Overview of the INEX 2009 Book Track

Gabriella Kazai[1], Antoine Doucet[2], Marijn Koolen[3], and Monica Landoni[4]

[1] Microsoft Research, United Kingdom
v-gabkaz@microsoft.com
[2] University of Caen, France
doucet@info.unicaen.fr
[3] University of Amsterdam, Netherlands
m.h.a.koolen@uva.nl
[4] University of Lugano
monica.landoni@unisi.ch

Abstract. The goal of the INEX 2009 Book Track is to evaluate approaches for supporting users in reading, searching, and navigating the full texts of digitized books. The investigation is focused around four tasks: 1) the Book Retrieval task aims at comparing traditional and book-specific retrieval approaches, 2) the Focused Book Search task evaluates focused retrieval approaches for searching books, 3) the Structure Extraction task tests automatic techniques for deriving structure from OCR and layout information, and 4) the Active Reading task aims to explore suitable user interfaces for eBooks enabling reading, annotation, review, and summary across multiple books. We report on the setup and the results of the track.

1 Introduction

The INEX Book Track was launched in 2007, prompted by the availability of large collections of digitized books resulting from various mass-digitization projects [1], such as the Million Book project[5] and the Google Books Library project[6]. 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. [3].

The overall goal of the INEX Book 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. Toward this goal, the track aims to provide opportunities for exploring research questions around three broad topics:

- Information retrieval techniques for searching collections of digitized books,
- Mechanisms to increase accessibility to the contents of digitized books, and

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

– Users' interactions with eBooks and collections of digitized books.

Based around these main themes, the following four tasks were defined:

1. The Book Retrieval (BR) task, framed within the user task of building a reading list for a given topic of interest, aims at comparing traditional document retrieval methods with domain-specific techniques, exploiting book-specific features, e.g., back-of-book index, or associated metadata, e.g., library catalogue information,
2. The Focused Book Search (FBS) task aims 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 aims at evaluating automatic techniques for deriving structure from OCR and layout information for building hyperlinked table of contents, and
4. The Active Reading task (ART) aims to explore suitable user interfaces enabling reading, annotation, review, and summary across multiple books.

In this paper, we report on the setup and the results of each of these tasks at INEX 2009. 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 two search tasks (BR and FBS), Section 5 reviews the SE task, and Section 6 discusses ART. We close in Section 7 with a summary and plans for INEX 2010.

2 Participating Organisations

A total of 84 organisations registered for the track (compared with 54 in 2008, and 27 in 2007), of which 16 took part actively throughout the year (compared with 15 in 2008, and 9 in 2007); these groups are listed in Table 1.

In total, 7 groups contributed 16 search topics comprising a total of 37 topic aspects (sub-topics), 4 groups submitted runs to the SE task, 3 to the BR task, and 3 groups submitted runs to the FBS task. Two groups participated in ART, but did not submit results. 9 groups contributed relevance judgements.

3 The Book Corpus

The track builds on a collection of 50,239 out-of-copyright books[7], digitized by Microsoft. 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. 50,099 of the books also come with an associated MAchine-Readable Cataloging (MARC) record, which contains publication (author, title, etc.) and classification information.

[7] Also available from the Internet Archive (although in a different XML format)

Table 1. Active participants of the INEX 2009 Book Track, contributing topics, runs, and/or relevance assessments (BR = Book Retrieval, FBS = Focused Book Search, SE = Structure Extraction, ART = Active Reading Task)

ID	**Institute**	**Topics**	**Runs**	**Judged topics (book/page level)**
6	University of Amsterdam	8, 11	2 BR, 4 FBS	Book: 3, 5, 7, 8, 11, 14, 15; Page: 8, 11, 14
7	Oslo University College	1, 2	10 BR, 10 FBS	Book 1, 2; Page: 1, 2
12	University of Granada			Book: 1, 16; Page: 1
14	Uni. of California, Berkeley		9 BR, ART	
29	Indian Statistical Institute			Book: 16
41	University of Caen	7, 9	3 SE	SE
43	Xerox Research Centre Europe		3 SE	SE
52	Kyungpook National Uni.	3, 4	ART	
54	Microsoft Research Cambridge	10, 16		Book: 3, 5, 7, 9, 10, 16; Page: 3, 5, 7, 9, 10, 16
78	University of Waterloo	5, 6	4 FBS	Book: 5, 6; Page: 5, 6
86	University of Lugano	12, 13, 14, 15		
125	Microsoft Dev. Center Serbia		1 SE	
335	Fraunhofer IAIS			SE
339	Universita degli Studi di Firenze			SE
343	Noopsis Inc.		1 SE	
471	Peking University, ICST			SE
	Unkown			Book: 13, 16

Each book in the corpus is identified by a 16 character long bookID – the name of the directory that contains the book's OCR file, e.g., A1CD363253B0F403.

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. The basic XML structure of a typical book in BookML is a sequence of pages containing nested structures of regions, sections, lines, and words, most of them with associated coordinate information, defining the position of a bounding rectangle ([coords]):

```
<document>
 <page pageNumber="1" 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="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 structures, such as a table of contents. For example, the first label attribute in the XML extract above signals the start of a new chapter on page 1 (label="PT_CHAPTER"). Other semantic units include headers (SEC_HEADER), footers (SEC_FOOTER), back-of-book index (SEC_INDEX), table of contents (SEC_TOC). 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, totaling around 400GB, was made available on USB HDDs. 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), and 2) Focused Book Search (FBS). Both these tasks used the corpus described in Section 3, and shared the same set of topics (see Section 4.3).

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, similar to those appended to an academic publication or a Wikipedia article. The reading 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, each run containing the results for all the 16 topics (see Section 4.3). A single run could be the result of either a generic (non-specific) or a book-specific IR approach. 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 topic for searching and without any human intervention) was compulsory. A run could contain, for each topic, a maximum of 1,000 books (identified by their bookID), ranked in order of estimated relevance.

A total of 21 runs were submitted by 3 groups (2 runs by University of Amsterdam (ID=6); 9 runs by University of California, Berkeley (ID=14); and 10 runs by Oslo University College (ID=7)), see Table 1. The 21 runs contained a total of 316,000 books, 1,000 books per topic (4 runs from Oslo University College only contained results for 11 of the 16 topics).

4.2 The Focused Book Search (FBS) 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, where 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 of the 37 topic aspects (see Section 4.3), a maximum of 1,000 books estimated relevant to the given aspect, ordered by decreasing value of relevance. For each book, a ranked list of non-overlapping book parts, i.e., XML elements or passages, 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 result, i.e., only XML elements or only passages.

A total of 18 runs were submitted by 3 groups (4 runs by the University of Amsterdam (ID=6); 10 runs by Oslo University College (ID=7); and 4 runs by the University of Waterloo (ID=78)), see Table 1. The 18 runs contained a total of 444,098 books and 2,638,783 pages; 5.94 pages per book. All runs contained XML elements, and in particular page level elements, with the exception of two runs by the University of Waterloo, which also contained title elements.

4.3 Topics

Topics are representations of users' information needs that may be more or less generic or specific. Reflecting this, a topic may be of varying complexity and may comprise one or multiple aspects (sub-topics). We encouraged participants to create multiple aspects for their topics, where aspects should be focused (narrow) with only a few expected relevant book parts (e.g., pages).

Participants were recommended to use Wikipedia when preparing their topics. The intuition behind the introduction of Wikipedia is twofold. First, Wikipedia can be seen as a real world application for both the BR and FBS tasks: articles often contain a reading list of books relevant to the overall topic of the article,

```
<topic id=‘‘10’’ cn_no=‘‘60’’>
<task>Find relevant books and pages to cite from the Wikipedia article on
      Cleopatra’s needle</task>
<title>Cleopatra needle obelisk london paris new york</title>
<description>I am looking for reference material on the obelisks known as
             Cleopatra’s needle, three of which have been erected: in London,
             Paris, and New York.</description>
<narrative>I am interested in the obelisks’ history in Egypt, their transportation,
           their physical descriptions, and current locations. I am, however, not
           interested in the language of the hieroglyphics.</narrative>
<wikipedia-title>Cleopatra’s needle</wikipedia-title>
<wikipedia-url>http://en.wikipedia.org/wiki/Cleopatra’s_Needle</wikipedia-url>
<wikipedia-text>Cleopatra’s Needle is the popular name for each of three Ancient
                Egyptian obelisks [...] </wikipedia-text>
<aspect aspect_id=‘‘10.1’’>
<aspect-title>Description of the London and New York pair</aspect-title>
<aspect-narrative>I am looking for detailed physical descriptions of the London and
                  New York obelisks as well as their history in Egypt. When and
                  where they were originally erected and what happened to them when
                  they were moved to Alexandria.</aspect-narrative>
<aspect-wikipedia-text>The pair are made of red granite, stand about 21 meters
                       (68 ft) high, weigh [...] </aspect-wikipedia-text>
</aspect>
<aspect aspect_id=‘‘10.2’’>
<aspect-title>London needle</aspect-title>
<aspect-narrative>I am interested in details about the obelisk that was moved to
                  London. When and where was it moved, the story of its
                  transportation. Information and images of the needle and the two
                  sphinxes are also relevant.</aspect-narrative>
<aspect-wikipedia-text>The London needle is in the City of Westminster, on the
                       Victoria Embankment [...] </aspect-wikipedia-text>
</aspect>
<aspect aspect_id=‘‘10.3’’>
<aspect-title>New York needle</aspect-title>
<aspect-narrative>I am looking for information and images on the obelisk that was
                  moved to New York. Its history, its transportation and
                  description of its current location.</aspect-narrative>
<aspect-wikipedia-text>The New York needle is in Central Park. In 1869, after the
                       opening of the Suez Canal, [...] </aspect-wikipedia-text>
</aspect>
<aspect aspect_id=‘‘10.4’’>
<aspect-title>Paris needle</aspect-title>
<aspect-narrative>Information and images on the Paris needle are sought. Detailed
                  description of the obelisk, its history, how it is different from
                  the London and New York pair, its transportation and current
                  location are all relevant.</aspect-narrative>
<aspect-wikipedia-text>The Paris Needle (L’aiguille de Cleopatre) is in the Place
                       de la Concorde. The center [...] </aspect-wikipedia-text>
</aspect>
</topic>
```

Fig. 1. Example topic from the INEX 2009 Book Track test set.

while they also often cite related books in relation to a specific statement in the article. Thus, we anticipated that browsing through Wikipedia entries could provide participants with suggestions about topics and their specific aspects of interest. Second, Wikipedia, can also provide participants with insights and relevant terminology to be used for better searches and refinements that should lead to a better mapping between topics and collection.

An example topic is shown in Figure 1. In this example, the overall topic includes all three Egyptian obelisks known as Cleopatra's needle, which were erected in London, Paris, and New York. The topic aspects focus on the history of the individual obelisks or on their physical descriptions. Paragraphs in the associated Wikipedia page (¡wikipedia-url¿) relate to the individual topic aspects, while the whole article relates to the overall topic.

Participants were asked to create and submit 2 topics, ideally with at least 2 aspects each, for which relevant books could be found in the corpus. To aid participants with this task, an online Book Search System (see Section 4.4) was developed, which allowed them to search, browse and read the books in the collection.

A total of 16 new topics (ID: 1-16), containing 37 aspects (median 2 per topic), were contributed by 7 participating groups (see Table 1). The collected topics were used for retrieval in the BR task, while the topic aspects were used in the FSB task.

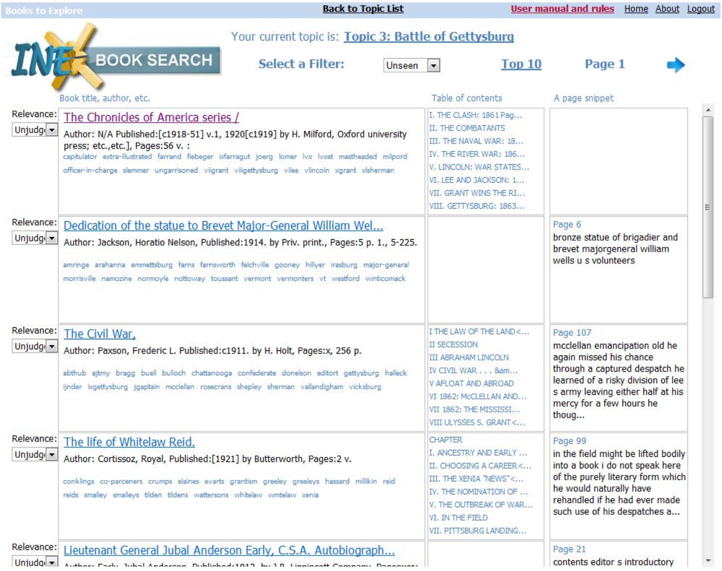
4.4 Relevance Assessment System

The Book Search System (`http://www.booksearch.org.uk`), developed at Microsoft Research Cambridge, is an online tool that allows participants to search, browse, read, and annotate the books of the test corpus. Annotation includes the assignment of book and page level relevance labels and recording book and page level notes or comments. The system supports the creation of topics for the test collection and the collection of relevance assessments. Screenshots of the relevance assessment module are shown in Figures 2 and 3.

In 2008, a game called the Book Explorers' Competition was developed to collect relevance assessments, where assessors competed for prizes [4]. The competition involved reading books and marking relevant content inside the books for which assessors were rewarded points. The game was based on two competing roles: *e*xplorers, who discovered relevant content inside books and *r*eviewers, who checked the quality of the explorers' assessments.

Based on what we learnt in 2008, we modified the game this year to consist of three separate, but interconnected 'Read and Play' games: In game 1, participants had the task of finding books relevant to a given topic and then ranking the top 10 most relevant books. In game 2, their task was to explore the books selected in game 1 and find pages inside them that are relevant to a given topic aspect. Finally, in game 3, their task was to review pages that were judged in game 2. Hence, we have, in essence, introduced a filtering stage (game 1) before the Book Explorer's Competition (game 2 and 3) in order to reduce the number of books to judge in detail.

The aim of game 1 was to collect book level judgements for the evaluation of the BR task, while page level assessments gathered in games 2 and 3 would be used to evaluate the FBS task.



Your current topic is: Topic 3: Battle of Gettysburg

Select a Filter: Unseen

Relevance	Book title, author, etc.	Table of contents	A page snippet
Relevance: Unjudg	The Chronicles of America series / Author: N/A Published:[c1918-51] v.1, 1920[c1919] by H. Milford, Oxford university press; etc.,etc.], Pages:56 v. : capitulator extra-illustrated farrand fiebeger ixfarragut joerg lomer lvx lvxet mastheaded milpord officer-in-charge slemmer ungarrisoned viigrant viigettysburg vilee vlincoln xgrant xlsherman	I. THE CLASH: 1861 Pag... II. THE COMBATANTS III. THE NAVAL WAR: 18... IV. THE RIVER WAR: 186... V. LINCOLN: WAR STATES... VI. LEE AND JACKSON: 1... VII. GRANT WINS THE RI... VIII. GETTYSBURG: 1863...	
Relevance: Unjudg	Dedication of the statue to Brevet Major-General William Wel... Author: Jackson, Horatio Nelson, Published:1914. by Priv. print., Pages:5 p. 1., 5-225. amringe arahanna emmettsburg farns farnsworth felchville gooney hillyer irasburg major-general morrisville namozine normoyle nottoway toussant vermont vermonters vt westford winticomack		Page 6 bronze statue of brigadier and brevet majorgeneral william wells u s volunteers
Relevance: Unjudg	The Civil War, Author: Paxson, Frederic L. Published:c1911. by H. Holt, Pages:x, 256 p. abthub ajtmy bragg buell bulloch chattanooga confederate donelson editort gettysburg halleck ijnder ixgettysburg jgaptain mcclellan rosecrans shepley sherman vallandigham vicksburg	I THE LAW OF THE LAND<... II SECESSION III ABRAHAM LINCOLN IV CIVIL WAR . . . &am... V AFLOAT AND ABROAD VI 1862: McCLELLAN AND... VII 1862: THE MISSISSI... VIII ULYSSES S. GRANT<...	Page 107 mcclellan emancipation old he again missed his chance through a captured despatch he learned of a risky division of lee s army leaving either half at his mercy for a few hours he thoug...
Relevance: Unjudg	The life of Whitelaw Reid. Author: Cortissoz, Royal, Published:[1921] by Butterworth, Pages:2 v. conklings co-parceners crumps elaines evarts grantism greeley greeleys hassard millikin reid reids smalley smalleys tilden tildens wattersons whitelaw wmtelaw xenia	CHAPTER I. ANCESTRY AND EARLY ... II. CHOOSING A CAREER<... III. THE XENIA "NEWS"<... IV. THE NOMINATION OF ... V. THE OUTBREAK OF WAR... VI. IN THE FIELD VII. PITTSBURG LANDING...	Page 99 in the field might be lifted bodily into a book i do not speak here of the purely literary form which he would naturally have rehandled if he had ever made such use of his despatches a...
Relevance: Unjudg	Lieutenant General Jubal Anderson Early, C.S.A. Autobiograph... Author: Early, Jubal Anderson, Published:1912. by J.B. Lippincott Company, Pages:		Page 21 contents editor s introductory

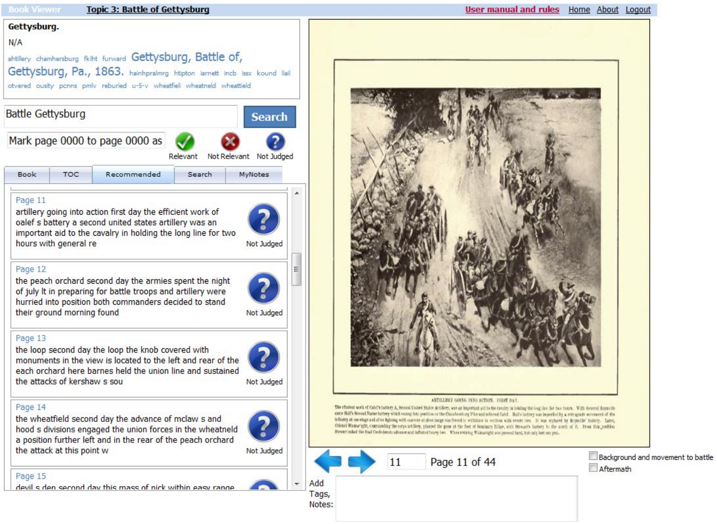
Fig. 2. Screenshot of the relevance assessment module of the Book Search System, showing the list of books in the assessment pool for a selected topic in game 1. For each book, its metadata, its table of contents (if any) and a snippet from a recommended page is shown.

4.5 Collected Relevance Assessments

We run the 'Read and Play' games for three weeks (ending on March 15, 2010), with weekly prizes of $50 worth of Amazon gift card vouchers, shared between the top three scorers, proportionate to their scores. Additional judgments were collected up to the period of April 15, 2010, with no prizes. Table 2 provides a summary of all the collected relevance assessments. The last column shows the implicit page level judgements, i.e., for pages in the assessment pool that are inside books that were judged irrelevant.

In total, we collected 4,668 book level relevance judgements from 9 assessors in game 1. Assessors were allowed to judge books for any topic, thus some books were judged by multiple assessors. The total number of unique topic-book pair judgements is 4,430.

In game 1, assessors could choose from 4 possible labels: "relevant", "top 10 relevant", "irrelevant" and "unsure". The latter label could be used either



Gettysburg.

N/A

ahtillery chamhersburg fkiht furward Gettysburg, Battle of, Gettysburg, Pa., 1863. hainhpralmrg htipton iarnett incb issx kound liail otvered ousity pcnns pmlv reburied u-5-v wheatfieli wheatneld wheattield

Battle Gettysburg

Mark page 0000 to page 0000 as

Page 11
artillery going into action first day the efficient work of oalef s battery a second united states artillery was an important aid to the cavalry in holding the long line for two hours with general re
Not Judged

Page 12
the peach orchard second day the armies spent the night of july lt in preparing for battle troops and artillery were hurried into position both commanders decided to stand their ground morning found
Not Judged

Page 13
the loop second day the loop the knob covered with monuments in the view is located to the left and rear of the each orchard here barnes held the union line and sustained the attacks of kershaw s sou
Not Judged

Page 14
the wheatfield second day the advance of mclaw s and hood s divisions engaged the union forces in the wheatneld a position further left and in the rear of the peach orchard the attack at this point w
Not Judged

Page 15
devil s den second day this mass of nick within easy range

ARTILLERY GOING INTO ACTION. FIRST DAY.

The efficient work of Calef's battery A, Second United States Artillery, was an important aid to the cavalry in holding the long line for two hours. With General Reynolds came Hall's Second Maine battery which swung into position on the Chambersburg Pike and relieved Calef. Hall's battery was imperilled by a retrograde movement of the infantry at one stage and after fighting with canister at close range was forced to withdraw in sections with severe loss. It was replaced by Reynolds' battery. Later, Colonel Wainwright, commanding the corps artillery, planted the guns at the foot of Seminary Ridge, with Stewart's battery to the north of it. From this position Stewart raked the final Confederate advance and inflicted heavy loss. When retiring Wainwright was pressed hard, but only lost one gun.

11 Page 11 of 44

Fig. 3. Screenshot of the relevance assessment module of the Book Search System, showing the Book Viewer window with Recommended tab listing the pooled pages to judge with respect to topic aspects in game 2. The topic aspects are shown below the page images.

to delay a decision on a given book, or when it was not possible to assess the relevance of a book due to language or technical reasons (e.g., the book was unreadable or could not be displayed). Books ranked in the top 10 most relevant books for a topic were labeled with "top 10 relevant". This was, however, seldom assigned, only in 34 cases across 10 topics.

Page level judgements could be contributed in all three games. However, in game 1, pages could only be judged with respect to the whole topic, while in games 2 and 3, pages were judged with respect to the individual topic aspects. The latter is required for the evaluation of the FBS task. For topics with a single aspect, i.e., 7, 9, 12, and 13, page level judgements could be collected in any of the games.

From the table, it is clear that game 1 proved much more popular than games 2 and 3. There are two principle reasons for this. On the one hand, games 2 and 3 can only start once books filtered through to them from game 1. On the other hand, in game 1, it is enough to find a single relevant page in a book to mark it relevant, while in games 2 and 3, judges need to read and judge a lot more of a book's content.

Out of the 4,430 books 230 was judged by 2 assessors and 4 by 3 judges. Judges only disagreed on 23 out of the 230 double-judged books, and 2 of the 4 triple-judged books.

Table 2. Collected relevance judgements per topic (up to April 15, 2010)

Topic	**Judged books**	**Rel. books**	**Judged pages**	**Rel. pages**	**Impl. irrel.**
	(game 1)	(game 1)	(games 1/2&3)	(games 1/2&3)	**(pages)**
1	61	10	628/0	602/0	1364
2	57	8	55/0	48/0	993
3	106	65	107/235	106/235	1850
5	1763	14	17/26	16/26	25074
6	90	9	192/0	20/0	4104
7	171	58	26/0(26)	25/0(25)	1608
8	471	155	12/0	1/0	9979
9	121	29	23/0(23)	23/0(23)	581
10	172	25	88/0	39/0	4526
11	1104	95	46/0	0/0	18860
13	9	7	0/0	0/0	19
14	195	18	3/0	1/0	3822
15	31	22	0/0	0/0	4
16	79	33	78/0	66/0	1200
Total	4,430	548	1,275/310	947/309	73,984

Due to the very few judgements available for topic aspects, we will only report results for the BR task in the next section.

4.6 Evaluation Measures and Results

For the evaluation of the BR task, we converted the book level assessments into binary judgements. Judgements labeled "relevant" or "top 10 relevant" were mapped to 1, and judgements labeled "irrelevant" or "unsure" were mapped to 0. If multiple assessors judged a book for a topic, a majority vote was used to determine whether a book is relevant or not. Ties were treated as relevant.

Table 3 shows the results for the BR task. Based on participants' descriptions of their retrieval methods, we marked runs that were book-specific in some way, e.g., used back-of-book index, with an * in the table. From these results, it appears that book-specific information is not yet incorporated into the retrieval approaches successfully, but it seems to hurt retrieval effectiveness in the current state of the art. Looking at the per topic results for MAP, see Figure 4, we found that only topic 2 had a book-specific approach as its best performance. For P10, book-specific retrieval strategies obtained best performance for topic 2, and tied with generic retrieval methods on topics 1, 5, 13, and 15. The MRR measure ties the two approaches on all but three topics: ge method is best on topics 1, and 11, and book-specific is best on topic 2. Bpref shows that generic IR methods are superior for all topics. For possible explanations into why book-specific methods do not improve on the traditional IR approaches, please refer to the respective papers, published by the participants of the book track, in the proceedings.

Table 3. Results for the Book Retrieval Task

Run id	**MAP**	**MRR**	**P10**	**bpref**	**Rel.Ret.**
p14_BR_BOOKS2009_FUS_TA*	0.1536	0.6217	0.3429	0.3211	200
p14_BR_BOOKS2009_FUS_TITLE*	0.1902	0.7907	0.4214	0.4007	310
p14_BR_BOOKS2009_OK_INDEX_TA*	0.0178	0.1903	0.0857	0.1737	127
p14_BR_BOOKS2009_OK_TOC_TA*	0.0185	0.1529	0.0714	0.1994	153
p14_BR_BOOKS2009_T2_INDEX_TA*	0.0448	0.2862	0.1286	0.1422	80
p14_BR_BOOKS2009_T2_TOC_TA*	0.0279	0.3803	0.0929	0.1164	75
p14_BR_BOOKS2009_OK_TOPIC_TA	0.0550	0.2647	0.1286	0.0749	41
p14_BR_BOOKS2009_T2FB_TOPIC_TA	0.2309	0.6385	0.4143	0.4490	329
p14_BR_BOOKS2009_T2FB_TOPIC_TITLE	0.2643	0.7830	0.4714	0.5014	345
p6_BR_inex09.book.fb.10.50	**0.3471**	**0.8507**	0.4857	**0.5921**	419
p6_BR_inex09.book	0.3432	0.8120	**0.5286**	0.5842	416
p7_BR_to_b_submit*	0.0915	0.4180	0.2000	0.2375	184
p7_BR_to_g_submit	0.1691	0.5450	0.3357	0.3753	325
p7_BR_tw_b3_submit*	0.0639	0.3984	0.1857	0.2015	164
p7_BR_tw_g3_submit	0.1609	0.5394	0.3214	0.3597	319
p7_BR_tw_b5_submit*	0.0646	0.4292	0.2000	0.1866	139
p7_BR_tw_g5_submit	0.1745	0.6794	0.3357	0.3766	326
p7_BR_wo_b3_submit*	0.0069	0.0333	0.0286	0.0422	70
p7_BR_wo_g3_submit	0.0272	0.2102	0.0786	0.1163	140
p7_BR_wo_b5_submit*	0.0108	0.0686	0.0500	0.0798	48
p7_BR_wo_g5_submit	0.0680	0.4779	0.1214	0.2067	173

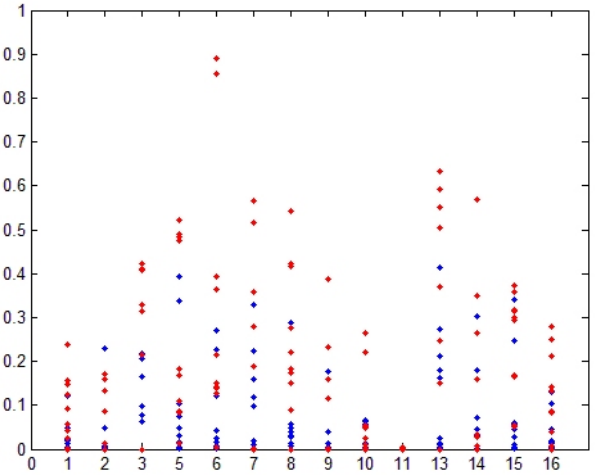
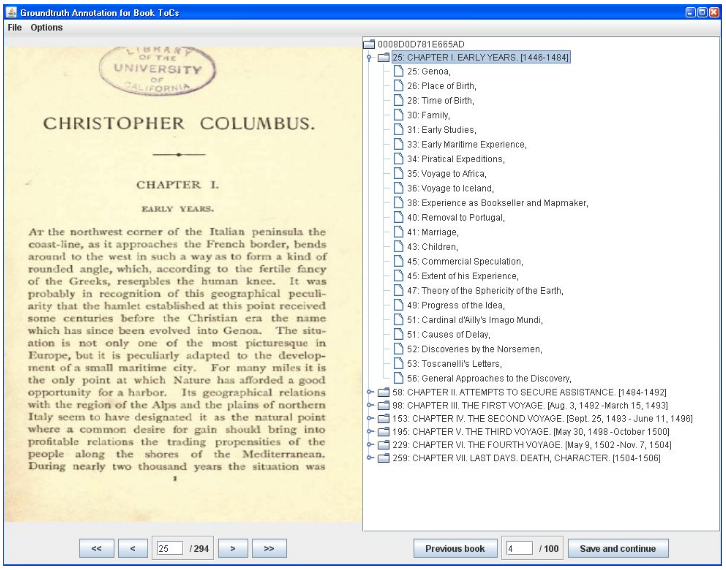


Fig. 4. Distribution of MAP scores across the 14 assessed topics in the BR task. Book-specific approaches are shown as blue dots, while generic IR approaches are shown as red dots.



CHRISTOPHER COLUMBUS.

CHAPTER I.

EARLY YEARS.

At the northwest corner of the Italian peninsula the coast-line, as it approaches the French border, bends around to the west in such a way as to form a kind of rounded angle, which, according to the fertile fancy of the Greeks, resembles the human knee. It was probably in recognition of this geographical peculiarity that the hamlet established at this point received some centuries before the Christian era the name which has since been evolved into Genoa. The situation is not only one of the most picturesque in Europe, but it is peculiarly adapted to the development of a small maritime city. For many miles it is the only point at which Nature has afforded a good opportunity for a harbor. Its geographical relations with the region of the Alps and the plains of northern Italy seem to have designated it as the natural point where a common desire for gain should bring into profitable relations the trading propensities of the people along the shores of the Mediterranean. During nearly two thousand years the situation was

1

0008D0D781E665AD

Fig. 5. A screenshot of the ground-truth annotation tool. In the application window, the right-hand side displays the baseline ToC with clickable (and editable) links. The left-hand side shows the current page and allows to navigate through the book. The JPEG image of each visited page is downloaded from the INEX server at www.booksearch.org.uk and is locally cached to limit bandwidth usage.

5 The Structure Extraction (SE) Task

The goal of the SE 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.

The first round of the structure extraction task, in 2008, ran as a pilot test and permitted to set up appropriate evaluation infrastructure, including guidelines, tools to generate ground-truth data, evaluation measures, and a first test set of 100 books. The second round was run both at INEX 2009 and at the International Conference on Document Analysis and Recognition (ICDAR) 2009 [2]. This round built on the established infrastructure with an extended test set of 1,000 digitized books.

Participants of the task were provided a sample collection of 1,000 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 distributed the PDF of books.

Participants could submit up to 10 runs, each containing the generated table of contents for the 1,000 books in the test set.

A total of 8 runs were submitted by 4 groups (1 run by Microsoft Development Center Serbia (MDCS), 3 runs by Xerox Research Centre Europe (XRCE), 1 run by Noopsis Inc., and 3 runs by the University of Caen).

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. This year, the annotation of a minimum number of books was required to gain access to the combined ground-truth set.

To make the creation of the ground-truth set for 1,000 digitized books feasible, we 1) developed a dedicated annotation tool, 2) made use of a baseline annotation as starting point and employed human annotators to make corrections to this, and 3) shared the workload across participants.

The annotation tool was specifically designed for this purpose and developed at the University of Caen, see Figure 5. The tool takes as input a generated ToC and allows annotators to manually correct any mistakes.

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. The ground-truth and the evaluation tool can be downloaded from http://users.info.unicaen.fr/~doucet/StructureExtraction2009/.

Table 4. Evaluation results for the SE task (complete ToC entries)

ParticipantID+RunID	Participant	Precision	Recall	F-measure
MDCS	MDCS	41.33%	42.83%	41.51%
XRCE-run1	XRCE	29.41%	27.55%	27.72%
XRCE-run2	XRCE	30.28%	28.36%	28.47%
XRCE-run3	XRCE	28.80%	27.31%	27.33%
Noopsis	Noopsis	9.81%	7.81%	8.32%
GREYC-run1	University of Caen	0.40%	0.05%	0.08%
GREYC-run2	University of Caen	0.40%	0.05%	0.08%
GREYC-run3	University of Caen	0.47%	0.05%	0.08%

The evaluation results are given in Table 4. The best performance ($F = 41.51\%$) was obtained by the MDCS group, who extracted ToCs by first recognizing the page(s) of a book that contained the printed ToC [5]. Noopsis Inc.

used a similar approach, although did not perform as well. The XRCE group and the University of Caen relied on title detection within the body of a book.

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, as well as benefiting from established methodology and guidelines for organising effective evaluation experiments.

ART is based on the evaluation experience of EBONI [6], 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, 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 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.

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.

The task has so far only attracted 2 groups, none of whom submitted any results at the time of writing.

7 Conclusions and plans

The Book Track this year has attracted considerable interest, cow from previous years. Active participation, however, remained a challenge for most of the participants. A reason may be the high initial setup costs (e.g., building infrastructure). Most tasks also require considerable planning and preparations, e.g., for setting up a user study. At the same time, the Structure Extraction task run at ICDAR 2009 (International Conference on Document Analysis and Recognition) has been met with great interest and created a specialist community. The search tasks, although explored real-world scenarios, were only tackled by a small set of groups. Since the evaluation of the BR and FBS tasks requires a great deal of effort, e.g., developing the assessment system and then collecting relevance judgements, we will be re-thinking the setup of these tasks for INEX 2010. For example, we plan to concentrate on more focused (narrow) topics for which only few pages in the corpus may be relevant. In addition, to improve the quality of the topics, we will look for ways to automate this process, hence also removing the burden from the participants.

To provide real value in improving the test corpus, we plan to run the SE task with the goal to use its results to convert the current corpus to an XML format that contains rich structural and semantic markup, which can then be used in subsequent INEX competitions.

Following the success of running the SE task in parallel at two forums, we will look for possible collaborators, both within and outside of INEX, to run ART next year.

Our plans for the longer term future are to work out ways in which the initial participation costs can be reduced, allowing more of the 'passive' participants to take an active role.

Acknowledgements

The Book Track is supported by the Document Layout Team of Microsoft Development Center Serbia, who developed the BookML format and a tool to convert books from the original OCR DjVu files to BookML.

References

1. Karen Coyle. Mass digitization of books. *Journal of Academic Librarianship*, 32(6):641–645, 2006.

2. Antoine Doucet, Gabriella Kazai, Bodin Dresevic, Aleksandar Uzelac, Bogdan Radakovic, and Nikola Todic. ICDAR 2009 Book Structure Extraction Competition. In *Proceedings of the Tenth International Conference on Document Analysis and Recognition (ICDAR'2009)*, pages 1408–1412, Barcelona, Spain, july 2009.
3. 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.
4. 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.
5. Aleksandar Uzelac, Bodin Dresevic, Bogdan Radakovic, and Nikola Todic. Book layout analysis: TOC structure extraction engine. In Shlomo Geva, Jaap Kamps, and Andrew Trotman, editors, *INEX*, Lecture Notes in Computer Science. Springer Verlag, Berlin, Heidelberg, 2009.
6. Ruth Wilson, Monica Landoni, and Forbes Gibb. The web experiments in electronic textbook design. *Journal of Documentation*, 59(4):454–477, 2003.