UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

Overview of the INEX 2010 Book Track: At the mercy of crowdsourcing

Kazai, G.; Koolen, M.; Doucet, A.; Landoni, M.

Publication date
2010

Document Version
Final published version

Published in
INEX 2010 workshop pre-proceedings

Link to publication

Citation for published version (APA):
Kazai, G., Koolen, M., Doucet, A., & Landoni, M. (2010). Overview of the INEX 2010 Book Track: At the mercy of crowdsourcing. In S. Geva, J. Kamps, R. Schenkel, & A. Trotman (Eds.), *INEX 2010 workshop pre-proceedings* (pp. 89-99). IR Publications. http://inex.otago.ac.nz/data/proceedings/INEX2010-preproceedings.pdf

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:09 Mar 2023

Overview of the INEX 2010 Book Track: At the Mercy of Crowdsourcing

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

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

Abstract. The goal of the INEX 2010 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 (Best Books to Reference) task aims at comparing traditional and book-specific retrieval approaches, 2) the Focused Book Search (Prove It) 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. [4].

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,

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

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

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

1. The Best Books to Referencel (BB) 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 Prove It (PI) 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 2010. 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 82 organisations registered for the track (compared with 84 in 2009, 54 in 2008, and 27 in 2007). As of the time of writing, we counted 10 active groups (compared with 16 in 2009, 15 in 2008, and 9 in 2007), see Table 1.

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. 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.

[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	19-20,22	2 BB, 4 PI	
7	Oslo University College	02-06	5 PI	
14	Uni. of California, Berkeley		4 BB	
54	Microsoft Research Cambridge	00-01,07-09,24-25		
86	University of Lugano	15-18,21,23		
98	University of Avignon		9 BB, 1 PI	
386	University of Tokyo			
662	Izmir Institute of Technology			
663	IIIT-H	10-14		
732	Wuhan University			

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) Best Books to Reference (BB), and 2) Prove It (PI). Both these tasks used the corpus described in Section 3, and shared the same set of topics (see Section 4.3).

4.1 The Best Books to Reference (BB) 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 83 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 only100 books (identified by their bookID), ranked in order of estimated relevance.

A total of 15 runs were submitted by 3 groups (2 runs by University of Amsterdam (ID=6); 4 runs by University of California, Berkeley (ID=14); and 9 runs by the University of Avignon (ID=98)), see Table 1.

4.2 The Prove It (PI) Task

The goal of this task was to investigate the application of focused retrieval approaches to a collection of digitized books. The scenario underlying this task is that of a user searching for specific information in a library of books that can provide evidence to confirm or reject a given factual statement. Users are assumed to view the ranked list of book parts, moving from the top of the list down, examining each result. No browsing is considered (only the returned book parts are viewed by users).

Participants could submit up to 10 runs. Each run could contain, for each of the 83 topics (see Section 4.3), a maximum of 1,000 book pages estimated relevant to the given aspect, ordered by decreasing value of relevance.

A total of 10 runs were submitted by 3 groups (4 runs by the University of Amsterdam (ID=6); 5 runs by Oslo University College (ID=7); and 1 run by the University of Avignon (ID=98)), see Table 1.

4.3 Topics

This year we explored the use of Amazon's Mechanical Turk (AMT) service to aid in the creation of topics for the test collection. This is motivated by the need to scale up the Cranfield method for constructing test collections where the significant effort required to create test topics and to collect relevance judgements is otherwise inhibiting. By harnessing the collective work of the crowds, crowdsourcing offers an increasingly popular alternative for gathering large amounts of data feasibly, at a relatively low cost and in a relatively short time. We are interested in using crowdsourcing to contribute to the building of a test collection for the Book Track, which has thus far struggled to meet this requirement by relying on its participants' efforts alone.

With this aim, we experimented with gathering topics both through Amazon's Mechanical Service and from the track participants. Our aim was to compare the quality of the collected topics and assess the feasibility of crowdsourcing topics (and relevance judgements later on). To this end, we first redefined the search tasks, simplifying them in order to make topic creation for them suitable as a Human Intelligent Task (HIT).

As mentioned already, in the Prove It task systems need to find evidence in books that can be used to either confirm or refute a factual statement given as the topic. In the Best Books task systems need to return the most relevant books on the general subject area of the topic. To collect the test topics for the two tasks, we created the following two HITs:

- Facts in books HIT (Book HIT): "Your task is to fnd a general knowledge fact that you believe is true in a book available at http://booksearch.org.uk. Both the fact and the book must be in English. The fact should not be longer than a sentence. For example, the fact that 'The first Electric Railway in London was opened in 1890 and run between the stations: Bank and Stockwell' can be found on page 187 of the book titled 'West London' by George Bosworth". Workers were asked to record the factual statement they found, the URL of the book containing the fact, and the page number.
- Facts in books and Wikipedia HIT (Wiki HIT): "Your task is to find a general knowledge fact that appears BOTH in a Wikipedia article AND in a

> book available at http://booksearch.org.uk. You can start either by finding a fact on Wikipedia first, then locating the same fact in a book, or you can start by finding a fact in a book and then in Wikipedia. For example, the Wikipedia page on Beethoven's Symphony No. 3 claims that 'Beethoven dedicated the symphony to Napoleon, but when Napoleon proclaimed himself emperor, Beethoven tore up the title'. Page 144 of the book titled Beethoven by Romain Rolland describes this very fact". Workers needed to record the factual statement, the URL and page number of the book where the fact is found, as well as the Wikipedia article's URL.

We created 10 Wiki HITs, paying $0.25 per HIT, and issued two batches of Book HITs, with 50 HITs in each batch, paying $0.10 per HIT in the first batch and $0.20 in the second batch. All 10 Wiki HITs were completed within a day, while only 32 Fact HITs were completed in 11 days out of the first batch. The second batch of 50 Book HITs was completed fully in 14 days. The average time required per Book HIT was 8 minutes in the first batch and 7 minutes in the second batch (hourly rate of $0.73 and $1.63, respectively), while Wiki HITs took on average 11 minutes to complete (hourly rate of $1.31). These statistics suggest that workers found the Wikipedia task more interesting, despite it taking longer. However, as we show later, the attractiveness of a HIT does not guarantee good quality topics.

At the same time, INEX participants were asked to create 5 topics each, 2 of which had to contain factual statements that appears both in a book and in Wikipedia. A total of 25 topics were submitted by 5 groups. Of these, 16 facts appear both in books and in Wikipedia.

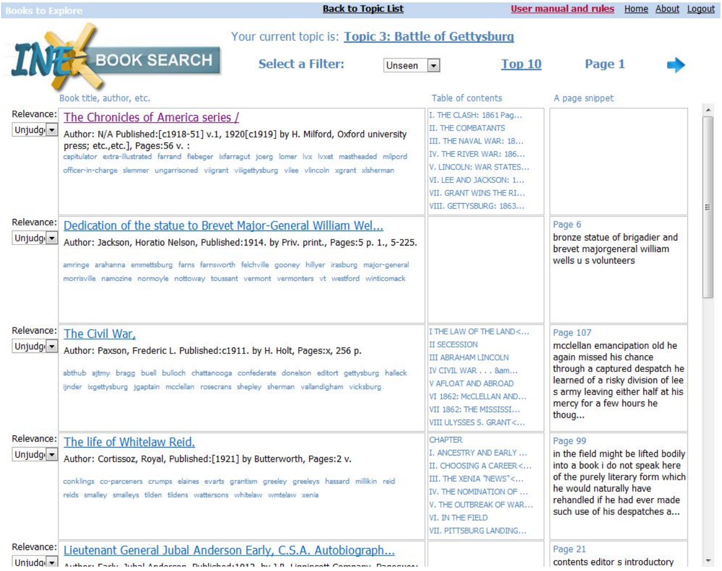
All collected topics were carefully reviewed and those judged suitable were selected into the set of test topics that is currently being used by the INEX Book Track. All topics contributed by INEX participants were selected, while filtering was necessary for topics created by AMT workers. Out of the 10 Wiki HITs, only 4 topics were selected. Of the 32 Book HITs in the first batch, 18 were acceptable, while 36 were selected from the 50 Book HITs in the second batch. HITs were rejected for a number of reasons: the information given was simply an extract from a book, rather than a fact (20), the fact was too specialised (5), or nonsensical (5), the HIT had missing data (3), or the worker submitted the example given in the task description (1). Of the total 58 accepted HITs, 18 had to be modified, either to rephrase slightly or to correct a date or name, or to add additional information. The remaining 40 HITs were high quality and reflecting real interest or information need.

From the above, it seems clear that crowdsourcing provides a suitable way to scale up test collection construction: MTurk workers contributed 58 topics, while INEX participants created only 25 topics. However, the quality of crowdsourced topics varies greatly and thus requires extra effort to weed out unsuitable submissions. We note that selecting workers based on their approval rate had a positive effect on quality: batch 2 of the Book HITs required workers to have a HIT approval rate of 95%. In addition, paying workers more also shows correlation with the resulting quality.

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 1 and 2.

In preparation for the relevance gathering stage, which will run in parallel, collecting judgements from INEX participants and from AMT workers, we simplified the assessment process from previous years.



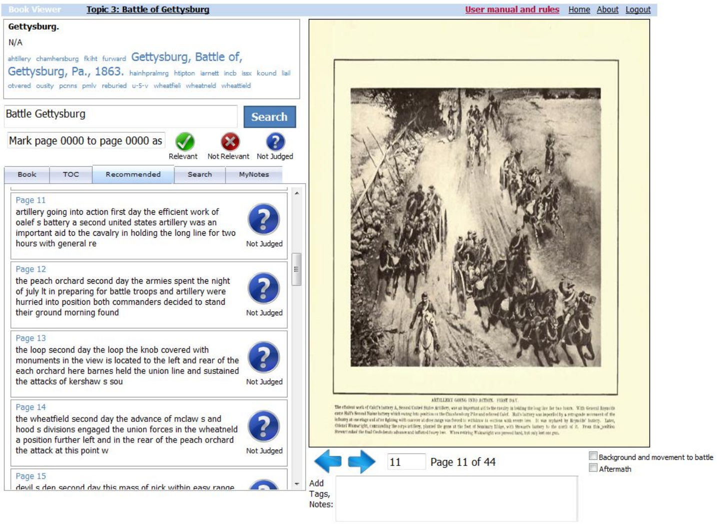
Your current topic is: Topic 3: Battle of Gettysburg

	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 viiigettysburg 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...		Page 21 contents editor s introductory

Fig. 1. 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

This is still in progress at the time of writing.



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

11 Page 11 of 44

□ Background and movement to battle
□ Aftermath

Fig. 2. 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.

4.6 Evaluation Measures and Results

We will report on these once sufficient amount of relevance labels have been collected.

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 2010 task was run as a follow-up of the conjoint INEX and ICDAR 2009 competition [2,3]. Participants were able to refine their approaches with the help of the groundtruth built in 2009.

Only one institution, the University of Caen, participated in this rerun of the 2009 task, while 2 contributed to the extension of the groundtruth data, since the University of Firenze joined the effort. The groundtruth now covers an additional 114 books and reaches a total of 641 annotated ToCs.

The performance of the 2010 run is given in Table 2 . A summary of the performance of the 2009 runs with the extended 2010 ground truth data is given in Table 3.

	Precision	Recall	F-measure
Titles	18.03%	12.53%	12.33%
Levels	13.29%	9.60%	9.34%
Links	14.89%	7.84%	7.86%
Complete except depth	14.89%	10.17%	10.37%
Complete entries	10.89%	7.84%	**4.86%**

Table 2. Score sheet of the run submitted by the University of Caen during the 2010 rerun of the SE competition 2009

RunID	Participant	F-measure (2010)	F-measure (2009)
MDCS	MDCS	43.39%	41.51%
XRCE-run2	XRCE	28.15%	28.47%
XRCE-run1	XRCE	27.52%	27.72%
XRCE-run3	XRCE	26.89%	27.33%
Noopsis	Noopsis	8.31%	8.32%
GREYC-run1	University of Caen	0.09%	0.08%
GREYC-run2	University of Caen	0.09%	0.08%
GREYC-run3	University of Caen	0.09%	0.08%

Table 3. Summary of performance scores for the 2009 runs with the extended 2010 groundtruth-rerun; results are for complete entries.

Naturally, the small increase in the size of the groundtruth does not make the results vary much (most of the groundtruth data was built for the 2009 experiments: 527 out 641 annotated books).

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 [5], 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

For the evaluation of our two search tasks (best books and prove it), we are currently collecting relevance assessments from INEX participants. This will be used as gold set for collecting judgements from workers on Amazon's Mechanical

Turk (AMT). Results will then be distributed around February 2011. The ART and SE tasks were offered as last year.

This year the focused search task (prove it) was based on factual statements for which systems were asked to find book pages that either confirmed or refuted the fact. 70

The SE task was run (though not advertised), using the same data set as last year. One institution participated and contributed additional annotations.

Unless we get a LOT more ACTIVE participants, 2011 will probably be the last year of the book track. We hope that with the burden of topic creation and relevance assessments removed, we will however get higher participation next year. We also plan to re-shape the research agenda by significantly increasing the size of the collection on the one hand, and by defining more challenging tasks that are focused on user interaction on the other hand, placing the ART in the centre.

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. Antoine Doucet, Gabriella Kazai, Bodin Dresevic, Aleksandar Uzelac, Bogdan Radakovic, and Nikola Todic. Setting up a competition framework for the evaluation of structure extraction from ocr-ed books. *International Journal on Document Analysis and Recognition*, pages 1–8, 2010.
4. 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.
5. Ruth Wilson, Monica Landoni, and Forbes Gibb. The web experiments in electronic textbook design. *Journal of Documentation*, 59(4):454–477, 2003.