The Financial Document Structure Extraction Shared Task (FinTOC 2021)

Ismail El Maarouf
Fortia Financial Solutions
Paris, France
ismail.elmaarouf@fortia.fr

Juyeon Kang
Fortia Financial Solutions
Paris, France
juyeon.kang@fortia.fr

Abderrahim Ait Azzi
Fortia Financial Solutions
Paris, France
abderrahim.aitazzi@fortia.fr

Sandra Bellato
Fortia Financial Solutions
Paris, France
sandra.bellato@fortia.fr

Mei Gan
Fortia Financial Solutions
Paris, France
mei.gan@fortia.fr

Mahmoud El-Haj
Lancaster University
Lancaster, UK
m.el-haj@lancaster.ac.uk

Abstract

This paper presents the FinTOC-2021 Shared Task on structure extraction from financial documents, its participants results and their findings. This shared task was organized as part of The 2nd Joint Workshop on Financial Narrative Processing (FNP 2021), held at the University of Lancaster. 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 third edition of this shared task, two subtasks were presented to the participants: one with English documents and the other one with French documents but with a different and revised dataset compared to FinTOC'2 edition.

1 Introduction

The use of PDF electronic documents is recurrent in the financial domain. They are used to share and broadcast information concerning investment strategies, policy and regulation. Even with a great layout, long documents can be hard to navigate, hence, the presence of a table-of-contents (TOC) can provide a valuable assistance for potential investors or regulators by increasing readability and facilitating navigation.

In this shared task, we focus on extracting the TOC of financial prospectuses. In these official documents, investment funds accurately depict their characteristics and investment modalities. Depending on their country of origin, they might be edited with or without a TOC, and they might follow a template as well. But even though their format is regulated, the choice of the text format, the layout, the graphics and tabular presentation of the data is in the hand of the editor. Thus, the TOC is of fundamental importance to tackle sophisticated NLP tasks such as information extraction or question answering on long documents.

In this paper, we report the results and findings of the FinTOC-2020 shared task.[1] The Shared Task was organized as part of The 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation (FNP-FNS 2020), to be held at The 28th International Conference on Computational Linguistics (COLING'2020).

A total of 5 teams submitted runs and contributed 5 system description papers. All system description papers are included in the FNP-FNS 2020 workshop proceedings and cited in this report.

2 Previous Work on TOC extraction

There are mainly two concepts in the literature to approach TOC extraction. The first one 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

[1] http://wp.lancs.ac.uk/cfie/fintoc2020/

INEX (Dresevic et al., 2009) and ICDAR competitions (Doucet et al., 2013; Beckers et al., 2010; Nguyen et al., 2018) 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, 2019; 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; 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).

3 Task Description

As part of the FNP-FNS Workshop, we present a shared task on Financial Document Structure Extraction.

Participants to this shared task were given two sets of financial prospectuses 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 two sets were specific to two 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.

For both sets, we observe that:

- some documents contain specific titles that do not appear in any other document

- the same title in two different documents can have a different position in the hierarchy

- two titles that follow each other can have the same layout but a different position in the TOC

- the font size of a higher-level title can be smaller than the font size of a lower-level one

- and a title can have the exact same layout as its associated paragraph.

For each subtask, all participating teams were provided with a training dataset which included the original PDFs alongside their corresponding JSON file representing the TOC of the document. This JSON represented the TOC by giving the titles, their pages, their depths and their IDs, as shown in Fig. 2. A private test set was used to evaluate the TOCs generated by the participants systems. As stated in Section 2, most of the previous research on TOC generation has focused on short papers such as research publications (*Arxiv* database), or weakly graphical material such as digitalized books. However, the task of extracting the TOC of commercial documents with a complex layout structure in the domain of finance is not much explored in the literature.

4 Shared Task Data

In this section, we discuss the corpus of documents used for the TOC extraction subtasks.

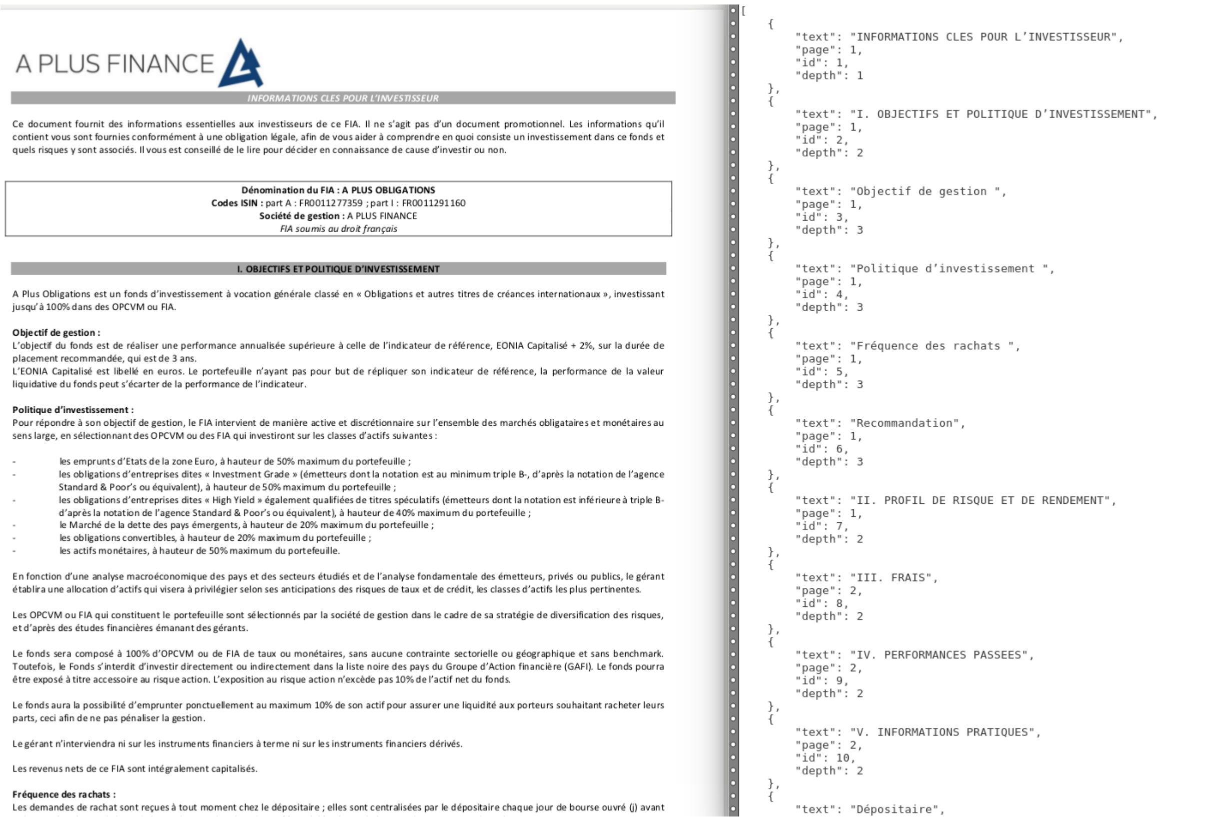
4.1 Corpus annotation

Investment documents can be accessed online in PDF format, and are also made available from asset managers. We compiled a list of 81 French documents, and 82 English documents from Luxembourg, to create the datasets of each subtask. We chose documents with a wide variety of layouts and styles.

We provided annotators with the original PDFs and a software that was developed internally to manually annotate the TOC of any PDF document. Once the annotator finishes their annotation task, the software produces a file containing the

Prospectus
DINVEST CONVICTIONS
AOÛT 2016
UBP
Groupe Union Bancaire Privée
OCPVM NON-CONFORME AUX NORMES EUROPEENNES
PROSPECTUS
COUPOLE IM EXCELLENCE MULTIGESTION
I CARACTERISTIQUES GENERALES
Amundi
ASSET MANAGEMENT
Informations clés pour l'investisseur
LCL TRESO 3 MOIS
Classe E - Code ISIN : (C) FR0007485743
Objectifs et politique d'investissement
Profil de risque et de rendement

Figure 1: Random pages from the shared task datasets. We observe a strong variability of complex layouts.



A PLUS FINANCE

INFORMATIONS CLES POUR L'INVESTISSEUR

Ce document fournit des informations essentielles aux investisseurs de ce FIA. Il ne s'agit pas d'un document promotionnel. Les informations qu'il contient vous sont fournies conformément à une obligation légale, afin de vous aider à comprendre en quoi consiste un investissement dans ce fonds et quels risques y sont associés. Il vous est conseillé de le lire pour décider en connaissance de cause d'investir ou non.

Dénomination du FIA : A PLUS OBLIGATIONS
Codes ISIN : part A : FR0011277359 ; part I : FR0011291160
Société de gestion : A PLUS FINANCE
FIA soumis au droit français

I. OBJECTIFS ET POLITIQUE D'INVESTISSEMENT

A Plus Obligations est un fonds d'investissement à vocation générale classé en « Obligations et autres titres de créances internationaux », investissant jusqu'à 100% dans des OPCVM ou FIA.

Objectif de gestion :
L'objectif du fonds est de réaliser une performance annualisée supérieure à celle de l'indicateur de référence, EONIA Capitalisé + 2%, sur la durée de placement recommandée, qui est de 3 ans.
L'EONIA Capitalisé est libellé en euros. Le portefeuille n'ayant pas pour but de répliquer son indicateur de référence, la performance de la valeur liquidative du fonds peut s'écarter de la performance de l'indicateur.

Politique d'investissement :
Pour répondre à son objectif de gestion, le FIA intervient de manière active et discrétionnaire sur l'ensemble des marchés obligataires et monétaires au sens large, en sélectionnant des OPCVM ou des FIA qui investiront sur les classes d'actifs suivantes :

- les emprunts d'Etats de la zone Euro, à hauteur de 50% maximum du portefeuille ;
- les obligations d'entreprises dites « Investment Grade » (émetteurs dont la notation est au minimum triple B-, d'après la notation de l'agence Standard & Poor's ou équivalent), à hauteur de 50% maximum du portefeuille ;
- les obligations d'entreprises dites « High Yield » également qualifiées de titres spéculatifs (émetteurs dont la notation est inférieure à triple B- d'après la notation de l'agence Standard & Poor's ou équivalent), à hauteur de 40% maximum du portefeuille ;
- le Marché de la dette des pays émergents, à hauteur de 20% maximum du portefeuille ;
- les obligations convertibles, à hauteur de 20% maximum du portefeuille ;
- les actifs monétaires, à hauteur de 50% maximum du portefeuille.

En fonction d'une analyse macroéconomique des pays et des secteurs étudiés et de l'analyse fondamentale des émetteurs, privés ou publics, le gérant établira une allocation d'actifs qui visera à privilégier selon ses anticipations des risques de taux et de crédit, les classes d'actifs les plus pertinentes.

Les OPCVM ou FIA qui constituent le portefeuille sont sélectionnés par la société de gestion dans le cadre de sa stratégie de diversification des risques, et d'après des études financières émanant des gérants.

Le fonds sera composé à 100% d'OPCVM ou de FIA de taux ou monétaires, sans aucune contrainte sectorielle ou géographique et sans benchmark. Toutefois, le Fonds s'interdit d'investir directement ou indirectement dans la liste noire des pays du Groupe d'Action financière (GAFI). Le fonds pourra être exposé à titre accessoire au risque action. L'exposition au risque action n'excède pas 10% de l'actif net du fonds.

Le fonds aura la possibilité d'emprunter ponctuellement au maximum 10% de son actif pour assurer une liquidité aux porteurs souhaitant racheter leurs parts, ceci afin de ne pas pénaliser la gestion.

Le gérant n'interviendra ni sur les instruments financiers à terme ni sur les instruments financiers dérivés.

Les revenus nets de ce FIA sont intégralement capitalisés.

Fréquence des rachats :
Les demandes de rachat sont reçues à tout moment chez le dépositaire ; elles sont centralisées par le dépositaire chaque jour de bourse ouvré (j) avant

```
[
  {
      "text": "INFORMATIONS CLES POUR L'INVESTISSEUR",
      "page": 1,
      "id": 1,
      "depth": 1
  },
  {
      "text": "I. OBJECTIFS ET POLITIQUE D'INVESTISSEMENT",
      "page": 1,
      "id": 2,
      "depth": 2
  },
  {
      "text": "Objectif de gestion ",
      "page": 1,
      "id": 3,
      "depth": 3
  },
  {
      "text": "Politique d'investissement ",
      "page": 1,
      "id": 4,
      "depth": 3
  },
  {
      "text": "Fréquence des rachats ",
      "page": 1,
      "id": 5,
      "depth": 3
  },
  {
      "text": "Recommandation",
      "page": 1,
      "id": 6,
      "depth": 3
  },
  {
      "text": "II. PROFIL DE RISQUE ET DE RENDEMENT",
      "page": 1,
      "id": 7,
      "depth": 2
  },
  {
      "text": "III. FRAIS",
      "page": 2,
      "id": 8,
      "depth": 2
  },
  {
      "text": "IV. PERFORMANCES PASSEES",
      "page": 2,
      "id": 9,
      "depth": 2
  },
  {
      "text": "V. INFORMATIONS PRATIQUES",
      "page": 2,
      "id": 10,
      "depth": 2
  },
  {
      "text": "Dépositaire",
```

Figure 2: A French prospectus with its JSON annotation file.

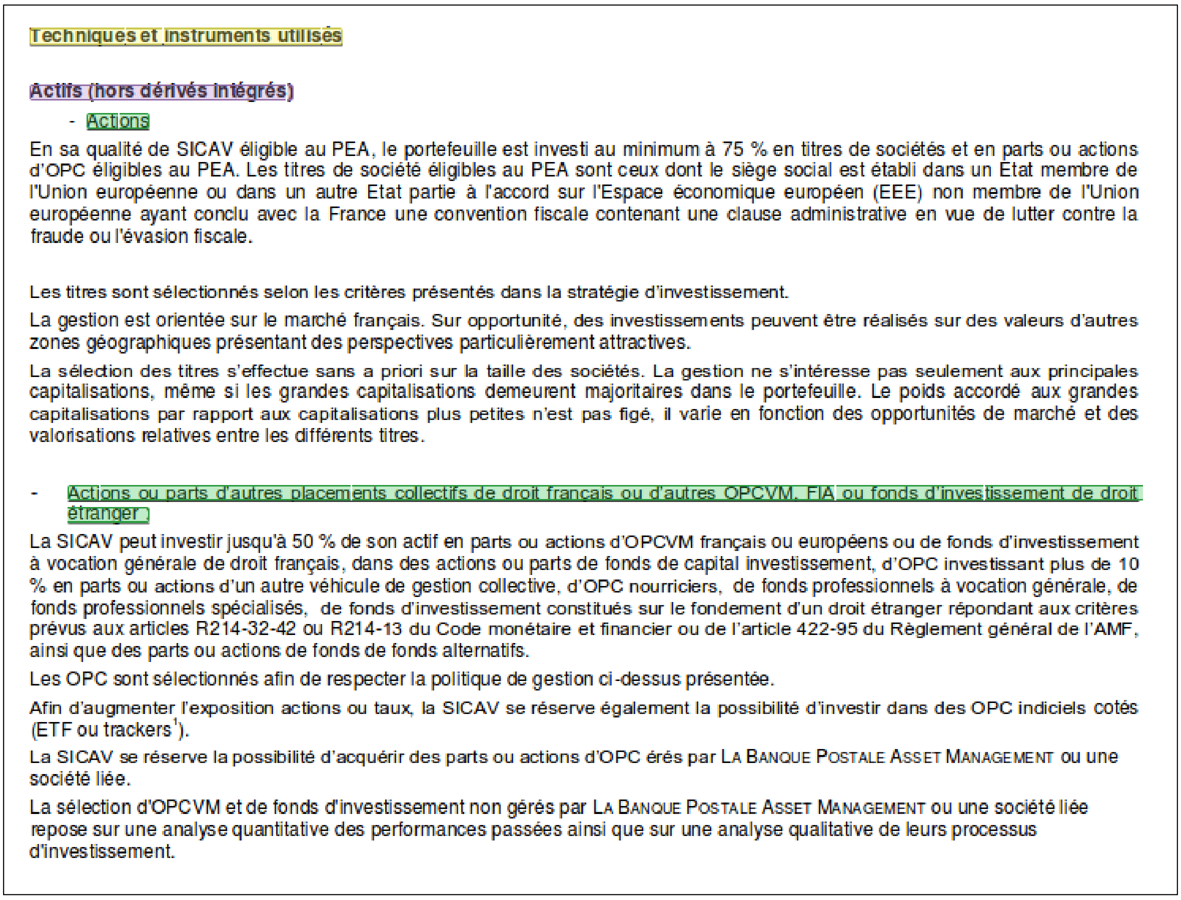
TOC-entries (title, page number, depth, and id) in a hierarchically structured format.

Each annotator was asked to:

1. Identify the title: Locate a title inside the PDF document.
2. Associate the entry level in the TOC: Every title is tagged with an integer representing the depth of the title in the TOC tree. The depth ranges from 1 to 10.
3. Tag the next title.

Each document was annotated independently by two people and a third person would review the annotations to resolve possible conflicts. For each dataset, the agreement scores between annotators are depicted in Table 1 and Table 2. We can observe high agreement scores, allowing us to be confident enough about the quality of our datasets.

Annotation Challenge: Title identification Investment prospectuses are commercial documents whose complex layout is optimized to highlight



Techniques et instruments utilisés

Actifs (hors dérivés intégrés)

- Actions

En sa qualité de SICAV éligible au PEA, le portefeuille est investi au minimum à 75 % en titres de sociétés et en parts ou actions d'OPC éligibles au PEA. Les titres de société éligibles au PEA sont ceux dont le siège social est établi dans un Etat membre de l'Union européenne ou dans un autre Etat partie à l'accord sur l'Espace économique européen (EEE) non membre de l'Union européenne ayant conclu avec la France une convention fiscale contenant une clause administrative en vue de lutter contre la fraude ou l'évasion fiscale.

Les titres sont sélectionnés selon les critères présentés dans la stratégie d'investissement.

La gestion est orientée sur le marché français. Sur opportunité, des investissements peuvent être réalisés sur des valeurs d'autres zones géographiques présentant des perspectives particulièrement attractives.

La sélection des titres s'effectue sans a priori sur la taille des sociétés. La gestion ne s'intéresse pas seulement aux principales capitalisations, même si les grandes capitalisations demeurent majoritaires dans le portefeuille. Le poids accordé aux grandes capitalisations par rapport aux capitalisations plus petites n'est pas figé, il varie en fonction des opportunités de marché et des valorisations relatives entre les différents titres.

- Actions ou parts d'autres placements collectifs de droit français ou d'autres OPCVM, FIA ou fonds d'investissement de droit étranger

La SICAV peut investir jusqu'à 50 % de son actif en parts ou actions d'OPCVM français ou européens ou de fonds d'investissement à vocation générale de droit français, dans des actions ou parts de fonds de capital investissement, d'OPC investissant plus de 10 % en parts ou actions d'un autre véhicule de gestion collective, d'OPC nourriciers, de fonds professionnels à vocation générale, de fonds professionnels spécialisés, de fonds d'investissement constitués sur le fondement d'un droit étranger répondant aux critères prévus aux articles R214-32-42 ou R214-13 du Code monétaire et financier ou de l'article 422-95 du Règlement général de l'AMF, ainsi que des parts ou actions de fonds de fonds alternatifs.

Les OPC sont sélectionnés afin de respecter la politique de gestion ci-dessus présentée.

Afin d'augmenter l'exposition actions ou taux, la SICAV se réserve également la possibilité d'investir dans des OPC indiciels cotés (ETF ou trackers[1]).

La SICAV se réserve la possibilité d'acquérir des parts ou actions d'OPC érés par LA BANQUE POSTALE ASSET MANAGEMENT ou une société liée.

La sélection d'OPCVM et de fonds d'investissement non gérés par LA BANQUE POSTALE ASSET MANAGEMENT ou une société liée repose sur une analyse quantitative des performances passées ainsi que sur une analyse qualitative de leurs processus d'investissement.

Figure 3: In this example, we can see that the titles tagged in green have the same style as the plain text of their paragraphs. Only the indentation is insightful to detect them.

	Xerox F1	Inex08 F1
tagger 1 & tagger 2	89.8%	77.0%
tagger 1 & reviewer	92.1%	82.8%
tagger 2 & reviewer	90.1%	79.6%

Table 1: Agreement scores between different annotators of the original French investment document training set (71 documents).

	Xerox F1	Inex08 F1
tagger 1 & tagger 2	87.7%	82.4%
tagger 1 & reviewer	95.6%	91.6%
tagger 2 & reviewer	91.8%	90.0 %

Table 2: Agreement scores between different annotators on a validation set of 62 documents from the original English investment document training set (69 documents).

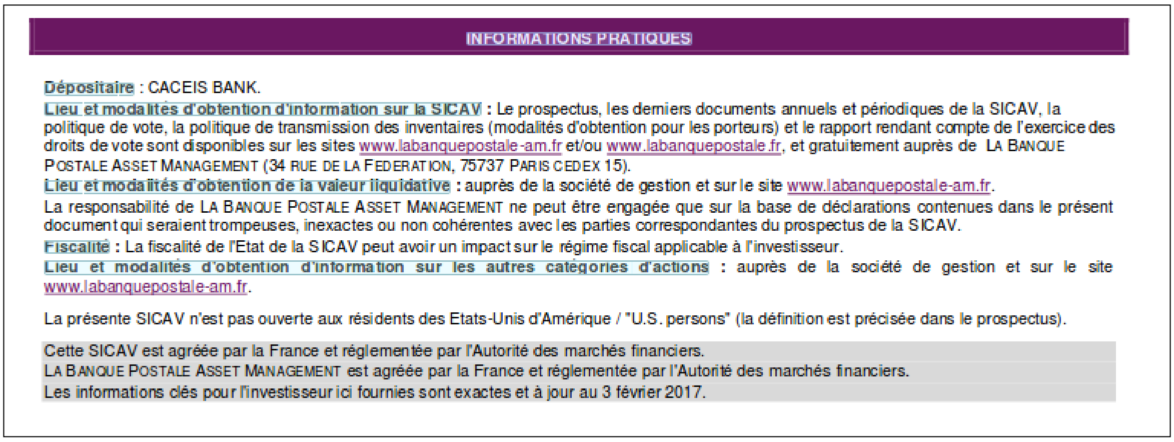
specific information such that a potential investor can identify it quickly. Hence, annotating a title and its level in the TOC hierarchy is a difficult task as one cannot rely on the visual appearance of the title to do so. Some examples can be observed in Fig. 3 and Fig. 4.

Annotation Challenge: Tagging PDF documents. The annotation of PDF documents is not an easy task since they are meant to be displayed. The tool we used for the annotations allows the annotators to directly tag on the PDF, however, the text selection relies on the HTML encoding of the PDF, where the text might slightly differ from what is actually displayed. For instance, it is possible that a piece of text is impossible to select if it is from an image. It is also possible that the tagged text has additional or missing characters.

4.2 Corpus Description

In the following, we provide an analysis of the data used for the shared task.

We simplified greatly the format of the annotation files compared to the first edition of the shared task (Juge et al., 2019). Instead of the XML format inherited from the Structure Extraction Competition (SEC) (Doucet et al., 2013) that implicitly



INFORMATIONS PRATIQUES

Dépositaire : CACEIS BANK.
Lieu et modalités d'obtention d'information sur la SICAV : Le prospectus, les demiers documents annuels et périodiques de la SICAV, la politique de vote, la politique de transmission des inventaires (modalités d'obtention pour les porteurs) et le rapport rendant compte de l'exercice des droits de vote sont disponibles sur les sites www.labanquepostale-am.fr et/ou www.labanquepostale.fr, et gratuitement auprès de LA BANQUE POSTALE ASSET MANAGEMENT (34 RUE DE LA FEDERATION, 75737 PARIS CEDEX 15).
Lieu et modalités d'obtention de la valeur liquidative : auprès de la société de gestion et sur le site www.labanquepostale-am.fr.
La responsabilité de LA BANQUE POSTALE ASSET MANAGEMENT ne peut être engagée que sur la base de déclarations contenues dans le présent document qui seraient trompeuses, inexactes ou non cohérentes avec les parties correspondantes du prospectus de la SICAV.
Fiscalité : La fiscalité de l'Etat de la SICAV peut avoir un impact sur le régime fiscal applicable à l'investisseur.
Lieu et modalités d'obtention d'information sur les autres catégories d'actions : auprès de la société de gestion et sur le site www.labanquepostale-am.fr.

La présente SICAV n'est pas ouverte aux résidents des Etats-Unis d'Amérique / "U.S. persons" (la définition est précisée dans le prospectus).

Cette SICAV est agréée par la France et réglementée par l'Autorité des marchés financiers.
LA BANQUE POSTALE ASSET MANAGEMENT est agréée par la France et réglementée par l'Autorité des marchés financiers.
Les informations clés pour l'investisseur ici fournies sont exactes et à jour au 3 février 2017.

Figure 4: In this second example, the identification of titles tagged in light blue is not evident because they might be followed by plain text in the same line.

encodes the title level, we used a simple JSON file containing a list of entries, where each entry has the following information: textual content, id, level, page number. An example of a JSON extract is provided in Fig. 2. In particular, the title level is explicitly stated. Statistics about levels on the French and English datasets are presented in Table 3.

In addition to the annotation files, the public dataset provided to the participants contained documents in PDF format. The private dataset on which participants were ranked contained documents in PDF format only.

5 Participants and Systems

A total of 30 teams registered in the shared task all from different institutions. Eventually, 6 teams participated and 5 teams submitted a paper with the description of their method, see Table 5 for more information about their affiliation. In Table 4, we show the details on the submissions per task. All the participants that submitted a standard run, sent a paper describing their approach as well.

Participating teams explored and implemented a wide variety of techniques and features. In this section, we give a brief description of each system. More details could be found in the description papers published in the proceedings of the FNP 2021 Workshop.

Christopher Bourez (Bourez, 2021): This team participated in both subtasks in both languages. They used ABBYY Finereader to extract text blocks and exclude tables from pages. The system leverages style properties such as font name, color, font type such as weight or italics, font size and computes a hash for the paragraph and the first line. Statistics are then computed on these style properties to feed a XGBoost classifier; these include font size ratio compared to common size of the document; indentation of the first line; local frequency of the style in a window of paragraphs; local and global frequency of each style feature; etc.

ISPRAS (Ilya et al., 2021): This team participated in both tasks in both languages. Their system uses PdfMiner to extract text, font and colors. They also attempt to extract the TOC page using regular expression and keyword matching techniques. Based on this preprocessing they build feature vectors including visual features (font, color, spacing), letter statistics, regex matches for line beginning and ending or content, presence in TOC, and the same features for a window of 3 lines around the candidate line to categorize. These features are fed to train an XGBoost classifier in various setups.

YSEOP (Gupta et al., 2021): This team participated in both tasks in both languages. Their system uses pdfminer to extract text lines and only extract features when they match the annotation or when the text line is a subset of the annotation. The features extracted include normalized coordinates, font statistics (%of bold and italic chars), page and line statistics, line beginning and line ending patterns and TFIDF of char ngrams. This was then used to identify bounding boxes in image-converted and fed into a Faster-RCNN classifier to fine tune the PubLayNet model. This was extract the IoU and probability of being a title of each text line which was in a 3rd step fed into a Gradient

	French dataset	**English dataset**
number of documents	71	72
average number of pages	28	91
level 1 (% of titles)	2%	5%
level 2 (% of titles)	11%	21%
level 3 (% of titles)	29%	30%
level 4 (% of titles)	24%	25%
level 5 (% of titles)	21%	11%
level 6 (% of titles)	13%	4%
level 7 (% of titles)	0%	2%
level 8 (% of titles)	0%	1%
level 9 (% of titles)	0%	1%
level 10 (% of titles)	0%	0%

Table 3: Statistics on the subtasks datasets.

	# teams	**# std runs**
French subtask	5	6
English subtask	6	8
papers	5	-

Table 4: Statistics on the participation on French and English subtasks

Boosting Classifier trained on the dataset. To obtain the TOC, the titles were ordered by size where the largest titles were attributed the highest level.

CILAB (Kim et al., 2021): This team participated in both subtasks in English. They used PDFminer to extract the text and its coordinates, as well as font style properties such as font size and font weight. They used a Random Forest model for title detection after experimenting with a total of 5 ML algorithms such as SVM. They also gathered additional data (400 prospectuses) which was pseudo-labeled and experimented on different splits of the data. For TOC extraction, a number of heuristics were designed based on a careful analysis of the data, using regular expressions font size and font style. The team observed that data augmentation was reflected by better performance.

Daniel (Giguet and Lejeune, 2021): This team participated in both tasks on both languages. They design a preprocessing pipeline which structures the document and extracts features from text (token, line and text block), vector shapes (rectangles and borders) and images (figures, but also small character shapes like arrows and checkboxes) from pdf2xml. The system performs generic Page Layout Analysis (PLA) of which title detection and TOC extraction is a step. It recognizes and labels content areas such as text, paragraphs, tables, figures, lists, headers and footers, and use a deterministic algorithm to detect the TOC page which is parsed and then linked to matching text lines and pages in the document which are then labelled as titles.

6 Results and Discussion

Evaluation Metric Since both subtasks tackle the same problem but on different corpora, we used the same evaluation metric as in FinTOC2020.

For the TOC generation part, we adapted the metrics proposed by the Structure Extraction Competition (SEC) held at ICDAR 2013 (Doucet et al., 2013): we adapted the script, replaced the customized Levenshtein distance specifically designed for SEC by a standard Levenshtein distance whose edit cost is 1 in all cases, and removed 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*. In the calculation of the *Inex F1 score*, correct entries in the predicted TOC are those which match the title of an entry in the groundtruth TOC *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:

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

Team	Affiliation	Tasks
Yseop Lab (Gupta et al., 2021)	Yseop, Paris, France	F and E
Daniel (Giguet and Lejeune, 2021)	Normandie Univ, UNICAEN, GREYC, Caen, France	F and E
ISPRAS (Ilya et al., 2021)	ISP RAS, Moscow, Russia	F and E
CILAB (Kim et al., 2021)	KIT, Gumi, Korea	E
Christopher Bourez (Bourez, 2021)	iValua, Paris, France	F and E

Table 5: List of the 5 teams that participated in Subtasks of the FinTOC2021 Shared Task. "F" refers to the French substask and "E" refers to the English subtask

We also provided scores for the title detection part separately: we used the F1 score, and considered as correct entries the predicted entries which match the titles of groudtruth entries according to the standard Levenshtein distance.

For both parts, the threshold on the Levenshtein *score* was set to 0.85[2]. Moreover, the Inex scores and title F1 score are calculated for each document and then averaged over the documents of the private set to produce two performance figures per team submission: one for TOC extraction, and another for title detection (TD).

Baseline For comparison purposes, we used the same baseline Title and TOC extractor used in FinTOC2020:

- extracting textual content from the PDF documents using `pdftohtml` utility from Poppler library[3]
- assigning groundtruth labels (title or non-title) to text segments by fuzzy string matching with the annotations
- vectorizing text segments into one-dimensional vectors of length 3 encoding the following features: is_bold, is_italic, is_all_capitalized
- training a SVM on the obtained dataset
- assigning to a predicted title the most frequent hierarchy level found in the training set

Table 6 (respectively Table 7) reports the results obtained by the participants and the baseline on TOC extraction from French documents (respectively English documents).

Discussion. Fior all tasks in all languages, we observe the same ranking: Christopher Bourez2, Christopher Bourez1 and ISP RAS; the scores between these 3 best systems are quite tight, particularly on Enligh TD (83, 82.2, 81.3 respectively) and the gap is much bigger on TOC tasks between the second and third (53.6, 52.5, 37.9 respectively). The other teams obtained much lower scores on TOC and online largely beat the baseline on English TD task, which probably means that French TD is an easier task than English TD. So the winning recipe seems to involve a lot of feature engineering, visual features, windowing techniques and Gradient Boosted trees (see (Bourez, 2021) and (Ilya et al., 2021) for more details).

Since TOC extraction task depends on TD task, participants have focused on improving their TD models, to the expense of TOC. We believe this partly explains why the scores are lower on TOC extraction compared to TD. Overall it seems that most systems have a much simpler approach to TOC compared to TD and they mostly fine-tuned their systems on TD.

7 Conclusions

In this paper we presented the setup and results for the Financial Document Structure Extraction task (FinToc) 2021, organized as part of The 2nd Joint Workshop on Financial Narrative Processing (FNP 2021). A total of 30 teams registered and 6 teams participated in the shared task with a wide variety of techniques. Five teams contributed with a paper describing their system.

This edition improved the datasets, composed of French investment documents, and annotated for the TOC extraction problem. A test set also supplements previously released datasets for both English and French (Bentabet et al., 2020) (Juge et al., 2019). TOC extraction on PDF documents is a realistic problem in everyday applications which explain the interest from and participation of both public universities and profit organizations.

[2] The script implementing these metrics can be found here: `https://drive.google.com/file/d/1HJDRRvzPiISvwUWv5aygn_kjxn1JqLG_/view?usp=sharing`

[3] see https://poppler.freedesktop.org/

Team	**TD**
Christopher Bourez2	**81.8**
Christopher Bourez1	81.7
ISP RAS	78.7
Yseop Lab	63.9
Daniel	60.6
Baseline	60.9

Team	**TOC**
Christopher Bourez1	**57.3**
Christopher Bourez2	57.3
ISP RAS	42.1
Baseline	36.5
Yseop Lab	22.4
Daniel	11.8

Table 6: Results obtained by the participants for the first FinTOC2021 subtask : TOC extraction from French documents. The title detection (TD) ranking is based on F1-score, while the Table-Of-Content (TOC) ranking is based on the harmonic mean between Inex F1 score and Inex level accuracy

Team	**TD**
Christopher Bourez2	**83**
Christopher Bourez1	82.2
ISP RAS	81.3
Yseop Lab	72.8
Cilab2	51.4
Daniel	46.5
Cilab1	45.6
Baseline	20.6

Team	**TOC**
Christopher Bourez2	**53.6**
Christopher Bourez1	52.5
ISP RAS	37.9
Cilab2	26.3
Cilab1	23.4
Yseop Lab	20.1
Daniel	NA
Baseline	13.2

Table 7: Results obtained by the participants for the second FinTOC2021 subtask : TOC extraction from English documents. The title detection (TD) ranking is based on F1-score, while the Table-Of-Content (TOC) ranking is based on the harmonic mean between Inex F1 score and Inex level accuracy

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, and Lidia Duarte, Bianca Chong, Marion Cargill, as well as Sandra Bellato and Mei Gan, authors of this paper.

References

Thomas Beckers, Patrice Bellot, Gianluca Demartini, Ludovic Denoyer, Christopher M. De Vries, Antoine Doucet, Khairun Nisa Fachry, Norbert Fuhr, Patrick Gallinari, Shlomo Geva, Wei-Che Huang, Tereza Iofciu, Jaap Kamps, Gabriella Kazai, Marijn Koolen, Sangeetha Kutty, Monica Landoni, Miro Lehtonen, Véronique Moriceau, Richi Nayak, Ragnar Nordlie, Nils Pharo, Eric Sanjuan, Ralf Schenkel, Xavier Tannier, Martin Theobald, James A. Thom, Andrew Trotman, and Arjen P. De Vries. 2010. Report on INEX 2009. *Sigir Forum*, 44(1):38–57.

Najah-Imane Bentabet, Rémi Juge, Ismail El Maarouf, Virginie Mouilleron, Dialekti Valsamou-Stanislawski, and Mahmoud El-Haj. 2020. The financial document structure extraction shared task (FinToc 2020). In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*, pages 13–22, Barcelona, Spain (Online). COLING.

Najah-Imane Bentabet, Rémi Juge, and Sira Ferradans. 2019. Table-of-contents generation on contemporary documents. In *Proceedings of ICDAR 2019*.

Christopher Bourez. 2021. FinTOC2021 - Document Structure Understanding. In *The Third Financial Narrative Processing Workshop (FNP 2021)*, Lancaster, UK.

Antoine Doucet, Gabriella Kazai, Sebastian Colutto, and Günter Mühlberger. 2013. Icdar 2013 competition on book structure extraction. In *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*, pages 1438–1443. IEEE.

Bodin Dresevic, Aleksandar Uzelac, Bogdan Radakovic, and Nikola Todic. 2009. Book layout analysis: Toc structure extraction engine. In *Advances in Focused Retrieval*, pages 164–171, Berlin, Heidelberg. Springer Berlin Heidelberg.

Mahmoud El-Haj, Paulo Alves, Paul Rayson, Martin Walker, and Steven Young. 2019. Retrieving, classifying and analysing narrative commentary in unstructured (glossy) annual reports published as pdf files. *Accounting and Business Research*, pages 1–29.

Mahmoud El Haj, Paul Rayson, Steven Young, and Martin Walker. 2014. *Detecting document struc-*

ture in a very large corpus of UK financial reports. LREC'14 Ninth International Conference on Language Resources and Evaluation. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014) . European Language Resources Association (ELRA), Reykjavik, Iceland, pp. 1335-1338.

Mahmoud El Haj, Paul Edward Rayson, Steven Eric Young, Paulo Alves, and Carlos Herrero Zorita. 2019. *Multilingual Financial Narrative Processing: Analysing Annual Reports in English, Spanish and Portuguese*. World Scientific Publishing.

Emmanuel Giguet and Gaël Lejeune. 2021. Daniel@FinTOC-2021: Taking Advantage of Images and Vectorial Shapes in Native PDF Document Analysis. In *The Third Financial Narrative Processing Workshop (FNP 2021)*, Lancaster, UK.

Abhijith Athreya Mysore Gopinath, Shomir Wilson, and Norman Sadeh. 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.

Anubhav Gupta, Hanna Abi Akl, and Hugues de Mazancourt. 2021. Not All Titles are Created Equal: Financial Document Structure Extraction Shared Task. In *The Third Financial Narrative Processing Workshop (FNP 2021)*, Lancaster, UK.

Kozlov Ilya, Oksana Belyaeva, Anastasiya Bogatenkova, and Andrew Perminov. 2021. ISPRAS@FinTOC-2021 Shared Task: Two-stage TOC generation model. In *The Third Financial Narrative Processing Workshop (FNP 2021)*, Lancaster, UK.

Remi Juge, Imane Bentabet, and Sira Ferradans. 2019. The FinTOC-2019 shared task: Financial document structure extraction. In *Proceedings of the Second Financial Narrative Processing Workshop (FNP 2019)*, pages 51–57, Turku, Finland. Linköping University Electronic Press.

Hyuntae Kim, Soyoung Park, Seongeun Yang, and Yuchul Jung. 2021. CILAB@FinTOC-2021 Shared Task: Title Detection and Table of Content Extraction for Financial Document. In *The Third Financial Narrative Processing Workshop (FNP 2021)*, Lancaster, UK.

Caihua Liu, Jiajun Chen, Xiaofeng Zhang, Jie Liu, and Yalou Huang. 2011. Toc structure extraction from ocr-ed books. In *International Workshop of the Initiative for the Evaluation of XML Retrieval*, pages 98–108. Springer.

Thi Tuyet Hai Nguyen, Antoine Doucet, and Mickael Coustaty. 2018. Enhancing table of contents extraction by system aggregation. In *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*.