Building a DDC-annotated Corpus from OAI Metadata

Mathias Lösch[1], Ulli Waltinger[2], Wolfram Horstmann[1], and Alexander Mehler[3]

[1] Bielefeld University Library
{Mathias.Loesch,Wolfram.Horstmann}@uni-bielefeld.de
[2] Faculty of Technology
Bielefeld University
uwalting@techfak.uni-bielefeld.de
[3] Department for Computer Science and Mathematics
Goethe University Frankfurt am Main
Mehler@em.uni-frankfurt.de

Abstract. Document servers complying to the standards of the *Open Archives Initiative* (OAI) are rich, yet seldom exploited source of textual primary data for research fields in text mining, natural language processing or computational linguistics. We present a bilingual (English and German) text corpus consisting of bibliographic OAI records and the associated full texts. A particular added value is that we annotated each record with at least one *Dewey Decimal Classification* (DDC) number, inducing a subject-based categorization of the corpus. By this means, it can be used as training data for machine learning-based text categorization tasks in digital libraries, but also as primary data source for linguistic research on academic language use related to specific disciplines. We describe the construction of the corpus using data from the *Bielefeld Academic Search Engine* (BASE), as well as its characteristics.

Keywords: Digital libraries, text mining, corpora, Dewey Decimal Classification

1 Introduction

The ongoing rise of digital libraries that store and disseminate the academic output of individual institutions—so-called *institutional repositories*—is vitally connected to the success of the *Open Archives Initiative* (OAI) and its *Protocol for Metadata Harvesting* (OAI-PMH). The key concept of OAI-PMH is to ease metadata exchange between digital libraries using a standardized XML-based format for encoding bibliographic records, thereby fostering the fast distribution of scholarly information (Lagoze and Van de Sompel, 2001).

Despite this success, however, little attention has been given to the OAI domain as a source of *primary data*. Since most of the metadata distributed via the protocol describes academic publications, it is potentially interesting to a range of fields including linguistics, natural language processing, and text mining. Recently, metadata records in the OAI *Dublin Core* (OAI DC) format

have already become subjects of interest to the area of machine learning. Text categorization and clustering techniques have been used to enhance search and browsing in digital libraries, with OAI DC records being used as surrogates for the actual documents in order to minimize computational costs (cf. Krowne and Halbert, 2005; Hagedorn et al., 2007; Newman et al., 2007; Mehler and Waltinger, 2009).

In this paper, we present a text corpus built from OAI DC metadata. The corpus comprises not only the actual bibliographic records, but also the underlying full texts. Furthermore, we annotate each record with at least one *Dewey Decimal Classification* (DDC, Dewey and Mitchell, 2003) number—this provides a subject-categorized view on the corpus making it suitable (though not exclusively) for experiments in text categorization. We choose the DDC as the target category system because it is one of the most widely used universal classification schemes worldwide,[4] and has recently been successfully tested as the target scheme for text categorization in digital libraries (Mehler and Waltinger, 2009; Wang, 2009).

The rest of this article is organized as follows. Starting from the description of the actual construction of the corpus (Section 2), we give insights into its properties in Section 3. Finally, we conclude and give a prospect on future work in Section 4.

2 Corpus Construction

In this section, we describe the aggregation and preprocessing of the input documents (Section 2.1), as well as their representation and organization in the corpus (Section 2.2).

2.1 Aggregation of Documents

Our starting point for constructing the corpus is the OAI DC metadata aggregated for the *Bielefeld Academic Search Engine* (BASE, cf. Pieper and Summann, 2006). This data basis currently comprises more than 26 million bibliographic records coming from over 1,700 repository servers. The data is available in the form of XML encoded files, as specified by OAI DC (Lagoze and Van de Sompel, 2001). See Figure 1 for the content-related part of a typical OAI DC record. Note that the fields are equivalent to the 15 metadata elements recommended by the *Dublin Core Metadata Initiative* (DMCI, cf. Dublin Core Metadata Initiative, 2008).

In order to transfer the data to the corpus, we have developed a software routine that automates most of the process. Since the raw OAI DC data is organized in a per-repository manner in the BASE environment, we can attune the program to the specific properties of individual repositories. This customization involves aspects like the location of the full text and the nature of the subject indexing information that is provided in the records.

[4] According to its maintainer *Online Computer Library Center, Inc.* (OCLC), 200,000 libraries use the classification system worldwide. See `http://www.oclc.org/dewey/` [accessed November 22, 2010].

```
<record>
  ...
  <metadata>
    <oai_dc:dc xmlns:xsi="...">
       <dc:title>
         Bielefeld Academic Search Engine (BASE): an end-user oriented
         institutional repository search service
       </dc:title>
       <dc:creator>Pieper, Dirk</dc:creator>
       <dc:creator>Summann, Friedrich</dc:creator>
       <dc:subject>LS. Search engines.</dc:subject>
       <dc:subject>HS. Repositories.</dc:subject>
       <dc:description>Purpose: This paper describes ...</dc:description>
       <dc:publisher>Emerald</dc:publisher>
       <dc:date>2006</dc:date>
       <dc:type>Journal Article (Print/Paginated)</dc:type>
       <dc:type>PeerReviewed</dc:type>
       <dc:format>application/pdf</dc:format>
       <dc:relation>
         http://conference.ub.uni-bielefeld.de/2006/proceedings/pieper_summann_final_web.pdf
       </dc:relation>
       <dc:identifier>http://eprints.rclis.org/9160/</dc:identifier>
       <dc:language>en</dc:language>
    </oai_dc:dc>
  </metadata>
</record>
```

Fig. 1: The content-related part of the OAI DC metadata record for a publication (Pieper and Summann, 2006). Dots indicate omitted content.

To be included in the corpus, a record has to fulfill two requirements: First, the underlying full text has to be freely downloadable (Open Access) in a machine-readable format (usually PDF) from an URL that is specified in the record. Second, we need to be able to determine its correct Dewey number. While the former issue depends solely on the policies of the hosting repository and/or the author of the document, the latter can be addressed in different ways. A number of repositories use DDC numbers in their records by default—those can simply be imported to the corpus directly. Others, however, use different knowledge representation systems for subject indexing, e.g., subject-specific classification schemes or subject headings. We capitalize on that by implementing an automatic mapping routine using cross-concordances between various subject indexing schemes and the DDC. The cross-concordances have been constructed manually.

To some degree, we can also identify the type of subject indexing scheme automatically by analyzing its notation structure. This becomes necessary if various schemes are used in the same repository and we need to dynamically select the appropriate cross-concordance table.

Once the DDC number is known and the PDF file has been downloaded, a plain text representation is generated and passed to a language identification routine. This step is indispensable because not every repository correctly employs the Dublin Core *language* field. Also, we currently only include documents in English and German.

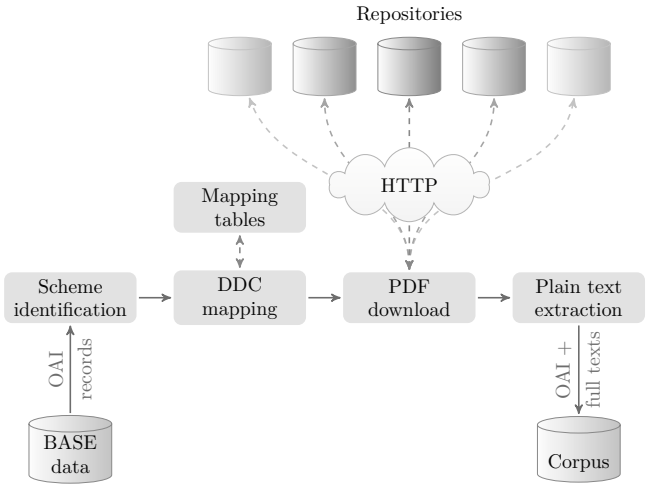

Fig. 2: The procedure of aggregating OAI DC records and full texts.

In order to avoid duplicates in the corpus, we compute an MD5 hash over the plain text representation of the full text document. The resulting 32-digit hexadecimal number serves as unique identifier for the record and the full text. Before storing a new document, its hash value is compared against all identifiers in the corpus and rejected in the case of a match.

Finally, both the record and the plain text are stored in association with their DDC numbers. Figure 2 schematically displays the process of aggregating documents for the corpus.

2.2 DDC-Annotation and Categorization

Although it is widely agreed in the area of corpus linguistics that annotations by the corpus creators should be kept separately from the primary data ("stand-off annotation", cf. Ide and Brew, 2000), we do not follow this principle here. The reason is that since an OAI DC record is already a well-structured representation of the underlying document, we can cleanly nest the annotations into the existing XML data (see Figure 3). An annotation consists of at least one DDC number and the unique identifier produced by the hashing algorithm (see previous section). Note that the added data resides in its own namespace so that it can be easily identified and filtered out by a standard XML parser.

While the full texts are stored as text files, the annotated OAI DC records are managed in an SQL database accessible via an HTTP interface—the rationale is to ease subject-based access to them by providing a DDC-categorized view on

```
<record>
 ...
 <metadata>
   <oai_dc:dc xmlns:xsi="...">
      <dc:title>
        Bielefeld Academic Search Engine (BASE): an end-user oriented
        institutional repository search service
      </dc:title>
      <dc:creator>Pieper, Dirk</dc:creator>
      <dc:creator>Summann, Friedrich</dc:creator>
      ...
   </oai_dc:dc>
 </metadata>
 <ubi:ubimeta xmlns:ubi="http://www.ub.uni-bielefeld.de/">
   <ubi:ddcnumbers>
     <ubi:ddc>025.04</ubi:ddc>
   </ubi:ddcnumbers>
   <ubi:uuid>55a1d25a8e3e65228a81d052ffefb570</ubi:uuid>
 </ubi:ubimeta>
</record>
```

Fig. 3: The record of Figure 1, annotated with a Dewey number and a unique identifier (dots indicate omitted content).

the corpus. The connection between a record and its full text is established by the unique identifier.[5]

3 Corpus Characteristics

In this section, we describe the properties of the corpus. We start by investigating general statistics including document counts and disk usage (Section 3.1). Next, we look more closely at the DDC structure in terms of feature selection on the lexical level (Section 3.2).

3.1 Corpus Statistics

Table 1 shows the general statistics of the corpus in terms of document numbers and disk space. Note that since we continually monitor the BASE data for new records, the collection is still growing and the values are subject to change.

When populating the DDC classes with documents, we pursue the goal of fully covering the semantic concept they represent by aggregating as many positive instances as possible. In the case of a text categorization scenario, such a corpus will enable the learning algorithm to generalize these concepts from the variety of training instances. However, since the number of such class instances is potentially very large for some classes (e.g., we could aggregate hundreds of thousands of documents for the subject of physics alone due to the strong Open Access tradition

[5] This implementation could be refined in the future by employing the RDF-based OAI-ORE protocol (Van de Sompel and Lagoze, 2007) for associating the metadata with their full texts and the Dewey numbers, utilizing the now available *Linked Data* web service for the DDC (cf. `http://dewey.info`).

Quantity	Value
No. of data sources (repositories)	101
No. of unique documents (English)	52,905
No. of unique documents (German)	37,228
Total size (OAI DC records)	439 MB
Total size (full texts)	13 GB

Table 1: General corpus statistics (disk space values refer to uncompressed UTF-8 encoded XML/text data).

DDC	English	German
000 Computer Science, information & general works	6847	3778
100 Philosophy & psychology	3536	2169
200 Religion	1123	1973
300 Social sciences	10948	8075
400 Language	1682	1297
500 Science	23989	6969
600 Technology	6669	5874
700 Arts & recreation	1280	3823
800 Literature	740	2063
900 History & geography	2226	2863

Table 2: Distribution of English and German documents across the top-level DDC classes. Note that a document can be a member of more than one class.

in this discipline), we need to introduce document limits per class. On the one hand this is necessary to keep the corpus computationally manageable, on the other hand we have to avoid unbalanced class sizes since this would cause bias towards the larger classes. We set the limit to 100 documents for third-level classes, resulting in limits of 1,000 documents for second-level, and 10,000 documents for top-level classes. If a class reaches its limit, we stop deliberately aggregating documents for it. Note that notwithstanding this condition, such a class can still grow through multiple categorized documents.

By this means, all documents in the corpus are categorized at least into one of the ten main classes of the DDC. Moreover, the documents also cover most of the second-level divisions and even some third-level sections of the DDC. It turns out, however, that we cannot balance the numbers of examples for all classes. This is mostly due to the lack of Open Access documents in some subjects (e.g., in the humanities), but also because of the structure of the DDC: although its decimal notation suggests a strictly hierarchical architecture, there are classes that violate this principle. The reasons for inconsistent structures in the DDC are historical as they result from its development according to the principle of the *literary warrant* (cf. Mitchell, 2001). That is, whether a new class is included in the classification system, and also its position in the hierarchy, are only determined by the literature that is to be indexed by the system at the given time (Beghtol,

1986). This procedure naturally leads to inconsistencies in the hierarchy, and again causes some classes to be less populated with example documents. Table 2 shows the distribution of documents across the DDC top-level classes and reveals its skewness.

3.2 Informative Terms

One of our purposes to build a DDC-related corpus is to provide training and test data for experiments in text categorization. Since most approaches to text categorization use lexical features to learn categories of documents, these features should reflect the structure of the underlying classification scheme. Otherwise, they would hardly enable the learning algorithm to effectively discriminate between the categories. In order to get a first idea of the data in this regard, we explore discriminative lexical features per top-level class of the DDC.

We measure, so to speak, the information value of a lexical feature or term t in relation to the DDC class c using the chi-square (χ^2) test (cf. Yang and Pedersen, 1997). The χ^2-test measures the lack of independence between t and c, and is computed as follows:[6]

$$\chi^2_{t,c} = \frac{N(AD - BC)^2}{(A + B)(A + C)(B + D)(C + D)} \tag{1}$$

where A is the number of documents of class c that contain t, B is the number of documents of any other class $d \neq c$, which contain t, C is the number of documents of class c that do not contain t, D is the number of documents of any other class $d \neq c$, which do not contain t, and, finally, N is the total number of documents in the corpus. The value of $\chi^2_{t,c}$ is zero if t and c are independent.

We compute the χ^2-scores on the English OAI DC corpus using the concatenated *title, subject* and *description* fields of the records as documents. However, we restrict the experiment to records containing more than 500 bytes of text (32,254 in total) in order to reduce computational cost.

Table 3 shows the five terms for each top-level class of the DDC that are top-scored by the χ^2-test. Note that we eliminate stop words and non-lexical items, transform all terms to lowercase, and apply stemming (Porter, 1980). Apart from some anomalies—for example there seems to be a prevalence of economic terms in the class 300 (Social sciences)—, the overall impression is that the selected terms are actually good candidates for representing their class, which suggests that the corpus is indeed suitable for DDC-based text categorization.

4 Conclusion

We presented the construction and the characteristics of a bilingual text corpus that we built using OAI DC metadata. Its DDC-based categorization makes the

[6] This is only done as a pretest as the chi-square statistics is known to be outperformed by other functions used for feature selection—see Sebastiani (2002) for an overview on such methods.

DDC	Top χ^2-scored terms
000	librari, comput, user, scienc, journal
100	psycholog, cognit, philosophi, mind, conscious
200	religion, church, religi, christian, theolog
300	market, polici, countri, wage, firm
400	languag, linguist, english, semant, syntact
500	physic, mathemat, energi, librari, polici
600	engin, biolog, cell, machineri, fibr
700	music, art, design, architectur, theatr
800	literari, australian, fiction, poetri, literatur
900	archaeolog, histori, songster, geographi, war

Table 3: Top χ^2-scored terms in the DDC top-level classes (note that the terms are in their stemmed forms)

.

corpus suitable for research tasks like automatic text classification. Despite some drawbacks regarding an imbalanced distribution of documents across the classes, a preliminary experiment using the χ^2-test indicated that the class structure is reflected by linguistic properties of the documents in the top-level classes.

Future work will include filling sparsely populated DDC classes with more documents. This can be accomplished by including more repositories and/or simply by continually monitoring the BASE data, since new academic texts are being constantly published. Another aspect could be the inclusion of additional annotations, for example part-of-speech tags, to make the corpus more suitable for linguistic research questions. We plan to make the corpus available to researchers on request.

Acknowledgement

We gratefully acknowledge financial support of 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.*

References

Beghtol, C. (1986). Semantic validity: Concepts of warrant in bibliographic classification systems. *Library resources & technical services 30*(2), 109–125.

Dewey, M. and J. S. Mitchell (2003). *Dewey Decimal Classification and relative index* (22 ed.). Dublin, Ohio: OCLC Online Computer Library Center.

Dublin Core Metadata Initiative (2008). Dublin core metadata element set, version 1.1. `http://dublincore.org/documents/2008/01/14/dces/` [accessed November 22, 2010].

Hagedorn, K., S. Chapman, and D. Newman (2007). Enhancing search and browse using automated clustering of subject metadata. *D-Lib Magazine 13*(7/8).

Ide, N. and C. Brew (2000). Requirements, tools, and architectures for annotated corpora. In *Proceedings of data architectures and software support for large corpora*, Paris (France), pp. 1–5. European Language Resources Association.

Krowne, A. and M. Halbert (2005). An initial evaluation of automated organization for digital library browsing. In *Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries*, pp. 246–255. ACM.

Lagoze, C. and H. Van de Sompel (2001). The Open Archives Initiative: Building a low-barrier interoperability framework. In *Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries*, pp. 54–62. ACM.

Mehler, A. and U. Waltinger (2009). 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 27*(4), 520–539.

Mitchell, J. S. (2001). Relationships in the Dewey Decimal Classification system. In C. A. Bean and R. Green (Eds.), *Relationships in the organization of knowledge*, pp. 211–226. Boston: Kluwer.

Newman, D., K. Hagedorn, C. Chemudugunta, and P. Smyth (2007). Subject metadata enrichment using statistical topic models. In *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, pp. 366–375. ACM.

Pieper, D. and F. Summann (2006). Bielefeld Academic Search Engine (BASE): An end-user oriented institutional repository search service. *Library Hi Tech 24*(4), 614–619.

Porter, M. (1980). An algorithm for suffix stripping. *Program 14*(3), 130–137.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys 34*(1), 1–47.

Van de Sompel, H. and C. Lagoze (2007). Interoperability for the discovery, use, and re-use of units of scholarly communication. *CTWatch Quarterly 3*(3), 32–41.

Wang, J. (2009). An extensive study on automated Dewey Decimal Classification. *Journal of the American Society for Information Science and Technology (JASIST) 60*(11), 2269–2286.

Yang, Y. and J. O. Pedersen (1997). A comparative study on feature selection in text categorization. In *Proceedings of the 14th International Conference on Machine Learning (ICML'97)*, Nashville, TN, USA, pp. 412–420.