2011 International Conference on Document Analysis and Recognition

ICDAR 2011 Book Structure Extraction Competition

Antoine Doucet
University of Caen Lower-Normandy
Campus Côte de Nacre
F-14032 Caen, France
antoine.doucet@unicaen.fr

Gabriella Kazai
Microsoft Research
7 JJ Thomson Ave
Cambridge, United Kingdom
gabkaz@microsoft.com

Jean-Luc Meunier
Xerox Research Center Europe
6 chemin de Maupertuis
F-38240 Meylan, France
jean-luc.meunier@xrce.xerox.com

***Abstract*—In this paper, we summarize the 2nd Book Structure Extraction competition run at ICDAR 2011. Its goal is to evaluate and compare automatic techniques for deriving structure information from digitized books, which could then be used to aid navigation inside the books. More specifically, the task that participants are faced with is to construct hyperlinked tables of contents for a collection of 1,000 digitized books. This paper reviews the setup of the competition, the book collection used in the task, and the measures used for the evaluation. It further presents the outcome of the competition: an additional ground truth of 513 book tables of contents, contributed by 6 institutions, and the result performance of the 4 participating research teams.**

***Keywords*-component; formatting; style; styling;**

I. Introduction

Mass-digitization projects, such as the Million Book project[1], efforts of the Open Content Alliance[2], and the digitization work of Google[3], are converting whole libraries by digitizing books on an industrial scale [1]. The process involves the efficient photographing of books, page-by-page, and the conversion of each page image into searchable text through the use of optical character recognition (OCR) software.

Current digitization and OCR technologies typically produce the full text of digitized books with only minimal structure information. Pages and paragraphs are usually identified and marked up in the OCR, but more sophisticated structures, such as chapters, sections, etc., are currently not recognized. In order to enable systems to provide users with richer browsing experiences, it is necessary to make available such additional structures, for example in the form of XML markup embedded in the full text of the digitized books.

The Book Structure Extraction competition aims to address this need by promoting research into automatic structure recognition and extraction techniques that could complement or enhance current OCR methods and lead to the availability of rich structure information for digitized books. Such structure information can then be used to aid user navigation inside books as well as to improve search performance [2].

The paper is structured as follows. We start by placing the competition in the context of the work conducted at the INEX evaluation forum (Section II). In Section III, we describe the setup of the competition, including its goals and the task that has been set for its participants. The book collection used in the task is detailed in Section IV. The ground truth creation process and its outcome are described in Section V, while the subsequent metrics and results are presented in Section VI. We conclude with a summary of the competition and our future plans in Section VII.

II. Background

Motivated by the need to foster research in areas relating to large digital book repositories, see e.g., [3], the Book Track was launched in 2007 as part of the Initiative for the Evaluation of XML retrieval (INEX)[4]. INEX was chosen as a suitable forum, as searching for information in a collection of books can be seen as one of the natural application areas of focused retrieval approaches, which have been investigated at INEX since 2002 [4]. In particular, focused retrieval over books presents a clear benefit to users, enabling them to gain direct access to parts of books (of potentially hundreds of pages in length) that are relevant to their information needs.

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. In 2007, the track focused on information retrieval (IR) tasks [5]. In 2008, two new tasks were introduced, including the book structure extraction task [6]. The structure extraction task was set up with the aim to evaluate automatic techniques for deriving structure from the OCR texts and page images of digitized books. The first round of the structure extraction task at INEX, in 2008, permitted to set up appropriate evaluation infrastructure, including guidelines, tools to generate ground truth data, evaluation measures, and a test set of 100 books.

[1] http://www.ulib.org/
[2] www.opencontentalliance.org/
[3] http://books.google.com/
[4] http://www.inex.cs.otago.ac.nz/

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.298

The second round was run both at INEX 2009 and at the International Conference on Document Analysis and Recognition (ICDAR) 2009. It allowed to extend the competition setup, develop an evaluation methodology [7] and to produce a ground truth of 527 manually annotated tables of contents.

The arrival of the competition at ICDAR 2009 triggered the expression of interest of 11 institutions, 7 of which participated in the evaluation phase, and 4 of which could build a system in due time.

The 2011 competition builds on the established infrastructure and a new test set of 1,000 digitized books. The main novelty since the first competition is that participants were given the possibility to train their systems on the 2009 ground truth data set.

III. Competition Setup

A. Goals

The goal of the book structure extraction competition is to test and compare automatic techniques for deriving structural information from digitized books in order to build hyperlinked tables of contents (ToC) that could then be used to navigate inside the books.

Example research questions whose exploration is facilitated by this competition include, but are not limited to:

- Can a ToC be extracted from the pages of a book that contain the actual printed ToC (where available) or could it be generated more reliably from the full content of the book?
- Can a ToC be extracted only from textual information or is page layout information necessary?
- What techniques provide reliable logical page number recognition and extraction and how logical page numbers can be mapped to physical page numbers?

B. Task Description

Given the OCR text and the PDF of a sample set of 1,000 digitized books of different genre and style, the task is to build hyperlinked tables of contents for each book in the test set. The OCR text of each book is stored in DjVu XML format (see Section IV). Participants may employ any techniques and can make use of either or both the OCR text and the PDF images to derive the necessary structure information and generate the ToCs.

Participating systems were asked to output an XML file (referred to as a "run") that contains the generated hyperlinked ToC for each book in the test set. The document type definition (DTD) for the XML output is given in Figure 1.

Participants were invited to submit up to 10 runs, each run containing the ToC for all 1,000 books. The ToCs created by participants are then compared to a manually built ground truth.

```
<!ELEMENT bs-submission
  (source-files, description, book+)>
<!ATTLIST bs-submission
  participant-id  CDATA  #REQUIRED
  run-id  CDATA  #REQUIRED
  task  (book-toc) #REQUIRED
  toc-creation (automatic |
    semi-automatic) #REQUIRED
  toc-source (book-toc | no-book-toc |
    full-content | other) #REQUIRED>
<!ELEMENT source-files EMPTY>
<!ATTLIST source-files
  xml  (yes|no) #REQUIRED
  pdf  (yes|no) #REQUIRED>
<!ELEMENT description (#PCDATA)>
<!ELEMENT book (bookid, toc-entry+)>
<!ELEMENT bookid (#PCDATA)>
<!ELEMENT toc-entry(toc-entry*)>
<!ATTLIST toc-entry
  title (#PCDATA) #REQUIRED
  page  (#PCDATA) #REQUIRED>
```

Figure 1. DTD of the XML output ("run") that participating systems are expected to submit to the competition, containing the generated hyperlinked ToC for each book in the test set.

C. Participating Organizations

Following the call for participation issued in January 2011, 11 organizations registered. They are listed in Table I. Several organizations have expressed interest but renounced participation due to time constraints. Of the 11 organizations that signed up, 5 dropped out, that is, they neither submitted runs, nor participated in the ground truth annotation process.

Four teams submitted runs, while two contributed to the ground truth creation even though they were not able to submit runs this year. They expressed their intent to participate in forthcoming rounds of the competition.

Contributing to ground truth creation was the sole condition upon which access to the compiled ground truth is granted. This condition was imposed with the aim to incentivize participants and increase the number of fully annotated ToCs, which in turn would lead to more reliable evaluation results. The observed community interest is a good indicator of the relevance of this new competition, and an encouragement to pursue it in coming years, as was already requested by several of the participants.

IV. Book Collection

The corpus of the INEX book track contains a collection of 50,239 digitized out-of-copyright books, provided by Microsoft Live Search and the Internet Archive [6].

The set of books used in the book structure extraction competition comprises 1,000 books selected from the INEX book corpus. It contains books of different genre, among which history books, biographies, literary studies, religious texts and teachings, reference works, encyclopedias, essays, proceedings, novels, and poetry.

Organization	Submitted runs	Ground truthing
IBM Tokyo (Japan)	0	n
INRIA (France)	0	n
Microsoft Development Center (Serbia)	1	y
Nankai University (PRC)	4	y
NII Tokyo (Japan)	0	y
Oslo University College (Norway)	0	n
Queensland University of Technology (Australia)	0	n
Thiagarajar College of Engineering (India)	0	n
University of Caen (France)	3	y
University of Innsbrück (Austria)	0	y
Xerox Research Centre Europe (France)	2	y

Table I
REGISTERED PARTICIPANTS AND ACTIVITY.

To facilitate the separate evaluation of techniques based on the analysis of book pages that contain the printed ToC versus techniques that are based on deriving structure information from the full book content, we selected 200 books into the total 1,000 that do not contain a printed ToC. To do this, we used a tool developed by Microsoft Development Center Serbia, which converts the DjVu XML OCR text into BookML, a format in which ToC pages are explicitly marked up. We then selected a set of 800 books with detected ToC pages, and a set of 200 books without detected ToC pages into the full test set of 1,000 books. We note that this ratio of 80:20% of books with and without printed ToCs is proportional to that observed over the whole INEX corpus of 50,239 books.

The uncompressed size of the structure extraction corpus is around 33GB.

Each book is provided in two different formats: portable document format (PDF), and DjVu XML containing the OCR text and basic structure markup as illustrated below:

```
<DjVuXML>
 <BODY>
  <OBJECT data="file..." [...]>
   <PARAM name="PAGE" value="[...]">
   [...]
   <REGION>
    <PARAGRAPH>
     <LINE>
      <WORD coords="[...]"> Moby </WORD>
      <WORD coords="[...]"> Dick </WORD>
      <WORD coords="[...]"> Herman </WORD>
      <WORD coords="[...]"> Melville </WORD>
      [...]
     </LINE>
     [...]
    </PARAGRAPH>
   </REGION>
   [...]
  </OBJECT>
  [...]
 </BODY>
</DjVuXML>
```

An <OBJECT> element corresponds to a page in a digitized book. A page counter, corresponding to the physical page number, is embedded in the @value attribute of the <PARAM> element, which has the @name="PAGE" attribute. The logical page numbers (as printed inside the book) can be found (not always) in the header or the footer part of a page. Note, however, that headers/footers are not explicitly recognized in the OCR, i.e., the first paragraph on a page may be a header and the last one or more paragraphs may be part of a footer. Depending on the book, headers may include chapter/section titles and logical page numbers (although due to OCR error, the page number is not always present).

Inside a page, each paragraph is marked up. It should be noted that an actual paragraph that starts on one page and ends on the next is marked up as two separate paragraphs within two page elements. Each paragraph element consists of line elements, within which each word is marked up separately. Coordinates that correspond to the four points of a rectangle surrounding a word are given as attributes of word elements.

V. GROUND TRUTH CREATION

The process of manually building the ToC of a book is very time-consuming. Hence, as in 2009, to make the creation of the ground truth for 1,000 digitized books feasible, we resorted to 1) facilitating the annotation task with a dedicated tool, 2) making use of a baseline annotation as starting point and employing human annotators to make corrections, and 3) sharing the workload.

An annotation tool was specifically designed for this purpose and developed at the University of Caen. The tool takes as input a generated ToC and allows annotators to manually correct any mistakes (see [7] for further details on the tool).

Naturally, to compare the submitted runs to a ground truth necessitates the construction of such a ground truth. Given the burden that this task may represent, we chose to split it between participating institutions, and rather than forcing participants to do annotations (which may trigger hasty and careless work), we encouraged them with an incentive: we

limited the distribution of the resulting ground truth set to those who contributed a minimum number of annotations.

Using the submitted ToCs as starting points of the annotation process greatly reduces the required effort, since only the missing entries need to be entered. Others simply need to be verified and/or edited, although even these often require annotators to skim through the whole book.

An important side-effect of making use of a baseline ToC is that this may trigger a bias in the ground truth, since annotators may be influenced by the ToC presented to them. To reduce this bias (or rather, to spread it among participating organizations), we chose to take the baseline annotations from the participants' submissions in equal shares.

Finally, the annotation effort was shared among all participants. Teams who submitted runs were required to contribute a minimum of 50 books, while others were required to contribute a minimum of 100 books (20% of which are books without a printed ToC). The created ground truth was made available to all contributing participants for use in future evaluations.

A. Collected Ground Truth Data

6 teams participated in the ground truth annotation process, 2 of which did not submit runs.

This joint effort resulted in a set of 649 annotated books. To ensure the quality and internal consistency of the collected annotations, each of the annotated ToCs were reviewed by the organizers, and any incorrect TOCs were removed. Any ToC with annotation errors were then removed. Errors were most of the time due to failure to follow the annotation guidelines or producing incomplete annotations.

Following this cleansing step, 513 annotated books remain to form the ground truth file that was distributed to each contributing organization.

B. Freely Available Ground Truth.

While the 2011 ground truth is only accessible to institutions that participated to the annotation process (see Figure I), it was decided to release the the ground truth set built in 2009. This is meant to facilitate the participation of other institutions in the future. This set contains 527 books and ToCs, and was fully described in an IJDAR article [7].

VI. Results

The book structure extraction competition relies on two complementary metrics: a title-based measure and a link-based measure. Both of them were extensively described in earlier papers ([7], [8]), and the corresponding software is available for download on the competition's web site[5]. For the purpose of evaluation, both techniques compare participant submissions to the ground truth. The fundamental difference is that the first measure does this primarily based on the similarity of titles, while the latter is based on equivalent page links (links to the same physical page).

[5]http://www.info.unicaen.fr/~doucet/StructureExtraction/

	Precision	Recall	F-measure
Titles	52.44%	56.96%	53.11%
Levels	43.82%	47.00%	44.17%
Links	48.09%	52.07%	48.72%
Complete entries	40.40%	43.17%	**40.75%**

Table II
AN EXAMPLE SCORE SHEET SUMMARIZING THE TITLE-BASED PERFORMANCE OF THE "MDCS" RUN.

RunID	Participant	F-measure
MDCS	MDCS	40.75%
Nankai-run1	Nankai U.	33.06%
Nankai-run4	Nankai U.	33.06%
Nankai-run2	Nankai U.	32.46%
Nankai-run3	Nankai U.	32.43%
XRCE-run1	XRCE	20.38%
XRCE-run2	XRCE	18.07%
GREYC-run2	University of Caen	8.99%
GREYC-run1	University of Caen	8.03%
GREYC-run3	University of Caen	3.30%

Table III
SUMMARY OF TITLE-BASED PERFORMANCE SCORES FOR THE STRUCTURE EXTRACTION COMPETITION 2011 (F-MEASURE FOR COMPLETE ENTRIES).

A. Official title-based measure

The title-based evaluation works as follows; to compare the ToC of a submission to that of the ground truth, it primarily searches the ground truth for entries with a "sufficiently similar" title. Whether two titles are "sufficiently similar" is measured in terms of string-edit distance, so as to take into account the possible variations of a same title, for instance, within the ToC ("3 His Birth and First Years") and within the book content ("Chapter 3: His Birth and First Years").

Once these **matching titles** have been collected, it is possible to check whether the ToC entry has a **matching link** (if the ground truth contains a matching title linking to the same physical page as the ToC entry) and whether it has a **matching depth level** (if the ground truth contains a matching title at the same depth level as that of the ToC entry). A ToC entry is a full match (**complete entry**) when there is a matching title in the groundtruth that links to the physical page and lies at the same depth.

Whether the ToC entries of a book match the groundtruth or not allows to compute recall, precision and F-measure values for each submission. These book-wise values are then averaged over the full ground truth set to produce score sheets such as the one shown in Table II.

A summary of the performance of all the submitted runs, based on F-measure for complete entries (see entry in bold in Table II) is given in Table III. The score sheets

RunID	Precision	Recall	F-measure
MDCS	64.5%	70.2%	65.1%
Nankai-run1	67.6%	67.4%	63.2%
Nankai-run4	67.6%	67.4%	63.2%
Nankai-run2	66.0%	60.3%	59.8%
Nankai-run3	65.8%	60.3%	59.8%
XRCE-run2	75.9%	55.1%	58.1%
XRCE-run1	79.3%	52.5%	57.6%
GREYC-run1	65.2%	49.9%	50.7%
GREYC-run2	65.2%	49.9%	50.7%
GREYC-run3	32.5%	24.5%	24.4%

Table IV
PERFORMANCE SCORES FOR THE STRUCTURE EXTRACTION COMPETITION 2011 BASED ON THE XRCE LINK-BASED METRICS.

corresponding to each of the runs are available online[6].

B. Alternative link-based measure

Since 2009, we have also relied on a complementary measure introduced by Meunier and Déjean [8]. The so called "XRCE link-based measure" aims to take into account the quality of the links directly, rather than conditionally to the title's validity.

The XRCE link-based measure permits to evaluate the performance of systems by matching ToC entries primarily based on links rather than titles. The corresponding results are given in Table IV. As it can be seen, the results improve as possible errors in the titles no longer lead to whole ToC entries being discounted.

As in the past, the best performing methods are those that focus specifically on ToC pages. For both measures, we can observe that, as in 2009, Microsoft Development Center Serbia produced the best run. Nankai University, although a newcomer, was second best.

VII. SUMMARY AND FUTURE PLANS

After its first two rounds at ICDAR, the book structure extraction competition gathered in a collaborative effort the ground truth for the tables of contents of a total of 1,040 books. This is thanks to the joint effort of a total of 10 institutions who participated to the ground truth creation in the 2009 and 2011 competitions.

This joint effort is only one expression of the renewed interest in the competition, already expressed by several participants. We therefore plan to continue running the competition in the coming years.

In future years, we aim to investigate the usability of the extracted ToCs. In particular, we will explore the use of qualitative evaluation measures in addition to the current precision/recall measures. This would enable us to better understand what properties make a ToC useful and which are important to users engaged in reading or searching. Such insights are expected to contribute to future research into providing better navigational aids to users of digital book repositories.

In spite of the tremendous efforts of participants to build the ground truth, we would like to experiment with crowdsourcing methods in the future. This may offer a natural solution to the evaluation challenge posed by the massive data sets handled in digitized libraries.

The experience of one of the organizers in using crowdsourcing for relevance assessments over the same data set suggests the feasibility of using crowdsourcing reliably in high cognitive tasks such as that of labelling ToCs [9].

[6]http://www.info.unicaen.fr/~doucet/StructureExtraction/2011/

REFERENCES

[1] K. Coyle, "Mass digitization of books," *Journal of Academic Librarianship*, vol. 32, no. 6, pp. 641–645, 2006.

[2] R. van Zwol and T. van Loosbroek, "Effective use of semantic structure in XML retrieval," in *ECIR*, ser. Lecture Notes in Computer Science, G. Amati, C. Carpineto, and G. Romano, Eds., vol. 4425. Springer, 2007, pp. 621–628.

[3] P. Kantor, G. Kazai, N. Milic-Frayling, and R. Wilkinson, Eds., *BooksOnline '08: Proceeding of the 2008 ACM Workshop on Research Advances in Large Digital Book Repositories*. New York, NY, USA: ACM, 2008.

[4] S. Geva, J. Kamps, and A. Trotman, Eds., *Advances in Focused Retrieval: 7th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2008)*, ser. Lecture Notes in Computer Science, vol. 5613. Springer Verlag, Berlin, Heidelberg, 2009.

[5] G. Kazai and A. Doucet, "Overview of the INEX 2007 Book Search Track (BookSearch'07)," *ACM SIGIR Forum*, vol. 42, no. 1, pp. 2–15, 2008.

[6] G. Kazai, A. Doucet, and M. Landoni, "Overview of the INEX 2008 Book Track," in *INEX*, ser. Lecture Notes in Computer Science, S. Geva, J. Kamps, and A. Trotman, Eds., vol. 5613. Springer Verlag, Berlin, Heidelberg, 2009.

[7] A. Doucet, G. Kazai, B. Dresevic, A. Uzelac, B.Radakovic, and N. Todic, "Setting up a competition framework for the evaluation of structure extraction from ocr-ed books," *International Journal of Document Analysis and Recognition (IJDAR), Special Issue on Performance Evaluation of Document Analysis and Recognition Algorithms.*, vol. 14, no. 1, pp. 45–52, 2011.

[8] H. Déjean and J.-L. Meunier, "Reflections on the inex structure extraction competition," in *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*, ser. DAS '10. New York, NY, USA: ACM, 2010, pp. 301–308. [Online]. Available: http://doi.acm.org/10.1145/1815330.1815369

[9] G. Kazai, J. Kamps, M. Koolen, and N. Milic-Frayling, "Crowdsourcing for book search evaluation: Impact of quality on comparative system ranking," in *Proceedings of the 34th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, New York NY, 2011.