Universität Bielefeld

University Library
&
Text Technology

Building a DDC-annotated Corpus from OAI Metadata

Mathias Lösch[1], Ulli Waltinger[2], Wolfram Horstmann[1], & Alexander Mehler[2]
1 Bielefeld University Library
2 Text Technology/Applied Computational Linguistics (Bielefeld University)
Contact: Mathias.Loesch@uni-bielefeld.de

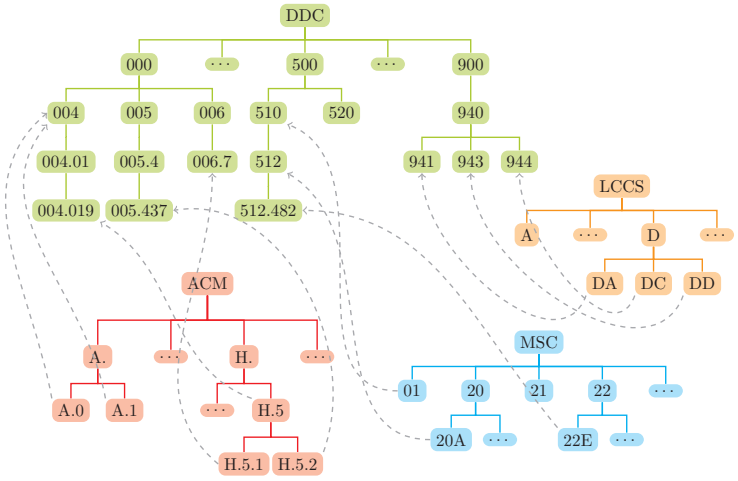
Motivation

HAVING subject-specific access to documents stored across different repositories would be a great advantage. However, this is not possible with the current OAI metadata, as there is no consistent subject indexing across repositories. Our aim is to address this problem using automatic document classification.

Method

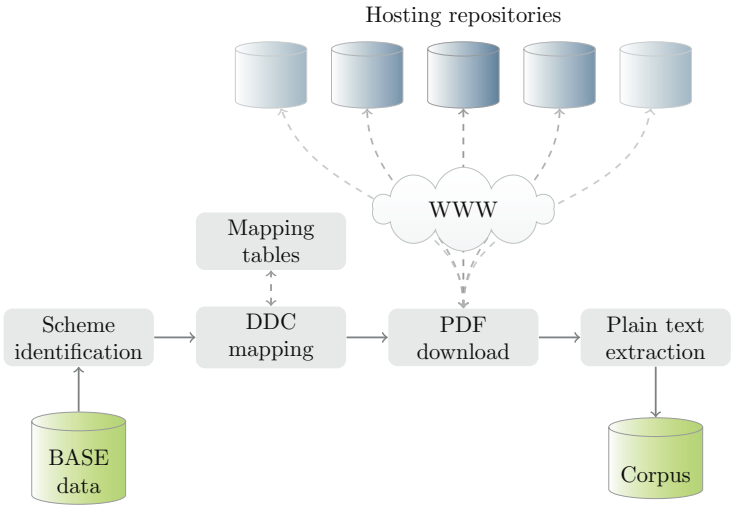
We built a bilingual text corpus of OAI records and the underlying *Open Access* full texts. Every document is annotated with a *Dewey Decimal Classification* (DDC) number. The corpus will be exploited as training data for text classification tasks to automatically enrich the subject indexing in OAI records with Dewey numbers.

Mapping to the DDC



In order to augment the OAI records with Dewey numbers, we map all sorts of classification schemes (both universal and subject-specific) to the DDC. Among the schemes we consider are the *Mathematics Subject Classification* (MSC), the *Library of Congress Classification Scheme* (LCCS), and the scheme of the *Association for Computing Machinery* (ACM).

Aggregating Documents



As a staple, we use the over 24,000,000 OAI records currently harvested for the *Bielefeld Academic Search Engine* (BASE). After determining the correct DDC number using our mappings, we download the full text from the hosting repository, generate a plain text version, and store everything in a database.

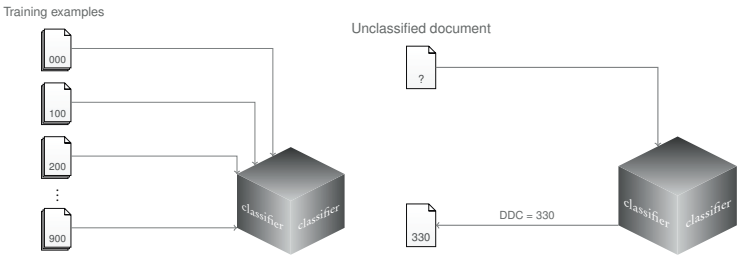
Corpus Statistics

DDC Class	English documents	German documents
000 Computer Science, information	6,836	3,645
100 Philosophy & psychology	3,444	1,952
200 Religion	1,075	1,859
300 Social sciences	10,445	7,323
400 Language	1,429	1,027
500 Science	23,884	6,694
600 Technology	6,219	5,596
700 Arts & recreation	1,094	3,438
800 Literature	674	1,917
900 History & geography	2,000	2,558
Total	**57,100**	**36,009**

Results & Discussion

- We were able to completely populate the first level of the DDC with records and documents.
- Large portions ($> 50\%$) of the second DDC level could be covered.
- Some classes (Mathematics, Physics, Computer Science) could even be covered with sample documents on the third DDC level.
- Our corpus still shows deficits especially in the humanities.
- This is mostly due to the lack of subject-specific classification schemes and *Open Access* full texts.

Future Application: Text Classification



Future Work

- We plan to use the aggregated DDC-classified records and documents as training data for automatic text classification.
- Thereby we want to automatically enhance the subject indexing of OAI records in the BASE data and provide new services using the so-enriched metadata, e.g. DDC-based browsing.
- We are going to overcome the current shortcomings of our corpus using additional data from *Wikipedia*.

References

[1] A. Mehler and U. Waltinger, "Enhancing Document Modeling by Means of Open Topic Models: Crossing the Frontier of Classification Schemes in Digital Libraries by Example of the DDC," *Library Hi Tech*, vol. 27, no. 4, pp. 520–539, 2009.

[2] F. Summann, "Open Access and Institutional Repositories: From Local Initiatives to Global Solutions," in *Proc. of CASLIN 2009*, Plzeň (Czech Repuplic), pp. 39–42, 2009.

Funding

This work was supported by the German Research Foundation (DFG) through the DFG-Project *Automatic enhancement of OAI metadata by means of computational linguistics methodology and development of services for a content-based network of repositories.*

DFG