Universität Bielefeld

INFORMATION.plus!
Universitätsbibliothek

BASE
Bielefeld Academic Search Engine

10 years of „Bielefeld Academic Search Engine“ (BASE)

Dirk Pieper - Friedrich Summann
Bielefeld University Library

OR2015 | 10th International Conference on Open Repositories
June 8-11, 2015, Indianapolis, Indiana, USA

Universität Bielefeld

INFORMATION.plus!
Universitätsbibliothek

10 years of „Bielefeld Academic Search Engine“ (BASE)

(www.base-search.net)

Looking at the past and future of the world wide repository landscape from a service providers perspective

BASE

Overview

- **Retrospect (2001 – 2014)**
- Current Situation (2015)
- Future prospective (2015 - …)

Some Milestones (I)

- 2001 Starting point as a search engine follow-up for a metasearch system
- 2004 Official Start (FAST Data Search)
- 2005 search history, sorting
- 2006 starting participation in EU projects (DRIVER), BASE-APIs
- 2007 Introducing multilingual search
- 2008 Index > 10 million records

Some Milestones (II)

- 2009-2011 Automatic classification of OAI-DC-Metadata (funded by German Research Foundation)
- 2011 Switch to open source (Lucene/Solr, VuFind)
- 2011 Index > 30 million records, Iphone-App
- 2012 Web-App, OAI-Interface, data delivery of subject sections
- 2014 OA-boosting

BASE

Universität Bielefeld

INFORMATION.plus! Universitätsbibliothek

BASE 2004

BASE Digital Collections

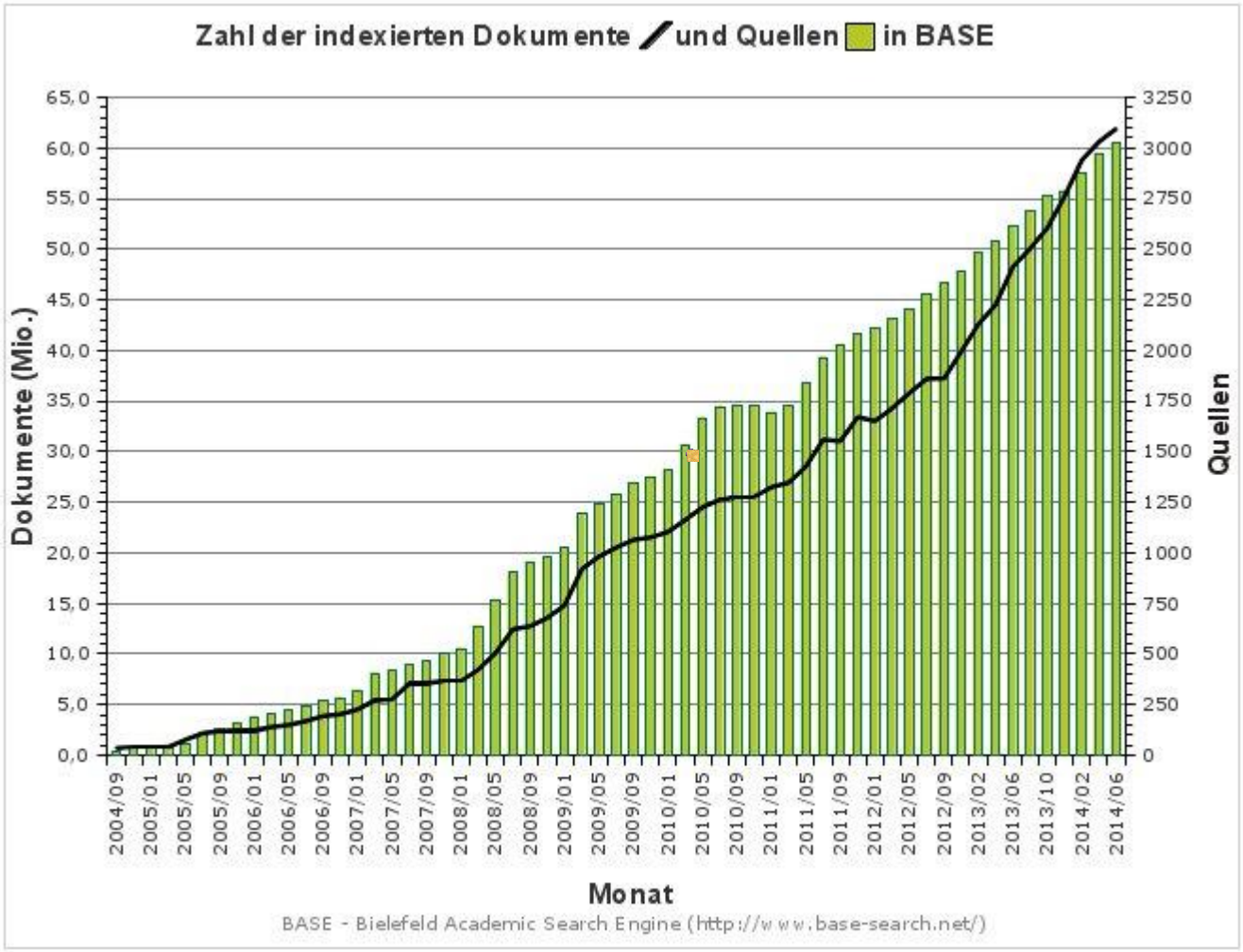
Demonstr

Basic Sea
Deutsche \

free con

BASE includes the following content sources

Mathematical Preprint Server, Bielefeld University, Faculty for Mathematics (complete)
E-journal Documenta Mathematica, Bielefeld University, Faculty for Mathematics (complete)
Online Publications, Bielefeld University (complete)
Journals of the German Enlightenment, Bielefeld University Library (complete)
Library Catalogue, Bielefeld University Library (partial)
Electronic Dissertations, Bochum University (complete)
Project Euclid, Cornell University (partial)
Historical Math Monographs, Cornell University Library (partial)
Internet Library of Early Journals, Host: Oxford University Library Services (complete)
Mathematical Collection, Springer Publishing House Heidelberg, Germany (partial)
Digital Collection Mathematica, SUB Goettingen/GDZ (complete)
Research Reports, BMBF, TIB/UB Hannover (partial)
Historical Mathematics Collection, The University of Michigan (partial)
Zentralblatt MATH, Host: FIZ Karlsruhe / Zentralblatt für Mathematik (partial)



Tabelle

BASE

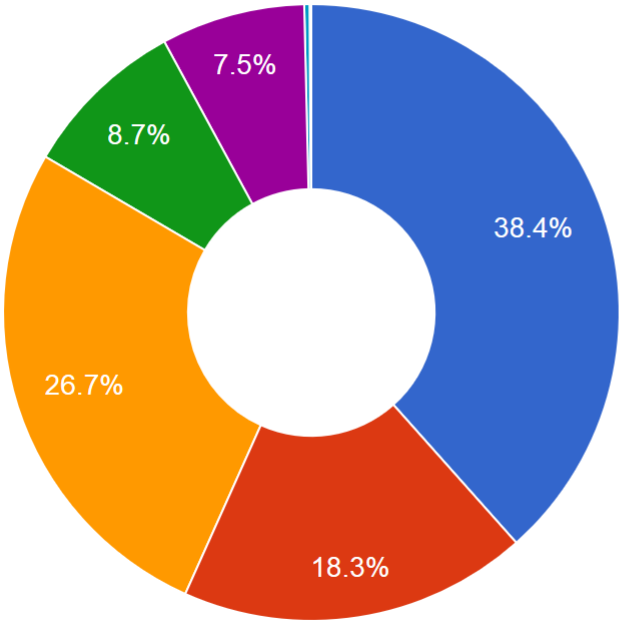
Overview

- Retrospect (2001 – 2014)
- **Current Situation (2015)**
- Future prospective (2015 - …)

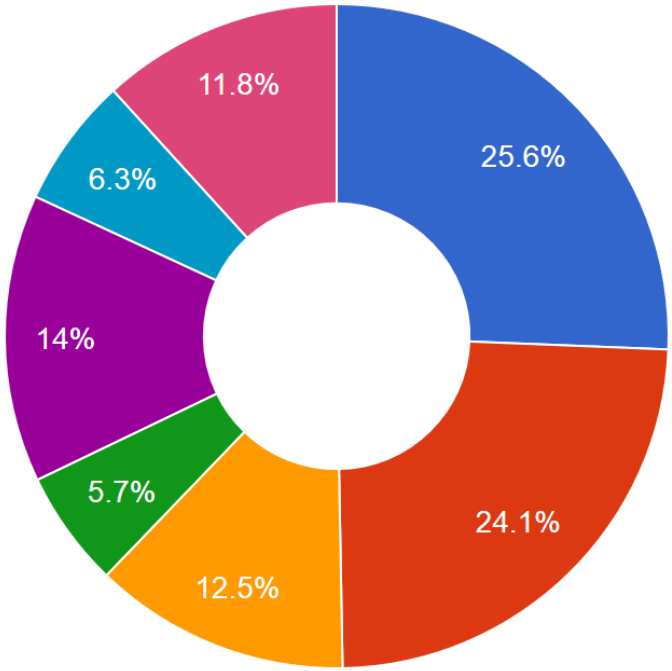
The BASE scope

- OA Repositories world-wide (institutional and subject reps)
- Academic Valuable Content
- Electronic Journals
- Aggregators (RePEc, Virtual Libraries, etc.)
- Digital Collections
- Dataset Repositories

Repository Types covered in BASE



- Institutional Repositories
- Publication Server
- Electronic Journals
- Thesis/Dissertation Servers
- Digital Collections
- Research Data
- Sonstiges



Universität Bielefeld

INFORMATION.plus!
Universitätsbibliothek

Contents:

Global Scientific Metadata harvested from Repositories

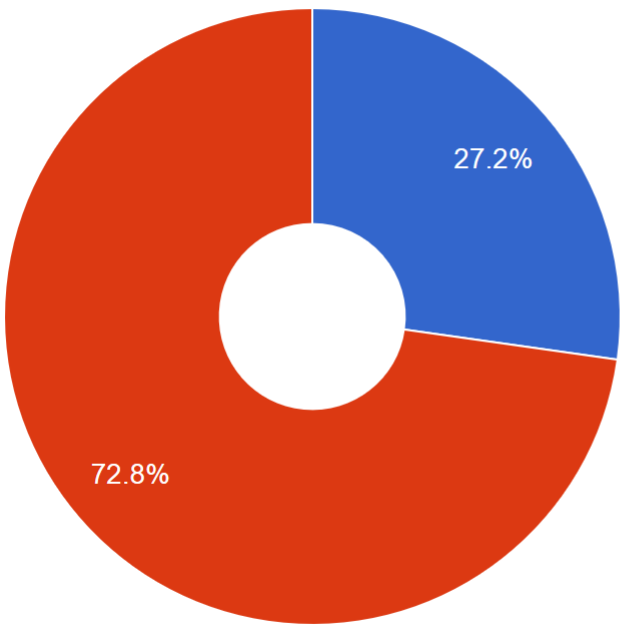
Technology:

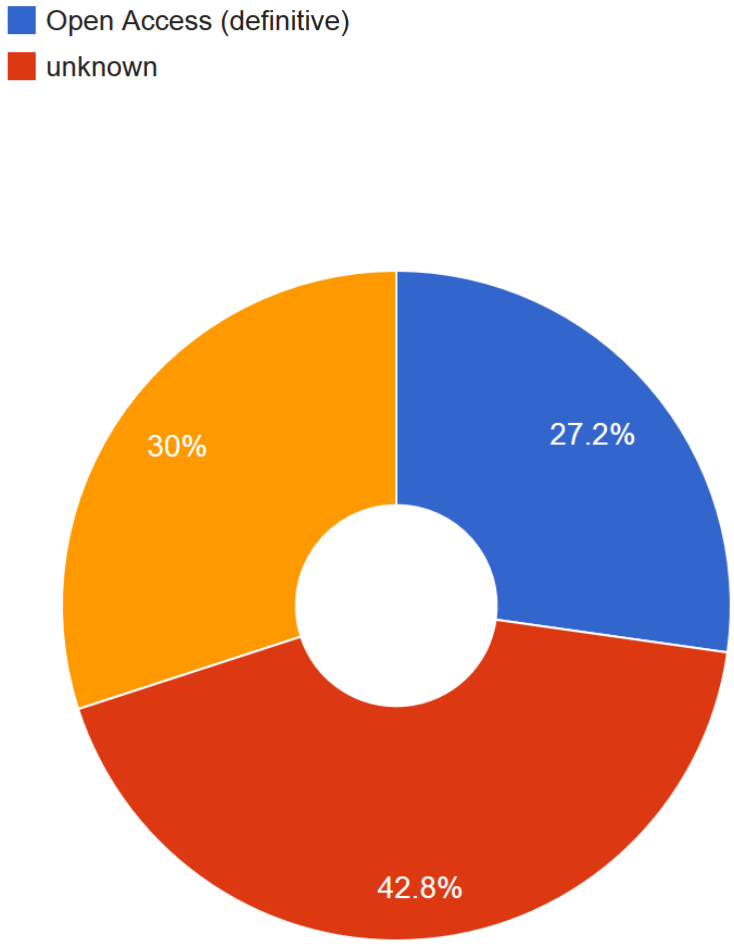
- Index (Solr)
- Search Interface (VuFind)
- APIs (HTTP, OAI)

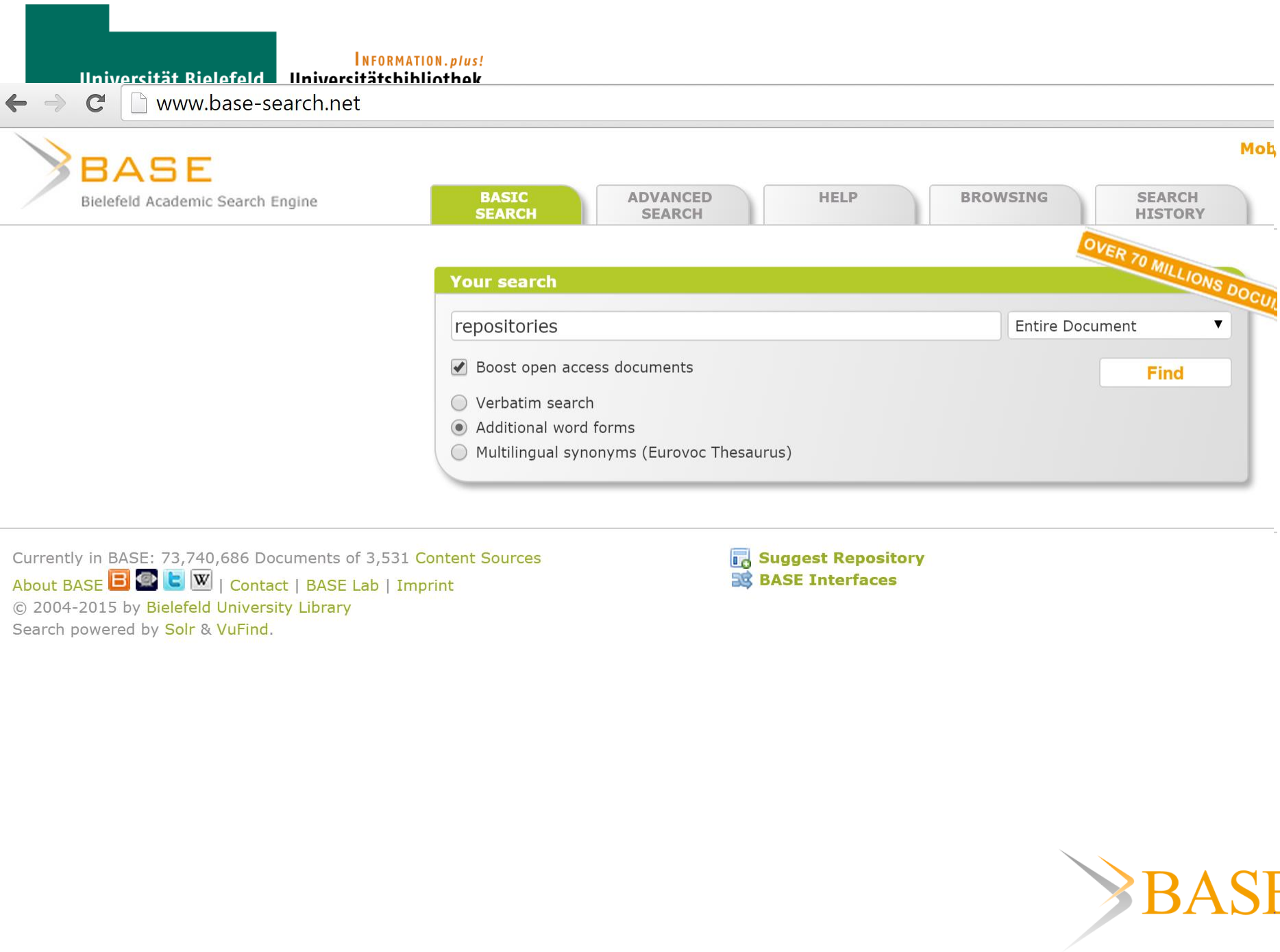
BASE harvests, aggregates, enriches and exposes OAI Metadata

BASE

Open Access Status






Universität Bielefeld
INFORMATION.plus!
Universitätsbibliothek
www.base-search.net
BASE
Bielefeld Academic Search Engine
Mob
BASIC SEARCH
ADVANCED SEARCH
HELP
BROWSING
SEARCH HISTORY
OVER 70 MILLIONS DOCU
Your search
repositories
Entire Document
Boost open access documents
Find
Verbatim search
Additional word forms
Multilingual synonyms (Eurovoc Thesaurus)
Currently in BASE: 73,740,686 Documents of 3,531 Content Sources
About BASE | Contact | BASE Lab | Imprint
© 2004-2015 by Bielefeld University Library
Search powered by Solr & VuFind.
Suggest Repository
BASE Interfaces
BASE

Universität Bielefeld

INFORMATION.plus!
Universitätsbibliothek

Hit List

1. HKUST Institutional Repository

Title: HKUST Institutional Repository

Hit List

1. Wikibooks: Git/Repository on a USB stick

Open Access

Title: Wikibooks: Git/Repository on a USB stick

Description: Instead of having to resort to a hosting company to store your central repository or to rely on a central server or internet connection to contribute changes to a project it s quite possible to use removeable memory to exchange and update local **repositories**. The basic steps are Mount the removable memory on a pre determined path Setup a bare re... Show all

Document Type: Book

Language: eng

URL: http://en.wikibooks.org/wiki/Git/Repository_on_a_USB_stick

Content Provider: WikiBooks - Open-content textbooks

Detail View | Email this | Export Record | Add to Favorites | Check in Google Scholar

2. RepositoriUM. Institutional Repository. Minho University, Braga, Portugal: A university repository where a mandate to deposit, financial incentives and strong advocacy can...

Open Access

Title: RepositoriUM. Institutional Repository. Minho University, Braga, Portugal: A university repository where a mandate to deposit, financial incentives and strong advocacy can transform an Institutional Repository's population

Author: Proudman, V.M.

Universität Bielefeld

INFORMATION.plus!
Universitätsbibliothek

Facts about BASE

- 3531 Repositories included
- From 102 Countries world-wide
- Ca. 73 Mill. Documents/Objects
- Ca. 70 % Open Accessible
- Ca. 10.3 Mill. Documents enriched with DDC-Codes (Dewey)

BASE

Universität Bielefeld

INFORMATION.plus! Universitätsbibliothek

Harvesting Environment
(Based on OAI-PMH)

- 5979 Repositories harvested
  - 4832 active
  - 3531 indexed
  - 1147 deprecated
- 184 Mill. Records
  - 111 Mill. unique
  - 73 Mill. indexed
- 1.03 Terabyte of Data
- 2712 Cronjobs (weekly)

BASE

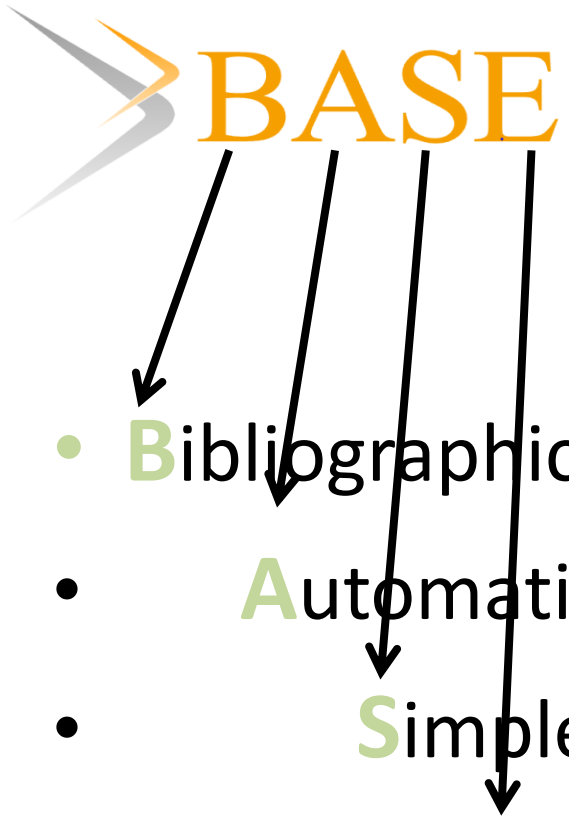
Khabsa M, Giles CL (2014) The Number of Scholarly Documents on the Public Web. PLOS ONE 9 (5) DOI: 10.1371/journal.pone.0093949

Figure 2. Relative number of documents by scholarly search engines and databases. Total and Google Scholar are estimates.
doi:10.1371/journal.pone.0093949.g002

Universität Bielefeld

INFORMATION.plus! Universitätsbibliothek

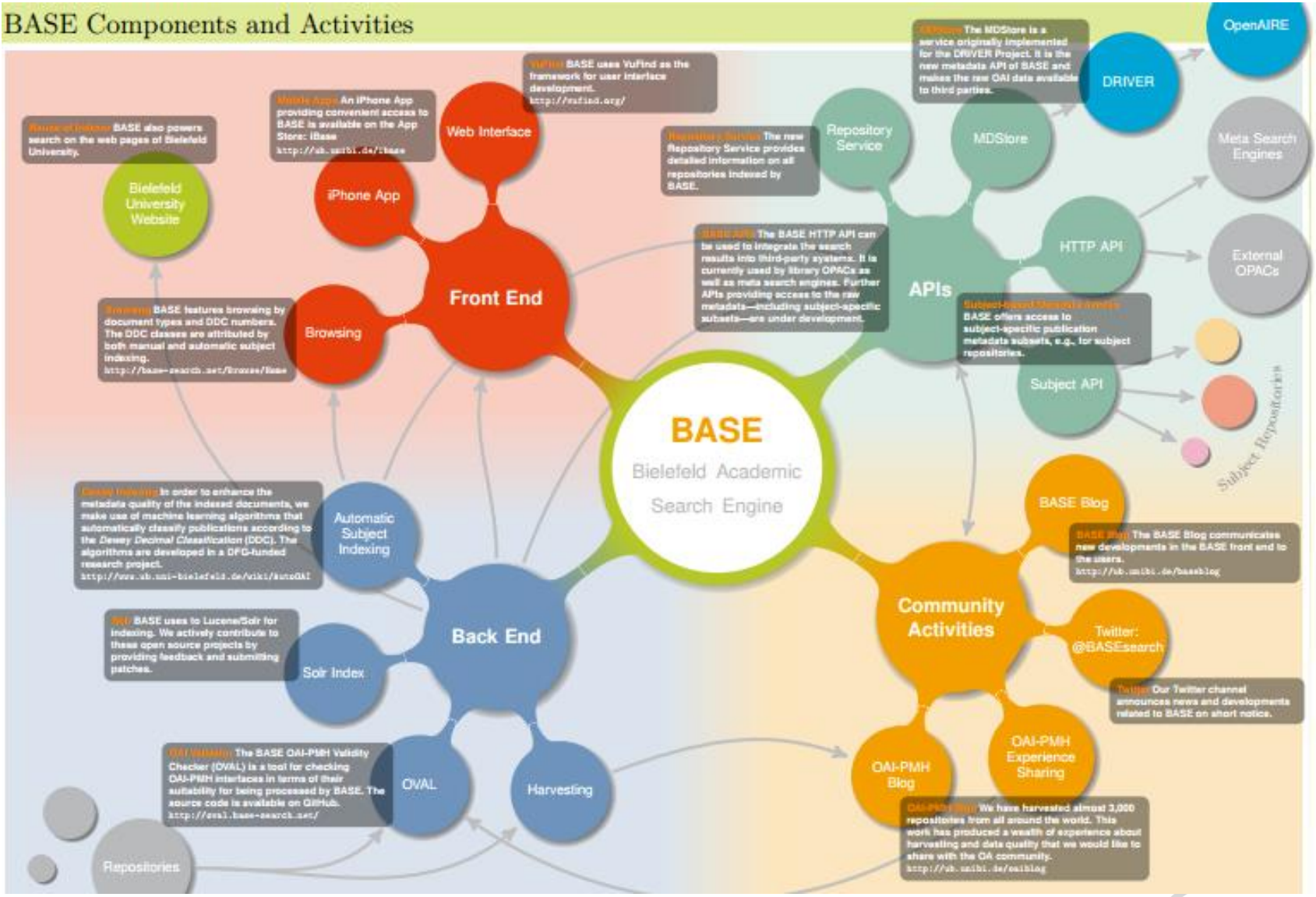
Basic Principles

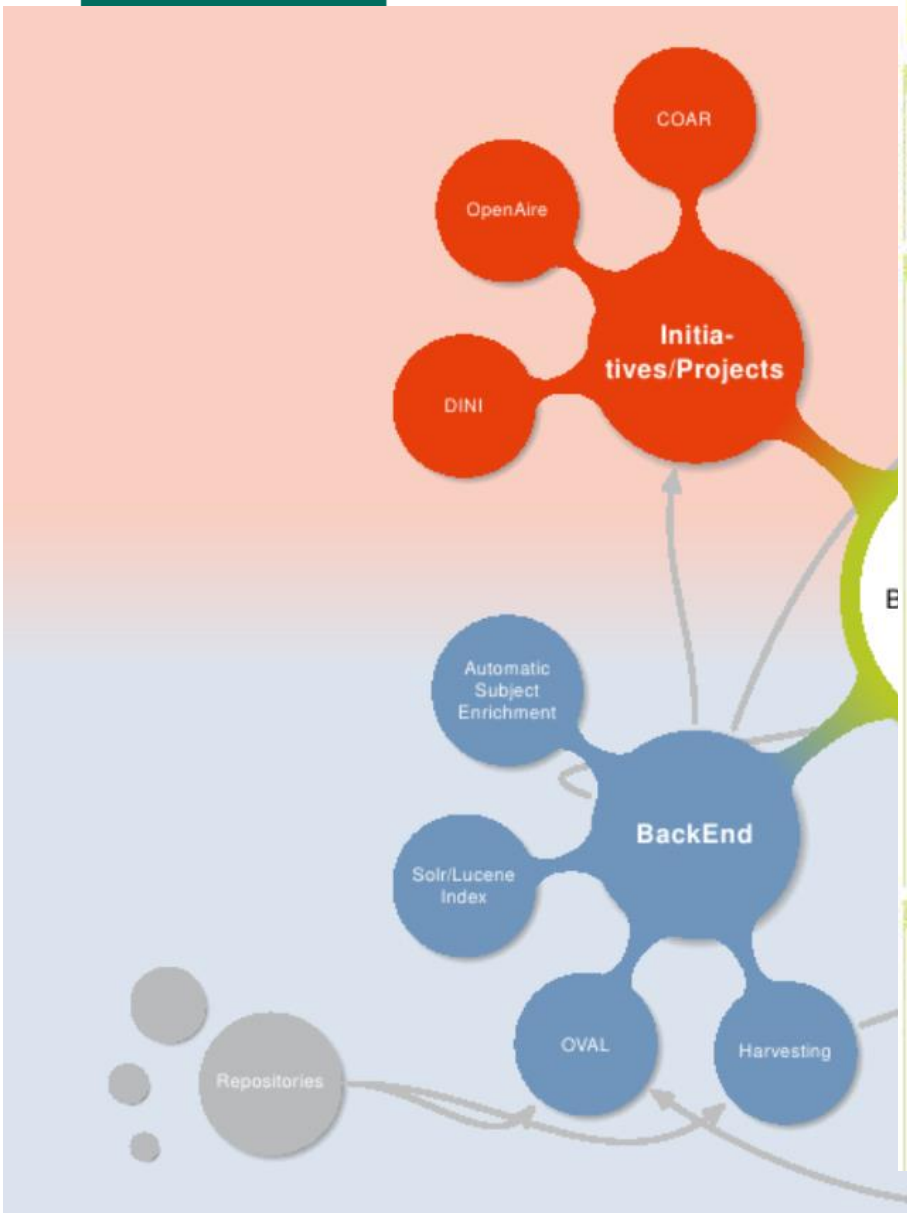


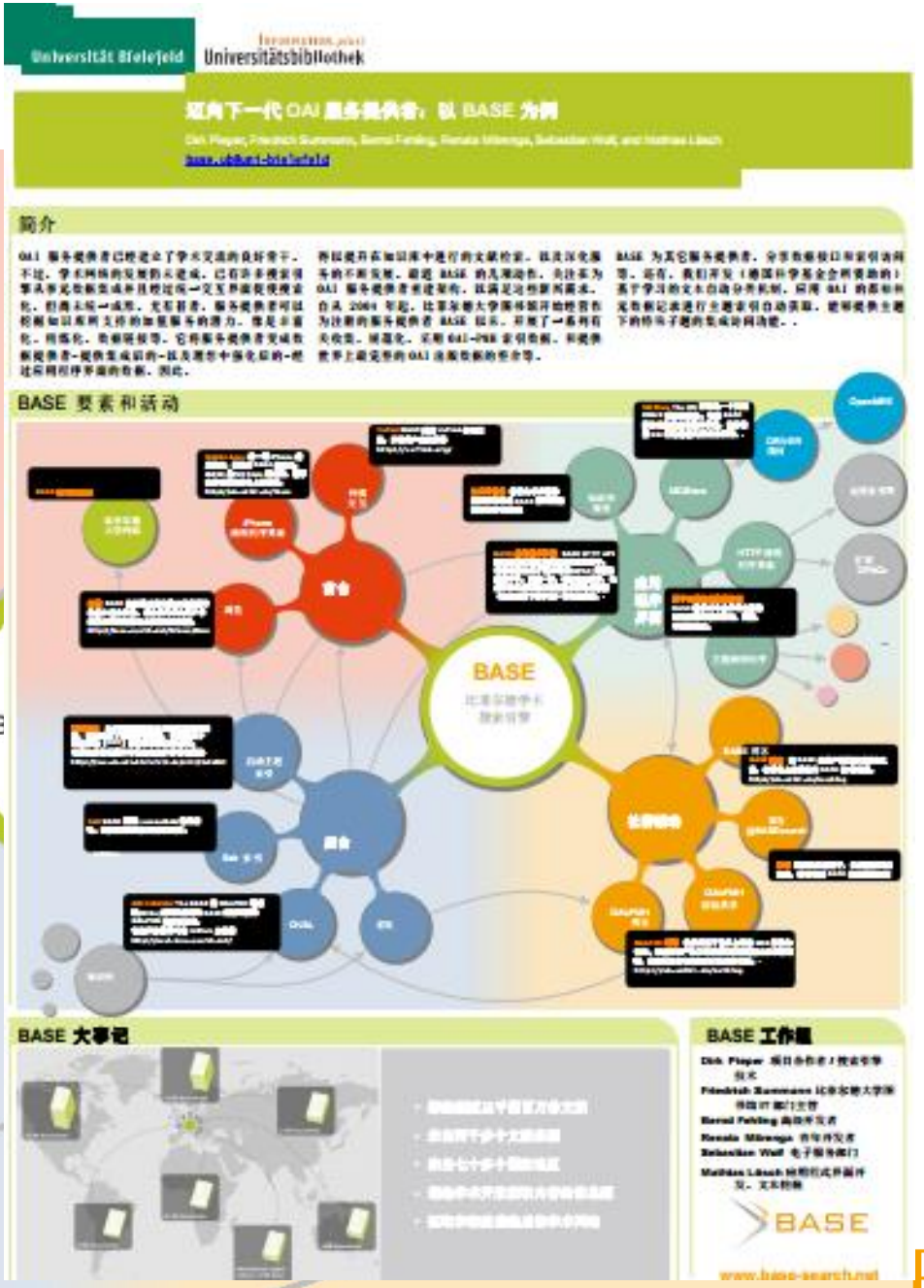
- **B**ibliographic Expertise
- **A**utomatic methods
- **S**imple Infrastructure
- **E**fficient Strategies

BASE

BASE Components and Activities




COAR
OpenAire
Initia-
tives/Projects
DINI
Automatic
Subject
Enrichment
BackEnd
Solr/Lucene
Index
OVAL
Harvesting
Repositories


Universität Bielefeld
Universitätsbibliothek
简介
BASE 要素和活动
BASE
BASE 大事记
BASE 工作组
BASE
www.base-search.net

BASE restrictions for including repositories

- OAI-PMH interface (available)
- OA is supported (at least some documents should be open accessible world-wide!)
- Valid DC Metadata (basic set of field contents)
- Operable URL for the document landing page

Main Issues to Avoid

- Wrong Document URLs
- Empty Records
- Invalid XML delivered
- Crashing Harvesting Processes
- Resumption Token Handling Fails
- No Incremental Delivery
- Changing OAI-PMH or homepage URLs without dissemination/redirect
- No OAI-PMH Interface Configuration

Main Issues Preferred

- OA Status delivered (on repository or document level)
- Metadata Guidelines compatible
  - Vocabularies used (for type, language, date, classification etc.)
- (Parallel) English end-user interface
- Citation/Abstract information delivered
- Available repository contact information
- Visible in Registries (OpenDOAR, openarchives …)

Universität Bielefeld

INFORMATION.plus!
Universitätsbibliothek

The **10** Biggest Misunderstandings around OAI-PMH

- OAI-PMH means ‚Everything is Open Access‘
- Persistent Identifiers are persistent
- Link to the Document page is not necessary
- Configuration is not needed
- Checking the Service is needless
- DublinCore is simple but sufficient
- OpenAccess Status Information is needless
- End-User Interface is not necessary
- Personal Email Adress is not needed
- Incremental Harvesting is sufficient

BASE

Universität Bielefeld

INFORMATION.plus! Universitätsbibliothek

Network Communication (example)

- Dear Sys Admin,

  we have harvested and indexed the OAI-PMH interface of your repository
  since longer.
  But when checking t
  problem:
  The oai_dc records (
  for example

BASE
Bielefeld Academic Search Engine

Mail-Prepper

- Std
- No records match
- Tomcat-Error
- Error ResumptionToken Handling
- Error NoOA Documents
- Error Empty Responses
- Handle not registered (DSpace)
- Handle not configured (DSpace)

Start

Dear Friedrich,
Thank you very mucl
me if it is ok for you.
an absolute one.
Regards
Stéphanie ….

BASE

Universität Bielefeld

INFORMATION.plus! Universitätsbibliothek

BASE Usage (May 2015)

Länder		Seiten	Zugriffe	Bytes
Germany	de	154,631	665,922	19.12 GB
China	cn	80,942	193,681	6.23 GB
United States	us	78,107	142,989	7.40 GB
France	fr	56,941	168,232	4.79 GB
Canada	ca	42,242	59,315	2.16 GB
Great Britain	gb	11,401	36,616	1.51 GB
Poland	pl	9,618	42,138	1.39 GB
Netherlands	nl	9,378	17,740	761.28 MB
Switzerland				
Austria				
Peru				
Spain				
Belgium				
Mexico				
Ukraine				
India				
Colombia				
Brazil				
Italy				
Argentina	ar	4,457	20,568	558.23 MB

Monat	Unterschiedliche Besucher	Anzahl der Besuche	Seiten	Zugriffe	Bytes
Jan 2015	39,030	75,098	670,532	2,108,535	67.32 GB
Feb 2015	37,677	72,774	645,761	1,976,945	63.28 GB
März 2015	44,295	86,937	805,643	2,375,123	81.41 GB
Apr 2015	44,017	84,034	882,180	2,416,771	86.39 GB
Mai 2015	42,028	81,204	584,379	1,961,255	59.67 GB
Juni 2015	290	308	1,368	4,624	123.26 MB
Juli 2015	0	0	0	0	0
Aug 2015	0	0	0	0	0
Sep 2015	0	0	0	0	0
Okt 2015	0	0	0	0	0
Nov 2015	0	0	0	0	0
Dez 2015	0	0	0	0	0
Total	207,337	400,355	3,589,863	10,843,253	358.18 GB

BASE

Overview

- Retrospect (2001 – 2014)
- Current Situation (2015)
- **Future prospective (2015 - …)**

Current developments

- Analysis of alternative (OAI-PMH) metadata formats
- Prototypical expansion of metadata workflows (use case: SCOAP3-repository)
- Increasing the OA percentage within BASE by detailed analysing of metadata (setSpecs, rights-information)
- Working on grant proposal with Helmholtz Association and German National Library for an ORCID Claiming Service
- Preparing resolving service for several identifiers

BASE Future Strategies

- Optimizing the Technical Platform
- Link Resolver Service
- Data Enrichment (Linked Open Data Strategies)
- Big Data Activities

Optimizing the Technical Platform

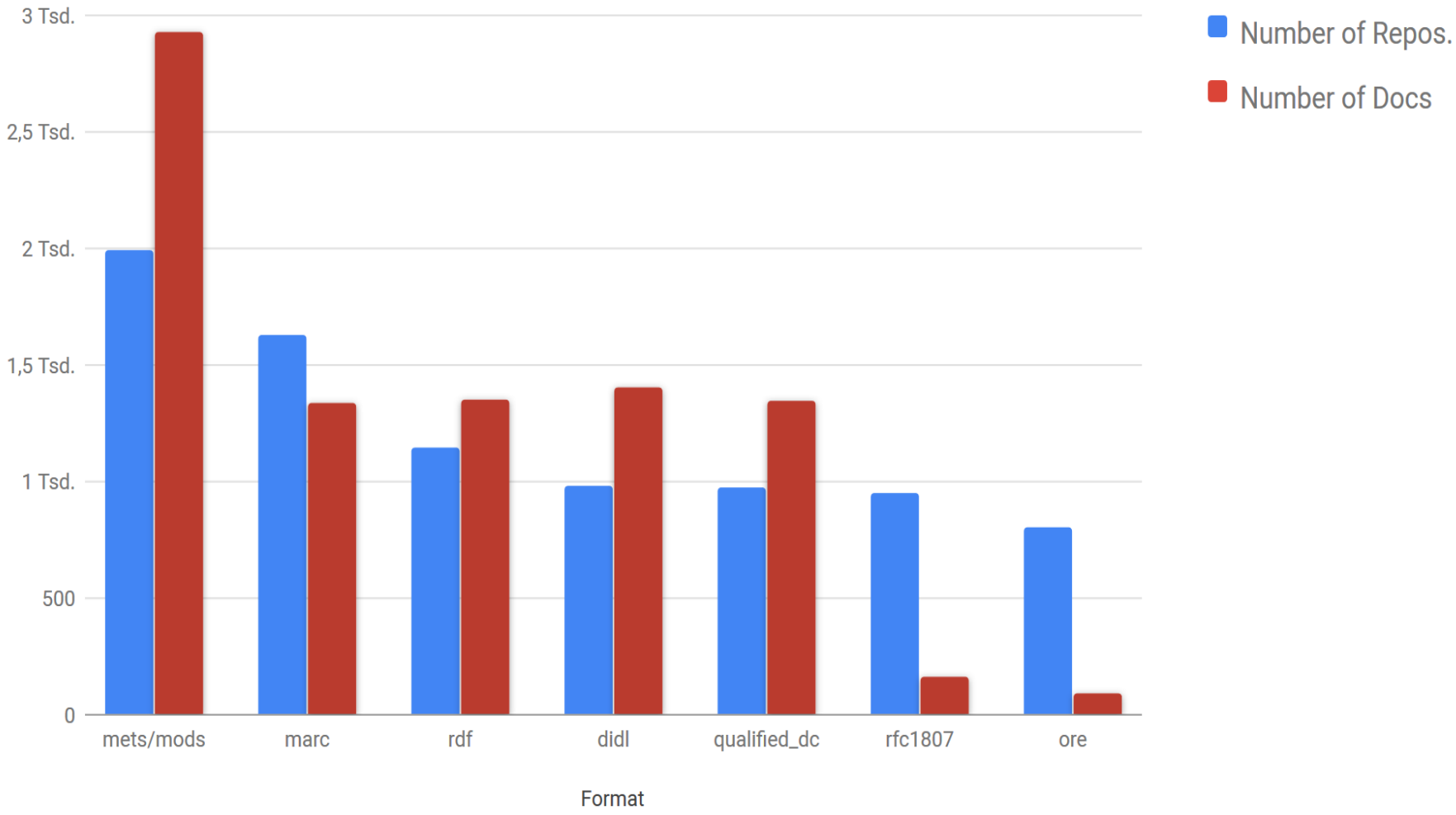
- Optimizing the VM Server Structure
- Distribution of Applications
- Optimizing the Lucene/Solr-Search Engine Environment
- Multi-node System

Data Enrichment (Linked Open Data Strategies)

Fundamental Aspect:

- More **Detailed Metadata Formats** have the potential to provide more detailed information
- But: It depends on the background quality

Metadata Formats in OAI-PMH Repositories



Link Resolver Service

- OAI Identifier
- DOI/Handle/ISSN/ISBN/URN/PMCID
- Author/Organization/Funder IDs
- Bibliographic Metadata

- Issue: Rights and Licences Normalization

SUBGoettinger	DE	ceu	181	Scan	**5x** http://creativecommons.org/licenses/by-sa/3.0 **5x** http://creativecommons.org/licenses/by-nc/2.5/ **5x** http://creativecommons.org/licenses/by-nc/2.0/ **4x** http://goedoc.uni-goettingen.de **4x** info:eu-repo/semantics/http://creativecommons.org/licenses/by-nd/3.0/de/ **4x** https://creativecommons.org/licenses/by/4.0 **4x** http://creativecommons.org/licenses/by/2.0/uk/legalcode **4x** http://creativecommons.org/licenses/by-nc-nd/3.0/deed.de **4x** http://creativecommons.org/licenses/by-nc/3.0/us/ **3x** http://iopscience.iop.org/1367-2630/14/10/103012/ **3x** http://creativecommons.org/licenses/by/2.5/http://creativecommons.org/license **3x** 719460905 **3x** 9 **3x** info:eu-repo/semantics/openAccess **3x** http://creativecommons.org/licenses/by-nc-nd/2.5/ch/deed.en **3x** http//creativecommons.org/licenses/by/2.0 **3x** http://creativecommons.org/licenses/by-nc-sa/3.0 **3x** http://creativecommons.org/licenses/by-nc/3.0/deed.en_US **3x** creativecommons.org/licenses/by-nc-sa/3.0/ **3x** info:eu-repo/semantics/http://creativecommons.org/licenses/by/2.5/ **3x** http://creativecommons.org/licenses/by-sa/3.0/de/deed.en **3x** http://creativecommons.org/licenses/by-sa/3.0/deed.de **2x** https://creativecommons.org/licenses/by-nc-nd/3.0/ **2x** http://creativecommons.org/licenses/by-nc/2.0/de/deed.en **2x** http://creativecommons.org/licenses/by-nc-nd/3.0/at/ **2x** http:// http://creativecommons.org/licenses/by/2.0 **2x** 594430

Big Data Activities

- Data Enrichment
  (Linked Open Data Strategies)
- Automatic Classification
- De-Duplication/Version Detection
- Fulltext Indexing

Universität Bielefeld

INFORMATION.plus!
Universitätsbibliothek

Thank you for your attention!

- friedrich.summann@uni-bielefeld.de
- dirk.pieper@uni-bielefeld.de

BASE