MARCH 2025

AI AND SUPERVISION: INSIGHTS LEARNED FROM AMF EXPERIMENTS ON THE POSSIBILITY OF AUTOMATED PROCESSING OF REGULATORY REPORTING

Application to two reports defined by the SFDR and Taxonomy regulations

AUTORITÉ
DES MARCHÉS FINANCIERS
AMF

OVERVIEW

For several years now, the AMF has been faced with a **constant expansion of its field of supervision**, with the adoption of new legislative and regulatory measures. Managing this growth in a tight budgetary context is a major challenge, prompting the AMF to constantly strive for greater efficiency and to **focus on the development of automatic data processing** as part of its Data strategy[1]. The use of its ICY[2] platform has enabled the AMF to continue its work in artificial intelligence (AI), particularly in natural language processing (NLP) and image processing, to automate manual tasks in the analysis of two new reports linked to sustainable finance, thereby **increasing the supervisory capabilities** of its teams.

Work on various publications imposed by regulations, as illustrated in this note by the Taxonomy[3] reports published in 2022 and the SFDR[4] appendices in 2023, has highlighted how the **level of standardization of a reporting** and the **technical choices of publication** can affect a **machine's ability to extract data automatically**[5].

These insights are particularly valuable in a context where the European legislator is increasingly incorporating provisions aimed at facilitating access to information and its automated processing, especially as the entry into application of the ESAP[6] regulation approaches. This regulation stipulates that published documents (including Taxonomy reports and SFDR annexes) must, at a minimum, be data-extractable and, for the most part, machine-readable. However, regulatory requirements remain, to date, insufficiently prescriptive in their definitions: only images would not be considered data-extractable, and certain documents currently deemed machine-readable[7] in the regulatory sense are not well-suited for automated processing.

In theory, the use of the XHTML format, combined with compliance with strict rules such as W3C standards, should ensure optimal machine processing of documents, thereby facilitating precise and reliable data extraction. However, the experiments conducted by the AMF did not validate this hypothesis—not because it was refuted, but because the documents analyzed did not conform to these standards. The work focused on portions of XHTML documents whose drafting is not subject to precise regulatory requirements, as well as on PDF documents, which are generally considered more complex to process but are not necessarily so, particularly when they are standardized.
Based on the difficulties encountered, these tests highlighted the obstacles to effective automated processing and underscored the need to **further harmonize formats and standards to improve both human and machine accessibility, as well as the usability of documents for the benefit of the public.**

[1] The AMF's Data strategy is based on three major challenges: making AMF data a shared asset, automatically processing data, and developing tools that extract value from data. https://www.amf-france.org/en/news-publications/news-releases/amf-continues-its-data-strategy-release-short-selling-data-public

[2] https://www.amf-france.org/fr/actualites-publications/actualites/icy-la-nouvelle-plateforme-de-surveillance-de-lamf-est-operationnelle

[3] Regulation (UE) 2020/852 : link

[4] Regulation (UE) 2019/2088 : link

[5] When a document is read by a human, attention is focused on its visual presentation, logical structure, and the meaning of its content. In contrast, a machine reads a document in a completely different way. Rather than "understanding" the content, it analyzes raw data through the underlying technical encoding. The encoding of a document involves transforming information, such as text, symbols, or images, into a format that computers can understand and display. Whether in a PDF or XHTML format, appropriate encoding ensures that the document is not only readable for the user but also exploitable by the machine for data extraction. The encoding process is generally transparent to humans and is managed by software applications that create a document and save it in a specific format.

[6] Regulation (EU) 2023/2859 (link), also known as ESAP for "European Single Access Point," aims to centralize and provide access to all regulated documents within the European financial market.

[7] See subsection « 3.1 Regulatory concepts machine readable / data extractable » for a clarification of these terms.

The AMF's AI initiatives have also highlighted a **significant lag in AI research advancements in the financial sector for the French language**, partly due to the technical costs associated with accessing a centralized and queryable database of documents in French. A glossary is available on page 10 to clarify or recall the abbreviations used in this document.

INTRODUCTION

Last years, the AMF continued its work in Artificial Intelligence (AI) to develop, among other things, two analytical support tools for the new sustainability-related reporting requirements that came into effect during those years.

The proposed solutions aim to save time for supervisory teams by automating the extraction of relevant information, a manual task that is often lengthy and tedious. To achieve this, the implemented tools provide two main functionalities:

- The centralization of relevant data extracted from numerous documents, and
- A visual interface enabling quick and easy consultation of this data.

The performance of the AI systems developed within these solutions (hereinafter referred to as the "model" or "machine") is measured by their ability to extract all relevant data from the processed documents with precision and reliability. Their effectiveness is inherently linked to the format and quality of the documents used.

The lessons learned from this work highlight the correlation between the performance of these AI systems (and therefore their ability to reduce low-value-added tasks) and the machine-readability of documents. While these findings are illustrated here through sustainability reporting, they are also applicable to many other types of reporting.

1. OVERVIEW OF THE 2022 IA PROJECT ON TAXONOMY REPORTING FOR NON-FINANCIAL ISSUERS

As a reminder, the European taxonomy constitutes a shared classification system within the European Union. Its objective is to identify economic activities considered sustainable, particularly from an environmental perspective. The taxonomy also establishes specific reporting obligations for issuers listed on financial markets, whether they are financial or non-financial entities. In 2022 (for the 2021 financial year), these entities were required to disclose indicators measuring the extent of their activities, investments, or operational expenditures eligible under the taxonomy[8].

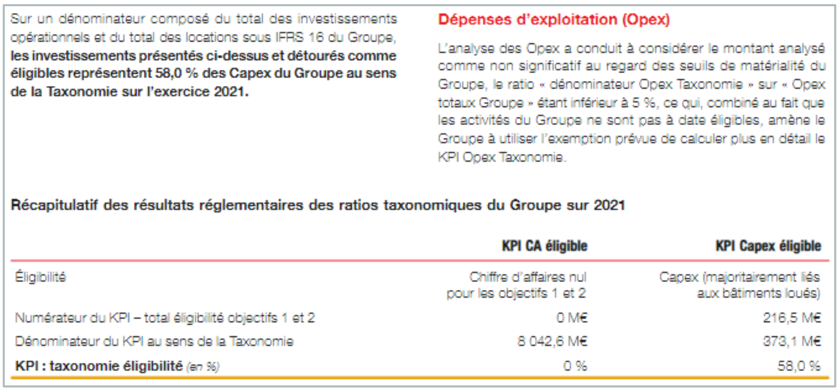
Based on a sample of 96 reports published in 2022 by non-financial issuers, the AMF conducted a project that same year aimed at automatically building a consolidated and reliable database of key performance indicators (KPIs) related to capital expenditures (CAPEX), operational expenditures (OPEX), and revenue (CA) eligible under the taxonomy[9].

Figure 1 provides an example of how these KPIs are presented in an annual report.

[8] Since 2023, the obligations have been extended to include alignments, i.e. compliance with minimum sustainability criteria.

[9] As the indicators in the taxonomy reports of financial issuers differ from those of non-financial issuers, the experiment was restricted to non-financial companies eligible for the taxonomy.

Figure 1 : example of a KPI presentation in table and text format



Sur un dénominateur composé du total des investissements opérationnels et du total des locations sous IFRS 16 du Groupe, **les investissements présentés ci-dessus et détourés comme éligibles représentent 58,0 % des Capex du Groupe au sens de la Taxonomie sur l'exercice 2021.**

Dépenses d'exploitation (Opex)

L'analyse des Opex a conduit à considérer le montant analysé comme non significatif au regard des seuils de matérialité du Groupe, le ratio « dénominateur Opex Taxonomie » sur « Opex totaux Groupe » étant inférieur à 5 %, ce qui, combiné au fait que les activités du Groupe ne sont pas à date éligibles, amène le Groupe à utiliser l'exemption prévue de calculer plus en détail le KPI Opex Taxonomie.

Récapitulatif des résultats réglementaires des ratios taxonomiques du Groupe sur 2021

	KPI CA éligible	**KPI Capex éligible**
Éligibilité	Chiffre d'affaires nul pour les objectifs 1 et 2	Capex (majoritairement liés aux bâtiments loués)
Numérateur du KPI – total éligibilité objectifs 1 et 2	0 M€	216,5 M€
Dénominateur du KPI au sens de la Taxonomie	8 042,6 M€	373,1 M€
KPI : taxonomie éligibilité *(en %)*	0 %	58,0 %

Given that the 2022 reports contain information to be extracted from both text and tables, the AMF adopted various AI techniques, combining natural language processing (NLP) and image processing[10]. These approaches were also supplemented by a set of rules, particularly to manage the consolidation of identical information extracted from both text and tables.

Furthermore, the model was trained to search for:

- The values of key performance indicators (KPIs), whether expressed as percentages (e.g., "10%"), in numerical format (e.g., "320 million"), or as quantifiers (e.g., "totality"); and
- The qualitative information associated with these values (e.g., "non-significant" or "non-material").

In the example provided in Figure 1 above, the machine extracts the following information:

Table 2 : data automatically extracted by the machine from the example in Figure 1

ICP	**Value**	**Quatlitative information if any**
CAPEX	58%*	
OPEX	No value**	Non-significant
CA	0%	

*This data appears twice: in the text and in the table. The system of rules that has been put in place makes it possible to manage this type of case.

** There is no figure related to the OPEX, which is qualified as non-significant by the sender. In this case, the machine should not return anything.

Across all the documents processed as part of this work, the results obtained are relatively satisfactory (see the focus on performance evaluation below). Furthermore, the visual interface[11] developed within this project also allows for the verification of the results produced by the model and, if necessary, manual correction, with direct access to the specific sections of the processed reports[12].

The investment required to further improve performance levels was deemed too significant to continue the project toward potential deployment. It is therefore crucial to examine the limiting factors identified during this work and to draw lessons from them. To this end, Section 3.2, titled "SYNTHESIS OF INSIGHTS

[10] Image processing techniques were used to extract information from the tables. For more details on the work carried out in AI on the Taxonomy report, please refer to Appendix 4.

[11] See Appendix 1 for a screenshot of the developed interface.

[12] A data correction by a user generates an automatic update of the database.

LEARNED FROM THE AMF'S WORK", provides an analysis of the relationship between the observed performance and the quality of the processed documents, highlighting the challenges related to machine-readability.

Focus on performance evaluation of the AI solution for Taxonomy

Two approaches can be used to assess the performance of the solution:

- **Evaluation based on the proportion of issuers**: This first approach measures the machine's ability to correctly extract the required information for each issuer. In other words, it calculates the number of reports where the machine successfully extracted all necessary information with precision. Among the sample of 96 reports analyzed, this approach indicates a success rate of 49%, meaning that the machine perfectly processed nearly half of the documents. Additionally, a partial success rate of 26% was observed, corresponding to reports where only some of the key performance indicators (KPIs) were correctly extracted. However, in 25% of the documents, the prototype failed to correctly extract at least one required piece of information.

- **Evaluation based on the number of KPIs extracted**[13]: The second approach assesses the machine's precision in terms of the number of KPIs for which the correct value was identified. With three KPIs per report, the dataset comprises a total of 288 indicators to be extracted. In this context, the model demonstrates an average accuracy of 70%, meaning that the machine correctly extracts a KPI in 7 out of 10 cases. However, it fails to find the value in 19% of cases and returns an incorrect value in 11% of cases.

2. SUMMARY OF THE IA PROJECT TO BE CARRIED OUT IN 2023 ON THE SFDR APPENDICES OF THE FUNDS

Between 2021 and 2023, the Sustainable Finance Disclosure Regulation (SFDR) came into effect. This regulation requires financial actors marketing or advising on financial products within the European Union to provide more transparent disclosures on the extent to which these financial products incorporate environmental or social characteristics. The regulation introduces a classification system with two levels of sustainability commitment[14].

Based on 6,300 fund prospectuses submitted to the AMF in early 2023 (for publication or modification), AMF services conducted a project that same year aimed at automating:

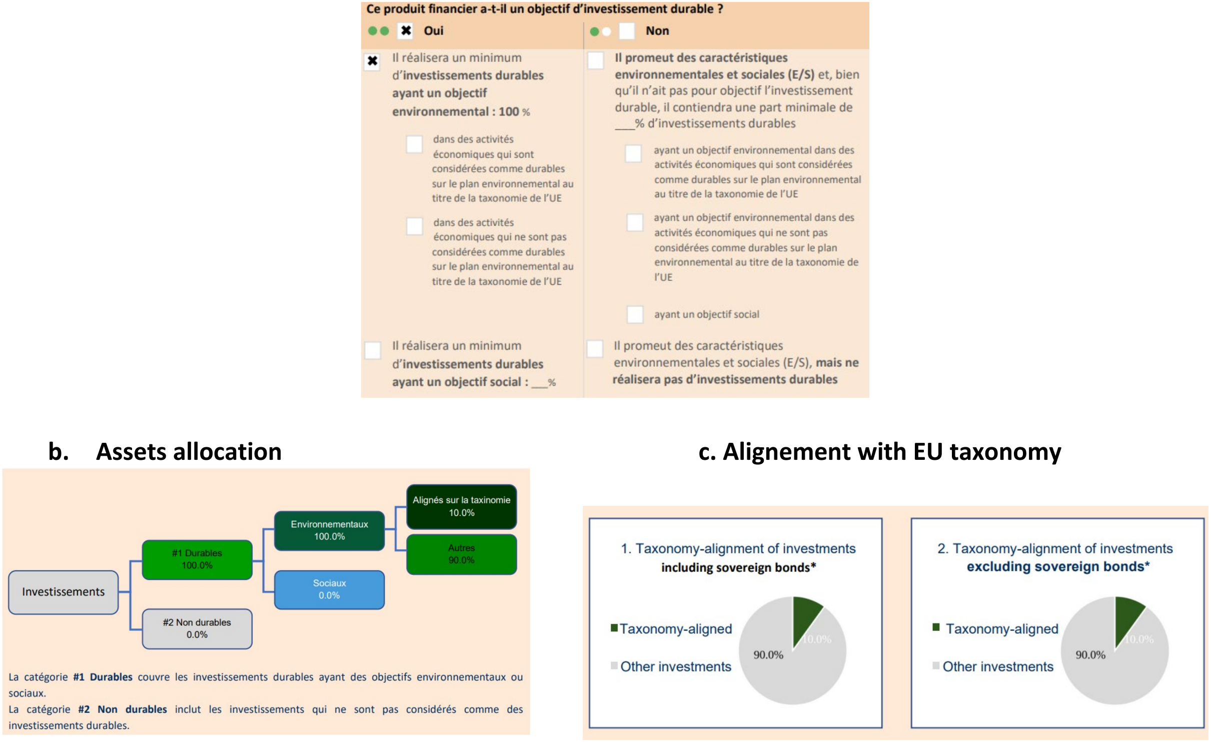
- the construction of a consolidated and reliable database of specific data extracted from the SFDR annexes, including: the fund's classification (Article 8 or Article 9 under SFDR[15]), information on sustainable investment objectives, the planned asset allocation, and the extent to which sustainable investments are aligned with the EU Taxonomy ;
- 11 basic compliance tests[16].

[13] To this extent, the correct extraction of qualitative information associated with a KPI is not taken into account.
[14] https://www.amf-france.org/en/news-publications/amfs-eu-positions/proposal-minimum-environmental-standards-financial-products-belonging-art9-and-8-categories-sfdr
[15] Funds that do not have an SFDR appendix are by default classified as article 6 within the meaning of the SFDR.
[16] These will not be detailed in this note, but by way of example, the solution identifies if any information is missing from the reporting.

The SFDR annexes take the form of a relatively standardized questionnaire-based form[17], where the list of questions and answers varies according to the fund category. Figure 2 below illustrates an example of the main sections of these annexes that contain the data AMF services aim to automate for extraction.

Figure 2 : illustration of the main parts of the SFDR appendix covered by the project

a. Information on sustainable investment objective



Ce produit financier a-t-il un objectif d'investissement durable ?

●● ☒ **Oui**

☒ Il réalisera un minimum d'**investissements durables ayant un objectif environnemental : 100** %

□ dans des activités économiques qui sont considérées comme durables sur le plan environnemental au titre de la taxonomie de l'UE

□ dans des activités économiques qui ne sont pas considérées comme durables sur le plan environnemental au titre de la taxonomie de l'UE

□ Il réalisera un minimum d'**investissements durables ayant un objectif social :** ___%

●○ □ **Non**

□ **Il promeut des caractéristiques environnementales et sociales (E/S)** et, bien qu'il n'ait pas pour objectif l'investissement durable, il contiendra une part minimale de ___% d'investissements durables

□ ayant un objectif environnemental dans des activités économiques qui sont considérées comme durables sur le plan environnemental au titre de la taxonomie de l'UE

□ ayant un objectif environnemental dans des activités économiques qui ne sont pas considérées comme durables sur le plan environnemental au titre de la taxonomie de l'UE

□ ayant un objectif social

□ Il promeut des caractéristiques environnementales et sociales (E/S), **mais ne réalisera pas d'investissements durables**

b. Assets allocation

La catégorie **#1 Durables** couvre les investissements durables ayant des objectifs environnementaux ou sociaux.
La catégorie **#2 Non durables** inclut les investissements qui ne sont pas considérés comme des investissements durables.

c. Alignement with EU taxonomy

1. Taxonomy-alignment of investments
including sovereign bonds*
Taxonomy-aligned
Other investments
90.0%
10.0%
2. Taxonomy-alignment of investments
excluding sovereign bonds*
Taxonomy-aligned
Other investments
90.0%
10.0%

The information to be extracted is published in various formats, including checkboxes, text, and graphics, which may sometimes be embedded as images within the documents. As with the previous project on Taxonomy reports, the AMF also employed various AI techniques, such as NLP and image processing[18].

In the example provided in Figure 2.a above, which pertains to sustainable investment objectives, the machine extracts the following information:

[17] https://www.esma.europa.eu/document/sfdr-templates
[18] For more details on the work carried out in IA on SFDR appendices, please refer to Appendix 4.

Table 2: data automatically extracted by the machine on the example in Figure 2.a Information on sustainable investment objectives

Informations on objectives		**Extracted data**
Does the financial product have a sustainable investment objective ?		Yes
% sustainable investments with an environmental objective	Total	100%
	Qualified as sustainable by the EU taxonomy	No information*
	Not qualified as sustainable by the UE taxonomy	No information*
% sustainable investments with a social objective		0%

NB : *no box is ticked, in which case the machine should return nothing

In the remainder of the example provided in Figure 2 above, the machine extracts all numerical data related to asset allocation and potential alignments with the EU Taxonomy[19].

For the documents processed in this project, the results obtained were generally very satisfactory, although a specific type of information yielded lower performance results (see the focus on performance evaluation below). Moreover, the relevant supervisory teams confirmed their interest in using the tool as it stands. As a result, it was deployed in its current state to allow for further in-depth testing of its practical application[20].

The insights gained regarding the quality of the results observed based on the format of the information extracted from the SFDR annexes have been incorporated into the analysis presented in Section 3.2, titled "SYNTHESIS OF INSIGHTS LEARNED FROM THE AMF'S WORK."

Focus on assessing the performance of the AI solution for SFDR

The performance of the solution developed according to the information to be extracted:
- Classify a financial product according to its sustainability objectives (Article 6, 8 or 9 within the meaning of the SFDR regulation): 95%.
- Automatically extract answers to a number of questions about the product, in particular its :
    - sustainable investment objectives: 81
    - Asset allocation: 80%.
    - alignment with the European green taxonomy: 30

3. OVERALL CONCLUSION ON THE FORMATS IMPOSED BY THE REGULATIONS IN THE VARIOUS EU FINANCE TEXTS

3.1. REGULATORY CONCEPTS MACHINE READABLE/DATA EXTRACTABLE

[19] Given the number of cases to be detailed, the solution's interface provides a number of tabs and tables to display the results, which are not shown here for ease of reading.

[20] The recipe is still in progress at the time of publication. See APPENDIX 2: INTERFACE OF THE DEPLOYED TOOL FOR EXPLORING SFDR APPENDICES

To facilitate access to data and improve information transparency, the issue of machine-readability of documents has taken a central role in European regulations in recent years. The Open Data Directive[21] was the first to provide a definition in 2020: a **machine-readable** format is "*a file format structured in such a way that software applications can easily identify, recognize, and extract specific data, including each statement of fact and its internal structure.*" In practice, this includes formats such as XML, XBRL, CSV, or JSON, which share the ability to represent information in an organized and hierarchical manner (for example, through the use of tags for XML and XBRL). This structure not only enables the automation of data processing but also allows for the reliable exchange of information between different systems. In 2023, Level 1 of the ESAP regulation (centralized electronic system for regulated documents) introduced the term **data extractable**[22] for the first time, defining it as "*a format allowing data extraction […] by a machine and not just human readability.*" Unlike the term machine-readable, this second definition is much more permissive and currently accepts the PDF format[23], which can be very difficult to process unless the document content is standardized.

ESAP represents a major step forward in regulatory efforts to promote machine-readability, not only because it mandates that the data covered by the 37 legislative acts[24] within its scope be published in a machine-readable or data-extractable format[25], but also because it is tasked with providing an indicative list of these formats and their characteristics (in Level 2 texts[26]). Subsequently, Level 3 texts will allow omnibus regulations and directives to specify, on a case-by-case basis, machine-readability requirements and the accepted formats.

Prior to ESAP, the ESEF regulation had already addressed machine-readability concerns by requiring the publication of consolidated IFRS financial statements in a machine-readable format. However, while ESEF mandates that issuers publish annual financial reports (AFR) in XHTML format and tag consolidated IFRS financial statements using iXBRL specifications[27], these requirements mainly focus on accounting data. In the absence of specific provisions covering all AFR information, only the data that is explicitly required to be tagged remains truly machine-readable at present.

Throughout this document, the term "tag" by default refers to XHTML tags; any references to iXBRL tags will be explicitly specified where applicable.

Focus on the difference between XHTML and iXBRL tags

XHTML and iXBRL tags are used in distinct contexts, even though both are built on XML. XHTML structures the content of the document, while iXBRL marks financial data so that it is directly accessible by machines.

For example, issuers subject to ESEF must publish their AFRs in XHTML. This format requires the use of XHTML tags, which allow machines to distinguish between section titles, text paragraphs, tables, etc.

Example of an XHTML tag defining a title: `<h1>Annual Financial Report 2023</h1>`

[21] Directive (UE) 2019/1024, Article 2, Paragraph 13 : link
[22] Regulation (EU) 2023/2859, Article 2 Paragraph (3) : link
[23] As long as it is not composed of an image, for example when scanned.
[24] 21 regulations and 16 directives
[25] Regulation (EU) 2023/2859, Article 5
[26] Validation of the level 2 texts by the European Commission is expected by the end of 2024.
[27] These include both a basic taxonomy and an 'extension' taxonomy to allow a degree of flexibility.

Under ESEF, iXBRL is used to tag specific financial elements (such as revenue, profits, etc.). With these specific tags, machines can extract all tagged information.

Example of an iXBRL tag indicating the LEI to identify the issuer:

```
<xbrli:entity>
 <xbrli:identifier
scheme=”http://standards.iso.org/iso/17442”>KGCEPHLVVKVRZYO1T647</xbrli:identifier>
<xbrli:entity>[28]
```

Without these tags, it would either be necessary to manually collect this information one by one or train an AI to extract data from the text, which is more costly and carries a higher risk of errors.

3.2. SYNTHESIS OF INSIGHTS LEARNED FROM THE AMF'S WORK

In the context of the AI projects presented in the first two sections of this note, it has become particularly evident that the performance of the developed tools is intrinsically linked not only to the format (which recent regulatory texts aim to define) but also, and above all, to the quality, encoding, and level of standardization of the processed documents.

The table below summarizes the key lessons learned from these projects, highlighting the challenges encountered regarding machine-readability. These findings are further enriched by similar conclusions drawn from work conducted to extract tagged information from IFRS consolidated financial statements using more traditional Data tools, as required by ESEF.

Table 3: Insights learned in AI work on Taxonomy & SFDR reports

Regulation	Concerned reporting	Regulatory requirements on machine-readability	Processed reportings format	Insights
ESEF	IFRS consolidated financial statements in AFR/URD	Machine-readable format required with data tagging	XHTML and iXBRL tags	- Extraction facilitated by the format and proper use of tags - However, due to the flexibility offered by the extension taxonomy, data consolidation is costly as it requires business experts to provide all mapping and calculation rules (e.g., for the calculation of net debt)
EU Green Taxonomy (in 2022)	Taxonomy report in ARF/URD	None	XHTML[29]	Extraction made difficult and costly due to: - Encoding issues caused by improper use of tagsdes soucis d’encodage liés à une mauvaise utilisation des balises :

[28] Example from the following link.

[29] All the documents from the issuers selected for the Taxonomy reports project were published in XHTML format. In 2022 these reports were poorly formatted overall, so it would probably have been easier to process reports in PDF format.

				▪ Inability to distinguish the documents structure through XHTML tagging, ▪ Absence of XHTML tags to isolate tables from text paragraphs or to extract information from tables. - lack of standardization in content presentation, with hightly heterogeneous tables.
SFDR	SFDR appendices in funds prospectuses	No strict requirement as such, but a relatively standardized form (for example, with an imposed document structure)	PDF	Extraction is relatively facilitated by the standardized form in terms of content but limited by : - the absence of bookmarks simplifying navigation within document sections, - the need of process graphics and images without being able to rely on text or tables, - encoding issues due to the lack of technical standards.

When a document is read by a human, attention is focused on visual presentation, logical structure, and the meaning of the content. However, a machine reads a document in a completely different way. Rather than "understanding" the content, it analyzes raw data through the underlying technical encoding.

The encoding of a document consists of transforming information, such as text, symbols, or images, into a format that computers can understand and display. Whether in a PDF or XHTML format, proper encoding ensures that the document is not only readable for the user but also exploitable by the machine for data extraction. The encoding process is generally transparent to humans and is managed by software applications that create and save a document in a specific format.

Two key factors affect encoding quality:

- The choices made during document drafting (for example, embedding an image to represent a table instead of creating it within the document);
- The application used to save or convert a document into a given format (for example, the tools used by issuers to publish their AFR/URD in XHTML do not allow for optimal encoding quality).

To improve the quality of document encoding and facilitate machine readability, it is necessary to technically standardize document drafting. This includes both unifying human choices in content construction (for example, banning images to represent tables) and specifying the standards that tools should follow to produce machine-readable documents (for example, adherence to World Wide Web Consortium (W3C) [30] standards for XHTML document creation).

Readers are invited to refer to Annex 3 for further details on the challenges encountered with document formats and encoding in projects related to Taxonomy reports and SFDR annexes.

[30] W3C HTML standards : link

It should be noted that, in both experiments conducted, one of the main challenges encountered was the ability to navigate the document structure, which is crucial for the machine to extract relevant information within the correct context. If the document structure is poorly defined or overly complex, it can lead to interpretation errors, where important data is misassociated or completely overlooked. This issue is particularly prevalent in dense documents or those with numerous nested sections.
The experiments conducted by the AMF confirmed these difficulties. Although content standardization, as applied to SFDR annexes, helps reduce errors even in a PDF format, it has proven insufficient to ensure optimal automated processing. Additionally, tests on Taxonomy reports highlighted the need to impose strict usage rules, even for formats considered machine-readable, such as XHTML. These findings indicate that without rigorous tagging, the potential of XHTML remains limited and does not fully meet the objectives of automated exploitation. Thus, although the experiments did not allow for practical verification, the methodical use of tags in an XHTML format appears to be the best solution to minimize errors and ensure reliable and precise data extraction.

3.3. WAYS FOR IMPROVEMENT FOR EXISTING AND FUTURE REGULATIONS

As part of the AI projects conducted on Taxonomy reports, it appears that most of the difficulties outlined in the previous section stem from the lack of XHTML technical specifications to guide the construction of AFR/URD. Unlike the SEC[31] and its EDGAR system (Electronic Data-Gathering, Analysis, and Retrieval), the European Commission has not required issuers to follow the XHTML standards developed by W3C.

However, in 2023, two developments that followed the AMF's AI work on this topic have improved the machine-readability of the Taxonomy sections within AFR/URD:

- the standardization of Taxonomy report tables
- guidance 2.2.6 of the ESEF Reporting Manual[32], which clarified expectations regarding the use of XHTML semantic tags[33].

These new standards could be sufficient to build a "reliable" database of information from the Taxonomy report while awaiting its regulatory transition to a machine-readable format[34].

However, this is not enough to ensure the full machine-readability of AFR/URD (particularly for the Taxonomy report). To achieve this, it would be necessary to propose compliance with W3C standards in the drafting of AFR/URD, especially the use of semantic tags to identify headings, paragraphs, and tables.

Furthermore, the AI work conducted on SFDR annexes highlighted the benefits of standardized reporting, particularly because standardization enhances the reliability of results while reducing development costs. However, standardization is currently limited in the case of SFDR annexes and should be strengthened to improve their machine-readability:

- by dual-publishing key data contained in an image, in the form of a text paragraph or a table[35] ;

[31] Filer Manual for 10-K filings, section 5.2.2 : link
[32] Reporting Manual ESEF, Guidance 2.2.6, page 30 : link
[33] The *guidance* is aimed particularly at the parties concerned by ESEF, but it has nevertheless led to an overall increase in the use of tags for all reports.
[34] https://www.esma.europa.eu/issuer-disclosure/electronic-reporting
[35] As far as the usability of graphs is concerned, the most recent AI models are still a long way off the human capacity to read this type of data. The implication is that even with considerable resources, it is not possible to automatically extract and structure data from graphs with satisfactory performance : link

- by establishing a technical standard to be followed when filling out the form (see examples in Annex 3: formatting and encoding issues) ;
- by adding bookmarks to facilitate navigation within document sections.

More broadly, the improvement paths identified earlier are relevant to all other regulations requiring the reporting of unstructured data, whose processing would necessitate automation. While the systematic use of genuinely machine-readable formats is not always justified in terms of cost-benefit analysis, it is important to anticipate potential needs for automated data collection and define reporting rules accordingly.

At the conclusion of its work, the AMF identifies five key lessons to facilitate machine-readability without compromising human readability in future regulations:

- **The use of a machine-readable format alone is not sufficient and must be accompanied by specific rules to achieve an optimal level of exploitability.**
- **Any image containing data intended for machine processing must be accompanied by a descriptive text or table containing the same information.**
- **The standardization of the reporting structure and the information it contains.**
- **The technical standardization of documents.**
- **The widespread adoption of the XHTML format with compliance to W3C standards, even for future regulations that do not provide for the integration of iXBRL tags in a machine-readable format.**

***

GLOSSARY

AI system : a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments (Article 3, AI Act).

HTML : mark-up language used to create web pages and documents, in particular to define hypertext links.

HTML specifications: technical specifications proposed, for example, by the World Wide Web Consortium (W3C) to lay down the rules for using the HTML language to create web pages so that they are consistent and more easily read by a machine.

XBRL (eXtensible Business Reporting Language) is an XML-based computer language designed specifically for automating business information requirements, such as preparing, sharing and analysing financial reports and statements.

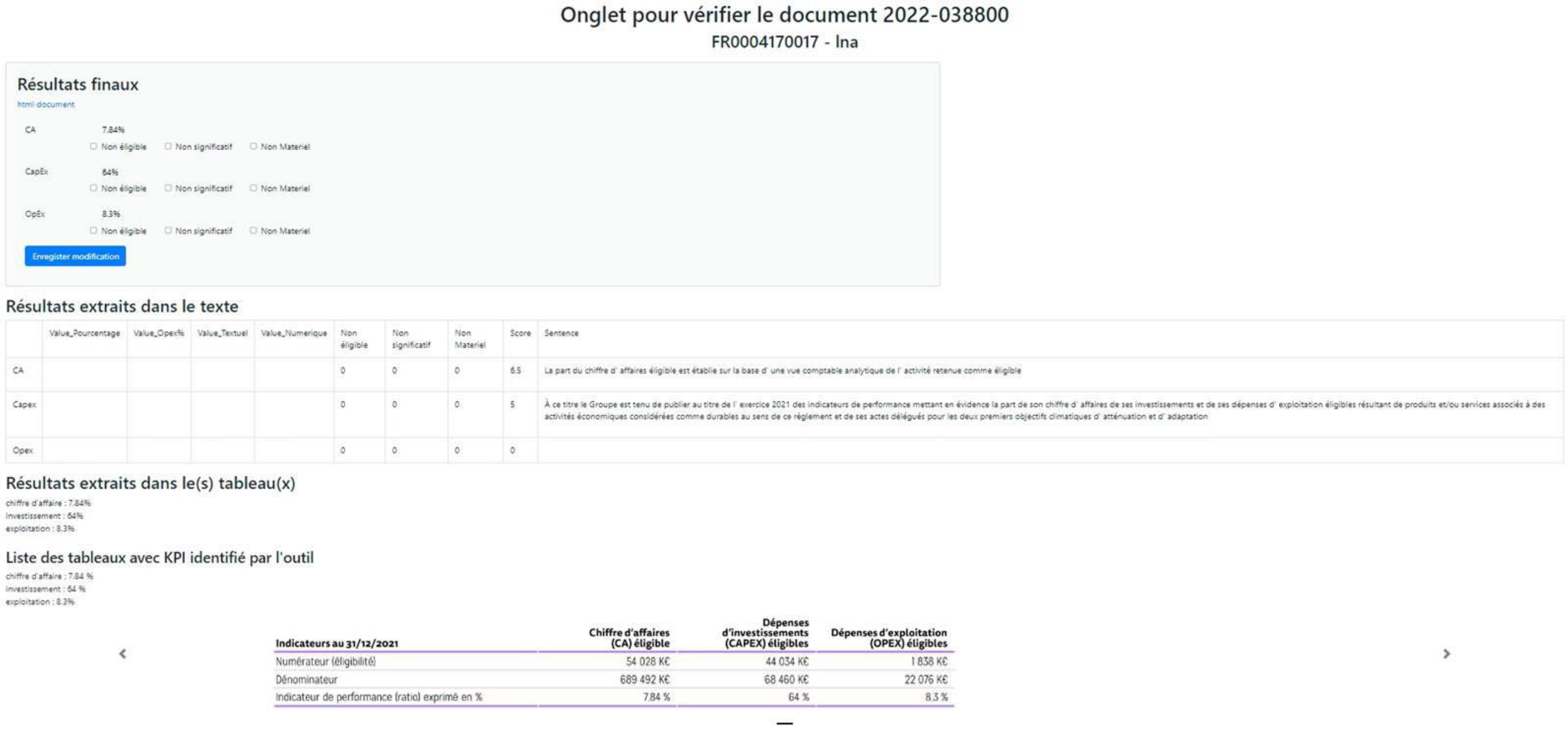
Inline-XBRL : iXBRL, or Inline XBRL, is an open standard that allows a single document to provide both human-readable data and machine-readable structured data.

Machine-readable: *a file format structured in such a way that software applications can easily identify, recognise and extract specific data, in particular each statement of fact and its internal structure.*

Data Extractable: *a platform-independent open electronic file format made available to the public without any restrictions preventing the documents from being re-used. The format is widely used or required by law, allows data to be extracted by a machine and is not just human-readable.*

NLP: Natural Language Processing is a multidisciplinary field involving linguistics, computer science and artificial intelligence. Its aim is to create tools capable of interpreting and synthesising text for various applications.

APPENDIX 1: INTERFACE FOR CONSULTING THE RESULTS OF THE EXPERIMENT ON TAXONOMY REPORTS



Onglet pour vérifier le document 2022-038800

FR0004170017 - Ina

Résultats finaux

html document

CA 7.84%
□ Non éligible □ Non significatif □ Non Materiel

CapEx 64%
□ Non éligible □ Non significatif □ Non Materiel

OpEx 8.3%
□ Non éligible □ Non significatif □ Non Materiel

Résultats extraits dans le texte

	Value_Pourcentage	Value_Opex%	Value_Textuel	Value_Numerique	Non éligible	Non significatif	Non Materiel	Score	Sentence
CA					0	0	0	6.5	La part du chiffre d' affaires éligible est établie sur la base d' une vue comptable analytique de l' activité retenue comme éligible
Capex					0	0	0	5	À ce titre le Groupe est tenu de publier au titre de l' exercice 2021 des indicateurs de performance mettant en évidence la part de son chiffre d' affaires de ses investissements et de ses dépenses d' exploitation éligibles résultant de produits et/ou services associés à des activités économiques considérées comme durables au sens de ce règlement et de ses actes délégués pour les deux premiers objectifs climatiques d' atténuation et d' adaptation
Opex					0	0	0	0	

Résultats extraits dans le(s) tableau(x)

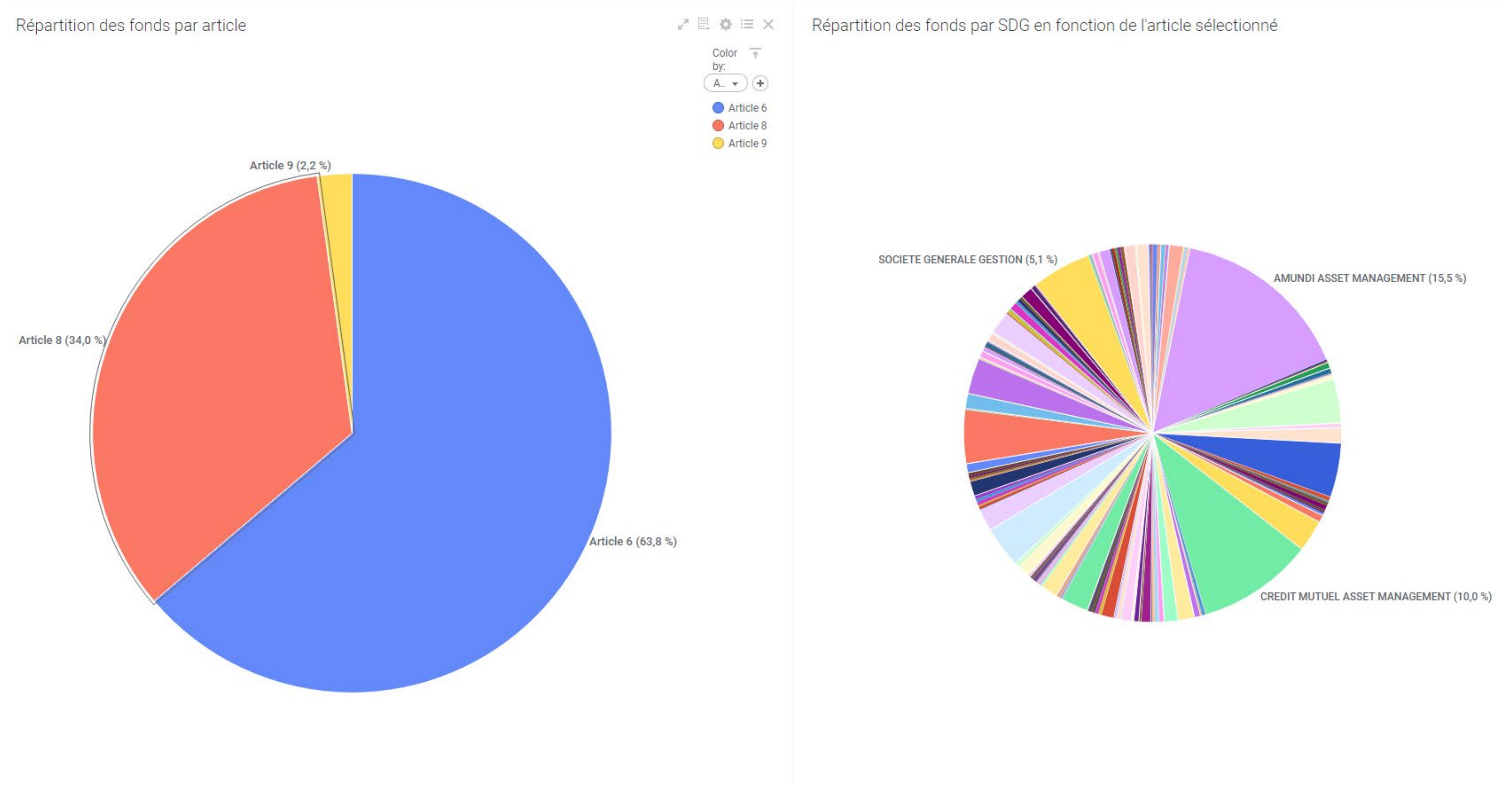
chiffre d'affaire : 7.84%
investissement : 64%
exploitation : 8.3%

Liste des tableaux avec KPI identifié par l'outil

chiffre d'affaire : 7.84 %
investissement : 64 %
exploitation : 8.3%

Indicateurs au 31/12/2021	Chiffre d'affaires (CA) éligible	Dépenses d'investissements (CAPEX) éligibles	Dépenses d'exploitation (OPEX) éligibles
Numérateur (éligibilité)	54 028 K€	44 034 K€	1 838 K€
Dénominateur	689 492 K€	68 460 K€	22 076 K€
Indicateur de performance (ratio) exprimé en %	7.84 %	64 %	8.3 %

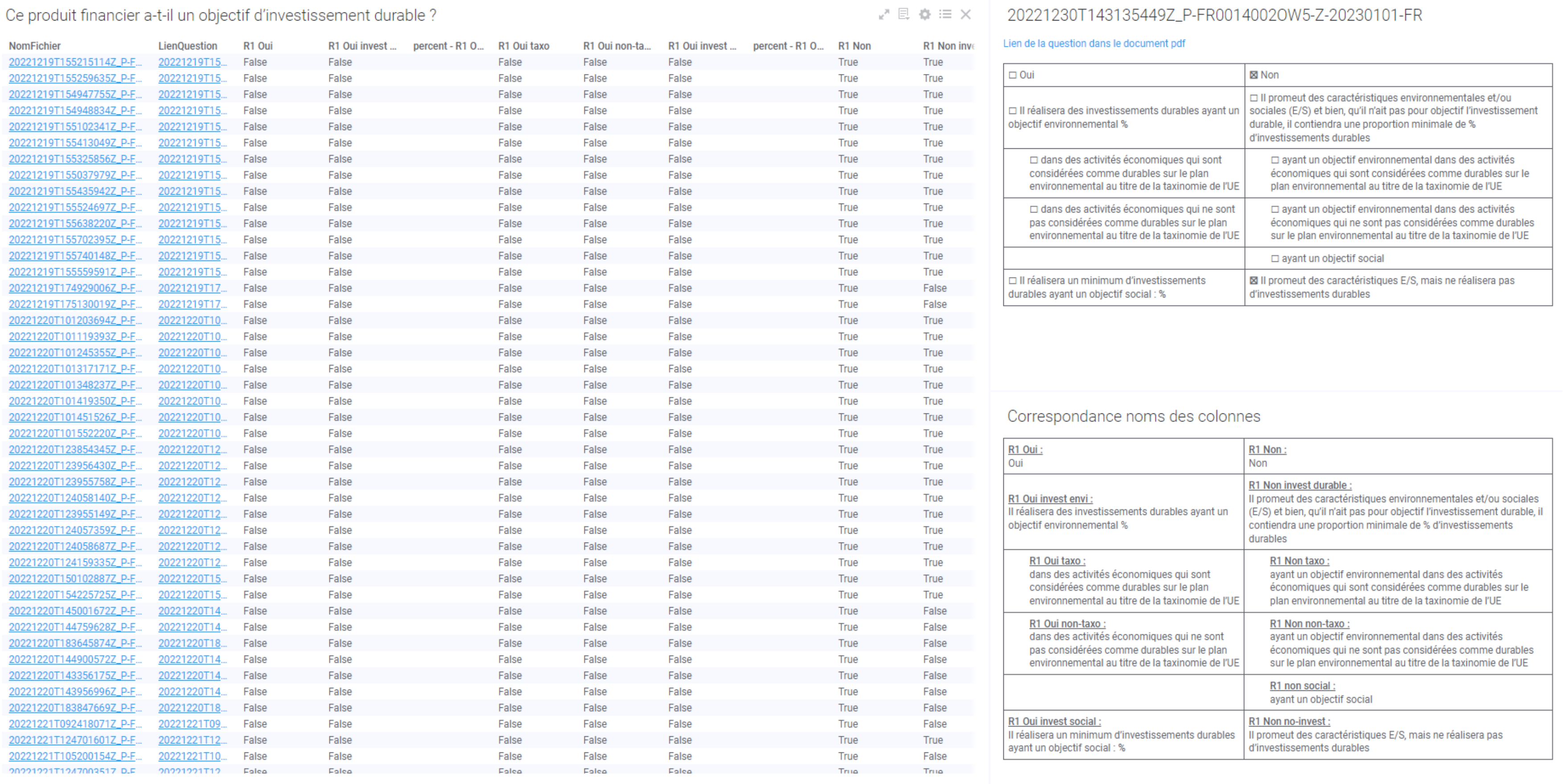
APPENDIX 2: INTERFACE FOR THE TOOL DEPLOYED TO EXPLORE SFDR APPENDICES



Répartition des fonds par article

Répartition des fonds par SDG en fonction de l'article sélectionné

Color by:
A..
● Article 6
● Article 8
● Article 9
Article 9 (2,2 %)
Article 8 (34,0 %)
Article 6 (63,8 %)
SOCIETE GENERALE GESTION (5,1 %)
AMUNDI ASSET MANAGEMENT (15,5 %)
CREDIT MUTUEL ASSET MANAGEMENT (10,0 %)



Ce produit financier a-t-il un objectif d'investissement durable ?

NomFichier	LienQuestion	R1 Oui	R1 Oui invest ...	percent - R1 O...	R1 Oui taxo	R1 Oui non-ta...	R1 Oui invest ...	percent - R1 O...	R1 Non	R1 Non inv
20221219T155215114Z_P-F...	20221219T15...	False	False		False	False	False		True	True
20221219T155259635Z_P-F...	20221219T15...	False	False		False	False	False		True	True
20221219T154947755Z_P-F...	20221219T15...	False	False		False	False	False		True	True
20221219T154948834Z_P-F...	20221219T15...	False	False		False	False	False		True	True
20221219T155102341Z_P-F...	20221219T15...	False	False		False	False	False		True	True
20221219T155413049Z_P-F...	20221219T15...	False	False		False	False	False		True	True
20221219T155325856Z_P-F...	20221219T15...	False	False		False	False	False		True	True
20221219T155037979Z_P-F...	20221219T15...	False	False		False	False	False		True	True
20221219T155435942Z_P-F...	20221219T15...	False	False		False	False	False		True	True
20221219T155524697Z_P-F...	20221219T15...	False	False		False	False	False		True	True
20221219T155638220Z_P-F...	20221219T15...	False	False		False	False	False		True	True
20221219T155702395Z_P-F...	20221219T15...	False	False		False	False	False		True	True
20221219T155740148Z_P-F...	20221219T15...	False	False		False	False	False		True	True
20221219T155559591Z_P-F...	20221219T15...	False	False		False	False	False		True	True
20221219T174929006Z_P-F...	20221219T17...	False	False		False	False	False		True	False
20221219T175130019Z_P-F...	20221219T17...	False	False		False	False	False		True	False
20221220T101203694Z_P-F...	20221220T10...	False	False		False	False	False		True	True
20221220T101119393Z_P-F...	20221220T10...	False	False		False	False	False		True	True
20221220T101245355Z_P-F...	20221220T10...	False	False		False	False	False		True	True
20221220T101317171Z_P-F...	20221220T10...	False	False		False	False	False		True	True
20221220T101348237Z_P-F...	20221220T10...	False	False		False	False	False		True	True
20221220T101419350Z_P-F...	20221220T10...	False	False		False	False	False		True	True
20221220T101451526Z_P-F...	20221220T10...	False	False		False	False	False		True	True
20221220T101552220Z_P-F...	20221220T10...	False	False		False	False	False		True	True
20221220T123854345Z_P-F...	20221220T12...	False	False		False	False	False		True	True
20221220T123956430Z_P-F...	20221220T12...	False	False		False	False	False		True	True
20221220T123955758Z_P-F...	20221220T12...	False	False		False	False	False		True	True
20221220T124058140Z_P-F...	20221220T12...	False	False		False	False	False		True	True
20221220T123955149Z_P-F...	20221220T12...	False	False		False	False	False		True	True
20221220T124057359Z_P-F...	20221220T12...	False	False		False	False	False		True	True
20221220T124058687Z_P-F...	20221220T12...	False	False		False	False	False		True	True
20221220T124159335Z_P-F...	20221220T12...	False	False		False	False	False		True	True
20221220T150102887Z_P-F...	20221220T15...	False	False		False	False	False		True	True
20221220T154225725Z_P-F...	20221220T15...	False	False		False	False	False		True	True
20221220T145001672Z_P-F...	20221220T14...	False	False		False	False	False		True	False
20221220T144759628Z_P-F...	20221220T14...	False	False		False	False	False		True	False
20221220T183645874Z_P-F...	20221220T18...	False	False		False	False	False		True	False
20221220T144900572Z_P-F...	20221220T14...	False	False		False	False	False		True	False
20221220T143356175Z_P-F...	20221220T14...	False	False		False	False	False		True	False
20221220T143956996Z_P-F...	20221220T14...	False	False		False	False	False		True	False
20221220T183847669Z_P-F...	20221220T18...	False	False		False	False	False		True	False
20221221T092418071Z_P-F...	20221221T09...	False	False		False	False	False		True	False
20221221T124701601Z_P-F...	20221221T12...	False	False		False	False	False		True	True
20221221T105200154Z_P-F...	20221221T10...	False	False		False	False	False		True	False
20221221T124700351Z_P-F	20221221T12	False	False		False	False	False		True	True

20221230T143135449Z_P-FR0014002OW5-Z-20230101-FR

Lien de la question dans le document pdf

☐ Oui	☒ Non
☐ Il réalisera des investissements durables ayant un objectif environnemental %	☐ Il promeut des caractéristiques environnementales et/ou sociales (E/S) et bien, qu'il n'ait pas pour objectif l'investissement durable, il contiendra une proportion minimale de % d'investissements durables
☐ dans des activités économiques qui sont considérées comme durables sur le plan environnemental au titre de la taxinomie de l'UE	☐ ayant un objectif environnemental dans des activités économiques qui sont considérées comme durables sur le plan environnemental au titre de la taxinomie de l'UE
☐ dans des activités économiques qui ne sont pas considérées comme durables sur le plan environnemental au titre de la taxinomie de l'UE	☐ ayant un objectif environnemental dans des activités économiques qui ne sont pas considérées comme durables sur le plan environnemental au titre de la taxinomie de l'UE
	☐ ayant un objectif social
☐ Il réalisera un minimum d'investissements durables ayant un objectif social : %	☒ Il promeut des caractéristiques E/S, mais ne réalisera pas d'investissements durables

Correspondance noms des colonnes

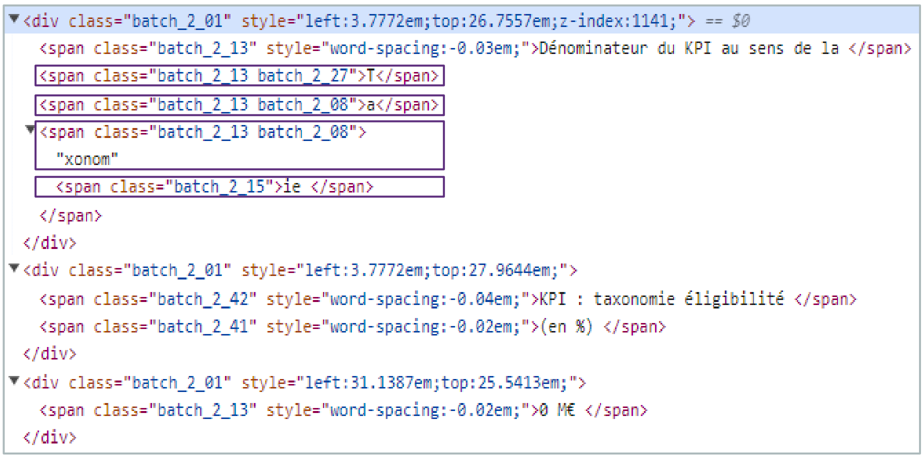
R1 Oui : Oui	R1 Non : Non
R1 Oui invest envi : Il réalisera des investissements durables ayant un objectif environnemental %	R1 Non invest durable : Il promeut des caractéristiques environnementales et/ou sociales (E/S) et bien, qu'il n'ait pas pour objectif l'investissement durable, il contiendra une proportion minimale de % d'investissements durables
R1 Oui taxo : dans des activités économiques qui sont considérées comme durables sur le plan environnemental au titre de la taxinomie de l'UE	R1 Non taxo : ayant un objectif environnemental dans des activités économiques qui sont considérées comme durables sur le plan environnemental au titre de la taxinomie de l'UE
R1 Oui non-taxo : dans des activités économiques qui ne sont pas considérées comme durables sur le plan environnemental au titre de la taxinomie de l'UE	R1 Non non-taxo : ayant un objectif environnemental dans des activités économiques qui ne sont pas considérées comme durables sur le plan environnemental au titre de la taxinomie de l'UE
	R1 non social : ayant un objectif social
R1 Oui invest social : Il réalisera un minimum d'investissements durables ayant un objectif social : %	R1 Non no-invest : Il promeut des caractéristiques E/S, mais ne réalisera pas d'investissements durables

APPENDIX 3: FORMAT PROBLEMS AND ENCODING

The visual aspect, which is directly understandable for a human, can be encoded in various ways depending on the chosen format (Word, PDF, XHTML, etc.). These differences can also be observed within a given format, such as XHTML, where the method of construction chosen by one author may differ from that of another[36]. Figure 3 illustrates how part of the table from the document in Figure 1 is presented in XHTML format. The encoding of the table at the bottom of Figure 3 does not use the appropriate tags[37] to indicate that the content is within a table with rows and columns. The order in which the elements appear is unclear and does not match the visual order, with the word "Taxonomie" being split into four separate parts across multiple tags ("T," "a," "xonom," and "ie")[38].

For the information in this table to be considered sufficiently data-extractable, it should have been enclosed within « `table` » tags, the headers within « `thead` » tags, each row within a « `tbody` » tag, and each column belonging to a row within « `tr` » tags. Making these elements machine-readable would have required the creation of a dedicated XBRL taxonomy for the Taxonomy regulation to assign a label to each value present in the table.

Figure 3 : part of the content of the previous table in xhtml



```
▼<div class="batch_2_01" style="left:3.7772em;top:26.7557em;z-index:1141;"> == $0
   <span class="batch_2_13" style="word-spacing:-0.03em;">Dénominateur du KPI au sens de la </span>
   <span class="batch_2_13 batch_2_27">T</span>
   <span class="batch_2_13 batch_2_08">a</span>
  ▼<span class="batch_2_13 batch_2_08">
     "xonom"
     <span class="batch_2_15">ie </span>
   </span>
 </div>
▼<div class="batch_2_01" style="left:3.7772em;top:27.9644em;">
   <span class="batch_2_42" style="word-spacing:-0.04em;">KPI : taxonomie éligibilité </span>
   <span class="batch_2_41" style="word-spacing:-0.02em;">(en %) </span>
 </div>
▼<div class="batch_2_01" style="left:31.1387em;top:25.5413em;">
   <span class="batch_2_13" style="word-spacing:-0.02em;">0 M€ </span>
 </div>
```

Figures 4 and 5 below provide an example of a correctly structured XHTML document: they present a table of Taxonomy CAPEX published in 2023 by an issuer (Figure 4) and a portion of its XHTML code (Figure 5). Unlike Figure 3 above, the table in Figure 4 is properly structured using "semantic" tags, as specified by W3C standards. As a result, it is automatically detectable by a machine and can be converted into a data table that can be processed by any structured data analysis tool.

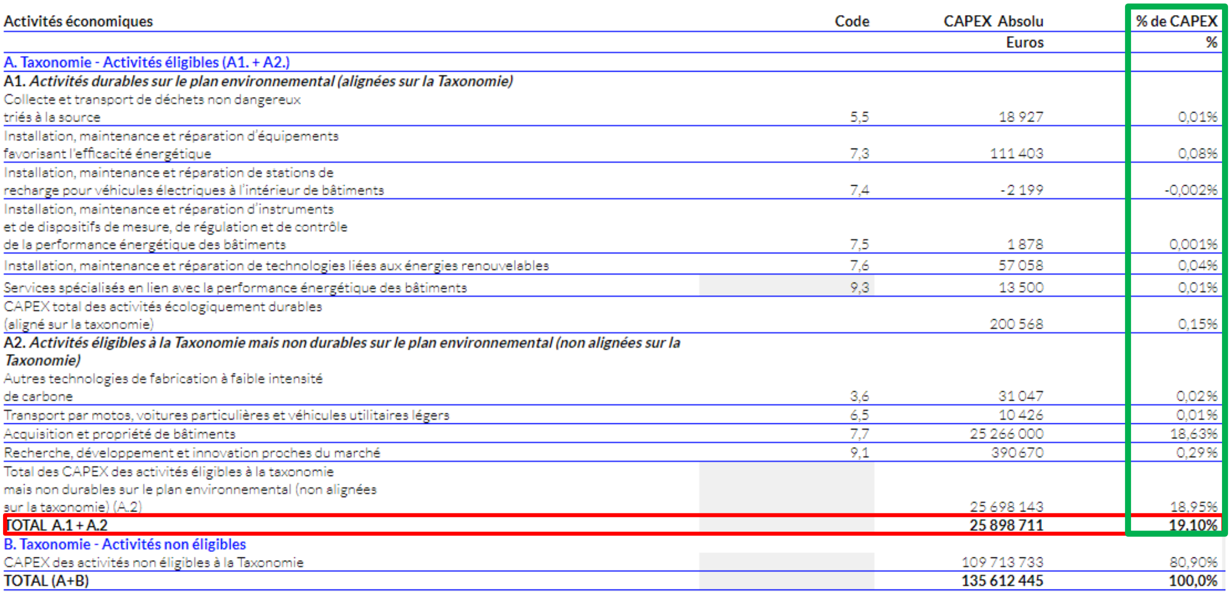
For example, it is instantly possible to isolate the row "TOTAL A.1 + A.2" (red box) and the column "% of CAPEX" (green box) to extract the KPI value, which is "19.10%" (at the intersection of the two).

[36] Le rédacteur peut par exemple rédiger son document directement en XHTML, ou faire une conversion d'un document Word vers le format XHTML. Selon l'approche choisie le document ne sera pas constitué de la même manière, ce qui peut complexifier le traitement automatique du document.

[37] Les balises adéquates sont les balises dites « sémantiques » et sont référencées dans les lignes de conduite W3C. Elles permettent par exemple d'indiquer un titre (et de spécifier son niveau), un paragraphe, une image ou encore une table.

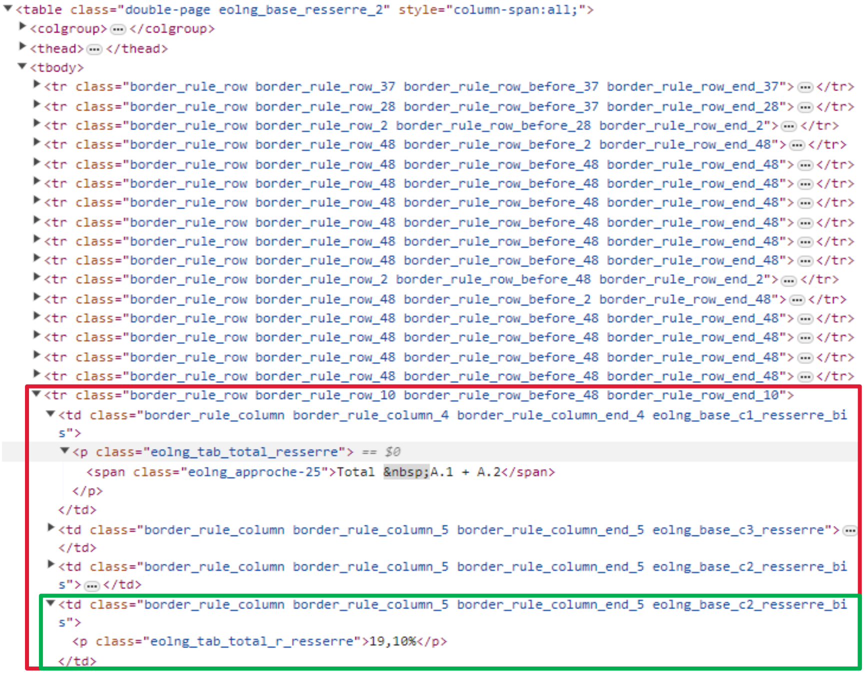
[38] Ceci n'est qu'un exemple, la DDS a observé de très nombreuses variantes allant de certaines presque machine-readable à certaines totalement illisibles.

Figure 4 : example of a standardized table relating to CAPEX and published in 2023



Activités économiques	Code	CAPEX Absolu	% de CAPEX
		Euros	%
A. Taxonomie - Activités éligibles (A1. + A2.)			
A1. *Activités durables sur le plan environnemental (alignées sur la Taxonomie)*			
Collecte et transport de déchets non dangereux triés à la source	5,5	18 927	0,01%
Installation, maintenance et réparation d'équipements favorisant l'efficacité énergétique	7,3	111 403	0,08%
Installation, maintenance et réparation de stations de recharge pour véhicules électriques à l'intérieur de bâtiments	7,4	-2 199	-0,002%
Installation, maintenance et réparation d'instruments et de dispositifs de mesure, de régulation et de contrôle de la performance énergétique des bâtiments	7,5	1 878	0,001%
Installation, maintenance et réparation de technologies liées aux énergies renouvelables	7,6	57 058	0,04%
Services spécialisés en lien avec la performance énergétique des bâtiments	9,3	13 500	0,01%
CAPEX total des activités écologiquement durables (aligné sur la taxonomie)		200 568	0,15%
A2. *Activités éligibles à la Taxonomie mais non durables sur le plan environnemental (non alignées sur la Taxonomie)*			
Autres technologies de fabrication à faible intensité de carbone	3,6	31 047	0,02%
Transport par motos, voitures particulières et véhicules utilitaires légers	6,5	10 426	0,01%
Acquisition et propriété de bâtiments	7,7	25 266 000	18,63%
Recherche, développement et innovation proches du marché	9,1	390 670	0,29%
Total des CAPEX des activités éligibles à la taxonomie mais non durables sur le plan environnemental (non alignées sur la taxonomie) (A.2)		25 698 143	18,95%
TOTAL A.1 + A.2		**25 898 711**	**19,10%**
B. Taxonomie - Activités non éligibles			
CAPEX des activités non éligibles à la Taxonomie		109 713 733	80,90%
TOTAL (A+B)		**135 612 445**	**100,0%**

Figure 5 : XHTML code for the table in Figure 4



```
▼<table class="double-page eolng_base_resserre_2" style="column-span:all;">
  ▶<colgroup>⋯</colgroup>
  ▶<thead>⋯</thead>
  ▼<tbody>
    ▶<tr class="border_rule_row border_rule_row_37 border_rule_row_before_37 border_rule_row_end_37">⋯</tr>
    ▶<tr class="border_rule_row border_rule_row_28 border_rule_row_before_37 border_rule_row_end_28">⋯</tr>
    ▶<tr class="border_rule_row border_rule_row_2 border_rule_row_before_28 border_rule_row_end_2">⋯</tr>
    ▶<tr class="border_rule_row border_rule_row_48 border_rule_row_before_2 border_rule_row_end_48">⋯</tr>
    ▶<tr class="border_rule_row border_rule_row_48 border_rule_row_before_48 border_rule_row_end_48">⋯</tr>
    ▶<tr class="border_rule_row border_rule_row_48 border_rule_row_before_48 border_rule_row_end_48">⋯</tr>
    ▶<tr class="border_rule_row border_rule_row_48 border_rule_row_before_48 border_rule_row_end_48">⋯</tr>
    ▶<tr class="border_rule_row border_rule_row_48 border_rule_row_before_48 border_rule_row_end_48">⋯</tr>
    ▶<tr class="border_rule_row border_rule_row_48 border_rule_row_before_48 border_rule_row_end_48">⋯</tr>
    ▶<tr class="border_rule_row border_rule_row_48 border_rule_row_before_48 border_rule_row_end_48">⋯</tr>
    ▶<tr class="border_rule_row border_rule_row_2 border_rule_row_before_48 border_rule_row_end_2">⋯</tr>
    ▶<tr class="border_rule_row border_rule_row_48 border_rule_row_before_2 border_rule_row_end_48">⋯</tr>
    ▶<tr class="border_rule_row border_rule_row_48 border_rule_row_before_48 border_rule_row_end_48">⋯</tr>
    ▶<tr class="border_rule_row border_rule_row_48 border_rule_row_before_48 border_rule_row_end_48">⋯</tr>
    ▶<tr class="border_rule_row border_rule_row_48 border_rule_row_before_48 border_rule_row_end_48">⋯</tr>
    ▶<tr class="border_rule_row border_rule_row_48 border_rule_row_before_48 border_rule_row_end_48">⋯</tr>
    ▼<tr class="border_rule_row border_rule_row_10 border_rule_row_before_48 border_rule_row_end_10">
      ▼<td class="border_rule_column border_rule_column_4 border_rule_column_end_4 eolng_base_c1_resserre_bi
       s">
        ▼<p class="eolng_tab_total_resserre"> == $0
           <span class="eolng_approche-25">Total &nbsp;A.1 + A.2</span>
         </p>
       </td>
      ▶<td class="border_rule_column border_rule_column_5 border_rule_column_end_5 eolng_base_c3_resserre">⋯
       </td>
      ▶<td class="border_rule_column border_rule_column_5 border_rule_column_end_5 eolng_base_c2_resserre_bi
       s">⋯</td>
      ▼<td class="border_rule_column border_rule_column_5 border_rule_column_end_5 eolng_base_c2_resserre_bi
       s">
         <p class="eolng_tab_total_r_resserre">19,10%</p>
       </td>
```

Finally, for PDFs, particularly SFDR annexes, data extractability is not necessarily achievable at a low cost and without errors.

Figure 2.a presents a form where checkboxes and percentages had to be identified automatically. The most efficient and widely used approach for processing PDF documents is to convert them into a text format[39]. However, the way the form is completed can vary depending on the asset management company that produced the annex, to the point of preventing the converter from extracting all the information.

Figure 6 below illustrates the issues caused by the heterogeneity of form completion methods. It shows that after converting the SFDR annex from which Figure 3 is taken, the checked and unchecked boxes disappeared, indicating that they were present as images or drawings[40] rather than as characters such as "☑" or "☐". The AMF also observed that some asset management companies used the character "X" or any other character that a custom font transforms into a checked box or a cross.

Figure 6 : exemple de conversion en texte d'une partie de la figure 3

```
Does this financial product have a sustainable investment objective?
Yes
No
It will make a minimum of sustainable investments with an environmental objective: 100%

in economic activities that are considered environmentally sustainable under the EU Taxonomy

in economic activities that are not considered environmentally sustainable under the EU
Taxonomy

It will make a minimum of sustainable investments with a social objective: ___%
```

[39] Open-source conversion tools exist for this purpose and offer relatively satisfactory performance. For example, the AMF used PyMuPDF and Unstructured.

[40] Drawings are objects specific to PDF documents that editors can add to create shapes such as circles and rectangles.

APPENDIX 4 : AUTOMATIC PROCESSINGS

TAXONOMY

The tool for extracting information related to the EU Green Taxonomy from the 2022 reports of non-financial companies is based on a sequence of modules incorporating artificial intelligence systems (AIS). This structured sequence enables the tool to organize information within the Universal Registration Documents (URD), identify the relevant section, extract key performance indicators (KPIs) from paragraphs or tables, aggregate the results, and finally conduct verification and inference processes.

To assist the AMF in saving time during the extraction of Taxonomy-related information, the tool must be able to:

- □ extract the eligibility shares of CapEx, OpEx, and revenue in each document;
- □ identify any reference to a materiality exemption clause or a non-eligibility statement for one or more of these KPIs; and,
- □ reference the paragraphs and/or tables containing information related to the Taxonomy.

Universal Registration Documents (URD) are very dense (several hundred pages) and contain various sections, including those related to Taxonomy information. The documents analyzed in this study are published in XHTML, a format that should have enabled automatic processing through the use of a tag-based system to define the content structure (title, section, subsection, etc.) and reference specific information. However, due to the lack of specifications on tag usage, significant additional development was required. These preliminary developments had to be completed before building the Taxonomy-specific content extraction modules and resulted in the creation of technical components (which can be reused beyond this study) that allow the system to navigate through URDs[41]. It should be noted that some of these developments would not have been necessary if the documents had still been published in PDF format, despite its lower data extractability.

Once the tool can isolate the sections of a document related to the Taxonomy report (see the focus on automatic detection of the "Taxonomy" section in URDs below), it must then extract KPIs (and associated narratives) from these sections. Two main cases arise: If the issuer has exclusively published its KPIs in text paragraphs, the tool simply applies text-processing techniques. If all or part of the required information is presented in a table, image-processing techniques have shown to yield the best results for extracting KPIs from tables. Additionally, in most documents, both configurations are combined. Therefore, in most URDs, the tool must process both text and tables before performing a final consolidation of the results. This step is particularly important to handle cases where the level of granularity of the KPIs[42] differs between the two representations.

The approach developed for text processing in the extraction of Taxonomy-related information relies on a set of techniques based on both unsupervised and supervised learning[43] (detailed in Annex II):

- □ Named entity recognition, which is tasked with identifying (and extracting) mentions of KPIs in the text (such as "CapEx," "investment expenditures," or "revenue") as well as their quantitative values (for example, "35%," "249 million," or "zero"), different activities (such as "extraction activities" or "energy production activities"), or organizations (such as "the Group" or "its subsidiaries").

[41] All the building blocks developed as part of the study are presented in Appendix II.

[42] While KPIs must be presented at group level, they can be broken down by entity or by activity/business group.

[43] Unsupervised learning is a branch of machine learning characterised by the analysis and clustering of unlabelled data, whereas supervised learning uses labelled data to learn how to predict these labels.

- □ Entity resolution, which aims to determine for each mention of an entity in a text (such as "investment expenditures" or "operating expenditures") the exact indicator to which the issuer is referring[44]. In other words, the algorithm learns to differentiate nuances in the text, such as distinguishing between a reference to OpEx as defined by the Taxonomy and OpEx as defined by IFRS standards. This step also enables the tool to differentiate between the meanings of "eligibility" and "alignment" of activities, as well as to identify whether the text refers to the activities of a specific subsidiary or the entire issuer group.

For example, in the following excerpt:

"The amount of OpEx as defined by the Taxonomy Regulation represents less than 3% of the Group's total operating expenditures for the 2021 fiscal year and is not considered significant."

The simple mentions of "OpEx" (1) and "operating expenditures" (2) do not allow for a definitive conclusion about the type of OpEx to which the issuer is referring. However, entity resolution enables the tool to automatically understand that:

- ○ refers to OpEx as defined by the Taxonomy (the KPI the tool seeks to extract).
- ○ refers to the Group's IFRS OpEx (on which the eligibility percentage is calculated).

By perceiving these nuances, the system can deduce that the issuer's narrative, justifying that its Taxonomy OpEx is not significant, is correct. It is also possible to infer that these narratives apply to the entire group and not just to a subsidiary.

- □ attribute detection, which aims to identify specific characteristics of certain types of entities, for example, whether a KPI is described by the issuer as "Not calculated," "Not significant," "Not material," or "Not eligible."
- □ relationship extraction, which links various entity mentions to each other, for example, associating a KPI with the correct corresponding amount or percentage within the paragraph, linking an activity to its related amount, percentage, or organization (issuer or subsidiary).

The approach developed for table processing consists of three main steps (detailed in Annex IV) [45]:

- □ first, the tool must be able to detect the presence of tables. To achieve this, the system converts the previously extracted "Taxonomy" section into an image and applies a pre-trained table detection algorithm[46]. A new image is generated for each identified table.
- □ in the second step, the tool must analyze the structure of each table to determine where the information to be extracted is located. In other words, using the image of each table identified in the previous step, a combination of algorithms and rules is applied to detect the outlines of the cells.
- □ the final step consists of extracting the information. Once the table structure's contours are clearly identified, the table is converted into a structured data table[47]. Then, a set of rules identifies the exact cell containing the value of each KPI based on column names, row labels, table format, and cell content[48].

[44] For the purposes of the experiment, these real entities are limited to the following types: 'Eligible CapEx', 'CapEx as defined by IFRS', 'Eligible OpEx', 'OpEx as defined by Taxonomy', 'OpEx as defined by IFRS', 'Eligible Income', 'Income as defined by IFRS', 'Specific activities' and 'Organisation'.

[45] The few URDs that correctly defined their tables using specific XHTML tags did not require image processing.

[46] Detection is based on the Paddle framework (link) - the model is described in this paper: link.

[47] The conversion is based on the PaddlOCR framework – Link

[48] Research has developed advanced approaches based on neural networks, but these have not yet been available in French. For reference, see: 'TAPAS: Weakly Supervised Table Parsing via Pre-training', J. Herzig et al. – Link

SFDR

The tool for assisting in the supervision of SFDR annexes in fund prospectuses consists of a sequence of modules incorporating artificial intelligence systems (AIS). This structured sequence allows for organizing information within the SFDR annexes by locating the annexes and identifying question-answer pairs where applicable, determining the article to which the annex is subject, extracting specific information, and conducting a series of automated checks.

To help the AMF save time in supervising SFDR annexes, the tool must be able to:

- □ determine the corresponding article for the prospectus (Article 6, 8, or 9);
- □ locate and reference the question-answer pairs within the annexes;
- □ extract information from the objectives form and the asset allocation and alignment charts with the EU Green Taxonomy.

Prospectuses (Regulation (EU) 2017/1129) are long and dense documents (several dozen pages) containing various sections and annexes, including those related to the extra-financial information required by SFDR. The annexes analyzed in this study are published in a relatively standardized PDF format (see Annexes II and III of Regulation (EU) 2019/2088).

The standardization of the document limits the freedom of authors by constraining both the form and content, which simplifies the reconstruction of the document structure and the referencing of specific information. However, the PDF format is designed to facilitate human readability rather than machine exploitation. For instance, when a machine processes the document, most of the structure is lost. This loss of structure necessitates specific preprocessing developments to ensure the annexes are of sufficient quality for machine processing. The required preprocessing developments include converting the PDF into machine-readable data, cleaning the extracted data, and referencing all questions and answers in the document.

The approach developed for preprocessing SFDR annexes follows a sequence of successive tasks primarily based on open-source tools, algorithm implementations, and human engineering:

- □ PDF file conversion, which aims to transform human-readable content into machine-exploitable data. This conversion is performed using the open-source tool PyMuPDF, allowing the machine to interact with the content of the PDF.
- □ Data cleaning, which significantly improves the quality of extracted data by correcting conversion errors inherent to this format.
- □ Referencing all questions, which indexes all portions of the document. The identification of questions is done using the fuzzysearch library, where an implemented algorithm searches for the presence and location of predefined questions in the document based on regulatory texts and the relevant article.
- □ Extracting answers provided by the asset management company, which is done mechanically by determining the coordinates of each answer based on the location of the current question and the following question in the document.

Once the document is preprocessed, the process continues with four main steps: a module for identifying the applicable article and three extraction modules for retrieving information related to objectives, asset allocation, and alignment with the Taxonomy. These modules analyze both text and the document's visual representation by leveraging the coordinates of characters, words, and paragraphs as well as embedded images.

The data processing methods developed to extract the required information and support SFDR annex supervision are detailed below:

- □ Fund classification, which identifies the article number (6, 8, or 9) applicable to the prospectus. If no SFDR annex is detected, the fund is classified as Article 6. Otherwise, it is categorized as Article 8 or 9 based on whether most of the identified questions belong to the Article 8 or Article 9 annexes.
- □ Analysis of sustainable investment objectives, which extracts information from the checkbox forms on the first page of the annexes, whether they are indicated by a marked or unmarked box or a stated percentage. As previously noted in this document, part of the content is lost or obfuscated during PDF conversion. This module applies several deterministic processing steps: identifying checkboxes, identifying narratives, linking checkboxes to corresponding narratives, classifying checkboxes as empty or checked, and extracting percentages if present.
- □ Extraction of asset allocations, which retrieves asset allocation percentages within each required classification component (e.g., "#1 Aligned with environmental and social characteristics," "#2 Other," "#1A Sustainable"). Percentages are extracted either from graphs using regular expressions or from text using SpaCy, which decomposes responses into sentences to identify components and their associated percentages.
- □ Analysis of investment alignment with the EU Taxonomy, which extracts alignment percentages from pie charts. This module leverages visual information from the PDF (coordinates, colors, proximity) and text to identify percentages, determine their corresponding legend in each graph, and associate the percentage with its correct label (aligned, non-aligned, sovereign bonds, etc.).

Finally, the last processing step detects inconsistencies and raises alerts regarding the quality of the document’s reporting:

- □ This module checks several aspects, such as the document's language, consistency between the listed questions and the article indicated in the sustainable investment objectives section, missing answers to required questions, the document’s machine-readability, and certain compliance checks based on the specific question analyzed.