

Setting up a Competition Framework for the Evaluation of Structure Extraction from OCR-ed Books

Antoine Doucet¹, Gabriella Kazai², Bodin Dresevic³, Aleksandar Uzelac³,
Bogdan Radakovic³, and Nikola Todic³

¹ University of Caen, France
doucet@info.unicaen.fr

² Microsoft Research Cambridge, United Kingdom
gabkaz@microsoft.com

³ Microsoft Development Center Serbia
{bodind,aleksandar.uzelac,bogdan.radakovic,nikola.todic}@microsoft.com

Abstract. This paper describes the setup of the Book Structure Extraction competition run at ICDAR 2009. 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 faced was to construct hyperlinked tables of contents for a collection of 1,000 digitized books. This paper describes the setup of the competition and its challenges. It introduces and discusses the book collection used in the task, the collaborative construction of the ground truth, the evaluation measures and the evaluation results. The paper also introduces a data set to be used freely for research evaluation purposes.

1 Introduction

Mass-digitization projects, such as the Million Book project⁴, efforts of the Open Content Alliance⁵, and the digitization work of Google⁶ 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

⁴ <http://www.ulib.org/>

⁵ <http://www.opencontentalliance.org/>

⁶ <http://books.google.com/>

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 [15].

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 2). In Section 3, 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 4, while the proposed measures for the evaluation of the participating systems' performance are described in Section 5. The approach chosen to construct the ground truth and obtained data are presented in Section 6. The results of the Structure Extraction competition are given in Section 7. We conclude with a summary of the competition and our future plans in Section 8.

2 Background

Motivated by the need to foster research in areas relating to large digital book repositories, see e.g., [8], the Book Track was launched in 2007 as part of the Initiative for the Evaluation of XML retrieval (INEX)⁷. 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 [7], which have been investigated at INEX since 2002 [4,5]. 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 need.

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 [9]. In 2008, two new tasks were introduced, including the book structure extraction task [11]. 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, in 2008, ran as a “beta” 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 [10] and at the International Conference on Document Analysis and Recognition (ICDAR) 2009 [3].

⁷ <http://www.inex.cs.otago.ac.nz/>

This round builds up on the established infrastructure with an extended test set of 1,000 digitized books.

3 Competition Setup

3.1 Goals

The goal of the structure extraction competition at ICDAR 2009 was 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?

3.2 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 tables of contents for each book in the test set. The OCR text of each book is stored in DjVu XML format (see Section 4). 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.

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.

Participants were invited to submit up to 10 runs, each run containing the ToCs for all 1,000 books in the test set.

The ToCs created by participants were compared to a manually built ground truth; see Sections 5 and 6 for details on the annotation process and the evaluation measures.

3.3 Participating Organizations

Following the call for participation issued in April 2009, 11 organizations registered. They are listed in Table 1. Several organizations have expressed interest but renounced participation due to time constraints. Of the 11 organizations that

```

<!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 were expected to submit to the competition, containing the generated hyperlinked ToC for each book in the test set.

| Organization | Submitted runs | Ground truthing |
|-------------------------------------|----------------|-----------------|
| Dublin City University | 0 | n |
| Exalead Inc. | 0 | n |
| Fraunhofer Institute | 0 | y |
| Microsoft Development Center Serbia | 1 | y |
| Noopsis Inc. | 1 | n |
| Peking University | 0 | y |
| Texas A&M | 0 | y |
| University of Amsterdam | 0 | n |
| University of Caen | 3 | y |
| University of Firenze | 0 | y |
| Xerox Research Centre Europe | 3 | y |

Table 1. Registered participants and activity.

signed up, 3 dropped out, that is, they neither submitted runs, nor participated in the ground truth annotation process.

Interestingly, among the organizations that signed up and did not manage to send runs, more than half (4 out of 7) still contributed to the ground truth creation, possibly suggesting their intent to participate in forthcoming rounds of the competition. However, contribution to the ground truth was also the sole condition upon which access to the compiled ground truth was then given. 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.

4 Book Collection

The corpus of the INEX book track contains 50,239 digitized, out-of-copyright books, provided by Microsoft Live Search and the Internet Archive [11].

The set of books used in the book structure extraction competition comprises 1,000 books selected from the INEX book corpus. It consists of 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 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 selected 200 books that do not contain a printed ToC into the total set of 1,000. 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 pages that contain the printed ToC (so called 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 any 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 was 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.

5 Measures

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:

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 in “3. His Birth and First Years” or even the word “chapter” as in “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 non-alphanumeric substitution, deletion and insertion remains 1) [12]:

Two strings A and B are “sufficiently similar” if

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

is less than 20% and if the distance between their first and last five characters (or less if the string is shorter) 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 [14] 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 (see Section 4), 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 2).

| | 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% |

Table 2. An example score sheet summarizing the performance evaluation of the “MDCS” run.

6 Ground Truth Creation

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. This section describes the ground truth annotation process and its outcomes.

6.1 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 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. A screen capture of the tool is shown in Figure 2. In the application window, the right-hand side displays the baseline ToC with clickable (and editable) links. The left-hand side shows the current page and allows to navigate through the book. The JPEG image of each visited page is downloaded from the INEX server at www.booksearch.org.uk and is locally cached to limit bandwidth usage.

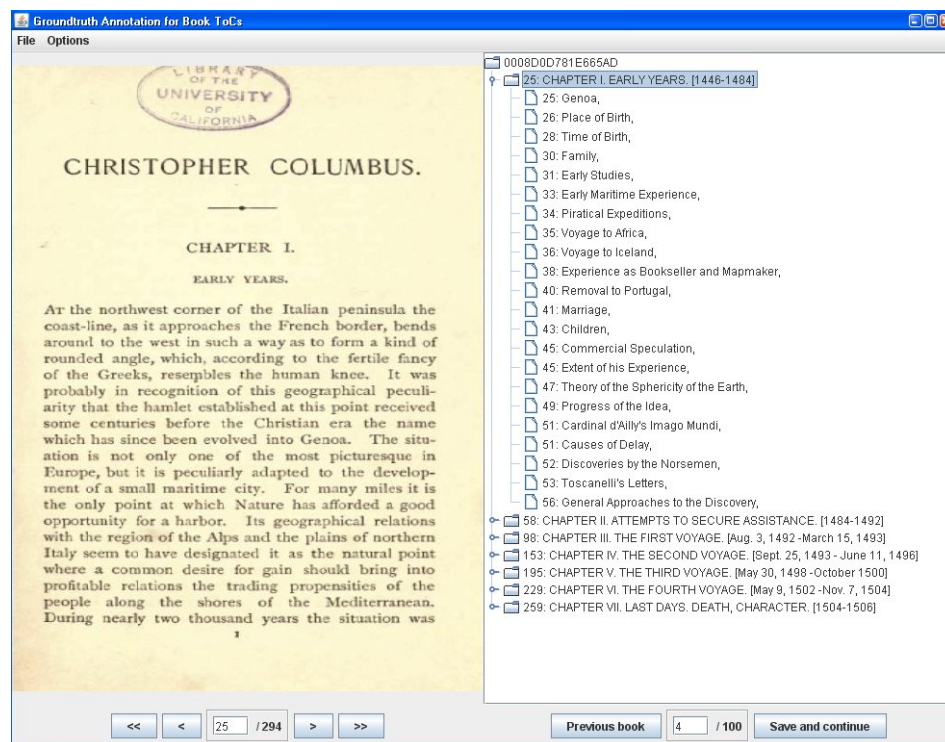


Fig. 2. A screen shot of the ground truth annotation tool.

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 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.

6.2 Collected Ground Truth Data

7 teams participated in the ground truth annotation process, 4 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 ToC was reviewed by the organizers, and a significant number had to be removed. Any ToC with annotation errors were then removed. Errors were most of the time 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. 97 of the annotated books are ones for which no ToC pages were detected.

Freely available ground truth. To facilitate the participation of other institutions in the future, it was decided to make available the ground truth set of the 100 ToCs that were built during the first Book Structure Extraction task at INEX 2008 [11]. This ground truth set is available from the competition’s website, together with the corresponding evaluation software⁸.

Consistency of the annotation. As this competition is in its early years, 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 (see Section 5). The result of this comparison is given in Table 3).

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 disagreement stems 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 as an excessive

⁸ <http://www.info.unicaen.fr/~doucet/StructureExtraction/#training>

| | 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% |

Table 3. The score sheet measuring annotator agreement for the 61 books that were assessed independently by two distinct institutions.

increase would lead to uniform results (more duly distinct titles would be deemed equivalent).

7 Results

A summary of the performance of all the submitted runs, based on F-measure for complete entries (see entry in bold in Table 2) is given in Table 4.

| RunID | Participant | F-measure (complete entries) |
|------------|--------------------|------------------------------|
| MDCS | MDCS | 41.51% |
| XRCE-run2 | XRCE | 28.47% |
| XRCE-run1 | XRCE | 27.72% |
| XRCE-run3 | XRCE | 27.33% |
| Noopsis | Noopsis | 8.32% |
| GREYC-run1 | University of Caen | 0.08% |
| GREYC-run2 | University of Caen | 0.08% |
| GREYC-run3 | University of Caen | 0.08% |

Table 4. Summary of performance scores for the Structure Extraction competition 2009.

The score sheets corresponding to each of the runs is available online⁹.

7.1 Approaches presented

Descriptions of the approaches by the participants revealed that MDCS [13] and Noopsis focused on exploiting the contents of ToC pages. They made no use of the rest of the books, except for the purpose of page linking (that is, to find the right page number corresponding to a ToC entry). The technique employed by MDCS consists of three steps: recognizing ToC pages, assigning

⁹ <http://www.info.unicaen.fr/~doucet/StructureExtraction/2009/#results>

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.

On the other hand, the technique followed by the University of Caen (GREYIC) [6] 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. They report that a bug in the assignment of the physical page numbers unfortunately made their results hard to interpret.

XRCE’s approach is entirely unsupervised [2]. An interesting first step is the removal of all headers and footers which are said to be a common cause of error. Each page in a book is then assigned a physical page number. The detection of ToC and index pages is unsupervised and keyword-based. Each ToC page is then “segmented” into ToC entries using the references to page numbers. The runs also experiment with the use of “trailing whitespaces”, a feature very similar to that used by the University of Caen. The impact of this feature is said to be very promising.

7.2 Detailed Results

The fact that an identified portion of the 1,000 books does not contain a ToC section allows for a separate evaluation over this subcollection. This part of the corpus is of particular interest, since ToC extraction in this case permits a clear improvement for the digitized version of a book compared to its printed version: the addition of a table of contents, which does not exist in the printed version.

Results for subset with books without ToC pages

The results, calculated over the set of 97 books in the ground truth that did not contain ToC pages are presented in Table 5.

| RunID | Participant | F-measure (titles) | F-measure (complete entries) |
|-------------|-------------|--------------------|------------------------------|
| XRCE-run3 | XRCE | 10.79% | 7.81% |
| XRCE-run2 | XRCE | 10.69% | 7.63% |
| XRCE-run1 | XRCE | 5.07% | 3.55% |
| Noopsis | Noopsis | 1.47% | 0.87% |
| MDCS | MDCS | 0.71% | 0.13% |
| GREYIC-run1 | U of Caen | 15.65% | 0.13% |
| GREYIC-run2 | U of Caen | 15.72% | 0.13% |
| GREYIC-run3 | U of Caen | 16.09% | 0.13% |

Table 5. Summary of performance scores over books without a detected ToC.

Focusing the evaluation on this subset highlights the difference between the techniques that exploit the presence of the printed ToCs to techniques that extract structure from the inner-content of the books, e.g., by searching for titles throughout the book and not only within the ToC pages. From the results it is clear that much remains to be done before these approaches can be deemed reliable. Compared with the best performance of 41.51% achieved over the whole test set (see Table 4), the best performance on the subset of 97 books without a printed ToC present is only 7.81% for complete entries, achieved by XRCE. When the F-measure is calculated over titles only, the best performance of 16.09% is obtained by the University of Caen. Noopsis and MDCS perform relatively poorly on this subset, stemming from their sole focus on extracting structure from ToC pages. their page linking process.

Results for the books with an identified ToC

We also compiled results based on the subset of 430 books that contained ToC pages. These are presented in Table 6. The results are in line with the observations obtained for the whole ground truth set, although the scores are naturally increased due to the removal of the set of books with no ToC pages (which is typically associated with poor performance).

| RunID | Participant | F-measure (complete entries) |
|------------|-------------|------------------------------|
| MDCS | MDCS | 50.84% |
| XRCE-run2 | XRCE | 33.17% |
| XRCE-run1 | XRCE | 33.17% |
| XRCE-run3 | XRCE | 31.73% |
| Noopsis | Noopsis | 10.00% |
| GREYC-run1 | U of Caen | 0.07% |
| GREYC-run2 | U of Caen | 0.07% |
| GREYC-run3 | U of Caen | 0.07% |

Table 6. Summary of performance scores over books with a detected ToC.

7.3 Alternative Measure

Participants were encouraged to propose alternative metrics, and Meunier and Déjean introduced the XRCE link-based measure 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 [2].

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 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 to evaluate the performance of systems works by matching ToC entries primarily based on links rather than titles. The corresponding results are given in Table 7. As it can be seen, the results improve as possible errors in the titles no longer lead to whole ToC entries being discounted.

| RunID | Precision | Recall | F-measure |
|------------|-----------|--------|-----------|
| MDCS | 65.9% | 70.3% | 66.4% |
| XRCE-run3 | 69.7% | 65.7% | 64.6% |
| XRCE-run2 | 69.2% | 64.8% | 63.8% |
| XRCE-run1 | 67.1% | 63.0% | 62.0% |
| Noopsis | 46.4% | 38.0% | 39.9% |
| GREYC-run1 | 6.7% | 0.7% | 1.2% |

Table 7. Performance scores for the Structure Extraction competition 2009 based on the XRCE link-based metrics.

The “complete entries” measure, used as a reference in most of this paper 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 in Section 5 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 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 at the competition’s web site¹⁰.

8 Summary and future plans

A strength of the conjoint organization between INEX and ICDAR 2009, is that while the results were announced at the ICDAR conference, 26-29 July 2009, in Barcelona, Spain, the participants had the opportunity to submit papers describing their approaches to the INEX 2009 Workshop, which was held in December 2009 in Brisbane, Australia.

¹⁰ <http://www.info.unicaen.fr/~doucet/StructureExtraction/>

We plan to continue running the competition in the coming years. This was also requested by several participants intending to return, as well by several other institutions who were developing their structure extraction systems this year. Some of these groups participated this year by contributing to the building of the ground truth set, even though they did not manage to submit any runs.

Another motivating reason to continue the competition is evidenced by the current results, indicating that much could still be improved upon, especially in the case of books that do not contain ToC pages.

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.

Acknowledgements

The Structure Extraction Task is supported by the Document Layout Team of Microsoft Development Center Serbia, who notably developed the evaluation software.

We would also like to thank student Paul Cercueil who implemented most of the ground truth annotation tool over a work placement at the University of Caen.

Finally, we are grateful to participants for their annotation effort, as well as for their numerous questions and suggestions.

References

1. K. Coyle. Mass digitization of books. *Journal of Academic Librarianship*, 32(6):641–645, 2006.
2. Hervé Déjean and Jean-Luc Meunier. XRCE participation to the book structure task. In Geva et al. [5].
3. Antoine Doucet, Gabriella Kazai, Bodin Dresevic, Aleksandar Uzelac, Bogdan Radakovic, and Nikola Todic. ICDAR 2009 Book Structure Extraction Competition. In *Proceedings of the Tenth International Conference on Document Analysis and Recognition (ICDAR'2009)*, pages 1408–1412, Barcelona, Spain, July 2009.
4. Shlomo Geva, Jaap Kamps, and Andrew Trotman, editors. *Advances in Focused Retrieval: 7th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2008)*, volume 5613 of *Lecture Notes in Computer Science*. Springer Verlag, Berlin, Heidelberg, 2009.
5. Shlomo Geva, Jaap Kamps, and Andrew Trotman, editors. *Advances in Focused Retrieval: 8th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2009)*, *Lecture Notes in Computer Science*. Springer Verlag, Berlin, Heidelberg, 2010.

6. Emmanuel Giguet, Alexandre Baudrillart, and Nadine Lucas. Resurgence for the book structure extraction competition. In Geva et al. [5].
7. Jaap Kamps, Shlomo Geva, and Andrew Trotman. Report on the SIGIR 2008 workshop on focused retrieval. *SIGIR Forum*, 42(2):59–65, 2008.
8. Paul Kantor, Gabriella Kazai, Natasa Milic-Frayling, and Ross Wilkinson, editors. *BooksOnline '08: Proceeding of the 2008 ACM Workshop on Research Advances in Large Digital Book Repositories*, New York, NY, USA, 2008. ACM.
9. Gabriella Kazai and Antoine Doucet. Overview of the INEX 2007 Book Search Track (BookSearch'07). *ACM SIGIR Forum*, 42(1):2–15, 2008.
10. Gabriella Kazai, Antoine Doucet, Marijn Koolen, and Monica 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, page 16. Springer, 2010.
11. Gabriella Kazai, Antoine Doucet, and Monica Landoni. Overview of the INEX 2008 Book Track. In Geva et al. [4].
12. V. I. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10:707+, February 1966.
13. Aleksandar Uzelac, Bodin Dresevic, Bogdan Radakovic, and Nikola Todic. Book layout analysis: TOC structure extraction engine. In Geva et al. [4].
14. C. J. van Rijsbergen. *Information retrieval*. Butterworths, London, 2 edition, 1979.
15. Roelof van Zwol and Tim van Loosbroek. Effective use of semantic structure in XML retrieval. In Giambattista Amati, Claudio Carpineto, and Giovanni Romano, editors, *ECIR*, volume 4425 of *Lecture Notes in Computer Science*, pages 621–628. Springer, 2007.