# Logical Structure Extraction from Digitized Books

Antoine Doucet

# Chapter 1

# Logical Structure Extraction from Digitized Books

*Antoine Doucet*

## 1.1 Introduction

Mass digitization projects, such as the Million Book Project, efforts of the Open Content Alliance, and the digitization work of Google, are converting whole libraries by digitizing books on an industrial scale [5]. The process involves the efficient photographing of books, page-by-page, and the conversion of the image of each page 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 not recognized. In order to enable systems to provide users with richer browsing experiences, it is necessary to make such additional structures available, 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 [35].

The chapter is structured as follows. We start by placing the competition in the context of the work conducted at the Initiative for the Evaluation of XML Retrieval (INEX) Evaluation Forum [22]. We then 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 also detailed. The ground-truth-creation process and its outcome are next described, together with the corresponding evaluation metrics used and the final results, alongside brief descriptions of the participants' approaches. We conclude with a summary of the competition and how it could be built upon.

### 1.1.1 *Background*

Motivated by the need to foster research in areas relating to large digital book repositories (see e.g., [21]), the Book Track was launched in 2007 [22] as part of the INEX. Founded in 2002, INEX is an evaluation forum that investigates focused retrieval approaches [14] 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.

One major limitation of digitized books is the fact that their structure is physical, rather than logical. Following this, the evaluation and relevance judgments based on the book corpus have essentially been based on whole books and selections of pages. This is unfortunate considering that books seem to be the key application field for structured information retrieval (IR). The fact that, for instance, chapters, sections, and paragraphs are not readily available has been a frustration for the structured IR community gathered at INEX, because it does not allow us to test the techniques created for collections of scientific articles and for Wikipedia.

Unlike digitally born content, the logical structure of digitized books is not readily available. A digitized book is often only split into pages with

possible paragraphs, lines, and word markup. This was also the case for the 50,000 digitized book collection of the INEX Book Search track [22]. The use of more meaningful structure, e.g., chapters, table of contents (ToC), bibliography, or back-of-book index, to support focused retrieval has been explored for many years at INEX and has been shown to increase retrieval performance [35].

To encourage research aiming to provide the logical structure of digitized books, we created the Book Structure Extraction competition, which we later brought to the community of document analysis.

Starting from 2008, within the second round of the INEX Book Track, we entirely created the methodology to evaluate the Structure Extraction process from digitized books: problem description, submission procedure, annotation procedure (and corresponding software), metrics, and evaluation.

### 1.1.2 *Context and Motivation*

The overall goal of the INEX Book Track is to promote interdisciplinary 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 IR tasks [24].

However, since the collection was made of digitized books, the only structure that was readily available was that of pages, each page being easily identified from the fact that it corresponds to one and only one image file, as a result of the scanning process. In addition, a few other elements can easily be detected through OCR, as we can see with the DjVu file format (an example of which is given in Figure 1.1). This markup denotes pages, words (detected as regions of text separated by horizontal space), lines (regions of text separated by vertical space), and "paragraphs" (regions of text separated by a significantly wider vertical space than other lines). Those paragraphs, however, are only defined as internal regions of a page (by definition, they cannot span over different pages).

Hence, there is a clear gap to be filled between research in structured IR, which relies on a logical structure (chapters, sections, etc.), and the digitized book collection, which contains only the physical structure. From

```
<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>
```

Figure 1.1. A sample DjVu XML document.

a cognitive point of view, retrieving book pages may be sensible with a paper book, but it is nonsense with a digital book. The BookML format, of which we give an example in Figure 1.2, is a better attempt to grasp the logical structure of books, but it remains clearly insufficient.

#### 1.1.2.1 *Structured Information Retrieval Requires Structure*

In the context of e-readers, even the concept of a page actually becomes questionable: What are pages if not a practical arrangement to avoid printing a book on a single 5 squared meter sheet of paper? For the moment, it seems, however, that users are still attached to the concept of a page, mostly as a convenient marker of "where did I stop last?", but when they can actually bookmark any word, line, or fragment of the book, how long will users continue to bookmark pages?

It is important to remember that books as we know them are only a step in the history of reading devices, starting from the papyrus, a very long scroll containing a single sequence of columns of text, used during 3 millennia until the Roman codex brought up the concept of a page.

```
<document>
 <page pageNumber=''I-N'' label=''PT_CHAPTER'' [...]>
  <region regionType=''text'' [...]>
   <section label=SEC_BODY'' [...]>
    <line [...]>
     <word val=''Moby'' [...]/>
     <word val=''Dick'' [...]/>
    </line>
    <line [...]>
     <word val=''Herman'' [...]/>
     <word val=''Melville'' [...]/>
    </line>    [...]
   </section>  [...]
  </region>    [...]
 </page>       [...]
</docment>
```

Figure 1.2. A sample BookML document.

The printing press in the 15th century allowed the shift from manual to mechanical copying, bringing books to the masses [36]. Because of reading devices, after switching from papyrus to paper, we are now living another dramatic change from the paper to the digital format; it is to be expected that the unnecessary implications of the paper format will disappear in the long run. All physical structure is bound to disappear or become widely unstable. For instance, should pages remain, the page content will vary widely every time the font size is changed, something that most e-readers allow.

What shall remain, however, is the logical structure, whose reason to be is not practical motivations but an editorial choice of the author to structure his works and to facilitate the readers' access. Unfortunately, it is exactly this part of the structure that the digitized book collection of INEX missed. On one hand, it seemed to be an ideal framework for structured IR, while on the other, the collection's logical structure was hardly usable. This motivated the design of the Book Structure Extraction competition, to bridge the gap between the digitized books and the (structured) IR research community.

#### 1.1.2.2 *Context*

In 2008, during the second year of the INEX Book Track, the Book Structure Extraction task was introduced [25] and 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 was "beta" run in 2008 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 built by the organizers. The second round was run both at INEX 2009 [26] and additionally at the *International Conference on Document Analysis and Recognition* (ICDAR) [9] where it was accepted as an official competition. This allowed us to reach the document analysis community and bring a bigger audience to the effort while inviting competitors to present their approaches at the INEX workshop. This further allowed one to build up on the established infrastructure with an extended test set and a procedure for collaborative annotation that greatly reduced the effort needed for building the ground-truth. The competition was run again in 2010 at INEX [27] and in 2011 and 2013 at ICDAR [11, 12] (INEX runs every year, whereas ICDAR runs every second year).

In the next section, we will describe the full methodology that we put in place from scratch to evaluate the performance of Book Structure Extraction systems, as well as the challenges and contributions that this work involved.

## 1.2 Book Collection

The INEX Book Search corpus contains 50,239 digitized, out-of-copyright books, provided by Microsoft Live Search and the Internet Archive [22]. It consists of books of different genres, including history books, biographies, literary studies, religious texts and teachings, reference works, encyclopedias, essays, proceedings, novels, and poetry.

Each book is available in three different formats: image files as portable document format (PDF), DjVu XML containing the OCR text and basic structure markup as illustrated in Figure 4.1, and BookML, containing a more elaborate structure constructed from the OCR and illustrated in Figure 1.2.

**DjVu format.** 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.

**BookML format.** The OCR content of the books has further been converted from the original DjVu format to an XML format, referred to as BookML, developed by the Document Layout Team of Microsoft Development Center, Serbia. BookML provides richer structure information, including markup for ToC entries. Most books also have an associated metadata file (*.mrc), which contains publication (author, title, etc.) and classification information in the MAchine-Readable Cataloging (MARC) record format.

The basic XML structure of a typical book in BookML (ocrml.xml) is a sequence of pages containing nested structures of regions, sections, lines, and words (Figure 1.2).

BookML provides a rich set of labels indicating structure information and additional marker elements for more complex texts, such as a ToC. For example, the label attribute of a section indicates the type of semantic unit that the text contained in the section is likely to be a part of, e.g., a table of contents (SEC_TOC), a header (SEC_HEADER), a footer (SEC_FOOTER), or the body of the page (SEC_BODY).

The original corpus totals 400 GB, while the reduced version is only 50 GB (and 13 GB compressed). The reduced version was created by removing the word tags and their attributes (coordinates, etc.) and propagating the values of the val attributes as content into the parent line elements.

## 1.3 Setting Up the Competition

The goal of the competition was 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 face is to construct hyperlinked ToCs for a collection of digitized books. As the name "Structure Extraction competition" suggests, the long-term goal of this effort is to extract the whole logical structure of documents, but the extraction of ToC has been planned as a significant milestone, unexpectedly difficult to reach. The next steps will be discussed in perspectives.

To evaluate the quality of extracted ToCs, we had to construct an appropriate book collection, define a format for ToCs, define metrics to compare extracted ToCs to a ground-truth, and last but not least, define ways to build such a ground-truth in a reasonable time, while still constructing a ground-truth that is large enough to allow for significant results, but without compromising quality and consistency.

### 1.3.1 *Defining the Evaluation Corpus*

In 2009, 2011, and 2013, the Book Structure Extraction evaluation corpus consisted of 1,000 distinct book subsets of the Book Search Track's 50,239 book corpus. Therefore, it consisted of a representative set of books of different genres, including history books, biographies, literary studies, religious texts and teachings, reference works, encyclopedias, essays, proceedings, novels, and poetry.

To facilitate the separate evaluation of Structure Extraction 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 made sure to include 200 books that did not contain a printed ToC into the total set of 1,000. To do this, we used the BookML format where pages that contain the printed ToC (the so-called ToC pages) are explicitly marked up. We then selected a set of 800 books with detected ToC pages, and added a set of 200 books without any detected ToC pages into the full test set of 1,000 books. Please note that this ratio of 80.20% of books with and without printed ToCs is proportional to that observed over the whole corpus of 50,239 books.

### 1.3.2 *Sample Research Questions*

To motivate the community of researchers, we provided a sample of open research questions that the competition shall help address. 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 contains 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?

### 1.3.3 *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 was to build hyperlinked ToCs for each book in the test set. The OCR text of each book is stored in the DjVu XML format (see once more Figure 1.1). Participants could employ any techniques and could make use of either or both the OCR text and the PDF images to derive the necessary structure information and generate the ToCs. Giving the possibility of using the OCR text (DjVu format) was meant to facilitate access for participants with no experience of OCR, and let them start from a preprocessed common-ground format. Participating systems were expected 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) of a run is given in Figure 1.3.

### 1.3.4 *Annotation of ToCs: Methodology and Software*

Naturally, comparing 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

```
<!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 1.3. DTD of the XML output (run) that participating systems were expected to submit to the competition, containing the generated hyperlinked ToC for each book in the test set.

rather than forcing participants to perform 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. This was pretty much in line with INEX practice. However, placing the burden on participants is evidently a hindrance, and so the effort must be as limited as possible. This section describes the ground-truth annotation process we designed and its outcomes.

#### 1.3.4.1 *Annotation Process*

The process of manually building the ToC of a book is very time-consuming. Hence, to make the creation of the ground-truth for 1,000 digitized books

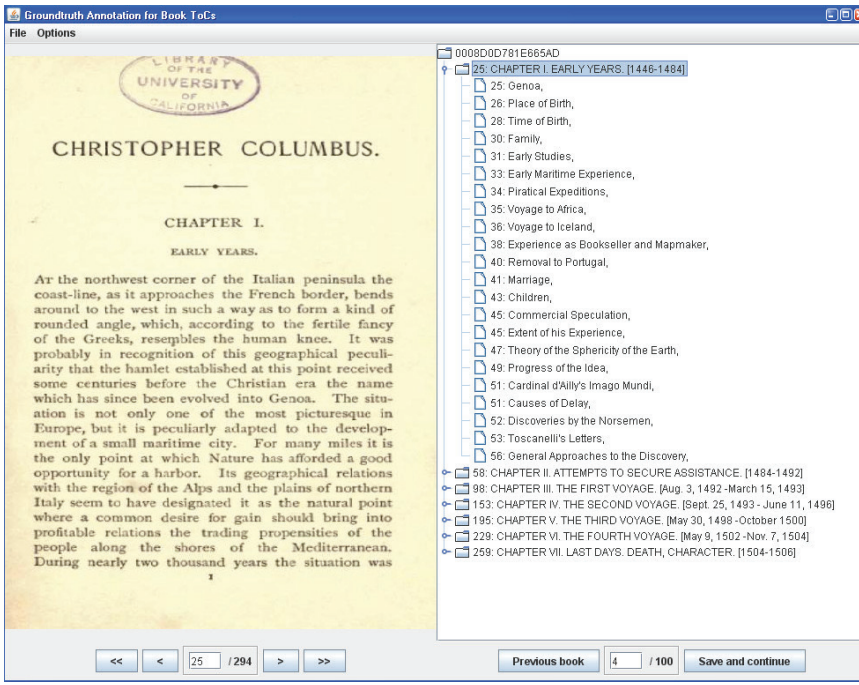


Figure 1.4. A screenshot of the ground-truth annotation tool.

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 and implemented by the organizers for this purpose. The tool takes as input a generated ToC and allows annotators to manually correct any mistakes. A screenshot of the tool is shown in Figure 1.4. 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 one to navigate through the book. The JPEG image of each visited page was downloaded from a former INEX server at www.booksearch.org.uk and locally cached to limit bandwidth usage.

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 participant submissions in equal shares.

Finally, the annotation effort was shared among all participants. Teams that submitted runs were required to contribute a minimum of 50 books, while others were required to contribute a minimum of 100 books (20% of those books did not contain a printed ToC). The created ground-truth was made available to all contributing participants for use in future evaluations.

#### 1.3.4.2 *Collected Ground-truth Data*

In 2009, seven teams participated in the ground-truth annotation process, four of which did not submit runs.

This joint effort resulted in a set of 2,004 annotated books. To ensure the quality and internal consistency of the collected annotations, each of the annotated ToCs was reviewed by the organizers, and a significant number had to be removed. Any ToC with annotation errors was then removed. Most of the time, errors were due to failure to follow the annotation guidelines or incomplete annotations.

Following this cleansing step, 527 annotated books remain to form the ground-truth file that was distributed to each contributing organization. Around 97 of the annotated books are those for which no ToC pages were detected.

In 2011, the output was very similar with six teams participating to the annotation phase, two of which did not submit run, and a total number of 513 annotated books brought out to form the 2011 ground-truth.

In 2013, there were again six participating teams; this time all submitted runs. In 2013, we were able to outsource the evaluation, thanks to partial funding from the Seventh Framework Programme (FP7) of the European Union (EU) Commission. This allowed one to annotate a total of 967 books using the Aletheia tool [4].

### 1.3.5 *Validation of the Annotation Procedure*

To validate the methodology, and as the evaluation is based on manually built ground-truth, it was crucial to validate the approach by verifying the consistency of the gathered ToC annotations.

To do this, we assigned the same set of books to two different institutions. This resulted in 61 books being annotated twice. We measured annotator agreement by using one of these sets as a run and the other as the ground-truth and calculating our official evaluation metrics. The result of this comparison is given in Table 1.1.

We can observe an agreement rate of over 70% for complete entries based on the *F*-measure. It is important to observe that most of the disagreements stem from title matching, which makes us question whether the 20% tolerance utilized when comparing title strings with the Levenshtein distance may need to be increased, so as to lower the impact of annotator disagreement on the evaluation results. However, this requires further investigation because an excessive increase would lead to uniform results (more duly distinct titles would be deemed equivalent).

### 1.3.6 *Metrics*

The automatically generated ToCs submitted by participants were evaluated by comparing them to a manually built ground-truth. The evaluation required the definition of a number of basic concepts.

Table 1.1. The score sheet measuring annotator agreement for the 61 books that were assessed independently by two distinct institutions.

| | **Precision (%)** | **Recall (%)** | ***F*-measure (%)** |
|---|---|---|---|
| Titles | 83.51 | 83.91 | 82.86 |
| Levels | 74.32 | 75.00 | 74.04 |
| Links | 82.45 | 82.87 | 81.83 |
| Complete, except depth | 82.45 | 82.87 | 81.83 |
| Complete entries | 73.57 | 74.25 | 73.31 |

**Definitions.** We define the atomic units that make up a ToC as ToC entries. A ToC entry has the following three properties: title, link, and depth level. For example, given a ToC entry corresponding to a book chapter, its title is the chapter title, its link is the physical page number at which the chapter starts in the book, and its depth level is the depth at which the chapter is found in the ToC tree, where the book represents the root.

Given the above definitions, the task of comparing two ToCs (i.e., comparing a generated ToC to one in the ground-truth) can be reduced to matching the titles, links, and depth levels of each ToC entry. This is, however, not a trivial task as we explain next.

**Matching titles.** A ToC title may take several forms, and it may only contain, e.g., the actual title of a chapter, such as "His Birth and First Years," or it may also include the chapter number as "3. His Birth and First Years" or even the word "chapter" as "Chapter 3. His Birth and First Years". In addition, the title that is used in the printed ToC may differ from the title, which then appears in the book content. It is difficult to differentiate between the different answers as all of them are in fact correct titles for a ToC entry.

Thus, to take into account not only OCR errors but also the fact that many similar answers may be correct, we adopt vague title matching in the evaluation. We say that two titles match if they are "sufficiently similar," where similarity is measured based on a modified version of the Levenshtein algorithm (where the cost of alphanumeric substitution, deletion, and insertion is 10, and the cost of nonalphanumeric substitution, deletion, and insertion remains 1) [32].

Two strings $A$ and $B$ are "sufficiently similar" if

$$D = \frac{\text{Levenshtein Dist} * 10}{\text{Min}(\text{length}(A), \text{lengh}(B))}$$

is less than 20% and if the distance between their first and last (up to) five characters is less than 60%.

**Matching links.** A link is said to be correctly recognized if there is an entry with matching title linking to the same physical page in the ground-truth.

**Matching depth levels.** A depth level is said to be correct if there is an entry with matching title at the same depth level in the ground-truth.

**Matching complete ToC entries.** A ToC entry is entirely correct if there is an entry with matching title and same depth level, linking to the same physical page in the ground-truth.

**Measures.** For a given book ToC, we can then calculate precision and recall measures [34] for each property separately and for complete entries. Precision is defined as the ratio of the total number of correctly recognized ToC entries and the total number of ToC entries in a generated ToC, 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 is then calculated as the harmonic mean of precision and recall. Each of these values was computed separately for each book and then averaged over the total number of books (macro-average).

The measures were computed over the two subsets of the 1,000 books, as well as the entire test set to calculate overall performance. The two subsets, originally comprising of 800 and 200 books, respectively, that do and do not have a printed ToC, allowed us to compare the effectiveness of techniques that do or do not rely on the presence of printed ToC pages in a book.

**Results.** For each submission, a summary was provided in two tables, presenting general information about the run as well as a corresponding score sheet (see an example in Table 1.2).

#### 1.3.6.1 *Alternative Measure and Discussion*

Participants were encouraged to propose alternative metrics, and thus the XRCE link-based measure [6] was introduced to complement the official measures with the aim to take into account the quality of the links directly, rather than conditionally to the title's validity.

Indeed, the official measure works by matching ToC entries primarily based on their title. Hence, the runs that incorrectly extract titles will be penalized with respect to all the measures presented in the score sheet of Table 1.2. For instance, a system that incorrectly extracts titles, while correctly identifying links will obtain very low scores (possibly 0%). The XRCE link-based measure permits one to evaluate the performance of systems works by matching ToC entries primarily based on links rather than titles.

Table 1.2. An example score sheet summarizing the performance evaluation of the "MDCS" run.

| | **Precision (%)** | **Recall (%)** | ***F*-measure (%)** |
|---|---|---|---|
| Titles | 57.90 | 61.07 | 58.44 |
| Levels | 44.81 | 46.92 | 45.09 |
| Links | 53.21 | 55.53 | 53.62 |
| Complete, except depth | 53.21 | 55.53 | 53.62 |
| Complete entries | 41.33 | 42.83 | 41.51 |

The "complete entries" measure, used as a reference in most of this chapter, is a global, cumulative measure. Because an entry must be entirely correct, i.e., title, link, etc., to be counted as a correct entry, an error in any of the criteria implies a complete error.

While the various measures presented previously have in common a sensitivity to errors in the titles of ToC entries, the alternative measure in turn is strongly dependent on the correctness of page links.

We do not claim that success with respect to one metric is more important than that with another, but believe that the measures presented should be seen as complementary. Depending on the application or situation, one metric may be preferred over another. For example, if navigation is key, then being able to land the user on a page where a chapter starts may be more important than getting the title of the chapter right.

One of our goals in the future is to provide a toolbox of metrics, to be used by researchers enabling them to analyze and better understand the outcome of each of their approaches. The current version of this toolkit is available on the competition's website.

### 1.3.7 *Approaches Presented*

Throughout the campaigns of the Book Structure Extraction competition, several approaches have been presented. They can be categorized into three families.

**ToC recognition and analysis.** Most of the approaches of the state of the art rely on the detection of ToC pages within the book, and their detailed

analysis for listing all ToC entries and linking them to the corresponding pages. To extract ToC entries and link them to the right page, the most effective technique to date remains the one developed by Dresevic *et al*. [13] which consists in recognizing ToC pages and then processing them so as to extract all ToC entries using a supervised method relying on pattern occurrences from an external training set. However, it is worth noting that the approach of Gander *et al*. [16] performs better for the sole ToC entry extraction (not taking page-linking into account). This rule-based technique is meant to mimic the ToC reading behavior of users, incorporating fuzzy logic so as to handle variations in the style of books.

Further work was presented by Wu *et al*. [37] who introduced a taxonomy of ToC styles requiring distinct system behavior: "flat," "ordered," and "divided" ToC. The reported scores seem to overcome the state of the art, although unfortunately they are reported over a combination of data sets, with no direct mention of the prior results of the Book Structure Extraction competition, which renders comparison difficult.

**Book content analysis.** Rather than searching for ToC pages as a prior, Giguet and Lucas [20] search for chapter beginnings through the book content, using a 4-page window and aiming to spot large whitespaces as strong indicators of the end of a chapter and the beginning of a new one. Unlike most other approaches, their method is fully unsupervised.

**Hybrid approaches.** Given that 20% of the books of the book collection have no physical ToC, the combination of techniques using ToC recognition and book content analysis seems a reasonable solution. Liu *et al*. [33] proposed exactly that, relying on the book content when no ToC was detected, which allowed for improvement over using ToC recognition only. Still, their method fell short of outperforming the top ToC page-based approaches [13, 16]. Surprisingly, no approach has been presented, fully combining the features of ToC pages and the book content.

### 1.3.8 *Summary*

Starting from scratch, a complete framework was created for the evaluation of Structure Extraction from digitized books.

Unlike book retrieval, where adjustments had to be made to existing IR evaluation techniques, everything had to be done to be able to evaluate Book Structure Extraction. Hence, every step of the competition setup is a contribution: the problem definition, the compilation of the collection, the task description and submission procedure, the definition of evaluation metrics, the annotation format and procedure, etc.

Clearly, in a similar fashion as with book retrieval, the most important challenge is (and remains) to be able to annotate such massive collections. Being able to do it was the first challenge that we have managed to address. The next challenge is to increase significantly the amount of collected ground-truth. One obvious way, in the light of the latest development in Book Search, is to rely on crowdsourcing.

The setup of this competition, initially sketched and tested on a low scale in INEX 2008, was validated by the community for the first time when the competition was accepted by the ICDAR 2009 Program Committee as a conjoint event between INEX and ICDAR. The contributions were recognized through the publication of papers, the main one being an article in the *International Journal of Document Analysis and Recognition* (IJDAR) describing the framework in 2010 [10].

However, the most crucial acknowledgment is that of participants of the Structure Extraction competition. Around 22 institutions have expressed interest, 10 have participated to the ground-truth creation, and 8 submitted runs. This adhesion to the competition is not only supportive, but it is also the only way that we could provide a decent-sized collection to the community.

**Standalone test data.** Indeed, to facilitate the participation of other institutions in the future, it was decided in 2010 to always make available the second to last ground-truth set. We then distributed the initial set of 100 ToCs built during the first Book Structure Extraction task at INEX 2008 [25]. Following this, as soon as the 2011 competition started, the data collected in 2009 (527 book ToCs) was made available online. Similarly, for the ICDAR 2013 competition, the 2011 ground-truth set was released (until then, its access remained restricted to sufficient contributors of the 2011 ground-truth set). Finally, thanks to dedicated funding for its construction, the 2013 ground-truth was openly distributed from the start, providing an additional 967 annotated books.

Effectively, these ground-truth sets, distributed together with the document collection and the evaluation software, are forming standalone evaluation packages, freely available to the research community on the competition's website.[a]

**ICDAR competition and INEX workshop.** A strength of the conjoint organization between INEX and ICDAR is the effective bridging of two communities: the competition was labeled by and presented at ICDAR, but at the same time, participants were invited to write papers presenting their approaches at the INEX workshop, a selection of which have been published by *Springer Lecture Notes in Computer Science* (LNCS) within the INEX workshop proceedings (see, e.g., [6–8, 20, 31, 33]).

The respective ICDAR and INEX schedules facilitated this, since the ICDAR result deadline is around the middle of the year, whereas the INEX workshop is generally held in December, with paper submission deadlines at the end of October.

**Future of the Structure Extraction competition.** The data sets are now freely available. This was requested by several participants intending to run further experiments, as well by several other institutions which were still developing their Structure Extraction systems at submission time.

Another important reason to open access to the data set is the current results, indicating that much could still be improved upon, especially in the case of books that do not contain ToC pages. This underlines how much remains to be done in the field of Book Structure Extraction.

The possible directions of future rounds of the Structure Extraction competition are discussed in Section 1.5. Before that, we will summarize related publications and put them in context.

## 1.4 Related Publications

The general background of the INEX workshop is best summarized in INEX reports published within INEX workshops [24–30] and the SIGIR Forum [1, 2, 22], the biannual publication of the ACM Specific Interest Group on Information Retrieval (SIGIR). The evolution of the share of

[a] http://pageperso.univ-lr.fr/antoine.doucet/structureExtraction/training.

the reports dealing with the Book Track is a good indicator of its growing importance within the INEX Framework.

The initial setup of the book retrieval task was presented within INEX 2007 [24], while a more extensive standalone description was published in the SIGIR Forum in 2008 [22]. A number of potential new user tasks were exposed in a short position paper presented at the European Conference on Digital Libraries (ECDL) 2008 [23]. This is where the idea of the Structure Extraction was proposed for the first time. The later rounds of the Book Track introduced new tasks as well as variations of existing ones. All these tasks, as well as the participants' approaches, are described in the corresponding Book Track overviews [23–30].

The Structure Extraction competition is also briefly overviewed within each of these papers. However, extensive description and discussion are rather found in publications of the document analysis community. Indeed, following the acceptance of the 1st, 2nd, and 3rd Structure Extraction competitions at the International Conference on Document Analysis and Recognition (ICDAR), respectively, in 2009, 2011, and 2013, their setups, overviews, and results were published in corresponding ICDAR proceedings [9, 11, 12]. The contribution to the evaluation of Book Structure Extraction was most extensively described in a longer article, published in the IJDAR in 2011 [10].

In addition, selected papers describing participant approaches were published yearly within the INEX workshop proceedings by Springer, as different volumes of the LNCS series [15, 17–19]. These volumes contain descriptions of participant approaches to both the book search and the Structure Extraction task.

## 1.5 Conclusions and Perspectives

Starting from the distribution of a digitized book collection in 2007 and the subsequent very first Book Search Track run at INEX [24], progress has been steady. We have designed techniques to gather a sufficient number of relevance assessments and evaluate Book IR. We also fostered renewed interest and designed evaluation methods for the problem of the extraction of the logical structure from digitized books, opening the way for applications of structured IR in a motivating application setting.

The number of registered participants of the Book Track has grown from 27 in 2007 to 54 in 2008, and 84 in 2009 and 2010. In 2011, at the 10th anniversary of INEX, the Book Track became the main track of INEX, replacing the *ad hoc* track which evaluated structured IR with collections of scientific articles (IEEE) and Wikipedia articles, from 2002 to 2010. For both the Book Search and the Structure Extraction tasks, participants have been invited to present their approaches at the INEX workshop, with proceedings published by Springer LNCS.

Starting form 2012, the INEX workshop, with the Book Track as its main track, will be co-located with the Cross-Language Evaluation Forum (CLEF), which recently grew from a forum on cross-language evaluation to a full-scale conference focused on multilingual and multimodal information access.

A number of improvements shall be considered for the future of the Book Structure Extraction task, e.g., crowdsourcing the ground-truth of Book Structure. In spite of the tremendous efforts of participants to build the ground-truth, we shall 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 step was successfully made in the Book Search task, and it is now natural for the Structure Extraction competition to follow a similar path.

Investigating the usability of the extracted ToCs also seems worthwhile. 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. This effort shall be led through crowdsourcing.

Both these extensions offer interesting questions in terms of quality control, a key issue to make the output of crowdsourcing useful. These shall be relevant internship topics for Master students.

**Further structure.** As the name "Book Structure Extraction" competition suggests, ToCs are not the sole objective, but rather a first milestone that proves to be far more difficult to reach than expected.

In the future, however, we plan to expand the task to include the identification of more exhaustive structure information, e.g., header/footer, bibliography, etc.

**Human–computer interaction (HCI).** An important fact about e-readers is that they deprive readers from a lot of context. Being returned only a fragment of text is not the same as being given a pointer into a printed book. The ability to search for keywords within an eBook is depriving readers from context that is intrinsically available with paper books [3]. This poses many questions in HCI.

## References

1. P. Bellot, T. Chappell, A. Doucet, S. Geva, S. Gurajada, J. Kamps, G. Kazai, M. Koolen, M. Landoni, M. Marx, A. Mishra, V. Moriceau, J. Mothe, M. Preminger, G. Ramírez, M. Sanderson, E. Sanjuan, F. Scholer, A. Schuh, X. Tannier, M. Theobald, M. Trappett, A. Trotman, and Q. Wang, Report on INEX 2012. *SIGIR Forum*, vol. 46, no. 2, pp. 50–59, December 2012.
2. P. Bellot, A. Doucet, S. Geva, S. Gurajada, J. Kamps, G. Kazai, M. Koolen, A. Mishra, V. Moriceau, J. Mothe, M. Preminger, E. SanJuan, R. Schenkel, X. Tannier, M. Theobald, M. Trappett, A. Trotman, M. Sanderson, F. Scholer, and Q. Wang, Report on INEX 2013. *SIGIR Forum*, vol. 47, no. 2, pp. 21–32, January 2013.
3. R. Chartier and A. Paire, *Pratiques de la lecture/sous la direction de Roger Chartier et à l'initiative d'Alain Paire*. Payot & Rivages, Paris, 2003. (Publication originale chez Rivages en, 1985.)
4. C. Clausner, S. Pletschacher, and A. Antonacopoulos, "Aletheia — an advanced document layout and text ground-truthing system for production environments," in *Proceedings of the 11th International Conference on Document Analysis and Recognition* (ICDAR'2011), IEEE, Beijing, China, pp. 48–52, September 2011.
5. K. Coyle, "Mass digitization of books," *Journal of Academic Librarianship*, vol. 32, no. 6, pp. 641–645, 2006.
6. H. Déjean and J.-L. Meunier, "XRCE participation to the 2009 book structure task," in *Focused Retrieval and Evaluation: 8th International Workshop of the Initiative for the Evaluation of XML Retrieval* (INEX'2009), *Lecture Notes in Computer Science*, vol. 6203, pp. 160–169. Berlin, Heidelberg, Springer, 2010.

7. 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 Series* (DAS'2010). New York, NY, ACM, pp. 301–308, 2010.
8. H. Déjean, "Using page breaks for book structuring," in *Focused Retrieval of Content and Structure: 10th International Workshop of the Initiative for the Evaluation of XML Retrieval* (INEX'2011), *Lecture Notes in Computer Science*, vol. 7424, pp. 57–67, Berlin, Heidelberg, Springer, 2012.
9. A. Doucet, G. Kazai, B. Dresevic, A. Uzelac, B. Radakovic, and N. Todic, "ICDAR 2009 Book Structure Extraction Competition," in *Proceedings of the 10th International Conference on Document Analysis and Recognition (ICDAR'2009)*, IEEE, Barcelona, Spain, pp. 1408–1412, July 2009.
10. 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*, vol. 14, no. 1, pp. 45–52 (Special Issue on Performance Evaluation of Document Analysis and Recognition Algorithms), 2011.
11. A. Doucet, G. Kazai, and J.-L. Meunier, "ICDAR 2011 Book Structure Extraction Competition," in *Proceedings of the 11th International Conference on Document Analysis and Recognition* (ICDAR'2011), IEEE, Beijing, China, pp. 1501–1505, September 2011.
12. A. Doucet, G. Kazai, S. Colutto, and G. Mühlberger, "Overview of the ICDAR 2013 Competition on Book Structure Extraction," in *Proceedings of the 12th International Conference on Document Analysis and Recognition* (ICDAR'2013), Washington DC, USA, pp. 1438–1443, August 25–28, 2013.
13. B. Dresevic, A. Uzelac, B. Radakovic, and N. Todic, "Book layout analysis: TOC structure extraction engine," in S. Geva, J. Kamps, and A. Trotman, Eds. *Advances in Focused Retrieval* (INEX'2008), *Lecture Notes in Computer Science Series*, vol. 5613, pp. 164–171. Berlin, Heidelberg, Springer-Verlag, 2009.
14. N. Fuhr, N. Gövert, G. Kazai, and M. Lalmas, Eds., in *Proceedings of the 1st Workshop of the Initiative for the Evaluation of XML Retrieval* (INEX'2002), Schloss Dagstuhl, Germany, December 9–11, 2002.
15. N. Fuhr, J. Kamps, M. Lalmas, and A. Trotman, Eds., "Focused access to XML documents," in *6th International Workshop of the Initiative for the Evaluation of XML Retrieval* (INEX'2007), Dagstuhl Castle, Germany, December 17–19, 2007. (*Selected Papers*, *Lecture Notes in Computer Science*, vol. 4862, New York, NY, Springer, 2008.)
16. 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* (HIP'2011). New York, NY, ACM, pp. 91–97, 2011.

17. S. Geva, J. Kamps, and A. Trotman, Eds., "Advances in focused retrieval," in *7th International Workshop of the Initiative for the Evaluation of XML Retrieval* (INEX'2008), Dagstuhl Castle, Germany, December 15–18, 2008. (*Revised and Selected Papers*, *Lecture Notes in Computer Science*, vol. 5631. New York, NY, Springer, 2009.)
18. S. Geva, J. Kamps, and A. Trotman, Eds., "Focused retrieval and evaluation," in *8th International Workshop of the Initiative for the Evaluation of XML Retrieval* (INEX'2009), Brisbane, Australia, December 7–9, 2009. (*Revised and Selected Papers*, *Lecture Notes in Computer Science*, vol. 6203. New York, NY, Springer, 2010.)
19. S. Geva, J. Kamps, R. Schenkel, and A. Trotman, Eds., "Comparative evaluation of focused retrieval," in *9th International Workshop of the Inititative for the Evaluation of XML Retrieval* (INEX'2010), Vught, The Netherlands, December 13–15, 2010. (*Revised and Selected Papers*, *Lecture Notes in Computer Science*, vol. 6932. Berlin, Heidelberg, Springer, 2011.)
20. E. Giguet and N. Lucas, "The book structure extraction competition with the resurgence software at CAEN university," in S. Geva, J. Kamps, and A. Trotman, Eds. *Focused Retrieval and Evaluation, Lecture Notes in Computer Science Series*, vol. 6203, pp. 170–178, Berlin, Heidelberg, Springer, 2010.
21. P. Kantor, G. Kazai, N. Milic-Frayling, and R. Wilkinson, Eds., BooksOnline'08: in *Proceedings of the 2008 ACM Workshop on Research Advances in Large Digital Book Repositories*. New York, NY, ACM, 2008.
22. G. Kazai and A. Doucet, "Overview of the INEX 2007 Book Search Track" (Book Search'07), *ACM SIGIR Forum*, vol. 42, no. 1, pp. 2–15, 2008.
23. G. Kazai, A. Doucet, and M. Landoni. "New tasks on collections of digitized books," in *Proceedings of the 12th European Conference on Research and Advanced Technology for Digital Libraries* (ECDL'2008). Berlin, Heidelberg, Springer-Verlag, pp. 410–412, 2008.
24. G. Kazai and A. Doucet, "Overview of the INEX 2007 Book Search Track (Book Search'07)," in *Focused Access to XML Documents: 6th International Workshop of the Initiative for the Evaluation of XML Retrieval* (INEX'2007), *Lecture Notes in Computer Science*, vol. 4862, pp. 148–161. Berlin, Heidelberg, Springer, 2008.
25. G. Kazai, A. Doucet, and M. Landoni, "Overview of the INEX 2008 Book Track," in S. Geva, J. Kamps, and A. Trotman, Eds. *Advances in Focused Retrieval* (INEX'2008), *Lecture Notes in Computer Science*, vol. 5613. Berlin, Heidelberg, Springer-Verlag, 2009.

26. G. Kazai, A. Doucet, M. Koolen, and M. Landoni, "Overview of the INEX 2009 Book Track," in *Advances in Focused Retrieval: 8th International Workshop of the Initiative for the Evaluation of XML Retrieval* (INEX'2009), *Lecture Notes in Computer Science*, vol. 6203, pp. 145–159. Berlin, Heidelberg, Springer, in press, 2010.
27. G. Kazai, M. Koolen, J. Kamps, A. Doucet, and M. Landoni, "Overview of the INEX 2010 Book Track: Scaling up the evaluation using crowdsourcing," in *Advances in Focused Retrieval: 9th International Workshop of the Initiative for the Evaluation of XML Retrieval* (INEX'2010), *Lecture Notes in Computers Science*, vol. 6932 pp. 98–117. Berlin, Heidelberg, Springer, 2011.
28. 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*. New York, NY, ACM, 2011.
29. G. Kazai, M. Koolen, J. Kamps, A. Doucet, and M. Landoni, "Overview of the INEX 2011 Books and Social Search Track," in *Focused Retrieval of Content and Structure: 10th International Workshop of the Initiative for the Evaluation of XML Retrieval* (INEX'2011), Lecture Notes in Computer Science, vol. 7424, pp. 1–29. Berlin, Heidelberg, Springer, 2012.
30. M. Koolen, G. Kazai, M. Preminger, and A. Doucet, "Overview of the INEX 2013 Social Book Search Track," in *Information Access Evaluation Meets Multilinguality, Multimodality, and Visualization: 4th International Conference on the Cross-Language Evaluation Forum* (CLEF'2013), Valencia, Spain, pp. 1–26, September 23–26, 2013.
31. 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 11th International Conference on Document Analysis and Recognition* (ICDAR'2011), IEEE, Beijing, China, pp. 252–258, September 2011.
32. V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions and reversals," *Soviet Physics Doklady*, vol. 10, no. 8, pp. 707–710, 1966.
33. J. Liu, C. Liu, X. Zhang, J. Chen, and Y. Huang, "TOC Structure Extraction from OCR-ed books," in, S. Geva, J. Kamps, and R. Schenkel, Eds. *Focused Retrieval of Content and Structure: 10th International Workshop of the Initiative for the Evaluation of XML Retrieval* (INEX'2011), *Lecture Notes in Computer Science Series*, vol. 7424, pp. 80–89. Berlin, Heidelberg, Springer, 2012.
34. C. J. Van Rijsbergen, *Information Retrieval*, 2nd Edition. London, Butterworths, 1979.

35. R. Van Zwol and T. Van Loosbroek, "Effective use of semantic structure in XML retrieval," in G. Amati, C. Carpineto, and G. Romano, Eds. *Advances in Informal Retrieval* (ECIR'2007), *Lecture Notes in Computer Science Series*, vol. 4425, pp. 621–628. Berlin, Heidelberg, Springer, 2007.
36. C. Vandendorpe. Du papyrus à l'hypertexte. *Essai sur les mutations du texte et de la lecture*. Boréal, Montréal, 1999. nt2 hypertexte.
37. Z. Wu, P. Mitra and C. Lee Giles. "Table of contents recognition and extraction for heterogeneous book documents," in *Proceedings of the 12th International Conference on Document Analysis and Recognition* (ICDAR'2013), Washington, DC, USA, pp. 1205–1209, August 25–28, 2013.