Overview of the ICDAR 2013 Competition on Book Structure Extraction

Antoine Doucet, Gabriella Kazai, Sebastian Colutto, Günter Mühlberger

Overview of the ICDAR 2013 Competition on Book Structure Extraction

Antoine Doucet
University of Normandy – Unicaen
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

Sebastian Colutto and Günter Mühlberger
University of Innsbruck
Innrain 52
6020 Innsbruck, Austria
first.last@uibk.ac.at

***Abstract*—This paper summarizes the 3rd Book Structure Extraction competition that was run at the ICDAR 2013. 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. The main novelty of the 2013 competition is that we were able to rely on an external provider for the ground truthing phase, hence granting the consistency of the evaluation. In addition, this allowed us to nearly double the number of annotated books from the 1,040 books annotated in 2009 and 2011 to over 2,000 books. The paper further presents the resulting performance of the 6 participating research teams, and briefly summarizes their approaches.**

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. The evaluation metrics used and the final results, alongside brief descriptions of the participants' approaches, are presented in Section VI. We conclude with a summary of the competition and plans for the future 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, founded in 2002, is an evaluation forum that investigates focused retrieval approaches [4], where structure information is used to aid the retrieval of parts of documents, relevant to a search query. Focused retrieval over books presents a clear benefit to users, enabling them to gain direct access to those parts of books (of potentially hundreds of pages in length) that are relevant to their information needs.

The overarching 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 2008, the book structure extraction task [5], was introduced as part of the INEX Book Track. The 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 2008, allowed us to set up appropriate evaluation infrastructure, including guidelines, tools to generate ground truth data, evaluation measures, and a test set of 100 books. The second and third rounds were run both at INEX and at ICDAR in 2009 and 2011. They allowed us to extend the competition setup, develop an evaluation methodology [6] and to produce a ground truth for a total of 1,037 manually annotated ToCs.

The arrival of the competition at ICDAR 2009 triggered the expression of interest of 11 institutions, 7 of which participated

[1] http://www.ulib.org/

[2] www.opencontentalliance.org/

[3] http://books.google.com/

[4] https://inex.mmci.uni-saarland.de/tracks/books/

in the evaluation phase, and 4 of which had a ToC extraction system built in time to submit runs. Similar numbers were observed in 2011, with again 11 institutions expressing interest, most of them new to the competition. Six of those institutions participated, with 4 of them submitting runs. The main novelty in 2011 was the possibility for participants to train their systems using the 2009 ground truth data set.

The 2013 competition built on the established infrastructure and a new test set of 1,000 digitized books. However, a key stage of the evaluation, namely the process by which the ground truth ToCs are obtained, has changed this year. Thanks to the involvement of the University of Innsbruck, the process was revamped, see Section III, alleviating any possibilities of evaluation bias and removing the burden on the participants. Unlike previous years, the ground truth data can be freely distributed this year to the wider research community (in previous years, the main incentive to contribute to the collaborative ground truth creation process was the exclusive access to the data for a limited time).

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 can logical page numbers be mapped to physical page numbers?

B. Task Description

As in previous years, 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 ToCs for each book in the test set. The OCR text of each book is stored in 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 needed to output an XML file (referred to as a "run") containing 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 were then compared to the ground truth ToCs during evaluation (see Sections V and VI).

```
<!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>
```

Fig. 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 2013, 9 organizations registered, see Table I. Several organizations have expressed interest but renounced participation due to time constraints. Of the 9 organizations that signed up, 6 submitted runs. This promising increase in active participants (6 out of 9), compared with previous years (4 out of 11), is likely a result of available training data[5] and the removed obligation on creating ground truth ToCs.

The increase in the number of participants, and the fact that half of the competition's participants submitted their first runs in 2013 provide support for continuing the competition in coming years.

IV. Book Collection

The corpus of the INEX book track contains a collection of 50,239 digitized out-of-copyright books, provided by Microsoft and the Internet Archive [5].

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, including history books, biographies, literary studies, religious texts and teachings, reference works, encyclopedias, essays, proceedings, novels, and poetry.

To facilitate the separate evaluation of techniques that are 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 books' 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

[5] https://doucet.users.greyc.fr/StructureExtraction/training/

Organization	Submitted runs	First registration	First submission
Elsevier	0	2013	-
EPITA (France)	1	2013	2013
INRIA (France)	0	2011	-
Microsoft Development Center (Serbia)	1	2009	2009
Nankai University (PRC)	1	2011	2011
NII Tokyo (Japan)	0	2011	-
University of Caen (France)	5	2009	2009
University of Innsbruck (Austria)	1	2011	2013
University of Würzburg (Germany)	1	2013	2013

TABLE I. REGISTERED PARTICIPANTS AND ACTIVITY.

ToC pages, and a set of 200 books without detected ToC pages into the full 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 25GB.

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>
```

The DjVu format is a physical description of the digitized book. An <OBJECT> element corresponds to a page of a digitized book. A page counter, corresponding to the physical page number, is embedded in the @value attribute of the <PARAM> element, having the @name="PAGE" attribute. The logical page numbers, corresponding to those printed inside the book, form part of the content and can often be found inside the first or the last paragraphs of a page. Depending on the book, these paragraphs may include chapter/section titles as well as the logical page numbers (although due to OCR error, the page number is not always present or correct).

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 object 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 construction of a ground truth for Optical Character Recognition tasks is usually a very time consuming and intense process. Many great tools like the freely available Aletheia [7] exist for this purpose. For the task of building the ground truth ToCs, however, requires a specialized tool that combines *structure analysis* and *correction*. To this end, the FEP (Functional Extension Parser) was developed by the University of Innsbruck during the 7^{th} Framework Programme (FP7) project IMPACT[6] (Improving Access To Text). Using the FEP, the work-flow for the ground truth creation process involves the following steps.

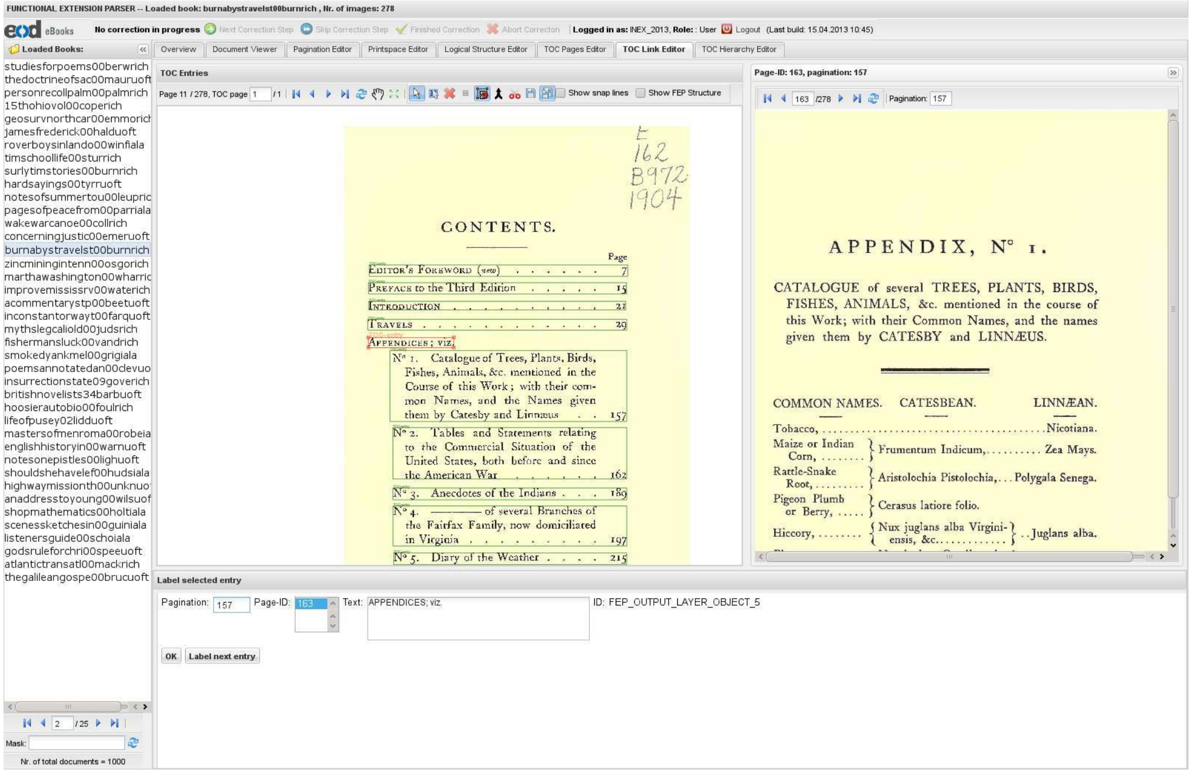
a) Starting phase.: First, all books, in PDF and DjVu XML format, were uploaded to the central database at the University of Innsbruck. An automatic structure analysis step, including the creation of the ToCs, was then carried out using an approach based on fuzzy learning in rule systems and context-free-grammars, as described by Gander et al. [8]. Its results could then be viewed and corrected using a web application that was developed in the course of the IMPACT project. The web application was written using the Google Web Toolkit (GWT) and thus consists of Javascript code on the client side and Java-Servlets on server side.

b) Manual correction.: The correction of the ToCs consists of three sub-steps: in the first step, the user has to select the images where the ToC pages are printed (if they exist). If no ToC page exists, the user needs to browse through the book, and mark up every desired ToC entry. Because the whole book needs to be browsed exhaustively, this process is the most demanding. If ToC pages were detected, all entries found are first segmented by a covering rectangle which can be constructed by the corresponding tool in the editor. At this stage, the user also edits the underlying text of the ToC entry (title and the page number). A screenshot of the FEP web application at this stage can be seen in Figure 2. In the final step of the correction, the hierarchy of all entries can be changed using drag and drop operations on the ToC entries.

To construct an unbiased ground truth of high quality for all participants, this year, unlike as in the previous competitions, the ground truth was not built collaboratively by the participants themselves [9], [10], but by an independent third party, a Vietnamese digitizing company called DIGI-TEXX[7], who were subcontracted by the University of Innsbruck. Their task was to perform the manual correction described above, using the FEP web application. This relieved participants from

[6] http://www.impact-project.eu/

[7] http://www.digi-texx.com.vn



FUNCTIONAL EXTENSION PARSER -- Loaded book: burnabystravelst00burnrich , Nr. of images: 278

No correction in progress

Logged in as: INEX_2013, Role: : User (Last build: 15.04.2013 10:45)

Loaded Books:

studiesforpoems00berwrich
thedoctrineofsac00mauruoft
personrecollpalm00palmrich
15thohiovol00coperich
geosurvnorthcar00emmorich
jamesfrederick00halduoft
roverboysinlando00winfiala
timschoollife00sturrich
surlytimstories00burnrich
hardsayings00tyrruoft
notesofsummertou00leupric
pagesofpeacefrom00parriala
wakewarcanoe00collrich
concerningjustic00emeruoft
burnabystravelst00burnrich
zincminingintenn00osgorich
marthawashington00wharric
improvemississrv00waterich
acommentarystp00beetuoft
inconstantorwayt00farquoft
mythslegcaliold00judsrich
fishermansluck00vandrich
smokedyankmel00grigiala
poemsannotatedan00clevuo
insurrectionstate09goverich
britishnovelists34barbuoft
hoosierautobio00foulrich
lifeofpusey02lidduoft
mastersofmenroma00robeia
englishhistoryin00warnuoft
notesonepistles00lighuoft
shouldshehavelef00hudsiala
highwaymissionth00unknuo
anaddresstoyoung00wilsuof
shopmathematics00holtiala
scenessketchesin00guiniala
listenersguide00schoiala
godsruleforchri00speeuoft
atlantictransatl00mackrich
thegalileangospe00brucuoft

2 / 25

Mask:

Nr. of total documents = 1000

TOC Entries

Page 11 / 278, TOC page 1 / 1

E
162
B972
1904

CONTENTS.

Page
EDITOR'S FOREWORD (*new*) 7
PREFACE to the Third Edition 15
INTRODUCTION 21
TRAVELS 29
APPENDICES; viz.
Nº 1. Catalogue of Trees, Plants, Birds, Fishes, Animals, &c. mentioned in the Course of this Work; with their common Names, and the Names given them by Catesby and Linnæus . . . 157
Nº 2. Tables and Statements relating to the Commercial Situation of the United States, both before and since the American War 162
Nº 3. Anecdotes of the Indians 189
Nº 4. ———— of several Branches of the Fairfax Family, now domiciliated in Virginia 197
Nº 5. Diary of the Weather 215

Page-ID: 163, pagination: 157

163 /278 Pagination: 157

APPENDIX, Nº 1.

CATALOGUE of several TREES, PLANTS, BIRDS, FISHES, ANIMALS, &c. mentioned in the course of this Work; with their Common Names, and the names given them by CATESBY and LINNÆUS.

COMMON NAMES.	CATESBEAN.	LINNÆAN.
Tobacco, ...		Nicotiana.
Maize or Indian Corn, ...	Frumentum Indicum, ...	Zea Mays.
Rattle-Snake Root, ...	Aristolochia Pistolochia, . . .	Polygala Senega.
Pigeon Plumb or Berry, ...	Cerasus latiore folio.	
Hiccory, ...	Nux juglans alba Virginiensis, &c.	. . Juglans alba.

Label selected entry

Pagination: 157 Page-ID: 163 Text: APPENDICES; viz. ID: FEP_OUTPUT_LAYER_OBJECT_5

Fig. 2. FEP web application for the ground truth construction process. The tab *TOC Link Editor* is selected indicating that the user is able to add and edit links on table of contents pages which are indicated by the segmenting rectangles. On the right side the linked page of the selected entry is shown and the bottom widget is used to edit the text of an entry and the page it is linking to.

the burden of ground truth creation and ensured unbiased evaluation.

c) Outcome.: During the ground truth creation process, two books were removed from the collection of 1,000 books: one book in Chinese (named *beikokunouraomo00miyarich* with id *5E7142B1957E157F*) and one duplicate (*romeinireland00mccauoft*, with id *0FCBA37050AF4762*). 31 further books contained no ToC entry. The ground truth was built for the remaining 967 books of the evaluation set, almost doubling the total size of the evaluation set accumulated since 2009. To ensure the quality of the data set, all of the books went through an second-level verification by DIGI-TEXX, following the initial manual correction.

In previous competitions, access to the ground truth was temporarily restricted to institutions that participated to its construction. Another benefit of the ground truth being created by a third party is that, for the first time in 2013, the ground truth will be immediately freely available on the competition's web site [8] to be used for research purposes by the wider research community.

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 ([6], [11]), and the corresponding software is available for download on the competition's web site[9]. Both techniques compare participants' submitted runs to the ground truth. The fundamental difference is that the first measure does this primarily based on the similarity of the titles in the ToC entries, while the latter is based on equivalent page links (links to the same physical page).

A. *Official title-based measure*

The title-based evaluation compares the ToC of a submission to that of the ground truth by first matching the ToC entry titles in the runs to the ground truth. This is done by calculating edit distance between two entry titles, thus allowing for possible variations, for instance, "3 His Birth and First Years" vs. "Chapter 3: His Birth and First Years".

For each ToC entry title, the depth level and the page number references are then checked (i.e., whether the ground truth ToC entry is at the same level in the ToC hierarchy and whether it links to the same physical page as the ToC entry being evaluated). A ToC entry is considered a full match (or **complete entry**) when both the depth and page number information are correct as per the ground truth data.

Based on the number of matching ToC entries, recall, precision and the F-measure can be computed for each submission. These per-book values are then averaged over the full ground truth set, producing score sheets such as the one shown in Table II.

[8] https://doucet.users.greyc.fr/StructureExtraction/training/

[9] https://doucet.users.greyc.fr/StructureExtraction/

	Precision	Recall	F-measure
Titles	58.38%	63.06%	59.59%
Levels	46.42%	50.01%	47.36%
Links	53.48%	57.49%	54.54%
Complete entries	42.77%	45.92%	**43.61%**

TABLE II. An example score sheet summarizing the title-based performance of the "MDCS" run.

RunID	Participant	F-measure
MDCS	MDCS	43.61%
Nankai	Nankai U.	35.41%
Innsbruck	Innsbruck U.	31.34%
Würzburg	Würzburg U.	19.61%
Epita	Epita	14.96%
GREYC-run-d	University of Caen	8.81%
GREYC-run-c	University of Caen	7.91%
GREYC-run-a	University of Caen	6.21%
GREYC-run-e	University of Caen	4.71%
GREYC-run-b	University of Caen	3.79%

TABLE III. Summary of title-based performance scores for the Structure Extraction competition 2013 (F-measure for complete entries).

A summary of the performance of all the submitted runs, based on the F-measure calculated for complete entries only is given in Table III. The score sheets corresponding to each of the runs are available online[10].

B. Alternative link-based measure

In 2009, a complementary measure was introduced by Meunier and Déjean [11]. 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 allows us 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.

C. Approaches presented

MDCS is used as the reference, as it always gave the best results in the previous rounds of the competition. The approach is unchanged since ICDAR 2009 [12]. While its performance keeps dominating in terms of title-based evaluation, another institution has for the first time performed slightly better in terms of the link-based evaluation: the University of Innsbruck.

RunID	Precision	Recall	F-measure
Innsbruck	75.7%	68.9%	67.2%
MDCS	64.9%	71.5%	66.6%
Nankai	69.5%	61.4%	62.4%
GREYC-run-d	59.5%	45.3%	45.0%
Würzburg	44.1%	54.7%	44.7%
GREYC-run-c	60.5%	38.8%	41.8%
GREYC-run-a	59.2%	36.3%	38.7%
Epita	37.8%	35.0%	35.0%
GREYC-run-b	32.9%	21.7%	23.9%
GREYC-run-e	32.9%	21.7%	23.9%

TABLE IV. Performance scores for the Structure Extraction competition 2013 based on the XRCE link-based metrics.

1) Approaches based on the exploitation of printed ToC pages: The provided descriptions of the approaches revealed that most of the participants focused on detecting ToC pages and exploiting their content. They made no use of the rest of the contents of the books, except for the purpose of page linking (i.e., to find the right page number corresponding to a ToC entry). The technique employed by **MDCS** consists of three steps: recognizing ToC pages, assigning every page in the book to a physical page number, and finally processing each ToC page to extract all ToC entries through a supervised method relying on pattern occurrences detected in a training set.

The approach of the University of **Innsbruck** relies on machine learning for the detection of ToC areas [13]. The approach also uses fuzzy logic to handle the variations in the style of books. A number of features are exploited in a rule-based fashion, aiming to reproduce the way human readers handle book structure. Those features are the coordinates of lines, blocks, and strings, the distances between consecutive physical layout elements, the line indents.

The approach of the University of **Würzburg** relies on the OCR-ed data (the DjVu XML files). It combines formal information from the optical character recognition process to functional information following a text analysis. This first participation is a result of a modular system, that shall ease further experiments.

The implementation of **Epita** relies on the text boxes provided in PDF documents and is meant for books containing a physical table of contents [14]. It first locates the ToC areas and then reconstructs and groups ToC entries with respect to features of text boxes: alignments, lines ending with numbers or not, text containing specific words, etc. Page linking is performed using the difference between the effective page number of a page in the middle of the book and its page number in the book. This difference is applied throughout the book.

2) Approaches based on full book content: The technique followed by the University of Caen (**GREYC**) [15] works on full documents, with no particular focus on ToC pages (with no attempt to detect them). Their goal is to detect chapter beginnings with a 4-page window that aims to spot large whitespaces as strong indicators of the end of a chapter and the beginning of a new one. Unlike other approaches, the method is totally unsupervised.

3) Hybrid approaches: The University of **Nankai** is the group that attempted to extract book structure both by analysing ToC areas and book content [16]. Interestingly, ToC and content analysis are used as two parallel alternatives: If a ToC is identified, the ToC area is exploited and the headlines found in the book are ignored. In others words, the book content is used if and only if no ToC area is detected. This is justified by empirical evidence that the method exploiting the analysis of ToC areas performs best.

Surprisingly, no successful attempt has been reported to combine the analysis of ToC area and book content over the same document. This however seems to be a natural way to improve upon the state of the art of the methods described above.

[10] https://doucet.users.greyc.fr/StructureExtraction/2013/

VII. Summary and Future Plans

During its first two rounds at ICDAR, the book structure extraction competition gathered, in a collaborative effort by 10 institutions, ground truth ToC annotations for a total of 1,037 books. In 2013, the 6 participating institutions could enjoy an evaluation with almost twice as many books as either of the previous years (967), without having to get involved in the ground truth creation process.

In future years, we aim to investigate the usability of the extracted ToCs, both for readers in navigating books and systems that index and search parts of books. 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.

To be able to build even larger evaluation sets, we hope to experiment with crowdsourcing methods, which have been shown to offer a reliable method for high cognitive tasks [17]. This may offer a natural solution to the evaluation challenge posed by the massive data sets handled in digitized libraries.

Acknowledgment

The ground truth production was partly funded by the Seventh Framework Programme (FP7) of the EU Commission.

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] 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.

[5] 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.

[6] 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.

[7] C. Clausner, S. Pletschacher, and A. Antonacopoulos, "Aletheia - an advanced document layout and text ground-truthing system for production environments," *IEEE Xplore Digital Library*, pp. 48–52, 2011, proceedings of the 2011 International Conference on Document Analysis and Recognition (ICDAR), 18-21 September 2011, Beijing. [Online]. Available: http://usir.salford.ac.uk/26267/

[8] L. Gander, C. Lezuo, and R. Unterweger, "Rule based document understanding of historical books using a hybrid fuzzy classification system," in *Proceedings of the 2011 Workshop on Historical Document Imaging and Processing*, ser. HIP '11. New York, NY, USA: ACM, 2011, pp. 91–97. [Online]. Available: http://doi.acm.org/10.1145/2037342.2037358

[9] A. Doucet, G. Kazai, B. Dresevic, A. Uzelac, B. Radakovic, and N. Todic, "ICDAR 2009 Book Structure Extraction Competition," in *Proceedings of the Tenth International Conference on Document Analysis and Recognition (ICDAR'2009)*, Barcelona, Spain, july 2009, pp. 1408–1412.

[10] A. Doucet, G. Kazai, and J.-L. Meunier, "ICDAR 2011 Book Structure Extraction Competition," in *Proceedings of the Eleventh International Conference on Document Analysis and Recognition (ICDAR'2011)*, Beijing, China, September 2011, pp. 1501–1505.

[11] 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

[12] A. Uzelac, B. Dresevic, B. Radakovic, and N. Todic, "Book layout analysis: TOC structure extraction engine," in *INEX*, ser. Lecture Notes in Computer Science, S. Geva, J. Kamps, and A. Trotman, Eds., vol. 5613. Springer Verlag, Berlin, Heidelberg, 2009.

[13] L. Gander, C. Lezuo, and R. Unterweger, "Rule based document understanding of historical books using a hybrid fuzzy classification system," in *Proceedings of the 2011 Workshop on Historical Document Imaging and Processing*, ser. HIP '11. New York, NY, USA: ACM, 2011, pp. 91–97. [Online]. Available: http://doi.acm.org/10.1145/2037342.2037358

[14] G. Lazzara, R. Levillain, T. Géraud, Y. Jacquelet, J. Marquegnies, and A. Crepin-Leblond, "The scribo module of the olena platform: A free software framework for document image analysis," in *Proceedings of the Eleventh International Conference on Document Analysis and Recognition (ICDAR'2011)*. IEEE.

[15] E. Giguet and N. Lucas, "The book structure extraction competition with the resurgence software at caen university," in *Focused Retrieval and Evaluation*, ser. Lecture Notes in Computer Science, S. Geva, J. Kamps, and A. Trotman, Eds. Springer Berlin / Heidelberg, 2010, vol. 6203, pp. 170–178. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-14556-8_18

[16] C. Liu, J. Chen, X. Zhang, J. Liu, and Y. Huang, "TOC Structure Extraction from OCR-ed Books," in *Focused Retrieval of Content and Structure : 10th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2011*, ser. Lecture Notes in Computer Science, S. Geva, J. Kamps, and R. Schenkel, Eds., vol. 7424. Springer, 2012, pp. 80–89.

[17] 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.

[18] 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.