Proceedings of the 4th Financial Narrative Processing Workshop @ LREC 2022, pages 83–88
Marseille, 24 June 2022

The Financial Document Structure Extraction Shared Task (FinTOC 2022)

Abderrahim Ait Azzi[1], Sandra Bellato[1], Blanca Carbajo Coronado[2], Mahmoud El-Haj[3], Ismail El Maarouf[1], Mei Gan[1], Ana Gisbert[2], Juyeon Kang[1], Antonio Moreno Sandoval[2]

Fortia Financial Solutions[1], Paris, France
Universidad Autónoma de Madrid[2], Madrid, Spain
Lancaster University[1], Lancaster, UK
{abderrahim.aitazzi, sandra.bellato, mei.gan, ismail.elmaarouf, juyeon.kang}@fortia.fr[1]
{blanca.carbajo, ana.gisbert, antonio.msandoval}@uam.es[2]
m.el-haj@lancaster.ac.uk[3]

Abstract

This paper describes the FinTOC-2022 Shared Task on the structure extraction from financial documents, its participants results and their findings. This shared task was organized as part of The 4th Financial Narrative Processing Workshop (FNP 2022), held jointly at The 13th Edition of the Language Resources and Evaluation Conference (LREC 2022), Marseille, France (El-Haj et al., 2022). This shared task aimed to stimulate research in systems for extracting table-of-contents (TOC) from investment documents (such as financial prospectuses) by detecting the document titles and organizing them hierarchically into a TOC. For the forth edition of this shared task, three subtasks were presented to the participants: one with English documents, one with French documents and the other one with Spanish documents. This year, we proposed a different and revised dataset for English and French compared to the previous editions of FinTOC and a new dataset for Spanish documents was added. The task attracted 6 submissions for each language from 4 teams, and the most successful methods make use of textual, structural and visual features extracted from the documents and propose classification models for detecting titles and TOCs for all of the subtasks.

Keywords: Financial Data Annotation, Document Structure Extraction, Table-Of-Contents Extraction, Machine Learning

1. Introduction

A vast amount of financial documents are created and published constantly in machine-readable formats (generally PDF file format), with only minimal structure information. Firms use such documents to report their activities, financial situation or potential investment plans to shareholders, investors and the financial markets, basically corporate annual reports containing detailed financial and operational information.

In some countries as in the US or in France, regulators such as EDGAR SEC or AMF require firms to follow a certain template when reporting their financial results to ensure standardization and consistency across firms' disclosures. In other European countries, on the other hand, the management usually has more discretion on what, where and how to report resulting in lack of standardization between financial documents published within the same market.

Existing work on book and document table of contents (TOC) recognition has been almost all on small size, application-dependent, and domain-specific datasets. However, TOC of documents from different domains differ significantly in their visual layout and style, making TOC recognition a challenging problem for a large scale collection of heterogeneous documents and books. Compared to regular books (mostly provided in a full text format with limited structural information such as pages and paragraphs), Financial documents, containing textual and non textual content, have a more sophisticated structure including, parts, sections, subsections, sub-sub-sections.

In this shared task, we focus on analyzing two types of financial documents: 1) Fund Prospectuses, official PDF documents in which investment funds precisely describe their characteristics and investment modalities, and 2) financial annual reports, publicly available PDF documents on which firms publish a year-end summary of their operations and financial conditions. In the case of the fund prospectuses, although the content they must include is often regulated, their format is not standardized and displays a great deal of variability ranging from plain text format, towards more graphical and tabular presentation of data and information. The layout information becomes more heterogeneous from a company to another in the case of the annual reports as there is no regulations on their document structure. While the majority of annual reports often contain a simplified table of contents (TOC), the majority of prospectuses are published without a TOC, which is usually needed to help readers to navigate within the document by following a simple outline of headers and page numbers, and assist legal teams in checking if all the contents required are fully included in both cases. Thus, automatic analyses of those documents to ex-

tract their structure is becoming more and more vital to many firms across the world.

Thanks to the contribution of the Autonomous University of Madrid (UAM, Spain) (Moreno-Sandoval et al., 2020), the fourth edition of the FinTOC shared task proposes the same welcomes a new track for Spanish documents in addition to English and French, and it will score systems on both Title detection and TOC generation performance as has been the practice from previous editions.

In this paper, we report the results and findings of the FinTOC-2022 shared task[1]. The Shared Task was organized as part of The 4th Financial Narrative Processing Workshop (FNP 2022)[2], to be held at The 13th Edition of the Language Resources and Evaluation Conference (LREC 2022)[3].

The shared task attracted 6 system submissions from 4 teams for each language and for the Title Detection and TOC extraction tasks. In general, the systems which make use of textual, structural and visual features, and exploit observed features during classification models training for the Title Detection and TOC extraction, perform better.

2. Previous Work on Document structure extraction

Previous work can be divided into two approaches for the TOC extraction. The first approach parses the hierarchical structure of sections and subsections from the TOC pages embedded in the document. This area of research was mostly motivated by the INEX ((Dresevic et al., 2009)) and ICDAR competitions ((Doucet et al., 2013), (Beckers et al., 2010); (Nguyen et al., 2017)) which aim at extracting the TOC of old and lenghtly OCR-ised books. The documents we target in this shared task are very different: they contain graphical elements, and the text is not displayed to respect a linear reading direction but is optimized to condense information and catch the eye of the reader. Apart from these competitions, we find the methods proposed by El-Haj et al. ((El-Haj et al., 2014),(El-Haj et al., 2019)), also based on the parsing of the TOC page.

In the second category of approaches, we find algorithms that detect the titles of the document using learning methods based on layout and text features. The set of titles is then hierarchically ordered according to a predefined rule-based function ((Doucet et al., 2013); (Liu et al., 2011); (Mysore Gopinath et al., 2018)). Lately, we find systems that address the hierarchical ordering of the titles as a sequence labelling task, using neural networks models such as Recurrent Neural Networks and LSTM networks ((Bentabet et al., 2019)). We also see that the large dataset like PubLayNet (Zhong et al., 2019) which contains various annotated elements in a page such as text, list, figure etc. is created based on over 1 million PDF articles and published allowing to lead interesting experiments on the document layout analysis.

3. Task Description

As part of the FNP 2022 Workshop, we present a shared task on Financial Document Structure Extraction. Participants to this shared task were given three sets of financial prospectuses and annual reports with a wide variety of document structure and length. Their systems had to automatically process the documents to extract their document structure, or TOC. In fact, the three sets were specific to three different subtasks:

TOC extraction from French documents The set of French documents is rather homogeneous in terms of structure, due to the existence of a common template. However, the words and phrasing can differ from one prospectus to another. Also, French prospectuses never include a TOC page that could be parsed.

TOC extraction from English documents English prospectuses are characterized by a wide variety of structures as there is no template to constrain their format. Contrary to the French documents, there is always a TOC page but the latter is usually highly incomplete as only the higher level section titles are displayed.

TOC extraction from Spanish documents This year we have introduced the set of documents in Spanish. The reports were chosen for their availability to annotate the titles in the pdf. However, they varied in size and structure, with little uniformity in structure. In this sense, the Spanish reports resemble the English ones. They tend to have TOC and many levels of nesting in the titles (up to 7). In addition, half of the reports do not follow a coherent structure in the section numbering.

3.1. Shared Task Data

In this section, we describe the datasets prepared for the shared task.

Dataset FinToc 2022 proposes enriched datasets for English and French and a new dataset for Spanish financial documents. As the previous editions, we carefully selected documents for each language with a large variety of structures and layouts, see the Figure 1 for a comparative layouts of the documents in different language.

The table 1 shows the statistics of the elaborated datasets for this edition. The average number of titles are 134 for French, 225 for English and 150 for Spanish and the maximum depth of the tiles are 9 for English and French datasets and 7 for Spanish.

The English and French datasets are composed of the financial prospectuses of different companies, published between 2010 and 2021. The Spanish dataset is taken from the FinT-esp corpus (Moreno-Sandoval et al., 2020) and consists of 90 documents with a distribution similar to the French and English datasets for

[1] http://wp.lancs.ac.uk/cfie/fintoc2022/

[2] http://wp.lancs.ac.uk/cfie/fnp2022/

[3] https://lrec2022.lrec-conf.org/en/

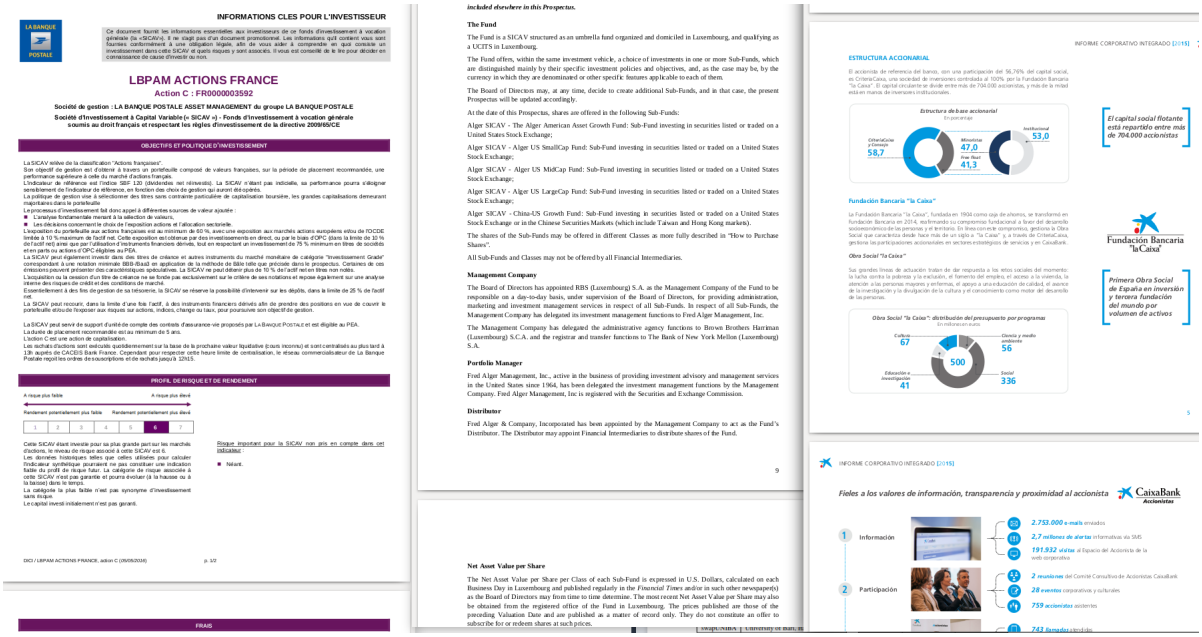

Figure 1: Pages randomly selected from the datasets in French, English and Spanish

	French	English	Spanish
training set	81	79	80
test set	10	10	10
average number of pages	24	90	158

Table 1: Statistics on Dataset

development, validation and test. The dates of the annual reports range from 2014 to 2018. The source is in PDF format, with a total number of pages between 40 and 400. In plain text, the files have an average of 36,285 words. The total number of tags noted in the 90 reports is 10,842, with an average of 148 tags per document.

All the annotated datasets are proposed in simple JSON files containing a list of entries, where each entry has the following information: textual content, id, level, page number (See the example of a JSON in the Figure 2).

Data Annotation Datasets were annotated by the way that the annotators first locate the position of the titles inside each PDF document, then link the title to the entry level in the TOC and give a depth level to each title ranging from 1 to 10. For each of the datasets, three annotators including one as reviewer collaborated to avoid the possible problems like inconsistencies and resolve the possible conflicts during the data annotation.

3.2. Evaluation metrics

FinTOC 2022 uses the evaluation metric as in the previous edition (Maarouf et al., 2021) since the proposed tasks tackle the same problem on different datasets: *Inex F1 score* and *Inex level accuracy*.

We propose two different metrics for each subtask. We use the F1 score for the title detection, meaning that we consider as correct entries the predicted entries which match the titles of groudtruth entries according to the standard Levenshtein distance.

For the TOC extraction, we adapt the metrics proposed by the Structure Extraction Competition (SEC) held at ICDAR 2013 (Doucet et al., 2013) by replacing the customized Levenshtein distance specifically designed for SEC by a standard Levenshtein distance whose edit cost is 1 in all cases, and removing the constraint on first and last 5 characters. The final ranking is based on the harmonic mean between *Inex F1 score* and *Inex level accuracy*. The *Inex F1 score* considers as correct entries in the predicted TOC those which match the title of an entry in the TOC groundtruth and have the same page number as this entry. The *Inex level accuracy* evaluates the hierarchy of the predicted TOC. If we denote by E_{ok} an entry in the predicted TOC with a correct page number, and by E'_{ok} an entry in the predicted TOC with a correct page number and a correct hierarchical level, then the *Inex level accuracy* is:

$$\frac{\sum E'_{ok}}{\sum E_{ok}}$$

For both tasks, the threshold on the Levenshtein score was set to 0.85.

4. Participants and Systems

A total of 24 teams registered this year to FinTOC Shared Task from different academic and private institutions. 4 teams submitted the systems results all for

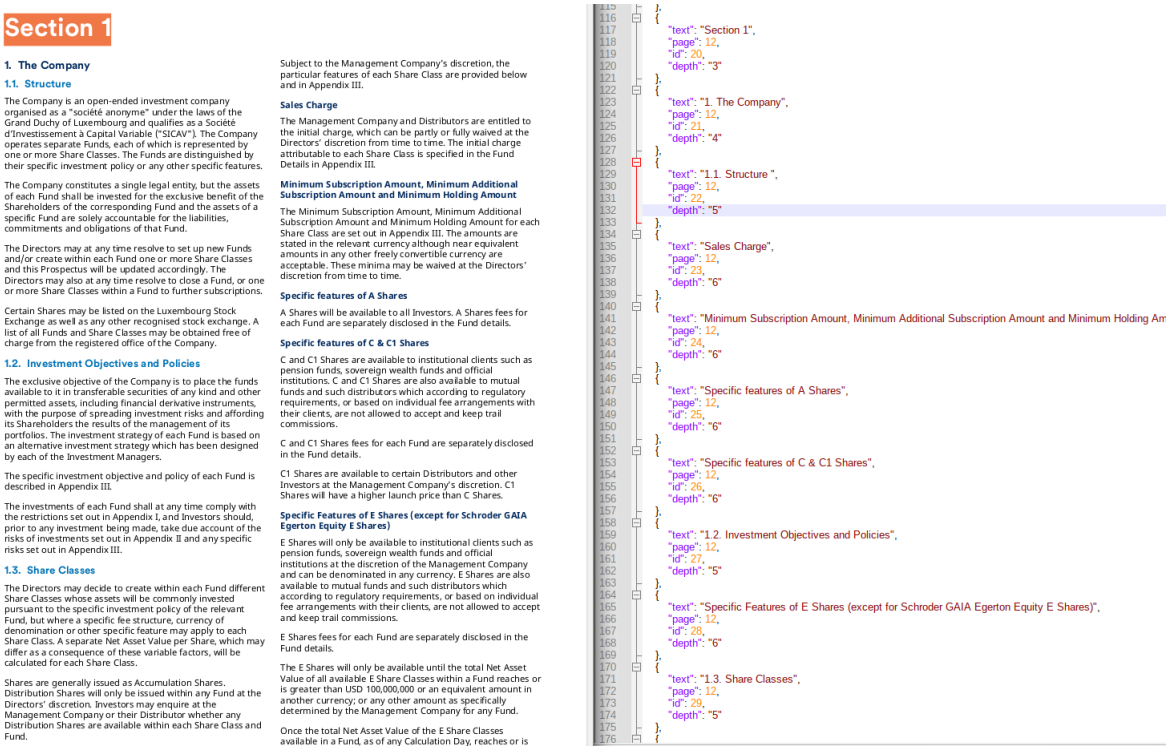

Figure 2: Example of a labeled document in a JSON format with its original PDF document.

three subtasks and 3 teams submitted a system description paper on their method and results as shown in the table 2.

Team	Affiliation
CILAB	KIT, Gumi, Korea
GREYC	CNRS, France
ISPRAS	ISP RAS, Moscow, Russia
swapUNIBA	University of Bari, Italy

Table 2: Participants and affiliations

GREYC ((Giguet and Lucas, 2022)) submitted the results of 2 standard runs on each of the datasets for Title Detection and TOC structure extraction. They propose an end-to-end pipeline which processes documents to first extract textual and visual information of the documents such as token, line, text block, text background, framed content, underline, table grid, bounding boxes of figures and of graphical bullets. Then, using those extracted features, the pipeline performs the document delimitation from the bundle of the datasets, detects header and footer areas from the individual document and applies the Page Layout Analysis which recognizes and labels content areas like texts, tables, figures, lists, headers and footers. They use predefined heuristics to detect a TOC from each document and link the TOC entries to its matching text lines and corresponding page number in the document where the detected line is considered as title.

ISP RAS ((Kozlov et al., 2021)) submitted also the results of 2 standard runs on each of the datasets for the Title Detection and the TOC extraction. They design a full pipeline including two main stages of classification using a decision tree-based algorithm, XGBoost classifier to classify a line as title or not and for each detected title, to find its depth. The PDF documents are preprocessed by PDFMiner and they trained a binary classifier for the first stage and a multiclass classifier for the second, based on the extracted textual, visual and structural features such as color, font style, indentation, list, line depth, letters, words and line statistics, etc. The first run for each language is the result of the classifiers which were trained on the dataset of a specific language, separately, and the second runs are the results of the classifiers trained on all of the datasets.

swapUNIBA ((Cassotti et al., 2022)) submitted the results of 1 standard run on each of the datasets for the Title detection and the TOC extraction. They design a system of Document Image Analysis by exploiting layout features like title, table, list and texts along with an object detector, Faster R-CNN. A pretrained Faster R-CNN model on the PubLayNet dataset was finetuned on the datasets of the shared task for the titles detection, which was preprocessed by pdfplumber. Then level classification module performs the inference of hierarchical level of each title using a multiclass Random Forest classifier trained on the given datasets. At this level, they consider in input the features of a single TOC entry detected by the Title detection module

like first five and last two characters of the text title, font name and size, bounding boxes normalized by the document width and height, etc.

5. Results and Discussion

The scores, based on the metrics described in the Section 3.2, are calculated for each document and then averaged over the documents for each language to produce two performance figures per team submission: one for Title Detection, and another for TOC Extraction. The title detection ranking is based on F1-score, while the TOC extraction ranking is based on the harmonic mean between *Inex F1 score* and *Inex level accuracy*.

Table 3 compares the results of both tasks in terms of the *F1 score* and *Inex level accuracy* on French data. We have two different winning systems for each subtask: ISP RAS1 for the Title Detection and ISP RAS2 for the TOC Extraction. The binary classifier trained only on the French data performs better for the Title detection, while the classifier trained on all the datasets performs better for the TOC extraction.

Team	Title Detection	TOC Extraction
CILAB	0.304	12,90
GREYC1	0.669	7,24
GREYC2	0.671	6,95
ISP RAS1	**0.778**	38,93
ISP RAS2	0.758	**41,58**
swapUNIBA	0.695	34,08

Table 3: Results obtained by the participants for the subtask on French data

Table 4 compares the results of both tasks on English data. Similarly to the results on French data, we also have two different winning systems: ISP RAS1 for the first task and ISP RAS2 for the second, showing that a multilingual dataset can be helpful for improving the overall results.

Team	Title Detection	TOC Extraction
CILAB	0.738	36,99
GREYC1	0.790	0,20
GREYC2	0.793	0,20
ISP RAS1	**0.900**	62,16
ISP RAS2	0.876	**63,17**
swapUNIBA	0.838	51,24

Table 4: Results obtained by the participants for the subtask on English data

Table 5 compares the results of both tasks on Spanish data. We have one winning system for both tasks: swapUNIBA. The best system achieved the F1 score of 0.569% for the title detection and 43,01 for the TOC extraction, indicating that the task needs to be more investigated to solve the problem. But knowing that the Spanish dataset is composed of the annual reports which contain more complex layouts comparing to the fund prospectus documents used in English and French datasets, the produced scores by the systems remain encouraging.

Team	Title Detection	TOC Extraction
CILAB	0.077	8,63
GREYC1	0.196	5,10
GREYC2	0.206	5,22
ISP RAS1	0.554	40,80
ISP RAS2	0.558	40
swapUNIBA	**0.569**	**43,01**

Table 5: Results obtained by the participants for the subtask on Spanish data

Teams submitting multiple systems were able to slightly improve their score within their own submissions, but we did not find that the individual submissions were statistically significantly different. And interestingly, we observe a trade-off from the results of the winning systems on English and French data according to the way that they exploit the datasets as a single dataset or a multilingual dataset (see (Kozlov et al., 2021) for more details.). Since the TOC extraction task depends on the results of the Title detection, the system with a high performance on the Title detection step achieves a high accuracy on the TOC extraction. For English data, the scores for both tasks were significantly improved comparing to those of the previous edition (Maarouf et al., 2021)[4]. Otherwise, both tasks on French and Spanish data are still far from solved.

6. Conclusions

This paper describes the fourth edition of the FinTOC shared task on extraction of the document structure from financial documents. The 6 system submissions from 4 teams for each of the languages, English, French and Spanish, showed that they all exploit textual and visual features extracted from the PDF documents using different text preprocessing tools. Interestingly, the best systems for the Title detection and the TOC extraction on English and French data achieved a good accuracy for the Title detection with a classifier trained on a single dataset while they perform better for the TOC extraction with a classifier trained on a multilingual dataset. More investigation on the error analysis will allow to clarify those impacts. For the Spanish data, the Object Detection approach using a pretrained deep neural model on the large dataset, PubLayNet,

[4]The scores published in the shared task description paper of FinTOC 2021 were miscalculated for the submissions Christopher Bourez1 and 2. The harmonic means are relatively 43,10 and 39 for the TOC extraction on English data and 46,20 and 39 on French data.

performs slightly better than a decision tree-based algorithm. It can be explained by the fact that the datasets used for English and French, and the dataset used for Spanish are quite different in terms of its type (fund prospectuses vs. annual reports), consequently, their structures and layouts are different and the annual reports contain much more visual elements like figures, graphs, tables, bulleted lists, etc. Introducing Spanish fund prospectuses in the shared task data and/or enriching the English and French datasets by adding annual reports would be interesting for the next edition of FinTOC.

7. Acknowledgements

We would like to thank our dedicated annotators who contributed to the building of the corpora used in this Shared Task over the years: Anais Koptient, Aouataf Djillani, Lidia Duarte, Bianca Chong, Marion Cargill, Sandra Bellato, Mei Gan, Anaïs Lhuissier for English and French data, Ana Gisbert, Blanca Carbajo, Ana García, Andrea Castillo, Victoria Matínez, Kateryna Sushkova and Antonio Moreno for Spanish data, and authors of this paper.

8. Bibliographical References

Beckers, T., Bellot, P., Demartini, G., Denoyer, L., De Vries, C. M., Doucet, A., Fachry, K. N., Fuhr, N., Gallinari, P., Geva, S., et al. (2010). Report on inex 2009. In *ACM SIGIR Forum*, volume 44:1, pages 38–57. ACM New York, NY, USA.

Bentabet, N.-I., Juge, R., and Ferradans, S. (2019). Table-of-contents generation on contemporary documents. *arXiv preprint arXiv:1911.08836*.

Cassotti, P., Musto, C., de Gemmis, M., Lekkas, G., and Semeraro, G. (2022). swapuniba@fintoc2022: Fine-tuning pre-trained document image analysis model for title detection on the financial domain. In *Proceedings of Language Resources and Evaluation (LREC'22)*, Marseille, France, June. European Language Resources Association (ELRA).

Doucet, A., Kazai, G., Colutto, S., and Mühlberger, G. (2013). Icdar 2013 competition on book structure extraction. In *2013 12th International Conference on Document Analysis and Recognition*, pages 1438–1443. IEEE.

Dresevic, B., Uzelac, A., Radakovic, B., and Todic, N. (2009). Book layout analysis: Toc structure extraction engine. In Shlomo Geva, et al., editors, *Advances in Focused Retrieval*, pages 164–171, Berlin, Heidelberg. Springer Berlin Heidelberg.

El-Haj, M., Rayson, P., Young, S., and Walker, M. (2014). Detecting document structure in a very large corpus of UK financial reports. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1335–1338, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

El-Haj, M., Alves, P., Rayson, P., Walker, M., and Young, S. (2019). Retrieving, classifying and analysing narrative commentary in unstructured (glossy) annual reports published as pdf files. *Research Methods & Methodology in Accounting eJournal*.

Mahmoud El-Haj, et al., editors. (2022). *Proceedings of the 4th Financial Narrative Processing Workshop*, Marseille, France, 24 June. The 13th Language Resources and Evaluation Conference, LREC 2022.

Giguet, E. and Lucas, N. (2022). Greyc@fintoc-2022: Handling document layout and structure in native pdf bundle of documents. In *Proceedings of Language Resources and Evaluation (LREC'22)*, Marseille, France, June. European Language Resources Association (ELRA).

Kozlov, I., Belyaeva, O., Bogatenkova, A., and Perminov, A. (2021). Ispras@ fintoc-2021 shared task: Two-stage toc generation model. In *Proceedings of the 3rd Financial Narrative Processing Workshop*, pages 81–85.

Liu, C., Chen, J., Zhang, X., Liu, J., and Huang, Y. (2011). Toc structure extraction from ocr-ed books. In *INEX*.

Maarouf, I. E., Kang, J., Azzi, A. A., Bellato, S., Gan, M., and El-Haj, M. (2021). The financial document structure extraction shared task (FinTOC2021). In *Proceedings of the 3rd Financial Narrative Processing Workshop*, pages 111–119, Lancaster, United Kingdom, 15-16 September. Association for Computational Linguistics.

Moreno-Sandoval, A., Gisbert, A., and Montoro, H. (2020). Fint-esp: a corpus of financial reports in spanish. In Fuster, et al., editors, *Multiperspectives in analysis and corpus design*, pages 89–102, Granada. Comares.

Mysore Gopinath, A. A., Wilson, S., and Sadeh, N. (2018). Supervised and unsupervised methods for robust separation of section titles and prose text in web documents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 850–855, Brussels, Belgium, October-November. Association for Computational Linguistics.

Nguyen, T.-T.-H., Doucet, A., and Coustaty, M. (2017). Enhancing table of contents extraction by system aggregation. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 01, pages 242–247.

Zhong, X., Tang, J., and Yepes, A. J. (2019). Publaynet: largest dataset ever for document layout analysis. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1015–1022. IEEE, Sep.