Supporting Book Search

A Comprehensive Comparison of Tags vs. Controlled Vocabulary Metadata

Bogers, Toine; Petras, Vivien

Published in:
Journal of Data and Information Management

DOI (link to publication from Publisher):
10.1515/dim-2017-0004

Creative Commons License
CC BY-NC-ND 4.0

Publication date:
2017

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (APA):
Bogers, T., & Petras, V. (2017). Supporting Book Search: A Comprehensive Comparison of Tags vs. Controlled Vocabulary Metadata. *Journal of Data and Information Management*, *1*(1), 17-34. https://doi.org/10.1515/dim-2017-0004

Research Article Open Access

Toine Bogers*, Vivien Petras

Supporting Book Search: A Comprehensive Comparison of Tags vs. Controlled Vocabulary Metadata

https://doi.org/10.1515/dim-2017-0004
received May 16, 2017; accepted June 20, 2017.

Abstract: Book search is far from a solved problem. Complex information needs often go beyond bibliographic facts and cover a combination of different aspects, such as specific genres or plot elements, engagement or novelty. Conventional book metadata may not be sufficient to address these kinds of information needs. In this paper, we present a large-scale empirical comparison of the effectiveness of book metadata elements for searching complex information needs. Using a test collection of over 2 million book records and over 330 real-world book search requests, we perform a highly controlled and in-depth analysis of topical metadata, comparing controlled vocabularies with social tags. Tags perform better overall in this setting, but controlled vocabulary terms provide complementary information, which will improve a search. We analyze potential underlying factors that contribute to search performance, such as the relevance aspect(s) mentioned in a request or the type of book. In addition, we investigate the possible causes of search failure. We conclude that neither tags nor controlled vocabularies are wholly suited to handling the complex information needs in book search, which means that different approaches to describe topical information in books are needed.

Keywords: book search, controlled vocabularies, social tagging, query analysis, failure analysis

***Corresponding author: Toine Bogers:** Science and Information Studies, Department of Communication & Psychology, Aalborg University Copenhagen, Denmark, E-mail: toine@hum.aau.dk
Vivien Petras: Berlin School of Library and Information Science, Humboldt-Universität zu Berlin, Germany

1 Introduction

In determining which book to read on a certain topic or for a specific audience, we have long relied on human expertise—be it a librarian, book seller or friend. Today, book search has been transformed into a more convenient affair, for example, with sophisticated book search engines such as Google Books that allow the full-text search of millions of books.

Still, the full-text of a book may not be sufficient to satisfy every information need, such as “Is this book satirical?”. Even more, the full-text is simply not available in many book search applications such as book seller or library catalogs. Even Amazon relies on information about the book (metadata) more than the book content and complements its search functionality with additional features, such as a recommendation system based on purchase histories.

This article explores the effectiveness of different book metadata elements in satisfying book search requests[1]. Different types of metadata elements, such as bibliographic metadata, controlled vocabulary (CV) terms[2], and user-generated content (e.g., reviews and tags) contain different information and should therefore be able to satisfy different information needs.

Our particular focus here is on the effectiveness of CVs vs. tags in book search. The discussion about the relative merits of CVs vs. uncontrolled text for search is an old one. Even the availability of full-text did not close the discussion as advances in text processing and semantic applications make CV a viable option for precise searching. User-generated content in the form of tags or reviews added a new, uncontrolled text variant to this discussion. Tags, if

[1] This text is an adapted and extended version of two conference papers: [5] and [6].

[2] In this paper, we use the term controlled vocabulary (CV) to denote any form of taxonomy, categorization or language-controlled terminology (e.g., subject headings) that prescribes the form or term for a certain concept that is described [13].

applied at scale, provide more keywords than CVs in many settings and could therefore improve search. However, the synonymy and polysemy vocabulary problems inherent in tags could also render search more imprecise, finding more irrelevant documents.

Comparative research on the advantages and disadvantages of tags vs. CVs has largely remained theoretical. Exploratory studies studied the potential of these metadata elements for search, but used small datasets that may not be realistic. Koolen's empirical comparison of different book metadata elements is a notable large-scale exception [22]. He found that reviews outperformed both bibliographic and CV metadata for real-world information needs. In this paper, we extend his research by delving deeper into the following problem statement:

> **PS** *How does the retrieval performance of tags and CVs compare under carefully controlled circumstances and what are the causes for the differences?*

We present an empirical comparison of CVs vs. tags in the book search domain using data provided from LibraryThing (LT), Amazon, the Library of Congress (LoC), and the British Library (BL). Real information needs are used to determine the search success for either book metadata element. Our contributions are:

1. A comparative analysis of the contributions of different metadata element sets for book search using a large-scale test collection.
2. A comparative analysis of tags and CVs, focusing on the complementarity of these metadata elements for search and potential popularity effects of tags.
3. A detailed request-by-request analysis based on the requested book type or relevance aspect searched for, that shows which request types work better with tags or CVs.
4. A failure analysis to determine why certain book search requests succeed while others fail based on request types and other collection features.

The remainder of this article is organized as follows. Section 2 presents related research on tag and CV searching, book search, book information needs and search request analysis. Section 3 describes the data, search requests and experimental methodology used in this study. We start our experiments in Section 4 by comparing the different types of book metadata and and their usefulness for book search. Section 5 zooms in on the comparison between tags and CV terms, discussing the popularity effect for tags and the complementarity of these metadata elements. Finally, in Sections 6 and 7 we perform an in-depth analysis of the book search requests themselves to uncover likely explanations for our results. This analysis includes a closer look at the types of books requests and at the possible causes for search failure. We conclude with a discussion of our findings and their wider implications in Section 8.

2 Background

In this chapter, we briefly discuss the relevant related research on (1) searching with natural language tags or CVs in general, (2) using tags and CVs in book search, and (3) book search information needs and search request analysis.

2.1 Uncontrolled Text, Controlled Vocabularies and Tags for Search

The advantages and disadvantages of CVs vs. uncontrolled text seem to be in balance with each other [1, 13, 14]. CVs provide synonym and homonym control and the expression of semantic relationships between concepts. As subject headings, thesauri or classifications, they represent concepts for describing a document's content. In search, the vocabulary control ensures both high recall and precision as a concept represented in a document can be found with whichever search term and will not be confused with polysemous terms. Conversely, CVs have large development costs and may use outdated terminology, which is harmful when searching newer documents. The natural or uncontrolled language in titles, abstracts or the full-text of a document and later in user-generated content (such as tags) is more varied and represents the author or user terminology, but can also lead to fewer or irrelevant search results.

Experiments comparing the effectiveness of CVs vs. uncontrolled text have as varied results as the pro and con arguments for either of them. Some showed the same effectiveness for CVs and uncontrolled text [42], some an advantage for CVs [7, 29], some an advantage for uncontrolled text [11, 32]. Surprisingly, the Cranfield experiments, which established the standard search evaluation processes still used today, showed that individual natural language terms performed better than CVs, which in turn performed better than the full-text [12]. Other studies found that CVs and uncontrolled text complement each other and add different aspects to an improved search performance [16, 40, 45].

Tags have been criticized for the same lack of vocabulary control as other uncontrolled metadata elements, even though the vocabulary variation in a massive dataset may be negligible in search [39, 49]. Tags were also found to be easier to apply than CVs in content descriptions [15]. In web search, studies found that tags help in finding relevant documents, but that tag terms were also included elsewhere in documents (e.g., the title) possibly rendering tags unnecessary [4, 18]. Other studies found a complementary effect for CVs and tags for search as the terms did not overlap [31, 46].

The complementarity of tags and CVs may be due to their different characteristics. For example, LibraryThing tags have been found to contain subjective, contextual, and personal descriptions [30, 50], whereas CVs such as the Library of Congress Subject Headings (LCSH) are required to be more abstract, objective, and impersonal. Whereas tags will cover whatever comes into a user's mind about the document content, rule-based CVs may only represent specific aspects (such as the main topics of a document) [33].

2.2 Tags and Controlled Vocabularies for Book Search

While searching the full-text of books [20, 51] has been neglected in research, book search in the metadata of library catalogs has been studied well over the last few decades [21, 43, 47]. Several studies have focused on the search effectiveness of different metadata elements.

Experimenting on short queries from the 2007 INEX Book Search Track, Magdy and Darwish demonstrated that just using the titles and chapter headings was almost as successful as using the full-text of a book for a search [35]. This indicates that information contained in these specific metadata elements could be more helpful in search.

For book search, the complementarity of tags and CVs could be shown in some studies [2, 48], whereas others found them to be equivalent or tags as more helpful in providing more terms [17, 34]. Conversely, in the same studies, self-referential tags introduced noise into the search, making tags less effective for some search requests. Tags also did not work as well for less popular (and thereby less tagged) books. Our study provides an in-depth analysis of the complementarity of tags and CV in book search delving deeper into the aspects that make either metadata element more successful.

In our study, we use the the INEX Social Book Search Track book collection, which has been studied before [26–28], allowing for some comparisons across experiments. The INEX Social Book Search Track[3] researches book search from different perspectives, for example, focusing on automatically detecting and categorizing book search requests, improving system algorithms for search or investigating how people interact with book search interfaces [23]. For judging the topical relevance of book suggestions for 24 book search requests in this collection, the reviews turned out to be more important than the core bibliographical elements or the tags [25]. As was already mentioned, Koolen [22] found that user reviews added most to retrieval success compared to core bibliographic metadata or CV in this collection (see Section 3.2).

Compared to the Koolen study [22], this paper concentrates on the relative effectiveness of tags vs. CVs. This study therefore uses a subset book collection from the INEX Social Book Search Track, where, differently from the previous studies, each document contains CVs as well as tags so that each has the same chance to contribute to the search performance.

2.3 Book Search Information Needs and Search Request Analysis

Research on book search information needs focuses on the search aspects that are combined in book search requests. Book search requests are rooted in a cultural context and should be treated as such [8]. A study on fiction book requests found that they included aspects of familiarity besides bibliographic information [36].

The book search requests used in our experiments come from the LibraryThing forums. They are real search requests and represent complex information needs. In contrast to search requests sent to web search engines or conventional book search engines like Amazon or Google Books, they are much longer and richer. A previous analysis of these book search requests found different relevance aspects such as novelty, engagement, and familiarity that are not represented in traditional book metadata [24]. We will discuss the impact of different relevance aspects in (Section 6.1).

Failure analysis attempts to identify the reasons why search requests fail in particular collections. For example, an important failure analysis of the standard TREC search requests found that semantic relationships represented in the search request may not be interpreted correctly by the search engine [9]. Other research tries to predict the difficulty of a search request based on linguistic features

[3] See http://social-book-search.humanities.uva.nl/#/overview (last accessed May 12, 2017)

such as the request terminology or the relationship between a request and the collection documents. Carmel and Yom-Tov offer a good summary of such approaches [10].

In this study, we combine several approaches for analyzing book search requests. We do not only analyze, which aspects in book search requests might be more effectively served by tags or by CVs, but we also perform a failure analysis to research why some requests are bound to fail.

3 Methodology

To study the contribution of different metadata elements to the search performance in book search, a document collection containing both controlled vocabulary metadata as well as tags is needed. The collection should be representative in terms of size, type and variety and include real-world information needs with relevance judgments. The INEX Amazon/LibraryThing collection meets these requirements. Section 3.1 describes the collection, while Section 3.2 explains how the collection was filtered to allow for controlled experiments on the retrieval effectiveness of tags and CVs. Finally, Section 3.3 describes the book search requests used for our experiments and Section 3.4 describes our experimental setup and evaluation protocol.

3.1 The Amazon/LibraryThing Book Collection

The Amazon/LT collection was adopted as the test collection for the INEX Social Book Search Track[4] and continues to be used in the SBS Workshops[5]. It was collected by Beckers et al. [3] and contains over 2.8 million book records in XML format aggregated from Amazon, the British Library (BL), the Library of Congress (LoC), and LibraryThing (LT). Book records (also referred to as 'documents') consist of over 40 different metadata elements from the different providers [27].

Core bibliographic metadata (e.g. author, title) are provided by Amazon, which also adds Dewey Decimal Classification (DDC) class numbers, Amazon subject headings, category labels from Amazon's category system, and the Amazon user reviews. BL and LoC provided CV terms (DDC and Library of Congress Subject Headings (LCSH)) for 1.15 million records and 1.25 million records respectively. LT provides all tags added to the books in the Amazon/LT collection. The book records were aggregated from the different providers by matching the ISBNs of the individual records.

3.2 Filtering

Filtering different Metadata Element Sets. In order to compare the search effectiveness of different metadata elements, we filtered the combined collection to only include a certain type of metadata element. These individual and different combinations of the element sets are called test collections in this paper. Metadata elements that were unlikely to contribute to the effectiveness of a search were removed (e.g., page numbers). The metadata elements used for our experiments are shown in Table 1.

Filtering for Tag and CV Equality. When comparing different metadata elements—tags and CV in particular—it is important to make the comparison as fair as possible. To give tags and CV a more equivalent chance in search and to be able to examine how they compare for individual documents, we filtered the original Amazon/LT collection so that all book records that did not contain at least one CV term and at least one tag were removed. This resulted in a test collection with 2,060,758 book documents.

Differences in CV Quality. There may be differences in quality of the CV content from the different providers. For example, the LCSH terms from BL or LoC could be of better quality than the subjects provided by Amazon. To examine this question, we conducted an experiment searching either in the CV terms from Amazon, BL or LoC. This required filtering the Amazon/LT to documents that included at least one CV term from each individual provider. This more restrictive filtering criterion reduced the number of searchable documents to 353,670. The experiments showed that there was no statistically significant difference between the CVs from different providers according to a repeated-measures ANOVA with a Greenhouse-Geisser correction using the Narrative representation of the search requests ($F(2.167, 465.921) = 2.050$, $p = .126$). Since the performance does not differ significantly between metadata providers, we do not distinguish between the different CV sources in the experiments and treat them as a combined CV metadata element.

Filtering for Tag Popularity. The social aspect of tagging (multiple annotators tag the same object) means that a tag

[4] http://inex.mmci.uni-saarland.de/data/documentcollection.html#books, last accessed May 12, 2017.

[5] http://social-book-search.humanities.uva.nl/#/overview, last accessed May 12, 2017.

Table 1: Overview of the Amazon/LT metadata element sets used in our experiments and their origins.

Provider	Bibliographic data (Core)	Controlled vocabulary content (CV)	User-generated content (UGC)
Amazon	Author, title, publication year, publisher	DDC class labels, Amazon subjects, geographic names & category labels	Reviews
BL		DDC class labels, LCSH topical, chronological & genre/form terms, geographic & personal names	
LoC		DDC class labels, LCSH topical, chronological & genre/form terms, geographic & personal names	
LT			Tags

can be assigned more than once to a book document. This is called the popularity effect [37]: popular books receive more (and more of the same) tags than unpopular books, whereas CV terms are more evenly distributed across all books. The Tags collection contains book documents with the original tag frequencies. In order to test the popularity effect of tags, we created an additional test collection called Unique tags. This test collection contains the same tags as the Tags collection, but each tag is reduced to a single appearance in a book record.

Tables 2 and 3 contain an overview of the test collections with the different metadata element combinations that were used in this study.

Collection Statistics. Table 4 shows type and token counts for the different test collections, both as total counts and averages per document. The numbers for the Tags collection show that there may be a popularity effect for tags, as the average number of tokens per document is much higher than the average number of types, at 119.5 vs. 13.1. Interestingly, CV elements have a higher average number of types per document at 36.5 than Tags, meaning that there are more unique CV terms than Tags in an average document. CV documents also have a higher average number of tokens at 53.3 than types. This means that some CV terms from the different providers overlap, giving the CV collection a slight popularity effect as well. The stricter filtering in the Unique tags collection gives both elements a fairer playing field with respect to the number of terms in the document, although technically, there are now more CV terms in a document than Tags. Unsurprisingly, reviews are the richest metadata element in textual content.

Table 2: Test collections with individual metadata elements.

Metadata element
Core
CV
Reviews
Tags
Unique tags

Table 3: Test collections with combined metadata elements.

Metadata elements
Core + CV
Core + Reviews
Core + Tags
Core + Reviews + Tags
Core + Reviews + Tags + CV
Reviews + Tags
Tags + CV
Unique tags + CV

3.3 Book Search Requests & Relevance Judgments

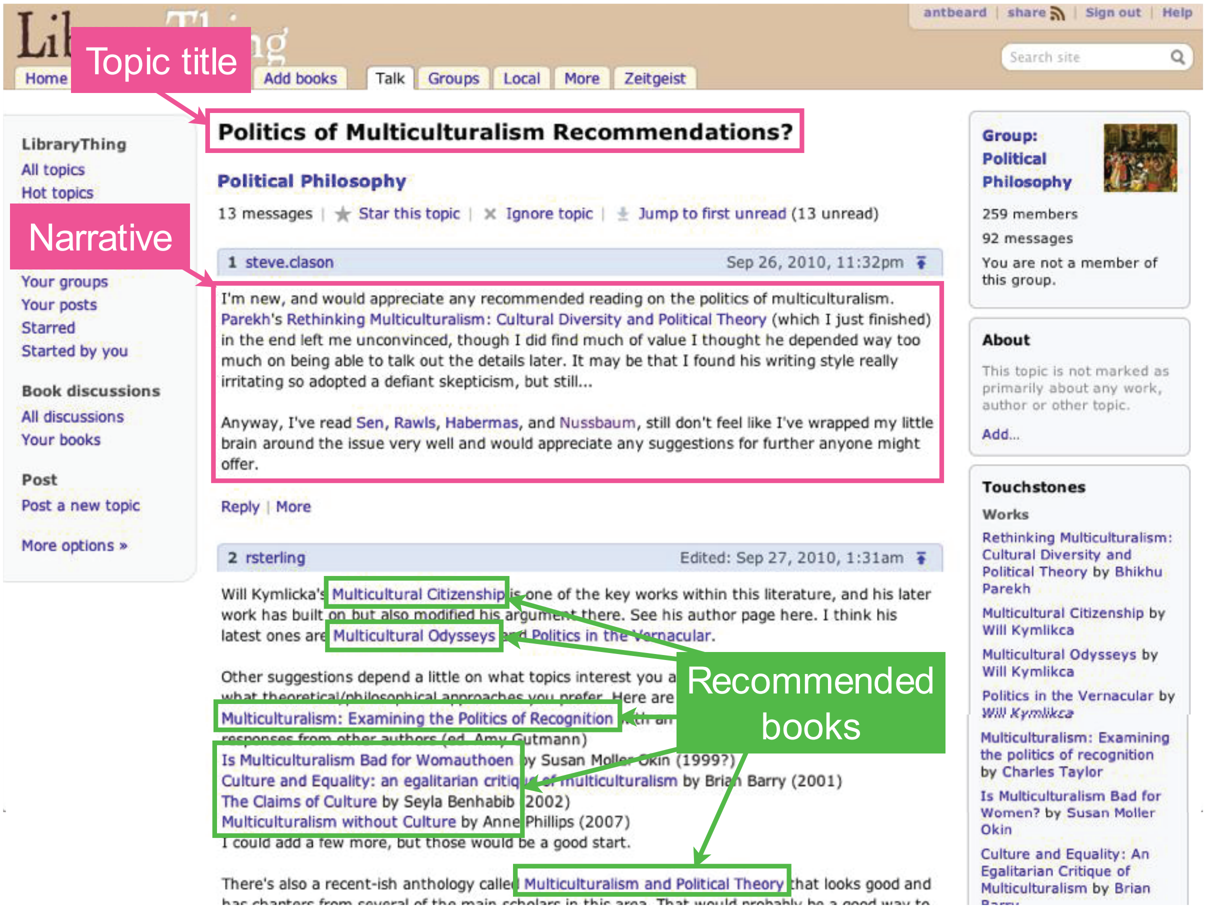
For the Amazon/LT test collection, a set of search requests representing actual book-related information needs along with relevance judgments are provided [25]. The search requests were harvested from the LT discussion forums where a member can ask for book recommendations and other members provide suggestions. Examples for such requests include (1) asking for suggestions on books to read about a certain topic or from a particular genre; (2) known-item requests where the user is looking for a book (s)he cannot remember the title by specifying plot details; and (3) book recommendations based on specific personal opinions. Frequently, the LT member requests include example books that the requester has already read and (dis)liked. Figure 1 shows an example book request[6].

Search requests from the Amazon/LT collection have different components, such as the Title and the Narrative of the original LT request—the title of the forum post and the text in the requester's post. The requester-provided narrative is usually longer and explains more about the context of the request including the books mentioned by

[6] Topic 99309, available at http://www.librarything.com/topic/99309, last accessed May 12, 2017.

Table 4: Type and token statistics for the different metadata element sets.

Metadata element(s)	#types	#tokens	avg. types/doc	avg. tokens/doc
Core	2,368,387	29,502,833	12.9	14.3
CV	2,208,694	109,793,695	36.5	53.3
Reviews	553,943,057	2,085,063,187	505.4	1902.4
Tags	2,272,393	246,313,480	13.1	119.5
Unique tags	2,272,393	47,253,002	13.1	22.9
Core + CV	2,427,963	137,235,770	28.8	66.6
Core + Tags	2,482,657	273,755,555	17.3	132.8
Core + Tags + CV	2,535,366	381,488,492	34.2	185.1
Core + Reviews + Tags + CV	1,282,874,111	4,792,960,555	622.5	2325.9
Tags + CV	2,353,659	354,046,417	29.4	171.8
Tags + Reviews	590,536,035	2,329,744,735	286.6	1130.7
Unique tags + CV	2,353,659	154,985,939	29.4	75.2



Topic title

Politics of Multiculturalism Recommendations?

Political Philosophy

13 messages | Star this topic | Ignore topic | Jump to first unread (13 unread)

Narrative

1 steve.clason Sep 26, 2010, 11:32pm

I'm new, and would appreciate any recommended reading on the politics of multiculturalism. Parekh's Rethinking Multiculturalism: Cultural Diversity and Political Theory (which I just finished) in the end left me unconvinced, though I did find much of value I thought he depended way too much on being able to talk out the details later. It may be that I found his writing style really irritating so adopted a defiant skepticism, but still...

Anyway, I've read Sen, Rawls, Habermas, and Nussbaum, still don't feel like I've wrapped my little brain around the issue very well and would appreciate any suggestions for further anyone might offer.

2 rsterling Edited: Sep 27, 2010, 1:31am

Will Kymlicka's Multicultural Citizenship is one of the key works within this literature, and his later work has built on but also modified his argument there. See his author page here. I think his latest ones are Multicultural Odysseys and Politics in the Vernacular.

Other suggestions depend a little on what topics interest you a
what theoretical/philosophical approaches you prefer. Here are
Multiculturalism: Examining the Politics of Recognition with an
responses from other authors (ed. Amy Gutmann)
Is Multiculturalism Bad for Womauthoen by Susan Moller Okin (1999?)
Culture and Equality: an egalitarian critique of multiculturalism by Brian Barry (2001)
The Claims of Culture by Seyla Benhabib (2002)
Multiculturalism without Culture by Anne Phillips (2007)
I could add a few more, but those would be a good start.

There's also a recent-ish anthology called Multiculturalism and Political Theory that looks good and has chapters from several of the main scholars in this area. That would probably be a good way to

Recommended books

Group: Political Philosophy

259 members

92 messages

You are not a member of this group.

About

This topic is not marked as primarily about any work, author or other topic.

Add...

Touchstones

Works

Rethinking Multiculturalism: Cultural Diversity and Political Theory by Bhikhu Parekh

Multicultural Citizenship by Will Kymlikca

Multicultural Odysseys by Will Kymlikca

Politics in the Vernacular by Will Kymlikca

Multiculturalism: Examining the politics of recognition by Charles Taylor

Is Multiculturalism Bad for Women? by Susan Moller Okin

Culture and Equality: An Egalitarian Critique of Multiculturalism by Brian Barry

Figure 1: An information need from the LibraryThing discussion forums.

the requester. We will use the combined Title+Narrative representations as search requests, if not reported otherwise[7].

The book suggestions in reply to the LT forum request were used as relevance assessments for the search requests. Based on additional criteria—such as whether the suggester had read the book or whether the book was then added to the book catalog of the requester—a graded relevance scheme was applied, making some books more relevant than others [27]. From the 680 provided search requests and relevance assessments from the 2014 edition of the INEX Social Book Search track, 340 randomly selected topics were used for training and 334 topics for testing purposes.

3.4 Experimental Setup

Retrieval Setup. For retrieval experiments, we used language modeling with Jelinek-Mercer smoothing as implemented in the Indri 5.4 toolkit[8]. Previous work has shown that for longer queries such as the rich Amazon/LT topic representations, JM smoothing outperforms Dirichlet smoothing [52]. We did not use any of the Indri-specific belief operators when constructing queries.

Ideally, a book search engine would be optimized for the specific combination of metadata elements indexed by the search engine. To emulate this situation and to avoid giving an unfair advantage to one collection over another, we optimized the retrieval performance of Indri for the test collections. We randomly split our original topic set of 680 into a training set and test set of 340 topics each. We used grid search to determine optimal parameter settings on our training topics. These optimal settings were then used on the 334 test topics to produce the results presented in the remainder of this paper.

We optimized three different parameters:

- **Degree of smoothing.** The λ parameter controls the influence of the collection language model, with higher values giving more influence to the collection language model. We varied λ in steps of 0.1, from 0.0 to 1.0.
- **Stopword filtering.** Either no filtering or using the SMART stop word list.
- **Stemming.** Either no stemming or Krovetz stemming.

This resulted in 44 (= 11 × 2 × 2) different possible combinations of these three parameters[9]. These optimal settings were then used on the 334 test book search requests to produce the results presented in the remainder of this paper.

Evaluation. To measure retrieval effectiveness, we use the NDCG@10 metric, which is also used in the INEX Social Book Search Track and thus enables comparability and replicability of our results. NDCG stands for Normalized Discounted Cumulated Gain and was proposed by Järvelin and Kekäläinen [19]. It is a metric that provides a single-figure measure of retrieval quality across recall levels and uses graded relevance judgments, preferring rankings where highly relevant books are retrieved before slightly relevant books. We use NDCG@10—NDCG cut off at rank 10—because most users do not inspect search results beyond the first page [38], so a high-quality results ranking in the top ten is most important.

As a result of the filtering we apply to the original Amazon/LT collection, occasionally relevant documents for certain topics are also filtered out. This is necessary to keep the evaluation process fair: penalizing a search engine for not retrieving documents that do not exist, is pointless. Consequently, the relevance assessments for those documents were also removed to avoid skewing the results.

When comparing the retrieval performance of different runs or results list subsets, we perform statistical significance testing. We use an α of 0.05 throughout this paper. In accordance with the guidelines proposed by Sakai [44], we use two-tailed paired t-tests when comparing the performance of pairs of retrieval runs and also report the effect size (ES) and the 95% confidence interval (CI). For comparisons between three or more retrieval runs, we use a repeated-measures ANOVA test. On occasion, other statistical tests will be used where relevant, such as the X^2-test.

4 Comparing Different Metadata Element Sets

This section compares the search effectiveness of the individual metadata elements and their different combinations. We report on experiments using the Narrative representation for search requests.

[7] Some experiments will report on just the Narrative representations.

[8] Available at http://sourceforge.net/projects/lemur/files/lemur/indri-5.4/, last accessed April 21, 2017.

[9] Readers interested in these optimal parameter settings are referred to http://toinebogers.com/?page_id=738 for a complete overview.

Question 1: Which of the individual metadata elements contributes most to search success?

Answer: Reviews and the Core bibliographic metadata provide the best performance.

Table 5 shows that Reviews provide the best retrieval performance. There is a statistically significantly difference between this element and the other elements (*F*(2.605, 794.414) = 18.770, *p* < .0005). The Core bibliographic metadata elements also significantly outperform the CV terms according to a two-tailed paired *t*-test (*t*(305) = 2.139, *p* < .05, ES = 0.122, 95% CI [0.0016, 0.0385]).

Question 2: Which combination of metadata elements achieves the best performance?

Answer: Any combination of the elements outperforms the equivalent individual metadata element. The combination of all metadata elements achieves the best results.

Combining all metadata elements into one set results in the best performance. Except for the test collections containing Reviews + Tags together, combining all elements significantly outperforms all other element configurations.

Question 3: Does the addition of Core bibliographical metadata change the search performance?

Answer: There is no significant difference when combining Core bibliographical metadata with CVs. Including Core bibliographical metadata in general achieves a better performance.

Any real-world book search engine would always include the core bibliographic data in its documents. The NDCG@10 scores seem to bene t from adding the Core elements to other metadata elements. These differences are significant according to a two-tailed paired *t*-test (*t*(1307) = 4.799, *p* < .0005, ES = 0.13, 95% CI [0.0083, 0.0199]).

However, an interesting result is that adding CV terms to Core bibliographic metadata results in only a very small improvement. Indeed, these improvements are not statistically significant according to a two-tailed paired *t*-test (t(333) = -0.140, *p* = .889, ES = 0.001, 95% CI [-0.0107, 0.0092]).

It is possible that a combination of the Core bibliographical metadata and CV collections could result in interaction effects of a complementary nature. However, a repeated-measures ANOVA with a Greenhouse- Geisser correction again showed no statistically significant differences between the configurations *F*(2.406, 517.282 = 0.973, *p* = .391).

Table 5: Results for the individual metadata elements.

Metadata element	NDCG@10
Core	0.0533
CV	0.0319
Review	**0.0993**
Tags	0.0395

Table 6: Results for the combined metadata elements.

Metadata elements	NDCG@10
Core + CV	0.0540
Core + Reviews	0.1063
Core + Tags	0.0610
Core + Reviews + Tags	0.1114
Core + Reviews + Tags + CV	**0.1115**
Reviews + Tags	0.1046

5 Comparing Tags and Controlled Vocabulary Terms

This section focuses on the comparison of the search effectiveness for Tags and CV terms[10]. First, we test which individual metadata element contributes most the search performance in Section 5.1. We then check whether Tags outperform CVs because of the popularity effect in Section 5.2. Finally, in Section 5.3, we show that the two metadata elements are complementary in search for the book search requests that we tested on.

5.1 Tags vs. Controlled Vocabulary Terms

Question 4: Is there a difference in performance between CVs and Tags in search?

Answer: Tags perform significantly better than CVs. The combination of both sources in Tags + CVs results in even better performance, but not significantly so.

Table 7 shows the results for the five collections with either Tags or CVs with Figure 2 representing the same information graphically. There is a statistically significant difference between the five element sets according to a repeated-measures ANOVA with a Greenhouse-Geisser correction (*F*(2.529, 842.144) = 4.650, *p* < .01).

Tags provide a significantly better performance for the 334 requests compared to CVs according to a two-

[10] The remainder of the experiments are based on the combined Title+Narrative request representation. These results differ from Section 4.

tailed paired *t*-test ($t(333) = 2.171$, $p < .05$, ES = 0.118, 95% CI [0.0160, 0.0325]). However, combining the two in Tags + CVs results in even better performance, which suggests they are complementary to a degree. While this combination also significantly outperforms the original CVs collection ($t(333) = 2.874$, $p < .05$, ES = 0.157, 95% CI [0.0069, 0.0368]), Tags + CVs does not perform significantly better than the Tags collection ($t(333) = 1.194$, $p = .253$, ES = 0.066, 95% CI [-0.0031, 0.1263]).

5.2 Popularity Effect

> **Question 5:** Do Tags outperform CVs because of the popularity effect?
>
> **Answer:** No. The Unique tags collection (without repeated tags) performs even better than Tags.

The popularity effect can occur in two different ways: the *unique frequency* and the *frequency of occurrence* effect. *Unique frequency* refers to the phenomenon that there could be more unique tags than unique CV terms assigned to any book in the Amazon/LT collection. However, when comparing the average number of types in the types and token statistic for the collection (Table 4 in Section 3.1), one can see that there are actually nearly three times as many tokens in the CV collection per document (36.5) than in the Tags collection (13.1). Because we aggregated all CV terms from the providers (see Table 1), there are actually many more CV terms to search. The better search performance for tags cannot be explained because of higher unique frequencies, so this aspect of the popularity effect does not apply to our collection.

The *frequency of occurrence* aspect of the popularity effect refers to the social phenomenon of tags, which can be assigned to a book several times on LibraryThing (as many as LT members assign the tag to the same book). This aspect of the popularity effect was given as a possible explanation for the performance difference between Tags and CV, because a change in term frequencies can impact the search performance [22]. In fact, there are many more tokens than types occurring in the Tags collection (119.5) compared to the CV collection (53.5) according to Table 4. Note that the average number of tokens for the CV collection is also higher than the average number of types. This indicates that certain terms are also repeated in documents in the CV collection. The Unique tags collection was created to test the *frequency of occurrence* aspect by removing tags that were repeated so that only unique tags remained. The average number of types now almost equals the average number of tokens. As with the CV collection, certain terms were repeated in different tags, leading to a higher token count. Table 7 shows that even with removing the frequency information, the Unique tags collection still performs significantly better than CVs ($t(333) = 2.135$, $p < .05$ (0.033), ES = 0.117, 95% CI [0.0014, 0.0338]). The Unique tags collection performs even slightly better than the original Tags collection, but this difference is not statistically significant according to a two-sided paired-samples t-test ($t(333) = 0.139$, $p = .890$, ES = 0.007, 95% CI [-0.0070, 0.0080]). The *frequency of occurrence* aspect also does not explain the better search performance for Tags.

Removing the popularity effect from the equation leaves us with the interpretation that Tags simply seem to match the search request vocabulary (or underlying information needs) better than CVs do. However, this is not true for all search requests as the next section will show.

Table 7: Results for the different test collections.

Metadata elements	NDCG@10
CV	0.0348
Tags	0.0519
Unique tags	0.0524
Tags + CV	0.0566
Unique tags + CV	**0.0583**

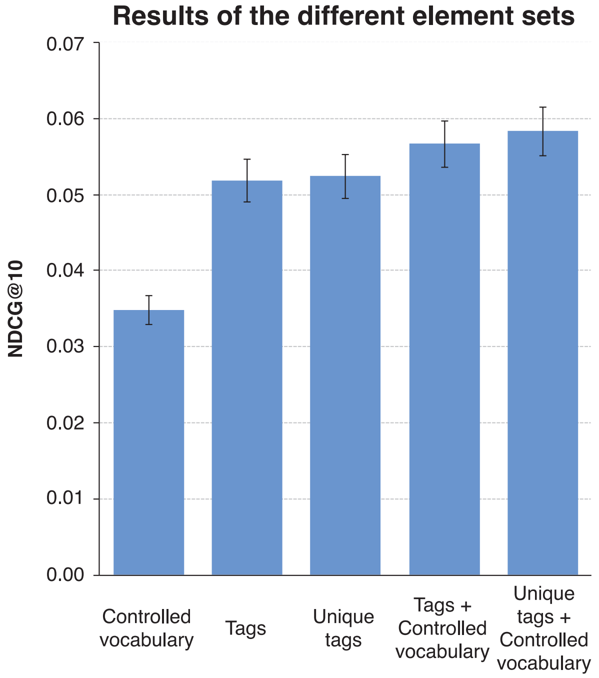


Figure 2: Results for the different test collections. Bars indicate average NDCG@10 scores over 334 topics, with error bars in black.

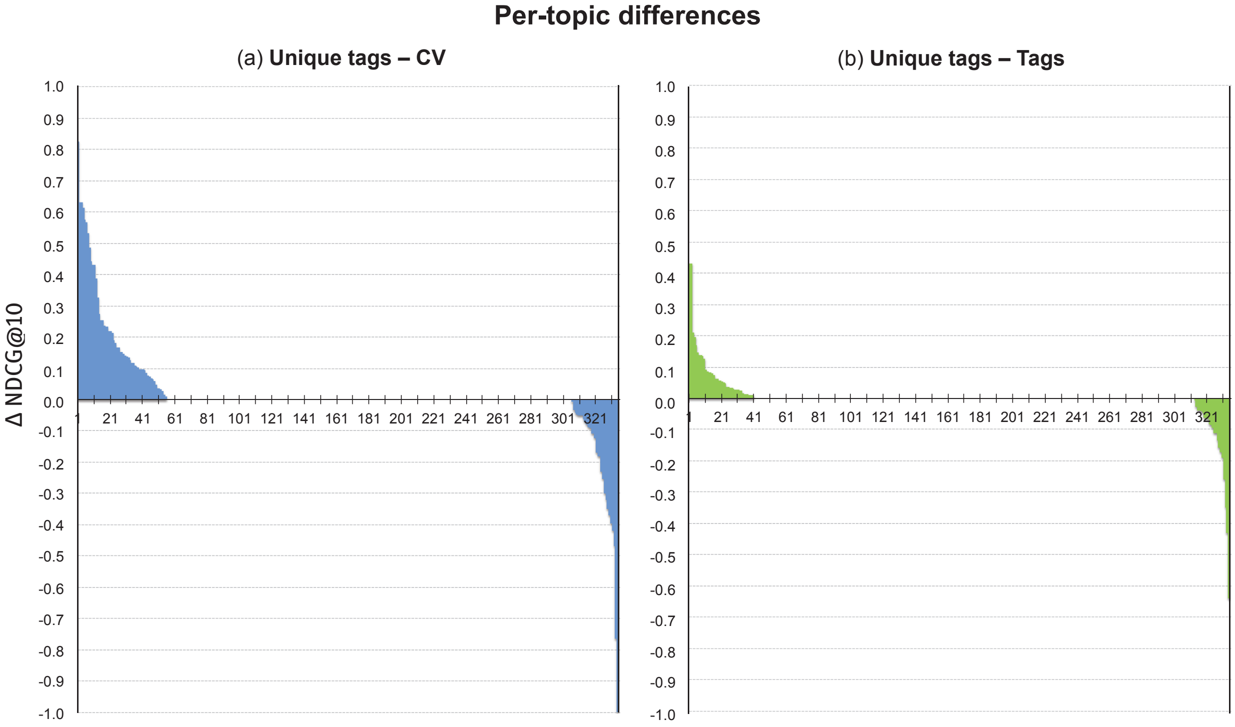


Figure 3: Differences in search performance ordered by per-request difference between (a) the Unique tags and CV, and (b) the Unique tags and Tags collections. Bars above the horizontal axis represent requests where Unique tags perform better, bars below the horizontal axis represent requests where the other collections perform better.

5.3 Complementarity

> **Question 6:** Do Tags, Unique tags, and CV complement or cancel each other out in terms of performance?
>
> **Answer:** Tags, Unique tags, and CV complement each other: they are successful on different sets of requests.

The best-performing of our five test collections is the Unique tags + CVs collection. A combination of Tags + CVs also leads to a better performance then either Tags or CVs. This indicates that for a number of search requests, a combination of the two metadata elements increases the search success, making them complementary.

It is important to note that a two-sided paired t-test for the difference in means between the Unique tags + CV and Unique tags collections showed no statistical significance ($t(333) = 1.062$, $p = .289$, ES = 0.058, 95% CI [-0.0051, 0.0169]. Surprisingly, neither did the combined Unique tags + CV collection improve significantly over the CV collection ($t(333) = 3.398$, $p < .05$ (.001), ES = 0.186, 95% CI [0.0099, 0.3712].

So while the combination did not significantly improve the overall results, a change in the average NDCG@10 number still meant that some search requests were improved, meaning that the other metadata element added information for some search requests. This effect can be seen in the per-request difference plots in Figure 3.

While Figure 3a shows the number of requests, which were better served by the Unique tags (bars above the horizontal line) or the CV (bars below the horizontal line) collection, Figure 3b shows the equivalent for the Unique tags vs. the Tags collections.

Both figures show that Unique tags perform better in a higher number of search requests compared to the other collections (based on the area above the horizontal axis compared to the area below in the figures)—this was already indicated by the average NDCG@10 numbers in Table 7. We can infer two conclusions from these results: (1) Unique tags and CVs have a complementary performance for some search requests (blue areas) and (2) the frequency of occurrence information in the Tags collection helps for some search requests after all, but for more requests, it actually hurts the performance (as seen by the better performance of the Unique tags collection).

The complementarity effect is especially important for those search requests where one of the individual metadata elements retrieves no relevant results, but the other would. Table 8 looks at these search requests. Out of the 334 tested search requests, only 94 book search requests found relevant documents in its top 10 results for either the UniqueTags, the CV or the UniqueTags+CV collections, these are compared in the table.

The upper half of Table 8 shows the absolute numbers for the search requests represented in Figure 3a. Unique tags achieved better results than CVs for 53 search requests, for 47 of them, the CV collection would not have found a single relevant document. Vice versa, using the CVs was better than the UniqueTags collection for 27 book search requests, but 11 of them would have found a relevant document also in the UniqueTags collection. Comparing these absolute numbers shows that there is a small complementarity effect, but that book search benefits more from adding Tags to the collection than CVs.

Nevertheless, CVs add to the search performance sometimes: the combination Unique tags + CV finds more relevant results than either individual metadata element in 24 cases. Most interestingly, 7 search requests would not have retrieved a relevant book in the top 10 search results if they had been searched alone in the individual element sets, but the combination of Unique tags + CV was more successful.

While we can find significant performance differences for the individual metadata element sets, Figure 3a shows that the majority of search requests appears to have no difference in performance between Unique tags[11] and CV. This not explained by similar good performance, but by the fact that 247 out of 334 search requests (or 74.0%) actually failed to find any relevant documents at all.
The next sections will try to identify those search request aspects that explain the performance differences between Unique tags and CV (Section 6) and also identify the reasons why many of the search requests are so difficult to fulfill (Section 7).

6 Search Request Analysis

In the previous section, we saw that Unique tags and CVs offer complementary performance, meaning some search requests are better served by Unique tags and some better by CVs. In this section, we focus on different aspects in the search requests to identify the aspects that make Unique tags or CV better for search. In Section 6.1, we analyze which relevance aspect is in a search request is better served by which metadata element. Section 6.2 performs a similar analysis for book types (fiction or non-fiction). Finally, Section 6.3 looks at the combination of both factors.

Table 8: Comparison of experimental configurations, showing the number of book search requests (out of 87) that achieve better search results and also where the other combination achieved zero results.

Metadata element configurations	# of book search requests
Unique tags better than CV:	53
CV achieved zero results	47
CV achieved non-zero results	6
CV better than Unique tags:	27
Unique tags achieved zero results	16
Unique tags achieved non-zero results	11
Unique tags + CV better than Unique tags and CV:	24
CV and Unique tags achieved zero results	7
CV and Unique tags achieved non-zero results	17

6.1 Relevance Aspects in Book Search Requests

> **Question 7:** What types of book requests are best served by the Unique tags and CV test collections?
>
> **Answer:** Requests that focus on content-based, familiarity, known-item, or socio-cultural aspects are best served by Unique tags. Engagement requests seem to be better served by CVs , but not significantly so.

We define as relevance aspects those aspects named in a search request that make a book relevant for the searcher. The LT forum requests differ widely in their relevance aspects: some requesters try to re-find a book by giving vague plot points or character indications, others try to match a specific mood or reading experience.

Koolen et al. [24] annotated a large set of SBS book requests (which include our 334 test requests as a subset) with one or more of a set of eight relevance aspects[12] inspired by Reuter [41]. Table 9 contains brief descriptions of these eight relevance aspects.

For the 87 search requests, which found relevant book documents in either the Unique tags or CV collection, Table 9 shows (in the column 'Requests overall') that most search requests (79.3%) contain at least one Content aspect (e.g. a certain topic, plot or genre is needed). A Familiarity

[11] In the remainder of this article, we will use Unique tags as our collection representing tags, because they provide the best individual performance.

[12] Available at http://social-book-search.humanities.uva.nl/#/data/suggestion, last visited May 14, 2017.

Table 9: Distribution of the relevance aspects over all 87 successful book requests, where either Unique tags or CV found relevant documents (column 1), the requests where Unique tags outperform CV terms by 120% or more (column 2), and the requests where CV terms outperform Unique tags by 120% or more (column 3). More than one aspect can apply to a single book request, so numbers to not add up to 100%.

Relevance aspect	Description	Requests overall (*N = 87)*	Unique tags > CV (*N* = 53)	CV > Unique tags (*N* = 27)
Accessibility	Language, length, or level of difficulty of a book	9.2%	7.5%	11.1%
Content	Topic, plot, genre, style, or comprehensiveness	79.3%	83.0%	70.4%
Engagement	Fit a certain mood or interest, are considered high quality, or provide a certain reading experience	25.3%	22.6%	33.3%
Familiarity	Similar to known books or related to a previous experience	47.1%	49.1%	37.0%
Known-item	The user is trying to identify a known book, but cannot remember the metadata that would locate it	12.6%	17.0%	7.4%
Metadata	With a certain title or by a certain author or publisher, in a particular format, or certain year	23.0%	24.5%	14.8%
Novelty	Unusual or quirky, or containing novel content	3.4%	3.8%	0%
Socio-cultural	Related to the user's socio-cultural background or values; popular or obscure	13.8%	15.1%	7.4%

aspect is included in almost half of the requests, meaning that searchers look for books, which are similar or related to previous reading experiences.

The next two columns in Table 9 show the distribution of relevance aspects where either Unique tags or CV outperform the other metadata element. We can now see element-based differences. For example, an expected outcome was that fewer requests with a Socio-cultural aspect could be fulfilled by the CV collection on average, while Accessibility requests were apparently better served with this collection. Other results were much more surprising, for example that requests with an Engagement aspect could be better fulfilled by CV.

Figure 4 provides some clarity to these distribution numbers showing that aside from Engagement requests, all other requests achieved better search performance (as measured by NDCG@10) with the Unique tags collection than the CV collection.

However, a statistical significance test shows that the difference between Unique tags and CV is not significant except for Familiarity according to a two-tailed paired-samples *t*-test ($t(35) = 2.268$, $p < .05$, ES = 0.377, 95% CI [0.0119, 0.2147]) and for Content ($t(62) = 3.489$, $p < .005$, ES = 0.440, 95% CI [0.0489, 0.1800]).

The difference in performance is usually explainable by the contents of the relevant book documents for either metadata element. Relevant documents for the search request #63529 ("*I just finished and enjoyed Climb the Wind by Pamela Sargent. Can anyone recommend other science fiction and or alternate history about Native Americans?*") for example just contain the term "*Science fiction*" in the CVs , but also contain the terms "*alternate history*" and "*native americans*" in Unique tags, which are more helpful in fulfilling this Content aspect search request. Requests that contain a Known-item aspect are difficult to fulfill just using CV terms, because the requests usually mention only vague plot elements and characters, not the topics CVs would cover. A good example is request #73796 ("*I read this book 5 to 10 years ago. It was like Francine Rivers, but doesn't seem to match any of her titles that I can find. It started with 3 older men of a small church searching for a new pastor and hiring a young man who seemed promising. The new pastor had great success but as the church grew into a mega church with building projects, etc, he strayed*

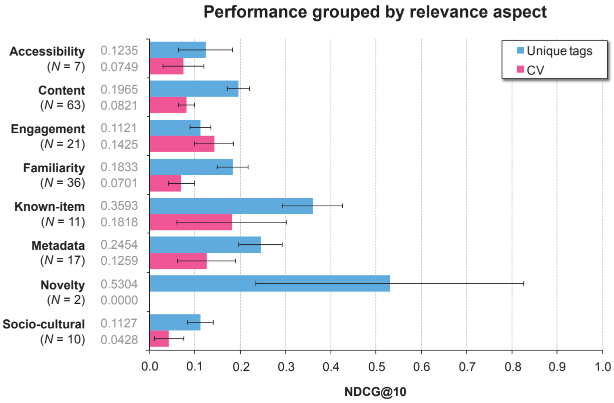


Figure 4: Results for the Unique tags and CV test collections, grouped by the eight relevance aspects expressed in the 87 successful book search requests, where either Unique tags or CV found relevant documents. Average NDCG@10 scores over all requests expressing a particular relevance aspect are shown in grey and as horizontal bars, with error bars in black.

away from the Word."), where relevant book documents only contained the generic CV terms "*Church buildings*" and "*Clergy*", but much more precise terms (such as "*church growth*", "*pastor*", "*mega churches*") in the Tags.

Aspects such as Socio-cultural and Novelty would not be expected to be covered by CV and they are indeed better searched using Unique tags. Accessibility and Content aspects in turn should be covered well by CV (and Accessibility requests occur more often in successful CV requests), but the Unique tags are usually still the better choice for covering such requests.

Engagement topics appear to be better served by CV than by Unique tags, but the difference is not significant ($t(20) = 0.767$, $p = .452$, ES = 0.167, 95% CI [-0.1132, 0.0524]). As a matter of fact, the relevant documents for requests with this aspect did not contain terms that were related to Engagement, so we consider the difference as coincidental.

6.2 Book Type: Fiction vs. Non-fiction

> **Question 8:** Does the type of book have an influence on performance for Unique tags or CV?
>
> **Answer:** Unique tags work better for fiction. CV work better for non-fiction requests, but the difference is not significant. Finding non-fiction books appears to be easier than fulfilling requests for fiction books.

The complementarity analysis showed that Tags and CV elements succeed on different groups of requests. What could explain this difference? The type of book requested—fiction or non-fiction—is one possibility. All 334 test requests were annotated whether they were requesting works of fiction or non-fiction. The first 100 topics were annotated by both authors, with resulting differences in only 5% of the cases, these were resolved through discussion. Because of the high agreement, the remaining topics were annotated by one author. The majority of requests (75.3%) were for works of fiction.

Figure 5 provides on overview of the average performance for Unique tags and CVs grouped by type of book. While Unique tags achieve a significantly better search performance for fiction books (t(58) = 3.571, p < .005, ES = 0.465, 95% CI [0.0568, 0.2016]), CVs appear to be better for non-fiction books, but this difference is not statistically significant (t(27) = 1,194, p = .243, ES = 0.226, 95% CI [-0.1699, 0.0449]).

These results are not surprising as we can expect Unique tags to contain the more contextualized, character, plot and personal information that would be needed to fulfill fiction book requests rather than CVs who may be more successful in objectively determining the topics of non-fiction books.

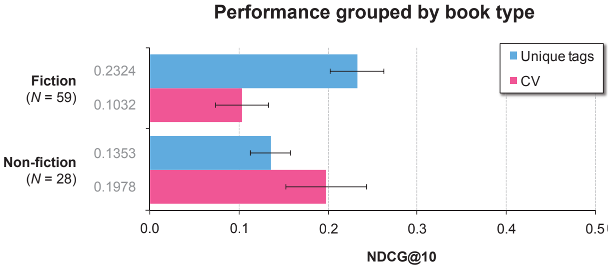


Figure 5: Results for the Unique tags and CV test collection, grouped by type of book(s) requested (fiction or non-fiction). Average NDCG@10 scores over all requests for a particular book type are shown in grey and as horizontal bars, with error bars in black.

6.3 Book Type and Relevance Aspects

> **Question 9:** Does the relevance aspect of a request have an influence on the performance by book type?
>
> **Answer:** No. Relevance aspects that are better addressed by Unique tags have a stronger presence in non-fiction books, which should be better addressed by CV on average. Other factors must play a role.

One possible explanation for the better search performance of Unique tags for fiction books might be that the relevance aspects asked for in book search requests for those books are better solved by Unique tags. Figure 6 shows the distribution of relevance aspects by the type of books requested. As expected, some of the aspects that Unique tags can cover better in search (see Figure 4) such as Familiarity also occur more often in fiction book search requests. Other results are more surprising. The Content, Engagement, Accessibility, Novelty and Socio-cultural aspect are more common in non-fiction requests. It appears that when requesters search for non-fiction, they are looking for particular style, degree of comprehensiveness, language or engagement level, that is, for a non-fiction book that is to their taste not only about the topic they are looking for.

Figure 4 showed surprisingly that Metadata requests are better served by Unique tags than by CV. Because they appear more often in fiction requests and Unique tags is generally better in serving fiction requests (Figure 5), this may be an explanation for this result. However, since this is a comparison just between Unique tags and CV metadata elements, there is also another explanation. Traditional CV terms do not contain bibliographical metadata, which is included in other metadata elements, while Unique tags terms may include them because no rules are applied. This may not be an argument for Unique tags, but an argument against analyzing these metadata elements in isolation without standard bibliographic metadata, which would always be included in any kind of information system.

Except for Engagement, all other aspects are better addressed by Unique tags on average, while the book type non-fiction is better addressed by CV on average (albeit not significantly so). It could be that the strong performance of CV in Engagement requests offsets the other results. However, another explanation would be that it is still another factor that impacts these results.

These analyses have shown that we can identify differences in performances based on particular search request aspects. However, more aspects could be looked at. The next section will look at the question why many search requests fail completely.

7 Failure Analysis

The absolute majority (74.0% of 334) of our tested book search requests fail when searching with Unique tags or CV. By combining the collections in Unique tags + CV, we find relevant books for 7 more requests, but 240 requests remain unsolved. This section tries to identify the causes for these failures. Sections 7.1 and 7.2 analyze, whether some relevance aspects or fiction or non-fiction books are harder to search for. Finally, Section 7.3 also studies whether data sparsity in requests or the book documents can explain search failure.

7.1 Relevance Aspects

Question 10: Do book search requests fail because their relevance aspects are difficult to search?

Answer: Accessibility and Known-item related search requests seem to fail more often, while Familiarity and Content related search requests seem to succeed more readily.

In Section 6.1, Figure 4, we can see that some relevance aspects—no matter what collection they are searched in—achieve better performance as expressed by higher NDCG@10 scores than other relevance aspects. For example, Known-item and Metadata requests achieve higher NDCG@10 scores on average than Accessibility and Socio-cultural requests. This first analysis determines whether some relevance aspects are inherently more difficult to search.

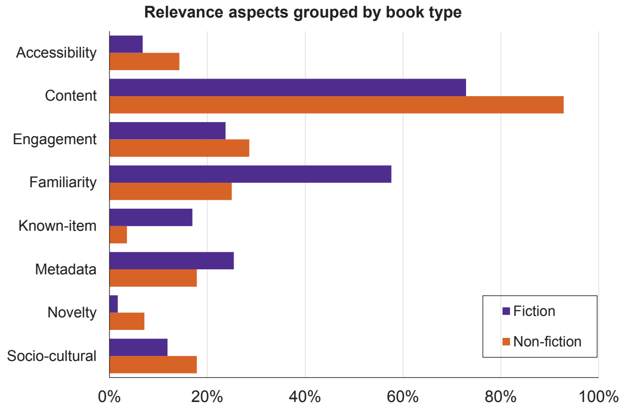


Figure 6: Distribution of relevance aspects by the type of book(s) requested (fiction (N = 59) vs. non-fiction (N = 28). Horizontal bars represent the percentage of all request of a particular book types that express a specific aspect. For example, 17.0% of all 59 fiction requests express a Known-item aspect.

Table 10 and Figure 7 show the distribution of relevance aspects over the successful and failed requests. Compare this also with Table 9, which shows which relevance aspects are better served by Unique tags or CV.

The Accessibility and Known-item related search requests occur more frequently in failed searches, while the Familiarity and Content related search requests seem to occur more frequently in successful searches. All these aspects are better addressed by Unique tags, so it may not necessarily be the collection that addresses these relevance aspects more successfully.

Surprisingly, the distribution of relevance aspects over successful and failed requests does not correlate with the performance numbers for successful requests as seen in Figure 4. Only because Known-item requests achieve a higher NDCG@10, they are not more successful. However, Metadata requests are indeed slightly more successful. Other factors may play a role why these successful requests achieve a higher performance.

7.2 Book Type

Question 11: Do book search requests fail because their relevance aspects are difficult to search?

Answer: Accessibility- and Known-item-related search requests seem to fail more often, while Familiarity- and Content-related search requests seem to succeed more readily.

Table 10: Tabular distribution of the relevance aspects over all 94 successful and 240 failed requests.

Relevance aspect	Successful (N = 94)	Failed (N = 240)
Accessibility	9.6%	15.4%
Content	79.8%	75.8%
Engagement	24.5%	21.3%
Familiarity	45.7%	35.8%
Known-item	13.8%	19.6%
Metadata	24.5%	22.5%
Novelty	4.3%	3.3%
Socio-cultural	14.9%	12.5%

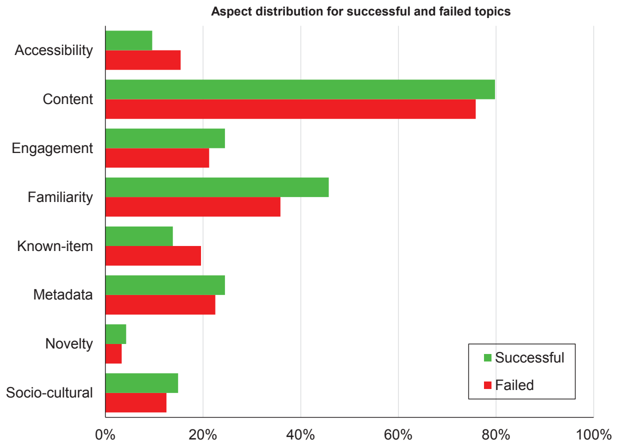


Figure 7: Visual distribution of the relevance aspects over all 94 successful and 240 failed requests.

Out of 240 failed requests, most (193 or 80.4%) were for fiction books. In contrast, only 63 (or 67.0%) fiction book requests were successful (out of 94). A statistical significance analysis showed that according to a Chi-square test ($X^2(1)$ = 6.771, p < .01), this difference is significant, meaning that fiction books are indeed harder to find.

7.3 Sparsity, Recall Base, and Example Books

Another possible explanation for the many failed search requests is data sparsity. Data sparsity could occur in book metadata or book search requests—if either is too short to contain adequate information for searching, then a search will be unsuccessful. A book search will also be unsuccessful, if the collection does not contain any or an inadequate number of relevant books for the request. This is called a low recall base. This section looks at these potential failure aspects.

> **Question 12:** Do book search requests fail because of data sparsity, a lower recall base, or a lack of examples?
>
> **Answer:** Sparsity does not appear to be a reason for search failure and neither is the size of the recall base. The number of examples provided by the requester does have a significant positive influence on performance.

In Table 11, the distribution numbers for the average lengths of book search requests and the relevant book documents in the Unique tags + CV collection shows the exact reverse relationship than expected. Both search requests and relevant documents are actually longer for failed searches than for successful ones. Both differences are statistically significant according to independent-samples t-tests for search request lengths (t(332) = 0.907, p = .365, 95% CI [9.915, 10.933]) and for relevant book document lengths (t(3889) = 6.257, p < .001, 95% CI [-5.580, 0.892]).

It appears, we may have to postulate the inverse assumption: a search may be unsuccessful, because there is too much information in either requests or documents and the search system is not able to adequately distinguish between important and unimportant terms for the search.

Table 11 also compares the available recall base for successful and unsuccessful search requests. There is no statistically significant difference between the average number of relevant books for either type of search request (t(332) = 1.269, p = .205, 95% CI [-2.301, 1.812]). An insufficient recall base does not explain the failed requests either.

A last analysis looks at a characteristic specific to LT forum requests, the number of example books that were provided by the requester in their original request. These are included in the requests sent to the system with the assumption that these might help the search. This assumption proves to be correct: more examples are provided for successful requests than unsuccessful requests (see Table 11) (t(332) = 4.638, p < .001, 95% CI [-1.098, 0.237]). A correlation analysis also showed a weak positive correlation (r = 0.175, p < .005) between the average NDCG@10 score for Unique tags + CVs and the number of provided examples, which means that the performance is better, the more examples are provided in the request.

8 Discussion & Conclusions

This paper presented a large-scale empirical analysis of different metadata elements for book search, with an emphasis on the comparison of the search effectiveness of Tags and CVs in this context. The analysis was carefully

Table 11: Breakdown of book search requests by request length, length of the relevant documents, size of the recall base, and the number of examples provided by the original requester.

	Avg. book search request length (in words)	Avg. relevant document length (in words)	Avg. no. of relevant documents	Avg. no of example books provided
Successful (*N* = 94)	86.7	73.9	13.3	1.63
Successful (*N* = 94)	96.6	79.5	11.0	0.54
Overall (*N* = 334)	93.8	77.7	11.7	0.84

controlled for a fair representation of Tags and CVs: at least one term from each metadata element needed to be present in a book document. The types statistic (Table 4) actually revealed that book documents contained more CV terms on average than Tags, which means that the more frequent CVs terms should have had a higher chance to successfully fulfill search requests. The opposite turned out to be true. This in-depth analysis of a large book record collection using 334 real-world search requests had the following results:

- When comparing Core bibliographic metadata to metadata elements such as CV, Tags or Reviews, Reviews will achieve the best search performance just using the individual metadata elements.
- Adding Core bibliographic metadata to another metadata element will improve the search results as expected.
- Combining all metadata elements will achieve the best performance.
- However, certain combinations may not outperform individual elements. Surprisingly, adding CV to Core bibliographic metadata will not significantly improve the results.
- When comparing just CVs with Tags, Tags will outperform CVs.
- Tags do not outperform CVs because of the popularity effect. The Unique tags collection, which containsless tags on average than the CV collection and only contains unique tags, will still outperform the CV collection.
- While not achieving a significant performance improvement, adding CV to Tags will improve theperformance, demonstrating a slight complementarity effect.
- Combining Tags and CVs may even find relevant documents for requests, which would fail for eitherindividual metadata collection.
- When grouping search requests by relevance aspect, Unique tags perform better on almost all of themexcept Engagement. Unique tags also perform better on aspects, which would have been in the domainof CVs such as Accessibility or Content.
- Unique tags seem to work better for fiction book requests, CVs seem to work better for non-fiction book requests.
- Even when people search for non-fiction books, more often than not they are not only interested in the content of the book, but also whether it fits their mood, reading style or books that they already liked.
- Search requests containing Accessibility or Known-Item relevance aspects are particularly hard to fulfill. Both aspects may be too vaguely expressed in the request to be able to distinguish relevant from non-relevant books.
- Searches for fiction books fail more often than searches for non-fiction books.
- In contrast to popular belief, longer requests or longer documents may not necessarily lead to more successful searches.
- Having more relevant books available in the book collection that is searched will also not necessarily improve the search performance.
- Mentioning a potentially relevant book in the search request, however, will improve the search performance.

What do we conclude from these results? A particularly compelling conclusion comes from the fact that even though the Tags collection contained fewer terms than the CV collection, a better search performance means that the tag vocabulary is richer and more attuned to the vocabulary used in the requests. We can hypothesize that this vocabulary phenomenon extends to the more general information need: Unique tags will better address these complex information needs than Tags.

It is still premature to give up on CVs. Circa a third of the successful requests found relevant documents only in to the CV element set, pointing towards a complementarity effect. Some requests are looking for information that a social tagging application cannot provide. The analysis of relevance aspects and book types in the search requests found some differences for these two metadata elements, but the predictiveness of these factors appears to be not very conclusive.

Most search requests fail when using Tags or CVs. Even when a search is successful, the performance is disappointingly low (in terms of NDCG@10), which indicates that neither metadata element may be adequate for the complex information needs that we studied here. The nature or quality of the terms in the Tags or CV collections may simply be inadequate in addressing the search requests.

The results from this study posit a number of new avenues for future work. One is to analyze other factors than the ones already included here: more combinations of metadata elements and search request representations, possibly developing a predictive model whether a particular search request can be fulfilled by existing metadata or would need to be restated.

Another avenue is to think about the nature of book metadata themselves. More plot and genre details, character and place names, mood or engagement categorizations as well as relationships to other books would equip book metadata to fulfill more of these complex information needs. Of course, these are subjective aspects, some of which cannot even be harvested from automatic full-text extraction.

Most interestingly, the issues and challenges we have discussed here and as potentials for future work will not all be solved by simply supplying the full-text of the book. Even in a fully digitized world, book search remains a research-worthy challenge.

References

[1] Aitchison, J.,Gilchrist, A. (1987). *Thesaurus Construction: A Practical Manual* (2nd ed.). ASLIB, London, 173.

[2] Bartley, P. (2009). Book Tagging on LibraryThing: How, Why, and What are in the Tags? *Proceedings of the American Society forInformation Science and Technology* 46, 1, 1–22.

[3] Beckers, T., Fuhr, N., Pharo, N., Nordlie,R., and Fachry,K. N. (2010). Overview and Results of the INEX 2009Interactive Track. In *Research and Advanced Technology for Digital Libraries*, M. Lalmas, J. Jose, A. Rauber, F. Sebastiani, and I. Frommholz (Eds.). Lecture Notes in Computer Science, Vol. 6273. Springer, 409–412.

[4] Bischoff, K., Firan, C.S., Nejdl, W. and Paiu, R. (2008). Can All Tags be Used for Search? In *CIKM '08: Proceedings of the 17th ACM conference on Information and Knowledge Management*. ACM, 193–202.

[5] Bogers, T. and Petras, V. (2015). Tagging vs. Controlled Vocabulary: Which is More Helpful for Book Search? In *Proceedings of iConference 2015*.

[6] Bogers, T. and Petras, V. (2017). An In-Depth Analysis of Tags and Controlled Metadata for Book Search. In *Proceedings of iConference 2017*.

[7] Brooks, T. A. (1993). All the Right Descriptors – A Test of the Strategy of Unlimited Aliasing. *Journal of the American Society for Information Science* 44(3), 137–147.

[8] Buchanan, G. and McKa, D. (2011). In the Bookshop: Examining Popular Search Strategies. In *JCDL '11: Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries*. ACM, New York, NY, USA, 269–278.

[9] Buckley, C. (2009). Why Current IR Engines Fail. *Information Retrieval* 12(6), 652–665.

[10] Carmel, D. and Yom-Tov, E. (2010). Estimating the Query Difficulty for Information Retrieval. *Synthesis Lectures on Information Concepts, Retrieval, and Services* 2(1), 1–89.

[11] Choi,Y., Hsieh-Yee, I. and Kules, B. (2007). Retrieval Effectiveness of Table of Contents and Subject Headings. In *JCDL '07: Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries*. ACM, New York, NY, USA, 103–104.

[12] Cleverdon, C. W. and Mills, J. (1963). The Testing of Index Language Devices. In *Aslib Proceedings*, Vol. 15. 106–130.

[13] Dextre Clarke, S. (2008). The Last 50 Years of Knowledge Organization: A Journey through my Personal Archives. *Journal of Information Science* 34(4), 427–437.

[14] Dextre Clarke, S. and Vernau, J. (2016). The Thesaurus Debate Continues. *Knowledge Organization* 43(3), 135–137.

[15] Golub, K.. Moon, J., Tudhope, D. Jones, C., Matthews, B., Puzoń, B.,and Lykke Nielsen, M. (2009). EnTag: Enhancing Social Tagging for Discovery. In *JCDL '09: Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Librarie*s. ACM, New York, NY, USA, 163–172.

[16] Gross, T. and Taylor, A.G. (2005). What Have We Got to Lose? The Effect of Controlled Vocabulary on Keyword Searching Results. *College & Research Libraries* 66(3), 212–30.

[17] Heymann, P. and Garcia-Molina, H. (2009). Contrasting Controlled Vocabulary and Tagging: Do Experts Choose the Right Names to Label the Wrong Things? In *WSDM '09: Proceedings of the 2nd ACM International Conference on Web Search and Data Mining (Late Breaking Results Session)*. Stanford InfoLab, 1–4.

[18] Heymann, P., Koutrika, G. and Garcia-Molina, H. (2008). Can Social Bookmarking Improve Web Search?. In *WSDM '08: Proceedings of the 2008 International Conference on Web Search and Data Mining*. ACM, 195–206.

[19] Järvelin, K. and Kekäläinen, J. (2002). Cumulated Gain-based Evaluation of IR Techniques. *ACM Transactions on Information Systems* 20(4), 422–446.

[20] Kazai, G., Koolen, M., Kamps, J., Doucet, A. and Landoni, M. (2011). Overview of the INEX 2011 Book and Social Search Track. In *INEX 2011 Workshop pre-proceedings (INEX Working Notes Series)*. 11–36.

[21] Kim, J.Y., Feild,H. and Cartright, M.(2012). Understanding Book Search Behavior on the Web. In *CIKM '12: Proceedings of the 21st ACM International Conference on Information and Knowledge Management*. ACM, New York, NY, USA, 744–753.

[22] Koolen, M. (2014). User Reviews in the Search Index? That'll Never Work!. In *ECIR '14: Proceedings of the 36th European Conference on Information Retrieval*. 323–334.

[23] Koolen, M., Bogers, T., Gäde,M., Hall, M., Hendrickx, I., Huurdeman, H., Kamps, J., Skov, M., Verberne, S. and Walsh, D. (2016). Overview of the CLEF 2016 Social Book Search Lab. In *CLEF 2016: Proceedings of the 7th International Conference of the CLEF Association*, Vol. LNCS 9822. 351–370.

[24] Koolen, M., Bogers, T., van den Bosch, A. and Kamps, J. (2015). Looking for Books in Social Media: An Analysis of Complex Search Requests. *ECIR '15: Proceedings of the 37th European Conference on Information Retrieval.* Springer, 184–196.

[25] Koolen, M., Kamps, J. and Kazai, G. (2012). Social Book Search: Comparing Topical Relevance Judgements and Book Suggestions for Evaluation. In *CIKM '12: Proceedings of the 21st International Conference on Information and Knowledge Management*. 185–194.

[26] Koolen, M., Kazai, G., Kamps, J., Doucet, A. and Landoni, M. (2012). Overview of the INEX 2011 Books and Social Search Track. In *Focused Retrieval of Content and Structure*. Springer, 1–29.

[27] Koolen, M., Kazai, M., Preminger, M. and Doucet, A. (2013). Overview of the INEX 2013 Social Book Search Track. In *CLEF 2013: Proceedings of the Fourth International Conference of the Cross-Language Evaluation Forum*, Pamela Forner, Jussi Karlgren, and Christa Womser-Hacker (Eds.). 1–26.

[28] Koolen, M., Kazai, G., Preminger, M., Kamps, J., Doucet, A. and Landoni, M. (2012). Overview of the INEX 2012 Social Book Search Track. In *CLEF 2012: Proceedings of the Third International Conference of the Cross-Language Evaluation Forum*, Pamela Forner, Jussi Karlgren, and Christa Womser-Hacker (Eds.). 1–20.

[29] Lancaster, F. W., Connell, T. H., Bishop,N. and Mccowan, S. (1991). Identifying Barriers to Effective Subject Access in Library Catalogs. *Library Resources and Technical Services* 35(4), 377–391.

[30] Lawson, K.G. (2009). Mining Social Tagging Data for Enhanced Subject Access for Readers and Researchers. *The Journal of Academic Librarianship* 35(6), 574–582.

[31] Lee, D. H. and Schleyer, T. (2010). A Comparison of MeSH Terms and CiteULike Social Tags As Metadata for the Same Items. In *IHI '10: Proceedings of the 1st ACM International Health Informatics Symposium*. ACM, New York, NY, USA, 445–448.

[32] Liu, Y-H. (2010). On the Potential Search Effectiveness of MeSH (Medical Subject Headings) Terms. In *IIiX '10: Proceedings of the Third Symposium on Information Interaction in Context*. ACM, New York, NY, USA, 225–234.

[33] LoC Cataloging Policy and Support Office. 2016. Assigning and Constructing Subject Headings H180. In *Library of Congress Subject Heading Manual*. https://www.loc.gov/aba/publications/FreeSHM/H0180.pdf

[34] Lu, C., Park, J.-R. and Hu, X. (2010). User Tags versus Expert-assigned Subject Terms: A Comparison of LibraryThing Tags and Library of Congress Subject Headings. *Journal of Information Science* 36(6), 763–779.

[35] Magdy, W. and Darwish, K. (2008). Book Search: Indexing the Valuable Parts. In *BooksOnline '08: Proceedings of the 2008 ACM Workshop on Research Advances in Large Digital Book Repositories*. ACM, New York, NY, USA, 53–56.

[36] Mikkonen, A. and Vakkari, P. (2016). Readers' Interest Criteria in Fiction Book Search in Library Catalogs. *Journal of Documentation* 72(4), 696–715.

[37] Nolland, M.G. and Meinel, C. (2007). Authors vs. Readers: A Comparative Study of Document Metadata and Content in the WWW. In *DocEng '07: Proceedings of the 2007 ACM Symposium on Document Engineering*. ACM, New York, NY, USA, 177–186.

[38] Pass, G., Chowdhury, A. and Torgeson, C. (2006). A Picture of Search. In *InfoScale '06: The First International Conference on Scalable Information Systems.*

[39] Qin, J. (2008). Folksonomies and Taxonomies: Where the Two can Meet. *New Dimensions in Knowledge Organization Systems* 11.

[40] Rajashekar, T. B. and Croft, B. W. (1995). Combining Automatic and Manual Index Representations in Probabilistic Retrieval. *Journal of the American Society for Information Science* 46(4), 272–283.

[41] Reuter, K. (2007). Assessing Aesthetic Relevance: Children's Book Selection in a Digital Library. *Journal of the American Society for Information Science and Technology* 58(12), 1745–1763.

[42] Rowley, J. E. (1994). The Controlled versus Natural Indexing Languages Debate Revisited: A Perspective on Information Retrieval Practice and Research. *Journal of Information Science* 20(2), 108–19.

[43] Saarinen, K. and Vakkari, P. (2013). A Sign of a Good Book: Readers' Methods of Accessing Fiction in the Public Library. *Journal of Documentation* 69(5), 736–754.

[44] Sakai, T. (2014). Statistical Reform in Information Retrieval? *SIGIR Forum* 48(1), 3–12.

[45] Savoy, J. and Abdou, S. (2008). Searching in MEDLINE: Query Expansion and Manual Indexing Evaluation. *Information Processing & Management* 44(2), 781–789.

[46] Seki, K., Qin, H. and Uehara, K. (2010). Impact and Prospect of Social Bookmarks for Bibliographic Information Retrieval. In *JCDL '10: Proceedings of the 10th Annual Joint Conference on Digital Libraries*. ACM, New York, NY, USA, 357–360.

[47] Slone, D. J. (2000). Encounters with the OPAC: Online Searching in Public Libraries. *Journal of the American Society for Information Science* 51(8), 757–773.

[48] Smith, T. (2007). Cataloging and You: Measuring the Efficacy of a Folksonomy for Subject Analysis. In *Proceedings of the 18th Workshop of the ASIST Special Interest Group in Classification Research*, Joan Lussky (Ed.). http://dlist.sir.arizona.edu/2061/

[49] Spiteri, L.F. (2007). The Structure and Form of Folksonomy Tags: The Road to the Public Library Catalog *Information Technology and Libraries* 26(3), 13–25.

[50] Voorbij, H. (2012). The Value of LibraryThing Tags for Academic Libraries. *Online Information Review* 36(2), 196–217.

[51] Willis, C. and Efron, M. (2013). Finding Information in Books: Characteristics of Full-text Searches in a Collection of 10 Million Books. In *ASIST '13: Proceedings of the 76th ASIS&T Annual Meeting. American Society for Information Science*, Silver Springs, MD, USA, Article 84, 84:1–84:10 pages.

[52] Zhai, C. and Lafferty, J. (2004). A Study of Smoothing Methods for Language Models Applied to Information Retrieval. *ACM Transactions on Information Systems* 22(2) 179–214.