The Book Structure Extraction Competition with the Resurgence Software for Part and Chapter Detection 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 Island team participated in the Structure Extraction Competition part of the INEX Book track for the second time, with the Resurgence software. We used a minimal strategy primarily based on top-down document representation with two levels, part and chapter. The main idea is to use a model describing relationships for elements in the document structure. Frontiers between high-level units are detected, parts and then chapters. Page is also used. The periphery center relationship is calculated on the entire document and reflected on each page. The strong points of the approach are that it deals with the entire document; it handles books without ToCs, and titles that are not represented in the ToC (e. g. preface); it is not dependent on lexicon, hence tolerant to OCR errors and language independent; it is simple and fast.

1 Introduction

The GREYC Island team participated for the second time in the Book Structure Extraction Competition part of the INEX evaluations [6]. The INEX Resurgence software used at Caen University to structure voluminous documents was modified to handle book parts, on top of chapters. The original Resurgence software processes academic articles (mainly in pdf format) and news articles (mainly in HTML format) in various text parsing tasks [8].

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]. In the first experiment, we handled only the chapter level [7]. We still could not propagate our principles on all the levels of the book hierarchy at a time. We consequently focused on the higher levels of book structure, part and chapter detection.

In the following, we explain our strategy and we detail the actual results on the INEX book corpus, both the 2010 corpus and the 2009 one. In section 3, we discuss the advantages of our method and make proposals for future competitions. In the last section we sum up our contribution.

S. Geva et al. (Eds.): INEX 2010, LNCS 6932, pp. 128–139, 2011.

2 Our Book Structure Extraction Method

2.1 Challenges

In the first experiment, the huge memory needed to handle books was found to be indeed a serious hindrance, as compared with the ease in handling academic articles: pdf2xml required up to 8 Gb of memory and Resurgence required up to 2 Gb to parse the content of large books (> 150 Mb). This was due to the fact that the whole content of the book was stored in memory. The underlying algorithms did not actually require the availability of the whole content at a time. 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 previous program that detected the structure of electronic academic articles. A new branch in Resurgence had to be written in order to deal with scanned documents. We propagated our document parsing principles on two levels of the book hierarchy at a time, part (meaning here part including a number of chapters) and chapter, hoping for an improvement of the results, but two levels proved insufficient to boost the quality.

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 [8].

Implementation Rules. For this experiment, we focused on part (if any) and chapter title detection so that the program detects only two levels, i. e. part titles and chapter titles.

Chapter Title Detection throughout the document was conducted using a sliding window to detect chapter transitions with two patterns, as explained in [7].

Chapter title extraction is made from the first third of the top of the page body. The model assumes that the title begins at the top of the page. The title right end is detected, by calculating the line height disruption: a contrast between the would-be title line height and the rest of the page line height. A constraint rule allows a number of lines containing at most 40 words.

Parts Wrapping Chapters are detected throughout the document using a sliding window of one page. The idea is to detect a page with few written lines. The transition page between two parts is characterized as follows:

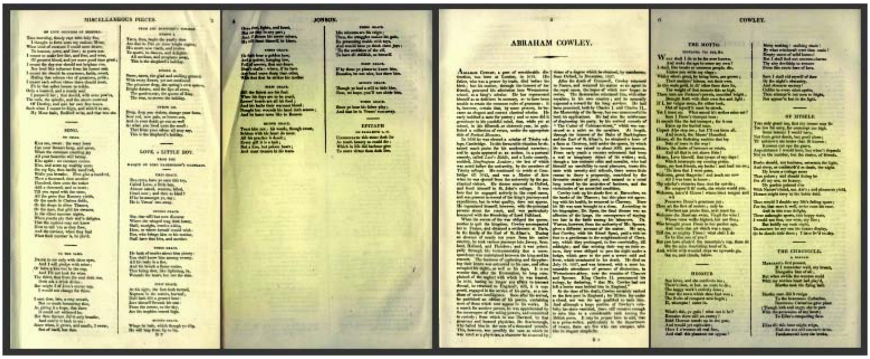


Fig. 1. View of the four-page sliding window to detect chapter beginning. Pattern 1 matches. Excerpt from 2009 book id = 00AF1EE1CC79B277.

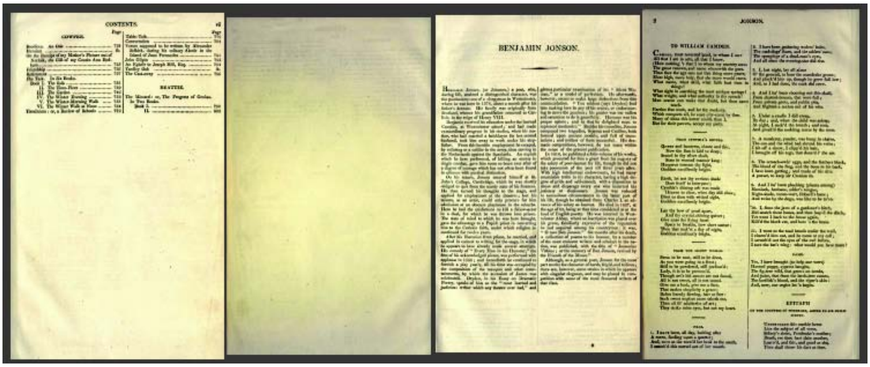


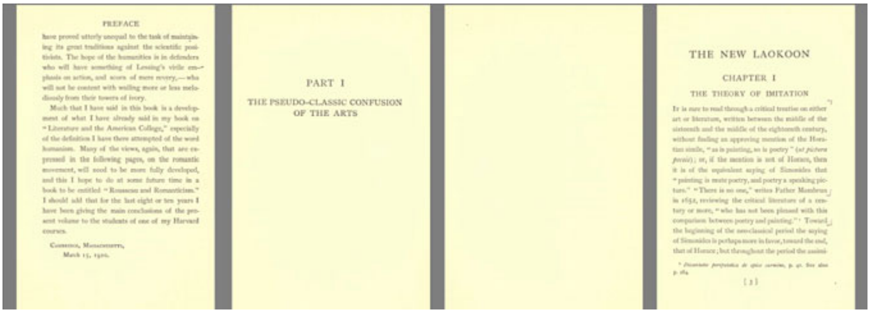
Fig. 2. View of the four-page sliding window to detect chapter announced by a blank page. Pattern 2 matches. Excerpt from 2009 book id= 00AF1EE1CC79B277.

- the text body in the page is mainly blank,
- with a blank line at least 5 times the standard line space height;
- followed by 1 to 3 written lines;
- with a blank line at least 5 times the standard line space height.

A global test checks if there are at least two successive parts in the book. Figure 3 illustrates the pattern. It applies on a single page. 3 context pages are given but are not used in the process.

2.3 Calibrating the System

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).



PREFACE

have proved utterly unequal to the task of maintaining its great traditions against the scientific positivists. The hope of the humanities is in defenders who will have something of Lessing's virile emphasis on action, and scorn of mere revery,—who will not be content with wailing more or less melodiously from their towers of ivory.

Much that I have said in this book is a development of what I have already said in my book on "Literature and the American College," especially of the definition I have there attempted of the word humanism. Many of the views, again, that are expressed in the following pages, on the romantic movement, will need to be more fully developed, and this I hope to do at some future time in a book to be entitled "Rousseau and Romanticism." I should add that for the last eight or ten years I have been giving the main conclusions of the present volume to the students of one of my Harvard courses.

Cambridge, Massachusetts,
March 15, 1910.

PART I

THE PSEUDO-CLASSIC CONFUSION OF THE ARTS

THE NEW LAOKOON

CHAPTER I

THE THEORY OF IMITATION

It is rare to read through a critical treatise on either art or literature, written between the middle of the sixteenth and the middle of the eighteenth century, without finding an approving mention of the Horatian simile, "as is painting, so is poetry" (*ut pictura poesis*); or, if the mention is not of Horace, then it is of the equivalent saying of Simonides that "painting is mute poetry, and poetry a speaking picture." "There is no one," writes Father Mambrun in 1652, reviewing the critical literature of a century or more, "who has not been pleased with this comparison between poetry and painting."[1] Toward the beginning of the neo-classical period the saying of Simonides is perhaps more in favor, toward the end, that of Horace; but throughout the period the assimi-

[1] *Dissertatio peripatetica de epico carmine*, p. 41. See also p. 184.

[3]

Fig. 3. View of the one-page sliding window to detect parts beginning. 3 context pages (1 page before, two pages after) are given but not used. Excerpt from 2009 book id = 2A5029E027B7427C.

The standard page structure recognition has been improved, by correcting a bug in the previous program that impaired the recognition of page header and footer [7]. It has also been improved by a better recognition of the shape of the body, which is not always rectangular in scanned books.

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.

The standard line height and standard line space height are computed in the following way. The most frequent representative intervals are computed to cope with OCR variation in line height. In the previous experiment, line height was calculated somewhat rigidly, after pdf2xml was used. Line recognition was dependent on bounding box heights. However, this is not very reliable for scanned text, and the program tended to create loose line segments. Moreover it also tended to artificially augment the line height, due to the presence of one capital letter for example, and thus the standard line was not contrasted against title lines, which are slightly bigger.

In the current experiment, the model drives the detection process. This means that unless there is a strong clue against it, the line is considered as continue. The line common characteristics are favored against occasional disruptions in bounding box height.

2.4 Experiment

The corpus provided in 2010 was extended as compared with the 2009 one. It comprised 1114 books instead of 1000 in 2009. The GREYC 2010 program detected only part and chapter titles. No effort was exerted to find the section titles and sub-titles. On the practical side, the team was interested in handling voluminous documents, such as textbooks and cultural heritage books, hence the interest in INEX. 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 techniques start from the lower levels. 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.

2.5 Results

GREYC Results. The entire corpus was handled, but 26 books were not analyzed. This was due to lack of time, since we could not use parallel computing this time, as we did in 2009 for the pdf2XML task.

The official results for 2010 are given in Table 1 and are compared with the first official evaluation in 2009 in Table 2. However, the very bad results in 2009 were due to a bug in page numbers.

Table 3 shows the complementary alternative evaluation based on links and provided by Xerox Research Center Europe (XRCE).

The 2010 results outperform the 2009 results as expected. This is mainly explained by improvements in the system calibration. Little gain is obtained from part detection. This is due to the fact that the number of parts is low (and even often null in individual books), as compared to the total number of sections

Table 1. Official evaluation 2010 on 2010 ground truth (641 books)

Results 2010	Precision	Recall	F-Measure
Titles	18,03%	12,53%	12,33%
Levels	13,29%	9,60%	9,34%
Links	14,89%	10,17%	10,37%
Complete entries	10,89%	7,84%	7,86%
Entries disregarding depth	14,89%	10,17%	10,37%

Table 2. Official evaluation 2009 against 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%

Table 3. Alternative 2010 evaluation with Xerox linked-based metrics

	XRCE Link-based Measure			
	Links			Accuracy (for valid links)
	Precision	Recall	F1	Title
GREYC 2010	63.9	39.5	42.1	47.6

and subsections to be found throughout the collection. The official evaluation does not distinguish false responses from nonresponses, so all sections titles are considered as false.

Comparison. GREYC was the sole participant in 2010 and was evaluated on a slightly extended book collection (1114 books). The ground truth contained 641 books. Hence, results could not be compared directly with the 2009 participants results on the 2009 corpus.

Results were therefore compared against the 2009 ground truth that comprised 527 books using the Python evaluation toolkit provided on sourceforge [3]. GREYC results were compared with results obtained by the 2009 official best run (run 3), 2009 results after correction of the page bug by Xerox (called GREYC-1 in [7,5] and GREYC-1C 2009 in the tables below to avoid confusion with runs), and the 2010 results. They were also compared with the results obtained on the same ground truth by other participants in 2009 (Table 4).

Table 4. Alternative evaluation comparing all participants against the 2009 ground truth

	XRCE Link-based Measure			
	Links			Accuracy (for valid links)
	Precision	Recall	F1	Title
MDCS	65.9	70.3	66.4	86.7
XRCE-run3	69.7	65.7	64.6	74.4
Noopsis	46.4	38.0	39.9	71.9
GREYC run 3	6.7	0.7	1.2	13.9
GREYC-1C 2009 page bug correction	59.7	34.2	38.0	42.1
GREYC 2010	64.4	38.9	41.5	47.6

Table 5 shows another tentative measure, provided by [4,3] to check if accuracy measured on title wording (INEX 08 like measure) could be useful, along with a level accuracy measure based on correct title retrieval. Results are given for the official best GREYC run (run 3) and for the results with page number correction GREYC-1C.

Table 5. Alternative accuracy evaluation with INEX 08 like measures for all participants, against the 2009 ground truth

	INEX08 like measure	
	Accuracy	
	Title	Level
MDCS	86.7	75.2
XRCE-run3	74.4	68.8
Noopsis	71.9	68.5
GREYC run 3	0.0	31.4
GREYC-1C 2009	42.1	73.2
GREYC 2010	22.3	64.2

3 Discussion

GREYC was the only candidate this year, but since official results in 2009 suffered from a gross error in page numbers, it was worth re-evaluating results on a comparable corpus. Moreover, the book part level was also tested. The low recall is still due to the fact that the full hierarchy of titles was not addressed as mentioned earlier. This will be addressed in the future.

3.1 Reflections on the Experiment

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

On the scientific side, some strong points of the Resurgence program, based on relative position and differential principles, were better implemented. We intend to further explore this way. 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 is dependent on typographical position, which is very stable in the corpus;
- It is not dependent on lexicon, hence tolerant to OCR errors and language independent.

Last, it is simple and fast.

The fact that typographical position is very stable in the corpus reflects real-life conventions in book printing.

The advantage of using the book body is clear when comparing two datasets, books without ToC and books with ToC [5,1]. The difference is clearer in the GREYC case with the link-based measure.

Another advantage is robustness. Since no list of expected and 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

Table 6. Comparison of results on two books datasets after [5]

	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

mention such as *chapter*, they are found as well, for instance a plain title stating “Christmas Day”. Resurgence did not rely on numbering of chapters: this is an important source of OCR errors. Hence they were retrieved as they were by our robust extractor. 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.

3.2 Reflections on Evaluation Measures

Concerning evaluation rules, the very small increment in quantified results did not reflect our qualitative assessment of a significant improvement. Though this is a subjective impression that seems fairly common, we were puzzled.

Generally speaking, the ground truth is still very coarse and it mostly relies on automated results depending on the ToC. 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]. In the INEX08-like measure on accuracy for title and level provided by XRCE, the figures decrease while precision and recall grow.

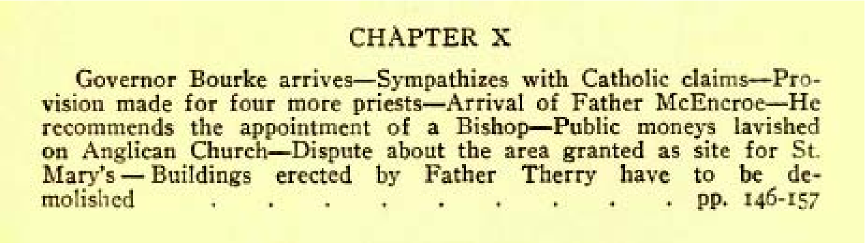
A test was made to evaluate level accuracy, since proceeding one level at a time allowed a relevance check on this measure. In 2009 GREYC calculated only chapters and the level accuracy was high, 73.2, in the GREYC results, after correction on the page bug. Scores in level accuracy in 2010 were calculated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

Level accuracy according to title wording was called Inex08 like measure after [4] The difference in level accuracy raises questions. Results in 2009 (GREYC-1C with page bug correction) were given for one level only, namely chapters. The submitted 2010 GREYC results with level information for part and chapter levels had a level accuracy of 64.2, but when level information was scrapped, it was better, 77.9. It should be the contrary. Another intriguing observation is that linked-based title accuracy and text-based level accuracy did not evolve together. Our guess is that level accuracy is not relevant, for it is calculated from the XML with relative depth for each book, and not against a standard layout scale for the entire collection.

Title accuracy improved in 2010 according to both XRCE link based measure and Inex08-like title wording based measure. This is explained by a better rightward segmentation, already tested in 2009 after the runs and mentioned in [7]. However, title accuracy according to the INEX08 like calculation sharply collapsed between 2009 and 2010.

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

	XRCE Link-based Measure					
	Links			Title accuracy	Inex08 like Accuracy	
	Precision	Recall	F1	for valid links	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



CHAPTER X

Governor Bourke arrives—Sympathizes with Catholic claims—Provision made for four more priests—Arrival of Father McEncroe—He recommends the appointment of a Bishop—Public moneys lavished on Anglican Church—Dispute about the area granted as site for St. Mary's — Buildings erected by Father Therry have to be demolished . . . . pp. 146-157

Fig. 4. View of the ToC entry. Excerpt from 2010 book id =A803EBAC7E50C7D0.

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 is used as the reference in the groundtruth. However, there is a significant difference between ToC and book titles. Sometimes, the mention *chapter* was not found in the book or was abbreviated as C in the ToC. Although differences in case, such as CHAPTER III in the book and Chapter 3 in the ToC are cared for in the evaluation, by using case insensitive option, differences in numbering and word segmentation penalized our results. The edit distance error margin seems to be wide, but we tried a Levenshtein distance error margin of 20% and found it is not sufficient, confirming other findings as suggested by [5,1].

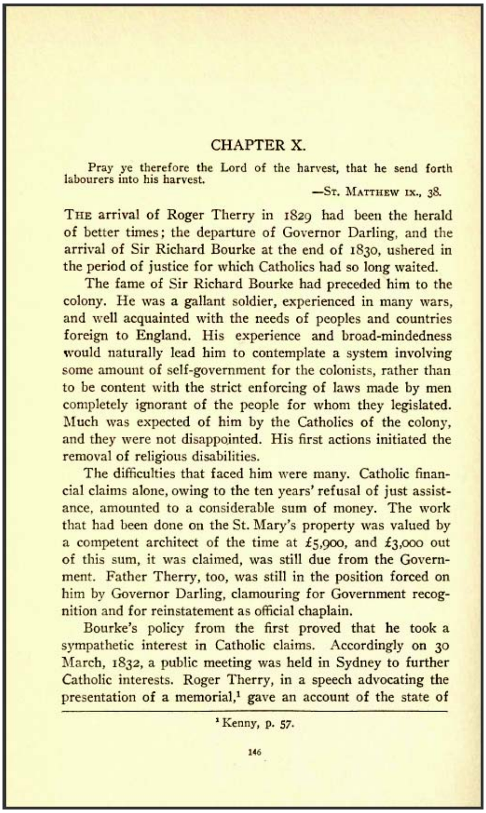
In some cases, detailed subentries were included in the chapter title, while they are not present in the ToC, or vice versa, as explained in [5]. All these details explain very low results in title accuracy. Figure 4 compared with Figure 5 shows an example where subentries are included in the ToC and in the reference deriving from it, but not on the corresponding page in the book body. Here is the ground truth entry for the book *Life of Archpriest Therry* (book id A803EBAC7E50C7D0) Chapter X, p. 146.

```
<toc-entry title="CHAPTER X Governor Bourke arrives Sympathizes
with Catholic claims Pro vision made for four more priests Arrival
of Father McEncroe He recommends the appointment of a Bishop
Public moneys lavished on Anglican Church Dispute about the area
granted as site for St. Mary s Buildings erected by Father Therry
have to be de molished pp." page="206" />
```

4 Proposals

The bias introduced by a semi-automatically constructed ground truth was salient as can be seen in Figure 4 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 [9,1] quantitative effort is also needed, but it is time-consuming. 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 valued in the performance measure.

One simple idea would be to consider equally right results for titles matching with either the ToC or the book body. It might be a good idea to give the bounding box containing the title as a reference for the ground truth. This solution would solve conflicts between manual annotation and automatic annotation,



CHAPTER X.

Pray ye therefore the Lord of the harvest, that he send forth labourers into his harvest.

—St. Matthew ix., 38.

The arrival of Roger Therry in 1829 had been the herald of better times; the departure of Governor Darling, and the arrival of Sir Richard Bourke at the end of 1830, ushered in the period of justice for which Catholics had so long waited.

The fame of Sir Richard Bourke had preceded him to the colony. He was a gallant soldier, experienced in many wars, and well acquainted with the needs of peoples and countries foreign to England. His experience and broad-mindedness would naturally lead him to contemplate a system involving some amount of self-government for the colonists, rather than to be content with the strict enforcing of laws made by men completely ignorant of the people for whom they legislated. Much was expected of him by the Catholics of the colony, and they were not disappointed. His first actions initiated the removal of religious disabilities.

The difficulties that faced him were many. Catholic financial claims alone, owing to the ten years' refusal of just assistance, amounted to a considerable sum of money. The work that had been done on the St. Mary's property was valued by a competent architect of the time at £5,900, and £3,000 out of this sum, it was claimed, was still due from the Government. Father Therry, too, was still in the position forced on him by Governor Darling, clamouring for Government recognition and for reinstatement as official chaplain.

Bourke's policy from the first proved that he took a sympathetic interest in Catholic claims. Accordingly on 30 March, 1832, a public meeting was held in Sydney to further Catholic interests. Roger Therry, in a speech advocating the presentation of a memorial,[1] gave an account of the state of

[1] Kenny, p. 57.

146

Fig. 5. View of first page of Chapter X (p. 146). Excerpt from 2010 groundtruth, book id = A803EBAC7E50C7D0.

leaving man or machine to read and interpret the content of the bounding box. It would also alleviate conflicts between ToC-based or text-based approaches.

Concerning details, it should be clear whether or when the prefix label indicating the book hierarchy level (Chapter, Section, and so on) and the numbering should be part of the extracted result. The chapter title is not necessarily preceded by such mentions, but in other cases there is no specific chapter title and only a number. The ground truth is not clear either on the extracted title case: sometimes the case differs in the ToC and in the actual title in the book.

It would be very useful to provide results by title depth (level) as suggested by [4,5], 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 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.

The time is ripe for eliciting effective face-to-face interaction between participants, to stimulate discussion and make the evaluation rules evolve faster. This should entice new participants to enter an important field of development with a lively discussion ahead.

5 Conclusion

The paper presents a strategy of detecting parts and chapters in a book without the use of the table of contents, using only layout features of the scanned pages. The strategy is mostly easy to follow and is reproduceable.

Acknowledgments. We thank student Alexandre Baudrillart who implemented new features in the program during his summer job at GREYC, Caen University. We also thank anonymous reviewers for their excellent comments and suggestions.

References

1. 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), Special Issue on Performance Evaluation of Document Analysis and Recognition Algorithms 14(1), 45–52 (2011)
2. Déjean, H.: pdf2xml open source software, http://sourceforge.net/projects/pdf2xml/ (last visited March 2010)
3. Déjean, H., Meunier, J.-L.: INEX structure extraction groundtruth, https://sourceforge.net/projects/inexse/ (last update 2010-09-29, last visited March 2011)
4. Déjean, H., Meunier, J.L.: XRCE Participation to the Book Structure Task, pp. 124–131. Springer, Heidelberg (2009), http://portal.acm.org/citation.cfm?id=1611913.1611928

5. Déjean, H., Meunier, J.L.: Reflections on the INEX structure extraction competition. In: Proceedings of the 9th IAPR International Workshop on Document Analysis Systems, DAS 2010, pp. 301–308. ACM, New York (2010), http://doi.acm.org/10.1145/1815330.1815369
6. Doucet, A., Kazai, G.: Icdar 2009 book structure extraction competition. In: IEEE (ed.) 10th International Conference on Document Analysis and Recognition ICDAR 2009, Barcelona, Spain, pp. 1408–1412 (2009)
7. 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)
8. Giguet, E., Lucas, N., Chircu, C.: Le projet Resurgence: Recouvrement de la structure logique des documents électroniques. In: JEP-TALN-RECITAL 2008, Avignon, France (2008)
9. Kazai, G., Doucet, A., Koolen, M., Landoni, M.: Overview of the INEX 2009 book track. In: Geva, S., Kamps, J., Trotman, A. (eds.) INEX 2009. LNCS, vol. 6203, pp. 145–159. Springer, Heidelberg (2010), http://portal.acm.org/citation.cfm?id=1881065.1881084
10. Vincent, L.: Google book search: Document understanding on a massive scale. In: Proceedings of the Ninth International Conference on Document Analysis and Recognition - Volume 02, pp. 819–823. IEEE Computer Society, Washington, DC, USA (2007), http://portal.acm.org/citation.cfm?id=1304596.1304903