

The Workshop on Classification and Subject Indexing in
Library and Information Science (LIS'2015)
Colchester Sept 2nd - 3rd 2015

The Role of Classification Information in Open Access Repositories

current status and future directions

Dirk Pieper ; Friedrich Summann
Bielefeld University Library

- **Overview**

- The Repository Landscape
 - Metadata Provision and Classification Information in Repositories
 - Classification-based Activities in Repositories
 - Future Directions
-

- Overview
 - **The Repository Landscape**
 - Metadata Provision and Classification Information in Repositories
 - Classification-based Activities in Repositories
 - Future Directions
-

The IR – **past**, present, future

- Started late nineties
 - Contents starting with thesis
 - OAI-PMH protocol definition 2001
 - Open Access movement
 - Establishing a global repository network
 - Extending
 - Size
 - Quality
 - Services
-

The IR – past, **present**, future

- More than 5000 repositories, more than 100 Mill. Objects
 - World-wide coverage
 - But: Institutional Repositories are at a turning-point:
 - More and more overlapping systems
 - local (CRIS, Publishing Platform, etc)
 - external (Subject Repositories, ResearchGate etc.)
 - Scholarly communication process changes
-

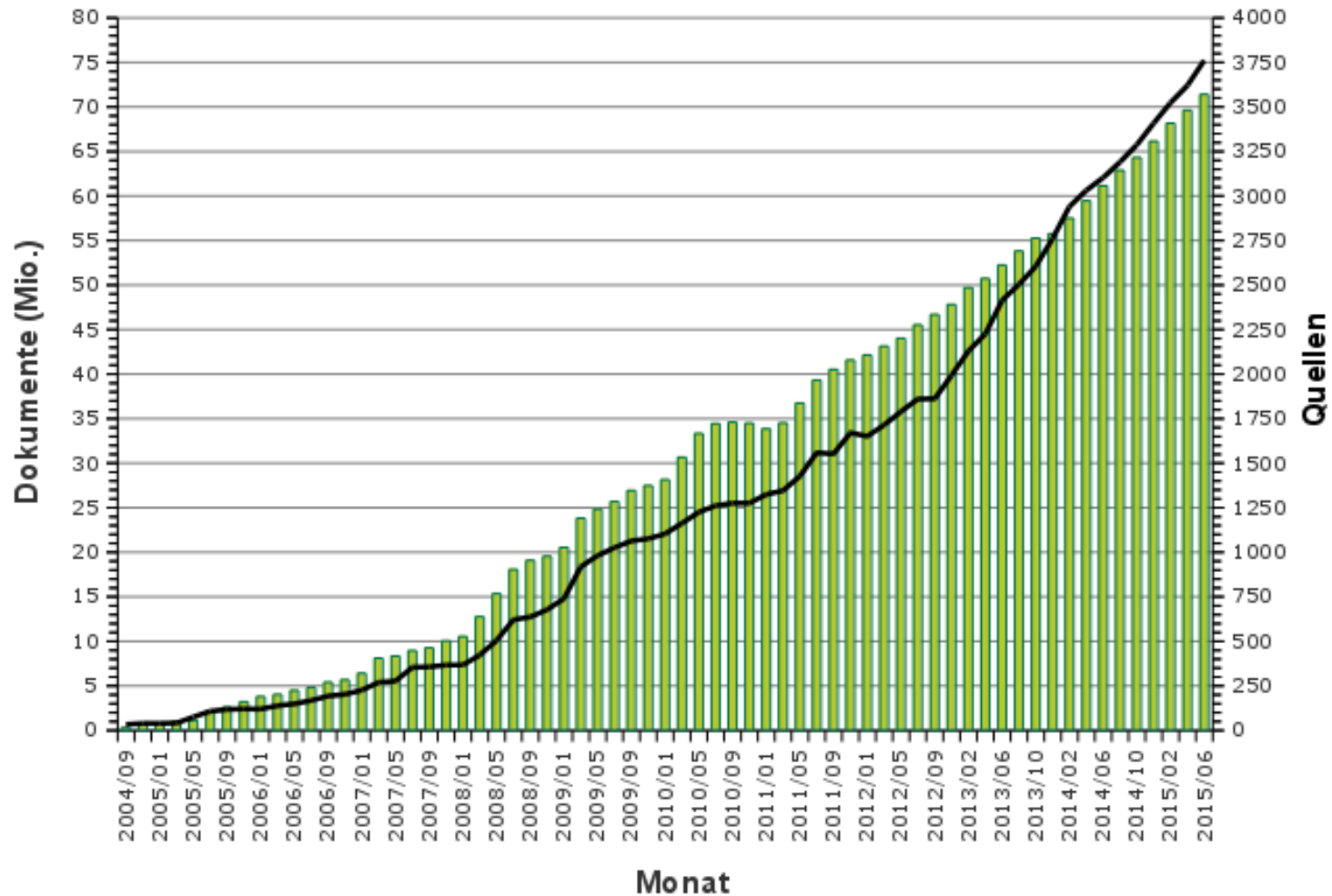
The BASE scope

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

Facts about BASE (Aug 20th 2015)

- 3644 Repositories included
 - From 102 Countries world-wide
 - Ca. 77 Mill. Documents/Objects
 - Ca. 70 % Open Accessible
 - Ca. 10.8 Mill. Documents enriched with DDC-Codes (Dewey)
-

Zahl der indexierten Dokumente / und Quellen ■ in BASE

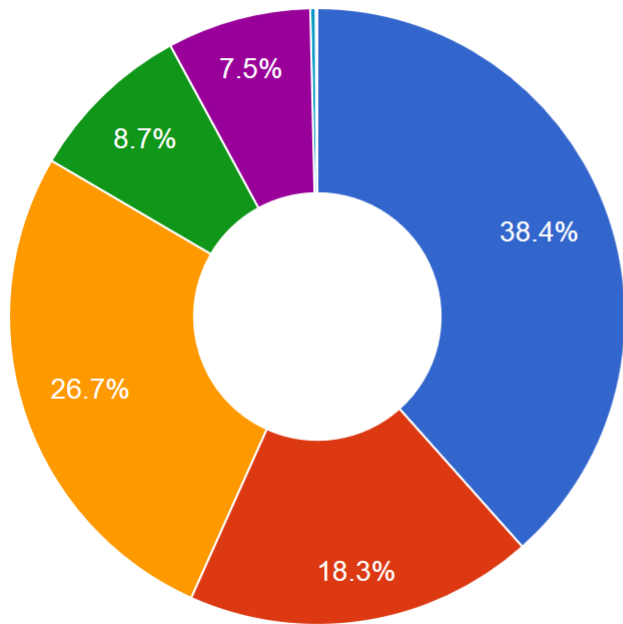


Harvesting Environment

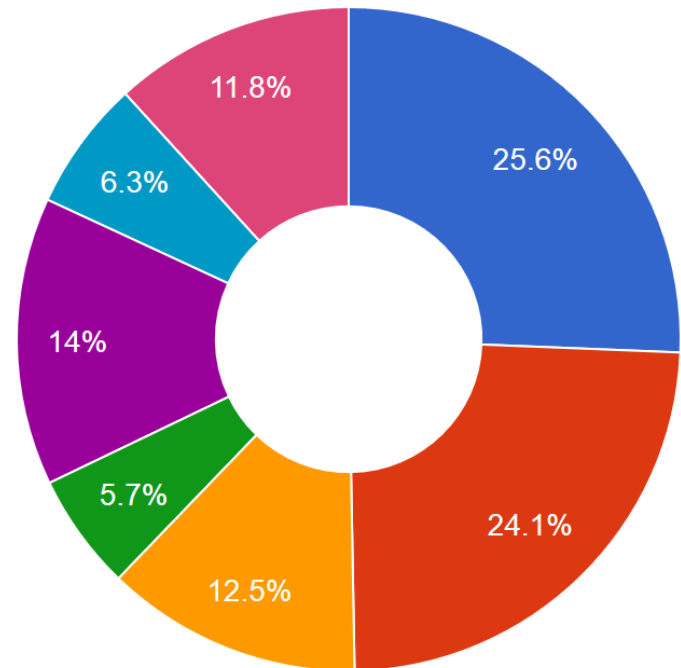
(Based on OAI-PMH)

- 5979 Repositories harvested
 - 4832 active
 - 3531 indexed
 - 1147 deprecated
 - 184 Mill. Records
 - 111 Mill. unique
 - 73 Mill. indexed
 - 1.03 Terabyte of Data
 - 2853 Cronjobs (weekly)
-

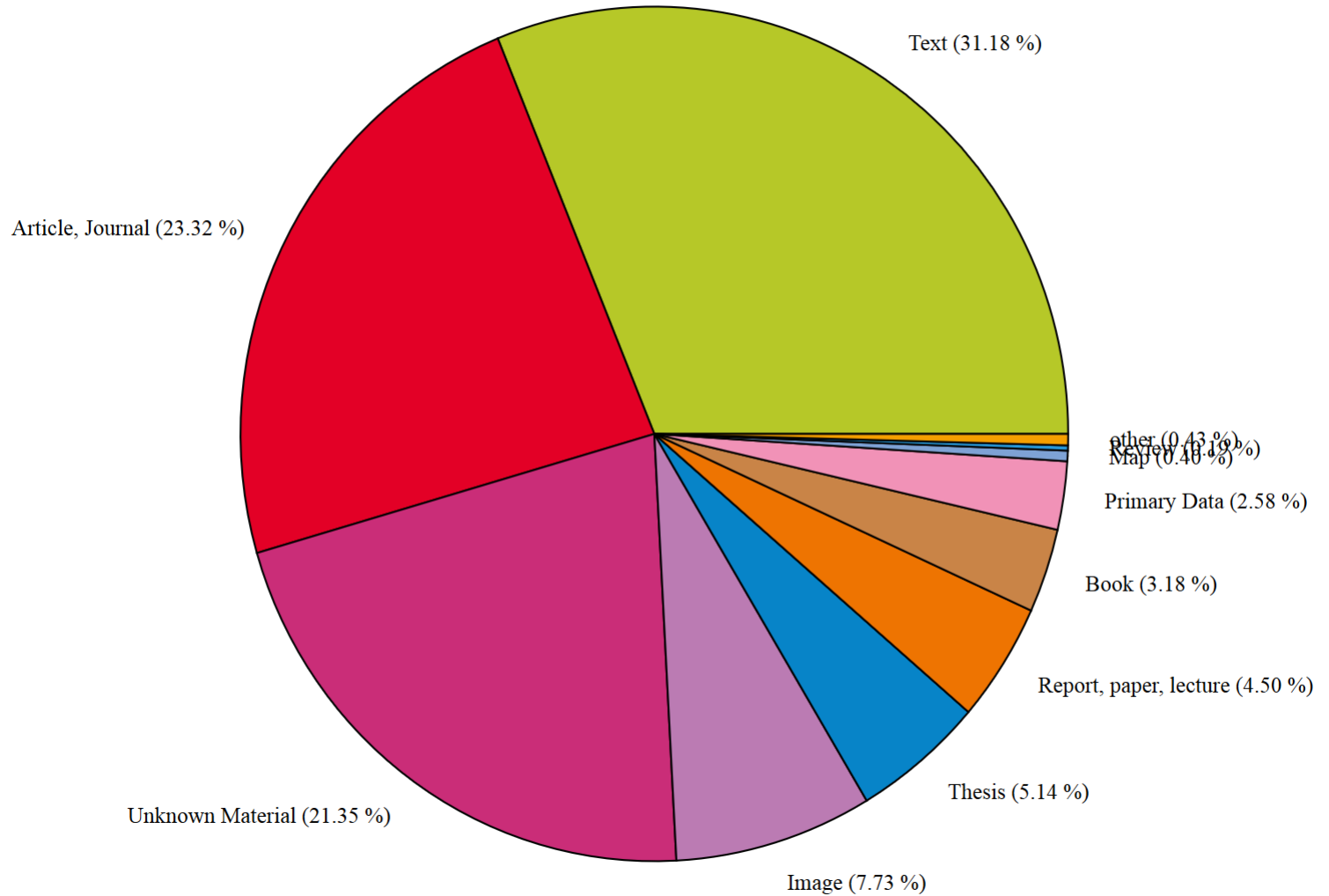
Repository Types covered in BASE



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



Distribution Publication Type



- Overview
 - The Repository Landscape
 - **Metadata Provision and Classification Information in Repositories**
 - Classification-based Activities in Repositories
 - Future Directions
-

Classification Information in Repositories

- OAI-PMH as Transport Layer
 - Dublin Core as the Mandatory Format
 - Set-Definition for Grouping Contents
 - Classification Information in dc:subject
 - Example:
 - `<dc:subject>LCSH:Ausdehnungslehre; LCCN QA205.H99</dc:subject>`
-

Record example

(DDC codes in setspec and dc:subject)

- ```
<record>
<header>
 <identifier>oai:pub.uni-bielefeld.de:2759012</identifier>
 <datestamp>2015-08-27T14:22:34Z</datestamp>
 <setSpec>journalArticle</setSpec>
 <setSpec>doc-type:article</setSpec>
 <setSpec>ddc:620</setSpec>
 <setSpec>journalArticleFtxt</setSpec>
 <setSpec>driver</setSpec>
 <setSpec>open_access</setSpec>
</header>
<metadata>
<oai_dc:dc xmlns="http://www.openarchives.org/OAI/2.0/oai_dc/"
 xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
 ...
xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/ http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
 <dc:title>An Oil-Based Lubrication System Based on Nanoparticulate TiO2 with Superior Friction and Wear
 Properties</dc:title>
 <dc:creator>Bogunovic, Lukas</dc:creator>
 ...
 <dc:subject>Lubrication</dc:subject>
 <dc:subject>Wear</dc:subject>
 <dc:subject>Titanium dioxide</dc:subject>
 <dc:subject>Friction</dc:subject>
 <dc:subject>DDC:620</dc:subject>
 <dc:description>We evaluated the perform
```

## Scanning the BASE Metadata Store:

- Listing dc:subject content from the BASE Data:
  - 3.1 GB of data
  - 49,947,721 Different terms
  - Very broad variety
  - Containing
    - Classification Codes
    - Subject Headings
    - All Kind of Text
-

# dc:subject Examples from Different Repositories

ddc:330: 23 ~ UnivEichstaett-Opus (23)

DDR; Doping; das Sportwunder DDR; Schweden und DDR; Diskurslinguistik: 1 ~  
UnivGoeteborg-OJS (1)

617 Chirurgie und verwandte medizinische Fachrichtungen: 2 ~ FUBerlin (2)

UDK 620.92.579.66: 1 ~ NAviationUniv-OJS (1)

MSC 15A24: 2 ~ MonarchChemnitz (2)

PACS 45.70.Cc 83.10.Rs 83.80.Fg ; Mathematics Subject Classification (2000) 65Y05 70E55: 1  
~ ARTXIKER (1)n

Primary 65Y05; Secondary 65Y10, 65F30: 1 ~ UnivNis-OJS (1)

65Y05: 1 ~ EduTice (1)

68T10 Pattern recognition, speech recognition: 4 ~ UnivKoeln-CSD (4)

rvk:AP 39800: 3 ~ TUDresden (3)



# Classification Information in Repository Metadata

## Universal Classifications

- DDC
- UDC
- LCC
- BK
- RVK

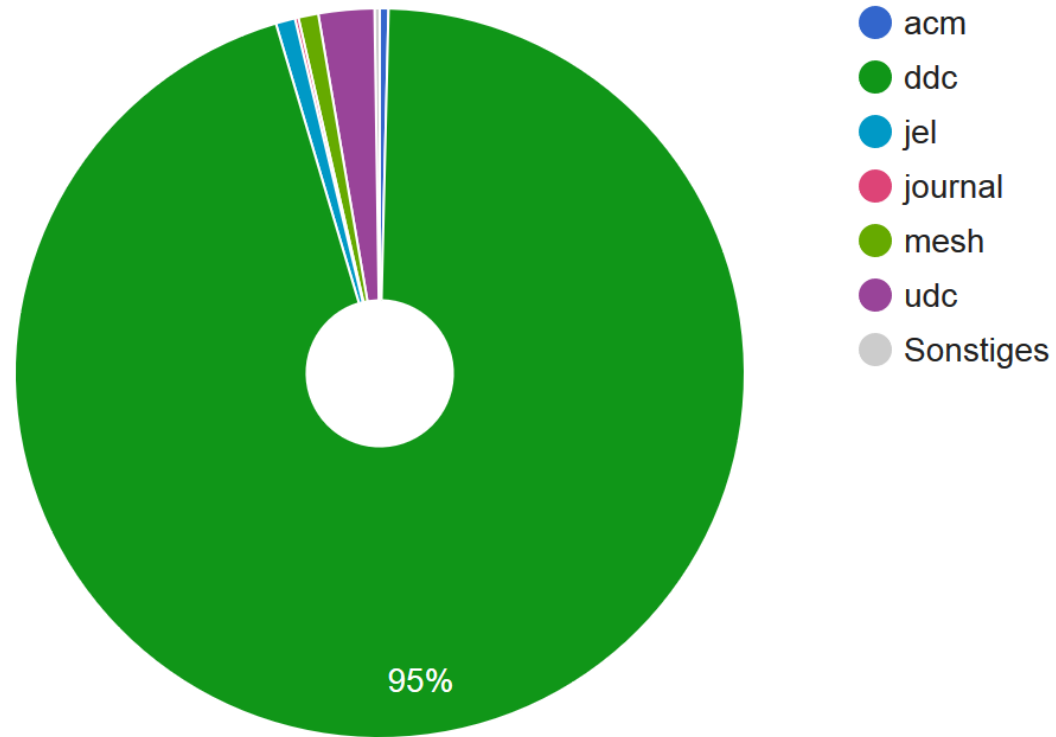
## Subject Classifications

- MSC
- PACS
- ACM

## Proprietary Classifications

- HEP
  - ELIS
-

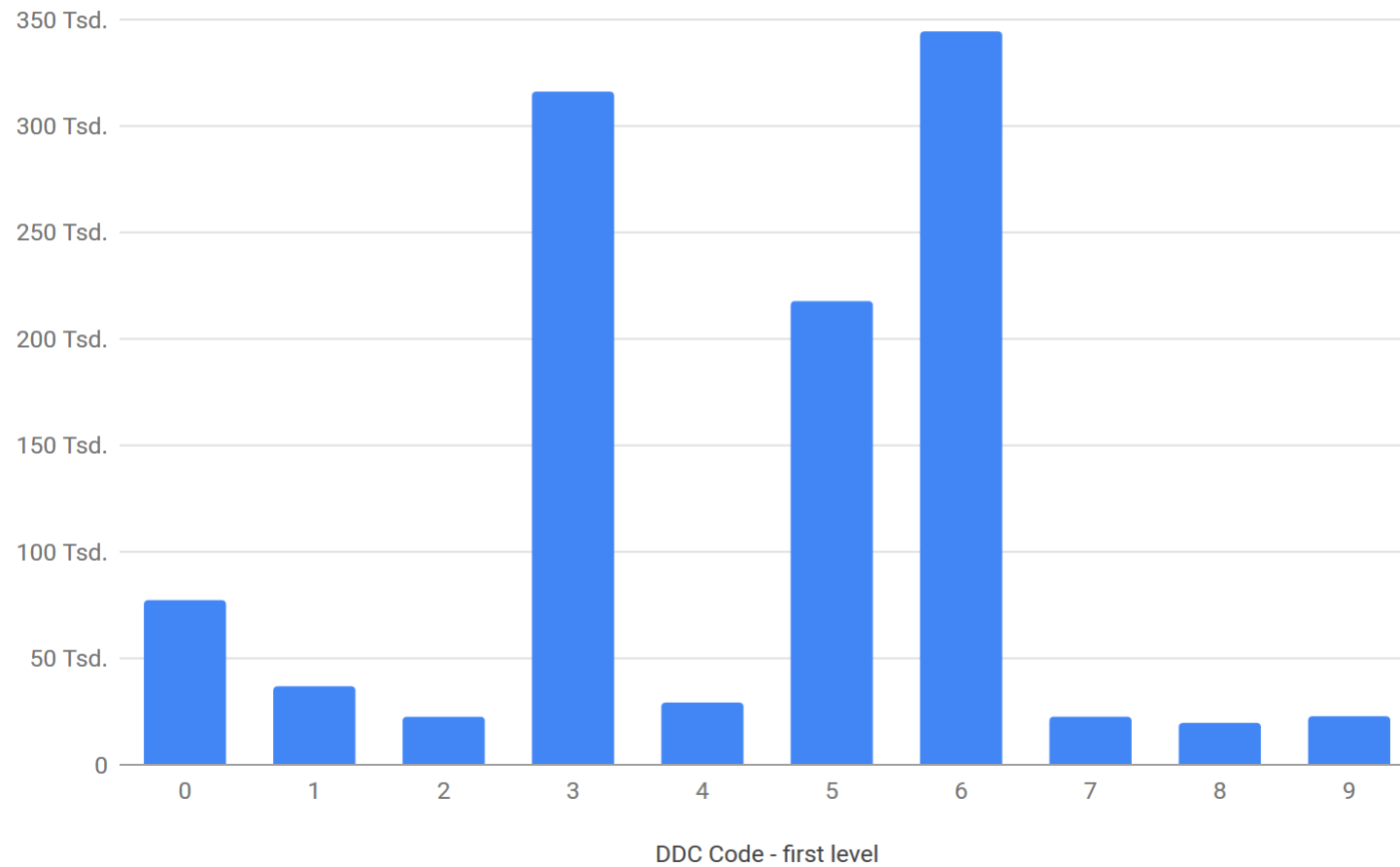
### Classification Distribution in Repository Metadata



# Most frequent DDC Terms (DDC Code/frequency)

- 330 ; 192152
- 610 ; 152522
- 540 ; 55136
- 620 ; 48548
- 530 ; 40073
- 570 ; 39509
- 370 ; 38975
- 616 ; 37556
- 000 ; 30778
- 150 ; 29505
- 300 ; 27044
- 320 ; 25230
- 004 ; 23331
- 616.07 ; 21350
- 510 ; 19179
- 200 ; 19157
- 615 ; 17031
- 340 ; 14973
- 550 ; 14265
- 616.8 ; 14159
- 617 ; 12947
- 500 ; 12711
- 580 ; 11851
- 430 ; 9431
- 230 ; 9312
- 302 ; 9276
- 621 ; 8983
- 900 ; 8601
- 590 ; 8558
- 658 ; 8490
- 100 ; 8378
- 400 ; 8137
- 070 ; 8007
- 199 ; 7999
- 720 ; 7722

### DDC Codes in OAI-PMH Repositories



# OAI-PMH Protocol Definition

## 2.6 Set

A *set* is an optional construct for grouping items for the purpose of [selective harvesting](#)

Repositories **may** organize items into sets

When a repository defines a set organization it **must** include set membership information in the [headers](#)

The following is an example of a possible set hierarchy in a repository:

Subjects

- Existential Kenesiology

- Quantum Psychology

For example, a group of cooperating e-print archives in a specific discipline may agree on sets that arrange metadata in their repositories based on a controlled subject classification.

# OAI-PMH Protocol Definition

## 2.7.2 Selective Harvesting and Sets

Harvesters may specify [set](#) membership as a criteria for selective harvesting. To specify set-based selective harvesting, a [setSpec](#) is included as the value of the **optional** set argument to the [ListRecords](#) and [ListIdentifiers](#) requests

---

- **Example**

- <record>

- <header>

- <identifier>oai:pub.uni-bielefeld.de:2759012</identifier>

- <datestamp>2015-08-27T14:22:34Z</datestamp>

- <setSpec>journalArticle</setSpec>

- <setSpec>doc-type:article</setSpec>

- <setSpec>ddc:620</setSpec>**

- <setSpec>journalArticleFtxt</setSpec>

- <setSpec>driver</setSpec>

- <setSpec>open\_access</setSpec>

- </header>

# Eprints LCC Usage (via Sets): Example

```
<set>
```

```
<setSpec>7375626A656374733D4A:4A31</setSpec>
```

```
 <setName>Subject = J Political Science: J General legislative and executive
papers</setName>
```

```
 </set>
```

```

```

```
<ListRecords>
```

```
 <record>
```

```
 <header>
```

```
 <identifier>oai:eprints.gla.ac.uk:106683</identifier>
```

```
 <datestamp>2015-05-26T15:04:44Z</datestamp>
```

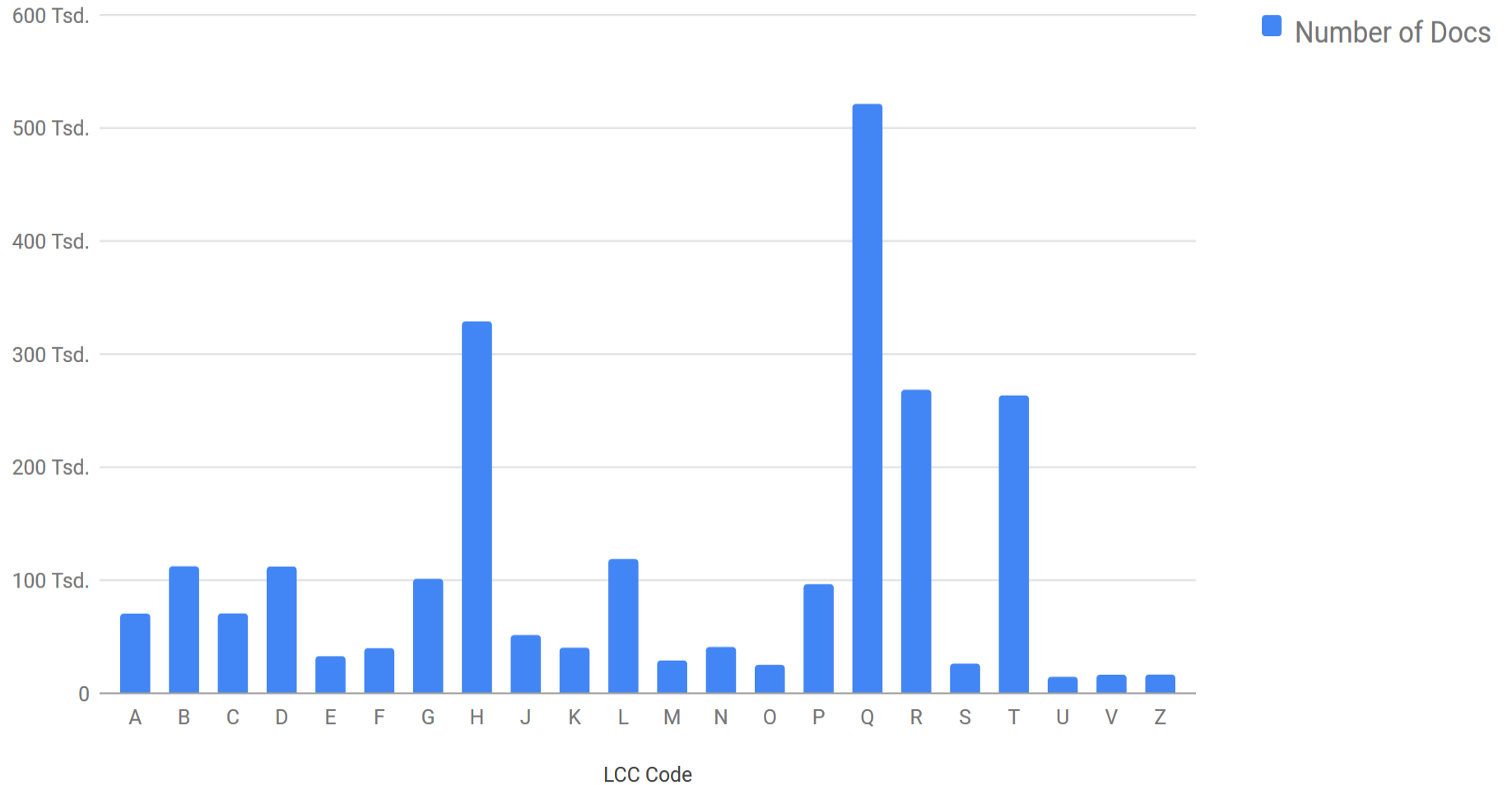
```
 <setSpec>7374617475733D707562</setSpec>
```

```
 <setSpec>74797065733D61727469636C65</setSpec></header>
```

```
 <metadata>
```



### LCC Codes in OAI-PMH Repositories



# Vocabulary Efforts and Effects

- DINI Certificate
  - DRIVER Guidelines
  - COAR Interest Group “Controlled Vocabularies for Repository Assets”
-

# DRIVER Guidelines 2.0

## Subject classification

Metadata delivered via OAI-PMH contain a broad range of subject headings and classification information. The used classification and subject heading systems and the presentation formats vary broadly

**It is recommended to use an URI when using classification schemes or controlled Vocabularies ...**

```
<dc:subject>info:eu-epo/classification/ddc/641</dc:subject>
```

---

# DRIVER Guidelines Vocabulary in Practice = 25998

## different terms, some examples

info:eu-repo/classification/ddc/020: 16 ~ UnivWien-DC (16)

info:eu-repo/classification/ddc/333.7: 341 ~ GSI-DE (341)

info:eu-repo/classification/udc/001.32: 2 ~ UnivMaribor (2)

info:eu-repo/classification/bk/50.33: 4 ~ SUBGoettingen (4)

info:eu-repo/classification/mesh/Actin Cytoskeleton: 4 ~ SUBGoettingen (4)

info:eu-repo/classification/acm/D.2 SOFTWARE ENGINEERING: 70 ~ PUMA (70)

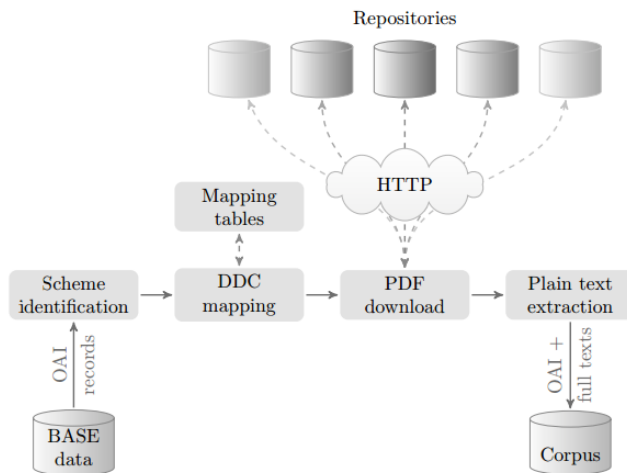
info:eu-repo/classification/ddc/333.7: 341 ~ GSI-DE (341)

info:eu-repo/classification/udc/614.47-053.4: 1 ~ EuropeanLibrary\_deweyfull:--  
base\_dc (1)

- Overview
    - The Repository Landscape
    - Metadata Provision and Classification Information in Repositories
    - **Classification-based Activities in Repositories**
    - Future Directions
-

# Use Case: Subject-Based Browsing (Automatic Classification of Repository Metadata)

The Bielefeld UL Approach:



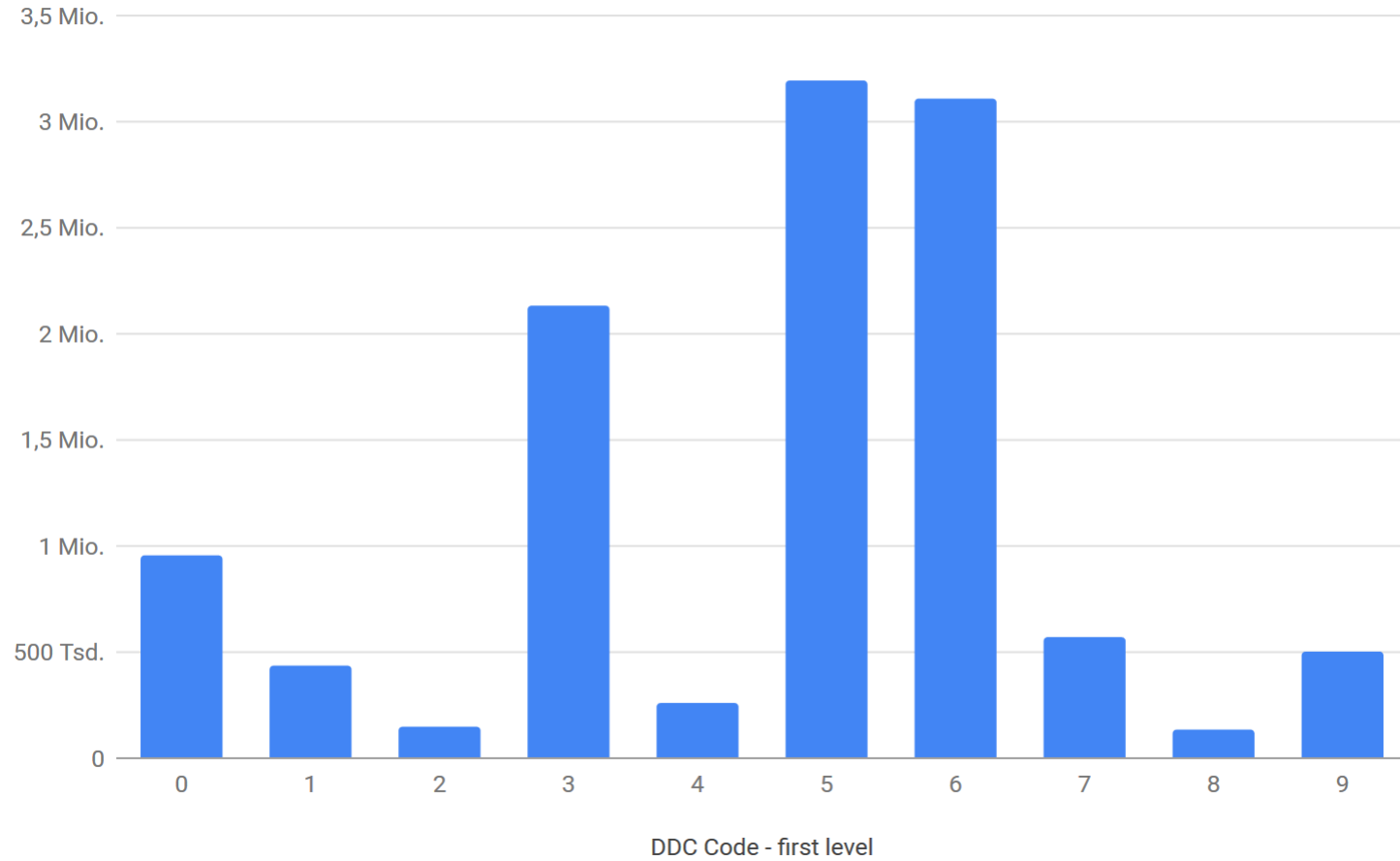
- Konstruktion eines DDC-kategorisierten Textkorpus aus der BASE-Datenbasis
- Metadaten + Volltexte
- ~ 100.000 Dokumente
- Deutsch und Englisch
- semi-automatische Vergabe von DDC-Nummern durch Konkordanzen zu Fachklassifikationen

# BASE feature: DDC Browsing

0 Computer science, information & general works (959742)	<a href="#">View Records</a>	50 Science (311720)	<a href="#">View Records</a>	540 Chemistry & allied sciences (259942)	<a href="#">View Records</a>
1 Philosophy & psychology (440218)	<a href="#">View Records</a>	51 Mathematics (322615)	<a href="#">View Records</a>	541 Physical chemistry (114720)	<a href="#">View Records</a>
2 Religion (153073)	<a href="#">View Records</a>	52 Astronomy (266159)	<a href="#">View Records</a>	542 Techniques, equipment & materials (17)	<a href="#">View Records</a>
3 Social sciences (2131477)	<a href="#">View Records</a>	53 Physics (514692)	<a href="#">View Records</a>	543 Analytical chemistry (44)	<a href="#">View Records</a>
4 Language (264598)	<a href="#">View Records</a>	<b>54 Chemistry (371809)</b>	<a href="#">View Records</a>	546 Inorganic chemistry (149)	<a href="#">View Records</a>
<b>5 Science (3191899)</b>	<a href="#">View Records</a>	55 Earth sciences & geology (309779)	<a href="#">View Records</a>	547 Organic chemistry (110)	<a href="#">View Records</a>
6 Technology (3102648)	<a href="#">View Records</a>	56 Fossils & prehistoric life (822)	<a href="#">View Records</a>	548 Crystallography (16)	<a href="#">View Records</a>
7 Arts & recreation (574437)	<a href="#">View Records</a>	57 Life sciences; biology (767769)	<a href="#">View Records</a>	549 Mineralogy (4)	<a href="#">View Records</a>
8 Literature (139299)	<a href="#">View Records</a>	58 Plants (Botany) (182629)	<a href="#">View Records</a>		
		59 Animals (Zoology) (224810)	<a href="#">View Records</a>		

At the moment, DDC numbers are assigned to **10,859,707 documents** in the BASE index

DDC Codes in BASE - computed





# Requirements for Feeding Metadata Records Into the Automatic Classifier

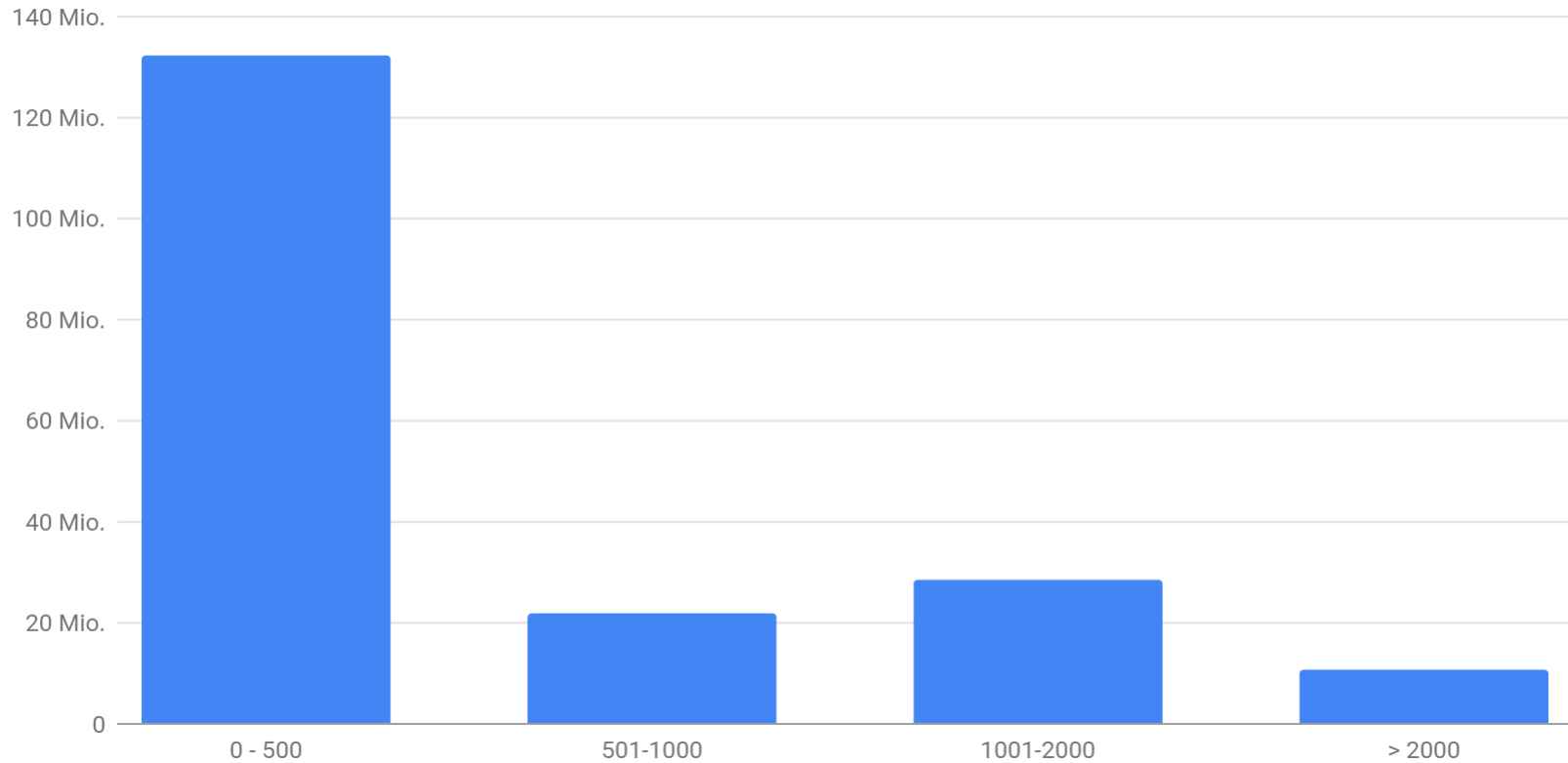
## Language of Text:

- Englisch
- German

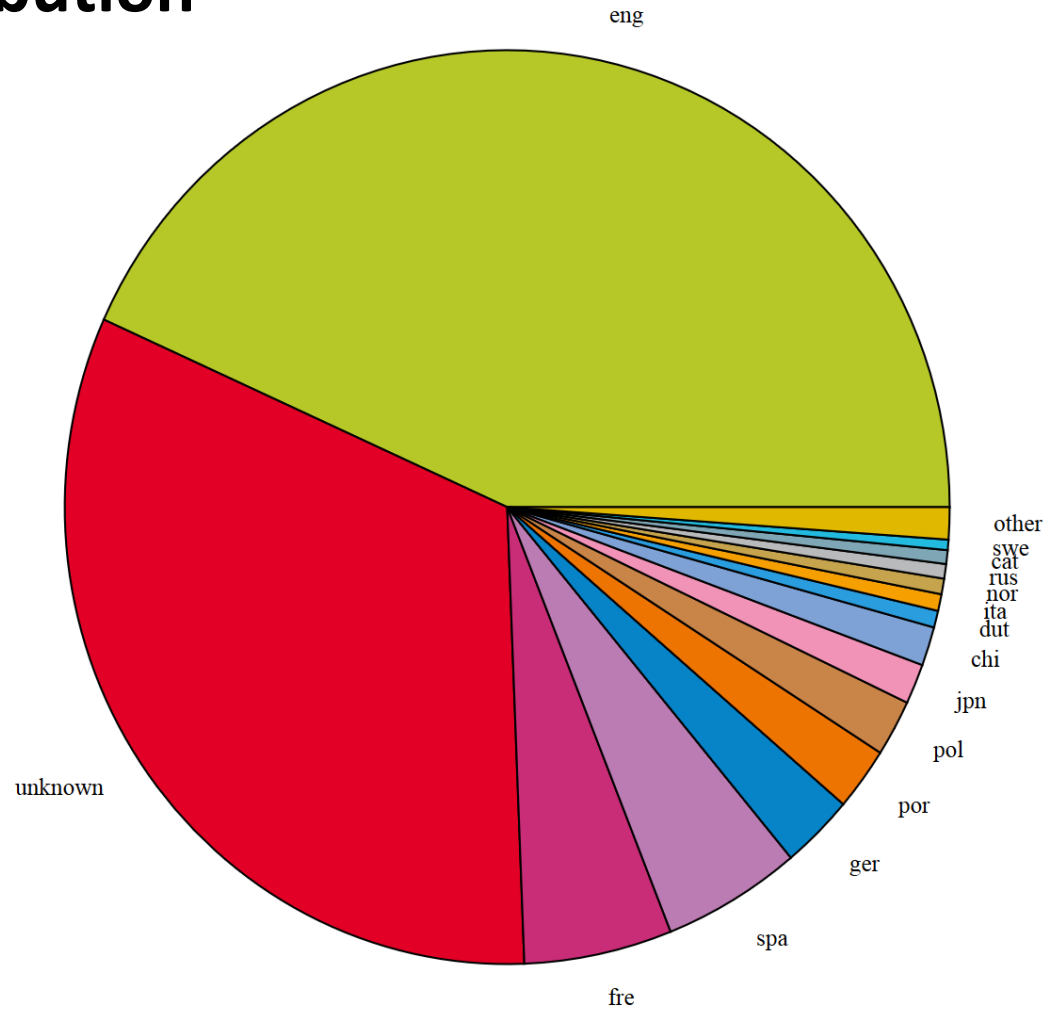
## Descriptive Text, no metadata as author and related

- At least 500 characters
-

## dc:description Character Length in Metadata



# Language Distribution



# Automatic Classification Processing Steps

Stream of Characters

Language Identification

Lower Casing  
Tokenization  
Stop Word Elimination

Dictionary

Document-Term Matrix

Thus, we demonstrate that Web services can be made efficient, certifiable, and self-learning.



Thus, we demonstrate that Web services can be made efficient, certifiable, and self-learning.

thus, we demonstrate that web services can be made efficient, certifiable, and self-learning.

thus we demonstrate that web services can be made efficient certifiable and self-learning

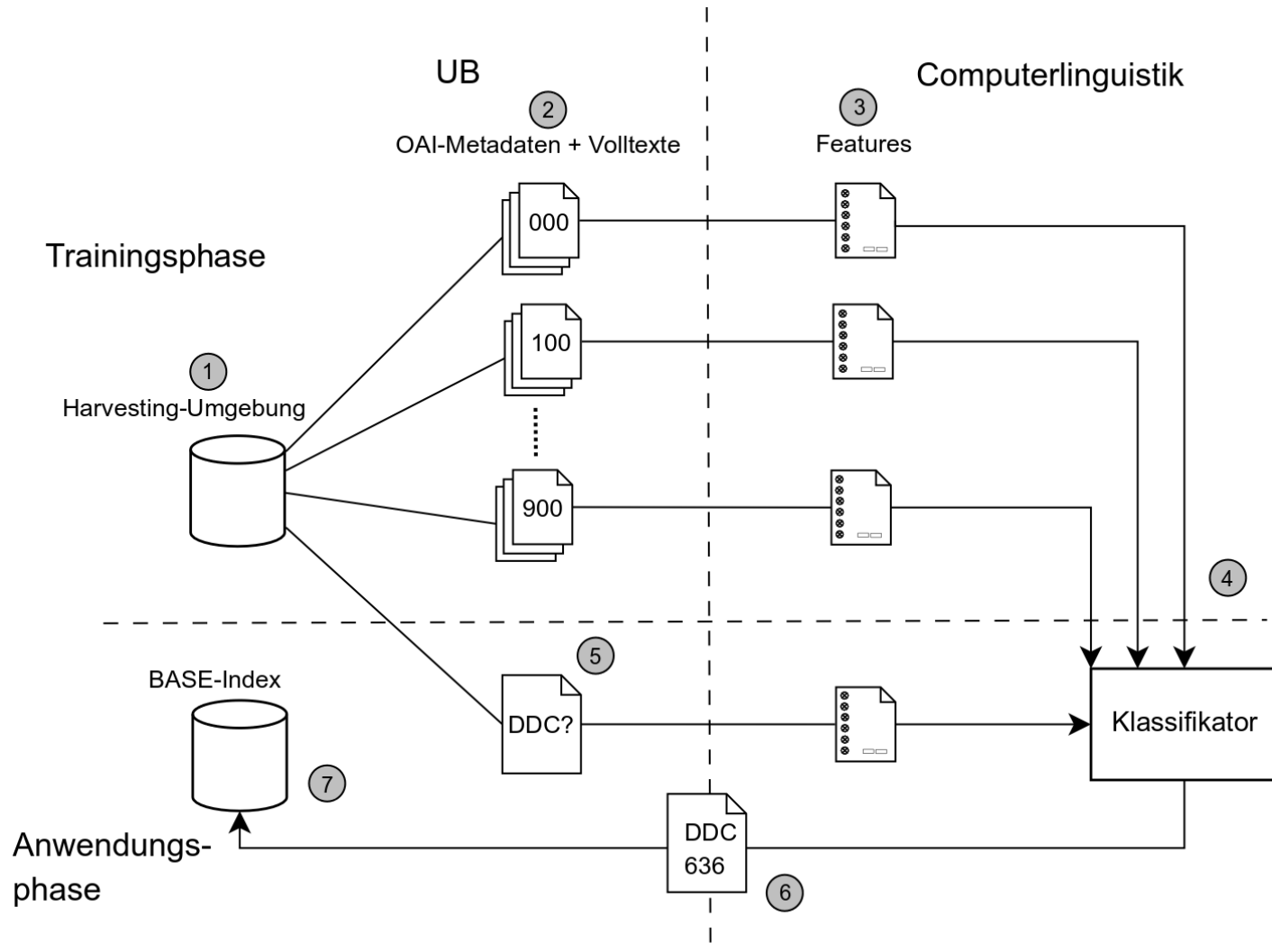
demonstrate web services efficient certifiable self-learning

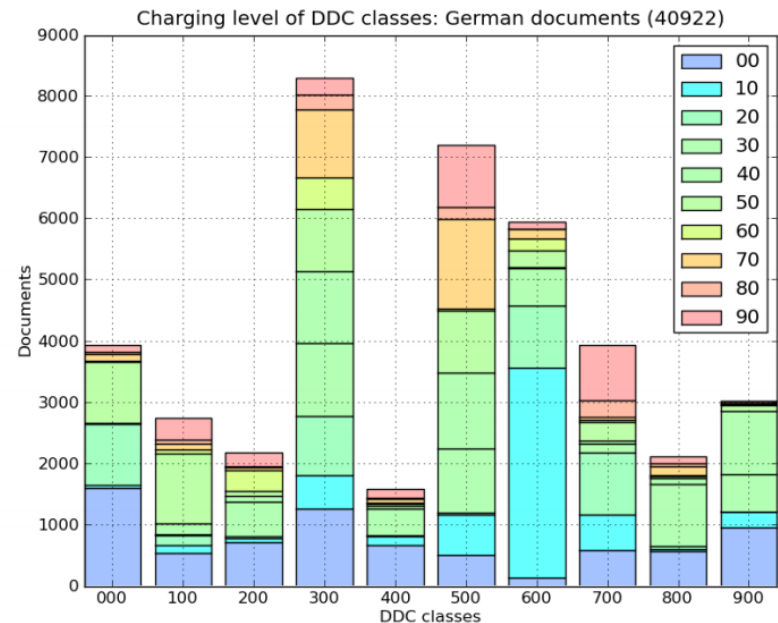
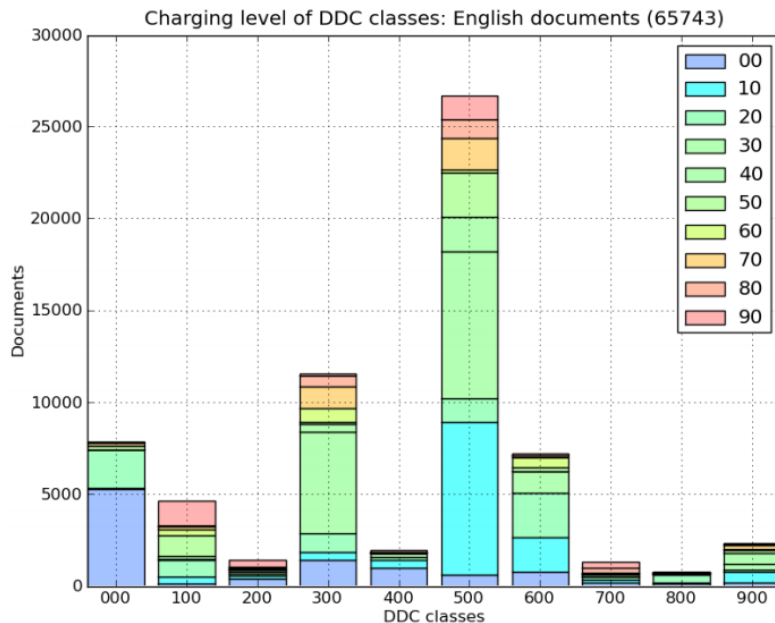
Token	ID
certifiable	0
demonstrate	1
efficient	2
services	3
web	4

```
-1 3:1 12:1 17:1 19:1
+1 0:1 1:1 2:1 3:1 4:1
-1 1:1 3:1 9:1 12:
+1 0:1 5:1 7:1 9:1
+1 0:1 5:1 18:1
-1 1:1 3:1 7:1 29:1
+1 3:1 9:1 37:1 109:1
+1 0:1 3:1 4:1 19:1
+1 0:1 17:1 36:1 61:1
-1 3:1 4:1 5:1 7:1 10:1
-1 1:1 3:1 4:1 7:1 9:1
-1 1:1 3:1 7:1 29:1
+1 0:1 1:1 2:1 3:1 18:1
```

Preprocessing Pipeline

# Data workflow Classifier





- Schiefe Verteilung der Dokumente über die DDC-Klassen
- Wenig Beispieldokumente in den Geisteswissenschaften
- Dokumentakquise ab der dritten DDC-Ebene (1.000 Klassen) extrem aufwändig mangels guter Sacherschließungsinformationen

# Use Case

## Data Provision based on Classification Sets

Dynamic BASE OAI-PMH interface with DDC as parameter

- Europeana
  - ZBW Kiel
  - Virtuelle Fachbibliotheken
  - ...
-

- Overview
    - The Repository Landscape
    - Metadata Provision and Classification Information in Repositories
    - Classification-based Activities in Repositories
    - **Future Directions**
-



# Future Directions

- Vocabulary Definition and Usage in Practice
  - More Detailed Classification Information  
(as part of Metadata Quality efforts)
  - Linked Open Data Integration
  - We expect more information, more vocabulary usage and more precision in prediction
-

## More Information:

- Vanderfeesten M, Summann F, Slabbertje M, eds. *DRIVER Guidelines 2.0 : Guidelines for content providers - Exposing textual resources with OAI-PMH.*; 2008.
  - Lösch M, Waltinger U, Horstmann W, Mehler A. Building a DDC-annotated Corpus from OAI Metadata. *Journal of Digital Information*. 2011;12(2).
  - Pieper D, Summann F. 10 years of „Bielefeld Academic Search Engine“ (BASE): Looking at the past and future of the world wide repository landscape from a service providers perspective. Presented at the OR2015. 10th International Conference on Open Repositories, Indianapolis
  - Summann F, Shearer K. *COAR Roadmap Future Directions for Repository Interoperability*. Göttingen: COAR Confederation of Open Access Repositories; 2015.
-

# Thank You!

[friedrich.summann@uni-bielefeld.de](mailto:friedrich.summann@uni-bielefeld.de)

[dirk.pieper@uni-bielefeld.de](mailto:dirk.pieper@uni-bielefeld.de)

---