Overview of the INEX 2008 Book Track

Gabriella Kazai[1], Antoine Doucet[2], and Monica Landoni[3]

[1] Microsoft Research Cambridge, United Kingdom
gabkaz@microsoft.com
[2] University of Caen, France
doucet@info.unicaen.fr
[3] University of Lugano
monica.landoni@unisi.ch

Abstract. This paper provides an overview of the INEX 2008 Book Track. Now in its second year, the track aimed at broadening its scope by investigating topics of interest in the fields of information retrieval, human computer interaction, digital libraries, and eBooks. The main topics of investigation were defined around challenges for supporting users in reading, searching, and navigating the full texts of digitized books. Based on these themes, four tasks were defined: 1) The Book Retrieval task aimed at comparing traditional and book-specific retrieval approaches, 2) the Page in Context task aimed at evaluating the value of focused retrieval approaches for searching books, 3) the Structure Extraction task aimed to test automatic techniques for deriving structure from OCR and layout information, and 4) the Active Reading task aimed to explore suitable user interfaces for eBooks enabling reading, annotation, review, and summary across multiple books. We report on the setup and results of each of these tasks.

1 Introduction

As a result of numerous mass-digitization projects [2], e.g., Million Book project[4], efforts of the Open Content Alliance[5], and the digitization work of Google[6], the full texts of digitized books are increasingly available on the Web and in digital libraries. The unprecedented scale of these efforts, the unique characteristics of the digitized material, as well as the unexplored possibilities of user interactions present exciting research challenges and opportunities, see e.g., [7].

Motivated by the need to foster research in this domain, the Book Track was launched in 2007 as part of the INEX initiative. The overall goal of the track is to promote inter-disciplinary research investigating techniques for supporting users in reading, searching, and navigating the full texts of digitized books and to provide a forum for the exchange of research ideas and contributions. In 2007, the track concentrated on identifying infrastructure issues, focusing on

[4] http://www.ulib.org/
[5] www.opencontentalliance.org/
[6] http://books.google.com/

information retrieval (IR) tasks. In 2008, the aim was to look beyond and bring together researchers and practitioners in IR, digital libraries, human computer interaction, and eBooks to explore common challenges and opportunities around digitized book collections. Toward this goal, the track set up tasks to provide opportunities for investigating research questions around three broad topics:

- IR techniques for searching collections of digitized books,
- Users' interactions with eBooks and collections of digitized books,
- Mechanisms to increase accessibility to the contents of digitized books.

Based around these main themes, four specific tasks were defined:

1. The Book Retrieval (BR) task, framed within the user task to build a reading list for a given topic, aimed at comparing traditional document retrieval methods with domain-specific techniques exploiting book-specific features, such as the back of book index or associated metadata, like library catalogue information,
2. The Page in Context (PiC) task aimed to test the value of applying focused retrieval approaches to books, where users expect to be pointed directly to relevant book parts,
3. The Structure Extraction (SE) task aimed to evaluate automatic techniques for deriving structure from OCR and layout information for building hyperlinked table of contents, and
4. The Active Reading task (ART) aimed to explore suitable user interfaces enabling reading, annotation, review, and summary across multiple books.

In this paper, we discuss the setup and results of each of these tasks. First, in Section 2, we give a brief summary of the participating organisations. In Section 3, we describe the corpus of books that forms the basis of the test collection. The following three sections detail the four tasks: Section 4 summarises the BR and PiC tasks, Section 5 reviews the SE task, and Section 6 discusses ART. We close in Section 7 with a summary and further plans.

2 Participating Organisations

A total of 54 organisations registered for the track (double from last year's 27), of which 15 took part actively throughout the year (up from 9 last year), see Tables 1 and 2. For active participants, the topics they created and assessed, and the runs they submitted are listed in Table 1. In total, 19 groups downloaded the book corpus, 11 groups contributed 40 search topics, 2 groups submitted runs to the Structure Extraction task, 4 to the Book Retrieval task, 2 to the Page in Context task, and 2 are currently participating in the Active Reading task. A total of 17 participants from 10 known[7] groups contributed to the relevance assessments.

[7] Three of the assessors did not provide an affiliation (topics assessed: 8, 60, 68).

ID	Organisation	Topics	Runs	Assessed topics
6	University of Amsterdam	51, 52, 65	3 BR, 7 PiC	8, 9, 21, 29, 51, 52, 57, 60
7	Oslo University College			12
14	University of California, Berkeley	66, 67	3 BR, ART	
17	University of Strathclyde			9, 21, 55
30	CSIR, Wuhan University	36, 38, 39, 42		
31	Faculties of Management and Information Technologies, Skopje	40, 46, 47, 48		
41	University of Caen	60, 61		31, 37, 60
43	Xerox Research Centre Europe		4 SE	
52	Kyungpook National University	44, 45, 49, 50	ART	1
54	Microsoft Research Cambridge	55, 56, 57, 58, 62, 63, 64, 70		1, 3, 5, 8, 21, 22, 27, 31, 36, 51, 53, 55, 56, 57, 62, 63, 64
56	JustSystems Corporation	53, 54, 59		53, 54
62	RMIT University	31, 37, 41, 43	10 BR	5, 8, 21, 27, 31, 36, 37, 39, 41, 57, 60, 64, 69
78	University of Waterloo	32, 33, 34, 35	2 BR, 6 PiC	12, 51, 53, 62
86	University of Lugano	68, 69		3, 15, 68, 70
125	Microsoft Development Center Serbia		3 SE	

Table 1. Active participants of the INEX 2008 Book Track, contributing topics, runs, and/or relevance assessments (BR = Book Retrieval, PiC = Page in Context, SE = Structure Extraction, ART = Active Reading Task)

3 The Book Corpus

The track builds on a collection of 50,239 digitized out-of-copyright books, provided by Microsoft Live Search and the Internet Archive. The corpus is made up of books of different genre, including history books, biographies, literary studies, religious texts and teachings, reference works, encyclopedias, essays, proceedings, novels, and poetry.

The OCR text of the books has been converted from the original DjVu format to an XML format referred to as BookML, developed by Microsoft Development Center Serbia. BookML provides additional structure information, including markup for table of contents entries. 50,099 of the books also come with an associated MAchine-Readable Cataloging (MARC) record, which contains publication (author, title, etc.) and classification information.

The basic XML structure of a typical book in BookML (ocrml.xml) is a sequence of pages containing nested structures of regions, sections, lines, and words ([coords] represents coordinate attributes, defining the position of a bounding rectangle for a region, line or word, or the width and height of a page):

```
<document>
```

ID	Organisation	ID	Organisation
Passive participants (Corpus download only)			
4	University of Otago	42	University of Toronto
10	Max-Planck-Institut Informatik	116	University of the Aegean
Passive participants			
5	Queensland University of Technology	104	UCLV
8	University College London	107	University of Sci. and Tech. of China
9	University of Helsinki	112	Hitachi, Ltd.
15	University of Iowa	115	IIT
19	University of Ca Foscari di Venezia	117	Iran
21	MPP	118	M.Tech Student
27	University at Albany (also ID=76)	127	UNICAMP
29	Indian Statistical Institute	148	UEA
32	CUHK	158	George Mason University
39	University of New South Wales	160	Universite Jean Monnet
51	Suny-Albany	161	University of California, Santa Cruz
60	Saint Etienne University	164	Isfahan University
66	University of Rostock	165	Universidad de Oriente
88	Independent	166	Drexel University
91	Auckland University of Technology	171	Chinese University of Hong Kong
93	Wuhan Institute of Technology	174	Alexandria University
96	Cairo Microsoft Innovation Center	181	COLTEC
100	Seoul National University		

Table 2. Passive participants of the INEX 2008 Book Track

```
 <page pageNumber=‘‘I-N’’ label=‘‘PT_CHAPTER’’ [coords] key=‘‘0’’ id=‘‘0’’>
  <region regionType=‘‘Text’’ [coords] key=‘‘0’’ id=‘‘0’’>
   <section label=SEC_BODY’’ key=‘‘408’’ id=‘‘0’’>
    <line [coords] key=‘‘0’’ id=‘‘0’’>
     <word [coords] key=‘‘0’’ id=‘‘0’’ val=‘‘Moby’’/>
     <word [coords] key=‘‘1’’ id=‘‘1’’ val=‘‘Dick’’/>
    </line>
    <line [...]>
     <word [...] val=‘‘Herman’’/>
     <word [...] val=‘‘Melville’’/>
    </line>     [...]
   </section>   [...]
  </region>     [...]
 </page>        [...]
</document>
```

BookML provides a set of labels (as attributes) indicating structure information in the full text of a book and additional marker elements for more complex texts, such as a table of contents. For example, a label attribute may indicate the semantic unit that an XML element is likely to be a part of, e.g., a section may be part of a header (SEC_HEADER), a footer (SEC_FOOTER), the back of book index (SEC_INDEX), the table of contents (SEC_TOC), or the body of the page

(SEC_BODY), etc. A page may be labeled as a table of contents page (PT_TOC), an empty page (PT_EMPTY), a back of book index page (PT_INDEX), or as a chapter start page (PT_CHAPTER), etc. Marker elements provide detailed markup, e.g., for table of contents, indicating entry titles (TOC_TITLE), and page numbers (TOC_CH_PN), etc.

The full corpus, which totals around 400GB, was distributed on USB HDDs (at a cost of 70GBP). In addition, a reduced version (50GB, or 13GB compressed) was made available for download. The reduced version was generated by removing the word tags and propagating the values of the val attributes as text content into the parent (i.e., line) elements.

4 Information Retrieval Tasks

Focusing on IR challenges, two search tasks were investigated: 1) Book Retrieval (BR), in which users search for whole books in order to build a reading list on a given topic, and 2) Page in Context (PiC), in which users search for information in books on a given topic and expect to be pointed directly at relevant book parts. Both these tasks used the corpus of over 50,000 books described in Section 3, and the same set of test topics (see Section 4.3). This was motivated by the need to reduce the relevance assessment workload and to allow possible future comparisons across the two tasks.

A summary of the tasks, the test topics, the online relevance assessment system, the collected assessments, and the evaluation results are described in the following sections. Further details and the various DTDs, describing the syntax of submission runs, are available online in the track's Tasks and Submission Guidelines at http://www.inex.otago.ac.nz/tracks/books/taskresultspec.asp.

4.1 The Book Retrieval (BR) Task

This task was set up with the goal to compare book-specific IR techniques with standard IR methods for the retrieval of books, where (whole) books are returned to the user. The user scenario underlying this task is that of a user searching for books on a given topic with the intent to build a reading or reference list. The list may be for research purposes, or in preparation of lecture materials, or for entertainment, etc.

Participants of this task were invited to submit either single runs or pairs of runs. A total of 10 runs could be submitted. A single run could be the result of either generic (non-specific) or book-specific IR methods. A pair of runs had to contain both types, where the non-specific run served as a baseline which the book-specific run extended upon by exploiting book-specific features (e.g., back-of-book index, citation statistics, book reviews, etc.) or specifically tuned methods. One automatic run (i.e., using only the topic title part of a test topic for searching and without any human intervention) was compulsory. A run could contain, for each test topic, a maximum of 1000 books (identified by their 16 character long bookID[8]), ranked in order of estimated relevance.

A total of 18 runs were submitted by 4 groups (3 runs by University of Amsterdam (ID=6); 3 runs by University of California, Berkeley (ID=14); 10 runs by RMIT University (ID=62); and 2 runs by University of Waterloo (ID=78)), see Table 1.

4.2 The Page in Context (PiC) Task

The goal of this task was to investigate the application of focused retrieval approaches to a collection of digitized books. The task was thus similar to the INEX ad hoc track's Relevant in Context task, but using a significantly different collection while also allowing for the ranking of book parts within a book. The user scenario underlying this task was that of a user searching for information in a library of books on a given subject. The information sought may be 'hidden' in some books (i.e., it forms only a minor theme) while it may be the main focus of some other books. In either case, the user expects to be pointed directly to the relevant book parts. Following the focused retrieval paradigm, the task of a focused book search system is then to identify and rank (non-overlapping) book parts that contain relevant information and return these to the user, grouped by the books they occur in.

Participants could submit up to 10 runs, where one automatic and one manual run was compulsory. Each run could contain, for each topic, a maximum of 1000 books estimated relevant to the given topic, ordered by decreasing value of relevance. For each book, a ranked list of non-overlapping XML elements, passages, or book page results estimated relevant were to be listed in decreasing order of relevance. A minimum of one book part had to be returned for each book in the ranking. A submission could only contain one type of results, i.e., only XML elements or only passages; result types could not be mixed.

A total of 13 runs were submitted by 2 groups (7 runs by the University of Amsterdam (ID=6); and 6 runs by the University of Waterloo (ID=78)), see Table 1. All runs contained XML element results (i.e., no passage based submissions were received).

4.3 Test Topics

The test topics are representations of users' informational needs, i.e, the user is assumed to search for information on a given subject. As last year, all topics were limited to deal with content only aspects (i.e., no structural query conditions).

Participants were asked to create and submit topics for which at least 2 but no more than 20 relevant books were found using an online Book Search system (see Section 4.4).

[8] The bookID is the name of the directory that contains the book's OCR file, e.g., A1CD363253B0F403

```
<?xml version=‘‘1.0’’ encoding=‘‘ISO-8859-1’’?>
<!DOCTYPE inex_topic SYSTEM ‘‘bs-topic.dtd’’>
<inex_topic track=‘‘book’’ task=‘‘book-retrieval/book-ad-hoc’’
topic_id=‘‘62’’  ct_no=‘‘2008-37’’>
 <title> Attila the hun </title>
 <description> I want to learn about Attila the Hun’s character, his way of
     living and leading his men, his conquests, and rule.
 </description>
 <narrative>
  <task> I was discussing with some friends about Attila the Hun. What I
     found interesting was the difference in our perceptions of Attila: As a
     great hospitable king vs. a fearsome barbarian. I want to find out more
     about Attila’s character, his way of living as well as about his wars to
     better understand what he and his era of ruling represents to different
     nations.
  </task>
  <infneed> Any information on Attila’s character, his treatment of others, his
     life, his family, his people’s and enemies’ view on him, his ambitions,
     battles, and in general information on his ruling is relevant, and so is any
     information that can shed light on how he is perceived by different nations.
     Poems that paint a picture of Attila, his court and his wars are also relevant.
  </infneed>
 </narrative>
</inex_topic>
```

Fig. 1. Example topic from the INEX 2008 Book Track test set.

A total of 40 new topics (ID: 31-70) were contributed by 11 participating groups (see Table 1), following the topic format described below. These were then merged with the 30 topics created last year for the PiC task (ID: 1-30). An example topic is shown in Figure 1.

Topic Format. The topic format remained unchanged from 2007, each topic consisting of three parts, describing the same information need, but for different purposes and at different level of detail:

<title>: represents the search query that is to be used by systems for the automatic runs. It serves as a short summary of the user’s information need.

<description>: is a natural language definition of the information need.

<narrative>: is a detailed and unambiguous explanation of the information need and a description of what makes a book part relevant or irrelevant. The narrative is taken as the only true and accurate interpretation of the user’s need. It consists of the following two parts:

- <task>: a description of the user task for which information is sought, specifying the context, background and motivation for the information need.

<infneed>: a detailed explanation of what information is sought and what is considered relevant or irrelevant.

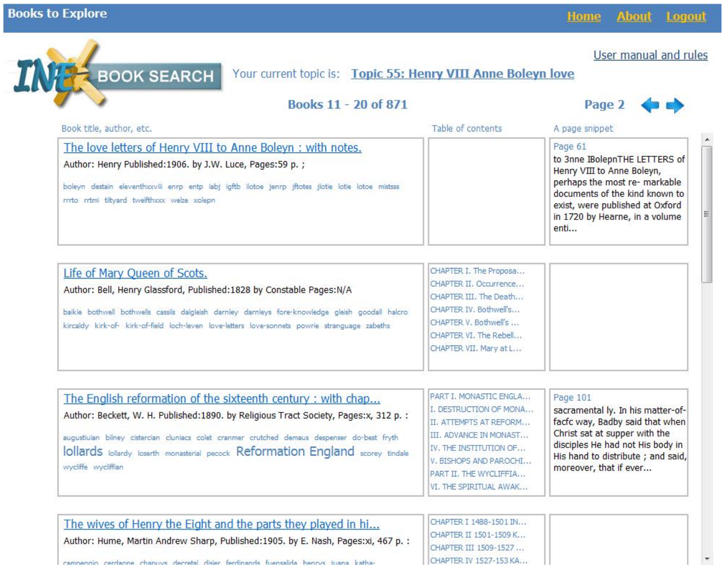
4.4 Relevance Assessment System

The Book Search system (http://www.booksearch.org.uk), developed at Microsoft Research Cambridge, is an online web service that allows participants to search, browse, read, and annotate the books of the test corpus.

For the collection of relevance assessments, a game called the Book Explorers' Competition was designed and deployed, where assessors (as individuals or as members of teams) competed for prizes sponsored by Microsoft Research. The competition involved reading books and marking relevant content inside the books for which assessors were rewarded points. Assessors with the highest scores at the close of the competition were pronounced the winners. The game was modeled as a two-player game with competing roles: explorer vs. reviewer. An explorer's task was to judge the set of pooled pages as well as to locate and mark additional relevant content inside books. Reviewers then had the task of checking the quality of the explorers' work by providing their own relevance assessments for each page that has been judged by at least one explorer. During this process, the reviewers could see the relevance assessments of all the explorers who assessed a particular page. In addition to the passage level exploration, both explorers and reviewers were required, independently (information was not shared), to assign a degree of relevance to the book as a whole (on a scale from 0 to 5, with 5 designating the highest degree of relevance). For further details on the relevance assessment gathering process, please refer to [8].

Screenshots of the assessment system are shown in Figures 2 and 3. Figure 2 shows the list of books in the assessment pool to be judged for a given topic. The list was built by pooling all the submitted runs, i.e., both BR and PiC runs, using a round robin process and merging additional search results from the Book Search system itself. Selecting a book from the list, opened the Book Viewer window (see Figure 3). There, assessors could browse through the book and search inside it, or go through the pages listed in the Assessment tab, which were pooled from the submitted PiC runs. Assessors could highlight text fragments on a page by drawing a highlight-box over the page image. They could also mark a whole page or a range of pages as relevant/irrelevant. A detailed user manual and system description is available at http://www.booksearch.org.uk/BECRulesAndUserManual.pdf.

Two rounds of the Book Explorers' Competition were run. The first round (run in Dec 2008) lasted two weeks and resulted in three winners. One of them participated as an individual assessor and the other two formed a team. The second round (run in Jan 2009) spanned four weeks and yielded four winners. All four assessors belonged to the same team; one among them also achieving the highest individual score.



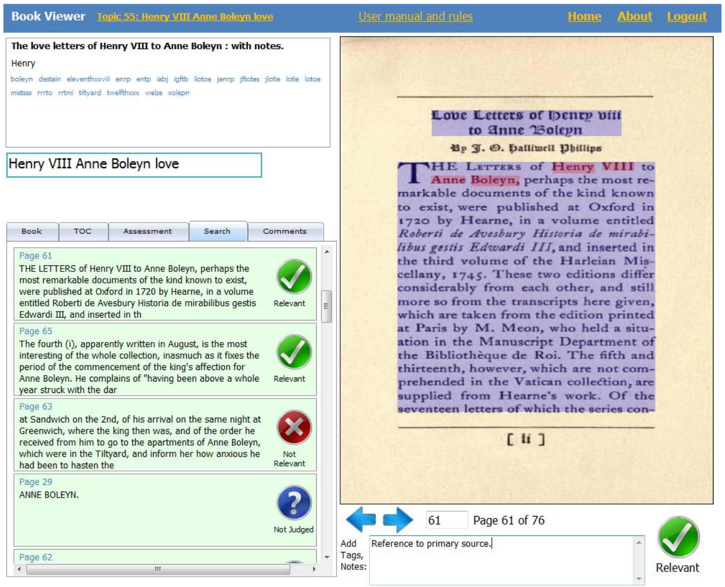
Your current topic is: Topic 55: Henry VIII Anne Boleyn love

Books 11 - 20 of 871

Page 2

Book title, author, etc.	Table of contents	A page snippet
The love letters of Henry VIII to Anne Boleyn : with notes. Author: Henry Published:1906. by J.W. Luce, Pages:59 p. ; boleyn destain eleventhxxviii enrp entp iabj igftb ilotoe jenrp jftotes jlotie lotie lotoe mistsss rrrto rrtmi tiltyard twelfthxxx welze xolepn		Page 61 to 3nne IBolepnTHE LETTERS of Henry VIII to Anne Boleyn, perhaps the most re- markable documents of the kind known to exist, were published at Oxford in 1720 by Hearne, in a volume enti...
Life of Mary Queen of Scots. Author: Bell, Henry Glassford, Published:1828 by Constable Pages:N/A baikie bothwell bothwells cassils dalgleish darnley darnleys fore-knowledge gleish goodall halcro kircaldy kirk-of- kirk-of-field loch-leven love-letters love-sonnets powrie stranguage zabeths	CHAPTER I. The Proposa... CHAPTER II. Occurrence... CHAPTER III. The Death... CHAPTER IV. Bothwell's... CHAPTER V. Bothwell's ... CHAPTER VI. The Rebell... CHAPTER VII. Mary at L...	
The English reformation of the sixteenth century : with chap... Author: Beckett, W. H. Published:1890. by Religious Tract Society, Pages:x, 312 p. : augustiuian bilney cistercian cluniacs colet cranmer crutched demaus despenser do-best fryth lollards lollardy loserth monasterial pecock Reformation England scorey tindale wycliffe wycliffian	PART I. MONASTIC ENGLA... I. DESTRUCTION OF MONA... II. ATTEMPTS AT REFORM... III. ADVANCE IN MONAST... IV. THE INSTITUTION OF... V. BISHOPS AND PAROCHI... PART II. THE WYCLIFFIA... VI. THE SPIRITUAL AWAK...	Page 101 sacramental ly. In his matter-of-facfc way, Badby said that when Christ sat at supper with the disciples He had not His body in His hand to distribute ; and said, moreover, that if ever...
The wives of Henry the Eight and the parts they played in hi... Author: Hume, Martin Andrew Sharp, Published:1905. by E. Nash, Pages:xi, 467 p. : campeggio cerdagne chapuys decretal disier ferdinands fuensalida henrys juana katha-	CHAPTER I 1488-1501 IN... CHAPTER II 1501-1509 K... CHAPTER III 1509-1527 ... CHAPTER IV 1527-153 KA...	

Fig. 2. Screenshot of the relevance assessment module of the Book Search system: List of books in the assessment pool for a selected topic.



The love letters of Henry VIII to Anne Boleyn : with notes.

Henry

boleyn destain eleventhxxviii enrp entp iabj igftb ilotoe jenrp jftotes jlotie lotie lotoe mistsss rrrto rrtmi tiltyard twelfthxxx welze xolepn

Henry VIII Anne Boleyn love

Page 61
THE LETTERS of Henry VIII to Anne Boleyn, perhaps the most remarkable documents of the kind known to exist, were published at Oxford in 1720 by Hearne, in a volume entitled Roberti de Avesbury Historia de mirabilibus gestis Edwardi III, and inserted in th
✓ Relevant

Page 65
The fourth (i), apparently written in August, is the most interesting of the whole collection, inasmuch as it fixes the period of the commencement of the king's affection for Anne Boleyn. He complains of "having been above a whole year struck with the dar
✓ Relevant

Page 63
at Sandwich on the 2nd, of his arrival on the same night at Greenwich, where the king then was, and of the order he received from him to go to the apartments of Anne Boleyn, which were in the Tiltyard, and inform her how anxious he had been to hasten the
✗ Not Relevant

Page 29
ANNE BOLEYN.
? Not Judged

Page 62

Love Letters of Henry viii
to Anne Boleyn

By J. O. Halliwell Phillips

THE LETTERS of Henry VIII to Anne Boleyn, perhaps the most remarkable documents of the kind known to exist, were published at Oxford in 1720 by Hearne, in a volume entitled *Roberti de Avesbury Historia de mirabilibus gestis Edwardi III*, and inserted in the third volume of the Harleian Miscellany, 1745. These two editions differ considerably from each other, and still more so from the transcripts here given, which are taken from the edition printed at Paris by M. Meon, who held a situation in the Manuscript Department of the Bibliothèque de Roi. The fifth and thirteenth, however, which are not comprehended in the Vatican collection, are supplied from Hearne's work. Of the seventeen letters of which the series con-

[ii]

61 Page 61 of 76

Add Tags, Notes: Reference to primary source.

✓ Relevant

Fig. 3. Screenshot of the relevance assessment module of the Book Search system: Book Viewer window with Assessment tab showing, listing pooled pages to judge.

4.5 Collected Relevance Assessments

The collection of relevance assessments was frozen on the 25th of February 2009. The data collected includes the highlight-boxes drawn by assessors on a page, the binary relevance labels assigned by judges to a page, any notes and comments added for a page, and the relevance degree assigned to the books. In total, 3674 unique books and 33,120 unique pages were judged across 29 topics, and 1019 highlight boxes were drawn by 17 assessors. Table 3 provides a breakdown of the assessments per topic. For more details on the collected data, please refer to [8].

From the collected assessments, separate book-level and page-level assessment sets (qrels) were produced, where multiple relevance labels assigned by multiple assessors were averaged. For example, a book with assigned relevance degrees of 3 and 5 (by two assessors from the multi-grade scale of 0-5) yielded an averaged score of 4. Note that the score that appears in the qrels is this value multiplied by 10. The page-level qrel set is similarly the average of the binary scores (0-1) assigned by multiple assessors to a page, multiplied by 10. For example, a page with scores of $\{0, 1, 1, 1\}$ yielded 0.75*10. A weighted version of the qrel sets was also released to participants, where assessors' topic familiarity was taken into account: $w = \frac{\sum f \cdot r}{\sum f}$, where w is the weighted average, r is the relevance score given to a page or book by the assessor, and f is the assessor's familiarity with the topic (as provided by the assessor on a seven point scale, where 1 meant practically no knowledge about the topic and 7 represented an expert on the area). For example, if a book was rated as 3 by an assessor with familiarity of 6, and rated as 5 by an assessor with familiarity of 1, then the weighted score is $(3 \cdot 6 + 5 \cdot 1)/(6 + 1) = 3.28$.

4.6 Evaluation Measures and Results

Both IR tasks were evaluated using standard IR measures reported by trec_eval v8.1[9]. The ranking of books in both the BR and PiC runs was evaluated as traditional document retrieval, by comparing the ranked list of books returned by systems to the book-level qrel set. To do this, the runs were first converted to TREC format, during which some runs were truncated at rank 1000; *rank* values were derived based on the ordering of book results in a run; and the *rsv* was set to $1000 - rank$.

The ranking of book parts in the PiC task was evaluated at page-level for each book, treating each page as a document and comparing the ranked list of pages returned by systems to the page-level qrels for that book, and then averaging over the run (where additional relevant, but not retrieved books were given 0 scores). Note that retrieved XML elements that were at a finer granularity level than page elements were converted to page-level results to match the qrel set granularity.

Tables 4, 5, and 6 show the results for the BR, PiC book-level, and PiC page-level evaluations, respectively. In addition, Figure 4 shows the recall/precision curves for BR runs.

Topic	Books				Pages		
ID	Total Judged	Relevant	Irrelevant	Skipped	Total Judged	Relevant	Irrelevant
25 topics - used in evaluation							
3	4	4	0	0	414	360	101
8	15	1	14	2	562	23	551
9	235	35	199	2	5991	285	5818
12	133	11	116	9	5660	48	5612
21	66	31	37	1	6026	1400	4696
22	30	12	18	0	956	244	712
27	35	21	14	1	365	101	274
31	18	9	17	0	129	46	97
36	9	7	2	0	1073	1043	30
37	15	7	11	0	120	34	99
39	25	7	18	0	358	27	331
41	14	9	5	0	370	276	94
51	135	15	107	14	1813	555	1270
52	41	23	18	0	1651	199	1456
53	1000	14	986	0	88	76	12
54	385	10	375	0	107	104	3
55	29	20	9	0	2108	397	1714
56	13	7	6	0	139	62	77
57	171	25	147	4	845	83	764
60	85	56	30	6	508	226	310
62	100	38	61	2	868	215	672
63	38	7	31	0	303	37	266
64	23	9	14	0	757	669	89
68	1	1	0	0	313	206	107
69	16	3	13	0	75	12	63
25	**2636**	**382**	**2248**	**41**	**31599**	**6728**	**25218**
Additional assessments - not used in evaluation							
1	999	0	999	0	55	0	54
5	33	0	33	2	495	145	495
15	1	0	1	0	0	0	0
29	5	0	4	1	421	0	421
70	0	0	0	0	495	7	488
5	**1038**	**0**	**1037**	**3**	**1466**	**154**	**1458**
29	**3674**	**382**	**3285**	**44**	**33120**	**6889**	**26724**

Table 3. Collected Relevance Assessments (25 February 2009)

We summarise below the main findings, but note that since the qrels vary greatly across topics, these should be treated more as preliminary observations.

For the BR task, the 4 submitting groups experimented with various techniques, e.g., using book content vs. MARC record information [9], ranking books by document score vs. best element score [5], or ranking books by the percentage

[9] http://trec.nist.gov/trec_eval/index.html

ParticipantID+RunID	MAP	iP[0.00]	iP[0.10]	P5	P10	P20
14_BOOKSONLY	0.0837	0.3761	0.3135	0.192	**0.136**	0.082
14_MARCONLY	0.0386	0.2302	0.1421	0.088	0.056	0.046
14_MERGEMARCDOC	0.0549	0.3076	0.2528	0.144	0.088	0.064
54_BSS	0.0945	0.3715	0.2484	0.168	**0.136**	0.09
6_BST08_B_clean_trec	0.0899	**0.4051**	0.2801	0.176	0.132	0.096
6_BST08_B_square_times_sim100_top8_fw_trec	0.0714	0.2771	0.223	0.152	0.12	0.088
6_inex08_BST_book_sim100_top8_forward_trec	0.0085	0.1058	0.0406	0.032	0.02	0.01
62_RmitBookTitle	0.0747	0.2469	0.2195	0.128	0.104	0.094
62_RmitBookTitleBoolean	0.0747	0.2469	0.2195	0.128	0.104	0.094
62_RmitBookTitleInfneed	0.067	0.331	0.1999	0.136	0.1	0.086
62_RmitBookTitleInfneedManual	0.0682	0.2757	0.1868	0.112	0.108	0.088
62_RmitConPageMergeTitle	0.05	0.2414	0.2017	0.104	0.072	0.064
62_RmitConPageMergeTitleBoolean	0.05	0.2414	0.2017	0.104	0.072	0.064
62_RmitConPageMergeTitleInfneedManual	0.0544	0.2786	0.2126	0.128	0.084	0.058
62_RmitPageMergeTitle	0.0742	0.3022	0.2601	0.144	0.116	0.084
62_RmitPageMergeTitleBoolean	0.0741	0.3022	0.2601	0.144	0.116	0.084
62_RmitPageMergeTitleInfneedManual	**0.1056**	0.3671	**0.3456**	**0.216**	0.132	**0.098**
78_1	0.0193	0.117	0.0683	0.024	0.012	0.01
78_2	0.0214	0.1162	0.0678	0.024	0.012	0.008

Table 4. Evaluation results for the BR runs

ParticipantID+RunID	MAP	iP[0.00]	iP[0.10]	P5	P10	P20
6_BST08_P_clean_trec	0.078	0.3359	0.2077	0.136	0.108	**0.096**
6_BST08_P_plus_B_trec	0.0761	0.3734	0.2028	0.136	**0.116**	0.078
6_BST08_P_plus_sim100_top8_fw_trec	0.0707	0.2794	0.1775	0.128	0.092	0.08
6_BST08_P_times_B_trec	0.0532	0.3179	0.1905	0.112	0.068	0.048
6_BST08_P_times_sim100_top8_fw_trec	0.0646	0.3408	0.1643	0.136	0.1	0.074
6_BST08_P_with_B_trec	**0.0785**	**0.3761**	**0.2189**	**0.152**	**0.116**	0.088
6_BST08_P_with_sim100_top8_fw_trec	0.053	0.2532	0.1645	0.128	0.096	0.062
78_3	0.0214	0.1162	0.0678	0.024	0.012	0.008
78_4	0.0513	0.278	0.2096	0.096	0.076	0.05
78_5	0.0214	0.1162	0.0678	0.024	0.012	0.008
78_6	0.0495	0.2744	0.205	0.096	0.076	0.048
78_7	0.0495	0.2744	0.205	0.096	0.076	0.048
78_8	0.0495	0.2744	0.205	0.096	0.076	0.048

Table 5. Book-level evaluation results for the PiC runs

of pages retrieved [12], as well as incorporating Wikipedia evidence [6]. The best performing run was a run submitted by RMIT (ID=62), ranking books by the percentage of pages retrieved using BM25 over a page level index (MAP=0.1056). The general conclusion, however, for the other 3 groups' experiments was that the simple book content based baseline performed better than any attempts to combine book-specific evidence to improve performance. This suggests that there is still plenty to be done in discovering suitable ranking strategies for books.

For the PiC task, the 2 submitting groups mostly experimented with ways of combining document and element level scoring methods [5,6]. The best performing runs, based on book-level scores, were submitted by the University of Amsterdam (ID=6), who found that while focused retrieval methods were able to locate relevant text within books, page level evidence was of limited use without the wider context of the whole book. The best page-level results were achieved by the University of Waterloo (ID=78), ranking book parts by element score and using no cutoff to limit the size of the ranked list (runs: 78_7 and 78_8).

ParticipantID+RunID	P	R	F
6_BST08_P_clean_trec	0.069	0.028	0.027
6_BST08_P_plus_B_trec	0.069	0.028	0.027
6_BST08_P_plus_sim100_top8_fw_trec	0.069	0.028	0.027
6_BST08_P_times_B_trec	0.069	0.028	0.027
6_BST08_P_times_sim100_top8_fw_trec	0.069	0.028	0.027
6_BST08_P_with_B_trec	0.064	0.027	0.025
6_BST08_P_with_sim100_top8_fw_trec	0.068	0.028	0.026
78_3	0.066	0.084	0.045
78_4	0.068	0.096	0.048
78_5	0.069	0.098	0.056
78_6	**0.070**	0.11	0.057
78_7	0.059	**0.14**	**0.065**
78_8	0.059	**0.14**	**0.065**

Table 6. Page-level evaluation results for the PiC runs (precision, recall and the harmonic mean of precision and recall (F-measure))

5 The Structure Extraction (SE) Task

The goal of this task was to test and compare automatic techniques for extracting structure information from digitized books and building a hyperlinked table of contents (ToC). The task was motivated by the limitations of current digitization and OCR technologies that produce the full text of digitized books with only minimal structure markup: Pages and paragraphs are usually identified, but more sophisticated structures, such as chapters, sections, etc., are typically not recognised.

Participants of the task were provided a sample collection of 100 digitized books of different genre and styles in DjVu XML format. Unlike the BookML format of the main corpus, the DjVu files only contain markup for the basic structural units (e.g., page, paragraph, line, and word); no structure labels and markers are available. In addition to the DjVu XML files, participants were

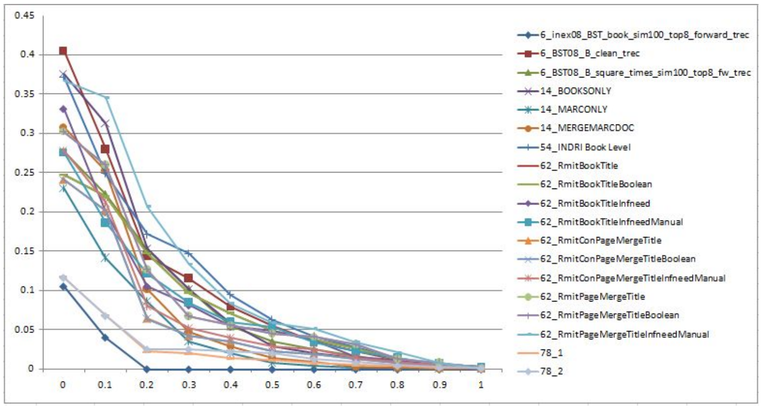


Fig. 4. Recall/precision curves for BR runs.

distributed the PDF of books or the set of JPEG image files (one per book page).

Participants could submit up to 10 runs, each containing the generated table of contents for the 100 books in the test set.

A total of 7 runs were submitted by 2 groups (3 runs by Microsoft Development Center Serbia (MDCS) (ID=125), and 4 runs by Xerox Research Centre Europe (XRCE) (ID=43)).

5.1 Evaluation Measures and Results

For the evaluation of the SE task, the ToCs generated by participants were compared to a manually built ground-truth, created by hired assessors, using a structure labeling tool built by Microsoft Development Center Serbia. The tool allowed assessors to attach labels to entries and parts of entries in the printed ToC of a book (using the PDF file as source).

Performance was evaluated using recall/precision like measures at different structural levels (i.e., different depths in the ToC). Precision was defined as the ratio of the total number of correctly recognized ToC entries and the total number of ToC entries; and recall as the ratio of the total number of correctly recognized ToC entries and the total number of ToC entries in the ground-truth. The F-measure was then calculated as the harmonic of mean of precision and recall. For further details on the evaluation measures, please see http://www.inex.otago.ac.nz/tracks/books/INEXBookTrackSEMeasures.pdf. The ground-truth and the evaluation tool can be downloaded from http://www.inex.otago.ac.nz/tracks/books/Results.asp#SE.

The evaluation results are given in Table 7. According to this, the best performance ($F = 53.47\%$) was obtained by the MDCS group (ID=125), who extracted ToCs by first recognizing the page(s) of a book that contained the printed ToC

[10]. The XRCE group (ID=43) relied on title detection within the body of a book and achieved a score of $F = 10.27\%$ [3].

ParticipantID+RunID	F-measure
125_MDCS	**53.47%**
125_MDCS_NAMES_AND_TITLES	52.59%
125_MDCS_TITLES_ONLY	23.24%
43_HF_ToC_prg_Jaccard	10.27%
43_HF_ToC_prg_OCR	10.18%
43_HF_TPF_ToC_prg_Jaccard	10.10%
43_HF_ToC_lin_Jaccard	5.05%

Table 7. Evaluation results for the SE task (complete ToC entries)

6 The Active Reading Task (ART)

The main aim of ART is to explore how hardware or software tools for reading eBooks can provide support to users engaged with a variety of reading related activities, such as fact finding, memory tasks, or learning. The goal of the investigation is to derive user requirements and consequently design recommendations for more usable tools to support active reading practices for eBooks. The task is motivated by the lack of common practices when it comes to conducting usability studies of e-reader tools. Current user studies focus on specific content and user groups and follow a variety of different procedures that make comparison, reflection, and better understanding of related problems difficult. ART is hoped to turn into an ideal arena for researchers involved in such efforts with the crucial opportunity to access a large selection of titles, representing different genres and appealing to a variety of potential users, as well as benefiting from established methodology and guidelines for organising effective evaluation experiments.

ART is based on the large evaluation experience of EBONI [11], and adopts its evaluation framework with the aim to guide participants in organising and running user studies whose results could then be compared.

The task is to run one or more user studies in order to test the usability of established products (e.g., Amazon's Kindle, iRex's Ilaid Reader and Sony's Readers models 550 and 700) or novel e-readers by following the provided EBONI-based procedure and focusing on INEX content. Participants may then gather and analyse results according to the EBONI approach and submit these for overall comparison and evaluation. The evaluation is task-oriented in nature. Participants are able to tailor their own evaluation experiments, inside the EBONI framework, according to resources available to them. In order to gather user feedback, participants can choose from a variety of methods, from low-effort online questionnaires to more time consuming one to one interviews, and think aloud sessions.

6.1 Task Setup

Participation requires access to one or more software/hardware e-readers (already on the market or in prototype version) that can be fed with a subset of the INEX book corpus (maximum 100 books), selected based on participants' needs and objectives. Participants are asked to involve a minimum sample of 15/20 users to complete 3-5 growing complexity tasks and fill in a customised version of the EBONI subjective questionnaire, usually taking no longer than half an hour in total, allowing to gather meaningful and comparable evidence. Additional user tasks and different methods for gathering feedback (e.g., video capture) may be added optionally. A crib sheet (see below) is provided to participants as a tool to define the user tasks to evaluate, providing a narrative describing the scenario(s) of use for the books in context, including factors affecting user performance, e.g., motivation, type of content, styles of reading, accessibility, location and personal preferences.

ART crib sheet. A task crib sheet is a rich description of a user task that forms the basis of a given user study based on a particular scenario in a given context. Thus, it aims to provide a detailed explanation of the context and motivation of the task, and all details that form the scenario of use:

- Objectives: A summary of the aims and objectives of the task from the users' point of view, i.e., what is it that users are trying to achieve in this task.
- Task: Description of the task.
- Motivation: Description of the reasons behind running the task.
- Context: Description of the context of the task in terms of time and resources available, emphasis and any other additional factors that are going to influence task performance.
- Background: Description of any background knowledge required to accomplish the task.
- Completion: Description of how to assess whether the task has been completed or not.
- Success: Description of whether the task has been completed successfully.

Participants are encouraged to integrate questionnaires with interviews and think aloud sessions when possible, and adapt questionnaires to fit into their own research objectives whilst keeping in the remit of the active reading task.

We also encourage direct collaboration with participants to help shape the tasks according to real/existing research needs. In fact one of the participants explained how English written material was not much use for their experiments as they were targeting Korean speaking users, so it was agreed that they would use their own book collection while still adopting the ART evaluation framework to ensure results were comparable at the end.

Our aim is to run a comparable but individualized set of studies, all contributing to elicit user and usability issues related to eBooks and e-reading.

Since ART is still ongoing, there is no data to be presented at this point.

7 Conclusions and plans

The Book Track this year has attracted a lot of interest and has grown to double the number of participants from 2007. However, active participation remained a challenge for most due to the high initial set up costs (e.g., building infrastructure). Most tasks also require advance planning and preparations, e.g., for setting up a user study. This, combined with the late announcement and advertising of some of the tasks has limited active participation this year. In particular, we received expressions of interest for the Structure Extraction and the Active Reading tasks, but the deadlines prohibited most people from taking part. We aim to address this issue in INEX 2009 by raising awareness early on in the start of the INEX year and by ensuring continuity with the tasks established this year.

As a first step in this direction, we are proposing to run the Structure Extraction task both at INEX 2009 and at ICDAR 2009 (International Conference on Document Analysis and Recognition) with an increased set of 1,000 books.

Both the Book Retrieval and Page in Context tasks will be run again in 2009, albeit with some modifications. The BR task will be shaped around the user task of compiling a reading list for selected Wikipedia articles, while we aim to expand the PiC tasks to tree retrieval [1].

The greatest challenge in running these two tasks has been the collection of relevance assessments. Due to the huge effort required, we decided to depart from the traditional method of relevance assessment gathering (i.e., one judge per topic), and designed a system where multiple judges assess the same topic. Implemented as an online game, assessors contributed relevance labels for passages, pages, and whole books on the topics they were interested in and for any number of books on that topic. This way of collecting judgements is aimed to provide a more realistic expectation on the assessors, but it also comes with its own risks. Attracting a sufficiently large group of dedicated assessors is one of the risks, for example. To address this issue, we are currently looking at using Amazon's Mechanical Turk service, as well as investigating the possibility of opening up the Book Search system and allowing users to create their own topics and saving their searches and book annotations for these. Other risks include the question of the quality of the collected relevance data due to a mixture of expert and non-expert judges. Working toward a solution, we introduced a number of measures, such as requiring assessors to specify their familiarity with their selected topics, as well as allowing users to quality check each other's work. We aim to explore additional measures in our future work.

We also plan to re-run this year's Active Reading task in 2009. We found that the introduction of ART was a challenge for number of reasons:

- Because of its original approach to evaluation, which is quite far away from the classic TREC paradigm, and the relative difficulty in framing ART in a formal way, the task organisation has suffered delays that have affected the availability of participants to get fully involved in it;
- User studies are per se risky and unpredictable and the idea of running a number of those in parallel in order to compare and combine results added

an extra layer of uncertainty to the task, somehow discouraging participants that were used to a more stochastic approach to evaluation;

- The formalisation of the procedure and protocols to be followed when running user studies was designed on purpose to be flexible and unconstructive in order to accommodate for participants' specific research needs. This flexibility, however, was interpreted by some as a lack in details that discouraged them from taking part.
- Opening up to different communities that were not yet involved in INEX required concentrated effort in order to advertise and raise awareness of what INEX's aims and objectives and in particular what ART's goals were. Some of this effort was simply too late for some interested parties.

The organisation of ART has proved a valuable experience though that has given us the opportunity to explore different research perspective while focusing on some of the practical aspects of the task. We believe that the effort that has gone into setting up ART this year will be rewarded by a more successful task next year.

Acknowledgements

The Book Track in 2008 was supported by the Document Layout Team of Microsoft Development Center Serbia. The team contributed to the track by providing the BookML format and a tool to convert books from the original OCR DjVu files to BookML. They also contributed to the Structure Extraction task by helping us prepare the ground-truth data and by developing the evaluation tools.

References

1. M S. Ali, Mariano P. Consens, Gabriella Kazai, and Mounia Lalmas. Structural relevance: a common basis for the evaluation of structured document retrieval. In *CIKM '08: Proceeding of the 17th ACM Conference on Information and Knowledge Management*, pages 1153–1162, New York, NY, USA, 2008. ACM.
2. K. Coyle. Mass digitization of books. *Journal of Academic Librarianship*, 32(6):641–645, 2006.
3. Hervé Déjean and Jean-Luc Meunier. XRCE participation to the book structure task. In Geva et al. [4].
4. Shlomo Geva, Jaap Kamps, and Andrew Trotman, editors. *Advances in Focused Retrieval: 7th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2008)*, Lecture Notes in Computer Science. Springer Verlag, Berlin, Heidelberg, 2009.
5. Kelly Itakura and Charles Clarke. University of Waterloo at INEX2008: Adhoc, book, and link-the-wiki tracks. In Geva et al. [4].
6. Jaap Kamps and Marijn Koolen. The impact of document level ranking on focused retrieval. In Geva et al. [4].

7. Paul Kantor, Gabriella Kazai, Natasa Milic-Frayling, and Ross Wilkinson, editors. *BooksOnline '08: Proceeding of the 2008 ACM workshop on Research advances in large digital book repositories*, New York, NY, USA, 2008. ACM.
8. Gabriella Kazai, Natasa Milic-Frayling, and Jamie Costello. Towards methods for the collective gathering and quality control of relevance assessments. In *SIGIR '09: Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, 2009.
9. Ray Larson. Adhoc and book XML retrieval with Cheshire. In Geva et al. [4].
10. Aleksandar Uzelac, Bodin Dresevic, Bogdan Radakovic, and Nikola Todic. Book layout analysis: TOC structure extraction engine. In Geva et al. [4].
11. R. Wilson, M. Landoni, and F. Gibb. The web experiments in electronic textbook design. *Journal of Documentation*, 59(4):454–477, 2003.
12. Mingfang Wu, Falk Scholer, and James A. Thom. The impact of query length and document length on book search effectiveness. In Geva et al. [4].