Bayerische StaatsBibliothek

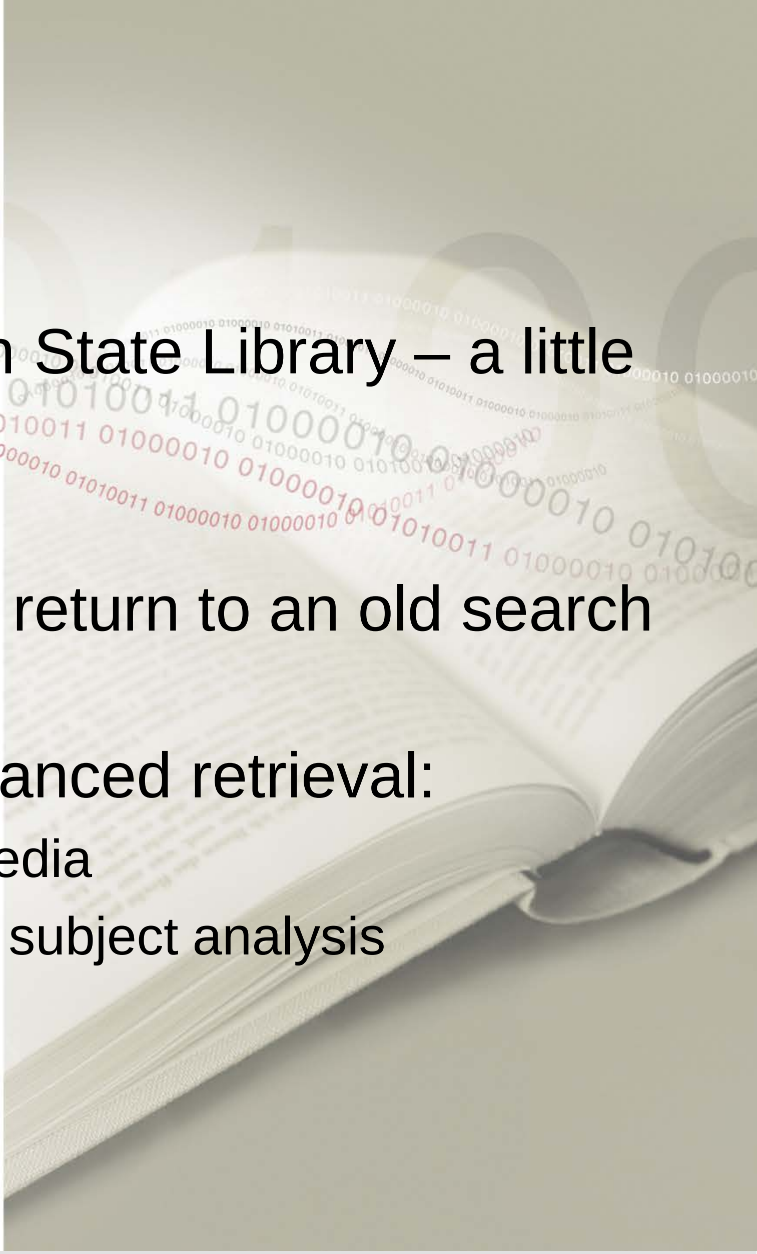
Information in erster Linie.

Enhanced retrieval using semantic technologies:

Ontology based retrieval as a new search paradigm? - Considerations based on new projects at the Bavarian State Library

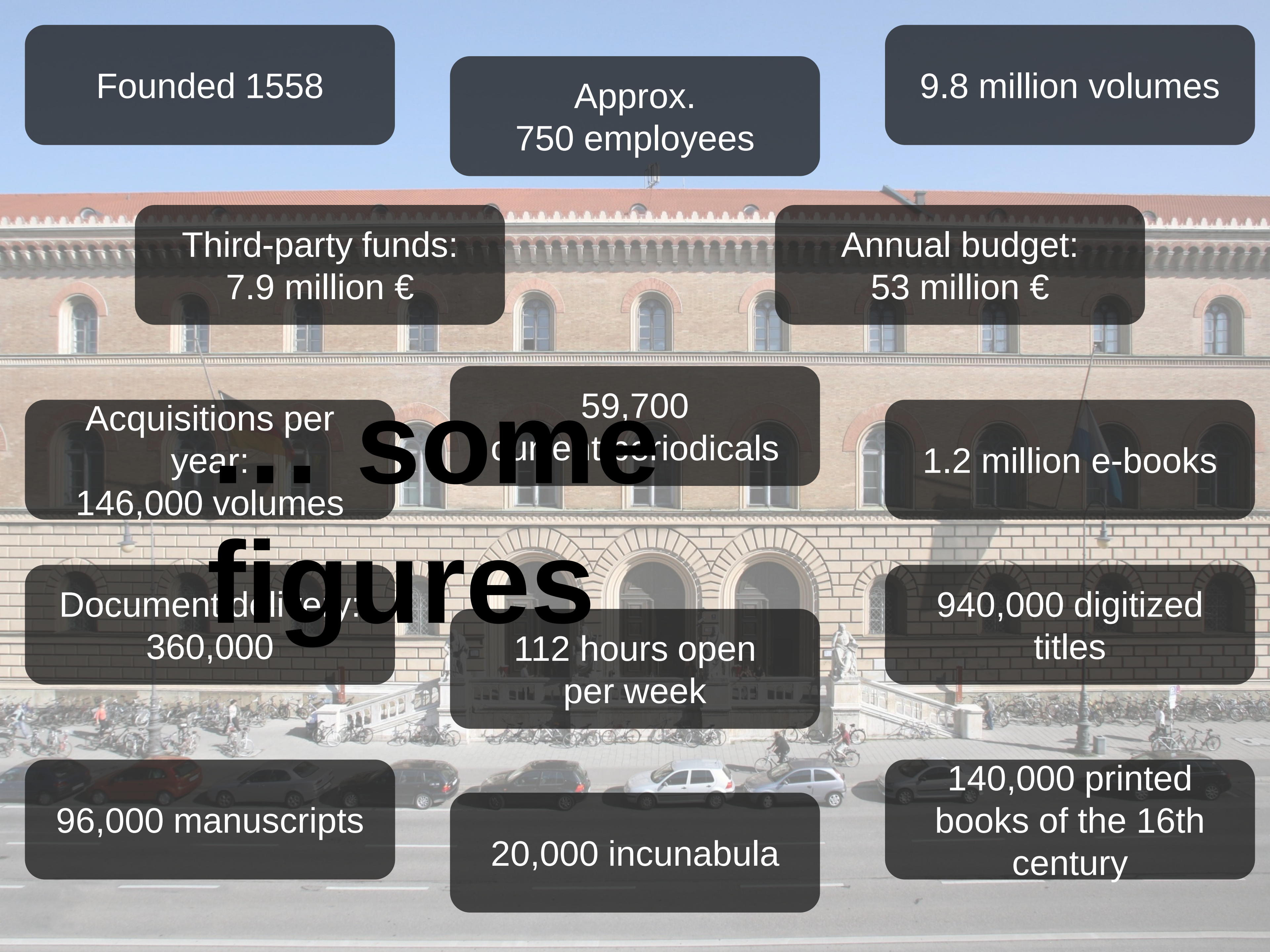
Dr. Berthold Gillitzer

28. Mai 2008



- Content:
  1. Some facts about the Bavarian State Library – a little bit of promotion
  2. Preliminary remarks
  3. A new search paradigm or the return to an old search paradigm?
  4. Two types of semantically enhanced retrieval:
     - A broad approach based on Wikipedia
     - A specialized approach for a deep subject analysis
  5. How to proceed?

... some figures



Founded 1558

Approx.
750 employees

9.8 million volumes

Third-party funds:
7.9 million €

Annual budget:
53 million €

Acquisitions per year:
146,000 volumes

59,700
current periodicals

1.2 million e-books

Document delivery:
360,000

112 hours open
per week

940,000 digitized titles

96,000 manuscripts

20,000 incunabula

140,000 printed books of the 16th century

The meaning of “semantic”, “ontology” and “concept”

- **Preliminary remarks**
  - The terms “semantic”, “ontology” and “concept” are not used in a narrow and technical sense but with a weak and a wide meaning
  - “Semantic search” is loosely interpreted as one that improves search accuracy and generates more relevant results by understanding the searcher’s intent and the contextual meaning of search terms
  - “ontology” means any kind of knowledge base with logically interconnected concepts, such as thesauri, authority files or DBpedia…

The meaning of “semantic”, “ontology” and “concept”

- **Preliminary remarks**
  - “Concept” is used for an unambiguous item in a system of knowledge representation with certain essential features and connected properties and relations apart from the various verbal expressions for this concept which are related to it
  - There are important relations between “semantic”, “ontology” etc. in a more technical sense as used in respect to linked open data eg. and these terms in a weaker sense as shown above but these relations are not to be discussed in this presentation

A new search paradigm or an old search paradigm?

- **Old fashioned end user catalogue retrieval in libraries:**
  - The user goes to the library and asks the librarian for help
  - A process of communication **based on shared common knowledge** helps clearing the subject
  - The librarian translates the description of the desired information into the catalogue description, identifies the relevant documents and provides the patron with desired books

A new search paradigm or an old search paradigm?

- **The second step: electronic catalogue with interface for the enduser**
  - Input of keywords by the enduser
  - String matching of the keywords in various categories (title, subject heading, classificatory systems, abstracts and so on)
  - Result list with the items where the string match was successful
  - Selection of the relevant title and ordering of the desired document
  - Advantages of this “new” method:
    - Enlargement of the “user interface” from one or a few librarians to an open web interface
    - Faster retrieval in more catalogue entries
    - Searching the catalogue independent from time and space

A new search paradigm or an old search paradigm?

- **The second step: electronic catalogue with interface for the enduser**
  - Disadvantages of this “new” method:
    - No dialogue to check the correct understanding of the subject
    - Ambiguity leads to false results
    - Synonyms are not taken into account as well as complex descriptions of a subject
    - Multilingual search is not possible

    ⇒The result list offers too many and at the same time too few results

A new search paradigm or an old search paradigm?

- **The next step –** enhanced retrieval using semantic technologies: Ontology based retrieval
  - The basic idea:
    - The catalogue search is first addressed to a knowledge base with logically interconnected unambiguous concepts
    - The user checks the relevance of his search terms in dialogue with the knowledge base
    - The catalogue items are connected to concepts of the knowledge base

    ⇒The result list is presented from the items connected with the concepts which are relevant for the user's search terms

    ⇒ modern catalogue retrieval is more similar to the old fashioned form of using the library implying a dialogue with the librarian

A new search paradigm or an old search paradigm?

- **The next step –** enhanced retrieval using semantic technologies: Ontology based retrieval
  - Is that new at all?
    - Doesn’t good old subject cataloguing provide the same service?
    - Your search is addressed to the subject headings which are logically interconnected in an authority file
    - The results are presented from the items connected to the identified subject headings

    ⇒The new enhanced retrieval does not compete with traditional subject cataloguing but is an improvement of its shortcomings:

A new search paradigm or an old search paradigm?

- **The next step –** enhanced retrieval using semantic technologies: Ontology based retrieval
  - Which shortcomings are to be cured?
    - Traditional subject cataloguing works with a controlled vocabulary which doesn’t match in every case with the expressions of the user even if different expressions are connected with a certain concept
    - The background knowledge in science evolves much faster than the library classification or authority files for subject cataloguing
    - Scientific language is multilingual as well as the user of the library but multilingualism is only a minor aspect of traditional subject cataloguing
    - Subject cataloguing is only related to the whole book or to specially relevant parts of at best but never addresses smaller parts of a document

A new search paradigm or an old search paradigm?

- **The next step –** enhanced retrieval using semantic technologies: Ontology based retrieval
  - Which shortcomings are to be cured?
    - The logical structure of the knowledge base in classificatory or subject cataloguing plays no substantial role in the catalogue search of the user
    - The scope of concepts describing the content of an item is limited to a few expressions contained in the title information together with the data from intellectual subject cataloguing
  - Consequences:

⇒ The background knowledge base needs a broad and multilingual vocabulary and up-to-date scientific concepts

⇒ We need a deeper and more detailed connection between the knowledge base and the respective documents not only on the level of titles but on the level of parts of the documents (single pages)

⇒ The logical structure of the knowledge base has to become an essential part of the bibliographic search engine

Consequences of the new search paradigm: two paradigmatical projects

- **The next step –** enhanced retrieval using semantic technologies - Ontology based retrieval

  Two different approaches:

  – Semantically enhanced search based on the broad knowledge base DBpedia:
    - SLUB semantics and related projects

  – Deeper subject analysis by collaboration between computer linguistics, librarians, special disciplines and information science
    - Specialized search engine for World War I in Eastern Europe

Semantical enhanced search based on the broad knowledge base DBpedia: SLUB semantics and related projects

- SLUB semantics and related projects – some basic information
  - Collaborative development from SLUB Dresden and Avantgard Labs GmbH Dresden
  - Enhancement of catalogue search in the respective OPAC using Wikipedia as an external data set to achieve multilingual and semantic search functionality
  - Project at the BSB to adopt and generalize the service for the enhancement of various OPACs

Semantical enhanced search based on the broad knowledge base DBpedia: SLUB semantics and related projects

- SLUB semantics and related projects
  - Working principle:

  1. Enrichment of bibliographic information with wikipedia concept-IDs taking different (analysis) steps:
     a. Enrichment of title information by collecting metadata and using interlinked identifiers in various bibliographic databases
     b. Assignment of Wikipedia concepts to bibliographic records by methods of data mining, disambiguation, language recognition etc.
     c. A score of (estimated) relevance is assigned to the concepts in relation to the bibliographic records
     d. Wikipedia concepts are part of DBpedia and as such unambigue and multilingual through the interlanguage links in Wikipedia
     e. The index of the search engine is enriched with the IDs of the Wikipedia concepts and the score of relevance

Semantical enhanced search based on the broad knowledgebase DBpedia: SLUB semantics and related projects

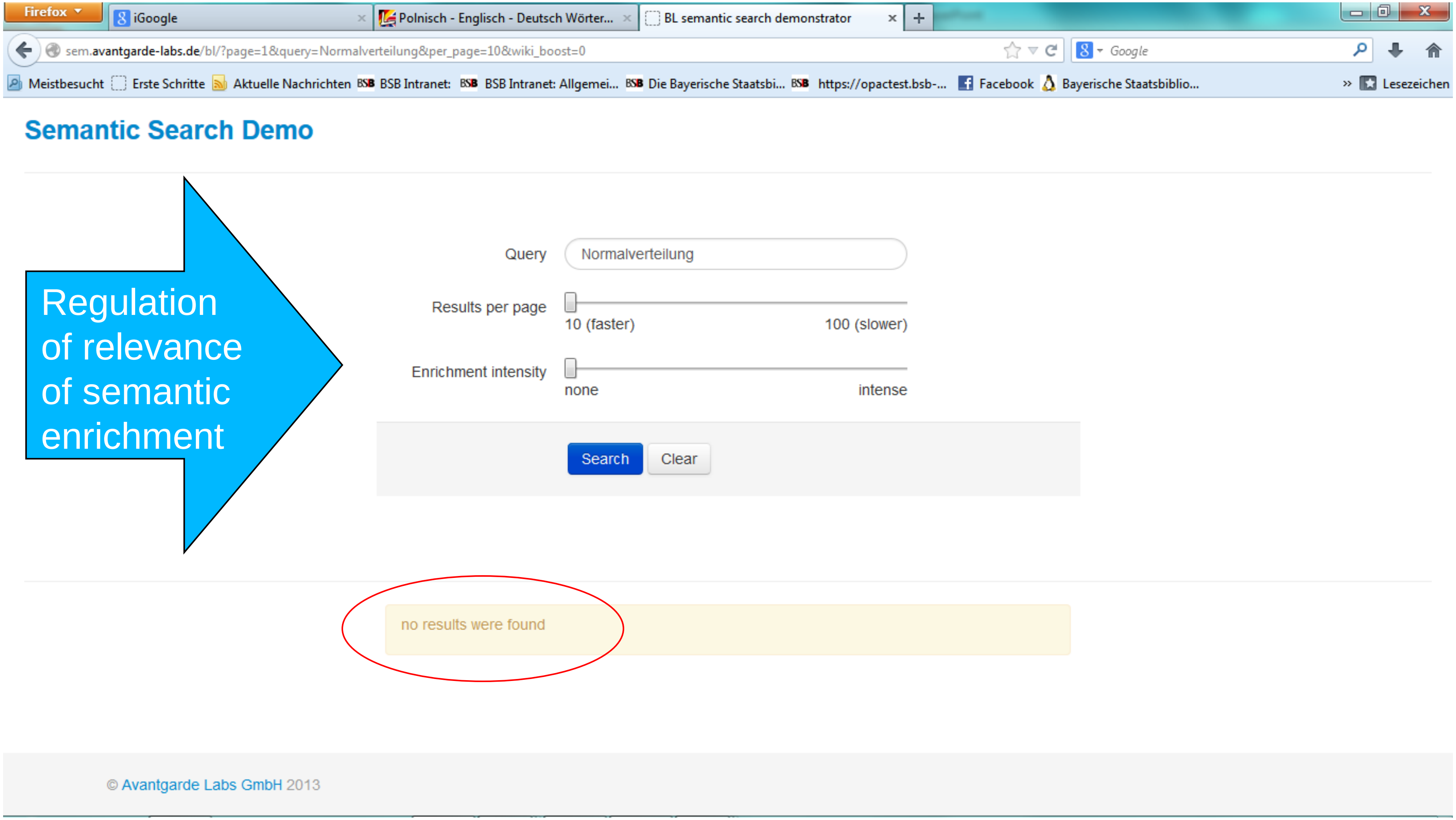
- SLUB semantics and related projects
  - Working principle:
    1. Enrichment of bibliographic information with wikipedia concept-IDs taking different (analysis) steps
    2. Expansion of the user query through assignment of Wikipedia concepts using similar methods
    3. Matching of Wikipedia concepts assigned to the user query and the catalogue entries additional to the string match between user query and the available metadata

Semantical enhanced search based on the broad knowledgebase DBpedia: SLUB semantics and related projects

- SLUB semantics and related projects
  - **Consequences**:

1. Multilinguality: Through inter language links in Wikipedia and a query expansion based on Wikipedia in various languages (German, English, Spain, Italian, Polish, Russian) multilingual retrieval is possible in a new way

Example from a demonstrator based on a random sample of data from the British Library:


Firefox
iGoogle
Polnisch - Englisch - Deutsch Wörter...
BL semantic search demonstrator
sem.avantgarde-labs.de/bl/?page=1&query=Normalverteilung&per_page=10&wiki_boost=0
Google
Meistbesucht
Erste Schritte
Aktuelle Nachrichten
BSB Intranet:
BSB Intranet: Allgemei...
Die Bayerische Staatsbi...
https://opactest.bsb-...
Facebook
Bayerische Staatsbiblio...
Lesezeichen

Semantic Search Demo

Query: Normalverteilung

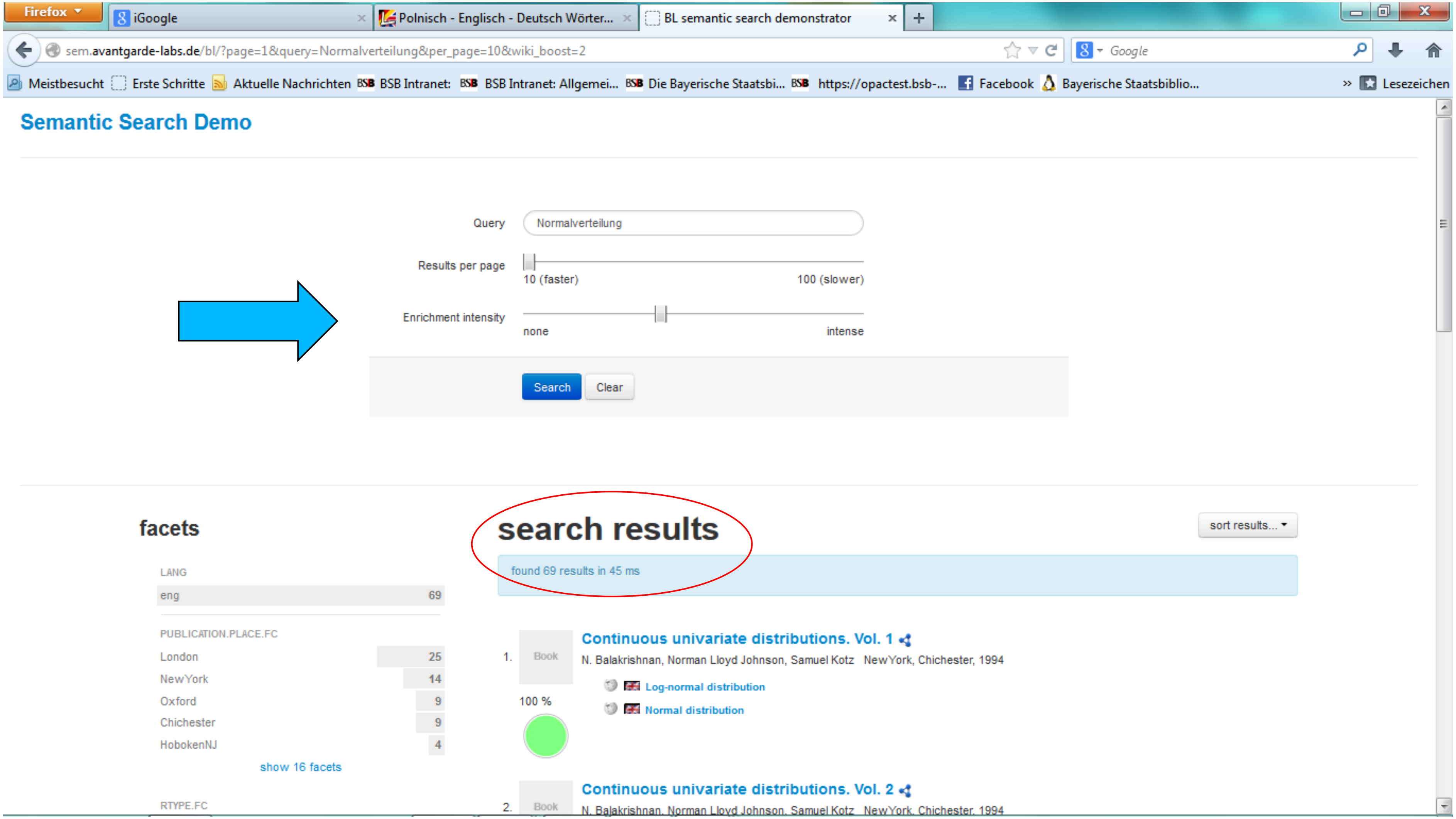
Results per page: 10 (faster) — 100 (slower)

Enrichment intensity: none — intense

no results were found

© Avantgarde Labs GmbH 2013

Regulation of relevance of semantic enrichment



Firefox
iGoogle
Polnisch - Englisch - Deutsch Wörter...
BL semantic search demonstrator
sem.avantgarde-labs.de/bl/?page=1&query=Normalverteilung&per_page=10&wiki_boost=2
Google
Meistbesucht
Erste Schritte
Aktuelle Nachrichten
BSB Intranet:
BSB Intranet: Allgemei...
Die Bayerische Staatsbi...
https://opactest.bsb-...
Facebook
Bayerische Staatsbiblio...
Lesezeichen

Semantic Search Demo

Query: Normalverteilung

Results per page: 10 (faster) – 100 (slower)

Enrichment intensity: none – intense

facets

LANG

eng	69

PUBLICATION.PLACE.FC

London	25
NewYork	14
Oxford	9
Chichester	9
HobokenNJ	4

show 16 facets

RTYPE.FC

search results

found 69 results in 45 ms

1. Book — Continuous univariate distributions. Vol. 1
N. Balakrishnan, Norman Lloyd Johnson, Samuel Kotz NewYork, Chichester, 1994
100 %
- Log-normal distribution
- Normal distribution

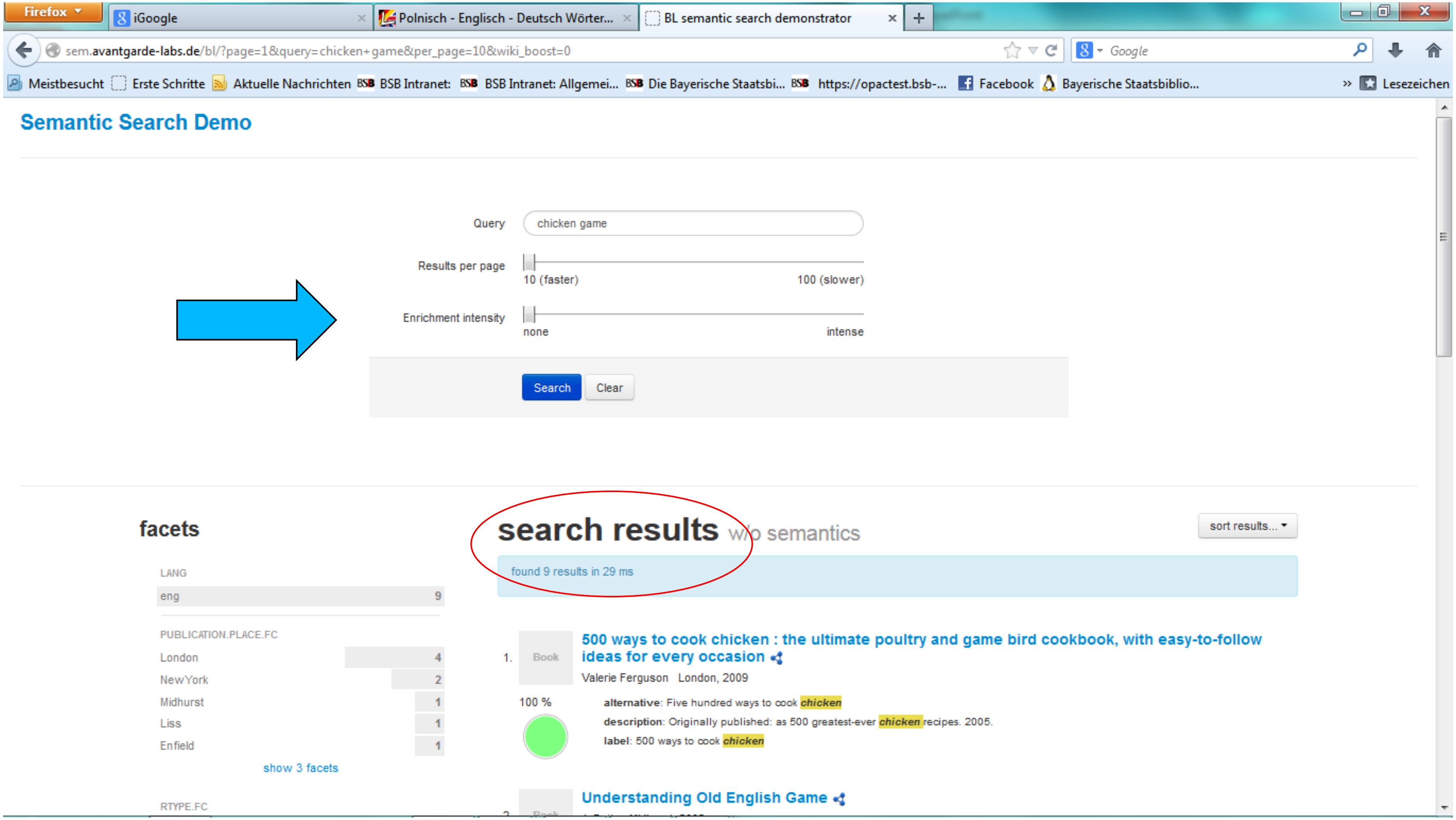
2. Book — Continuous univariate distributions. Vol. 2
N. Balakrishnan, Norman Lloyd Johnson, Samuel Kotz NewYork, Chichester, 1994

Semantical enhanced search based on the broad knowledgebase DBpedia: SLUB semantics and related projects

- SLUB semantics and related projects
  - **Consequences**:

2. Conceptual (semantic) search: The search based on the extraction and mapping of concepts as shown above yields more and in many cases more relevant results than a search based solely on string match between the user query and the bibliographic data

Example from a demonstrator based a on random sample of data from the British Library:


Firefox
iGoogle
Polnisch - Englisch - Deutsch Wörter...
BL semantic search demonstrator
sem.avantgarde-labs.de/bl/?page=1&query=chicken+game&per_page=10&wiki_boost=0
Google
Meistbesucht
Erste Schritte
Aktuelle Nachrichten
BSB Intranet:
BSB Intranet: Allgemei...
Die Bayerische Staatsbi...
https://opactest.bsb-...
Facebook
Bayerische Staatsbiblio...
Lesezeichen

Semantic Search Demo

Query: chicken game

Results per page: 10 (faster) — 100 (slower)

Enrichment intensity: none — intense

Search | Clear

facets

LANG

eng	9

PUBLICATION.PLACE.FC

London	4
NewYork	2
Midhurst	1
Liss	1
Enfield	1

show 3 facets

RTYPE.FC

search results w/o semantics

sort results...

found 9 results in 29 ms

1. Book — **500 ways to cook chicken : the ultimate poultry and game bird cookbook, with easy-to-follow ideas for every occasion**

Valerie Ferguson London, 2009

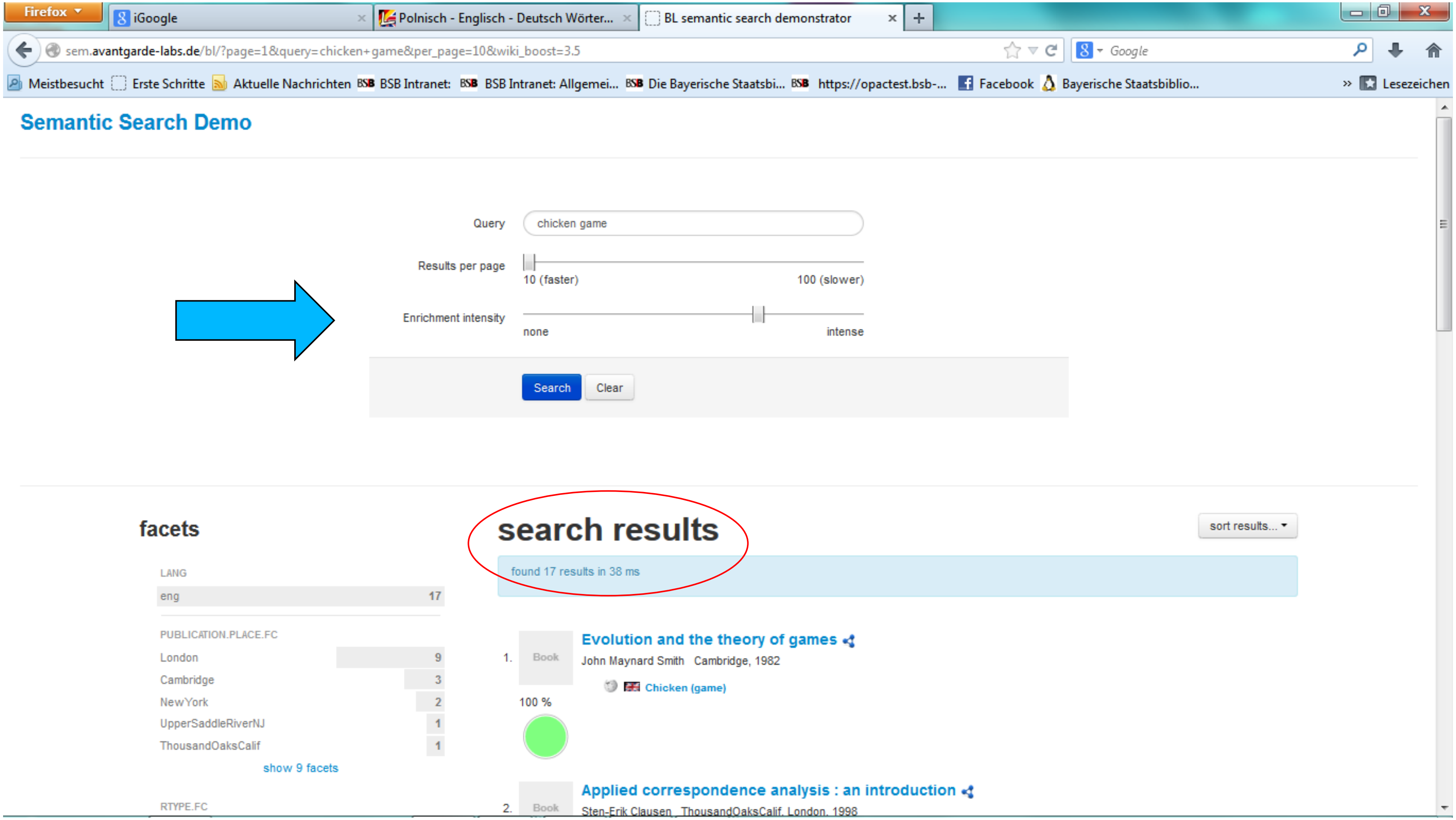
100 %

alternative: Five hundred ways to cook *chicken*

description: Originally published: as 500 greatest-ever *chicken* recipes. 2005.

label: 500 ways to cook *chicken*

2. Book — **Understanding Old English Game**


Firefox
iGoogle
Polnisch - Englisch - Deutsch Wörter...
BL semantic search demonstrator
sem.avantgarde-labs.de/bl/?page=1&query=chicken+game&per_page=10&wiki_boost=3.5
Google
Meistbesucht
Erste Schritte
Aktuelle Nachrichten
BSB Intranet:
BSB Intranet: Allgemei...
Die Bayerische Staatsbi...
https://opactest.bsb-...
Facebook
Bayerische Staatsbiblio...
Lesezeichen
Search
Clear
sort results...
show 9 facets

Semantic Search Demo

Query: chicken game

Results per page: 10 (faster) – 100 (slower)

Enrichment intensity: none – intense

facets

LANG

eng	17

PUBLICATION.PLACE.FC

London	9
Cambridge	3
NewYork	2
UpperSaddleRiverNJ	1
ThousandOaksCalif	1

RTYPE.FC

search results

found 17 results in 38 ms

1. Book — **Evolution and the theory of games**
John Maynard Smith Cambridge, 1982
Chicken (game)
100 %

2. Book — **Applied correspondence analysis : an introduction**
Sten-Erik Clausen , ThousandOaksCalif, London, 1998

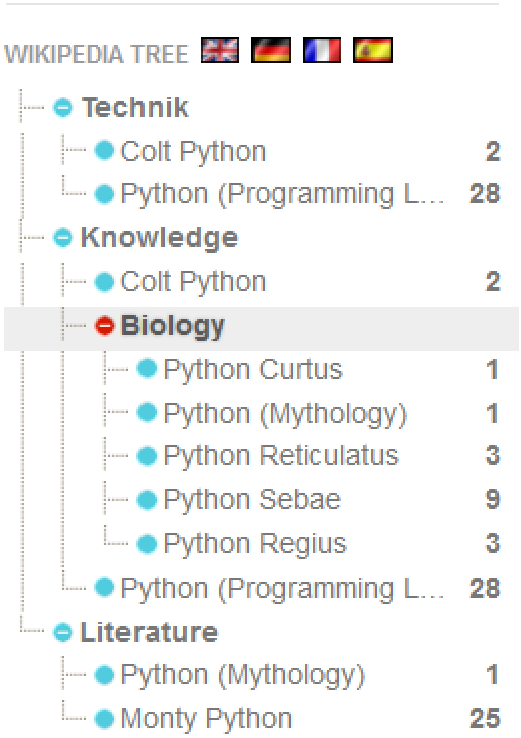
Semantical enhanced search based on the broad knowledgebase DBpedia: SLUB semantics and related projects

- SLUB semantics and related projects
  - **Consequences:**

    3. Navigation using the logical structure of the background knowledge base Wikipedia: the graph of the Wikipedia category system can be used to generate new subject related facets to filter the initial result list as well as to expand the search results to items connected to higher level concepts or concepts with other logical connections

Navigation based on the logical structure of wikipedia

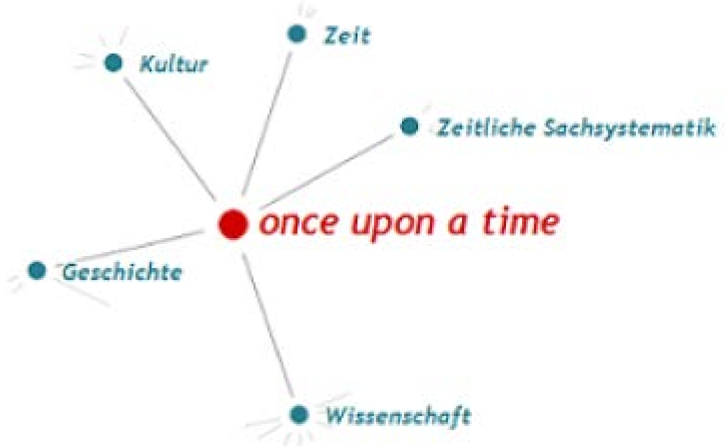
Wikipedia facets in the test application



WIKIPEDIA TREE

- Technik
  - Colt Python 2
  - Python (Programming L... 28
- Knowledge
  - Colt Python 2
  - Biology
    - Python Curtus 1
    - Python (Mythology) 1
    - Python Reticulatus 3
    - Python Sebae 9
    - Python Regius 3
  - Python (Programming L... 28
- Literature
  - Python (Mythology) 1
  - Monty Python 25

Navigation using a topic graph based on Wikipedia categories



Semantical enhanced search based on the broad knowledgebase DBpedia: SLUB semantics and related projects

- SLUB semantics and related projects
  - Remarks:
    1. Some of the results of this project would have been possible even on the basis of traditional subject cataloguing if only the logical structure of the authority files were used
    2. The semantic enrichment is seriously based on subject cataloguing for assignment of Wikipedia concepts to bibliographic records
    3. The quality of the results depends on the quality of the bibliographic records; if they are “poor” the assignment of Wikipedia concepts is difficult
    4. This method does not go deep into the documents

    ⇒ This method is not useful to obtain new “scientific” information about the documents in a library

Semantic social library search engine for special subject areas

- Specialized search engine for World War I in Eastern Europe – a complementary approach
- Project aims:
  - Development of a generic software for a “deep” subject oriented search in documents of a special subject area
  - Identification of unambiguous concepts (subjects) within the electronic fulltext, abstracts, table of contents and indexes of the documents of special subject area with computer linguistic methods vs. string matching in fulltext
  - Creation of a specialized ontology (knowledge base) concerning World War I in East-, Central East and South East European countries
  - Improvement of the semantic search by a social search: providing of an own workspace for the user with various possibilities of private and shared annotations which should lead to a folksonomy and help to improve the basic ontology

Semantic social library search engine for special subject areas

- Specialized search engine for World War I in Eastern Europe – a complementary approach
- Status of the project:
  - The raw concept is completed
  - Participating institutions:
    - Bavarian State Library
    - Institute for Informatics (Ludwigs-Maximilians-Universität Munich – LMU)
    - Center for Information and Language Processing – LMU (CIS)
    - Faculty of Linguistics and Literary Studies – Slavic Philology (LMU)
  - Providing of the documents and electronic fulltexts is in process
  - Start of the project probably in January 2014

Semantic social library search engine for special subject areas

- The key role of the ontology
- Creation of the ontology:
  - Based on special thesauri, encyclopedias, authority files (like GND) and other sources the ontology is built up in three dimensions: persons/institutions, places and events
  - The concepts of the ontology are unambiguous and logically interconnected (which persons are involved in which event, where does an event take place…)
  - The ontology is partly built automatically by processes such as named entity extraction from relevant texts but the whole ontology is intellectually controlled by specialist for language processing as well as specialists for Slavic studies
  - For each concept different and even multilingual verbal expressions are stored with the concept

Semantic social library search engine for special subject areas

- The key role of the ontology
- Ontology based (semantic) text analysis and indexing:
  - Identification of concepts from the ontology within the texts based on the variants of naming and verbal expressions of these concepts
  - The relations between the concepts are taken into account in the process of the calculation of the index
  - Concepts are detected within the whole text corpus and assigned not only to the whole title but also to parts or single pages of a document
  - The building of the ontology and the semantic text analysis are interdependent and iterative processes
- Ontology based (semantic) search:
  - The user’s query is primarily directed to the knowledge base
  - In the system’s first response to the user’s query information about the involved concepts and their content is presented additionally to the bibliographic data so that the user can check whether the system really understood what he meant and searched for

Semantic social library search engine for special subject areas

- The social search:
- Elements of the crowdsourcing component:
  - Creation of a simple but attractive workspace for the user to build private title lists, make annotations to the texts and integrate the results of his work with the search engine in reference management software or web 2.0 applications
  - Possibility of defining open or controlled groups to share the results of the user’s work as well as annotations with other scientists
  - Possibility of explicit improvement of the ontology and the index of the text analysis to remove ambiguities or falsely detected concept occurrences in the text due to ambiguity and to enlarge and improve the ontology
  - Semantic and statistical analysis of the user’s open shared annotations to build up an index for the social search

Semantic social library search engine for special subject areas

- Why also social search?
- Annotations of users often reflect current developments and interests of the scientific community and can therefore improve the search results when taken into account in the building of the search index
- Ontology based search presupposes a high quality ontology which can only be achieved by a significant amount of intellectual work
- To keep the ontology up to date and improve it continuously this process of development can not be based only on the work of “professional” librarians or scientist who are paid, but needs the collaboration of the scientific community working with the search engine

Semantic social library search engine for special subject areas

- Remarks:
- The project hasn’t started and the success has yet to be proved
- In relation to the number of documents and titles this kind of content analysis is complex and comparatively labour-intensive
- This approach seems to be useful for special topics and is aimed at researchers
- The main benefit of this kind of content analysis is the detection of relevant literature to a certain topic where this hardly would have been possible without semantic technologies

The future of subject oriented retrieval

- How to proceed?
- In times of Google Books, open archives and other publicly available services for digital information the support in subject oriented retrieval with a high quality “semantic” service might be a crucial factor of the future success of libraries as information providers
- Traditional subject cataloguing, a broad approach to automatic topic detection like SLUBsemantics and a specialized approach like the project of the specialized search engine World War I are not competing but complementary
- It is a major task for libraries to connect these services in order to make this kind of service available throughout on different levels of subject analysis for different needs
- A deep subject analysis as in the second project should be developed in intense collaboration with the respective researchers. Therefore this kind of service should be included in other services with an intense connection to researchers such as the development of research enviroments

Thank you for your attention!

Any further questions?

Ask now or don‘t be quiet for ever but write an e-mail:
berthold.gillitzer@bsb-muenchen.de

Dr. Berthold Gillitzer, Bavarian State Library