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

BASE - Bielefeld Academic Search Engine (http://www.base-search.net/)
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: 38.4%
- Publication Server: 26.7%
- Electronic Journals: 18.3%
- Thesis/Dissertation Servers: 11.8%
- Digital Collections: 11.8%
- Research Data: 6.3%
- Other: 12.5%
- Other: 14%
Distribution Publication Type

- Text (31.18%)
- Article, Journal (23.32%)
- Unknown Material (21.35%)
- Thesis (5.14%)
- Report, paper, lecture (4.50%)
- Book (3.18%)
- Primary Data (2.58%)
- Image (7.73%)
- Other (0.43%)
- Map (0.40%)
• 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
• Dublic 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>
An Oil-Based Lubrication System Based on Nanoparticular TiO2 with Superior Friction and Wear Properties

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)

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

- acm
- ddc
- jel
- journal
- mesh
- udc
- Sonstiges

95%
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
  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
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

Number of Docs

LCC Code

Universitätsbibliothek
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

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

- 0: 1 Mio.
- 1: 500 Tsd.
- 2: 500 Tsd.
- 3: 2 Mio.
- 4: 500 Tsd.
- 5: 3,5 Mio.
- 6: 3,0 Mio.
- 7: 700 Tsd.
- 8: 100 Tsd.
- 9: 100 Tsd.

DDC Code - first level
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 Distribution Character Length in Metadata
Language Distribution
Automatic Classification Processing Steps

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

Preprocessing Pipeline
Data workflow Classifier

Trainingsphase

1. Harvesting-Umgebung
2. UB
   - OAI-Metadaten + Volltexte
3. Computerlinguistik
   - Features

4. Klassifikator

Anwendungsphase

5. BASE-Index
6. DDC 636
7. DDC 636
   - DDC?
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:

- 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.
Thank You!

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