Tagging vs. Controlled Vocabulary: Which is More Helpful for Book Search?

Toine Bogers[1] and Vivien Petras[2]
[1]Department of Communication, Aalborg University Copenhagen, Copenhagen, Denmark
[2]Humboldt-Universität zu Berlin, Germany

Abstract
The popularity of social tagging has sparked a great deal of debate on whether tags could replace or improve upon professionally assigned metadata as descriptors of books and other information objects. In this paper we present a large-scale empirical comparison of the contributions of individual metadata elements like core bibliographic data, controlled vocabulary terms, reviews, and tags to the retrieval performance. Our comparison is done using a test collection of over 2 million book records with metadata elements from Amazon, the British Library, the Library of Congress, and LibraryThing. We find that tags and controlled vocabulary terms do not actually outperform each other consistently, but seem to provide complementary contributions: some information needs are best addressed using controlled vocabulary terms whereas other are best addressed using tags.

Keywords: book search; indexing; controlled vocabularies; social tagging
Citation: Editor will add citation.

Contact: toine@hum.aau.dk[1], vivien.petras@ibi.hu-berlin.de[2].

1 Introduction

Ever since the Cranfield experiments (Cleverdon & Mills, 1963), researchers have argued about Controlled vocabularies (CVs) for information retrieval. Even with the realization of full-text retrieval, the discussion continued with advances in text processing as well as semantic applications making either alternative better. User-generated content (UGC)—in the form of tags or reviews—has added another dimension to this discussion. Tags in particular have generated discussion whether they improve retrieval because of scale, or whether the vocabulary problem inherent to tagging means their potential will never be realized.

Most of the work comparing tags to CVs for book search has remained theoretical. Few exploratory studies have focused on the potential of these metadata elements for retrieval. The only notable exception is a large-scale empirical comparison by (Koolen, 2014), who found that UGC, in particular reviews, outperformed professionally assigned metadata. In this paper we delve deeper into this problem: which (combination of) metadata elements can best contribute to retrieval success, and how does the retrieval performance of tags and CVs compare under carefully controlled circumstances?

We present an empirical comparison in the book search domain using LibraryThing (LT), Amazon, the Library of Congress (LoC), and the British Library (BL) as data providers. The study uses a large-scale collection from the INEX Social Book Search Track, filtered to allow a fair comparison between tags and CVs. A substantial set of requests representing real information needs is used. The analysis focuses on the differences in using tags or CVs overall and distinguished by different book types or request types. Our contributions are:

- Empirical evidence on the contributions of different metadata element sets for book search based on a large-scale test collection.
- Analysis of the contributions of tags and CVs for book search.
- Insights on impact factors and suggestions for future work on improving book search.

The remainder of this paper is organized as follows. Section 2 presents research on natural language and CV searching, social tagging and book retrieval using UGC. Section 3 explains the methodology for experiments in this study. Section 4 describes the results, while Section 5 analyses the results for their impact of tags and CVs on book search respectively. The final section discusses the outcomes of this study and concludes with suggestions for future work.

2 Related Work

CVs such as subject headings, thesauri or classifications provide language-controlled keywords for describing a document's content. They are contrasted to the natural language in abstracts or the full-text of a document and later in UGC such as tags. UGC is defined here as user-provided natural language (keywords or text) for content descriptions.

2.1 Natural Language vs. Controlled Vocabularies

The arguments for natural or controlled language indexing have been enumerated often (Aitchison & Gilchrist, 1987; Clarke, 2008). Advantages of controlled indexing are synonym and homonym control and the expression of semantic relationships between concepts. The advantages of natural languages are the use of the users' vocabulary and the avoidance indexing errors. CVs have large development costs and often use outdated vocabulary. Natural language can lead to a loss in precision or recall because of vagueness.

2.2 Searching with Natural Language or Controlled Vocabularies

While many experiments showed early that natural language performs as well as CVs for searching (Rowley, 1994), others claimed that natural language can lead to a performance drop (Lancaster, Connell, Bishop, & Mccowan, 1991; Brooks, 1993). Notably, the Cranfield experiments showed that individual natural language terms performed best, but controlled indexing was better than full-text (Cleverdon & Mills, 1963). Several studies found that CVs and natural language complement each other (Rajashekar & Croft, 1995; Gross & Taylor, 2005; Savoy & Abdou, 2008), others find users are better served with the natural language (Choi, Hsieh-Yee, & Kules, 2007; Liu, 2010).

2.3 Social Tagging vs. Controlled Vocabularies

Social tagging systems have been criticized for the same lack of vocabulary control as natural language, even though the possible inclusion of more user-adequate vocabulary has been noted (Spiteri, 2007; Qin, 2008). Golub et al. (2009) found that most users preferred tagging to choosing from a CV when describing content and that more searches were successful when using the tags. This was also demonstrated for tag-assisted web search (Heymann, Koutrika, & Garcia-Molina, 2008; Bischoff, Firan, Nejdl, & Paiu, 2008), however a significant number of tags were also included in other metadata elements such as the title making the tags possibly unnecessary. Lee and Schleyer (2010) demonstrated that MeSH headings and CiteULike tags do not overlap. Seki, Qin, and Uehara (2010) then showed that both performed similarly when searched separately, but the performance increased significantly when combined. This indicates that tags and CV terms could be complementary in retrieval.

2.4 Searching for Books: Tags vs. Controlled Vocabularies

Magdy and Darwish (2008) demonstrated in just using the titles and chapter headings, a search (experimenting on short queries from the 2007 INEX Book Search Track) was almost as successful as using the full-text of a book. This corroborates the idea that certain metadata elements are more important in retrieval than others.

Several studies exploring LibraryThing tags and CVs (e.g. LCSH). found that the terms didn't overlap and that tags had a broader coverage including personal and contextual information while subject headings covered more abstract concepts (Smith, 2007; Bartley, 2009). These small-scale studies indicated that tags and CV complement each other for book search, however, two larger studies found that the tags provide a much richer vocabulary for searching (Heymann & Garcia-Molina, 2009; Lu, Park, & Hu, 2010).

The INEX Social Book Search Track (Koolen, Kazai, Kamps, Doucet, & Landoni, 2012; Koolen, Kazai, Preminger, et al., 2012; Koolen, Kazai, Preminger, & Doucet, 2013) evaluates book retrieval in Amazon, LibraryThing and libraries. Koolen, Kamps, and Kazai (2012) analyzed 24 book search queries in the Amazon/LibraryThing corpus and compared relevance judgments with the book suggestions from LT members. They found that for judging topical relevance, the reviews were more important than the core bibliographical elements or the tags.

Koolen (2014) is the closest study in comparison to this paper. He finds that the reviews in LT added most to the retrieval success compared to core bibliographic metadata or CV. While his study uses the complete Amazon/LibraryThing collection, this paper uses a subset, where each document contains CV as well as tags so that each has the same chance to contribute to the retrieval success.

3 Methodology

To study which metadata elements contribute most to retrieval performance in book search, a document collection containing both 'traditional' library metadata as well as UGC such as tags and reviews 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 (A/LT) collection meets these requirements. Sections 3.1-3.3 introduce it in more detail. Section 3.4 describes our experimental setup and evaluation protocol.

3.1 The Amazon/LibraryThing Collection

The A/LT collection was adopted as the test collection for the INEX Social Book Search Track[1]. It contains book records in XML format for over 2.8 million books. These records are aggregated from four providers (Koolen et al., 2013): Amazon, the British Library (BL), the Library of Congress (LoC), and LibraryThing (LT). Each contributes different metadata elements to the collection. Core bibliographic metadata such as author and title are provided by Amazon, which also contributes Dewey Decimal Classification (DDC) class numbers, Amazon subject headings, category labels from Amazon's category system, and user reviews. BL and LoC contribute CV (DDC and Library of Congress Subject Headings (LCSH)) to 1.15 million records and 1.25 million records respectively. Finally, LT contributes all tags added to the books in the A/LT collection. Matching the book records from the different providers was done based on their ISBNs. The book records (henceforth referred to as 'documents') in the A/LT collection contain over 40 different metadata elements, but not all of them are likely to contribute to effective retrieval, such as the number of pages. After removing those, we are left with the metadata elements shown in Table 1.

Table 1: Overview of the A/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, Amazon geographic names, Amazon category labels	Reviews
BL		DDC class labels, LCSH topical terms, geographic names, personal names, chronological terms, genre/form terms	
LoC		DDC class labels, LCSH topical terms, geographic names, personal names, chronological terms, genre/form terms	
LT			Tags

3.2 Filtering

When comparing different metadata elements—tags and CV in particular—it is important to make the comparison as fair as possible. The popularity effect in tagging systems (Noll & Meinel, 2007) causes popular books to receive more (and more of the same) tags than unpopular books, whereas CV terms are more evenly distributed across all books.

To ensure a fair comparison between tags and CV, we filtered out all books from the collection that did not contain *at least one* CV term and *at least one* tag. This addresses one aspect of the popularity effect and ensures that the distribution of element content over the document collection is less likely to be the cause of differences

[1] See https://inex.mmci.uni-saarland.de/tracks/books/ for more information, last accessed September 5, 2014.

in retrieval performance. However, since the *social* aspect of tagging—multiple annotators tagging the same object—is fundamental to its success, we did not reduce the tags assigned to a book to a set of unique tags. Instead, we treat the textual content of each metadata element as a bag-of-words representation, as opposed to a set of unique word types. The filtering process resulted in the Any-CV collection, containing 2,060,758 documents. Unless stated otherwise, this Any-CV collection will be used for all experiments reported in the remainder of this paper.

The effect of CVs on retrieval performance is not only contingent on their presence in the document representation: there may also be differences in quality. For instance, it is possible that professional metadata from BL or LoC is of better quality than that provided by Amazon. To examine this question, we performed an even more restrictive filtering. The resulting Each-CV collection contains documents that include at least one CV term *from each individual provider* and at least one tag from LT. This more restrictive filtering criterion reduces the number of documents searchable in the Each-CV collection to 353,670. This collection will help us determine which of the three data providers provides the CV metadata which contributes most to the retrieval success.

Table 2: Type and token statistics for the different element sets in the Any-CV and Each-CV collections.

	Element set	#types	#tokens	avg. types/doc	avg. tokens/doc
Any-CV	Core	26,533,832	27,541,834	12.9	13.4
	Controlled vocabulary	75,268,209	105,929,251	36.5	51.4
	Review	553,943,057	2,085,063,187	505.4	1902.4
	Tag	36,592,978	244,681,548	17.8	118.8
	User-generated content	590,536,035	2,329,744,735	286.6	1130.7
	All elements	1,282,874,111	4,792,960,555	622.5	2325.9
Each-CV	Controlled vocabulary (Amazon)	11,423,142	17,190,997	32.3	48.6
	Controlled vocabulary (BL)	3,541,891	4,502,813	10.0	12.7
	Controlled vocabulary (LoC)	3,886,196	5,009,686	11.0	14.2
	Controlled vocabulary (All)	18,851,229	26,703,496	53.3	75.5

Table 2 shows type and token counts for the different element sets in the Any-CV and Each-CV collections, both as total counts and averages per document. It shows that there is a partial popularity effect for tags, as the average number of tokens per document is much higher than the average number of types, at 118.8 vs. 17.8. Interestingly, CV elements have a higher average number of types per document at 36.5, but an expected lower average number of tokens at 51.4. The stricter filtering gives CV elements as fair a playing field as possible. Reviews are unsurprisingly the richest metadata element in textual content.

3.3 Book Requests & Relevance Judgments

The A/LT test collection provides a varied set of topics representing actual book-related information needs, along with relevance judgments (Koolen, Kamps, & Kazai, 2012). The topics are harvested from the LT discussion forums where members ask for book recommendations and others provide suggestions. Examples 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. Very often, these requests are accompanied by books that the requesters have already read and (dis)liked. Figure 1 shows an example book request[2].

Topics harvested from LT have different representations, such as the Title and the Narrative of the request—the text in the requester's post. For IR purposes, the information needs were annotated (Koolen et al., 2013) to add a Query representation. We will use the Query and Narrative representations. While the queries are succinct expressions of the information need, the requester-provided narrative is usually longer and explains more about the request's context. They also include books that are mentioned by the requester. For assessing the relevance of documents, the book suggestions in reply to the LT forum request were harvested. 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

[2]Topic 99309, available at http://www.librarything.com/topic/99309, last accessed September 5, 2014.

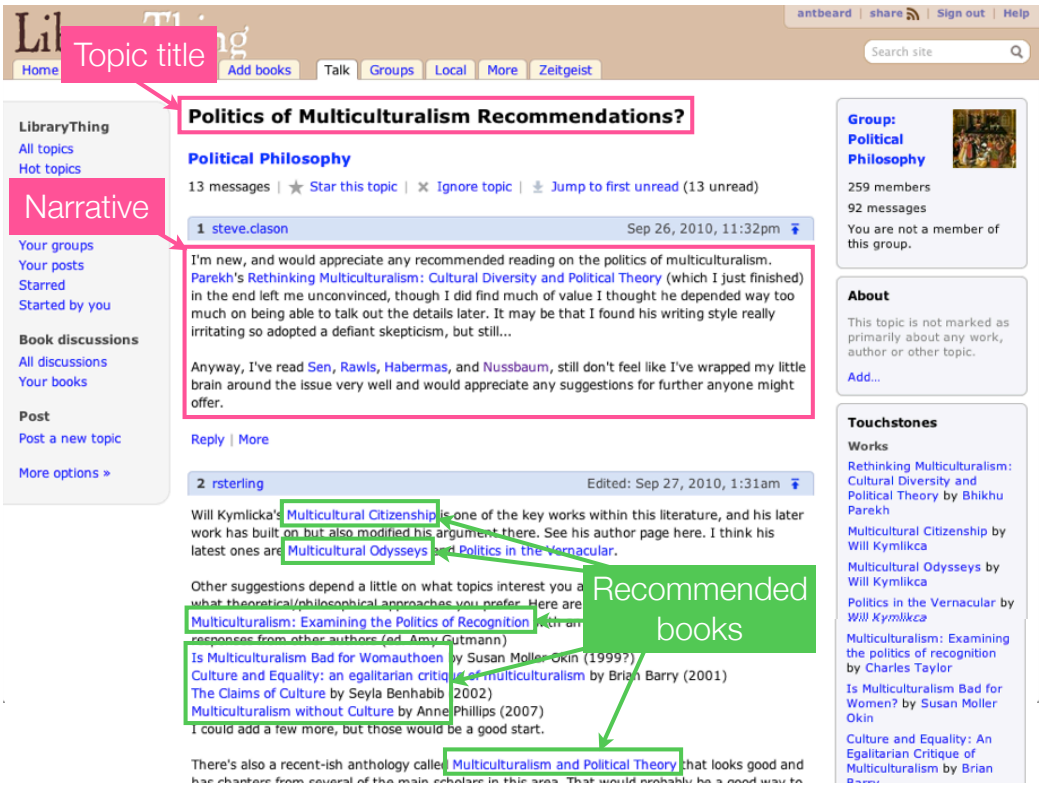

Figure 1: An information need from the LibraryThing discussion forums.

requester, a graded relevance scheme was applied, making some books more relevant than others (Koolen et al., 2013). For the 2014 edition of the INEX Social Book Search track, 680 topics representing information needs and their relevance assessments were provided.

3.4 Experimental Setup

In our experiments we aimed to compare the performance of individual element sets, specific combinations of elements, and all element sets combined. Table 3 shows the retrieval configurations that were used for both collections. In total, 36 different experimental configurations were tested. The Any-CV collection was used to answer our main research questions: (1) which of the metadata elements contributes most to the retrieval success; and (2) which combination of metadata elements achieves the best performance. The Each-CV collection was used to compare the contributions of the different CV providers. Because realistic book search engines always include the core bibliographical data, we added this in combination (Core + CV). The experiments done using the Each-CV collection ask (3) which providers of CV elements contribute most to retrieval performance, and (4) whether adding the core bibliographical data changes the results.

Retrieval Setup

For retrieval experiments, we used language modeling with Jelinek-Mercer smoothing as implemented in the Indri 5.4 toolkit.[3] Previous work has shown that for longer queries such as the rich A/LT topic representations, JM smoothing outperforms Dirichlet smoothing (Zhai & Lafferty, 2004). 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 avoid giving an unfair advantage to one collection over another, we have optimized the retrieval performance of Indri for each of our 36 collection-topic combinations. 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 340 test topics to

[3] Available at http://sourceforge.net/projects/lemur/files/lemur/indri-5.4/, last accessed September 5, 2014.

Table 3: Experimental configurations using different element sets using the Any-CV and Each-CV collections.

Collection	Element set
Any-CV **collection**	Core
	Controlled vocabulary (All)
	Reviews
	Tags
	User-generated (Reviews + Tags)
	Core + Controlled vocabulary (All)
	Core + Reviews
	Core + Tags
	Core + User-generated
	All elements
Each-CV **collection**	Controlled vocabulary (Amazon)
	Controlled vocabulary (BL)
	Controlled vocabulary (LoC)
	Controlled vocabulary (All)
	Core + Controlled vocabulary (Amazon)
	Core + Controlled vocabulary (BL)
	Core + Controlled vocabulary (LoC)
	Core + Controlled vocabulary (All)

produce the results presented in the remainder of this paper. We optimized three different parameters: smoothing, stopword filtering, and stemming. For the degree of smoothing, we varied the λ parameter, which controls the influence of the collection language model, in increments of 0.1, from 0.0 to 1.0. For stopword filtering we either did not filter or applied the SMART stopword list. For stemming we either did not perform stemming or applied the Krovetz stemming algorithm. This resulted in 44 different possible combinations of these three parameters, and $36 \times 44 = 1584$ training runs in total.[4]

Evaluation

To measure retrieval effectiveness, we use NDCG@10 (NDCG cut off at rank 10), which is also used in the INEX Social Book Search Track. It enables comparability and replicability of our results. NDCG stands for Normalized Discounted Cumulated Gain and was proposed by Järvelin and Kekäläinen (2002). 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.

The filtering applied to the original A/LT collection meant that occasionally relevant documents for certain topics also had to be filtered out to keep the evaluation fair. Consequently, the relevance assessments for those documents were also removed to avoid skewing in the results.

We perform statistical significance testing when comparing the retrieval performance of different runs and use an α of 0.05 throughout this paper. In accordance with the guidelines proposed by Sakai (2014), we use two-tailed paired t-tests when comparing the performance of two different 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.

4 Results

This section presents our main experimental results. Section 4.1 starts with the experiments comparing the quality of the controlled vocabulary terms from the different providers using the Each-CV collection. Section 4.2 describes the results for our main experiments comparing the benefits of different (combinations of) element sets on the

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

Any-CV collection.

4.1 Quality Comparison of Controlled Vocabulary Sources

Question 1: Is there a difference in performance between the CV from different providers?
Answer: There is no significant difference in performance between providers or the combination of all three.

The experiments with the Each-CV collection tested which source of CV terms resulted in the best retrieval performance: Amazon, BL, LoC, or a combination of all three. There was no statistically significant difference between these four element sets according to a repeated-measures ANOVA with a Greenhouse-Geisser correction using the Query representation ($F(2.605, 549{,}730) = 0.867$, $p = .445$) or the Narrative representation ($F(2.167, 465.921) = 2.050$, $p = .126$). This means that no matter which of the sources for CVs is used, performance does not change significantly.

Question 2: Does the addition of Core bibliographical data to the CV change retrieval performance?
Answer: There is no difference in provider quality when combining Core bibliographical data with CVs. Including Core bibliographical data in general *does* result in better performance compared to using *only* CVs.

Any real-world book search engine would always include the core bibliographic data in its document representation. It is possible that a combination of the Core bibliographical data and CVs 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 four configurations for either the Query representation,($F(1.667, 355.057) = 0.305$, $p = .697$) or the Narrative representation $F(2.406, 517.282 = 0.973$, $p = .391$).

Is the addition of Core bibliographical data a good idea in general? It is when using a richer information need representation: for the Narrative representation, a repeated-measures ANOVA with a Greenhouse-Geisser correction revealed a statistically significant difference between runs with and without Core bibliographical data ($F(1.305, 280.491) = 3.870$, $p < .05$). Post-hoc tests using the Bonferroni correction revealed that using the combination resulted in a higher NDCG@10 score (0.0773 ± 0.0213 vs. 0.0481 ± 0.0154). For the Query representation this difference was not significant.

4.2 Comparing Retrieval Performance of Different Metadata Element Sets

Table 4 shows the main results (NDCG@10 scores) of our experiments with the Any-CV collection comparing the different (combinations of) element sets for the Query and Narrative representations. Figure 2 represents the information graphically.

Table 4: Results for the different metadata elements and their combination on the Any-CV collection set using NDCG@10 as evaluation metric. Best-performing runs for the individual and combined element sets per topic representation are printed in bold.

Metadata element(s)	**Request representation**	
	Query	Narrative
Core	0.0249	0.0533
Controlled vocabulary	0.0205	0.0319
Reviews	**0.0361**	0.0993
Tags	0.0306	0.0395
User-generated content	0.0296	**0.1046**
Core + Controlled vocabulary	0.0241	0.0540
Core + Reviews	0.0366	0.1063
Core + Tags	0.0378	0.0610
Core + User-generated content	0.0369	0.1114
All metadata elements	**0.0435**	**0.1115**

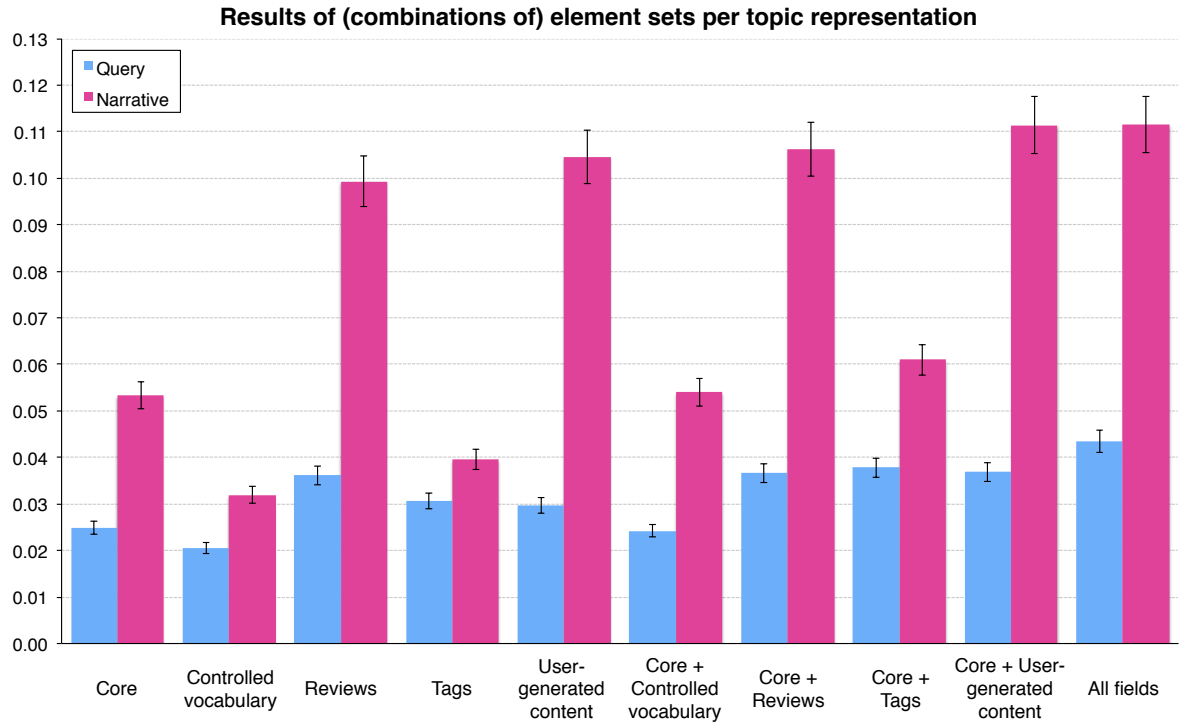

Figure 2: Results for the different metadata elements and their combination on the Any-CV collection set using NDCG@10 as evaluation metric. Bars indicate average NDCG@10 scores over all topics, with error bars in black.

Question 3: Which of the individual element sets contributes most to retrieval success?
Answer: Reviews provide the best retrieval performance, especially for rich information needs. The differences between the other element sets are not significant.

The top half of Table 4 shows that Reviews and UGC provide the best retrieval performance. Using the Query representation, the differences between the individual element sets Core, CV, Reviews, and Tags are not significant according to a repeated-measures ANOVA. However, for the Narrative representation there is a statistically significantly difference between the set ($F(2.605, 794.414) = 18.770$, $p < .0005$). Here, it is Reviews and UGC that significantly outperform the other individual element sets. Another interesting finding is that for the Narrative representation the Core element set significantly outperforms the CV set according to a two-tailed paired t-test ($t(305) = 2.139$, $p < .05$, ES = 0.122, 95% CI [0.0016, 0.0385]).

Question 4: Which combination of element sets achieves the best performance?
Answer: In general, any combination of element sets outperforms the equivalent individual metadata element(s) set(s). 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 element sets containing UGC elements, combining all fields significantly outperforms all other element set configurations. For both topic representations, retrieval scores seem to benefit from adding the Core elements to other metadata elements. For the Query representation these differences are significant according to a two-tailedpaired t-test ($t(1323) = 2.117$, $p < .05$, ES = 0.058, 95% CI [0.0003, 0.0081]) as well as for the Narrative representation ($t(1307) = 4.799$, $p < .0005$, ES = 0.13, 95% CI [0.0083, 0.0199]). Another interesting result is that adding Core bibliographic data to CV results in only a very small improvement. Indeed, these improvements are not statistically significant according to a two-tailed paired t-test for neither Query ($t(330) = 0.159$, $p = .874$, ES = 0.008, 95% CI [-0.0083, 0.0097]) nor Narrative ($t(333) = -0.140$, $p = .889$, ES = 0.001, 95% CI [-0.0107, 0.0092]). This suggests that CV is not complementary to Core bibliographic data.

Question 5: Which metadata element provides better performance: CV or Tags ?
Answer: Despite a slight advantage for Tags, in general neither outperforms the other significantly.

According to Table 4, Tags score higher than CV for both topic representations. However, according to a two-tailed paired t-test these differences are not statistically significat for either the Query representation ($t(329) = 1.518$, $p = .130$, ES = 0.083, 95% CI[-0.0030, 0.0235]) or for the Narrative representation ($t(305) = 1.256$, $p = .210$, ES = 0.071, 95% CI [-0.0056, 0.0255]). When we compare the performance of these two combined with Core bibliographic data, we again find higher scores for the runs that use Tags. For the Query representation, this difference is statistically significant according to a two-tailed paired t-test ($t(330) = 2.264$, $p < .025$, ES = 0.125, 95% CI [0.0018, 0.0255]). However, for the Narrative representation it is not ($t(333) = 1.125$, $p = .262$, ES = 0.062, 95% CI [-0.0052, 0.0192]). In general, this suggests there does not seem to be any meaningful difference between the two metadata elements.

5 Analysis

One of our most interesting findings is that, in general, Tags and CVs do not outperform each other. In this section we delve deeper into this and analyze whether performance differences can be observed under certain conditions. In Section 5.1, we investigate whether Tags and CVs are successful for the same topics (canceling each other out) or for different topics (complementing each other). In Sections 5.2 and 5.3, we analyze whether different types of books and different types of information needs influence the retrieval contribution of Tags and CVs.

5.1 Performance: Tags vs. Controlled Vocabularies

Question 6: Do Tags and CVs complement each other or cancel each other out in terms of retrieval performance?
Answer: Tags and CVs complement each other—both elements are successful with different sets of topics.

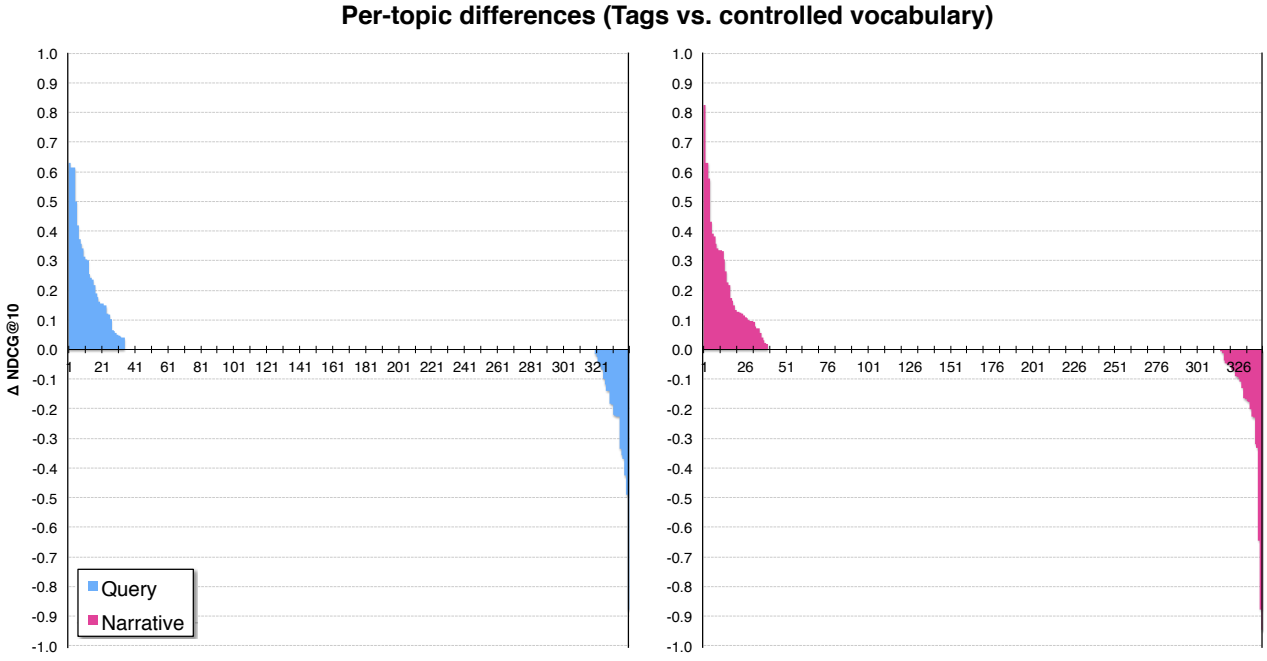

Figure 3: Differences in retrieval performance between the Tags and CV element sets, ordered by per-topic difference. Bars above the horizontal axis represent topics where Tags perform better, bars below the horizontal axis represent topics where CVs perform better.

Figure 3 shows how many of the requests were better fulfilled by Tags (bars above the horizontal line) and how many requests were better fulfilled by searching the CV (bars below the horizontal line). The area above the horizontal axes is larger than the area below, confirming that Tags show a small (non-significant) advantage over CV. Figure 3 also shows that there are different types of topics: for most topics, searching either Tags or the CV makes no difference, but for certain topics one of the two elements outperforms the other.

Table 5 shows the number of topics where either Tags or CVs outperform each other by more than 120%. For those cases, we postulate that one metadata element set contributes much more than the other. As a matter of fact, the retrieval success is almost unequivocally either based on Tags or CVs being searched, because the score for the other metadata element set is very often zero. One important observation is that for the overwhelming majority of requests (over 80%), neither metadata element set can help locate relevant documents. However, we have already seen that the other individual element sets do not fare much better and that the combination of element sets increases performance significantly.

Table 5: Comparison of retrieval performance of Tags or CV for book search

Performance group	**Query**	**Narrative**
Tags > CVs (by at least 120%)	34	37
CV > Tags (by at least 120%)	20	24
Tags = CV and non-zero	2	5
Both Tags and CV fail (NDCG@10 = 0)	284	274

For the remaining requests, where relevant documents were found, Tags contribute to the retrieval success more often than CVs. CVs still provide a substantial contribution: over one third of the requests where relevant documents were found at all could only be found due to this metadata element set. The number of requests where both elements perform similarly is very small (3% and 8% for Query and Narrative respectively). For almost all of the successful requests, retrieval of relevant documents depends either on the information found in the Tags or in the CV. We infer from these numbers that Tags and CVs do not overlap as much as assumed in previous studies (Heymann & Garcia-Molina, 2009; Lu et al., 2010), but actually complement each other.

5.2 Book Types: Fiction vs. Non-fiction

Question 7: Does the type of book have an influence on performance for Tags or CV?
Answer: Tags appear to perform better for retrieving fiction than CV elements, but not significantly so. Retrieving non-fiction books seems to be an easier task than fulfilling requests for fiction books in general.

The answer to the previous question showed that Tags and CV elements succeed on different groups of topics. A next question could be to determine the nature of these different groups. One possibility for dividing our topics into different groups is the type of book requested: fiction or non-fiction. To this end, we annotated all 340 topics as requesting works of fiction or non-fiction. The first 100 topics were annotated individually by the two authors, which resulted in an agreement of 95%. Any remaining differences for these 100 topics were resolved through discussion to arrive at perfect agreement. The remaining 240 topics were annotated by one of the authors, because of the high agreement. The majority of topics (76%) were requests for works of fiction.

Figure 4 shows the comparison of retrieval performance between Tags and CV elements, organized by book type. Both element sets achieve higher scores for non-fiction books in both topic representations, but the differences are not significant according to the one-way ANOVA tests. There is little difference between CV and Tags elements on non-fiction requests. Tags, however, scores higher on fiction requests, but the difference is again not significant when tested using a one-way ANOVA.

Table 6 compares the retrieval performance of Tags and CVs organized by the book type requested. The distinction between book types achieves no large differences in the distribution of the element set's contribution to retrieval success overall. We can see that for fiction books, Tags still contribute successfully to more requests. However, for non-fiction books, the number of requests where CV or Tags retrieve relevant books, is about even. Overall, Tags might be better for describing (and searching) fiction books while both element sets are about even

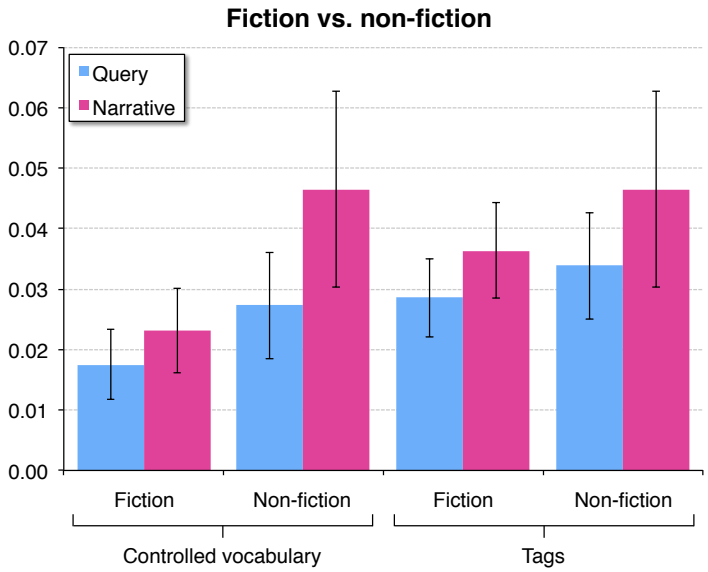

Figure 4: Results of the effect of book type on retrieval performance for Tags and CV elements.

for non-fiction books. The differences are very small, however, and only significant for the query representation of requests.

Table 6: Comparison of retrieval performance of Tags or CVs for book search organized by book type

Performance group	Query		Narrative	
	Fiction	Non-fiction	Fiction	Non-fiction
Tags > CVs (by at least 120%)	25	9	26	11
CV > Tags (by at least 120%)	10	10	17	7
Tags = CV and non-zero	1	1	3	2
Both Tags and CV fail (NDCG@10 = 0)	222	62	212	62

A possible explanation for the trend of Tags being better at dealing with fiction requests could be that the topics of non-fiction books can be determined objectively, whereas the many themes and subjects in works of fiction are harder to index completely using a CV like the LCSH, whereas a group of people tagging these works may able to describe them more comprehensively through their combined effort. Different users may recognize or resonate with different themes, so this results in a more varied description. Since the differences are small, this remains a hypothesis.

5.3 Request Types: Search vs. Recommendation

Question 8: Does the type of information need have an influence on performance for Tags or CV?
Answer: Tags are better for satisfying known-item needs, where only some plot details are remembered as well as needs incorporating aspects of both search and recommendation. CV elements are better for pure recommendation needs represented by past reading behavior. The differences are indicative but not significant.

Another possibility for dividing our topics into different groups is looking at the information need they represent. As discussed in Section 3, information needs from LT discussion forums vary in their intention or format. In our analyis, we distinguish between four types of information needs defined in Table 7.

All 340 test set topics were annotated with respect to whether the LT user's Narrative expressed one of these four, mutually exclusive types of information needs. Again, the first 100 topics were annotated individually by the two authors, resulting in an agreement of 71%. Any remaining differences were resolved through discussion

Table 7: Request types on LT discussion forums.

Information need type	Description
Search	Requesters explicitly ask for books about a topic or from a particular genre.
Search & Recommendation	Requesters explicitly ask for books about a topic or from a particular genre as well as provide information about (dis)similar books they read in the past.
Recommendation	Requesters ask for books that (dis)similar to other books they have read in the (recent) past. These topics often lack an explicit topical information need.
Known-item	Requesters want to re-find a book they cannot remember the title of and supply whatever details they remember about the plot.

to arrive at perfect agreement. The remaining 240 topics were annotated by one of the authors, owing to the relatively high agreement. About half of the requests are of the type 'Search & recommendation' (49%), followed by 'Search' (21%), 'Known-item' (16%) and 'Recommendation' (14%).

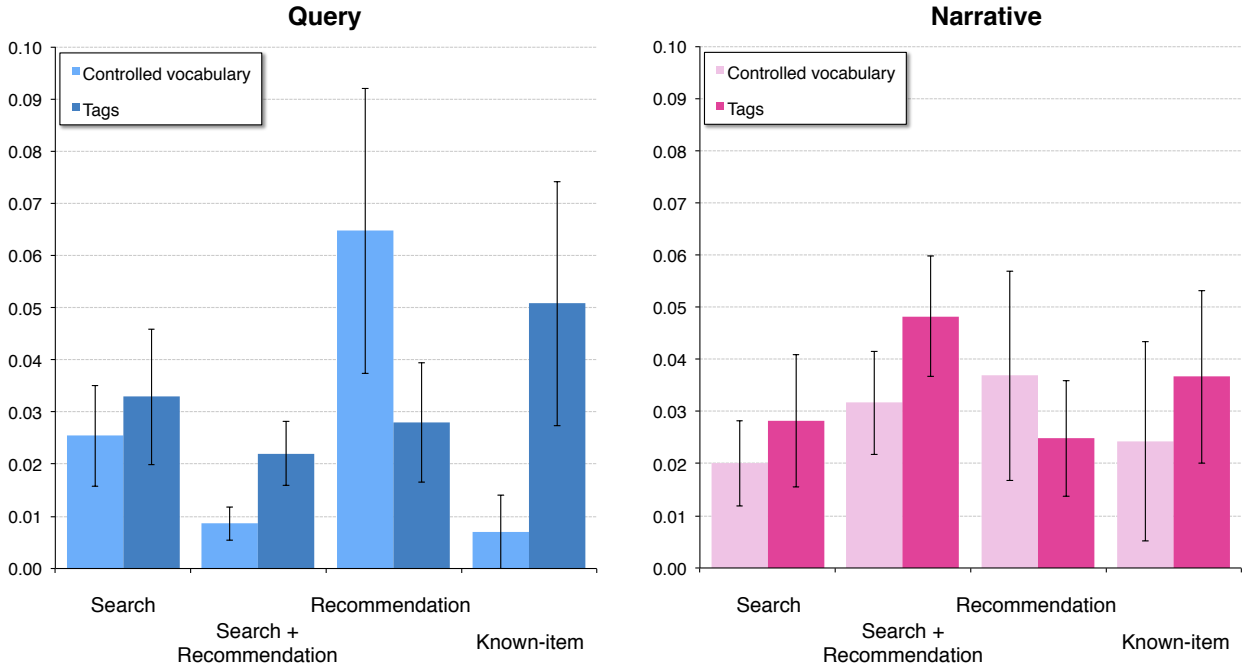

Figure 5: Results of the effect of request type on retrieval performance for Tags and CV elements.

Figure 5 shows the retrieval performance scores for Tags and CVs based on request types (as always both topic representations are analyzed). For both topic representations, Tags perform better than CVs for all but one request type: pure recommendation. The differences between request types are statistically significantly different from each other for the Query representation according to a one-way ANOVA ($F(3, 336) = 3.981$, $p < .01$). The differences for the Narrative representation are not statistically significant ($F(3, 336) = 0.584$, $p = .626$); here, more information about the user's need seems to equalize the differences between the two element sets.

Table 8 compares the retrieval performance of Tags and CVs organized by the request type. It also shows that both element sets perform equally well in search request types, while Tags contributes more for search & recommendation and known-item. These results can be expected when one follows the hypothesis that Tags are better suited for contextual descriptions that are needed for book-based recommendations and known-item searches, where only few details about the content of the book is remembered. Still, known-item search is particularly hard for both metadata element sets. This is to be expected, because often the request are either described in too general a way for the retrieval to be able to narrow it down or too specific in the details for the Tags and CV collections to be able to cover it.

Like known-item, recommendation topics are also very difficult—most result in zero relevant documents retrieved. This suggests that many topics could benefit from using specific recommendation algorithms like collaborative filtering. Recommendations often have an implicit information need, which is not expressed through the query or narrative, but rather the user's past reads. This information is not available to the search engine, so it usually fails on such topics. Perhaps surprisingly, CVs seem to be slightly better than Tags for the recommendation-like requests. None of these differences are significantly different: neither for Query according to a Chi-square test ($\chi^2(9) = 13.780$, $p = .130$), nor for Narrative ($\chi^2(9) = 11.438$, $p = .247$)).

Table 8: Comparison of retrieval performance of Tags and CVs, organized by by request type. (S = search; S & R = search & recommendation; R = recommendation; KI = known-item)

	Query				Narrative			
Performance group	**S**	**S & R**	**R**	**KI**	**S**	**S & R**	**R**	**KI**
Tags > CVs (by at least 120%)	6	20	2	6	4	24	5	4
CV > Tags (by at least 120%)	5	9	6	0	4	12	6	2
Tags = CV and non-zero	1	0	1	0	2	3	0	0
Both Tags and CV fail (NDCG@10 = 0)	60	136	39	49	62	126	37	49

6 Conclusions & Future Work

In this paper have presented a large-scale empirical comparison of different (combinations of) metadata elements for book search, with a emphasis on the comparison between Tags and CVs in particular. The most important conclusion from our study is that Tags and CVs achieve similar retrieval effectiveness in book search. These results were found after leveling the playing field for both as much as possible, by requiring both CV and Tag content to be present in every document. Still, significant differences exist in the distribution of CV terms and Tags. The average number of types is much larger for the CV than the Tags element set, whereas the average number of tokens is much larger for the Tags element set. This means that there are more unique terms in CV, but more repetition of them in Tags.

While differences in retrieval effectiveness are not statistically significant, tags do appear to achieve better scores overall. The differences in type/token averages could offer a possible explanation for this. Despite a lower number of word types for Tags, a similar retrieval performance compared to CVs could mean that the keywords contained in the Tags element set are qualitatively better, i.e., provide a better description of the books' content for the topic representations of the information need. Another, more roundabout explanation is that precision (i.e., finding a few, but highly relevant books at the top of the result list) is more important than recall (i.e., finding all relevant books) for book search. More terms to match on (more types) is likely to benefit recall whereas more repetitions of the same terms (higher token counts) could strengthen precision, because certain terms are more strongly associated with relevance then. This would suggest that CVs improve recall, while Tags have a precision-enhancing effect. Future work should investigate which of these two effects has a higher impact on the retrieval performance for book search. In comparing the retrieval performance, we also did not compare the quality of keywords in the tag or controlled vocabulary metadata elements, something that has been done in the many theoretical comparisons of tags vs. professional metadata. We leave it up to future work to zoom in on either metadata element set and determine the impact of individual keywords for search success.

We also found that Tags and CVs complement each other when contributing to retrieval success—typically, only one of the two metadata elements would contain the terms essential to search success. About a third of the successful requests could only be satisfied due to the CV element set, half due to Tags - not interchangeably. This means that removing either metadata element from the book search engine is likely to decrease retrieval performance. Neither the type of book requested nor the type of information need represented in the book request was able to adequately explain for what type of topics the CVs or Tags elements achieved better results, so more work is needed here to uncover other possible factors.

In comparing CV data from Amazon, the Libray of Congress and the British Library, all three data providers performed similarly, even though Amazon CV metadata elements contained a considerably higher number of

terms and tokens. This can probably be explained by Amazon using very broad and abstract keywords in their category systems such as *books* or *fiction* that do not contribute at all to the retrieval performance.

The overall NDCG@10 scores achieved are quite low. This means that for most of the information needs either represented as a query or a richer narrative, only few relevant books are ranked near the top by the search engine. This shows that book search is a difficult problem to solve! One explanation for this is the selection problem: for each information need, less than 10 books are commonly considered relevant, but the collection is very large. Vague topic representations make it even harder to select the most relevant books. Future work could focus on filtering and focusing information needs and their topic representations for retrieval. Many studies including this one have found that adding more terms to the query generally results in better performance, as the Query vs. Narrative comparison also demonstrates. Future work could focus on which topic representations are the most suitable for book search.

Another explanation for low scores could also be that the content in the metadata elements available for search is not appropriate for the information needs tested. The retrieval experiments have shown that while Tags and CVs are both adequate for book search, other elements, in particular Reviews, work even better. The combination of Core bibliographic data with other elements also has the potential to significantly increase retrieval effectiveness. Future research on book search engines should focus on how to combine metadata elements more effectively.

References

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

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

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

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.

Choi, Y., Hsieh-Yee, I., & 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* (pp. 103–104). New York, NY, USA: ACM.

Clarke, S. G. D. (2008). The Last 50 Years of Knowledge Organization: A Journey through my Personal Archives. *Journal of Information Science*, *34*(4), 427-437.

Cleverdon, C. W., & Mills, J. (1963). The Testing of Index Language Devices. In *ASLIB proceedings* (Vol. 15, pp. 106–130).

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

Gross, T., & 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.

Heymann, P., & Garcia-Molina, H. (2009, February). 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* (pp. 1–4). ACM.

Heymann, P., Koutrika, G., & 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* (pp. 195–206).

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

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

Koolen, M., Kamps, J., & 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* (pp. 185–194).

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

Koolen, M., Kazai, G., Preminger, M., & Doucet, A. (2013). Overview of the INEX 2013 Social Book Search Track. In P. Forner, J. Karlgren, & C. Womser-Hacker (Eds.), *"Information Access Evaluation meets Multilinguality, Multimodality, and Visual Analytics", CLEF '13: Fourth International Conference of the Cross-Language Evaluation Forum* (pp. 1–26).

Koolen, M., Kazai, G., Preminger, M., Kamps, J., Doucet, A., & Landoni, M. (2012). Overview of the INEX 2012 Social Book Search Track. In P. Forner, J. Karlgren, & C. Womser-Hacker (Eds.), *"Information Access Evaluation*

meets Multilinguality, Multimodality, and Visual Analytics", CLEF '12: Third International Conference of the Cross-Language Evaluation Forum (pp. 1–20).
Lancaster, F., Connell, T., Bishop, N., & Mccowan, S. (1991). Identifying Barriers to Effective Subject Access in Library Catalogs. *Library Resources and Technical Services*, *35*(4), 377-391.
Lee, D. H., & 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* (pp. 445–448). New York, NY, USA: ACM.
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 ontext* (pp. 225–234). New York, NY, USA: ACM.
Lu, C., Park, J.-R., & 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.
Magdy, W., & Darwish, K. (2008). Book Search: Indexing the Valuable Parts. In *Proceedings of the 2008 ACM Workshop on Research Advances in Large Digital Book Repositories* (pp. 53–56). New York, NY, USA: ACM.
Noll, M. G., & 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* (pp. 177–186). New York, NY, USA: ACM.
Qin, J. (2008). Folksonomies and Taxonomies: Where the Two Can Meet. *New Dimensions in Knowledge Organization Systems*, *11*.
Rajashekar, T., & 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.
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.
Sakai, T. (2014). Statistical Reform in Information Retrieval? *SIGIR Forum*, *48*(1), 3–12.
Savoy, J., & Abdou, S. (2008). Searching in MEDLINE: Query Expansion and Manual Indexing Evaluation. *Information Processing & Management*, *44*(2), 781-789.
Seki, K., Qin, H., & 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* (pp. 357–360). New York, NY, USA: ACM.
Smith, T. (2007). Cataloging and You: Measuring the Efficacy of a Folksonomy for Subject Analysis. In J. Lussky (Ed.), *Proceedings of the 18th Workshop of the American Society for Information Science and Technology Special Interest Group in Classification Research.*
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.
Zhai, C., & Lafferty, J. (2004). A Study of Smoothing Methods for Language Models Applied to Information Retrieval. *ACM Transactions on Information Systems*, *22*(2), 179–214.

Table of Figures

Table of Tables