OAI 7 - TUTORIAL 5

Harvesters and subject based repositories

Geneva, 22nd June 2011

Friedrich Summann
Bielefeld University Library

Tutorial Overview

- Harvesting Techniques
- Harvesting and its challenges
- Supporting Tools
- Metadata Aggregation and Data Qualit**y**
- Harvesting Subject Repositories

OAI-PMH – a ten years success story

The starting point: OAI-PMH

- Protocol for Metadata Harvesting

2001 OAI-PMH 1.0

2002 OAI-PMH 2.0

2008 OAI-ORE

My Conclusion:

Harvesting is easy

But:

Putting things (results) together
is the real challenge

OAI-PMH in practice:

- Transparent
- Easy to Use (related to the protocol simplicity)
- Robust
- Broad variety

Harvesting Background

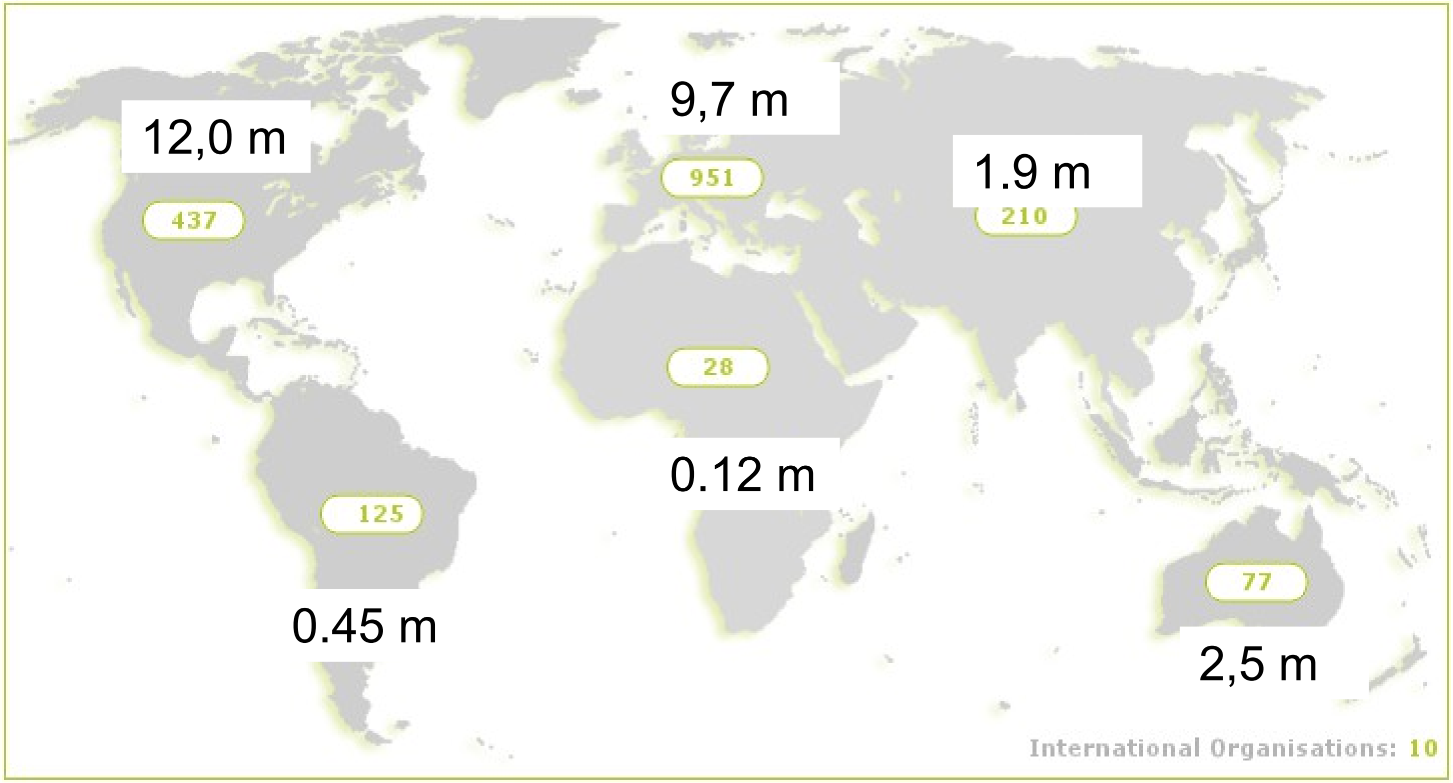
BASE (Bielefeld Academic Search Engine)
(around 2500 repositories)

DRIVER
(around 300 repositories)

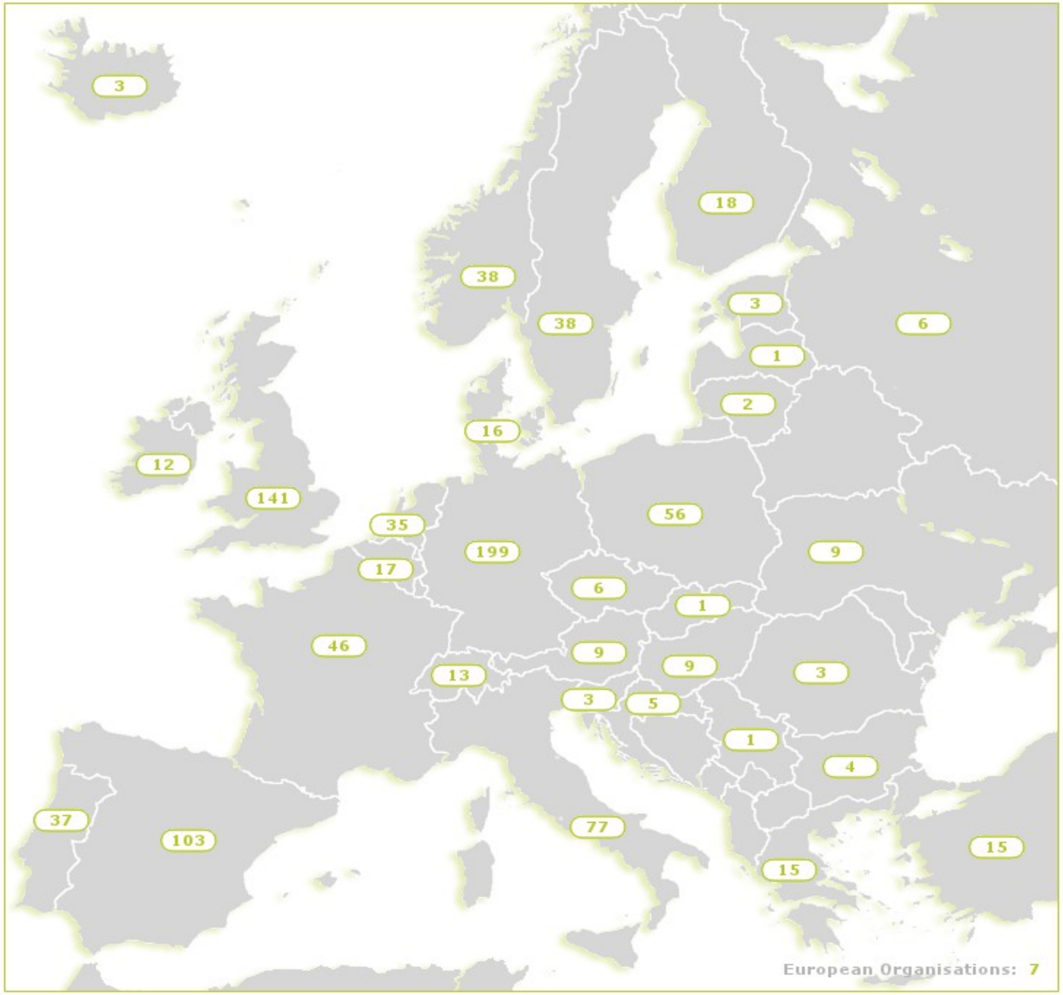
OpenAire
(in progress)

Repositories: Geographical Distribution (1)

BASE Repositories: (1837 in total)



Repositories: Geographical Distribution (2)



Harvester Technology

With your browser (manually)

Software packages

- PKP Harvester
- D-NET
- Other packages

Writing/Adopting your own harvester
(based on modules (CPAN for example)

Break: Harvesting in practice

The OAI verbs

<basicurl>?Identify
<basicurl>?ListSets
<basicurl>?ListMetadataFormats
<basicurl>?ListRecords

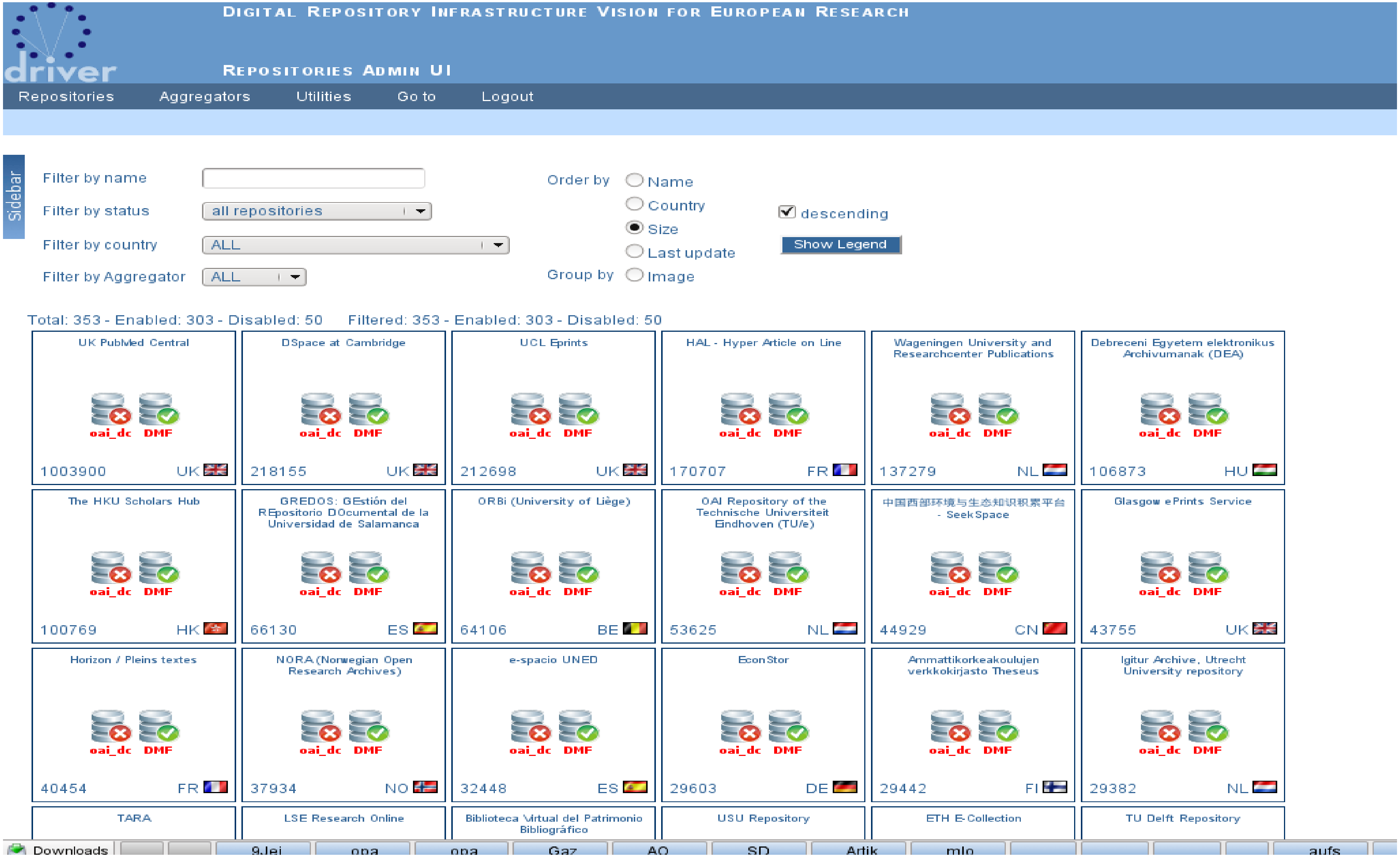
<basicurl>?ListIdentifiers
<basicurl>?GetIdentifier

Example:
http://repositories.ub.uni-bielefeld.de/escholarship/oai2/oai2.php

Harvesting in Practice: Command-Line

```
# Base URL [tab] Repository ID [tab] Repository Set [tab] Metadata Prefix [tab]
#http://edoc.hu-berlin.de/OAI-2.0      HUBerlin
##http://www.archive.org/services/oai2.php      InternetArchive collection:americana
#http://www.archive.org/services/oai2.php      InternetArchive collection:biodiversity
#http://rmrr.ro/index.php/rmrr/oai      JRMRR
#http://www.ammjournal.ro/index.php/AMM/oai     JAMM
#http://bibliography.ied.edu.hk:8080/TRSOAI/servlet/OAIDataProvider     HKEBD

wc
3179   6420 184430 harv_conf.txt

sh do_upd
Laden fuer UnivWesternCape-RR ergab 198 - 0
Laden fuer PolytechnicNamibia ergab 579 - 0 # (not responding)
Laden fuer CapePUT ergab 1221 - 0
Laden fuer AmericanUnivCairo ergab 1947 - 34
```

Harvesting in Practice: DRIVER



Filter by name

Filter by status: all repositories

Filter by country: ALL

Filter by Aggregator: ALL

Order by ○ Name ○ Country ● Size ○ Last update

☑ descending

Group by ○ Image

Total: 353 - Enabled: 303 - Disabled: 50   Filtered: 353 - Enabled: 303 - Disabled: 50

UK PubMed Central oai_dc DMF 1003900 UK	DSpace at Cambridge oai_dc DMF 218155 UK	UCL Eprints oai_dc DMF 212698 UK	HAL - Hyper Article on Line oai_dc DMF 170707 FR	Wageningen University and Researchcenter Publications oai_dc DMF 137279 NL	Debreceni Egyetem elektronikus Archivumanak (DEA) oai_dc DMF 106873 HU
The HKU Scholars Hub oai_dc DMF 100769 HK	GREDOS: GEstión del REpositorio DOcumental de la Universidad de Salamanca oai_dc DMF 66130 ES	ORBi (University of Liège) oai_dc DMF 64106 BE	OAI Repository of the Technische Universiteit Eindhoven (TU/e) oai_dc DMF 53625 NL	中国西部环境与生态知识积累平台 - SeekSpace oai_dc DMF 44929 CN	Glasgow ePrints Service oai_dc DMF 43755 UK
Horizon / Pleins textes oai_dc DMF 40454 FR	NORA (Norwegian Open Research Archives) oai_dc DMF 37934 NO	e-spacio UNED oai_dc DMF 32448 ES	EconStor oai_dc DMF 29603 DE	Ammattikorkeakoulujen verkkokirjasto Theseus oai_dc DMF 29442 FI	Igitur Archive, Utrecht University repository oai_dc DMF 29382 NL
TARA	LSE Research Online	Biblioteca Virtual del Patrimonio Bibliográfico	USU Repository	ETH E-Collection	TU Delft Repository

Tutorial Overview

- Harvesting Techniques
- **Harvesting and its challenges**
- Supporting Tools
- Metadata Aggregation and Data Quality
- Harvesting Subject Repositories

Harvesting : Challenges and pitfalls

- Repository does not respond (temporarily, specific verbs)
- Results are not xml-valid
- Harvesting breaks (especially big reps)
- Incremental Harvesting does not work
- No deleting information, added records
- Variety of Field Contents
- Change of behavior (basicurl, contents)
- Metadata point to reference or citation only
- Link to Document is not operable
- Fulltext access is restricted

Tutorial Overview

- Harvesting Techniques
- Harvesting and its challenges
- **Supporting Tools**
- Metadata Aggregation and Data Quality
- Harvesting Subject Repositories

Related to Harvesting: Questions to answer

- What to harvest?
- How to harvest?
- How to aggregate?

Registries (1)

Universal Registries

http://www.openarchives.org/Register/BrowseSites.pl (OpenArchives Initiative)

http://gita.grainger.uiuc.edu/registry/ (Univ. of Illinois)

http://roar.eprints.org/ (Eprints-Registry)

http://www.opendoar.org/ (OpenDOAR)

Registries (2) :Communities

- http://wiki.dspace.org/OaiInstallations (DSpace)
- http://pkp.sfu.ca/ojs-journals (OJS)
- http://digitalcommons.bepress.com/subscriber_gallery/all.html (Digital Commons)
- http://www.oclc.org/contentdm/collections/ (ContentDM)

Registries (3): National

Germany (DINI) http://www.dini.de/dini-zertifikat/liste-der-repositorien/

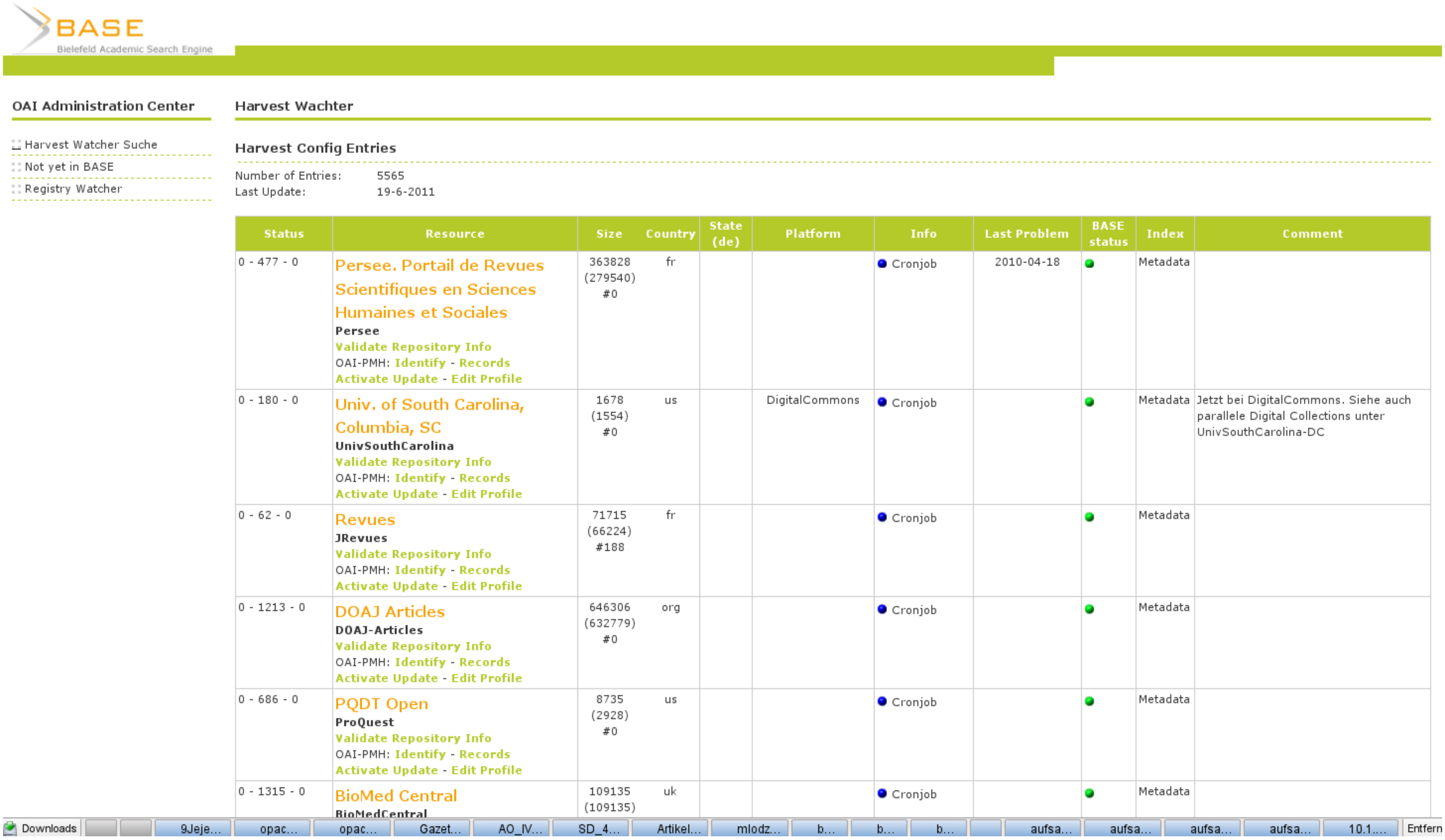
Netherlands (DARE/Narcis)
http://www.narcis.nl/repositories/Language/en

Spain (Hispana)
http://hispana.mcu.es/es/comunidades/directorio.cmd

Italy (Pleiade)
http://www.openarchives.it/pleiadi/

Poland (Dlibra)
http://dlibra.psnc.pl/index.php?option=com_content&task=view&id=12&Itemid=27&lang=pl

BASE Tools: Harvest Watcher



OAI Administration Center

- Harvest Watcher Suche
- Not yet in BASE
- Registry Watcher

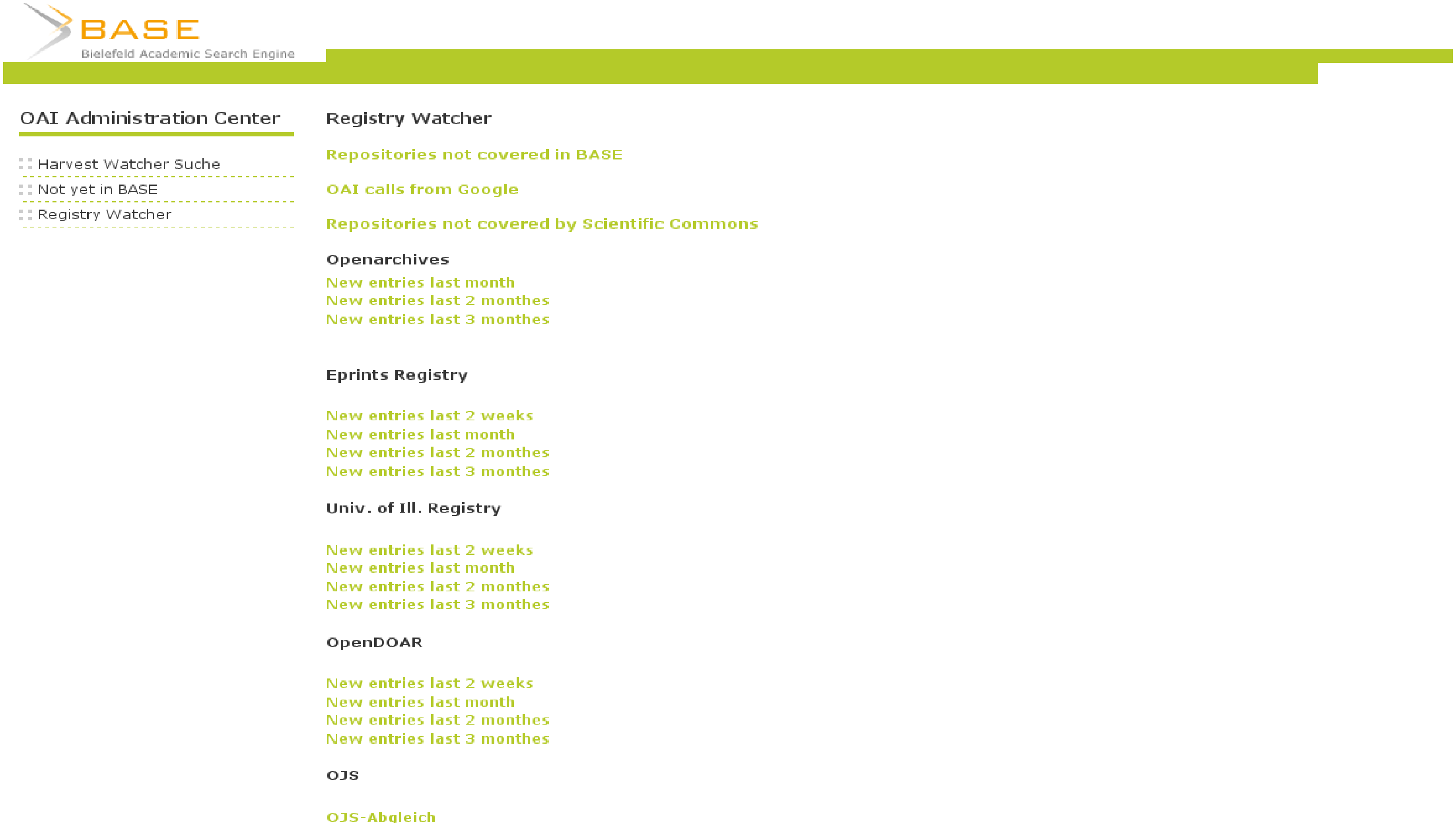
Harvest Wachter

Harvest Config Entries

Number of Entries: 5565
Last Update: 19-6-2011

Status	Resource	Size	Country	State (de)	Platform	Info	Last Problem	BASE status	Index	Comment
0 - 477 - 0	Persee. Portail de Revues Scientifiques en Sciences Humaines et Sociales Persee Validate Repository Info OAI-PMH: Identify - Records Activate Update - Edit Profile	363828 (279540) #0	fr			● Cronjob	2010-04-18	●	Metadata	
0 - 180 - 0	Univ. of South Carolina, Columbia, SC UnivSouthCarolina Validate Repository Info OAI-PMH: Identify - Records Activate Update - Edit Profile	1678 (1554) #0	us		DigitalCommons	● Cronjob		●	Metadata	Jetzt bei DigitalCommons. Siehe auch parallele Digital Collections unter UnivSouthCarolina-DC
0 - 62 - 0	Revues JRevues Validate Repository Info OAI-PMH: Identify - Records Activate Update - Edit Profile	71715 (66224) #188	fr			● Cronjob		●	Metadata	
0 - 1213 - 0	DOAJ Articles DOAJ-Articles Validate Repository Info OAI-PMH: Identify - Records Activate Update - Edit Profile	646306 (632779) #0	org			● Cronjob		●	Metadata	
0 - 686 - 0	PQDT Open ProQuest Validate Repository Info OAI-PMH: Identify - Records Activate Update - Edit Profile	8735 (2928) #0	us			● Cronjob		●	Metadata	
0 - 1315 - 0	BioMed Central BioMedCentral	109135 (109135)	uk			● Cronjob		●	Metadata	

BASE Tools: Registry Watcher



BASE
Bielefeld Academic Search Engine

OAI Administration Center

- Harvest Watcher Suche
- Not yet in BASE
- Registry Watcher

Registry Watcher

Repositories not covered in BASE

OAI calls from Google

Repositories not covered by Scientific Commons

Openarchives

New entries last month
New entries last 2 monthes
New entries last 3 monthes

Eprints Registry

New entries last 2 weeks
New entries last month
New entries last 2 monthes
New entries last 3 monthes

Univ. of Ill. Registry

New entries last 2 weeks
New entries last month
New entries last 2 monthes
New entries last 3 monthes

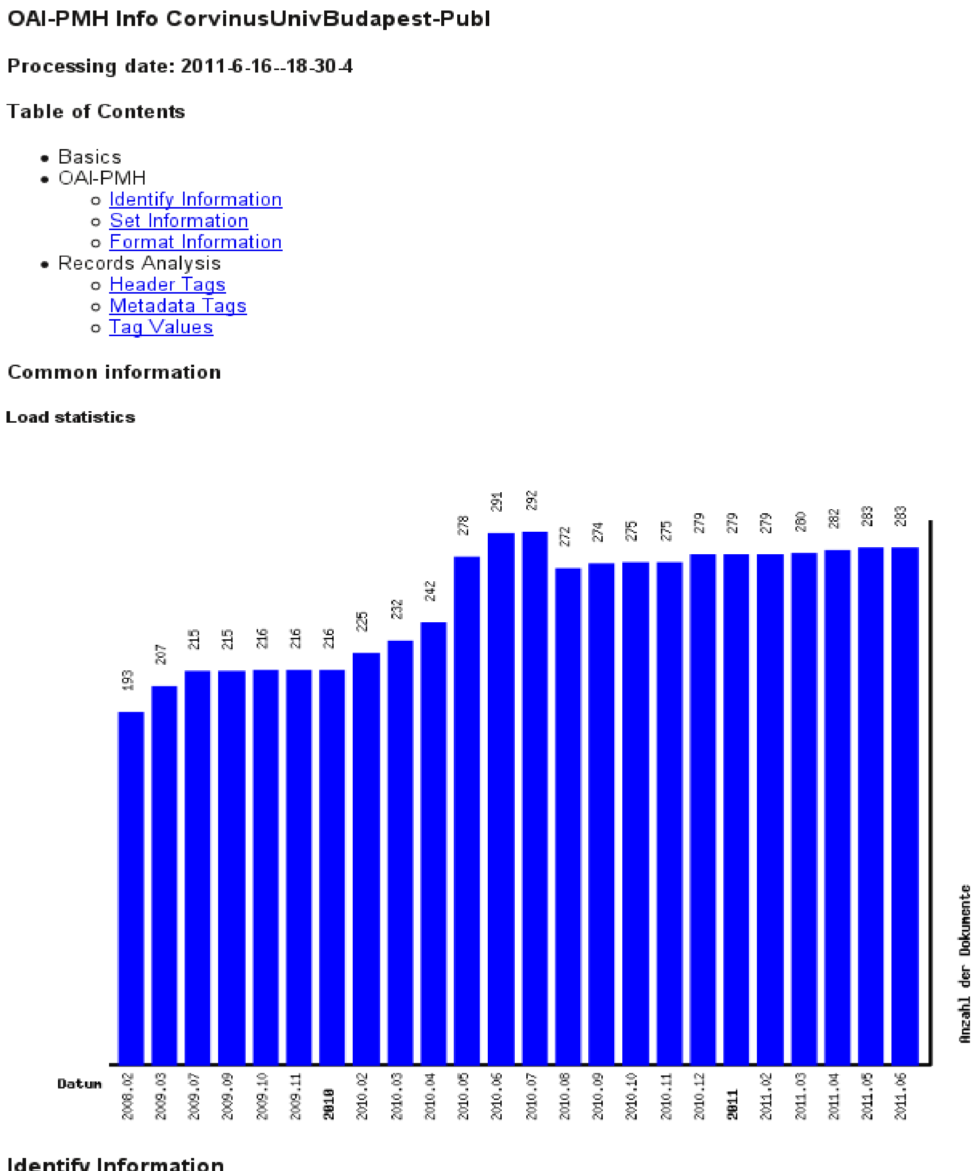
OpenDOAR

New entries last 2 weeks
New entries last month
New entries last 2 monthes
New entries last 3 monthes

OJS

OJS-Abgleich

BASE Tools: Repository Analyzer



OAI-PMH Info CorvinusUnivBudapest-Publ

Processing date: 2011-6-16--18-30-4

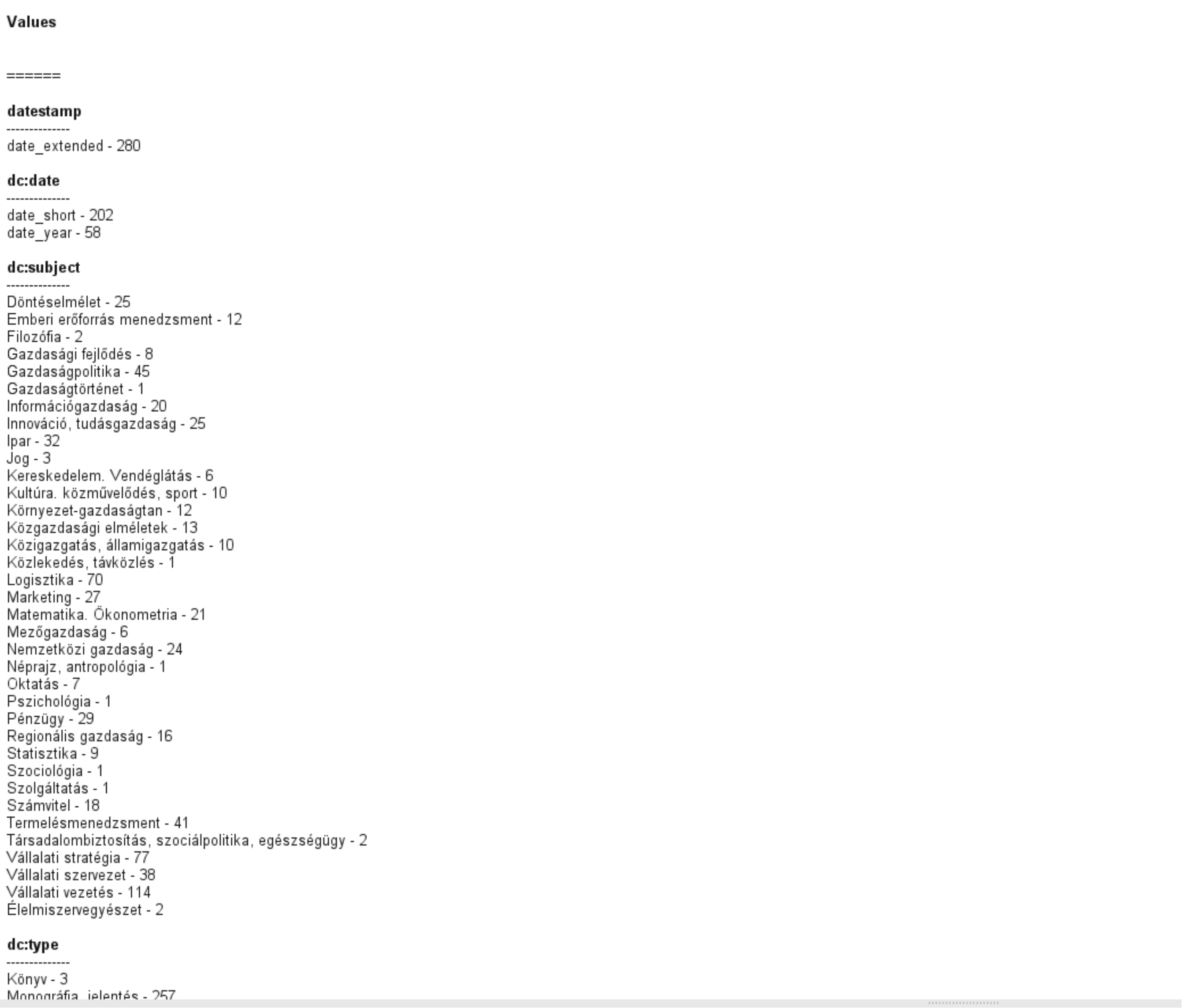
Table of Contents

Common information

Load statistics

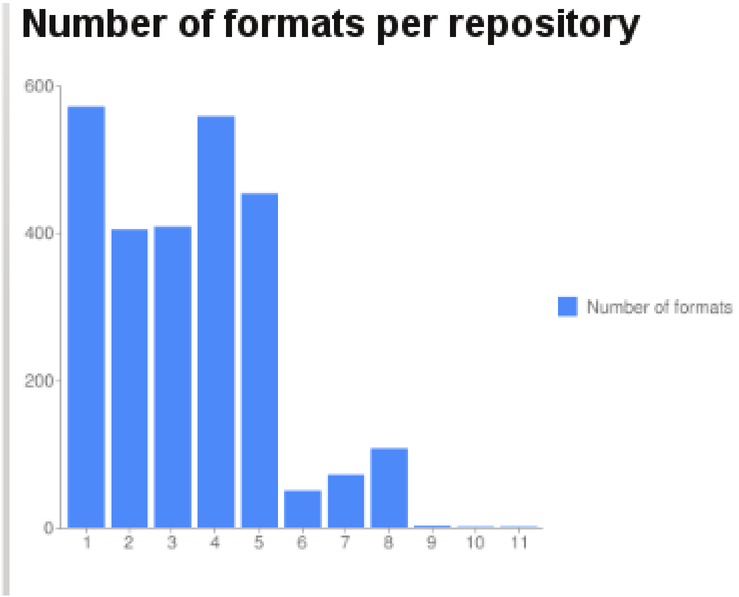
Identify Information

BASE Tools: Repository Analyzer (2)

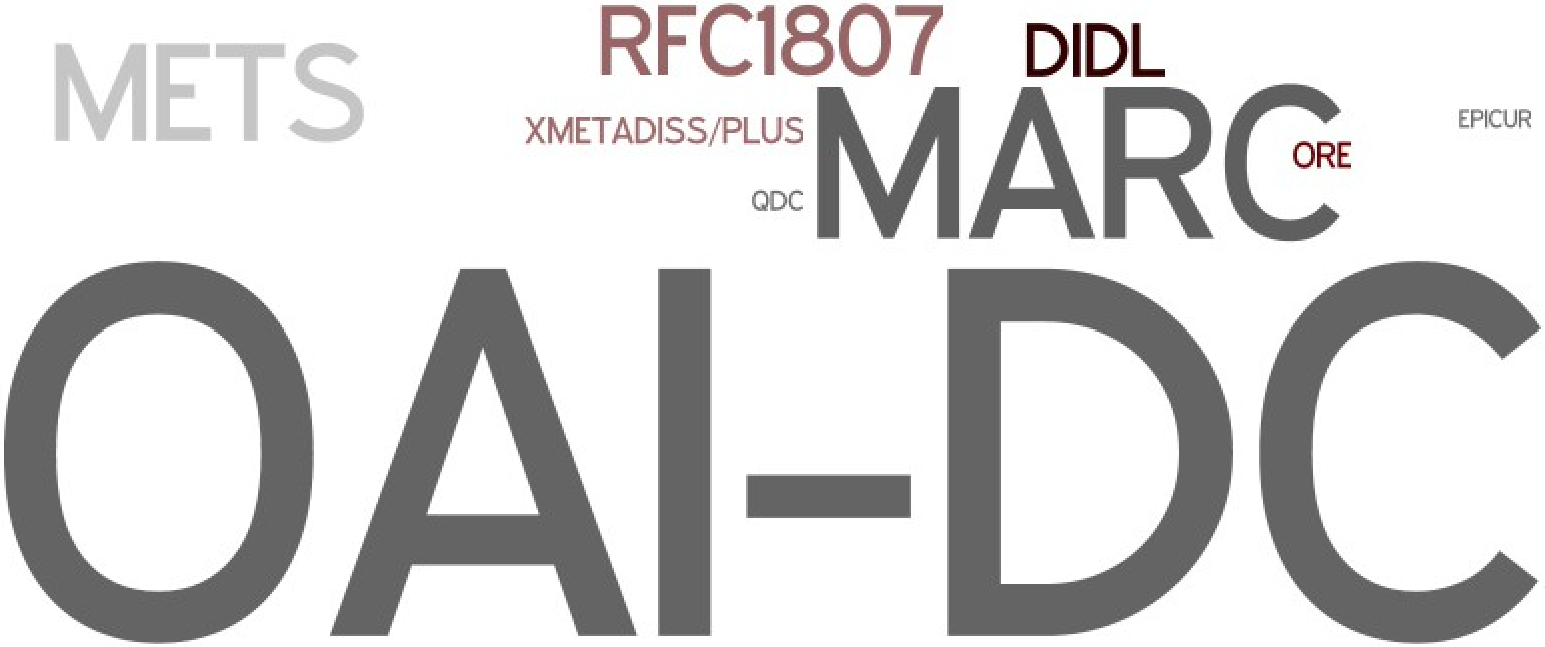


```
Values

======

datestamp
--------------
date_extended - 280

dc:date
--------------
date_short - 202
date_year - 58

dc:subject
--------------
Döntéselmélet - 25
Emberi erőforrás menedzsment - 12
Filozófia - 2
Gazdasági fejlődés - 8
Gazdaságpolitika - 45
Gazdaságtörténet - 1
Információgazdaság - 20
Innováció, tudásgazdaság - 25
Ipar - 32
Jog - 3
Kereskedelem. Vendéglátás - 6
Kultúra. közművelődés, sport - 10
Környezet-gazdaságtan - 12
Közgazdasági elméletek - 13
Közigazgatás, államigazgatás - 10
Közlekedés, távközlés - 1
Logisztika - 70
Marketing - 27
Matematika. Ökonometria - 21
Mezőgazdaság - 6
Nemzetközi gazdaság - 24
Néprajz, antropológia - 1
Oktatás - 7
Pszichológia - 1
Pénzügy - 29
Regionális gazdaság - 16
Statisztika - 9
Szociológia - 1
Szolgáltatás - 1
Számvitel - 18
Termelésmenedzsment - 41
Társadalombiztosítás, szociálpolitika, egészségügy - 2
Vállalati stratégia - 77
Vállalati szervezet - 38
Vállalati vezetés - 114
Élelmiszervegyészet - 2

dc:type
--------------
Könyv - 3
Monográfia, jelentés - 257
```

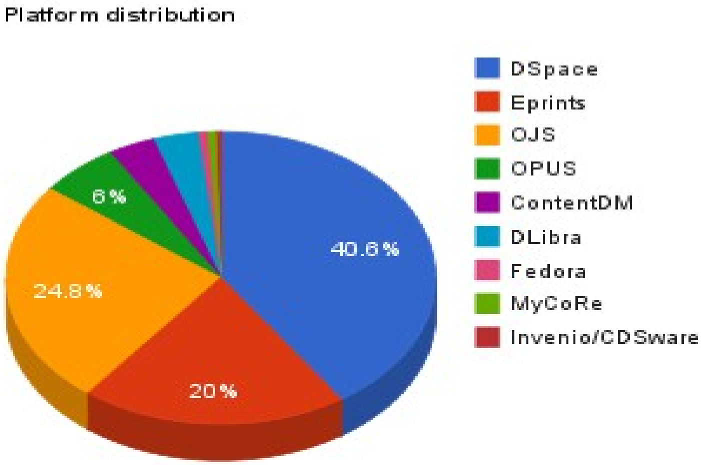
Metadataformats



OAI-PMH Metadata Formats



Repository Platform Software: Distribution

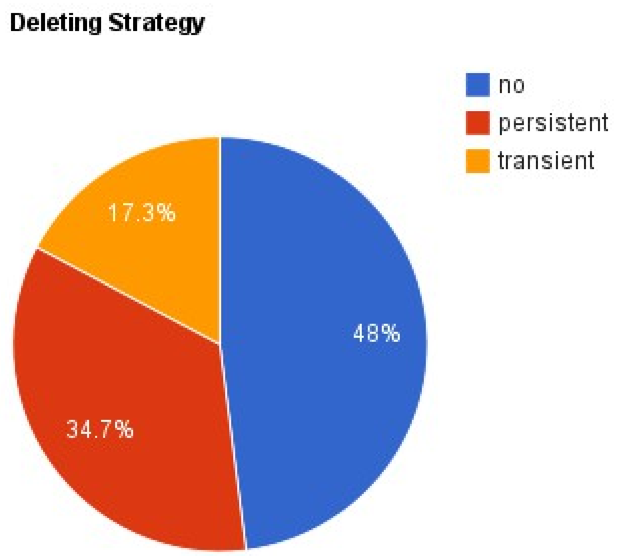


Repositories characteristics

Deleting Strategy

- no
- persistent
- transient

Deleting strategy: DIstribution

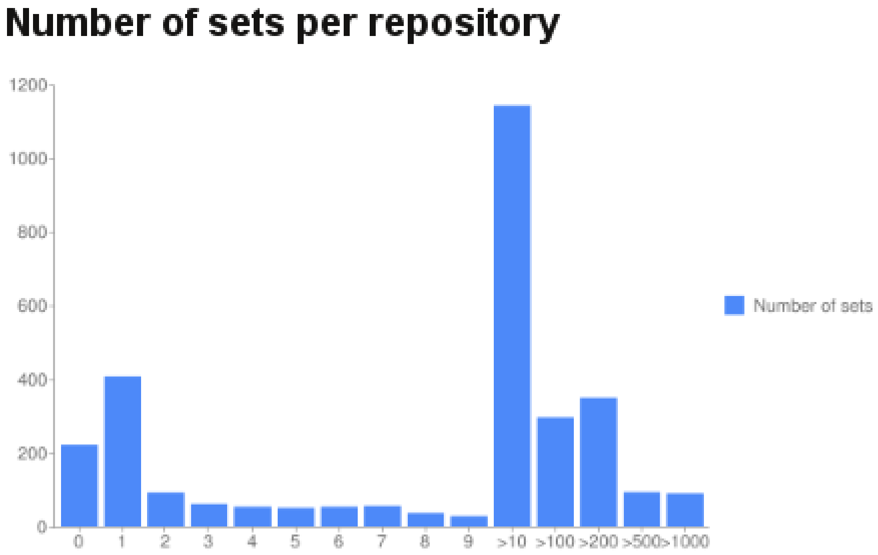


Set Definitions

Set Definitions: derived from ListSets (April 2011)

Total number: 464106

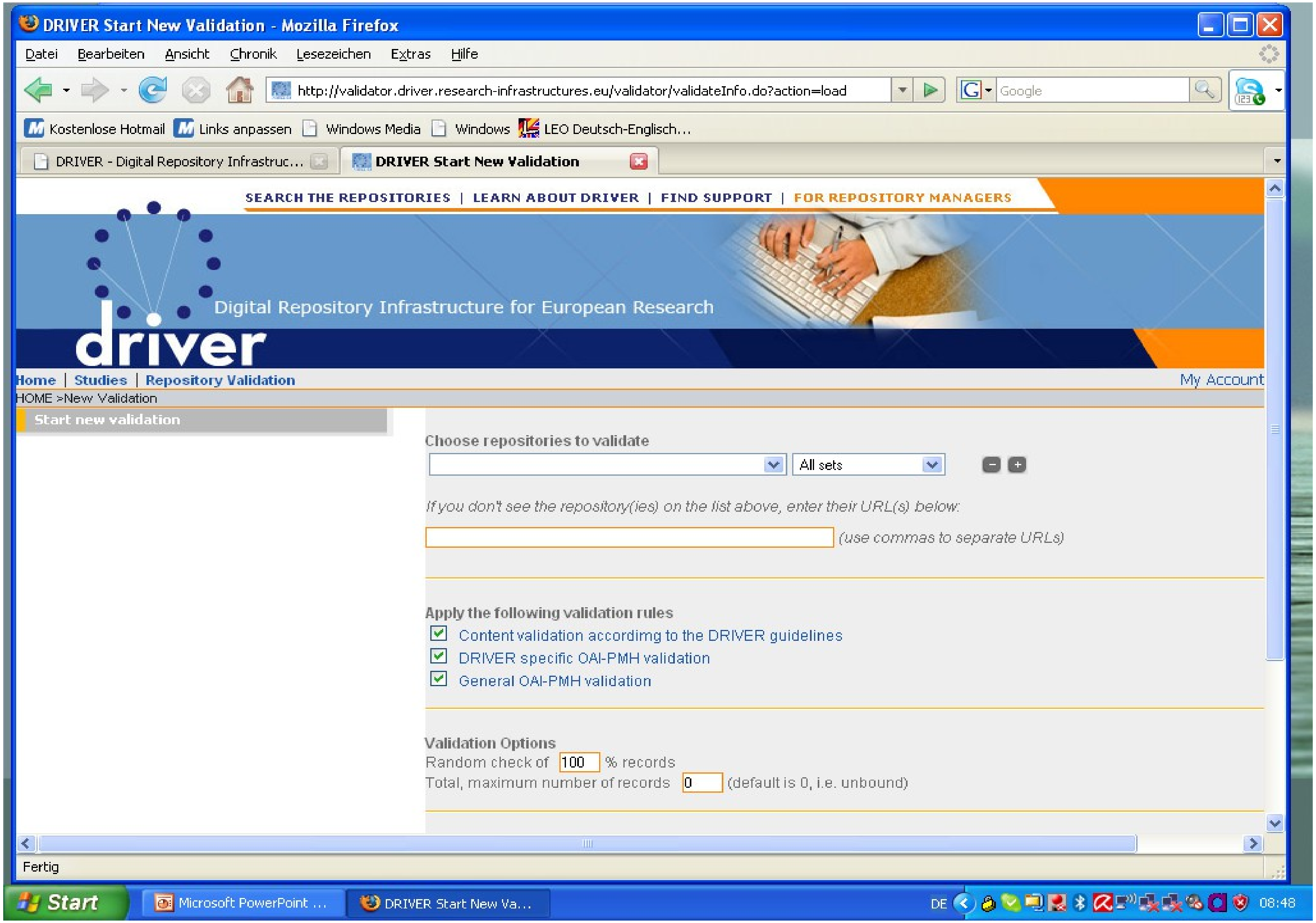
0 - 197
1 - 363
2 - 82
3 - 55
4 - 48
5 - 45
6 - 48
7 - 50
8 - 33
9 - 25
10 – 99 - 1016
100 – 199 - 264
200 – 499 - 311
500 – 999 - 83
>1000 - 80



Tools: Validator Services

- Openarchives Validation
- Repository Explorer (http://re.cs.uct.ac.za/)
- DRIVER Validator (http://validator.driver.research-infrastructures.eu/pages/validatorHome.jsp)
- BASE OAI-PMH Validity Checker

Tools: DRIVER Validator


DRIVER Start New Validation - Mozilla Firefox
Datei Bearbeiten Ansicht Chronik Lesezeichen Extras Hilfe
http://validator.driver.research-infrastructures.eu/validator/validateInfo.do?action=load
Google
Kostenlose Hotmail
Links anpassen
Windows Media
Windows
LEO Deutsch-Englisch...
DRIVER - Digital Repository Infrastruc...
DRIVER Start New Validation
SEARCH THE REPOSITORIES | LEARN ABOUT DRIVER | FIND SUPPORT | FOR REPOSITORY MANAGERS
Digital Repository Infrastructure for European Research
driver
Home | Studies | Repository Validation
My Account
HOME >New Validation
Start new validation
Fertig
Start
Microsoft PowerPoint ...
DRIVER Start New Va...
DE
08:48

Choose repositories to validate

All sets

If you don't see the repository(ies) on the list above, enter their URL(s) below:

(use commas to separate URLs)

Apply the following validation rules

☑ Content validation accordimg to the DRIVER guidelines

☑ DRIVER specific OAI-PMH validation

☑ General OAI-PMH validation

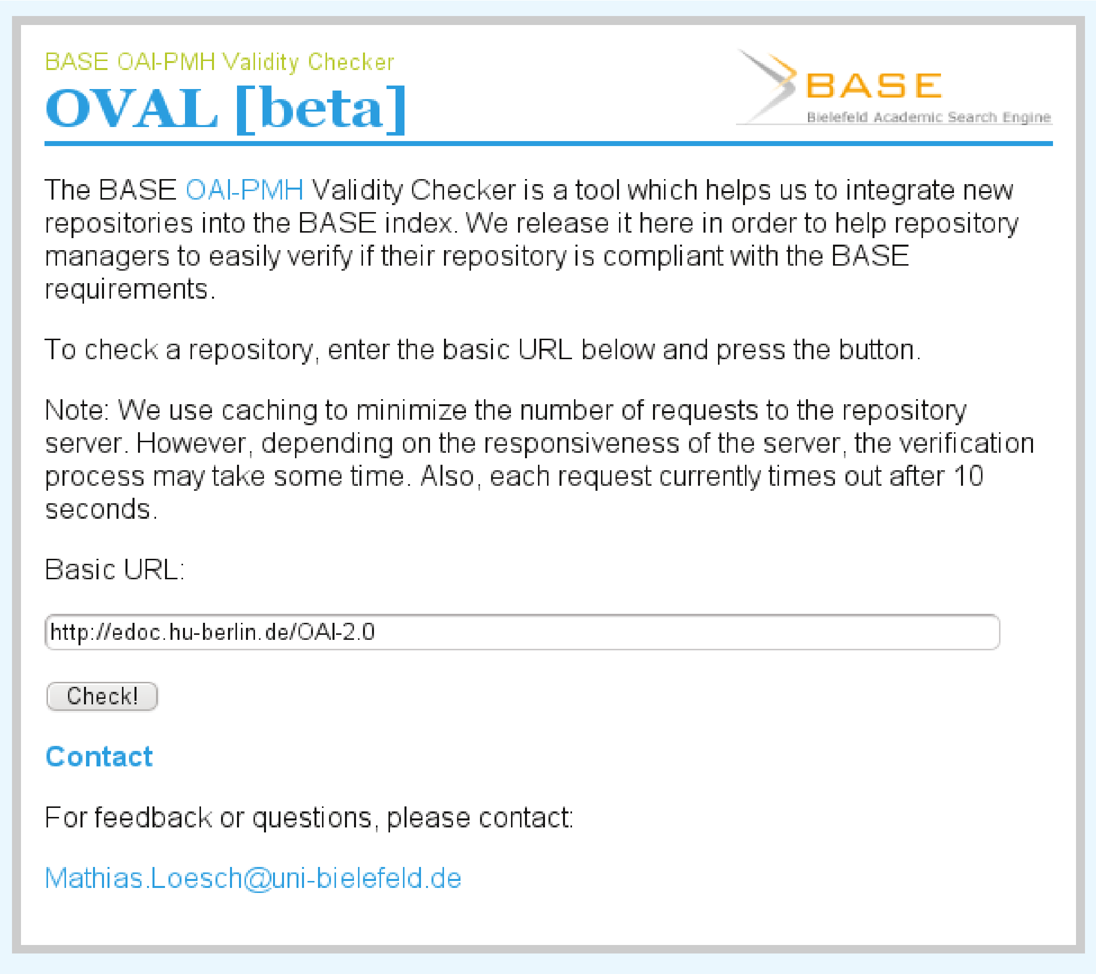
Validation Options

Random check of 100 % records

Total, maximum number of records 0 (default is 0, i.e. unbound)

Tools: BASE OAI-PMH Validity Checker

http://oval.base-search.net/oval



BASE OAI-PMH Validity Checker

OVAL [beta]

The BASE OAI-PMH Validity Checker is a tool which helps us to integrate new repositories into the BASE index. We release it here in order to help repository managers to easily verify if their repository is compliant with the BASE requirements.

To check a repository, enter the basic URL below and press the button.

Note: We use caching to minimize the number of requests to the repository server. However, depending on the responsiveness of the server, the verification process may take some time. Also, each request currently times out after 10 seconds.

Basic URL:

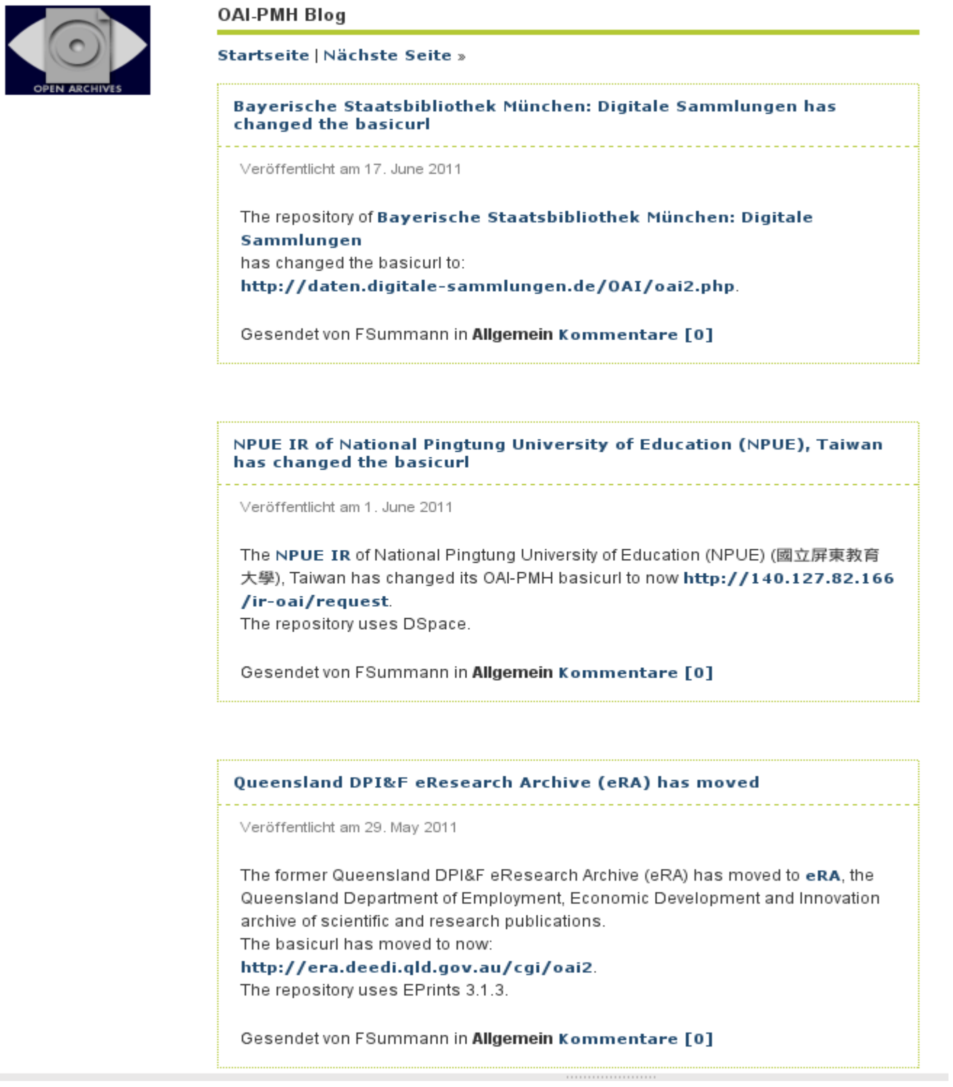
http://edoc.hu-berlin.de/OAI-2.0

Contact

For feedback or questions, please contact:

Mathias.Loesch@uni-bielefeld.de

Tools: OAI-PMH Blog



OAI-PMH Blog

Startseite | Nächste Seite »

Bayerische Staatsbibliothek München: Digitale Sammlungen has changed the basicurl

Veröffentlicht am 17. June 2011

The repository of **Bayerische Staatsbibliothek München: Digitale Sammlungen**
has changed the basicurl to:
http://daten.digitale-sammlungen.de/OAI/oai2.php.

Gesendet von FSummann in **Allgemein** **Kommentare [0]**

NPUE IR of National Pingtung University of Education (NPUE), Taiwan has changed the basicurl

Veröffentlicht am 1. June 2011

The **NPUE IR** of National Pingtung University of Education (NPUE) (國立屏東教育大學), Taiwan has changed its OAI-PMH basicurl to now **http://140.127.82.166/ir-oai/request**.
The repository uses DSpace.

Gesendet von FSummann in **Allgemein** **Kommentare [0]**

Queensland DPI&F eResearch Archive (eRA) has moved

Veröffentlicht am 29. May 2011

The former Queensland DPI&F eResearch Archive (eRA) has moved to **eRA**, the Queensland Department of Employment, Economic Development and Innovation archive of scientific and research publications.
The basicurl has moved to now:
http://era.deedi.qld.gov.au/cgi/oai2.
The repository uses EPrints 3.1.3.

Gesendet von FSummann in **Allgemein** **Kommentare [0]**

Tutorial Overview

- Harvesting Techniques
- Harvesting and its challenges
- Supporting Tools
- **Metadata Aggregation and Data Quality**
- Harvesting Subject Repositories

dc:language: Variety of MetadataValues

Analysis: European Repositories, Oct. 2009
804 different values in 4720585 tags

Top values

en – 1385175
eng – 511085
spa – 345658
de – 319937
en_GB - 178381
ger – 166587
eng; - 102678
FR – 95798
…
I

; - 3
? - 3
at;deu - 2
enm;eng - 2
FRA – 2
fr_BE - 2
Andere Sprache – 2
cat, spa, fra, eng. - 2

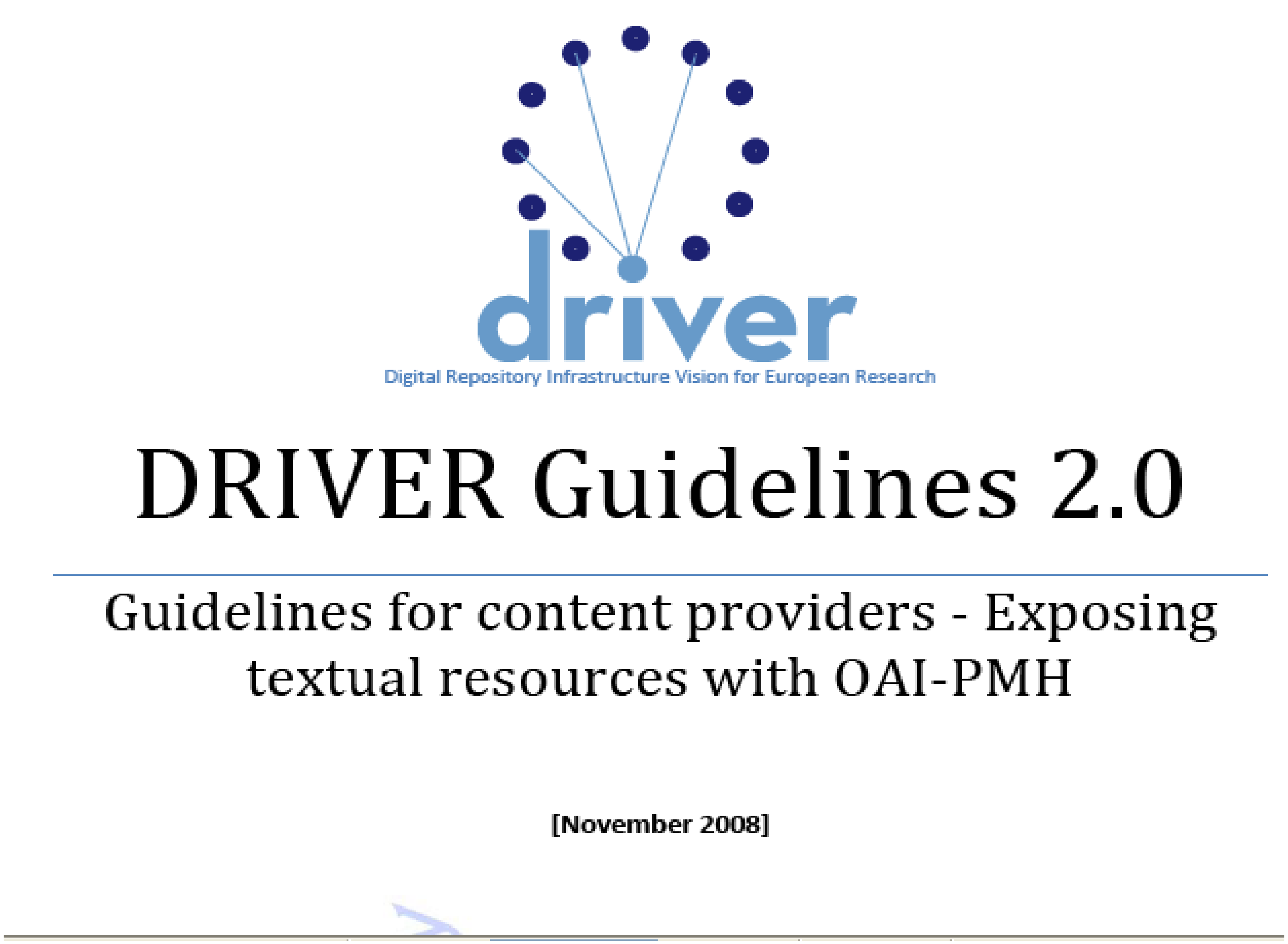
dc:type: Variety of Metadata Values

Analysis: German Repositories, Sept. 2009
2772 different values in 1394089 tags

Top values

Dataset – 588525
Artikel – 192306
Rezension – 113924
Text – 73210
Text.Thesis.Doctoral – 30201
Article – 29278
Miszelle – 27060
NonPeerReviewed – 24688
ResearchPaper – 16046
Dissertation - 15531
…
I

Software - 7
Kulturkarten - 7
Composition - 7
Interactive Resource - 4
Interview – 3
Media - 1
content analysis – 1
Anniversary Publication – 1
qualitative research -1



driver
Digital Repository Infrastructure Vision for European Research

DRIVER Guidelines 2.0

Guidelines for content providers - Exposing textual resources with OAI-PMH

[November 2008]

http://www.driver-support.eu/documents/DRIVER_Guidelines_v2_Final_2008-11-13.pdf

DRIVER Guidelines Recommendations (1)

- Deleting strategy
  - (persistent, transient)
- Set for OA documents
  - (driver)
- Batch size
  - (100 - 500 records per response)
- Resumption Token life span
  - (at least 24 hours)

DRIVER Guidelines Recommendations (2)

- Fulltext link
- Content Recommendations
  - identifier
  - creator
  - contributor
  - source (citation)
  - ddc as classification
- Standardized Contents
  - type
  - language
  - date

Tutorial Overview

- Harvesting Techniques
- Harvesting and its challenges
- Supporting Tools
- Metadata Aggregation and Data Quality
- **Harvesting Subject Repositories**

Subject Repositories: Registries

- Disciplinary repositories
  http://oad.simmons.edu/oadwiki/Disciplinary_repositories

- OpenDOAR

Subject Repositories

The Big Ones:

arXiv.org (Physics)
CERN Document Server (Physics)
PubMed Central (Life Sciences)
CiteSeer (Computer Science)
ELIS (Library Science)
REPEC (Economics)
EconStor (Economics)
SSOAR (Social Sciences)
. . .

Subject Repositories: Strategies

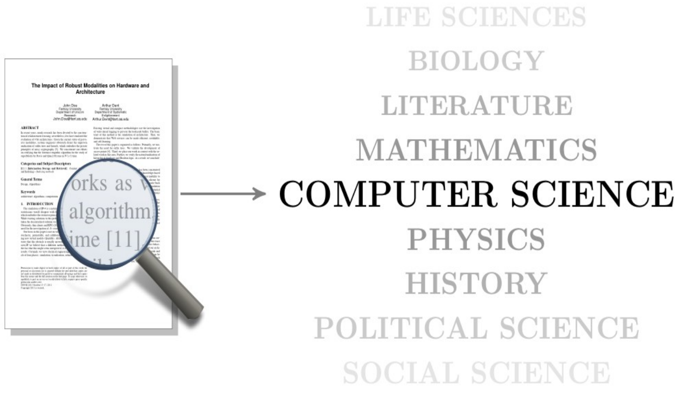
- Harvesting discipline repositories
- Filtering discipline-related documents from universal repositories
- Merging those two approaches
- Building subject-orientated services

dc:subjects values

dc:subject 8,585407 entries (all repositories, Nov. 2009)

1398907: Articles
1072722: Research Article
285041: Astrophysics
226740: High Energy Physics - Phenomenology
211920: Correspondence
207338: High Energy Physics - Theory
198582: Mathematics
139611: 33-00
139611: 00A22
138549: Article
130679: Book Review
124776: Functions
98114: Physics
92072: General Relativity and Quantum Cosmology
88285: Quantum Physics
84088: 26A09
81981: 33B10
78108: Elementary Functions
74320: Science
…
32451: 610

Approach: Automatic Classification

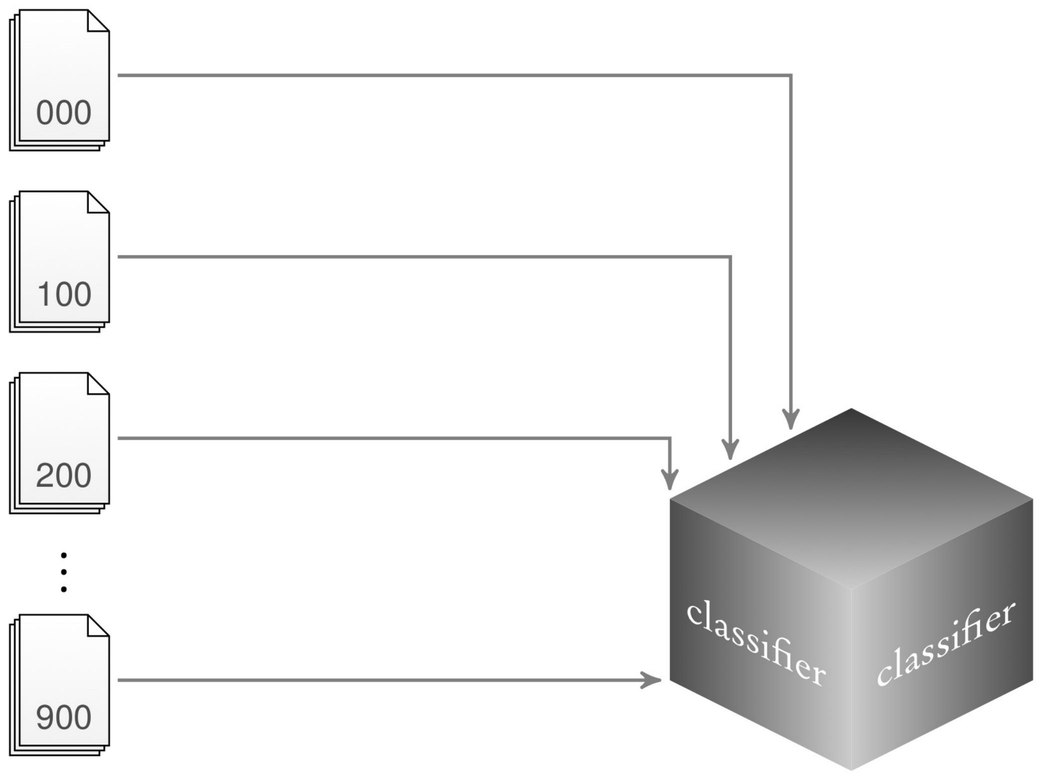

The Impact of Robust Modalities on Hardware and Architecture
orks as
algorithm
ime [11]

LIFE SCIENCES
BIOLOGY
LITERATURE
MATHEMATICS
COMPUTER SCIENCE
PHYSICS
HISTORY
POLITICAL SCIENCE
SOCIAL SCIENCE

Contents for Classifier Feed

- dc:description: 30 to 40 % of metadata records have dc:description
- Document fulltext (if accessible)
- Setspec contains ddc and lcc codes
- dc:subject contains lots of subject-orientated information

Training examples

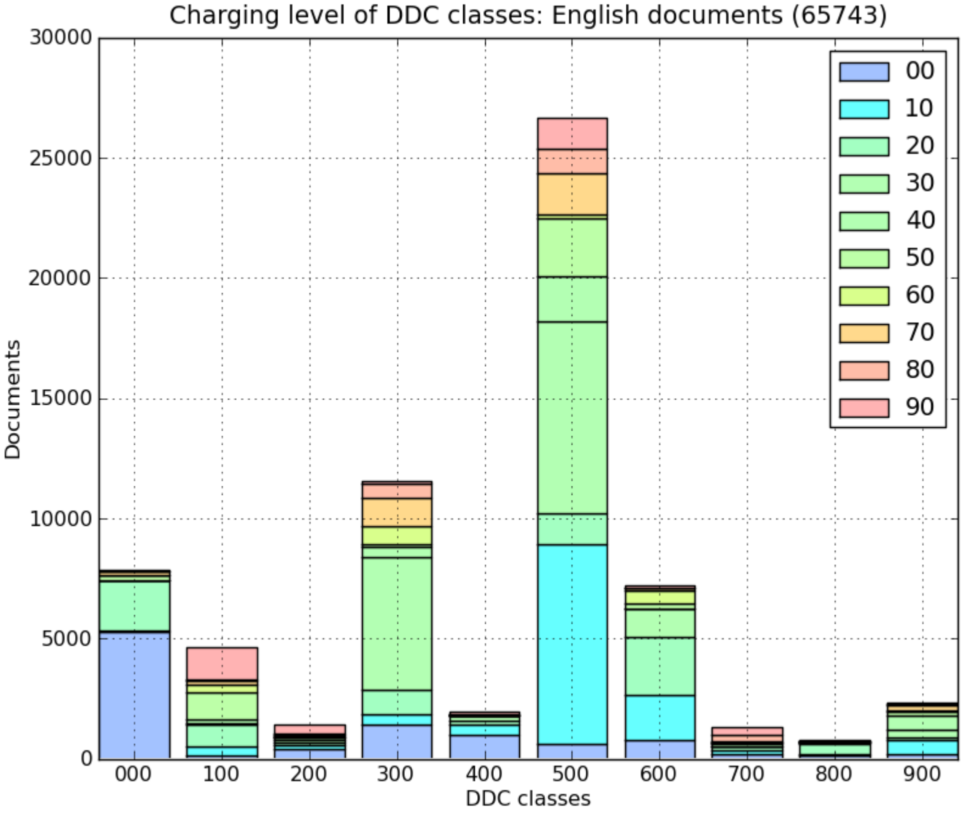


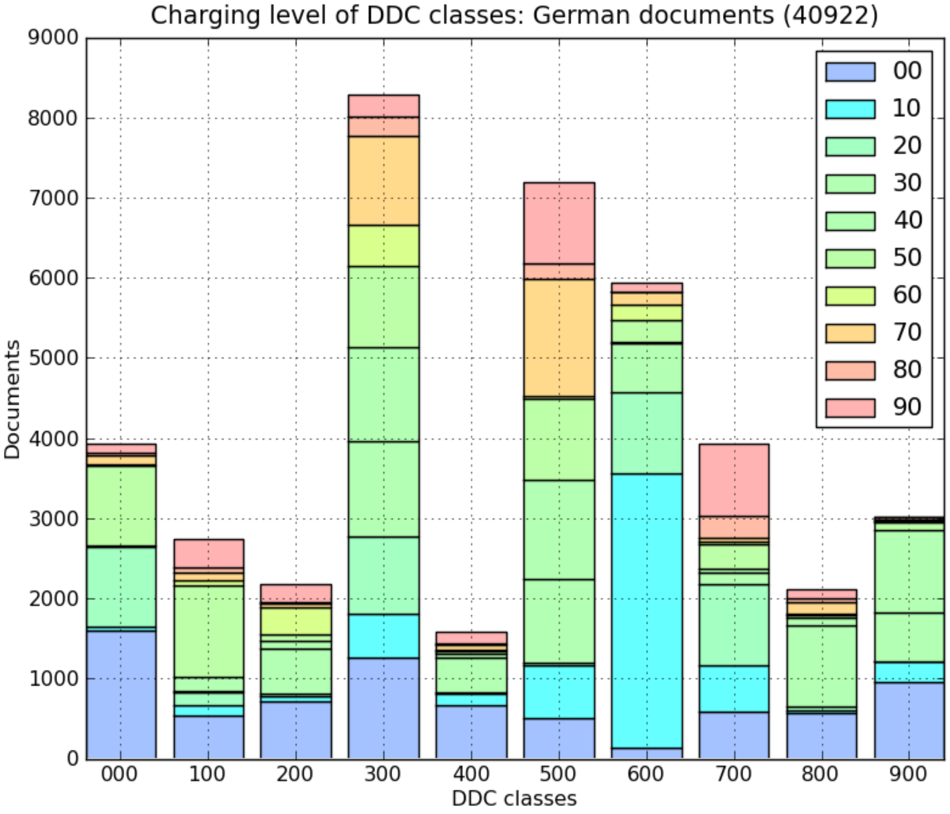
Mapping of frequently used classifications

LCC
ELIS classification
ArXiv classification

DDC codes: ~400.000 Documents = 1,4%

DDC classes distribution in Harvesting Results





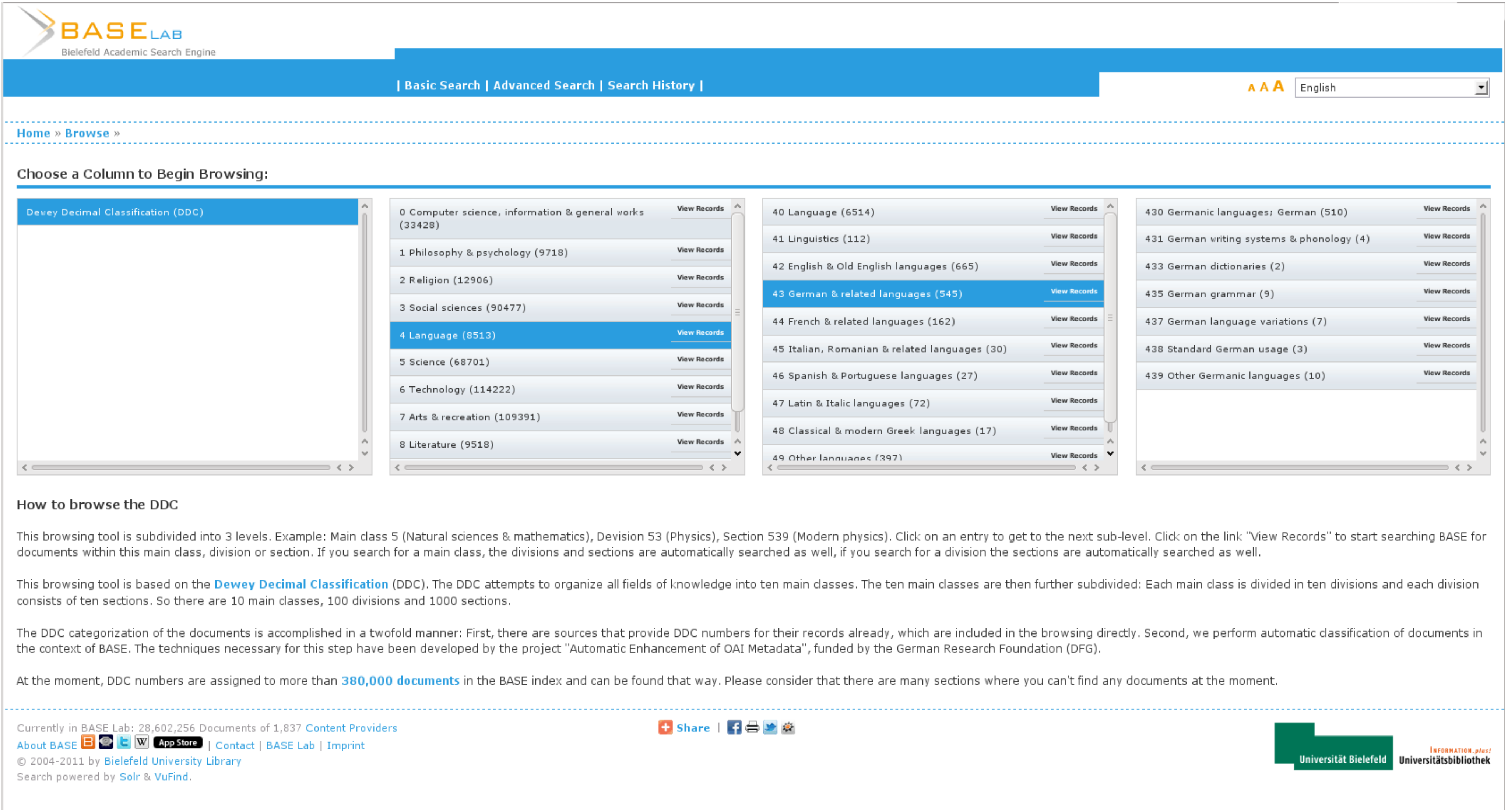
Universität Bielefeld Universitätsbibliothek

BASE Bielefeld Academic Search Engine

OpenAIRE

driver

Subject-based Browsing



BASE LAB Bielefeld Academic Search Engine

| Basic Search | Advanced Search | Search History |

A A A English

Home » Browse »

Choose a Column to Begin Browsing:

Dewey Decimal Classification (DDC)

0 Computer science, information & general works (33428) View Records
1 Philosophy & psychology (9718) View Records
2 Religion (12906) View Records
3 Social sciences (90477) View Records
4 Language (8513) View Records
5 Science (68701) View Records
6 Technology (114222) View Records
7 Arts & recreation (109391) View Records
8 Literature (9518) View Records

40 Language (6514) View Records
41 Linguistics (112) View Records
42 English & Old English languages (665) View Records
43 German & related languages (545) View Records
44 French & related languages (162) View Records
45 Italian, Romanian & related languages (30) View Records
46 Spanish & Portuguese languages (27) View Records
47 Latin & Italic languages (72) View Records
48 Classical & modern Greek languages (17) View Records
49 Other languages (397) View Records

430 Germanic languages; German (510) View Records
431 German writing systems & phonology (4) View Records
433 German dictionaries (2) View Records
435 German grammar (9) View Records
437 German language variations (7) View Records
438 Standard German usage (3) View Records
439 Other Germanic languages (10) View Records

How to browse the DDC

This browsing tool is subdivided into 3 levels. Example: Main class 5 (Natural sciences & mathematics), Devision 53 (Physics), Section 539 (Modern physics). Click on an entry to get to the next sub-level. Click on the link "View Records" to start searching BASE for documents within this main class, division or section. If you search for a main class, the divisions and sections are automatically searched as well, if you search for a division the sections are automatically searched as well.

This browsing tool is based on the Dewey Decimal Classification (DDC). The DDC attempts to organize all fields of knowledge into ten main classes. The ten main classes are then further subdivided: Each main class is divided in ten divisions and each division consists of ten sections. So there are 10 main classes, 100 divisions and 1000 sections.

The DDC categorization of the documents is accomplished in a twofold manner: First, there are sources that provide DDC numbers for their records already, which are included in the browsing directly. Second, we perform automatic classification of documents in the context of BASE. The techniques necessary for this step have been developed by the project "Automatic Enhancement of OAI Metadata", funded by the German Research Foundation (DFG).

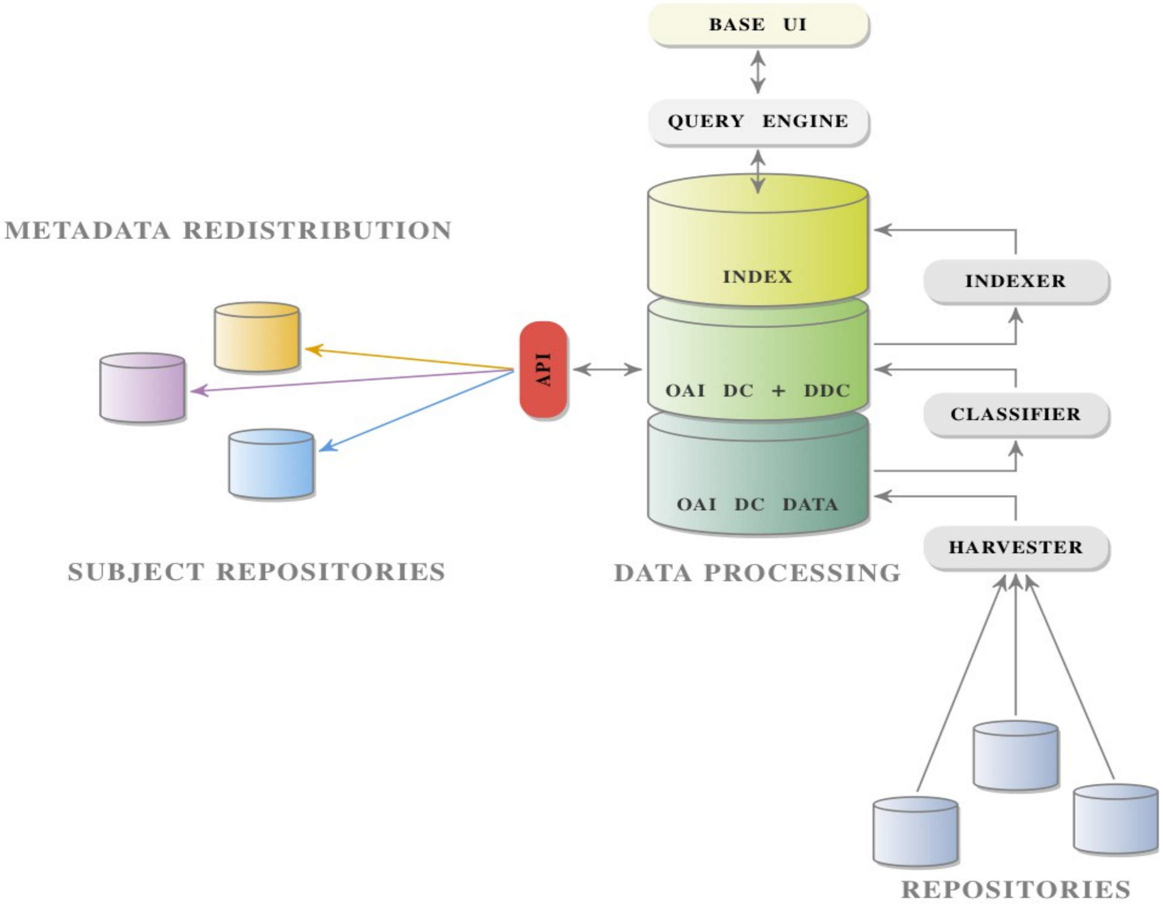
At the moment, DDC numbers are assigned to more than 380,000 documents in the BASE index and can be found that way. Please consider that there are many sections where you can't find any documents at the moment.

Currently in BASE Lab: 28,602,256 Documents of 1,837 Content Providers
About BASE App Store | Contact | BASE Lab | Imprint
© 2004-2011 by Bielefeld University Library
Search powered by Solr & VuFind.

Share |

Universität Bielefeld INFORMATION.plus! Universitätsbibliothek

Delivering Subject-Related Documents



The End. Thank you!

Questions, Comments etc.

Mail: friedrich.summann@uni-bielefeld.de