The Book Structure Extraction Competition with the Resurgence Full Content Software at Caen University

Emmanuel Giguet and Nadine Lucas

GREYC Cnrs. Caen Basse Normandie University
BP 5186 F-14032 CAEN Cedex France
name.surname@unicaen.fr

Abstract. The GREYC participated in the Structure Extraction Competition, part of the INEX/ICDAR Book track, for the third time, with the Resurgence software. We used a minimal strategy primarily based on full-content top-down document representation with two then three levels, part, chapter and section. The main idea is to use a model describing relationships for elements in the document structure. Frontiers between high-level units are detected. The periphery center relationship is calculated on the entire document and then reflected on each page. The weak points of the approach are that level hierarchy is implicit, and dependent on named levels. It does not fit with the chapter and section levels reflected in the ground-truth. The strong points are that it deals with the entire document; it handles books without ToCs, and extracts titles that are not represented in the ToC (e. g. preface); it is tolerant to OCR errors and language independent; it is simple and fast. A test on sections was run after the competition to help understand the evaluation issues with more than two levels.

1 Introduction

The GREYC laboratory participated for the third time in the Book Structure Extraction Competition part of the INEX ICDAR evaluations in 2011 [1]. The extraction software Resurgence used at Caen University does not rely on the ToC pages but on the full content of the books. The experiment was conducted from pdf documents to ensure the control of the entire process. The document content is extracted using the pdf2xml software [2]. The original Resurgence software processes small documents, academic articles (mainly in pdf format) and news articles (mainly in HTML format) in various information extraction tasks and text parsing tasks [3].

In 2009, Resurgence handled only the chapter level [4] and in 2010 it handled part and chapter levels [5]. Surprisingly, better results were obtained when parts and chapters were evaluated irrespective of level. Since GREYC was the only participant in 2010, the experiment was reiterated in 2011 in runs 1 and 2 for part and chapter levels. We studied the effect of a complex hierarchy on the ground truth and on evaluation. In run 3, a test was made on three levels including sections with numbered series.

In the following, we explain our method on the 2011 ICDAR book corpus challenge. Results are compared in the two evaluation grids, ICDAR and link-based (Xerox) in section 3. In section 4 we discuss ways to correct our system and better handle sections.

S. Geva, J. Kamps, and R. Schenkel (Eds.): INEX 2011, LNCS 7424, pp. 86–97, 2012.

After the competition, we tested the section level again with more success as expected. In the last section, we point at some inconsistencies or difficulties in the ground-truth constitution and make proposals for future competitions with an enhanced annotation tool.

2 A Differential Book Structure Extraction Method

2.1 Challenges

The size of the book corpus is the first challenge. Resurgence was modified in order to load the necessary pages only. The objective was to allow processing on usual laptop computers.

The fact that the corpus was OCR documents also challenged our original program that detects the structure of electronic academic articles. A new branch in Resurgence had to be written in order to deal with scanned documents. The document parsing principles were tested on two levels of the book hierarchy at a time, part (meaning here a book part including a number of chapters) and chapter. Experiments on sections in run 3 with a new method are also reported, and will be further explained in section 4 with a new corrected run outside the competition.

2.2 Strategy

The strategy in Resurgence is based on document positional representation. and does not rely on the table of contents (ToC). This means that the whole document is considered first. Then document constituents are considered top-down (by successive subdivision). with focus on the middle part (main body) of the book. The document is thus the unit that can be broken down ultimately to pages.

The main idea is to use a model describing relationships for elements in the document structure. The model is a periphery-center dichotomy. The periphery center relationship is calculated on the entire document and reflected on each page. The algorithm aims at retrieving the book main content bounded by annex material like preface and post-face with different layout. It ultimately retrieves the page body in a page, surrounded by margins [4].

We adopted the principle to get systematically down the book structure hierarchy one level at a time. For this experiment, we focused on part (if any) and chapter title detection, so that the program detects two levels. i. e. part titles and chapter titles in runs 1 and 2 as in 2010. The transition page between two parts or between two chapters is characterized in a sliding window of four pages as detailed in [5].

In run 3 some elaboration on title detection using “longitudinal” series at a given level was tempted as detailed below. In run 3 we also included three levels, chapter, part and section detection.

2.3 Title Extraction Strategy

Title extraction is conducted in four steps for all three levels. First, selection of would-be numbered titles; second, reconstruction of series names through creation of an equivalence table for each series; third, series validation through numbers; fourth, starting point detection.

2.3.1 Selection of Candidate Titles

A regular expression detects characters patterns:

a) sharing the same layout;
b) placed on the same line;
c) beginning with a capitalized word followed by a number (Arabic. roman or ordinal).

Note that in practice. extracted series may contain for example:

- CHAPTEE.
- THE.
- BOOK.
- Chapter

2.3.2 Series Names Reconstruction

Series names candidates are checked throughout the document. A global test checks if there are at least two successive series name candidates in the book for the same level (as derived from position and layout).

The "word" before the number, also called prefix, is kept in memory, if its frequency is above 1. The idea is to detect series names candidates as prefixes, such as *Chaptee, Book*, without being blocked by a strong expectation on a given wording. This is to avoid both OCR errors and misses when series wording varies from conventional use, with *Poem* or *Sermon* instead of *Chapter,* or *Book* instead of *Part.*

Thus, a series comprising some OCR errors like CHAPTER I ... CHAPTEK V … CHAPTEE XI is considered as a good text segment candidate provided the Levenshtein distance between two wordings is small (below 20%). The prefix variants will be considered as equivalent to the most frequent wording, thus CHAPTEK and CHAPTEE will be equivalent to CHAPTER. Note that in practice some extracted series may still be deemed incorrect, if there are more OCR errors than correct titles. This has no importance for the structuration task. A correction rule could be applied on the entire collection for search engines tasks, later on.

2.3.3 Series Name Validation

Once the series name is fixed, the numbers are checked with some tolerance. The idea is to find one or several grossly growing series in number with an equivalent series title, considered as a prefix. In the example below, some numbers are missing (typically first chapters are more difficult to detect). Some others have been sliced by return commands, so the series is awkward.

- CHAPTER: III IV V VI VII VIII IX X XII XIII XIV XV XVI XVII XVIII XIX XX XXI XXII XXIII XXIV XXV XXVI XXVII XXVIIL XXIX XXX XXXI XXXII XXXIII XXXIV II III IV V VI VII VIII IX X XL XII XIII XIV XV XVI XVII XVIII XIX XX XXI XXIV XXV XXV\

I XXVII XXVIII XXIX XXX XXXI XXXII II III IV V VI VII VIII IX X XL XII XIII XIV XV XVI XVII XVIII XIX XX XXL XXII XXIV XXV XXVI XXVII XXVIII II III IV V VI VII VIII IX X XII XIII XIV XV XVI XVIII XIX XX XXI XXII XXIII XXIV XXV XXVI XXVI\

I XXVIII XXIX XXX XXXI XXXII XXXIII XXXIV XXXV XXXVI XXXVII II III IV V VI VII VIII IX X XI XII XIII XV XVI XVII XVIII XIX XX XXL XXII XXIII XXIV XXV XXVI XXVII XXVIII XXIX XXX XXXI XXXII XXXIV XXXV XXXVI XXXVII XXXIX XL

However, increasing series are found no withstanding some holes or redundancies. One or more such series will be considered correct as a plausible level prefix, here chapter level series. The same will apply to a shorter series at another level, book parts.

- BOOK: II III IV IV V

On the contrary, some wordings selected as prefix at stage 1 will be forgotten, because they are not followed by a grossly growing series of numbers. It might have been a title such as *The second world war.*

- THE:

Last, a series of numbers without any increase will also be forgotten.

- Chapter: XX XVII XIX

2.3.4 Starting Point Detection

In order to find often overlooked chapters, mainly first chapters or sections, the starting point for titles series was established at the beginning of the main body, that is, after the ToC if any. Thus, a procedure to detect would-be ToCs was applied.

2.4 Calibrating the System

On the practical side, the team was interested in handling voluminous documents, such as textbooks and cultural heritage books. Working on the whole document requires the ability to detect and deal with possible heterogeneous layouts in different parts of the document (preface. main body. appendices). Layout changes can impact page formatting (e.g.. margin sizes. column numbers) as well as text formatting (e.g.. font sizes. text alignments) [6].

The standard page structure recognition has been improved by a better recognition of the shape of the body, which is not strictly rectangular in scanned books [5]. Line detection, standard line height and standard space height detection were also improved. They are important in our approach, because the standard line is the background against which salient features such as large blanks and title lines can be detected. The improvements in line computation improved the results in chapter detection as explained in [5].

However, the hierarchy consolidation was not implemented.

2.5 Experiment

The corpus provided in 2011 was similar in size to the 2009 one. It comprised 998 books (as compared with 1114 books in 2010 and 1000 in 2009, some empty) [1, 10].

The GREYC 2011 program detected only part and chapter titles in run 1 and 2. The top-down strategy and the highest levels in the book hierarchy were favoured because this is the most useful step when filtering large book collections, in text mining tasks for instance. Moreover, most if not all known techniques start from the lower levels [7]. Reasonable results can be obtained for those levels with existing programs once the relevant parts or chapters have been retrieved.

There was only one run to test section detection, run 3. However, due to a bug in document numbers, it ran astray. Run 4 was added after the competition to test the strategy explained in 2.2, at the section level as well. It will be discussed separately.

3 Results

3.1 General Results

The official results for 2011 are reproduced in Table 1, against a ground-truth of 513 books. GREYC missed one book of the ground-truth. It is at the fourth and last rank. The entire corpus was handled, with 60 misses. The very bad results in run 3 were due to a bug in document numbers.

Table 2 shows the F-link measure, with the same ranking.

Table 1. F-measure evaluation 2011 on 2011 ground-truth (513 books)

RunID	Participant	F-measure (complete entries)
MDCS	Microsoft Development Center Serbia	40.75%
Nankai-run1	Nankai University. China	33.06%
Nankai-run4	Nankai University	33.06%
Nankai-run2	Nankai University	32.46%
Nankai-run3	Nankai University	32.43%
XRCE-run1	Xerox Research Centre Europe	20.38%
XRCE-run2	Xerox Research Centre Europe	18.07%
GREYC-run2	GREYC University of Caen. France	8.99%
GREYC-run1	GREYC University of Caen. France	8.03%
GREYC-run3	GREYC University of Caen. France	3.30%

Table 2. F-link evaluation 2011 on 2011 ground-truth (513 books)

RunID	Participant	F-link
MDCS	Microsoft Development Center Serbia	65.1%
Nankai-run1	Nankai University, China	63.2%
Nankai-run4	Nankai University, China	63.2%
Nankai-run2	Nankai University	59.8%
Nankai-run3	Nankai University	59.8%
XRCE-run2	Xerox Research Centre Europe	58.1%
XRCE-run1	Xerox Research Centre Europe	57.6%
GREYC-run1	GREYC University of Caen, France	50.7%
GREYC-run2	GREYC University of Caen, France	50.7%
GREYC-run3	GREYC University of Caen, France	24.4%

3.2 Greyc Results Evolution

These results are compared with the GREYC official evaluation in 2009 best run and with 2010 in Table 3.

Table 3. Official evaluation 2009 to 2010 on the 2009 ground-truth (527 books)

Results 2009	**Precision**	**Recall**	**F-Measure**
Titles	19.83%	13.60%	13.63%
Levels	16.48%	12.08%	11.85%
Links	1.04%	0.14%	0.23%
Complete entries	0.40%	0.05%	0.08%
Entries disregarding depth	1.04%	0.14%	0.23%
Results 2010			
Titles	18.03%	12.53%	12.35%
Levels	13.29%	9.60%	9.34%
Links	14.89%	7.84%	7.86%
Complete entries	14.89%	10.17%	10.37%
Entries disregarding depth	10.89%	7.84%	4.86%

Table 4. GREYC 2010 and 2011 evaluation with Xerox linked-based metrics

	XRCE Link-based Measure			
	Links			Title accuracy (for valid links)
	Precision	Recall	F1	
GREYC 2010	63.9%	39.5%	42.1%	47.6%
GREYC 2011 - run2	65.2%	49.9%	50.7%	46.2%
GREYC 2011 – run3	32.5%	24.5%	24.4%	31.1%

Table 4 shows the evaluation based on links and initially provided by Xerox Research Center Europe (XRCE).

The 2011 results slightly outperform the 2010 results as expected for chapter detection. This is mainly explained by improvements in the system calibration. Little gain is obtained from part detection, as in 2010. This is due to the fact that most book parts are not signalled in the ground-truth. Even if it were, the number of parts is low (and even often null in individual books), as compared to the total number of titled sections to be found throughout the collection. But the main interest in this year evaluation was to assess the effect of multilevel description with series. The failure of run 3 was a bad blow. It was found that chapter and sections are the two levels on which annotation focuses, being mainly based on ToCs.

4 Discussion

GREYC was the only candidate in 2010, so comparison with others was not possible. It was worth re-evaluating results on roughly the same corpus, and the same method, through runs 1 and 2. Moreover, we tested a new method in run 3, based on numbered series, and including sections. It is level independent but not quite lexicon-free. It was corrected in post run 4, to evaluate the benefits of this strategy.

The ground-truth annotation was not easy since we had to browse entire books, which took an enormous time with slow response delays to “turn” pages. We worked with a Mac, which could be a plea. It was not possible to establish two levels and save them explicitly as such through the menu. Therefore, the reference cannot be deeper than two levels, try as we may. As far as we saw, chapters and sections were the only levels used by other participants. Parts including chapters would as a consequence be judged as false when detected, as well as titled sub-sections in chapters.

4.1 Reflections on the Experiment

4.1.1 Extra Run

GREYC corrected a bug concerning document id numbers in Run 3 including section level and using series. The corrected run is called Corrected Run-4 and it obtained significantly better results.

Table 5 shows results given for the official best GREYC run for two levels (part and chapter) in run 2, and the best results for three levels (part. chapter and section) with document number correction in the post competition GREYC-Corrected Run 4. Fusion of position clues with series validation proved efficient.

Table 5. Comparison of two-level and three-level results for GREYC 2011

	F1 Link	F1 Inex Link
GREYC-run 2	50.7%	10.8%
GREYC corrected run 4	58.8%	20.1%

However, the need to propagate these principles to all the sub-levels of the book hierarchy (such as sub-sections) was not felt. This is because the subsections are seldom accompanied by a prefix, which is part of the recognition pattern used by GREYC to extract title series. As a consequence, many numbered but un-titled sections and subsections will go unnoticed. A different strategy has to be found for deep subdivisions. Moreover, the subsections are seldom kept in ToCs and the ground-truth also ignores them.

4.1.2 Comparison

On the scientific side, some strong points of the Resurgence program were ascertained. They are based on relative position and differential principles. The advantages are the following:

- The program deals with the entire document body, not on the table of contents;
- It handles books without table of contents (ToC), and titles that are not represented in the ToC (e. g. preface). It would be most welcome if the annotated corpus could be checked directly inside the book when looking for errors;
- It is dependent on typographical position, which is very stable in the corpus, despite heterogeneous domains and styles;
- It is not dependent on lexicon, or very little in run 3. Hence it is tolerant to OCR errors and it is language independent;
- Last, it is simple and fast.

The advantage of using the book body is clear when comparing two datasets, books without ToC and books with ToC [6, 9]. The difference is clearer in the GREYC case with the link-based measure.

Table 6. Comparison of 2009 results on two books datasets after [6]

	whole dataset (precision / recall)	no-ToC dataset (precision/ recall)
MDCS	65.9 / 70.3	0.7 / 0.7
XRCE	69.7 / 65.7	30.7 / 17.5
NOOPSIS	46.4 / 38.0	0.0 / 0.0
GREYC-1C 2009	59.7 / 34.2	48.2 / 27.6

Another advantage is robustness. Since no list of memorized forms is used, but position and distribution instead, fairly common strings are extracted, such as CHAPTER or SECTION, but also uncommon ones, such as PSALM or SONNET. When chapters have no numbering and no explicit mention such as *chapter*, they are found as well, for instance a plain title stating “Christmas Day”.

Resurgence took advantage on numbering of titles series through many steps in 2011: since numbers are an important source of OCR errors, a tolerant pattern recogniser is used. This approach reflects an original breakthrough to improve robustness and proves very useful to generate ToCs to help navigate digitized books when none was provided in the printed version (20% of the corpus).

4.2 Reflections on Evaluation Measures

Concerning evaluation rules. the very small increment in quantified results did not reflect our qualitative assessment of a significant improvement in numbered series.

Generally speaking, the ground-truth is still very coarse and it mostly relies on automated results depending on the ToC [9. 1]. If the ToC is the reference, it is an error to extract prefaces, for instance, because they generally do not figure in ToCs. In the same way, most ToCs do not reflect the whole hierarchy of sections and subsections, but skip lower levels. The participants using the book body as main reference are penalized if they extract the whole hierarchy of titles as it appears in the book, when the ToC represents only higher levels.

For all participants, accuracy on titles seems to be a thorny question, because there is a huge difference in title accuracy as calculated by INEX organizers from the retrieval of the wording, and title accuracy as calculated by XRCE from the links [1, 7]. In the INEX08-like measure on accuracy for title and level provided by XRCE, the figures decrease while precision and recall grow.

A test was made to evaluate level accuracy, since proceeding one level at a time allowed a relevance check on this measure. In 2009 GREYC calculated only chapters and the level accuracy was high, 73.2. in the GREYC results, after correction on the document id bug. Scores in level accuracy in 2010 were calculated with part and chapter level information and then without part and chapter level information to check consistency (Table 7).

Table 7. GREYC link-based evaluation with and without level information against the 2009 ground-truth as compared with 2011 evaluation and ground-truth

	XRCE Link-based Measure				**Inex08 like Accuracy**	
	Links			**Accuracy** for valid links		
	Precision	**Recall**	**F1**	*Title*	*Title*	*Level*
GREYC-1C 2009	59.7	34.2	38.0	13.9	42.1	73.2
GREYC 2010	64.4	38.9	41.5	47.6	22.3	64.2
GREYC 2010 without level info	64.4	38.9	41.5	47.6	22.3	**77.9**
GREYC 2011 run 2	65.2	49.9	50.7	46.2	21.9	**80.4**

In 2011, title accuracy was lower but level accuracy was slightly better. GREYC reached the best official relative level accuracy among all participants with a 80.4% score, followed by MDC at 79.2%, as shown in Table 8.

Since GREYC was the only candidate working from the actual book body layout and not after the ToC, results suffered from the fact that ToC when present — in 80% of the cases — is used as the baseline reference in the ground-truth [1]. However, there are significant differences between ToC and book titles as reported in [5, 6].

Table 8. 2011 alternative link-based evaluation against the 2011 cleaned ground-truth (513 books), compared with Inex-like accuracy depending on title recognition

	F-link	Titl-acc	RelLevel-accuracy	F~Inex Link	Inex Titl-acc	Level-acc
MDCS	65.1%	83.7%	79.2%	47.6%	69.1%	79.6%
NANKAI-1	63.2%	74.4%	76.3%	40.9%	54.2%	77.2%
NANKAI-2	59.8%	75.9%	75.5%	40.1%	56.6%	76.4%
NANKAI-3	59.8%	75.9%	75.5%	40.1%	56.5%	76.3%
NANKAI-4	63.2%	74.4%	76.3%	40.9%	54.2%	77.2%
UNICAEN-1	50.7%	46.2%	61.4%	10.8%	21.9%	61.3%
UNICAEN-2	50.7%	46.2%	80.4%	10.8%	21.9%	80.4%
UNICAEN-3	24.4%	31.1%	64.0%	4.2%	11.2%	63.9%
XRCE-1	57.6%	60.9%	78.6%	24.8%	43.8%	78.6%
XRCE-2	58.1%	63.7%	77.9%	23.5%	40.1%	77.9%

Table 9. GREYC 2011 runs in the two measures (level accuracy in bold)

	F-link	Titl-acc	RelLvl-acc	F~Inex Link	Inex Titl-acc	Level-acc
UNICAEN-1	50.7%	46.2%	61.4%	10.8%	21.9%	61.3%
UNICAEN-2	50.7%	46.2%	**80.4%**	10.8%	21.9%	**80.4%**
UNICAEN-3	24.4%	31.1%	64.0%	4.2%	11.2%	63.9%
UNICAEN-Corrected 4	56.5%	56.8%		20.1%	33.1%	**78.4%**

The scores for corrected run 4 were calculated using the download package [8] but the new item Relative Level accuracy was not included.

5 Proposals

The bias introduced by a semi-automatically constructed ground-truth was salient as can be seen in the example above, where split words or added *pp.* at the end of the entry illustrate poor quality against human judgment. Manually corrected annotation is still to be checked to improve the ground-truth quality. As mentioned in [1] quantitative effort is also needed, but it is time-consuming. Crowdsourcing was considered a better solution to minimize annotator's discrepancies [10].

However, it might not be realistic to expect a clean unique reference for a large book collection. It might be better to handle parameters according to the final aim of the book processing, such as navigation or information filtering. Thus known automatic biases might be countered or even valued in the performance measure according to real use.

It would be very useful to provide results by normalized title depth (level) as suggested by [5, 7], because providing complete and accurate results for one or more levels would be more satisfying than missing some items at all levels. It is important to get coherent and comparable text spans for many tasks, such as indexing, helping navigation or pre-processing for text mining.

The reason why the beginning and end of the titles are overrepresented in the evaluation scores is not clear and a more straightforward edit distance for extracted titles should be provided.

One simple idea used in the 2011 evaluation was to consider equally results for titles matching with either the ToC or the book body, with or without a prefix such as *Chapter* [1].

Despite shortcomings, mostly due to early stage development, the book structure extraction competition was very interesting. The corpus provided for the INEX /ICDAR Book track is the best available corpus offering full books at document level [1, 9, 10]. Although it comprises mostly XIXth century printed books, it is very valuable, for it provides various types of layout. Besides, this corpus meets our requirements for electronic use of patrimonial assets. The ground-truth is manually corrected, so that the dataset is easier to work with than the dataset provided by [11].

Some efforts should be exerted to improve the interface used to annotate books, so that the whole title hierarchy can be clearly and conveniently marked. Accordingly, accurate level measures reflecting the human judgement could trigger better automatic recognition.

References

1. Doucet, A., Kazai, G., Meunier, J.-L.: ICDAR 2011 Book Structure Extraction Competition. In: 11th International Conference on Document Analysis and Recognition (ICDAR 2011), pp. 1501–1505 (2011)
2. Giguet, E., Lucas, N., Chircu, C.: Le projet Resurgence: Recouvrement de la structure logique des documents électroniques. In: JEP-TALN-RECITAL 2008 Avignon (2008)
3. Déjean, H., Giguet, E.: pdf2xml open source software, `http://sourceforge.net/projects/pdf2xml/` (last update February 25, 2011; last visited February 2012)
4. Giguet, E., Lucas, N.: The Book Structure Extraction Competition with the Resurgence Software at Caen University. In: Geva, S., Kamps, J., Trotman, A. (eds.) INEX 2009. LNCS, vol. 6203, pp. 170–178. Springer, Heidelberg (2010)
5. Giguet, E., Lucas, N.: The Book Structure Extraction Competition with the Resurgence Software for Part and Chapter Detection at Caen University. In: Geva, S., Kamps, J., Schenkel, R., Trotman, A. (eds.) INEX 2010. LNCS, vol. 6932, pp. 128–139. Springer, Heidelberg (2011)

6. Déjean, H., Meunier, J.-L.: Document: a useful level for facing noisy data. In: 4th Workshop on Analytics for Noisy Unstructured Text Data (AND 2010), Toronto, Canada, pp. 3–10 (2010)
7. Déjean, H., Meunier, J.-L.: Reflections on the INEX structure extraction competition. In: 9th IAPR International Workshop on Document Analysis Systems (DAS 2010), pp. 301–308. ACM, New York (2010), doi:10.1145/1815330.1815369
8. Source forge, `https://sourceforge.net/projects/inexse/`
9. Doucet, A., Kazai, G., Dresevic, B., Uzelac, A., Radakovic, B., Todic, N.: Setting up a Competition Framework for the Evaluation of Structure Extraction from OCR-ed Books. International Journal of Document Analysis and Recognition (IJDAR) 14(1), 45–52 (2010)
10. Kazai, G., Koolen, M., Kamps, J., Doucet, A., Landoni, M.: Overview of the INEX 2010 Book Track: Scaling Up the Evaluation Using Crowdsourcing. In: Geva, S., Kamps, J., Schenkel, R., Trotman, A. (eds.) INEX 2010. LNCS, vol. 6932, pp. 98–117. Springer, Heidelberg (2011)
11. Vincent, L.: Google Book Search: Document understanding on a massive scale. In: 9th International Conference on Document Analysis and Recognition (ICDAR 2007), pp. 819–823. IEEE (2007)