Towards an IR Test Collection for the German National Library

Johanna Munkelt[1], Philipp Schaer[2], and Klaus Lepsky[2]

[1] Fachhochschulbibliothek Dortmund, 44227 Dortmund, Germany
johanna.munkelt@fh-dortmund.de
[2] Technische Hochschule Köln, 50678 Cologne, Germany
firstname.lastname@th-koeln.de

1 Introduction

Automatic content indexing is one of the innovations that are increasingly changing the way libraries work. In theory, it promises a cataloguing service that would hardly be possible with humans in terms of speed, quantity and maybe quality. The German National Library (DNB) has also recognised this potential and is increasingly relying on the automatic indexing of their catalogue content. The DNB took a major step in this direction in 2017, which was announced in two papers. Since September 2017, the DNB has discontinued the intellectual indexing of their series B (monographs and periodicals outside the publishing industry) and H (university publications) and has switched to an automatic process for these series. The subject indexing of online publications (series O) has been purely automatical since 2010. This again raises the well-known question: What is the quality of the automatic indexing compared to the manual work or in other words to which degree can the automatic indexing replace people without a significant drop regarding quality?

As an argument for the conversion, the DNB primarily mentions the progressing modernisation and the currently prevailing heterogeneity in the subject indexing of the various document types. The DNB intends to make a large proportion of access to document types accessible in a uniform manner using a process that saves time and is intended to contrast this heterogeneity with a more homogeneous subject indexing.

A critical article by Klaus Ceynowa [1], General Director of the Bayerische Staatsbibliothek, which was published in the Frankfurter Allgemeine Zeitung at the end of July 2017, triggered a discussion about the DNB's decision. Heidrun Wiesenmüller, Professor for Library and Information Management at the Stuttgart Media University, also discusses the topic and shows some understanding for Ceynova's position, although she puts some of his criticisms into perspective and shows understanding for several arguments of the DNB [6].

A typical counterargument put forward by critics of automatic indexation is that the quality of content extraction suffers. A regulation which favours a uniform development of more documents than before is only justified if the general development standard does not fall to a lower, “unacceptable level” [1], without specifying the actual thresholds for such a measure, which Wiesenmüller recommends. She points out that there aren't any common quality measures at the moment accepted in the community. In order to set up certain quality standards a discussion that includes the concerned interest groups would be necessary.

The decision of the DNB and the various opinions on the same one raised the following question: Is automatic content indexing ready to satisfy the demands of librarians? We argue that it is necessary to carry out tests that examine the quality of the indexing by looking on the retrievability of the content. So far there is no evaluation of the subject indexing quality of the DNB data. It is impossible to make a comparison with previous practice because there is a lack of suitable tools, for example a test collection.

A test collection designed specifically for the purpose of testing the subject indexing quality in DNB data is intended to provide a remedy. With the help of this collection a retrieval test and thus an objective evaluation is made possible, which so far could not take place and is nevertheless of great importance for the further developments at the DNB.

Currently there is a project group working on constructing and working with such a test collection [3] at the Department of Information Science at the Technische Hochschule Köln. The test collection should be based on principles similar to those used by the TREC community. In the following, the development process, the most relevant steps of construction and the prospects of using a test collection are presented.

2 Designing a Test Collection with DNB Catalogue Data

Three basic building blocks are required for a standard retrieval test and its successful execution: (1) a document collection, (2) real information needs, which are imitated with the help of so-called topics with a short description and a more precise explanation (narrative), and (3) relevance assessments that are made for the result sets from the collection for the respective topics.

With reference to the change in the subject indexing practice of DNB explained at the beginning, some special features can be noted for a retrieval test and the underlying test collection:

- the test collection consists of sufficient data taken from the DNB database,
- the data must be processed accordingly and freed from unnecessary ballast for the sake of clarity,
- the content distribution of the functional areas is at best based on the distribution of functional areas within the overall data of the DNB,
- the topics for which relevance assessments are to be made are not linked to a specific subject area in terms of content, but are as broadly dispersed as possible,
- a suitable procedure must be chosen for the preparation of the relevance judgements, which is both practicable and expedient,
- the implementation of all planned steps must take place,
- the test collection must be tested for functionality and suitability.

2.1 The Document Corpus

The corpus supplied by DNB consists of 200,000 documents and is broadly diversified with a content distribution of over 100 subject groups. Of the data records, 131,538 document contents were intellectually indexed and 68,462 document contents machine-accessed; the share of machine-accessed document contents is therefore around 35 percent. DNB's general catalogue contains more than 32 million records. For technical reasons, the entire data pool of the DNB cannot be used as a corpus for the test collection. The supplied Pica+ format is an internal bibliographic data format of special software, which in its pure form has very poor readability. The Pica+ format had to be converted into a clearer format, in this case a Solr-friendly XML. In this transformation, categories containing unnecessary information such as internal DNB codes were filtered out of the metadata, for reasons of clarity and data record reduction. The categories may have a different value internally in the DNB, but in this case they lead to a very confusing view of the documents and therefore were removed.

2.2 Topics and Pooling

An existing topic pool was used as a basis for creating the topics. The topics that were already used in a retrieval test in the MILOS II project [4] were considered for this purpose. The MILOS II project dealt with the automatic indexing of catalogue data and was carried out in 1998 with data from the DNB. The data for the MILOS II retrieval test is comparable to the data of the test collection that is the subject of this work. The written results for the MILOS II test include 100 topics. These topics were created for a corpus consisting of title recordings of the DNB and covering all subject groups, with the exception of fiction, children and young people literature, and calendars. The database is therefore based on similar requirements. The number of topics to be created has been set to 50. The 100 topics from MILOS II were sifted and filtered. In particular, the topics were removed that had led to very few or no hits, since some of them are very specific. As with TREC, the topics consist of a topic ID, the topic title, a description that summarises in one sentence what is searched for, and the narrative, an explanation of what is relevant, partially relevant and not relevant for the topic. The description is on average 8.44 words long, the narrative is much more specific with 48.26 words on average. Description and narratives are also of great importance for the reuse of the test collection. One can use these topic components to find out why certain documents were evaluated as relevant or not relevant for a topic.

The corpus of the test collection consists of 200,000 documents and 50 topics have been selected, which form the foundation for the search queries. The ideal case, in which a relevance judgement is available for each individual document topic pair, would mean the production of 10 million relevance judgements, which must be assessed by human judges. To avoid this enormous amount of work, a document pool was created for each topic, which was used for the relevance assessments.

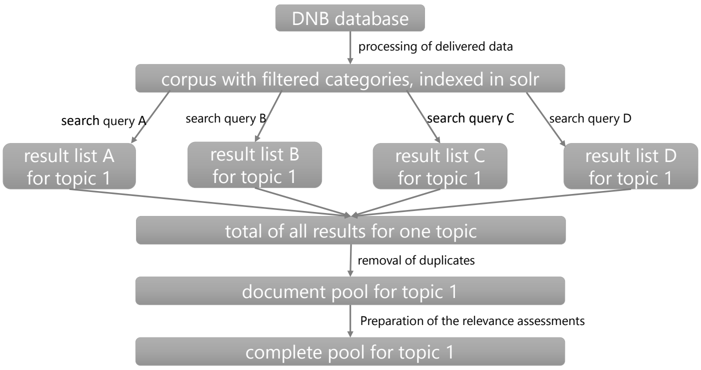
DNB database
processing of delivered data
corpus with filtered categories, indexed in solr
search query A
search query B
search query C
search query D
result list A for topic 1
result list B for topic 1
result list C for topic 1
result list D for topic 1
total of all results for one topic
removal of duplicates
document pool for topic 1
Preparation of the relevance assessments
complete pool for topic 1

Fig. 1. Overview on the pooling process.

The pooling procedure used for this (see Figure 1) is based on the procedure used at TREC for pool creation as part of the ad hoc task. At TREC, the systems of different participants search in the same corpus for the same topic. Since the systems usually work differently, the result sets differ. The top k ranked results are transmitted to TREC, from which a large pool is formed and this pool is evaluated. For the creation of these test collections, it was not possible to use different systems that searched the corpus. This results in slight changes in the pooling process. Four people with a background in library sciences and previous knowledge of the DNB corpus developed a search strategy for each topic and formulated a search query. This means that there are four search queries for each of the 50 topics, whose result sets can be combined into a pool. The four search queries for each topic are documented with the respective result sets. For each topic, the result sets are more or less similar. For topic 03 (Gerontology), for example, 31, 90, 179 and 89 records were found using different search strategies. The duplicates were removed and the remaining documents were combined in a pool. Such a pool was created for each topic. These pools are the basis for the relevance assessment. Some topics are characterised by a higher number of records than others, so the pools are of different sizes. The majority of the results per pool are between 50 and 150 documents. The smallest pool consists of 13 records (39 - Medicine in the Third Reich), the largest pool consists of 514 records (14 - Alternative energies), such outliers are exceptions.

2.3 Relevance Assessments

With the help of the open-source toolkit Relevation! [2] a total of 6,984 relevance assessments were made. Relevation! ensures a clearer presentation of the documents from the respective pools for a topic, simplifies the evaluation and finally outputs a result list of the relevance ratings in TREC format. The relevance was documented in graded form. A hit in a result set can therefore have three values: “0” for not relevant, “1” for partially relevant and “2” for relevant. From

Table 1. Examples of topics with topic ID, description and narrative (in German).

ID	Topic	Description	Narrative
24	Geschichte Israels	Finde Dokumente, die sich mit der Geschichte des Staates Israel beschäftigen.	Relevante Dokumente behandeln die Geschichte des Staates Israel und die territorialen Dispute, die damit einhergehen. Relevant sind auch Dokumente, die sich mit dem historischen, kulturellen und religiösen Israel befassen und dessen Bedeutung für den gegenwärtigen Staat und die Konflikte. Teilweise relevante Dokumente behandeln die Geschichte Israels im Hinblick auf bestimmte Einzelaspekte. Nicht relevant sind Dokumente, in denen das Wort Geschichte im Sinne von "Erzählung" vorkommt.
25	Heilfasten	Finde Dokumente, die sich mit Heilfasten beschäftigen.	Relevante Dokumente behandeln unterschiedliche Methoden von Nulldiäten und Heilfasten und der Wirkung auf die körperliche und seelische Gesundheit sowie den Risiken, die damit verbunden sind. Relevant sind auch Dokumente, die sich mit dem esotherischen Aspekt von Fasten und Diät beschäftigen. Teilweise relevant sind Dokumente, die das religiöse Fasten zum Thema haben.
26	Elektroautos	Finde Dokumente, die sich mit elektrisch betriebenen Fahrzeugen befassen.	Relevante Dokumente behandeln die Entwicklung von Kraftfahrzeugen, die mit Elektromotoren betrieben werden und die Bedeutung für die Energiewirtschaft. Relevant sind auch Dokumente, die sich mit dem Umweltaspekt elektrischer Motoren beschäftigen. Teilweise relevant sind Dokumente, die Elektromobilität im Allgemeinen mit dem Thema Elektroautos verbinden.
27	Homöopathische Mittel	Finde Dokumente, die sich mit homöopathischen Mitteln und ihren Wirkungen beschäftigen.	Relevante Dokumente behandeln die verschiedenen homöopathischen Wirkstoffe sowie die Diskussion um Wirksamkeit homöopathischer Mittel allgemein. Auch tiermedizinische Dokumente sind relevant.

the total number of 6,984 judgements 909 were relevant, 1,457 were partially relevant, and 4,616 were not relevant. In the preparation of the judgements, care was taken to ensure that they are based on the guidelines from the topics and the narratives. However, relevance assessments can never claim to be perfect because of their subjectivity. The use of a pooling procedure also does not guarantee the completeness of the relevant assessments. In this case, the pooling procedure is a compromise between the amount of work, the time available and the human and monetary resources.

All topics and all relevance assessments were prepared by one person or finally revised. This avoids that different relevance judgements of different persons can influence the results of possible retrieval tests. It is impossible to completely avoid subjectivity in the relevance judgements. However, if the judgements on a topic come from the same person, they are stringent in themselves and can be considered valid.

3 Conclusion and Future Work

The final test collection consists of the three essential components: (1) 200.000 revised document records containing all content-bearing fields of original DNB title recordings, (2) real information needs in the form of 50 topics with ID, title, description and narratives, and (3) relevance judgements that were made using a pooling process and the Relevation! toolkit.

The test collection can serve as a basis for retrieval tests to answer the question on the quality of the automatic indexing practices at DNB. By documenting all work steps and thought processes, the construction of the test collection as such has been made comprehensible. All data is static, the re-use of the collection with the contained documents, topics and relevance judgements is possible. The test collection can be reused, transferred to other systems or adapted to the requirements of other retrieval tests. The wide range of subject groups makes the test collection suitable as a general test collection with library metadata from the German National Library. The size of the collection also facilitates reuse. The collection itself is not published yet but should be available in autumn 2018.

The general feasibility of the IR collection was tested during the construction process. It was not yet used to evaluate the initial questions on indexing quality of the new DNB procedures. While this would be the primary area of application after the completion of this retrieval test, it should be assessed to what extent the performance of the test collection is convincing and which further areas of application are conceivable.

In the area of relevance assessments, there is still a need to expand the test collection. With 200,000 documents and 50 topics, there are 10 million possible relevance assessments, of which almost 7,000 were compiled in this case. An expansion of this quota is conceivable with a corresponding investment in working time. Also interesting can be the use of the document set with a new set of topics, to which relevance judgements can be made. New topics and more specific topics are an option for further areas of application of the test collection due to the deliberately general distribution of subject areas in the document quantity. Finally the test collection should be tested for inter-rater reliability by drawing up various relevance assessments for the existing topics [5].

References

1. Ceynowa, K.: Deutsche Nationalbibliothek: In Frankfurt lesen jetzt zuerst Maschinen. FAZ.NET (Jul 2017), `http://www.faz.net/1.5128954`
2. Koopman, B., Zuccon, G.: Relevation!: An open source system for information retrieval relevance assessment. In: SIGIR '14. pp. 1243–1244. ACM, New York, NY, USA (2014), `http://doi.acm.org/10.1145/2600428.2611175`
3. Munkelt, J.: Erstellung einer DNB-Retrieval-Testkollektion. Bachelor thesis, Technische Hochschule Köln (2018)
4. Sachse, E., Liebig, M., Gödert, W.: Automatische Indexierung unter Einbeziehung semantischer Relationen: Ergebnisse des Retrievaltests zum MILOS II-Projekt. Fachhochschule Köln, Fachbereich Bibliotheks- und Informationswesen, Köln (1998)
5. Schaer, P.: Better than their reputation? on the reliability of relevance assessments with students. In: Information Access Evaluation. Multilinguality, Multimodality, and Visual Analytics. pp. 124–135. Springer, Berlin, Heidelberg (2012)
6. Wiesenmüller, H.: Das neue Sacherschließungskonzept der DNB in der FAZ (Aug 2017), `http://www.basiswissen-rda.de/neues-sacherschliessungskonzept-faz/`