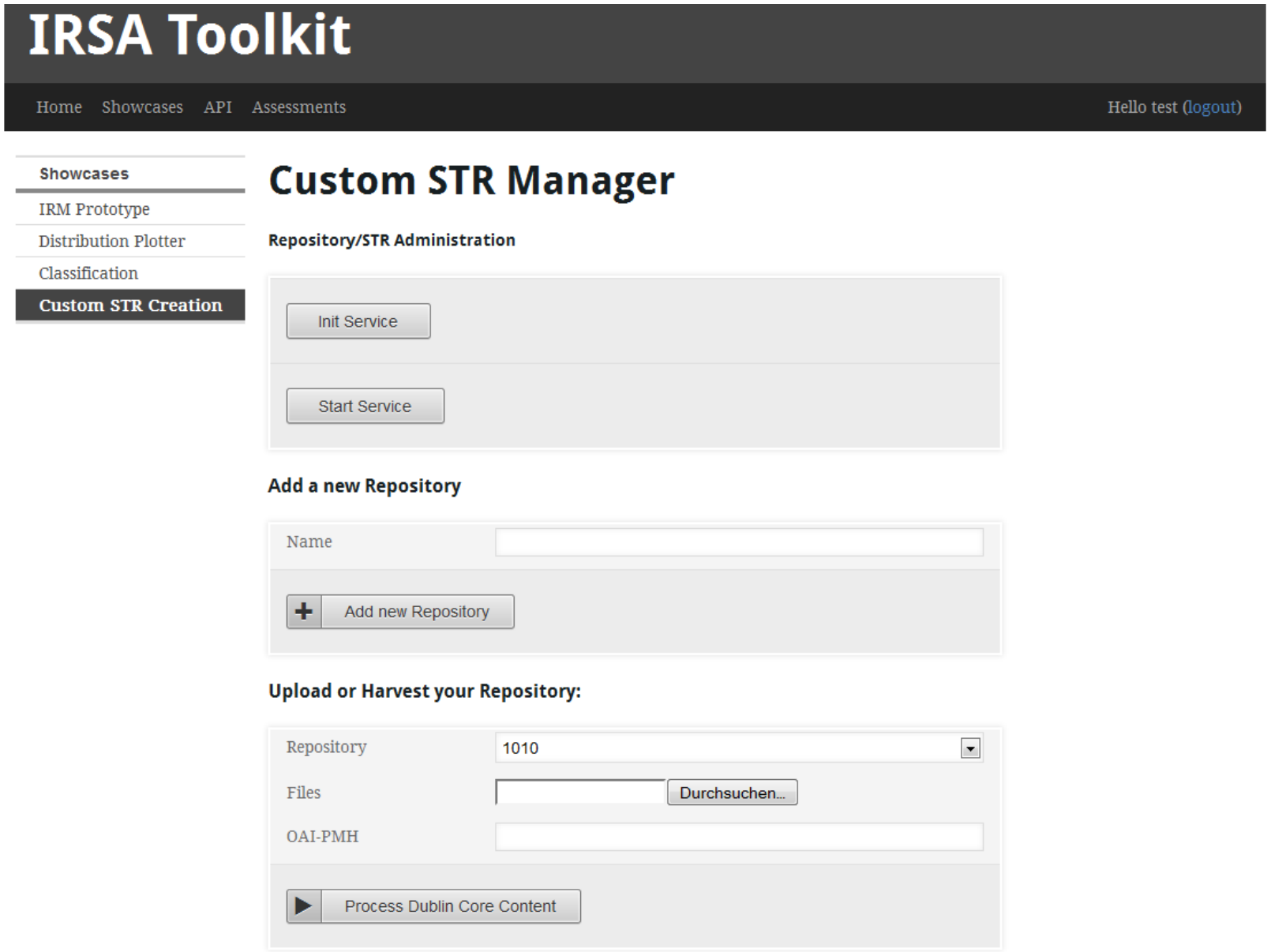
Free Terms in Description →

Co-Occurence Analysis of free and controlled vocabulary (e.g. using Jaccard, NWD, Dice etc.)

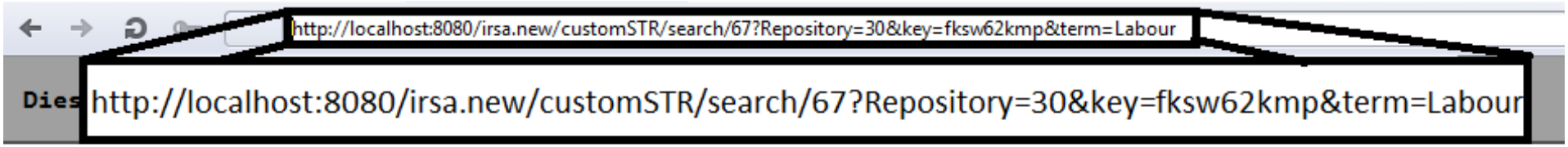
```
header:
  identifier : oai:gesis.izsoz.de:19389
  datestamp : 2011-01-10T13:46:00Z
  setSpec : SSOAR

metadata:
   dc:
      identifier: http://nbn-resolving.de/urn:nbn:de:0168-ssoar-193894
      title: How can international donors promote transboundary water management?
      creator: Mostert, Erik
      creator: Deutsches Institut für Entwicklungspolitik gGmbH
      subject: Political science (320)
      subject: Life sciences, biology (570)
      subject: International Relations, International Politics, Development Policy (10505)
      subject: Ecology, Environment (20900)
      subject: Management; Afrika; Entwicklung; Entwicklungsland; Akteur; Wasser
      source: Bonn
      source: DIE Discussion Paper (1860-0441) 8/2005
      description: "This paper discusses how international donors can promote the
          development of transboundary water management. It assumes, first, that
          cooperation will take place whenever the major stakeholders consider cooperation
          to be a better option than non-cooperation. The perceptions and motivations of
          the stakeholders are therefore crucial. Secondly, this paper assumes that the
          major stakeholders are not 'states', but specific groups and individuals:
          individual politicians, sectoral government bureaucracies, regional and local
          governments, farmers, electricity companies, etc. Some of these may be involved
          in the international negotiations themselves, others may be needed to get
          international agreements ratified or implemented, and still others may be
          affected by transboundary water management but lack the means to exert any
          influence." (author's abstract)
      language: English
      rights: Deposit Licence - No Redistribution, No Modifications
      contributor: SSOAR - Social Science Open Access Repository
      date: 10.01.2011 13:46
```

Workflow


Provider DL
Web Frontend
OAI harvester
Documents
Pre-Processing
Thesaurus (optional)
Repository
RESTfull API
Database


IRSA Toolkit
Home
Showcases
API
Assessments
Hello test (logout)
Showcases
IRM Prototype
Distribution Plotter
Classification
Custom STR Creation
Custom STR Manager
Repository/STR Administration
Init Service
Start Service
Add a new Repository
Name
Add new Repository
Upload or Harvest your Repository:
Repository
1010
Files
Durchsuchen...
OAI-PMH
Process Dublin Core Content

RESTful API Webservice




```
<searchtermrecommendations entryterm="Labour" limit="10" count="1613" sort="confidence" project="ssoar">
  <term name="Arbeitsmarktpolitik" confidence="0.7684294" type="subject" vendor="oaitester"/>
  <term name="Arbeitsmarktforschung" confidence="0.7186046" type="subject" vendor="oaitester"/>
  <term name="Arbeitsmarkt" confidence="0.71541077" type="subject" vendor="oaitester"/>
  <term name="Labor Market Research" confidence="0.71030426" type="subject" vendor="oaitester"/>
  <term name="Wirtschaft" confidence="0.69170296" type="subject" vendor="oaitester"/>
  <term name="Economics" confidence="0.6903163" type="subject" vendor="oaitester"/>
  <term name="Labor Market Policy" confidence="0.68868715" type="subject" vendor="oaitester"/>
  <term name="labor market" confidence="0.68090516" type="subject" vendor="oaitester"/>
  <term name="soziale Sicherung" confidence="0.63170177" type="subject" vendor="oaitester"/>
  <term name="EU" confidence="0.6113885" type="subject" vendor="oaitester"/>
</searchtermrecommendations>
```

Live Demo

Conclusion

As part of the IRM II project we have developed a system that

- is based on the **free Apache 2.0 License**
- may be used on **our servers or** can be set up on **your own system**
- uses the widely accepted **Dublin Core** standard via a **OAI-PMH** interface
- will now be beta-tested to estimate hardware requirements and further evaluate performance of custom sets

Got your attention? →
Thomas.Lueke@gesis.org or Philipp.Schaer@gesis.org for beta-test accounts

Further Information on our Project-Website:
http://www.gesis.org/en/research/external-funding-projects/projektuebersicht-drittmittel/irm2/

Thank you for your attention!

Any Questions?

The projects IRM I and IRM II

- DFG (German Research Foundation) Funding (2009-2013)
- IRM = Information Retrieval Mehrwertdienste
- Implementation and Evaluation of value added services for the retrieval in digital libraries
- Main idea: Usage of scientific models in IR
  - Bibliometrical analysis of core journals
  - Centrality scores in author networks
  - ➢ **Co-Occurence analysis of subjects**
- Our goal is the creation of reusable services

http://www.gesis.org/en/research/external-funding-projects/projektuebersicht-drittmittel/irm2/

Improvement in an individual query (GIRT 131). Original Query: *bilingual education.*

Table 1: Top 4 Recommendations of the 3 STRs

#	General (gSTR)	Topic-fitting (tSTR)	Best-performing (bSTR)
1	Multilingualism	Child	Multilingualism
2	Child	School	Speech
3	Speech	Multilingualism	Ethnic Group
4	Intercultural Education	Germany	Minority

Table 2: Statisics (bold font means further improvement)

Exp. Type	AP	rPrecison	p@5	p@10	p@20
No Exp.	0.039	0.127	0.4	0.3	0.2
gSTR	**0.072**	**0.144**	**0.6**	**0.6**	**0.4**
tSTR	**0.076**	**0.161**	**0.8**	0.6	**0.45**
bSTR	**0.147**	0.161	**1**	**1**	**0.85**

- A simple heuristic is used to select the best fitting STR for each topic (tSTR). We also list the general STR (gSTR) as baseline and the best-performing STR as comparison.
- To measure retrieval performance we use 100 topics from the GIRT corpus, measurements: MAP, rPrecision and p@{5,10,20}, * α = .05, ** α = .01

Exp. Type	MAP	rPrecision	p@5	p@10	p@20
gSTR	0.155	0.221	0.548	0.509	0.449
tSTR	0.159	0.224	0.578*	0.542**	0.460
bSTR	0.179**	0.233**	0.658**	0.601**	0.512**