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
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. This paper provides an overview of the INEX 2009 Book Track. The main goal of the 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 status of the track.

1 Introduction

The INEX Book Track was launched in 2007, prompted by the numerous mass-digitization projects [1], e.g., the Million Book project[5], the Open Content Alliance[6], and the Google Books Library project[7]. As a result of these efforts the full texts of digitized books have become available by the thousands 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. [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 set up tasks to provide opportunities for investigating research questions around three broad topics:

– IR techniques for searching collections of digitized books,

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

– 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 of interest, 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 Focused Book Search (FBS) 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 current status 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 BR and FBS 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 84 organisations registered for the track (up from 54 in 2008, and 27 in 2007), of which 15 took part actively throughout the year (same as in 2008, and up from 9 in 2007), see Table 1.For the full list of participants, please refer to the INEX web site at `http://www.inex.otago.ac.nz/people/participants.asp`.

In total, 7 groups contributed 16 search topics with a total of 37 aspects, 4 groups submitted runs to the Structure Extraction task, 3 to the Book Retrieval task, and 3 groups submitted runs to the Focused Book Search task. Two groups participated in the Active Reading task, but did not submit results.

3 The Book Corpus

The track builds on a collection of 50,239 digitized out-of-copyright books[8], 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

[8] The collection, although in a different XML format, can also be found on the Internet Archive.

ID	Organisation	Topics	Runs	Assessed topics
6	University of Amsterdam	8, 11	2 BR, 4 FBS	
7	Oslo University College	1, 2	10 BR, 10 FBS	
14	University of California, Berkeley		10 BR, ART	
41	University of Caen	7, 9	3 SE	SE
43	Xerox Research Centre Europe		3 SE	SE
52	Kyungpook National University	3, 4	ART	
54	Microsoft Research Cambridge	10, 16		
78	University of Waterloo	5, 6	4 FBS	
86	University of Lugano	12, 13, 14, 15		
125	Microsoft Development 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

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)

the books also come with an associated MAchine-Readable Cataloging (MARC) record, which contains publication (author, title, etc.) and classification information.

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 (ocrml.xml file extension) 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>
 <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 texts, such as a table of contents. For example, the label attributes in the XML extract above indicate that a new chapter starts on page 1 (label="PT_CHAPTER") and that the section element is part of the main body of text on the page (label="SEC_BODY"). Other semantic units include headers (SEC_HEADER), footers (SEC_FOOTER), back of book index (SEC_INDEX), table of contents (SEC_TOC). 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 new chapter 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) Focused Book Search (FBS), 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).

A summary of the tasks, the test topics, and the online relevance assessment systemare described in the following sections. The relevance assessment collection phase is not yet underway, thus evaluation results will be published only after the INEX workshop.

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, for example to append at the end of an article, such as 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 16 test topics. 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[9]), ranked in order of estimated relevance.

A total of 22 runs were submitted by 3 groups (2 runs by University of Amsterdam (ID=6); 10 runs by University of California, Berkeley (ID=14); and 10 runs by Oslo University College (ID=7)), see Table 1.

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. 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, a maximum of 1000 books estimated relevant to the given aspect, 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 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.

4.3 Test Topics

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

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

Participants were encouraged to use Wikipedia at different stages when preparing topics. The intuition behind the introduction of Wikipedia is twofold. First, Wikipedia articles often contain a reading list of books relevant to the general topic of the article, while they also often cite related books relevant to a specific statement in the article. Thus, topics linked to Wikipedia articles have a real world application. Second, we anticipated that browsing through Wikipedia entries could provide participants with suggestions about topics and their specific aspects of interest, and at the same time provide them with insights and relevant terminology to be used for better searches and refinements that should lead to a better mapping between topics and collection.

Participants were asked to create and submit 2 topics, ideally with at least 2 aspects each, using an online Book Search system (see Section 4.4).

A total of 16 new topics (ID: 1-16), containing 37 aspects, were contributed by 7 participating groups (see Table 1). An example topic is shown in Figure 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 web service that allows participants to search, browse, read, and annotate the books of the test corpus. Screenshots of the assessment system 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. The competition involved reading books and marking relevant content inside the books for which assessors were rewarded points [4].

Based on what we learnt in 2008, we are modifying the game this year to consist of two separate stages: 1) In the first stage assessors are asked to find books relevant to the 16 topics and rank the top 10 most relevant books for each topic, then 2) in the second stage, assessors will again compete as explorers and reviewers, providing page level judgements for the 37 topic aspects.

We expect the assessment phase to start in mid December and conclude by the end of January 2010. Results of the evaluation will be published soon after the assessments have been collected.

5 The Structure Extraction (SE) Task

As in 2008, 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.

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 1000 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 1000 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 4. 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. 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 2. 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.

ParticipantID+RunID	Participant	F-measure
MDCS	MDCS	41.51%
XRCE-run2	XRCE	28.47%
XRCE-run1	XRCE	27.72%
XRCE-run3	XRCE	27.33%
Noopsis	Noopsis	8.32%
GREYC-run1	University of Caen	0.08%
GREYC-run2	University of Caen	0.08%
GREYC-run3	University of Caen	0.08%

Table 2. 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 [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, 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 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.

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.

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 a lot of interest and has grown considerably from last year. However, active participation remained a challenge for most of the participants who signed up to the track. A reason for this may be the high initial set up costs (e.g., building infrastructure to search books). Most tasks also require advance 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.

Our immediate plans are to commence the relevance assessment gathering stage for the BR and FBS tasks from mid December. We aim to have the evaluation results published by mid February 2010.

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 in 2008 was supported by the Document Layout Team of Microsoft Development Center Serbia. The team contributed to the track by pro-

viding 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. K. 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. R. Wilson, M. Landoni, and F. Gibb. The web experiments in electronic textbook design. *Journal of Documentation*, 59(4):454–477, 2003.

```
<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.

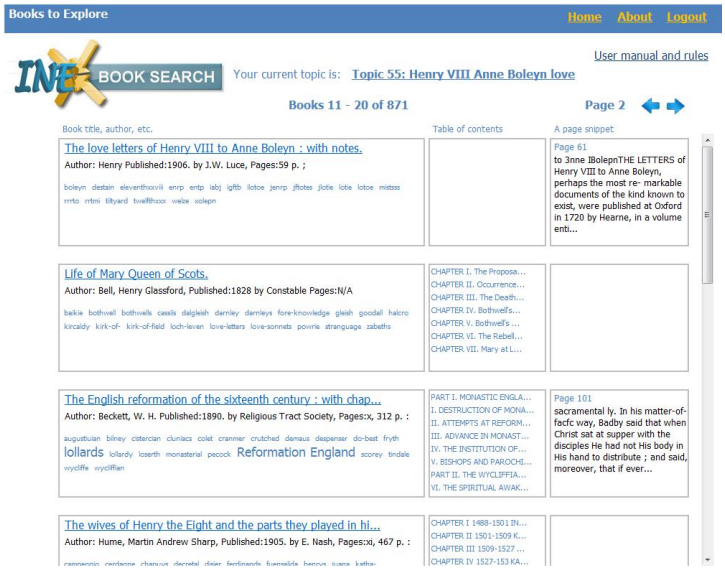

Fig. 2. Screenshot of the relevance assessment module of the Book Search system: List of books in the assessment pool for a selected topic.

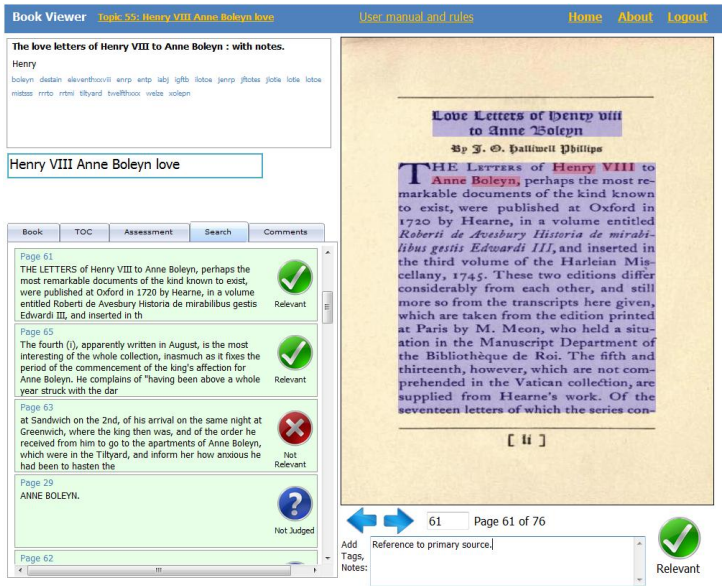

Fig. 3. Screenshot of the relevance assessment module of the Book Search system: Book Viewer window with Assessment tab showing, listing pooled pages to judge.

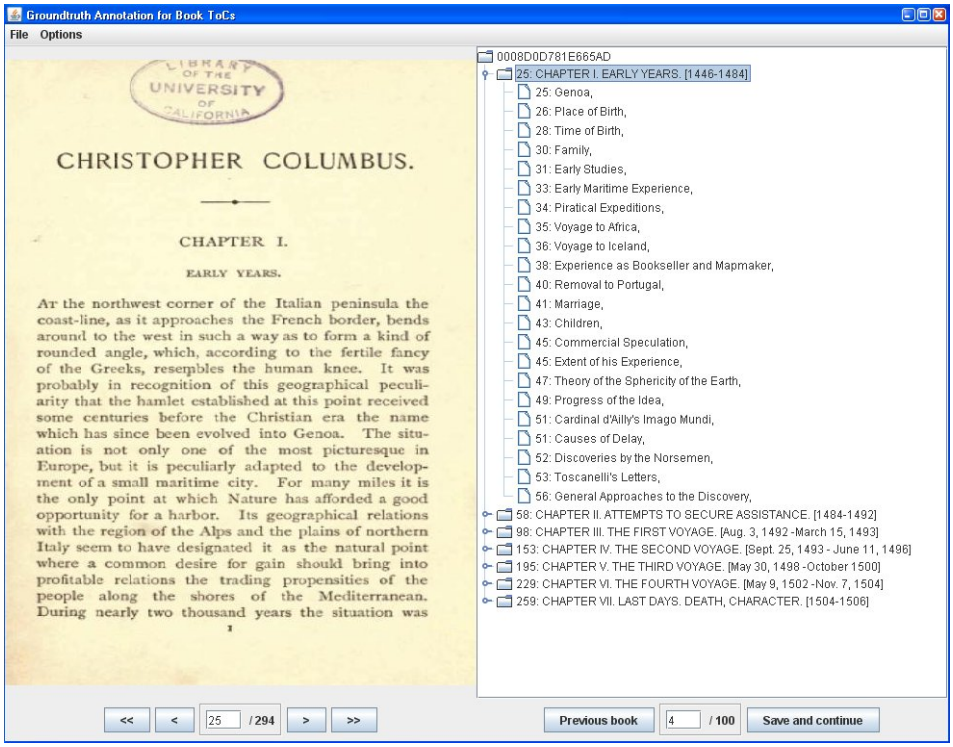

Fig. 4. A screenshot of the ground-truth annotation tool. In the application window, the right-hand side displays the baseline ToC with clickable (and editable) links. The left-hand side shows the current page and allows to navigate through the book. The JPEG image of each visited page is downloaded from the INEX server at `www.booksearch.org.uk` and is locally cached to limit bandwidth usage.