Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

Evaluation Methods within the LibRank Project

Working Paper
2016

www.librank.info

Working Paper

Version 1.2

Evaluation Methods within the LibRank Project

Christiane Behnert

2016

This working paper describes methodological issues that have been determined within the research project LibRank funded by the German Research Foundation (DFG – Deutsche Forschungsgemeinschaft). The final methods applied may slightly differ from the descriptions in this paper, as adjustments to the research design had been made based on results of a pretest and the first evaluation run. Project results will be published elsewhere. This working paper has not been peer-reviewed.

Author: Christiane Behnert

Institution: Hamburg University of Applied Sciences (HAW)
Faculty of Design, Media, Information
Department Information
Finkenau 35
22081 Hamburg

E-Mail: christiane.behnert@haw-hamburg.de

Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

Table of Contents

List of figures

List of tables

0. Definitions

Test system = mirrored EconBiz at time x (static)

Test system + data = **Test environment** (static)

Test environment + ranking parameters incl. weighting = **(Test) Ranking** (dynamic due to adjustment after every analysis)

Evaluation run: In one run, the same search queries will be submitted to each **Ranking**.

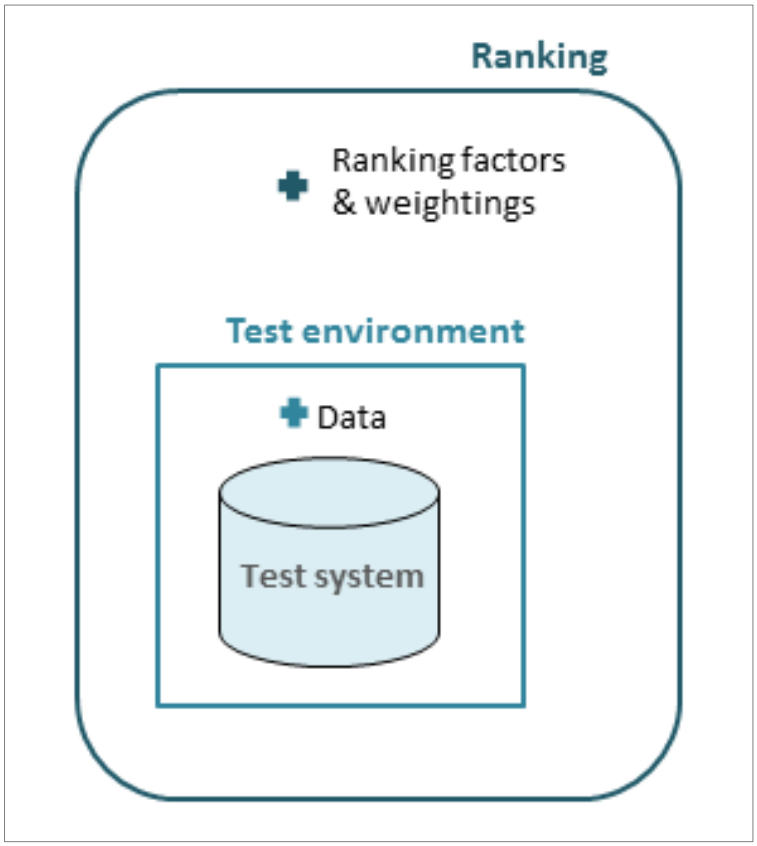

Figure 1: Definitions

1. Introduction

The aim of the evaluation phase is to test the impact of the implemented ranking methods on the retrieval effectiveness of the search system. The **research questions** (RQ) to be answered are:

I. What possible factors for relevance ranking exist with regard to library information systems?
II. Which factors or factor groups are promising for the evaluation?
III. Which factors or factor groups are feasible for the evaluation?
IV. Which factors show the most significant impact on the retrieval effectiveness:
    a) within a particular group,
    b) in combination with other factors within one group,
    c) in combination with factors of other groups?
    d) Which factors - individual or combined - increase, neutralize or contradict each other?

After the identification of suitable ranking factors (see S*tate of the Art report* as the result of WP1b and with regard to RQ I & II) and the data collection (see *Data Availability Report*[1] as the result of WP1a and with regard to RQ III), the test system which is a mirrored database of the ZBW provided information portal EconBiz has been implemented (as the result of WP2a).

Before the beginning of the evaluation phase, the test system will be updated to provide the current EconBiz data set. This is very important, also because the *Standard Thesaurus Wirtschaft* (STW) update with its restructured version may influence the retrieval performance due to revised subject terms. To start the evaluation phase, the methods to be used need to be analyzed and adapted to the EconBiz context (as the result of WP2b and with regard to RQ IV). In this study, we can adopt part of the evaluation methods of the "Framework for evaluating the retrieval effectiveness of search engines" proposed by Lewandowski (2012) to the EconBiz context. Since EconBiz is an information system which is based on search engine technology – the open source discovery software *VuFind* – there are evaluation issues compliable to the framework.

Although the following sections are not structured in accordance with this framework (see Figure 2), many points are taken from it and are categorized slightly different: Firstly, we describe the presentation, different types, the sources and number of search queries (section 2); then we present a concept for selecting the assessors (section 3), followed by the results that will be judged by the assessors using a special software tool (section 4). After these theoretical considerations that need to be made prior to the actual testing, including a pretest (section 5), the data gathered from the evaluation runs will be analyzed (section 6). It will be emphasized how queries, assessors, results and data analysis for evaluating factors for relevance ranking do not differ much from the web search engine perspective, but need to be altered appropriately for our library information system.

Framework for Evaluating the Retrieval Effectiveness of Search Engines				
1. QUERY SELECTION	**2. RESULTS COLLECTION**	**3. RESULTS PRESENTATION**	**4. RESULTS JUDGEMENT**	**5. DATA ANALYSIS**
Intents	Descriptions	Position of result description in SERP	Selection of jurors	Relevance of the results
Topics	Results	Screen real estate of description	Jurors per query	Results descriptions
Generating descriptions	Other elements (ads, one-box results)	Graphical elements	Scales	Diversity
Generating query / topic aspects	(Results classification)			Other analysis (based on results classification)
Query properties (demographic etc.)				

Figure 2: A model for evaluating the retrieval effectiveness of search engines (modified after Lewandowski, 2012, p. 465)

2. Search queries

2.1 Query presentation

Search queries or tasks must be presented to the assessors in a comprehensible way, so that the assessors are able to decide whether a certain result is relevant to the query or not. It is necessary to

[1] As the *Data Availability Report* is for internal use only, it will not be publicly available.

provide a sufficient amount of additional information for assessing results, because "obtaining multiple representations of a single information need is a better approach to representing user needs than relying on solitary, isolated queries" (Kelly & Fu, 2007, p. 31). Therefore, the queries will be enriched with a **description** of the particular underlying **problem-oriented information need**.

With regard to the query intent, there are two kinds of information needs: concrete vs. problem-oriented (Frants, Shapiro, & Voiskunskii, 1997). Compared to the concrete information need which usually can be satisfied with the one correct answer or fact (e.g., "Who is CEO of company X?"), the problem-oriented information need requires several documents and the search success is not objectively measurable (e.g., "What approaches are being evolved to avoid rising unemployment figures?"). As will be described in more detail in section 2.2, we will focus on queries intending to satisfy problem-oriented information needs, which means that "*the assessment of relevance will be different – also depending on who are asking and their respective context*" (Stock & Stock, 2013, p. 106). Taking this statement into account, we present a search task, for example, as follows:

- Search query: *Prinzipal-Agenten-Theorie*
- Description: *Was sagt diese Theorie aus und wo findet sie Anwendung?*

We can abstain from displaying further assessment information in the form of: "*Relevant documents are documents that contain information on…*" According to Saracevic (1996), the concept of relevance is an intuitive one, i.e. everyone understands the meaning of relevance intuitively. Therefore, we assume that within our study, it will not be necessary to state, what kind of documents the relevant ones should be. Only the query and description will be displayed to the user. This approach is different from the one pursuing within the TREC[2] runs, as they provide a "narrative section" describing the desired information, i.e., "what makes a document relevant" (D. K. Harman, 2005), so that the assessors are able to get the idea of the underlying information need themselves; whereas we reversely provide the description of the information need, without providing explicit information, what a relevant document would look like, so that the assessors get the idea of the desired relevant documents intuitively.

2.2 Query types

As mentioned above, the evaluation methods in this study are partially based on a framework for evaluating search engines (Lewandowski, 2012). In the framework, it is recommended to consider different query types. We can relate query types to web search queries, as they have been analyzed and categorized by Broder (2002). These are:

- Informational (e.g. "Prinzipal-Agenten-Theorie", as the example above shows),
- Navigational (e.g. "econbiz" for finding the website of the portal) and
- Transactional query types (e.g. "download anti-virus software free" for intending to download a program).

In the context of library information systems, informational queries can be equated with *topic searches,* navigational queries with *known item searches* and transactional queries with searches for further sources, e.g., databases (Lewandowski, 2010). Regarding OPAC searches, the distinction between known item searches and transactional queries is not simple, because in order to formulate a transactional query, one would assume to know the actual source searched for.

[2] Text Retrieval Conference, http://trec.nist.gov/

According to the different kinds of information needs as mentioned in section 2.1, the underlying information need of topic searches would be problem-oriented, i.e. several documents usually need to be retrieved, whereas the intent of known item searches is to satisfy concrete information needs, i.e. there is one relevant document containing the desired information.

Although, there is a certain "complexity of the issues in defining a concept of 'known-item search'" (Lee, Renear, & Smith, 2007, p. 14), it can commonly described as a search for an item a user has in mind because he or she knows of it or believes that such an item exists.

In contrary to topic searches, the identification of known item searches in discovery systems is not without difficulty. In traditional OPACs the selection of a search field helps to understand which type of search a user wants to conduct, for example by using the author field (the user believes the item exists and is held by the library), call number (the user knows that the item exists) or search for the ISBN (the user knows that the book exists and/or believes it has been purchased by the library). (Rulik, 2014, p. 19,20) Since the search behavior in the library context nowadays is strongly influenced by using web search engines (Lewandowski, 2010), library users prefer the single search interface, leaving the known-item query determination to automatic methods, for instance, by detecting certain keywords indicating a journal or book or determiners, such as "the" and "a" or in the German language "der" or "eine", that mostly occur due to copying and pasting a title (Kan & Poo, 2005).

When searching for information on the web, users often formulate navigational queries. Although the "numbers on the ratio of navigational queries differ from one study to another […], the clear result of all studies is that navigational queries account for a noteworthy number of queries" (Lewandowski, 2011b, p. 357). Queries entered into OPACs are often determined as known item searches, as well. (Kan & Poo, 2005, p. 91)

To understand the frequency of actual query types EconBiz processes, we classified 2,000 queries typed into the single search interface from EconBiz log files. The log files were extracted in July 2014.

The queries were analyzed and inductively categorized into the following 10 categories of query terms:

0. a) Only last name of author
   b) First and last name of author
      i. Category "only author" (see Figure 4) = (categories 1 a + 1 b)- (category ii.)
1. Author and title(keyword)
2. Author and year
3. Author and title and year
4. Author and title and edition
      ii. Category "total author+" = ∑ categories 2..5
5. Title only
6. Title and year
7. Other known item searches
8. Journal title
      iii. Category "total title+" = ∑ categories 6..9
9. Topic search, i.e. topic related terms or keywords

The categories 1 to 9 preliminarily identify known-item searches, only category no. 10 is the query type for topic searches. The results showed a nearly equal distribution of known item searches (50.25%) and topic searches (47.85 %), while a minority of queries (1.75%) could not been identified as either topic or known item search (Figure 3). The overall distribution of search frequencies is shown in Figure 4.

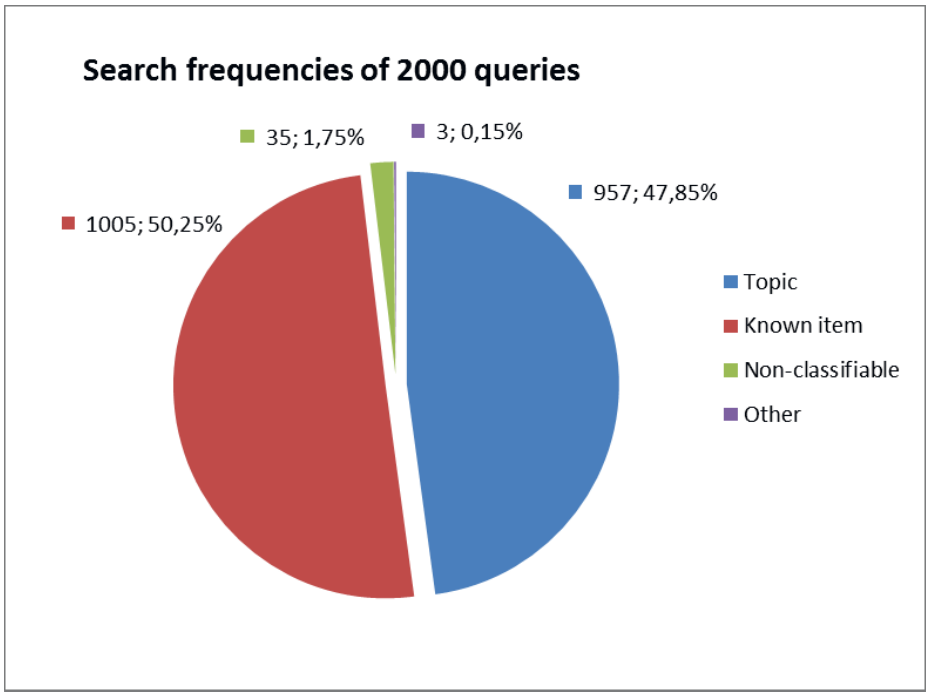

Figure 3: Proportion of topic and known-item searches

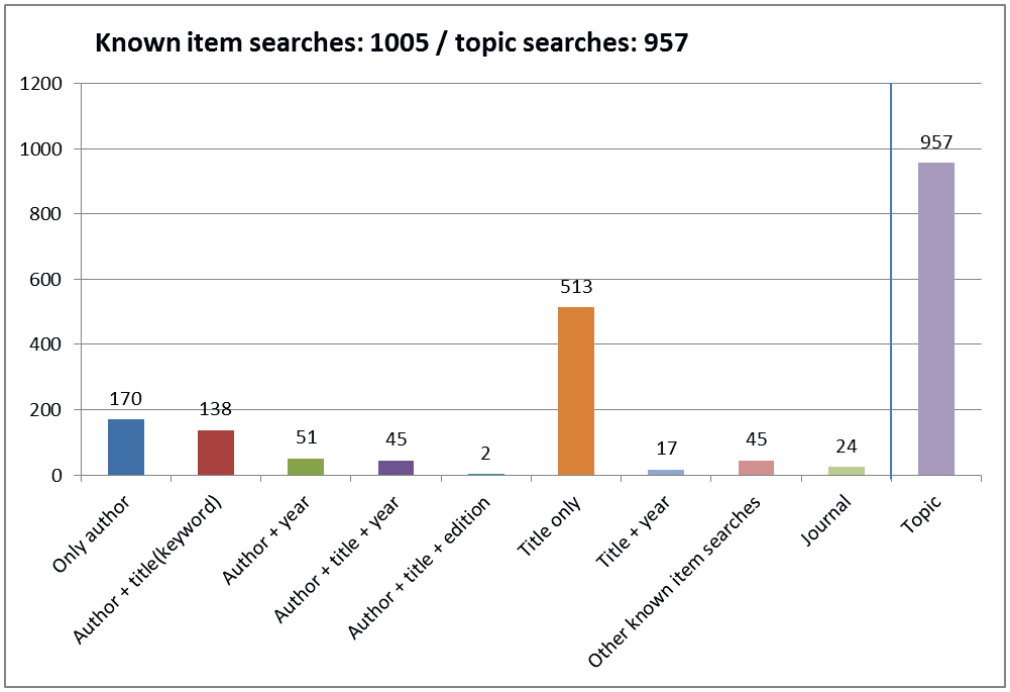

Figure 4: Distribution of search frequencies

About half of the entered queries are considered to be known-item searches. These findings are in overall consistency with the results of the analysis by Rulik (2014, p. 36). She analyzed queries from a Library Discovery system and found around 45% of the queries to be known items.

In this study we **focus on topic searches** because **known-item searches** only require relevance ranking to display the one correct search result on the top position of the search result list. Thus, we assess the ranking of known-item searches according to its success rate (see section 6) in a separate evaluation run, as the ranking performance for topic searches must not downgrade the performance for known-item searches. In contrast to topic search tasks, an explicit description of the concrete information need for relevance assessment is not necessary. Table 1 gives an overview on the query types and terms with regard to the runs within our project (see grey highlighted fields).

Query type in web search	**Query type in library context**	**Type of information need**	**Query terms in LibRank context**[3]
Informational	Topic search	Problem-oriented	Title keyword
			Topic keyword
			Full text
Navigational	Known-item search	Concrete	Title or part of title
			Author name
			Year of publication
Transactional	Search for further sources	Concrete	*Not to be considered*

Table 1: Overview of query types

2.3 Sources

Queries of the type **known-item search** can be obtained from the already classified 2,000 queries from the EconBiz log files. Figure 3 shows that 994 known-item searches (nearly 50%) could be identified, with a majority of title searches (503), as presented in Figure 4.

For **topic searches**, the first idea was to extract original user information needs by analyzing help requests to EconDesk[4] and derive search queries from them. These search queries would not be the original ones, but the information needs would be. We analyzed and categorized 918 EconDesk requests sent within 10 months (October 2013 until July 2014) into query intention types according to Broder (see overview in Table 1, column 1). In total, 390 informational requests could be identified, contrary to 74 navigational and 184 transactional requests.[5]

Thus, there are 390 described information needs for generating the query terms which represent topic searches. Unfortunately, a sample test showed that generating the appropriate search queries out of the described information needs does not produce an adequate (a minimum) number of results in EconBiz: Using keywords of the information need descriptions as search terms would often produce no result or too few results, for example the term “Dienstalterszulage” is neither found for

[3] Search in individual or combined fields.
[4] EconDesk is a research guide or helpdesk for EconBiz users, send via chat request or e-mail by users. http://www.econbiz.de/eb/hilfe/research-guide-econdesk/e-mail/
[5] The remaining 270 queries (ca. 30% of all queries) were not part of the query types according to Broder (2002), but could be identified as service oriented, e.g. requests for library processes like loaning a book. For details of the analysis see the internal report by Linhart (2014).

full text documents nor as a keyword in the STW (STW Thesaurus for economics[6]). However, some of the queries produce a very large result set, e.g., the query "Prinzipal-Agenten-Theorie" would lead to more than 7,000 results in EconBiz because it is also a standardized term in the STW.

One explanation for a result list being too small could be that users send their requests after they had entered search terms in EconBiz themselves without success, i.e. they got unwanted or no results. We came to the conclusion that the major part of requests sent to EconDesk cannot be answered using (only) EconBiz or without a certain amount of knowledge in economics which underlines the significance of EconDesk as an online help tool.

The **current approach** is to use the already classified topic searches of the 2,000 queries from the EconBiz log files. We intend presenting the queries to students of economics (e.g., at HAW) who will describe the possible associated information need. The idea is that three different students will work on each query to assure a good quality and, if needed, we can formulate one (combined) search query description based on the provided three afterwards. A similar approach to obtain information need statements based on the search query has been undertaken within a study by Huffman & Hochster (2007).

2.4 Number of queries

According to the project application and notice of granting, we have € 5,000 for assessors' remuneration, based on the assumption that we have € 5 for assessing one task and a total of 1,000 tasks being evaluated. These figures will have to be adjusted due to the fact that the number of results to one query exceeds the usual 10 documents (see section 4.1) and a remuneration of € 5 per task is not an adequate incentive. In addition, we will perform more than the originally intended two evaluation runs, starting with a low number of queries during one evaluation run (e.g., 10) and increasing the number of tasks per run after being able to make a valid recommendation based on the assessments of the prior runs.

3. Assessors

3.1 Selection of assessors (assessor groups)

Regarding user models, so far there has not been done enough research to identify user models in the context of all ZBW services. For the test design concept within LibRank, we created **user models** that refer exclusively to EconBiz users (see Table 2).

User model	Assumed document type preferences
Users are on-site	Printed and electronic materials
Users are not on-site	Electronic materials

Table 2: User models and their preferred document types

In the first instance, we divided users broadly into two groups of users that are "on-site", i.e., in the library building, and users that are "not on-site", i.e. at home, based on the assumption that the latter group prefers electronically available documents, whereas users on-site are interested in both

[6] http://zbw.eu/stw/version/latest/about.en.html

digital and non-digital documents, as they are within reach of printed library materials, as well.[7] We strongly assume that the pretest or first evaluation runs will show the validity of this distinction.

Since EconBiz is intended for end users seeking economic literature in an academic context, the tasks dealing with **topic searches** should be performed by users with an economic background or knowledge, e.g., researchers, professors or students. Due to the limited amount of money for the assessors' remuneration, we intend to collaborate with a couple of ZBW subject specialists for judging the *topical relevance*[8] of the results explicitly. As they are members of staff of one institution taking part in this research project, we are confident that contacting them will be more gainful than reaching out for less known experts at other research institutions or universities.

Situational relevance of documents could also be assessed by different status groups; in particular, students, e.g. students at the HAW, Department of Economics[9] or students at the Faculty of Business Administration[10] or Department of Economics[11] at the University of Hamburg. Details of contacting still need to be clarified.

In order to assess **known-item searches**, it is not necessary to make allowances for user models or assessors with subject knowledge, because those results only need to be judged to verify if the one correct answer is on top position of the results list, which can be accomplished by our student assistants.

3.2 Number of assessors

The total number of assessors to be involved in the evaluation runs depends on the number of search tasks. In general, one assessor would be able to work on more than one task, but the total number of assessors depends on the number of results per task (see section 4.1) or the amount of time needed. We assume that at some point in time during tasks performances, the ability to concentrate begins to decrease. To estimate the duration for assessing one task and the number of tasks one assessor can perform in a row, we will measure start and end time within the pretest (see section 5.1).

A solution to this problem would be to perform the assessment per document instead of performing it on the task level. However, currently, the evaluation with the Relevance Assessment Tool (RAT) is based on task level, not on document level, i.e. the results per task cannot be split and assessed by different participants nor is it possible to save assessments of one particular task and finish the remaining results at a later time. The required software alterations can be implemented after the description of the Relevance model is completed as far as possible, i.e., there may be alterations necessary depending on the relevance assessments after every evaluation run.

---

[7] These assumptions behind the user models are also mirrored by the ranking factors *physical location* and *availability.*

[8] *Topical relevance* as the subject matter or aboutness can be assessed due to intellectual interpretations by humans (Borlund, 2003, p. 916), i.e. in this research context by subject experts. In terms of *situational relevance,* other factors besides topical ones are taken into consideration, e.g. availability, length or language of the document, which are dynamic factors that depend on the concrete work task or information need (Borlund, 2003, pp. 921–922).

[9] http://www.haw-hamburg.de/ws-w.html

[10] http://www.bwl.uni-hamburg.de/en.html

[11] http://www.wiso.uni-hamburg.de/en/fachbereiche/vwl/home/

4. Search Results

In order to assess search results with the Relevance Assessment Tool, the specific elements of a result need to be considered, because not every detail displayed is necessary or useful to present them to the assessors. The following two sections provide information on what elements will be assessed and how they are to be displayed by the RAT.

4.1 What will be assessed?

In Web search we distinguish between the list of search results consisting, for example, of result descriptions (snippets) and the linked web page as the document to be assessed. This is a major difference to library information systems, because library materials are not only digital documents. Thus, due to the nature of (meta)data in EconBiz, judging **result descriptions** would not contribute to the data analysis, because the **results** being assessed technically are the descriptions (surrogates) of the actual documents, e.g. full text articles, book chapters, books. This means that further links to full texts will be made non-clickable, so that every assessor has the same prerequisites in terms of topical relevance assessment.

Comparability problem with EconBiz data

With regard to the evaluation framework for search engines (Lewandowski, 2012), in the EconBiz context the question arose: Are result descriptions on search engine result pages equal to metadata of surrogates in the EconBiz results list (short title record - 'Kurztitelanzeige') and the actual result would be the complete description / surrogate (full bibliographic record - 'Volltitelanzeige')? The metadata of the surrogates in the result lists differ, e.g. publisher details (depending on document type) or abstract (not every article is provided with an abstract, even if full text is accessible).

For comparability purposes, the first idea was to narrow the results for evaluation into different document types; e.g., only articles. However, this would not be expedient, because a mixed search result list is required. Another idea was to split the results into groups of different accessibility; e.g., with full text access or external link to table of content, and let these groups be assessed by different assessor groups. Since such approaches neglect the concept of reality, they have been rejected, because the problem of comparability could not be solved within this research project.

The search results in EconBiz contain several **elements**, but not all of them will be displayed by the RAT as the basis for assessment. With regard to the document types in EconBiz, we focus on books, articles and journals (see Table 3). Since Working Papers are treated like books during the indexing process, they are labeled as the same document type. In some cases, they are additionally provided with the metadata publication type (subcategory) - "Publikationsform (Subkategorien): Working Paper" or are identifiable due to the series title ("Schriftenreihe").

The elements in EconBiz to be displayed are shown in detail in Figure 7, Figure 8, Figure 9 and Figure 10 that can be found in the Appendix.

Document types in EconBiz	Assumed to be relevant for results assessment?
Book / Working Paper	yes
Article	yes
Journal	yes
Institution	no (only small proportion of records, thus negligible)
Internet source	no (only small proportion of records, thus negligible)
Portal	no (only small proportion of records, thus negligible)
Other	no (only small proportion of records, thus negligible)

Table 3: Document types in EconBiz

As mentioned above, we present one **document without a separate result description** as the search result to be assessed. Thus, we will not consider the data of the result list (see the green boxes in Figure 7) except for the bold marked terms that are the highlighted search terms occurring in the metadata; these terms will be transferred to the particular metadata of the result. An example for an **article** in EconBiz is shown by Figure 8: The red boxes are metadata elements to be neglected, the yellow boxes indicate elements to be altered, as the name of the holding library (e.g. ZBW) will be changed into the anonymous name of 'your library' ("Ihre Bibliothek"). The same alterations will be made to **books** (Figure 9); additionally, the text of the metadata field "Beschreibung" (see yellow box at the bottom) will be displayed as a complete description below the title field (see yellow box at the top, below the title and author names). It is also necessary to have the full description at the top of the result, because the field "Beschreibung" will be substituted with the more detailed information on available copies within the tab "Exemplare" (see Figure 10), as the number of copies might influence the relevance assessments. Again, the name of the holding library will be replaced by 'your library' ("Ihre Bibliothek").

The number of results per query (task) depends on the number of duplicate results and the number of test rankings. With a cut-off value of 20 we will have the top 20 results per test ranking, leaving out any duplicates. If we test, for example, 3 rankings with 60 results and 10 duplicates, there would be 50 documents left for assessment (see Figure 6 in section 5.2).

The **number of results per task** during one evaluation run can also differ from one run to another. For example, in one of the first evaluation runs ZBW subject specialists could judge 10 search queries with a maximum of 200 results each, and in the following run students would be assessing fewer documents per task because of the altered ranking weightings.[12]

4.2 Assessing with the Relevance Assessment Tool

The Relevance Assessment Tool is a web-based software application that was designed to assist researchers in conducting search engine retrieval effectiveness studies, with regard to reducing time and effort. It follows a modular approach and contains the following components: 1. Test design and project administration, 2. Search engine result scraping, 3. Collecting relevance judgements and 4.

[12] The benefit of having a pool of (a maximum of) 200 assessments per task is that the data can be used for future tests and "learning to rank" after project completion, as well as being part of the test data set for other interested institutions/researchers.

Results download.[13] Due to its flexibility and modular structure, it can be altered to meet study specific requirements. (Lewandowski & Sünkler, 2013) Thus, we are able to make use of this tool within the LibRank project to design and conduct the retrieval tests with the EconBiz test environment and instead of testing different search engines we evaluate different rankings.

Before using this tool for assessing results, some alterations must be implemented, according to the project specific methods and test design. These implementations are described below.

A binary assessment allows only two conditions: a document is either relevant or irrelevant. With user-based models for retrieval tests, this is not adequate, because "documents of different relevance grades are treated as equally important with relevance conflated into two categories" (Carterette, Kanoulas, & Yilmaz, 2012, p. 116). To allow a differentiated assessment, graded relevance can be observed by using a scale assessment with, for example, a 5-point scale. It is recommended that both **binary and scale assessment** is used. (Lewandowski, 2013, p. 346)

For scale assessments, the RAT has an implemented **slider**, which consists of a 0 – 100-point scale, whereas the particular relevance score will not be visible to the assessor. The start position of the slider button is invisible so that a possible influence of the assessors due to the pre-set position will be avoided.

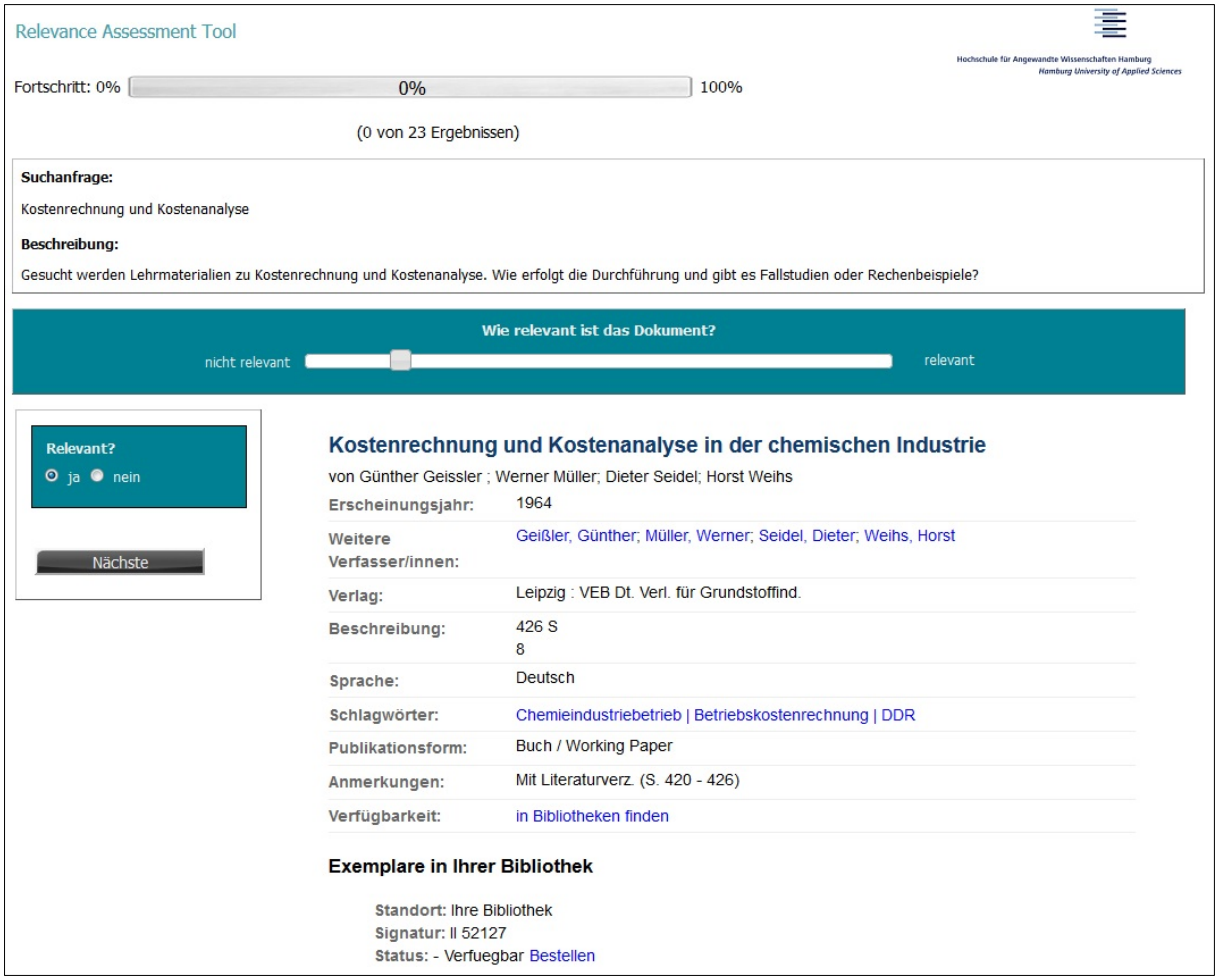

Figure 5: Display of a document with binary assessment and slider by the Relevance Assessment Tool

[13] A free trial demo is available at http://www.searchstudies.org/rat/

The original **order of the results** will not be visible to the assessors, as the results per task will be mixed randomly by the tool to avoid order effects. **Duplicates,** i.e. duplicate results that are produced by more than one ranking, will only be judged once per task per evaluation run. The assessment data of every document will be integrated in the analyzing process for every test ranking per task (for example, see Figure 6 in section 5.2). The **source of results** will be made anonymous, e.g. the EconBiz logo will not be displayed. Regarding our two user models, the **description of the usage situation** can be presented to the assessor on the RAT homepage, for example: „*Sie befinden sich zuhause und brauchen ganz dringend XY, aus diesem Grund sind Sie vorrangig/ausschließlich an elektronisch verfügbaren Dokumenten interessiert*.“) It can be assumed that the location of the user will not have a huge effect, which may transpire after a couple of evaluation runs or as a result of the pretest.

RAT criterion	Description
Scale	Binary (0; 1) and slider (0-100 points, displayed as non-relevant to relevant)
Order of results	Mixed order, i.e. original order will not be visible
Source of results	Anonymous
Duplicates	Duplicate results will be removed
Usage situation	Displayed description prior to assessment

Table 4: Overview on assessment details with RAT

5. Testing

5.1 Pretest

Besides testing the project specific implementations in the RAT, the pretest aims to answer two major questions: 1) How many **duplicate results** are produced with a cut-off value of 20 and 2) what is the **amount of time needed** for assessing one task? Based on this information we can determine the number of tasks per evaluation run and the number of assessors needed.

Search queries used in the pretest will consist of terms understandable to assessors without special economics or business knowledge, but for example terms of general academic or interdisciplinary context or simple business terms(e.g., statistical methods, academic writing, communication and presentation), so that the tasks can be **assessed by our student assistants**. Therefore, contacting the ZBW subject specialists is not an inevitable prerequisite for conducting the pretest.

5.2 Evaluation runs

In one evaluation run, the same search queries will be entered to every test ranking, as illustrated in Figure 6. Every ranking contains selected factors including their particular weightings. All search results produced by every test ranking (e.g., 60) will be displayed in a random order, after identifying and cleansing out the duplicate results (e.g., 10). Thus, we have one pool of search results (e.g., 50) for assessment. The number of rankings tested during one evaluation run depends on the results of the pretest, concerning the number of results and amount of time needed for assessing their relevance.

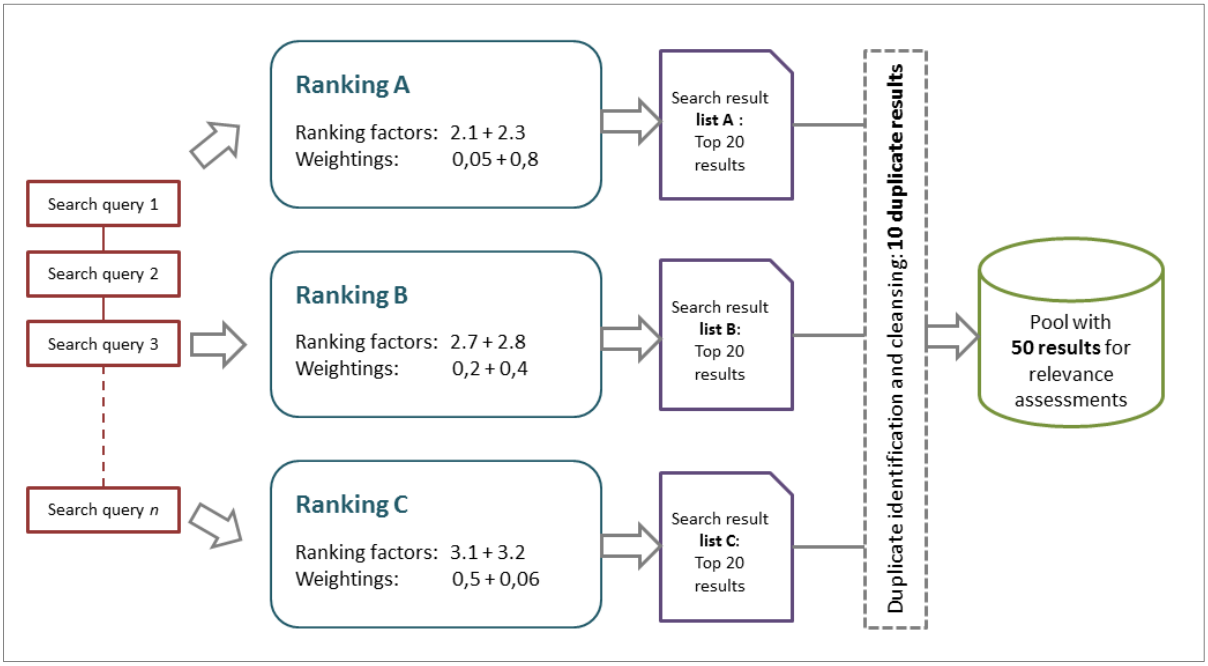

Figure 6: Example evaluation run (own illustration)

The aim of the evaluation runs is to observe individual ranking factors, as well as the combination of factors of the same group and in combination with factors of other groups (see research questions in section 1). The possible ranking factors identified are listed below in Table 5.

Ranking factor	Description
1. Text statistics	
Term frequency (TF)	Relative frequency of search term in a document
Inverse document frequency (IDF)	Relative frequency of term in all documents (rarely occurring terms are preferred)
Search term order	If a query consists of more than one search term, documents with the term at the beginning of the query will be ranked higher
Search term distance	If a query consists of more than one search term, documents with the terms closest to each other will be ranked higher
Position of search term	If term appears in beginning of document, it is ranked higher
Document length	Documents within a sudden length range are preferred (content must not be neither too short nor too long)
Emphasis on terms within a document	Terms that are emphasized are weighted higher
Anchor text	Terms appearing in anchor text are ranked higher
2. Popularity	
Click popularity	Documents that have been visited by many users are preferred (in conjunction with dwell time)
Dwell time	The amount of time the document has been viewed (amount of time that implies reading)
Usage frequency	Items that have been downloaded or loaned more often than others are ranked higher
Purchasing behavior	Works that have a large number of purchased copies (locally and globally) are weighted higher
Publisher authority	Works by a highly reputed publisher are weighted higher
Ratings	Documents rated or recommended by others are ranked higher

Reference counting	Document is ranked by the number of incoming links or citations
Reference popularity	Document is ranked by the number of links or citations in relation to other documents or entities
3. Freshness	
Publication date	Current documents are preferred
Accession date	Documents that have recently been accessioned are preferred
4. Locality & availability	
Physical location	Physical location of the item and the user
Availability of the webpage / item	Current availability of the item
5. Content properties	
Additional information	Documents with other information than basic descriptions available are weighted higher
File format	Documents written in a particular file format are preferred to other formats
Language	Documents in the preferred language(s) are ranked higher
6. User background	
User group	Ranking based on preferences of a particular user group
Usage data	Personalized ranking based on personal (social) profile of the individual user

Table 5: Overview on possible ranking factors; for green marked factors data for evaluation are available

The decision, **which ranking factors should be tested**, is based on one major aspect: The ranking factors identified and described within the first working package (see report) cannot entirely be tested, because we do not have the particular data for every individual factor (e.g. *ratings*). Thus, we should consider any ranking factors with the underlying data being (at least partially) available (see green marked factors in Table 5), according to the Data Availability Report as a result of WP1.[14] Due to this decision, we could reduce the number of factors to a total of 7 and 4 different groups, as listed in Table 6. As another result, the factors *click popularity* and *usage* could be merged together, because they both are based on the same underlying data. Additionally, some factors, e.g. physical location, have been integrated in another factor as a required part, so that the three groups *freshness*, *locality & availability* and *content properties,* eventually, consist of one factor, each.

The **7 factors will be systematically tested** during the evaluation runs considering three possible combinations (Table 6) and depending on the data availability per factor (see the project's internal data availability report, "Requirements specification for the test system"):

1. Every factor of every factor group will be tested individually (marked as 'i').
2. Every factor will be tested in combination with all other factors of the same group (marked as 'c').
3. Every factor will be tested in combination with all factors of other groups (marked as 'cg').

Since the **current state of the running version of EconBiz** cannot be evaluated, the currently applied ranking algorithms will be used as a test ranking in the first evaluation run, to provide a baseline for comparison with subsequent test rankings.

[14] Since *text statistics* are the prerequisites for producing any results in a text based retrieval system at all, there are no other data needed in this first group of ranking factors, contrary to the factors of the other groups.

During another evaluation run, one test ranking only consisting of *text statistics* factors should be applied to compare those **documents, which lack certain data**, e.g. citation data or circulation data for the group *popularity*. The second test ranking within the same run would include combined factors of text statistics and popularity, to be able to observe possibly existing correlations between *text statistics* and *popularity*.

Group	Ranking Factor	Subfactor	Abbreviation	CP-views	CP-contents	CP-clicks	UF-records	UF-clicks	UF-loans	PB-editions	PB-libraries	PA-publisher	PA-peer-rev.	RCP-item	RCP-journal	RCP-author	PD	AV	AI
2. Popularity	Click popularity = Usage	Number of clicks on bibl. record (views)	**CP-views**	*i*															
		Number of clicks on further content	**CP-contents**	c	*i*														
		Number of citation (record) downloads	**UF-records**	c	c	c	*i*												
		Number of full text downloads (clicks on availability button)	**UF-clicks**	c	c	c	c	*i*											
		Number of loans at the library	**UF-loans**	c	c	c	c	c	*i*										
	Purchasing behavior	Number of editions	**PB-editions**	c	c	c	c	c	c	*i*									
		Number of item owning libraries	**PB-libraries**	c	c	c	c	c	c	c	*i*								
	Publisher authority	Number of items by a certain publisher	**PA-publisher**	c	c	c	c	c	c	c	c	*i*							
		Peer-reviewed vs. non peer-reviewed journals	**PA-peer-rev.**	c	c	c	c	c	c	c	c	c	*i*						
	Reference counting & popularity	Number of citation counts for item	**RCP-item**	c	c	c	c	c	c	c	c	c	c	*i*					
		Citation impact for journal	**RCP-journal**	c	c	c	c	c	c	c	c	c	c	c	*i*				
		Citation impact for author	**RCP-author**	c	c	c	c	c	c	c	c	c	c	c	c	*i*			
3. Freshness	Publication date	Publication date of reviews, books, articles (print + electronic)	**PD**	cg	cg	cg	cg	cg	cg	cg	cg	cg	cg	cg	cg	cg	*i*		
4. Locality & availability	Availability	n.a.	**AV**	cg	cg	cg	cg	cg	cg	cg	cg	cg	cg	cg	cg	cg	cg	*i*	
5. Content properties	Additional information	Availability of abstracts, tables of contents, reviews	**AI**	cg	cg	cg	cg	cg	cg	cg	cg	cg	cg	cg	cg	cg	cg	cg	*i*

Table 6: Possible combinations of ranking factors to be tested in the evaluation runs

i..individual
c..in combination with another factor of the same group
cg..in combination with a factor of another group

6. Data analysis

The main measures in Information Retrieval evaluation are precision and recall, or precision and recall based measures. Precision measures the ability of an IR system to retrieve *only* relevant results, whereas recall measures the ability to retrieve *all* relevant results to a search query. The existence of an inverse relationship between recall and precision had been one result of the Cranfield experiments. In this research project, Cleverdon & Keen (1966) studied "factors determining the performance of indexing systems" in an experimental environment for the first time.[15] They laid the foundation for systematic IR evaluation based on a formal framework - the Text Retrieval Conference (TREC) that started 1992. TREC[16] aims at evaluating the quality of retrieval systems using specific test collections (including queries, information need descriptions, relevance judgements and documents). An overview on TREC measures[17] is provided by Buckley & Voorhees (2005). The major TREC-1 measures are applied to binary assessments:

- *Precision @ cut-off*
- *Recall @ cut-off* (which had been relinquished in TREC-2)
- *Interpolated precision at recall point x*
- *Eleven-Point Average*
- *Three-Point Average*

Two new measures were added for TREC-2:

- *R-Precision*, which was more applicable than Precision @ cut-off;
- *Non-interpolated Average Precision*, which replaced the eleven- and three-point average, and *Mean Average Precision (MAP)*.

These measures are based on the Cranfield paradigm. However, the Cranfield paradigm uses the following assumptions (Buckley & Voorhees, 2005, p. 68):

- *Judges can assess the relevance of a document from the document's content.*
- *All relevant documents are equally desirable.*
- *The relevance of one document is independent of the relevance of any other document.*
- *The user information need is static.*
- *A set of topics with corresponding judgment sets is representative of the user population.*
- *The list of relevant documents for each topic is complete (that is, all relevant documents are known).*

Based on these rather simplifying assumptions, there are two main issues with these measures: *inconsistency* and *incompleteness*. As relevance is subjective, dynamic, and full recall in web search is not feasible (Buckley & Voorhees, 2004), these assumptions cannot be seen as realistic, and they do not adequately take user behavior into account.

---

[15] A brief history on the Cranfield experiments is provided by Baeza-Yates & Ribeiro-Neto (2011, p. 132f.).
[16] http://trec.nist.gov/overview.html
[17] A TREC software tool can be used for analyzing data applying different measures: http://trec.nist.gov/trec_eval/

Within LibRank we do consider user behavior in searching and accessing search results, as presented in the State of the Art Report as a result of WP1. That is the reason for the use of a binary scale for assessment, i.e., non-relevant or relevant, and additionally graded relevance assessments, i.e., from non-relevant to relevant on a slider (see section 4.2). Therefore, we need to apply metrics that are suitable for graded relevance assessments, as well. Whereas binary based metrics had been used within TREC for many years, there have been several attempts to introduce graded relevance metrics (Kekäläinen, 2005), which are, for example:

- *sliding ratio*
- *relative relevance*
- *cumulated gain-based measures*

Cumulated gain-based measures are also integrated in the TREC tool *trec_eval*. We will describe Cumulated gain-based measures in more detail in section 6.2. In statistics, cumulated frequencies are the sum of the frequencies until a certain boundary (e.g., 1.0 or 100%). They consider human behavior when scanning search results top down, i.e., usually the first 10 results in a ranked list. Thus, it is reasonable to calculate cumulated precision scores instead of solely absolute precision scores (see, for example, Lewandowski, 2015; Lewandowski, 2008).

Metrics for IR evaluation are quite diverse. For instance, they can be categorized into system-oriented or user-oriented measures, binary or graded relevance measures that all have assets and drawbacks.[18] The choice of metrics does not only depend on the scale assessment, but the query type needs to be considered, too. As discussed in section 2.2, we evaluate informational and navigational queries. We can differentiate between relevance values and success rates (Lewandowski, 2011a). Therefore, with regard to Table 1, for topic search results we can evaluate relevance values and for known-item search results we measure success rates.[19]

6.1 Analysis of known-item search results

The evaluation of known-item searches answers the research question: Is the IR system or the ranking able to produce the one relevant (correct) result and put it on top position of the result list?

Mean Reciprocal Rank (MRR) is a standard measure used in TREC. It was introduced by Kantor & Voorhees (2000) to assess the performance of an IR system using OCR corrupted text versions. For instance, Craswell & Hawking (2005) applied MRR to navigational queries, as well as they did in TREC the year before (Craswell, Hawking, Wilkinson, & Wu, 2004).

Although MRR is applicable to measure retrieval effectiveness on navigational queries, one disadvantage can be recognized because it only “the first correct result is considered and it takes only a few discrete values, for instance, 1, 1/2, and 1/3, for rank positions 1, 2, and 3, respectively” (Baeza-Yates & Ribeiro-Neto, 2011, p. 143). Therefore, “the measure is insensitive to large difference in low rank” (Sanderson, 2010, p. 284).

With regard to user-orientation, a measure to examine the cost of the user on how many nonrelevant documents he or she has to look at in a ranked result list is the **Expected Search Length**

[18] An overview of IR metrics with calculation examples is provided by Baeza-Yates & Ribeiro-Neto (2011).
[19] For results of the evaluation runs see the Working Paper “Results of Evaluation Runs and Data Analysis” as a result of Working Package 4.

proposed by Cooper (1968). This measure was then used by the research community in order to enhance the Cranfield metrics (D. Harman, 2011, p. 25).

Another measure applicable to evaluation of known-item searches is the success rate or **Success @n**, i.e., to evaluate what proportion of queries produce the one relevant document until position *n* in the Ranking (Craswell & Hawking, 2005; Lewandowski, 2011b). For instance, the ranking position 1, 5 or 10 is considered.

A study on the ability of a library information system to deal with known-item searches has been done by Rulik (2014). She studied the retrieval effectiveness of the discovery system *beluga* on known-items and applied **success @n** and **MRR.**

In the evaluation run using known-item searches we focus on the **success rate** and **MRR** to answer the question, if the one correct result is one of the top ten or even on top position of the result list.

6.2 Analysis of topic search results

As mentioned above, we will assess the binary and graded relevance of a document. For the **binary** relevance scores, a **Precision Graph** of the top 20 results of all tasks for each of the test ranking, both **cumulated and non-cumulated**, will be created. We refrain from using MAP, because one disadvantage of MAP is that it makes "no distinction in pooled collections between documents that are explicitly judged as nonrelevant and documents that are assumed to be nonrelevant because they are unjudged." (Buckley & Voorhees, 2004, p. 26)

Precision graphs can also be created for the **graded relevance** assessments, showing the **Graded Average Precision (GAP)**, proposed by Robertson, Kanoulas, & Yilmaz (2010).This measure is based on Average Precision (AP), whereas the "AP of a ranked Iist is the average of the precisions at each relevant document in that Iist" (Carterette et al., 2012, p. 113). GAP is described as "the **cumulated product of graded precision** values and graded recall step values at documents of positive relevance grade, as average precision can be defined as the cumulated product of precision values and recall step values at relevant documents"(Robertson et al., 2010, p. 606). In our evaluation runs the *documents of positive relevance* are given by the binary assessments because we do not provide a scale of graded relevance levels. Instead, we analyze the **GAP based on all of the documents' relevance scores** (0 - 100), regardless of their binary assessments.

In contrast to precision, cumulated gain based measures "allow researchers to test different weighting schemes for relevant documents, which reflect different user scenarios" (Kekäläinen, 2005, p. 1022). To analyze the systems' ability to rank search results by relevance in descending order, the **Discounted Cumulative Gain (DCG)**, which was introduced as a novel measurement by Järvelin & Kekäläinen (2000), can be used. DCG is based on two observations:

1. highly relevant documents are preferable at the top of the ranking than mildly relevant ones;

2. relevant documents that appear at the end of the ranking are less valuable. (Baeza-Yates & Ribeiro-Neto, 2011, p. 146)

For direct comparison of all our different test rankings, the **Normalized Discounted Cumulative Gain (NDCG)** is needed. NDCG is a corrected version of DCG using normalized figures.[20] NDCG or its

[20] A step-by-step calculation example is given by Baeza-Yates & Ribeiro-Neto (2011, pp. 145–150).

variants are regularly used in IR evaluation, e.g., in TREC (Clarke, Craswell, & Soboroff, 2009; Collins-Thompson, Maconald, Bennett, Diaz, & Voorhees, 2015). It has also been analyzed in several studies that compare different IR evaluation measures (Sakai & Song, 2011). For example, Sakai (2007) argues that “AnDCG [Average Normalized Discounted Cumulative Gain] and nDCG are the best among the rank-based graded-relevance metrics (Sakai, 2007, p. 547)".

7. Further research

The evaluation of different rankings with human relevance assessments is based on the documents’ surrogates, i.e. metadata. The representation of a surrogate influences the relevance judgement, e.g., one record can lack certain metadata that another record lacks not or another result would be judged irrelevant “because of its misleading description” (Lewandowski, 2008, p. 931). The question remains, on what criteria the assessors judge whether the document or surrogate is relevant or to what degree of relevance it is to them, i.e., why do they think a document may be relevant? Such relevance clues can be categorized into topical and situational relevance clues (Saracevic, 2007, p. 2127f.). Within LibRank, for example, information on author or journal impact, or other popularity information could be an important criterion. Further, formal aspects, e.g., the presence of abstract or full text availability, could be decisive. Although, some of these criteria are, of course, individual ranking factors as well, an experiment may provide a deeper understanding on the particular relevance cues. The idea would be not to question the human assessors directly but, for example, to provide different surrogates containing different metadata to a document.

8. Acknowledgements

The research project LibRank is funded by the German Research Foundation (DFG – Deutsche Forschungsgemeinschaft) from 3/2014 until 2/2016. We thank Alexandra Linhart for her contribution within her “Research & Venture” project as part of her master studies at the Department of Information, University of Applied Sciences Hamburg.

9. References

Baeza-Yates, R., & Ribeiro-Neto, B. (2011). Retrieval evaluation. In *Modern Information Retrieval: the concepts and technology behind search* (2nd ed., pp. 131–176). Addison-Wesley.

Borlund, P. (2003). The concept of relevance in IR. *Journal of the American Society for Information Science and Technology*, *54*, 913–925. doi:10.1002/asi.10286

Broder, A. (2002). A taxonomy of web search. *ACM SIGIR Forum*, *36*(2), 3. doi:10.1145/792550.792552

Buckley, C., & Voorhees, E. M. (2004). Retrieval evaluation with incomplete information. In *Proceedings of the 27th annual international conference on Research and development in information retrieval - SIGIR ’04* (pp. 25–32). New York, New York, USA: ACM Press. doi:10.1145/1008992.1009000

Buckley, C., & Voorhees, E. M. (2005). Retrieval System Evaluation. In E. M. Voorhees & D. K. Harman (Eds.), *TREC: Experiment and Evaluation in Information Retrieval* (pp. 53–75). Cambridge, MA ;

London, UK: MIT Press.

Carterette, B., Kanoulas, E., & Yilmaz, E. (2012). Evaluating web retrieval effectiveness. In D. Lewandowski (Ed.), *Web search engine research* (pp. 105–137). Emerald Group Publishing.

Clarke, C. L., Craswell, N., & Soboroff, I. (2009). Overview of the TREC 2009 Web Track. In *The Eighteenth Text REtrieval Conference (TREC 2009) Proceedings*. National Institute of Standards and Technology. Retrieved from http://trec.nist.gov/pubs/trec18/papers/WEB09.OVERVIEW.pdf

Cleverdon, C. W., & Keen, M. (1966). Aslib Cranfield research project - Factors determining the performance of indexing systems; Volume 2, Test results. Retrieved from https://dspace.lib.cranfield.ac.uk/handle/1826/863

Collins-Thompson, K., Maconald, G., Bennett, P., Diaz, F., & Voorhees, E. M. (2015). TREC 2014 Web Track Overview. In *The Twenty-Third Text REtrieval Conference (TREC 2014) Proceedings*. National Institute of Standards and Technology. Retrieved from http://trec.nist.gov/pubs/trec23/papers/overview-web.pdf

Cooper, W. S. (1968). Expected search length: A single measure of retrieval effectiveness based on the weak ordering action of retrieval systems. *American Documentation*, *19*(1), 30–41. doi:10.1002/asi.5090190108

Craswell, N., & Hawking, D. (2005). Overview of the TREC-2004 Web Track. In *The Thirteenth Text Retrieval Conference (TREC 2004)*. National Institute of Standards and Technology. Retrieved from http://trec.nist.gov/pubs/trec13/papers/WEB.OVERVIEW.pdf

Craswell, N., Hawking, D., Wilkinson, R., & Wu, M. (2004). Overview of the TREC 2003 Web Track Non-interactive Experiments. In *The Twelfth Text Retrieval Conference (TREC 2003)*. National Institute of Standards and Technology. Retrieved from http://trec.nist.gov/pubs/trec12/papers/WEB.OVERVIEW.pdf

Frants, V., Shapiro, J., & Voiskunskii, V. (1997). *Automated information retrieval: theory and methods*. San Diego: Academic Press.

Harman, D. (2011). Information retrieval evaluation. *Synthesis Lectures on Information Concepts, Retrieval, and Services*. Morgan & Claypool Publishers.

Harman, D. K. (2005). The TREC Test Collections. In E. M. Voorhees & D. K. Harman (Eds.), *TREC: Experiment and Evaluation in Information Retrieval* (pp. 21–52). Cambridge, MA ; London, UK: MIT Press.

Huffman, S. B., & Hochster, M. (2007). How well does result relevance predict session satisfaction? *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '07*, 567. doi:10.1145/1277741.1277839

Järvelin, K., & Kekäläinen, J. (2000). IR evaluation methods for retrieving highly relevant documents. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '00* (pp. 41–48). New York, New York, USA: ACM Press. doi:10.1145/345508.345545

Kan, M.-Y., & Poo, D. C. C. (2005). Detecting and supporting known item queries in online public access catalogs. In *Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries - JCDL '05* (pp. 91–99). New York: ACM Press. doi:10.1145/1065385.1065406

Kantor, P. B., & Voorhees, E. M. (2000). The TREC-5 Confusion Track: Comparing Retrieval Methods for Scanned Text. *Information Retrieval*, *2*(2-3), 165–176. doi:10.1023/A:1009902609570

Kekäläinen, J. (2005). Binary and graded relevance in IR evaluations—Comparison of the effects on ranking of IR systems. *Information Processing & Management*, *41*(5), 1019–1033. doi:10.1016/j.ipm.2005.01.004

Kelly, D., & Fu, X. (2007). Eliciting better information need descriptions from users of information search systems. *Information Processing & Management*, *43*(1), 30–46. doi:10.1016/j.ipm.2006.03.006

Lee, J. H., Renear, A., & Smith, L. C. (2007). Known-Item Search: Variations on a Concept. *Proceedings of the American Society for Information Science and Technology*, *43*(1), 1–17. doi:10.1002/meet.14504301126

Lewandowski, D. (2008). The retrieval effectiveness of web search engines: considering results descriptions. *Journal of Documentation*, *64*(6), 915–937. doi:10.1108/00220410810912451

Lewandowski, D. (2010). Using search engine technology to improve library catalogs. *Advances in Librarianship*, *32*, 35–54. doi:10.1108/S0065-2830(2010)0000032005

Lewandowski, D. (2011a). Evaluierung von Suchmaschinen. In D. Lewandowski (Ed.), *Handbuch Internet-Suchmaschinen 2* (pp. 203–228). Heidelberg: AKA Verlag.

Lewandowski, D. (2011b). The retrieval effectiveness of search engines on navigational queries. *Aslib Proceedings*, *63*(4), 354–363. doi:10.1108/00012531111148949

Lewandowski, D. (2012). A Framework for Evaluating the Retrieval Effectiveness of Search Engines. In C. Jouis, I. Biskri, J.-G. Ganascia, & M. Roux (Eds.), *Next Generation Search Engines: Advanced Models for Information Retrieval* (pp. 456–479). Hershey, PA: IGI Global. doi:10.4018/978-1-4666-0330-1

Lewandowski, D. (2013). Verwendung von Skalenbewertungen in der Evaluierung von Web-Suchmaschinen. In H.-C. Hobohm (Ed.), *Informationswissenschaft zwischen virtueller Infrastruktur und materiellen Lebenswelten : Tagungsband des 13. Internationalen Symposiums für Informationswissenschaft (ISI 2013), Potsdam, 19. bis 22. März 2013* (pp. 339–348). Glückstadt: Hülsbusch.

Lewandowski, D. (2015). Evaluating the retrieval effectiveness of web search engines using a representative query sample. *Journal of the Association for Information Science and Technology*, *66*(9), 1763–1775. doi:10.1002/asi.23304

Lewandowski, D., & Sünkler, S. (2013). Designing search engine retrieval effectiveness tests with RAT. *Information Services and Use*, *33*(1), 53–59. doi:10.3233/ISU-130691

Linhart, A. (2014). EconDesk – Auswertung der Logs: Informations- und Servicebedürfnisse. ZBW.

Robertson, S. E., Kanoulas, E., & Yilmaz, E. (2010). Extending average precision to graded relevance

judgments. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval - SIGIR '10* (pp. 603–610). New York, New York, USA: ACM Press. doi:10.1145/1835449.1835550

Rulik, I. (2014). *Known-Item-Suchanfragen im Discoverysystem beluga: Retrievaleffektivität und Empfehlungen*. HAW Hamburg.

Sakai, T. (2007). On the reliability of information retrieval metrics based on graded relevance. *Information Processing & Management*, *43*(2), 531–548. doi:10.1016/j.ipm.2006.07.020

Sakai, T., & Song, R. (2011). Evaluating diversified search results using per-intent graded relevance. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information - SIGIR '11* (p. 1043). New York, New York, USA: ACM Press. doi:10.1145/2009916.2010055

Sanderson, M. (2010). Test Collection Based Evaluation of Information Retrieval Systems. *Foundations and Trends® in Information Retrieval*, *4*(4), 247–375. doi:10.1561/1500000009

Saracevic, T. (1996). Relevance reconsidered. In P. Ingwersen & N. O. Pors (Eds.), *Information science: Integration in perspectives. Proceedings of the Second Conference on Conceptions of Library and Information Science (CoLIS 2)* (pp. 201–218). Copenhagen: Royal School of Librarianship.

Saracevic, T. (2007). Relevance: A review of the literature and a framework for thinking on the notion in information science. Part III: Behavior and effects of relevance. *Journal of the American Society for Information Science and Technology*, *58*(13), 2126–2144. doi:10.1002/asi.20681

Stock, W. G., & Stock, M. (2013). Basic Ideas of Information Retrieval. In *Handbook of Information Science* (pp. 105–117). De Gruyter Saur.

Appendix

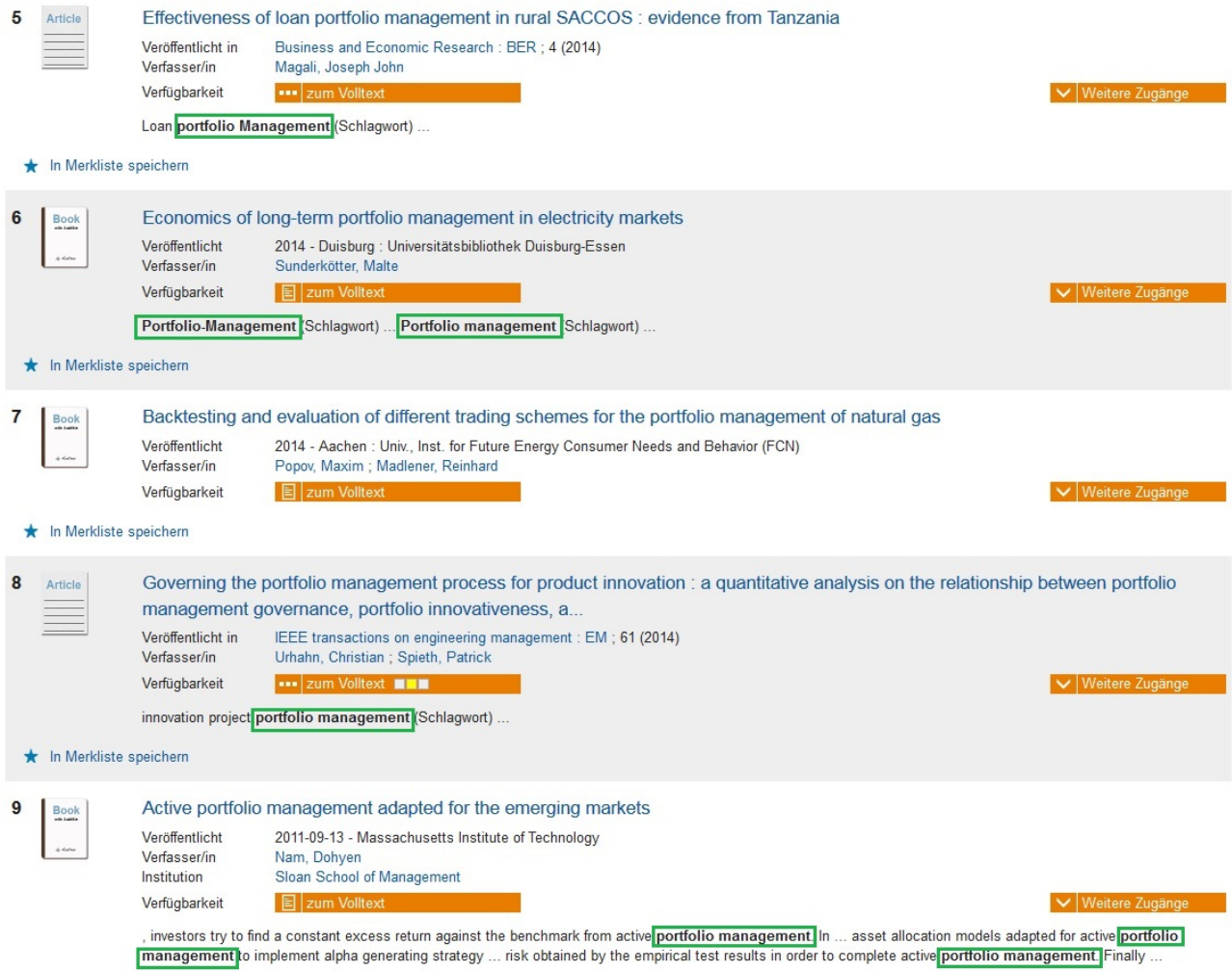

Figure 7: Result list in EconBiz

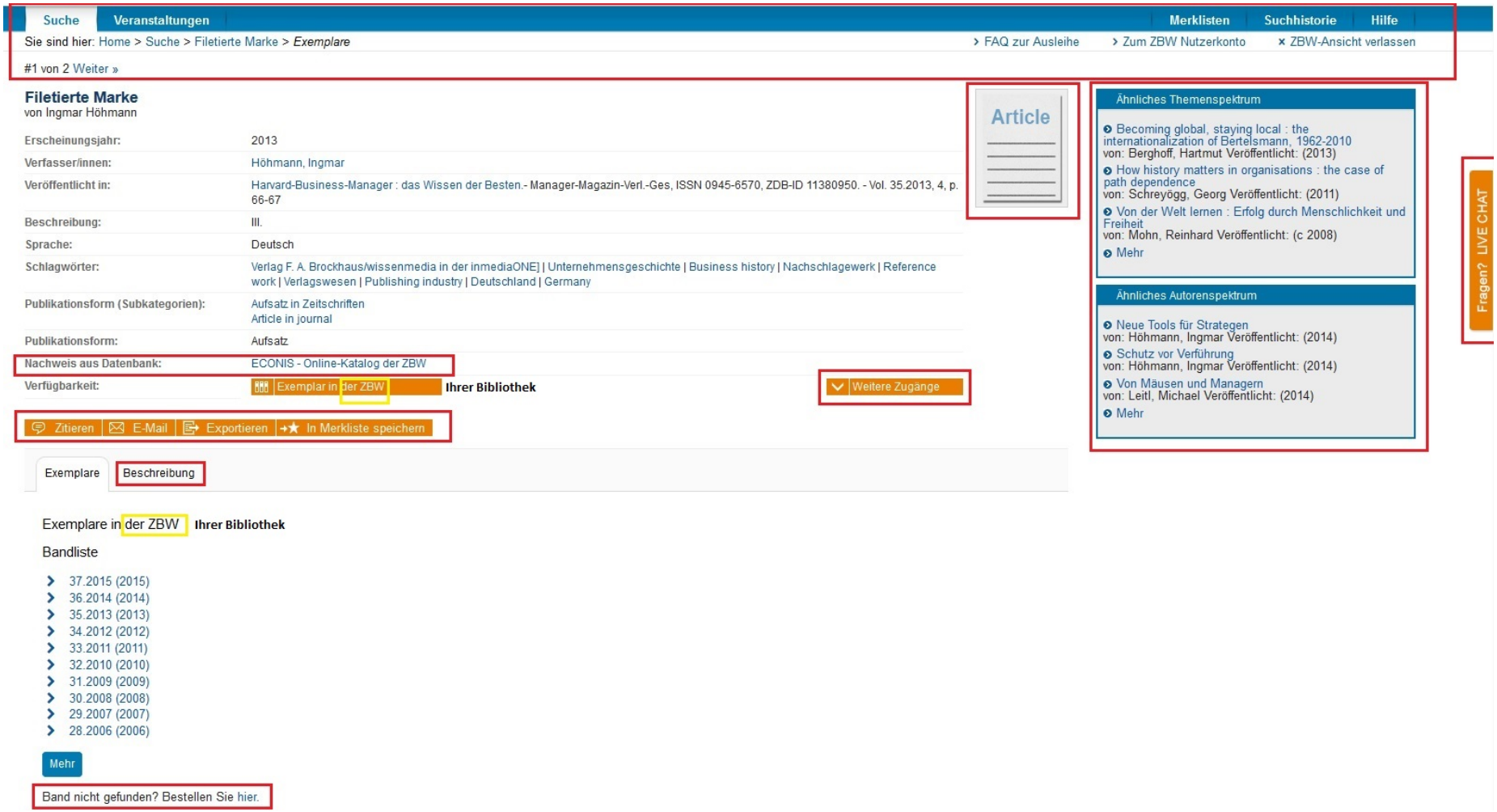

Figure 8: Result in EconBiz, document type 'article'

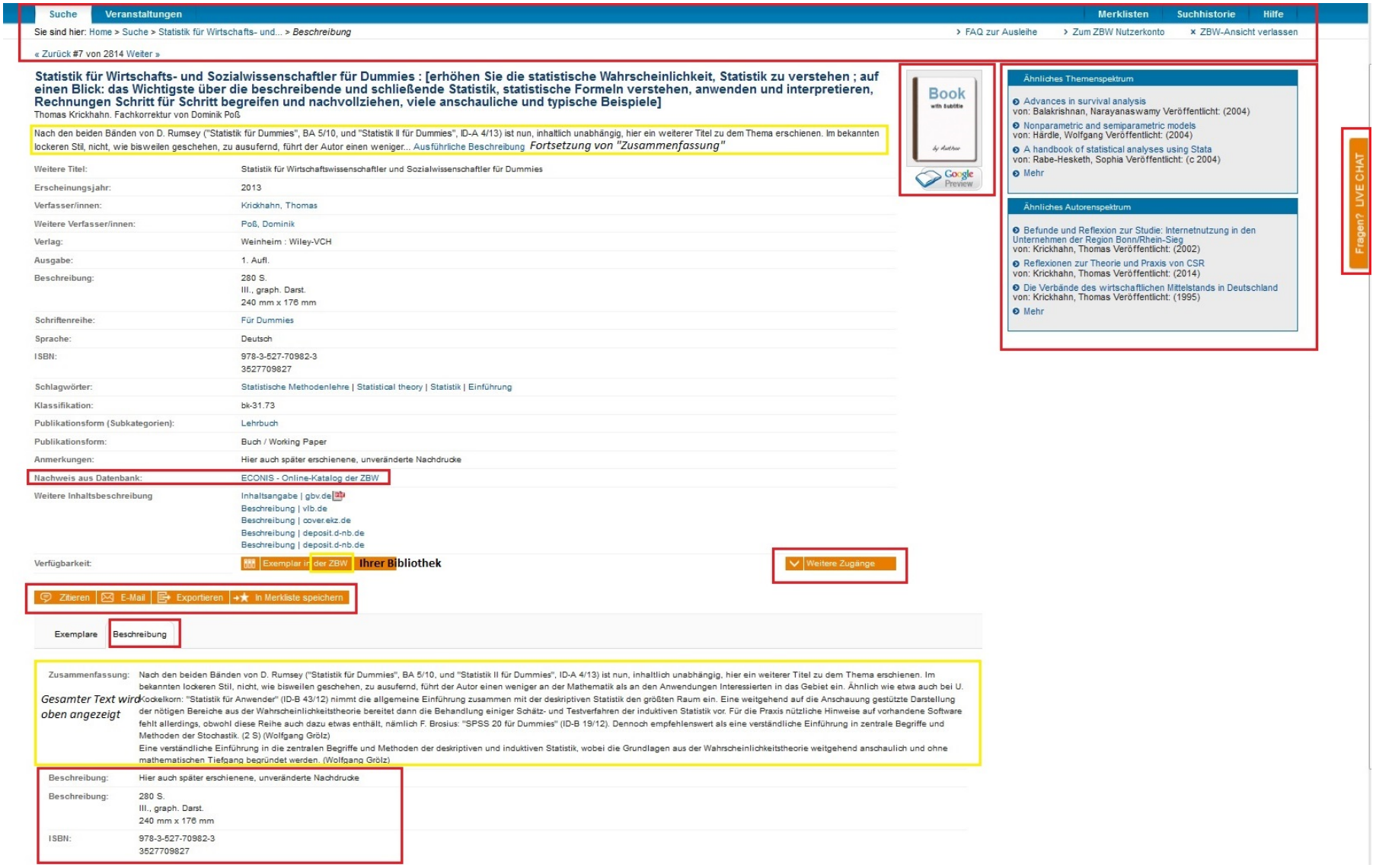

Figure 9: Result in EconBiz, document type 'book', with description

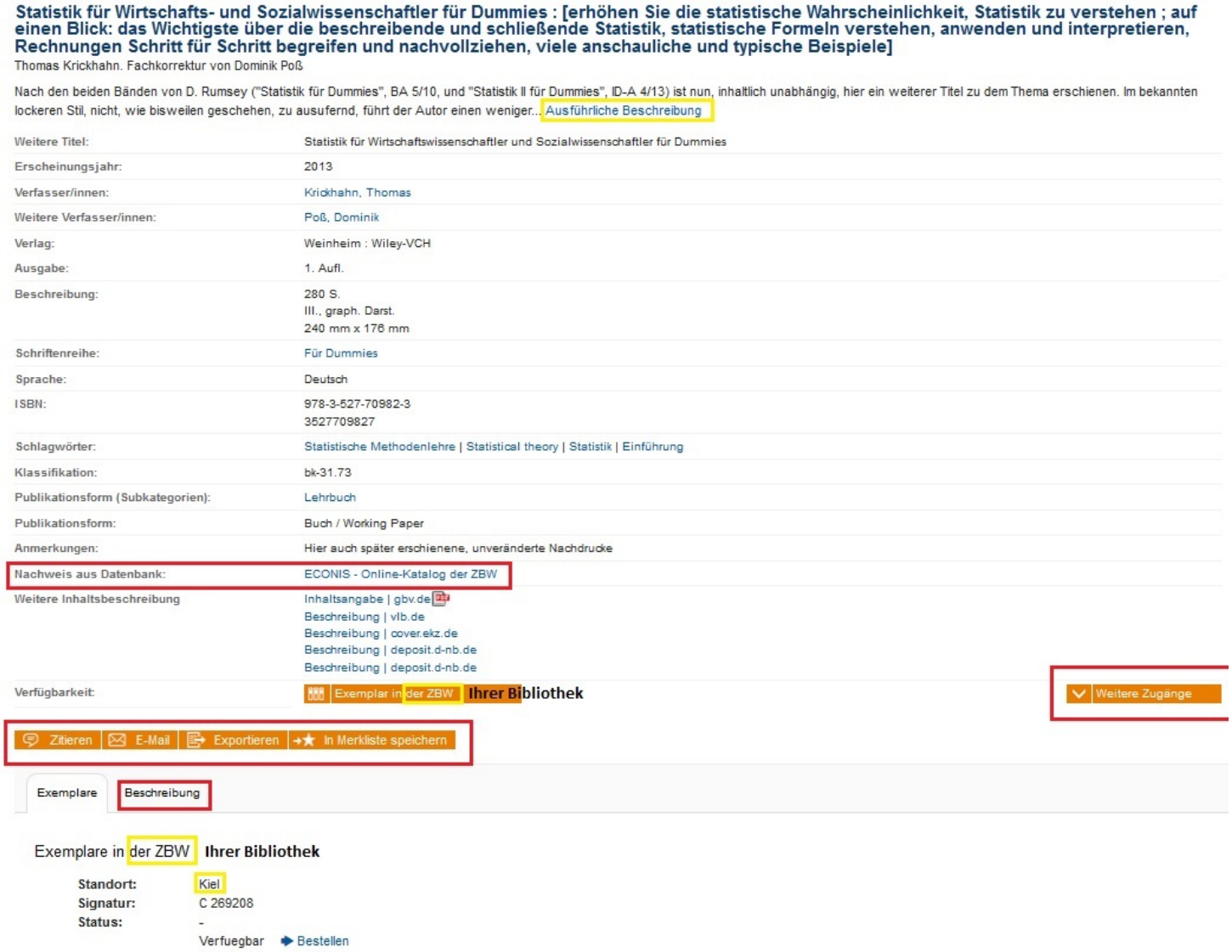

Figure 10: Result detail in EconBiz, document type ‘book’, with information on available copies