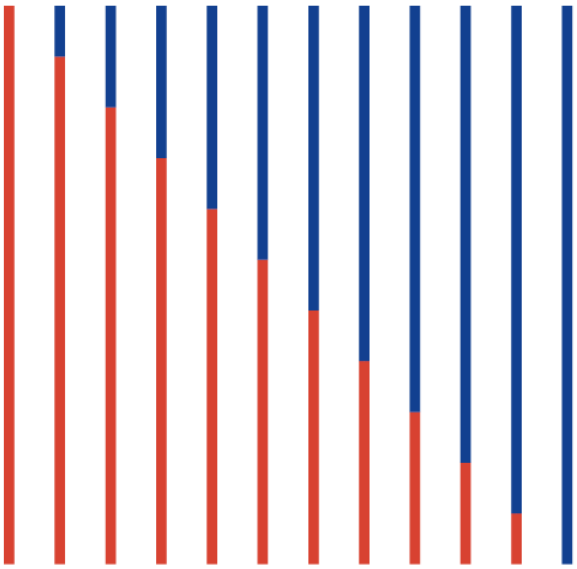


MARS 2025

IA ET SUPERVISION : ENSEIGNEMENTS DES EXPERIMENTATIONS DE L'AMF SUR LA POSSIBILITÉ D'UN TRAITEMENT AUTOMATISE DES REPORTINGS REGLEMENTAIRES

Application sur deux *reportings* définis par les règlements SFDR et Taxonomie

AUTORITÉ DES MARCHÉS FINANCIERS

AMF

AUTORITÉ
DES MARCHÉS FINANCIERS
AMF

SYNTHÈSE

Depuis plusieurs années, l'Autorité des marchés financiers (AMF) voit son **champ de supervision s'élargir** avec l'adoption de nouveaux dispositifs législatifs et réglementaires. Gérer cet élargissement dans un contexte budgétaire contraint constitue un défi important, incitant l'AMF à toujours rechercher une plus grande efficacité et à **mettre l'accent sur le développement de traitements automatiques des données** dans le cadre de sa stratégie sur la donnée[1]. L'utilisation de sa plateforme de surveillance ICY[2] a permis à l'AMF de poursuivre ses expérimentations en intelligence artificielle (IA), notamment en traitement du langage naturel (*Natural Language Processing* – NLP) et en traitement d'image, pour automatiser les tâches manuelles dans l'analyse de deux nouveaux *reportings* liés à la finance durable, **renforçant ainsi les capacités de supervision** des équipes.

Les travaux sur différentes publications imposées par la réglementation, telles qu'illustrées dans cette note par les rapports Taxonomie[3] publiés en 2022 et les annexes SFDR[4] en 2023, ont mis en évidence la manière dont **le niveau de standardisation d'un *reporting*** et les **choix techniques de publication** pouvaient affecter la **capacité d'une machine à extraire automatiquement des données**[5].

Ces enseignements sont particulièrement utiles dans un contexte où le législateur européen intègre de plus en plus de dispositions visant à faciliter l'accès à l'information et son traitement automatisé, et en particulier à l'approche de l'entrée en application du règlement visant à établir un point d'accès unique européen pour un accès centralisé aux informations publiées utiles pour les services financiers, les marchés des capitaux et la durabilité (ESAP[6]). Ce règlement prévoit que les documents publiés (dont feront partie les rapports Taxonomie et les annexes SFDR) devront *a minima* être dans un format permettant l'extraction des données et en majorité lisible par la machine[7]. Or, les exigences règlementaires restent à ce jour peu prescriptives dans leurs définitions : seules les images ne seraient pas considérées comme des données pouvant être extraites et certains documents considérés aujourd'hui comme lisible par la machine au sens règlementaire se prêtent mal à une exploitation automatique.

[1] La stratégie Data de l'AMF repose sur trois grands enjeux : faire des données AMF un patrimoine partagé, traiter automatiquement les données, et développer les outils qui extraient la valeur des données. https://www.amf-france.org/fr/actualites-publications/communiques/communiques-de-lamf/lamf-poursuit-sa-strategie-autour-de-la-donnee-avec-louverture-au-public-de-donnees-sur-les-ventes

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

[3] Règlement (UE) 2020/852 : lien

[4] Règlement (UE) 2019/2088 : lien

[5] Lorsqu'un document est lu par un humain, l'attention se porte sur la présentation visuelle, la structure logique, et la signification du contenu. En revanche, une machine lit un document de manière complètement différente. Plutôt que de « comprendre » le contenu, elle analyse les données brutes à travers l'encodage technique sous-jacent. L'encodage d'un document consiste à transformer les informations, comme le texte, les symboles ou les images, en un format que les ordinateurs peuvent comprendre et afficher. Que ce soit pour un format PDF ou XHTML, un encodage approprié assure que le document est non seulement lisible par l'utilisateur, mais aussi exploitable par la machine pour l'extraction de données. Le processus d'encodage est en général transparent pour l'humain et est géré par les applications informatiques qui permettent de créer un document et de l'enregistrer dans un certain format.

[6] Le Règlement (EU) 2023/2859 (lien), aussi nommé ESAP pour « *European Single Access Point* » a pour objectif de permettre la centralisation et l'accessibilité de l'ensemble des documents régulés de la place financière européenne.

[7] Les formats permettant l'extraction de données ne nécessitent pas nécessairement que chaque information d'intérêt soit balisée (ou « surlignée »), tandis que les formats lisibles par la machine sont des formats de fichier structurés de telle manière que des applications logicielles peuvent facilement identifier, reconnaître et extraire des données spécifiques, notamment des indicateurs quantitatifs, chaque énoncé d'un fait et la structure interne de ces données.

En théorie, l'utilisation du format XHTML, combinée au respect de règles strictes telles que les spécifications HTML du *World Wide Web Consortium* (W3C)[8], devrait garantir une exploitation optimale des documents par la machine, facilitant ainsi une extraction de données précise et fiable. Toutefois, les expérimentations menées par l'AMF n'ont pas permis de valider cette hypothèse, non pas parce qu'elle a été infirmée, mais parce que les documents analysés n'étaient pas conformes à ces standards. Les travaux ont porté sur des portions de documents XHTML dont la rédaction n'est pas soumise, réglementairement, à des règles précises, et sur des documents au format PDF, réputés plus complexes à exploiter, mais qui ne le sont pas toujours, en particulier lorsqu'ils sont standardisés. Sur la base des difficultés rencontrées, ces tests ont mis en évidence les obstacles à une exploitation automatisée efficace et soulignent la nécessité de **poursuivre l'harmonisation des formats et normes pour améliorer l'accessibilité humaine et machine, ainsi que l'utilisation des documents au profit du public.**

Les travaux en IA de l'AMF ont aussi souligné **un retard significatif dans les avancées de la recherche en IA sur le domaine financier pour la langue française**, partiellement dû aux coûts techniques d'accès à une base centralisée et requêtable de documents en français. *Un glossaire est disponible page 10 afin de préciser ou rappeler certains concepts et les abréviations utilisées dans ce document.*

[8] Le W3C est un organisme de standardisation à but non lucratif proposant un ensemble de standards, ou spécifications techniques, pour promouvoir la compatibilité des technologies du *World Wide Web* telles que le HTML et le XHTML. La spécification HTML à jour est disponible sur le lien suivant : lien.

INTRODUCTION

Ces dernières années, l'AMF a poursuivi ses travaux en Intelligence Artificielle (IA) pour développer, entre autres, deux outils d'aide à l'analyse des nouveaux *reportings* entrés en application ces années-là en matière de finance durable.

Les solutions proposées ont pour objectif de faire de gagner du temps aux équipes de supervision en automatisant l'extraction d'informations pertinentes, une tâche manuelle souvent longue et fastidieuse. Pour cela, les outils développés offrent deux fonctionnalités principales :

- la centralisation des données pertinentes extraites de nombreux documents, et,
- une interface visuelle permettant leur consultation de façon rapide et aisée.

La performance des systèmes d'IA développés au cœur de ces solutions (désignés ci-après par les termes « modèle » ou « machine ») se mesure par leur capacité à extraire avec précision et fiabilité l'ensemble des données des documents à traiter. Elle est intrinsèquement liée au format et à la qualité des documents exploités.

Les enseignements tirés de ces travaux mettent en évidence le lien entre les performances de ces systèmes d'IA, et donc leur capacité à être utilisés pour alléger les tâches à faible valeur ajoutée, et l'exploitabilité des documents par la machine. Bien qu'illustrées ici par des *reportings* en finance durable, ces enseignements s'appliquent également à de nombreux autres *reportings*.

1. PRÉSENTATION SYNTHÉTIQUE DU PROJET IA MENÉ EN 2022 SUR LES RAPPORTS TAXONOMIE DES ÉMETTEURS NON FINANCIERS

Pour rappel, la taxonomie européenne constitue un système de classification partagé au sein de l'Union européenne. Son objectif est d'identifier les activités économiques considérées comme durables, en particulier sur le plan environnemental. La taxonomie établit également des obligations de *reporting* spécifiques pour les sociétés cotées sur les marchés financiers, qu'il s'agisse de sociétés financières ou non. En 2022 (sur l'exercice 2021), ces entités étaient tenues de divulguer des indicateurs mesurant l'étendue de leurs activités, investissements ou dépenses opérationnelles éligibles à la taxonomie[9].

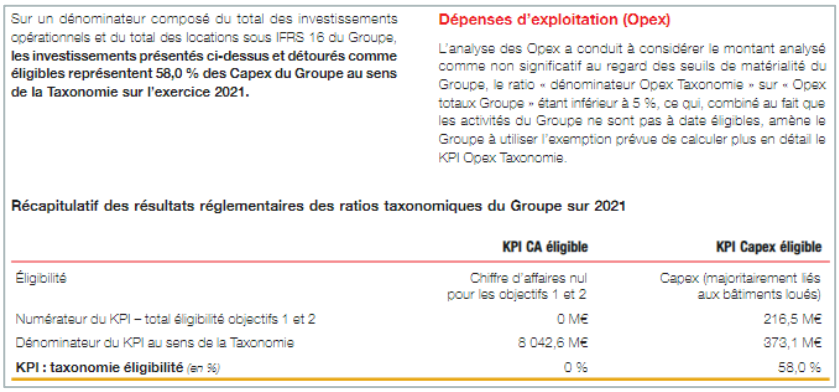
À partir d'un échantillon de 96 rapports publiés en 2022 par des sociétés non financières, l'AMF a conduit la même année un projet visant à construire automatiquement une base de données consolidées et fiables des indicateurs clés de performances (ICP) de dépenses d'investissement (CAPEX), de dépenses opérationnelles (OPEX) et de chiffre d'affaires (CA) éligibles à la taxonomie[10].

L'illustration 1 montre un exemple de la présentation de ces ICP dans un rapport annuel.

[9] Depuis 2023, les obligations ont été étendues aux alignements, c'est-à-dire au respect de critères minimaux de durabilité.

[10] Les indicateurs présents dans les rapports taxonomie des émetteurs financiers différant de ceux des émetteurs non-financiers, l'expérimentation s'est restreinte aux sociétés non-financières éligible à la taxonomie.

Illustration 1 : exemple d'une présentation d'ICP en tableau et en texte



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 %

Étant donné que les rapports 2022 contiennent des informations à extraire tant dans les textes que dans les tableaux, l'AMF a adopté diverses techniques en IA, combinant à la fois du traitement du langage naturel (NLP) et du traitement d'image[11]. Ces approches ont également été complétées par un ensemble de règles gérant notamment la consolidation des mêmes informations extraites à la fois du texte et des tableaux.

Par ailleurs, le modèle a été entraîné pour rechercher à la fois :

- les valeurs des indicateurs clés de performance, qu'elles soient en pourcentage (p.ex. « 10% »), en format numérique (p.ex. « 320 millions ») ou sous forme de quantifieur (p.ex. « totalité ») ; et,
- les informations qualitatives leur étant associées (p.ex. « non-significatif » ou « non-matériel »).

Dans l'exemple donné dans l'Illustration 1 ci-dessus, la machine extrait les informations suivantes :

Tableau 2 : données extraites automatiquement par la machine sur l'exemple de l'illustration 1

Indicateurs clés de performance	Valeur	Information qualitative si présente
CAPEX	58%*	
OPEX	Absence de valeur**	Non-significatif
CA	0%	

*Cette donnée apparaît deux fois : dans le texte et dans le tableau. Le système de règles mis en place permet de gérer ce type de cas.

**Il n'y a pas de chiffre lié à l'OPEX, qui est qualifié de non-significatif par l'émetteur. Dans ce cas la machine ne doit rien renvoyer.

Sur l'ensemble des documents traités dans le cadre de ces travaux, les résultats obtenus sont relativement satisfaisants (voir ci-après le focus sur l'évaluation des performances). Par ailleurs, l'interface visuelle[12] développée dans le cadre de ce projet permet également la vérification des résultats produits par le modèle et, le cas échéant, une correction manuelle, avec un accès direct aux sections spécifiques des rapports traités[13].

[11] Les techniques en traitement d'image ont été utilisées pour extraire les informations des tableaux. Pour plus de détails sur les travaux réalisés en IA sur le rapport Taxonomie, le lecteur est invité à se référer à l'annexe 4.

[12] Voir annexe 1 pour une capture d'écran de l'interface développée.

[13] Une correction des données par un utilisateur génère une mise à jour automatique de la base de données.

L'investissement nécessaire pour améliorer les niveaux de performances a été jugé trop important pour poursuivre le projet jusqu'à un déploiement éventuel. Il est donc important de s'intéresser aux facteurs limitants identifiés au cours de ces travaux et d'en tirer des enseignements. À cet effet, la section 3.2, intitulée « SYNTHÈSE DES ENSEIGNEMENTS TIRÉS DES TRAVAUX AMF » ci-après, présente une analyse de la relation entre les performances observées et la qualité des documents traités, soulignant les enjeux de l'exploitabilité machine.

Focus sur l'évaluation des performances de la solution IA pour les reportings Taxonomie

Deux approches sont possibles pour estimer les performances de la solution :

- l'**évaluation par rapport à la proportion de sociétés cotées :** cette première approche consiste à mesurer la capacité de la machine à extraire correctement les informations recherchées pour chaque émetteur. En d'autres termes, il s'agit de calculer le nombre de rapports pour lesquels la machine a réussi à extraire toutes les informations nécessaires de manière précise. Sur l'échantillon de 96 rapports étudiés, cette approche montre un taux de réussite de 49%, indiquant que la machine a parfaitement traité près de la moitié des documents. De plus, un taux de réussite partielle de 26% est observé, correspondant aux rapports pour lesquels une partie des indicateurs clés de performance a été correctement extraite. Cependant, pour 25% des documents, le prototype n'a pas réussi à extraire correctement au moins une information.

- **l'évaluation par nombre d'indicateurs clés de performance extraits**[14] : la seconde approche évalue la précision de la machine en termes de nombre d'indicateurs clés de performance pour lesquels la valeur correcte a été trouvée. Avec trois indicateurs par rapport, l'échantillon totalise 288 indicateurs à extraire. Dans ce cadre, le modèle affiche une précision moyenne de 70%, ce qui signifie que la machine extrait correctement un indicateur dans 7 cas sur 10. Toutefois, elle échoue à trouver la valeur dans 19 % des cas et renvoie une valeur erronée dans 11 % des cas.

2. PRÉSENTATION SYNTHÉTIQUE DU PROJET IA MENÉ EN 2023 SUR LES ANNEXES SFDR DES FONDS

Entre 2021 et 2023, le règlement sur la publication d'informations en matière de durabilité dans le secteur des services financiers (*Sustainable Finance Disclosure Regulation, SFDR*) est entré en application. Celui-ci exige aux acteurs financiers qui commercialisent ou conseillent des produits financiers dans l'Union européenne une plus grande transparence sur la manière dont ces produits financiers prennent en compte les caractéristiques environnementales ou sociales. Le règlement introduit une classification avec deux niveaux d'engagement en matière de durabilité[15].

À partir des 6 300 prospectus de fonds envoyés à l'AMF tout début 2023 (pour publication ou pour modification), l'AMF a également conduit la même année un projet visant à automatiser :

- la construction d'une base de données consolidées et fiables de certaines données issues des annexes SFDR : la catégorie du fond (article 8 ou article 9 au sens de SFDR[16]), les informations

[14] Dans cette mesure, la bonne extraction de l'information qualitative associée à un ICP n'est pas prise en compte.

[15] https://www.amf-france.org/fr/actualites-publications/positions-ue-de-lamf/proposition-de-criteres-minimaux-environnementaux-pour-les-produits-financiers-des-categories-art9

[16] Les fonds ne disposant pas d'annexe SFDR sont par défaut classé comme article 6 au sens de SFDR.

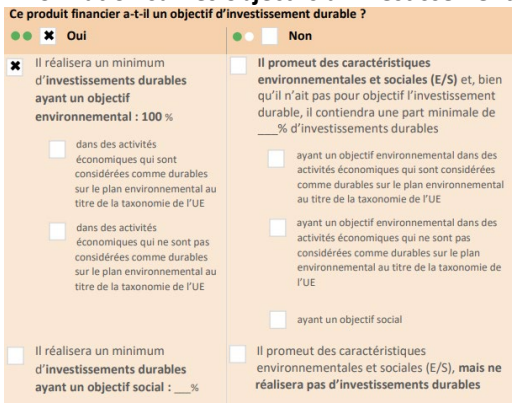
sur les objectifs d'investissement durable, l'allocation des actifs prévue, et, sur l'alignement des investissements durables sur la taxonomie de l'Union européenne) ;

- 11 tests basiques de conformité[17].

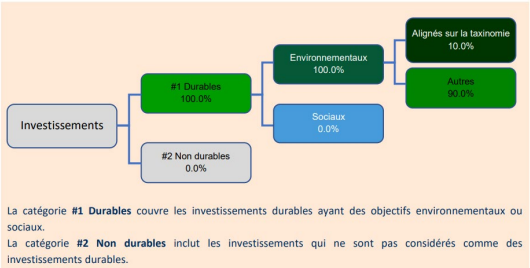
Les annexes SFDR prennent la forme d'un formulaire relativement standardisé[18] avec une liste de questions/réponses qui s'adapte à la catégorie du fond. L'illustration 2 ci-dessous montre un exemple des principales parties de ces annexes qui contiennent les données dont l'AMF cherche à automatiser l'extraction.

Illustration 2 : principales parties de l'annexe SFDR visées dans le projet

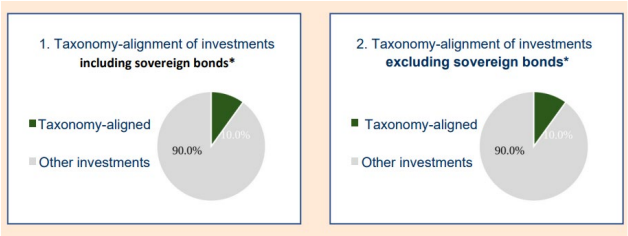
a. Information sur les objectifs d'investissement durable



b. Allocation des actifs



c. Alignement sur la taxonomie EU



Les informations à extraire sont publiées sous diverses formes, tels que des cases à cocher, du texte, et des graphiques, qui peuvent parfois être intégrés en tant qu'images dans les documents. Aussi, à l'instar du projet précédent sur les rapports de Taxonomie, l'AMF a également utilisé diverses techniques en IA telles que le traitement du langage naturel (NLP) et le traitement d'image[19].

Dans l'exemple donné dans l'illustration 2.a ci-dessus sur les objectifs d'investissement durable, la machine extrait les informations suivantes (cf. : Tableau 2) :

[17] Ceux-ci ne seront pas détaillés dans cette note, néanmoins à titre d'exemple, la solution identifie s'il manque une information attendue dans le reporting.

[18] https://www.esma.europa.eu/document/sfdr-templates

[19] Pour plus de détails sur les travaux réalisés en IA sur les annexes SFDR, le lecteur est invité à se référer à l'annexe 4.

Tableau 2 : données extraites automatiquement par la machine sur l'exemple de l'illustration 2

a. Information sur les objectifs d'investissement durable

Informations sur les objectifs		**Données extraites**
Le fond a-t-il un objectif financier ?		Oui
% investissement durable ayant un objectif environnemental	Total	100%
	Considéré comme durable selon la taxonomie UE	Absence d'information*
	Pas considéré comme durable selon la taxonomie UE	Absence d'information*
% investissement durable ayant un objectif social		0%

NB : *aucune case n'est cochée, dans ce cas la machine ne doit rien renvoyer

Dans le reste de l'exemple donné dans l'illustration 2 ci-dessus, la machine extrait l'ensemble des chiffres présents en lien avec l'allocation des actifs et des potentiels alignements avec la taxonomie européenne[20].

Dans le cadre du traitement des documents collectés pour ces travaux, les résultats obtenus sont globalement très satisfaisants, bien qu'un certain type d'information présente de moins bons résultats (voir ci-dessous le focus sur l'évaluation des performances).

Les enseignements tirés de la qualité des résultats observée en fonction du format des informations à extraire dans les annexes SFDR ont été intégrés à l'analyse proposée dans la section 3.2, intitulée « SYNTHÈSE DES ENSEIGNEMENTS TIRÉS DES TRAVAUX AMF ».

Focus sur l'évaluation des performances de la solution IA pour SFDR

Les performances de la solution développée en fonction des informations à extraire :

- Classer un produit financier en fonction de ses objectifs en matière de durabilité (Article 6, 8 ou 9 au sens du règlement SFDR) : 95%
- Extraire automatiquement les réponses apportées à un certain nombre de questions sur le produit, en particulier sur ses :
  - objectifs d'investissement durable : 81%
  - allocations d'actifs : 80%
  - alignements à la taxonomie verte européenne : 30%

3. CONCLUSION D'ENSEMBLE SUR LES FORMATS IMPOSÉS PAR LA RÉGLEMENTATION EUROPEENNE

[20] Compte tenu du nombre de cas à détailler, l'interface de la solution propose une restitution des résultats sous plusieurs onglets et tableaux qui ne sont pas repris ici pour des questions de lisibilité.

LES NOTIONS RÉGLEMENTAIRES SUR LES FORMATS PERMETTANT L'EXTRACTION DE DONNEES OU LISIBLES PAR LA MACHINE

Afin de permettre un accès facilité aux données et une meilleure transparence de l'information, la question de l'exploitabilité des documents par la machine occupe une place centrale dans les réglementations européennes ces derniers années. La directive sur les données ouvertes et la réutilisation des informations du secteur public (*Open Data* [21]) est la première à avoir proposé une définition en 2020 : un format lisible par la machine est « *un format de fichier structuré de telle manière que des applications logicielles puissent facilement identifier, reconnaître et extraire des données spécifiques, notamment chaque énoncé d'un fait et sa structure interne* ». En pratique, ce sont par exemple les formats XML, XBRL, CSV ou JSON, qui ont en point commun la capacité à présenter les informations de manière organisée et hiérarchisée (par exemple via l'utilisation de balises pour XML et XBRL). Cela permet non seulement d'automatiser le traitement des données, mais aussi de les échanger de manière fiable entre différents systèmes.

En 2023, c'est dans le texte du règlement ESAP (système électronique centralisé des documents régulés) que le concept de **format permettant l'extraction de données**[22] est introduit pour la première fois comme un « *format permettant l'extraction de données […] par une machine et qui n'est pas seulement lisible par l'être humain* ». Contrairement au concept de **format lisible par la machine**, cette deuxième définition est bien plus permissive et accepte, en état, le format PDF[23] qui peut s'avérer très compliqué à exploiter (sauf dans certains cas lorsque le contenu des documents en question est standardisé).

ESAP constitue une étape majeure dans le cadre des avancées réglementaires pour la promotion de l'exploitabilité par la machine, non seulement parce qu'il impose que les données visées dans les 37 actes législatifs[24] entrant dans son champ d'application soient publiées dans un format lisible par la machine ou permettant l'extraction de données[25], mais également parce qu'il a pour mandat de donner une liste indicative de ces formats et de leurs caractéristiques (dans les textes d'application du règlement[26]). Par ailleurs, les standards prévus dans les règlements/directives sectoriels permettront de préciser, au cas par cas, les exigences en matière d'exploitabilité par la machine et les formats acceptés.

Avant ESAP, le règlement ESEF avait déjà œuvré en faveur de l'exploitabilité par la machine avec la publication des états financiers consolidés IFRS dans un format lisible par la machine. Néanmoins, si ESEF prévoit la publication des rapports financiers annuels (RFA) des émetteurs en format XHTML et le balisage des états financiers consolidés IFRS selon des spécifications iXBRL[27], ces exigences se concentrent principalement sur les données comptables. En l'absence de dispositions spécifiques couvrant l'ensemble des informations des RFA, seules les données pour lesquelles un balisage est imposé sont réellement exploitables par les machines à ce jour.

Dans le reste du document, le terme « balise » désigne par défaut les balises XHTML. Toute référence aux balises iXBRL sera explicitement précisée lorsqu'elle s'applique.

[21] Directive (UE) 2019/1024, Article 2, Alinéa 13 : lien
[22] Règlement (EU) 2023/2859, Article 2 Alinéa (3) : lien
[23] Tant que celui-ci n'est pas composé d'une image, par exemple lorsqu'il est scanné.
[24] 21 règlements et 16 directives
[25] Règlement (EU) 2023/2859, Article 5
[26] Une validation des textes de niveau 2 par la Commission Européenne est attendue d'ici la fin 2024.
[27] Qui incluent à la fois une taxonomie de base et une taxonomie dite « d'extension » pour permettre une certaine flexibilité.

Focus sur la différence entre les balises XHTML et iXBRL

Les balises XHTML et iXBRL sont utilisées dans des contextes distincts, bien qu'elles soient toutes deux construites sur XML : XHTML structure le contenu du document, tandis qu'iXBRL marque les données financières pour qu'elles soient directement accessibles par la machine.

Par exemple, les émetteurs soumis à ESEF doivent publier leurs RFA en XHTML. Ce format impose l'utilisation de balises XHTML, qui permettent à la machine de distinguer dans le contenu les titres des sections, les paragraphes de texte, les tableaux, etc.

Exemple d'une balise XHTML qui définit un titre : <h1>Rapport financier annuel 2023</h1>

Dans le cadre d'ESEF, iXBRL est utilisé pour marquer les éléments financiers spécifiques (comme les chiffres d'affaires, les bénéfices, etc.). Grâce à ces balises spécifiques, il est possible pour la machine d'extraire toutes les informations balisées.

Exemple d'une balise iXBRL qui indique le LEI pour identifier l'émetteur :
<xbrli:entity>
<xbrli:identifier scheme="http://standards.iso.org/iso/17442">KGCEPHLVVKVRZYO1T647</xbrli:identifier>
<xbrli:entity>[28]

Sans elles, il faudrait soit collecter ces informations à la main une par une, soit entraîner une IA à extraire les informations dans le texte (plus coûteux et avec des risques d'erreurs).

3.1. SYNTHÈSE DES ENSEIGNEMENTS TIRÉS DES TRAVAUX AMF

Dans le cadre des travaux IA présentés dans les deux premières parties de cette note, il est apparu de façon très notable que les performances des outils développés sont intrinsèquement liées au format que les récents textes réglementaires essaient d'encadrer mais, aussi et surtout, à la qualité, à l'encodage et au niveau de standardisation des documents traités.

Le tableau ci-dessous récapitule les enseignements tirés de ces travaux, soulignant les difficultés rencontrées en lien avec les enjeux d'exploitabilité machine. Ceux-ci sont enrichis des conclusions de même nature issues des travaux réalisés pour extraire, via des outils de type Data plus traditionnels, les informations balisées dans les états financiers consolidés IFRS tel que prévu par ESEF.

Tableau 3 : enseignements tirés dans les travaux IA sur les rapports Taxonomie & SFDR

Réglementation	Document concerné	Exigences réglementaires sur la lisibilité par la machine	Format des documents exploités	Enseignements tirés

[28] Exemple tiré de https://www.amf-france.org/sites/institutionnel/files/private/2022-11/ESMA%20ESEF%20traduction%20Reporting%20Manual_correction%20nov%2022%20FR.pdf

ESEF	États financiers consolidés IFRS dans les RFA/DEU	Format lisible par la machine attendu avec un balisage des données	XHTML et balises iXBRL	- Extraction facilitée par le format et la bonne utilisation des balises - Néanmoins, du fait de la flexibilité offerte par la taxonomie d'extension, consolidation des données coûteuse car nécessite que les experts métier fournissent toutes les règles de mise en cohérence et de calcul (ex : par exemple pour le calcul de la dette nette)
Taxonomie verte EU (en 2022)	Rapport Taxonomie dans les RFA/DEU	Aucune	XHTML[29]	Extraction rendue difficile et coûteuse par : - des soucis d'encodage liés à une mauvaise utilisation des balises : ▪ incapacité à distinguer la structure du document par le balisage XHTML, ▪ absence de balises XHTML pour isoler les tableaux des paragraphes de texte ou pour extraire les informations des tableaux ; - une absence de standardisation dans la présentation du contenu avec la présence de tableaux très hétérogènes.
SFDR	Annexe SFDR dans les prospectus des fonds	Pas d'exigence à proprement parler mais un formulaire relativement standardisé (structure du document imposée par exemple)	PDF	Une extraction relativement facilitée par le formulaire standardisé au niveau de son contenu mais limitée par : - l'absence de signets simplifiant la navigation dans les sections des documents ; - la nécessite de traiter des graphiques et des images sans pouvoir s'appuyer sur le texte ou un tableau ; - des soucis d'encodage dus à l'absence de standardisation technique.

Lorsqu'un document est lu par un humain, l'attention se porte sur la présentation visuelle, la structure logique et la signification du contenu. En revanche, une machine lit un document de manière complètement différente. Plutôt que de « comprendre » le contenu, elle analyse les données brutes à travers l'encodage technique sous-jacent.

L'encodage d'un document consiste à transformer les informations, comme le texte, les symboles ou les images, en un format que les ordinateurs peuvent comprendre et afficher. Que ce soit pour un format

[29] Tous les documents des émetteurs sélectionnés pour le projet sur les rapports Taxonomie étaient publiés en format XHTML. En 2022 ces rapports étaient globalement mal formatés de sorte qu'il aurait été probablement plus simple de traiter des rapports en format PDF.

PDF ou XHTML, un encodage approprié assure que le document est non seulement lisible par l'utilisateur, mais aussi exploitable par la machine pour l'extraction de données. Le processus d'encodage est en général transparent pour l'humain et est géré par les applications informatiques qui permettent de créer un document et de l'enregistrer dans un certain format.

Deux éléments affectent notamment la qualité de l'encodage :

- les choix réalisés au moment de la rédaction du document (par exemple insérer une image pour représenter un tableau plutôt que de le créer dans le document) ;
- l'application utilisée pour enregistrer ou convertir un document dans un certain format (par exemple les outils utilisés par les émetteurs pour publier leur RFA/URD en XHTML ne permettent pas d'obtenir une qualité optimale, d'un point de vue encodage).

Afin d'améliorer la qualité de l'encodage d'un document pour en faciliter sa lecture par une machine, il est donc nécessaire de standardiser techniquement la rédaction des documents, de sorte à la fois d'uniformiser les choix humains dans la construction du contenu (par exemple bannir les images pour représenter un tableau), et de préciser les spécifications techniques auxquelles les outils doivent se référer pour publier des documents exploitables par la machine (par exemple le suivi des spécifications HTML du W3C[30] pour la construction des documents en XHTML).

Le lecteur est invité à se référer à l'annexe 3 pour plus de détails sur les difficultés rencontrées avec les formats et l'encodage des documents exploités pour les projets sur les rapports Taxonomie, et les annexes SFDR.

Il est à noter que, dans les deux expérimentations réalisées, une des principales difficultés rencontrées est la capacité à naviguer dans la structure des documents, qui est pourtant primordiale pour que la machine puisse extraire les informations pertinentes dans le bon contexte. Si la structure du document est mal définie ou complexe, cela peut entraîner des erreurs d'interprétation pour lesquelles des données importantes sont mal associées ou complètement ignorées. Ce problème se rencontre particulièrement lorsque les documents sont denses ou comportent de nombreuses sections imbriquées.
Les expérimentations menées par l'AMF ont confirmé ces difficultés. Si la standardisation du contenu, comme celle appliquée aux annexes SFDR, contribue à réduire les erreurs, même dans un format PDF, elle se révèle insuffisante pour assurer une exploitation automatisée optimale. De plus, les tests réalisés sur les rapports Taxonomie ont mis en évidence la nécessité d'imposer des règles strictes d'utilisation, même pour des formats considérés comme lisibles par la machine tels que le XHTML. Ces travaux montrent que, sans un balisage rigoureux, le potentiel du format XHTML reste limité et ne permet pas d'atteindre pleinement les objectifs d'exploitabilité automatisée. Ainsi, bien que les expérimentations n'aient pas permis de le vérifier en pratique, l'utilisation méthodique de balises dans un format XHTML semble être la meilleure solution pour minimiser les risques d'erreurs et garantir une extraction de données fiable et précise.

3.2. PISTES D'AMÉLIORATION POUR LES RÉGLEMENTATIONS EXISTANTES ET À VENIR

Dans le cadre des travaux IA réalisés sur les rapports Taxonomie, il apparaît que la majeure partie des difficultés énoncées dans la partie précédente viennent de l'absence de spécifications techniques XHTML pour guider la construction des RFA/DEU. En effet, à la différence de la *Securities & Exchange*

[30] Normes HTML du W3C : lien

Commission (SEC)[31] aux Etats-Unis et son système de collecte, d'analyse et de récupération de données électroniques (*Electronic Data-Gathering, Analysis and Retrieval* - EDGAR), la Commission européenne n'a pas demandé aux émetteurs de suivre les normes XHTML développés par W3C.

Cependant, en 2023, deux évènements postérieurs aux travaux IA de l'AMF sur ce sujet sont venus améliorer l'exploitabilité des sections Taxonomie dans les RFA/DEU :

- la standardisation des tableaux du rapport Taxonomie ;
- la ligne directrice 2.2.6 du Reporting Manual ESEF[32] de l'ESMA, qui est venue préciser les attentes concernant l'utilisation des balises sémantiques XHTML[33].

Ces nouvelles normes pourraient suffire pour construire une base de données « fiable » des informations issues du rapport Taxonomie en attendant leur passage réglementaire en format lisible par la machine[34].

Toutefois, cela n'est pas suffisant pour permettre l'exploitabilité des RFA/URD dans leur globalité (et notamment pour le rapport Taxonomie). Pour cela, il conviendrait de proposer le suivi des normes W3C dans la rédaction des RFA/DEU (en particulier l'usage des balises sémantiques pour identifier les titres, les paragraphes et les tableaux).

Par ailleurs, les travaux IA réalisés sur les annexes SFDR ont mis en avant les bénéfices d'un *reporting* standardisé, en particulier parce que la standardisation fiabilise autant les résultats qu'elle permet de réduire les coûts de développement. Néanmoins, celle-ci est limitée dans le cas des annexes SFDR, et devrait être renforcée pour améliorer leur exploitabilité par la machine :

- par la double-publication des données importantes contenues dans une image, sous forme d'un paragraphe de texte ou un tableau[35];
- par une standardisation technique à suivre pour remplir le formulaire (cf. les exemples donnés en annexe 3 : problèmes de formats et encodage) ;
- par l'ajout de signets facilitant la navigation dans les sections des documents.

Plus largement, les pistes d'amélioration identifiées précédemment sont pertinentes pour l'ensemble des autres réglementations prévoyant des *reportings* de données non structurées, dont l'exploitation nécessiterait une automatisation. Aussi, bien que l'utilisation systématique de formats effectivement lisibles par la machine ne soit pas toujours justifiée au regard du rapport coût/bénéfice, il est important d'anticiper les besoins éventuels en matière de collecte de données automatisée et de définir les règles de reporting en conséquence.

À l'issue de ses travaux, l'AMF souligne cinq enseignements principaux pour faciliter l'exploitation par la machine sans dégrader la lisibilité humaine dans les réglementations à venir :

- **l'utilisation d'un format lisible par la machine à lui seul n'est pas suffisant et doit être accompagné de règles spécifiques pour atteindre un niveau d'exploitabilité optimal ;**
- **l'obligation d'associer à toute image contenant des données à traiter par une machine, un texte ou tableau descriptif contenant ces mêmes informations ;**
- **la standardisation de la structure du *reporting* et des informations qu'il contient ;**

[31] Filer Manual for 10-K filings, section 5.2.2 : lien
[32] Reporting Manual ESEF, Guidance 2.2.6, page 30 : lien
[33] La *guidance* vise particulièrement les parties concernées par ESEF, néanmoins cela a permis d'augmenter globalement l'usage des balises pour l'entièreté des rapports.
[34] https://www.esma.europa.eu/issuer-disclosure/electronic-reporting
[35] Concernant l'exploitabilité des graphiques, les plus récents modèles d'IA sont encore loin des capacités humaines de lecture de ce type de données. L'implication est que même avec des moyens conséquents, extraire et structurer automatiquement les données présentes dans des graphiques n'est pas envisageable avec des performances satisfaisantes. Lien

- **la standardisation technique des documents ;**
- **la généralisation du format XHTML avec le suivi des normes W3C, et ce même pour les prochaines réglementations qui ne prévoient pas l'intégration de balises iXBRL dans un format lisible par la machine.**

***

AUTORITÉ
DES MARCHÉS FINANCIERS
AMF

GLOSSAIRE

Système d'IA : un système reposant sur une machine, conçu pour fonctionner avec différents niveaux d'autonomie, qui peut faire preuve d'adaptabilité après son déploiement, et qui, pour des objectifs explicites ou implicites, déduit, à partir des données qu'il reçoit, comment générer des résultats tels que des prédictions, du contenu, des recommandations ou des décisions qui peuvent influencer des environnements physiques ou virtuels (Article 3, AI Act).
HTML : langage de balisage utilisé pour la création de pages web et de documents, permettant notamment de définir des liens hypertextes.
Spécifications HTML : spécifications techniques proposées, par exemple, par le *World Wide Web Consortium* (W3C) afin de donner les règles d'usage du langage HTML pour la constitution de pages web afin que celles-ci qu'ils soient homogènes et plus aisément lisibles par une machine.
XBRL (eXtensible Business Reporting Language) est un langage informatique basé sur XML qui a été conçu expressément pour l'automatisation des exigences d'information des entreprises, telles que la préparation, le partage et l'analyse des rapports financiers, des états.
Inline-XBRL : iXBRL, ou Inline XBRL, est un standard ouvert permettant à un document unique de fournir à la fois des données lisibles par l'homme et des données structurées lisibles par une machine.
Corpus : ensemble fini de textes choisi comme base d'une étude.
Format lisible par la machine *(Machine-readable)* : un format de fichier structuré de telle manière que des applications logicielles puissent facilement identifier, reconnaître et extraire des données spécifiques, notamment chaque énoncé d'un fait et sa structure interne.
Format permettant l'extraction de données (*Data Extractable)* : un format de fichier électronique ouvert indépendant de la plateforme et mis à disposition du public sans aucune restriction empêchant la réutilisation des documents. Le format est largement utilisé ou requis par la loi, permet l'extraction de données par une machine et n'est pas seulement lisible par l'homme.
NLP : *Natural Language Processing*, ou traitement automatique du langage naturel en français, est un domaine multidisciplinaire impliquant la linguistique, l'informatique et l'intelligence artificielle. Il vise à créer des outils de capable d'interpréter et de synthétiser du texte pour diverses applications.

AUTORITÉ
DES MARCHÉS FINANCIERS

AMF

ANNEXE 1 : INTERFACE PERMETTANT LA CONSULTATION DES RESULTATS DE L'EXPERIMENTATION SUR LES RAPPORTS TAXONOMIE

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

Enregister modification

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%

‹

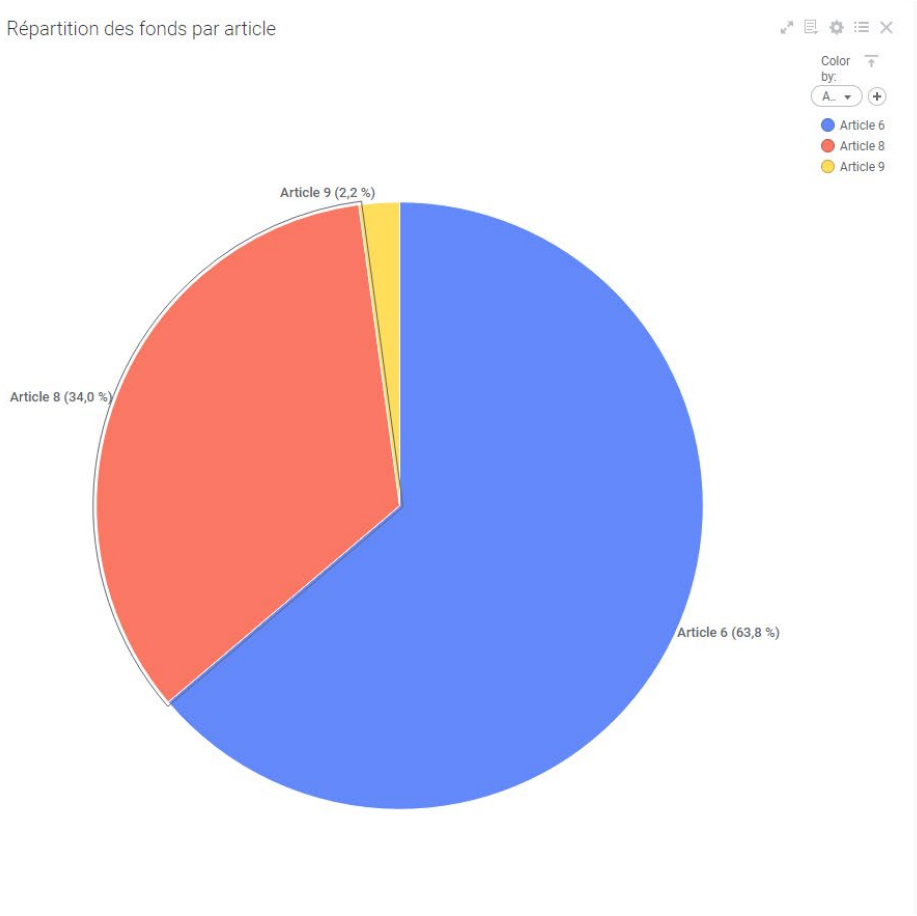
Indicateurs au 31/12/2021	Chiffre d'affaires (CA) éligible	Dépenses d'investissements (CAPEX) éligibles	Dépenses d'exploitation (OPEX) éligibles
Numérateur (éligibilité)	54 028 K€	44 034 K€	1 858 K€
Dénominateur	689 492 K€	68 460 K€	22 076 K€
Indicateur de performance (ratio) exprimé en %	7,84 %	64 %	8,3 %

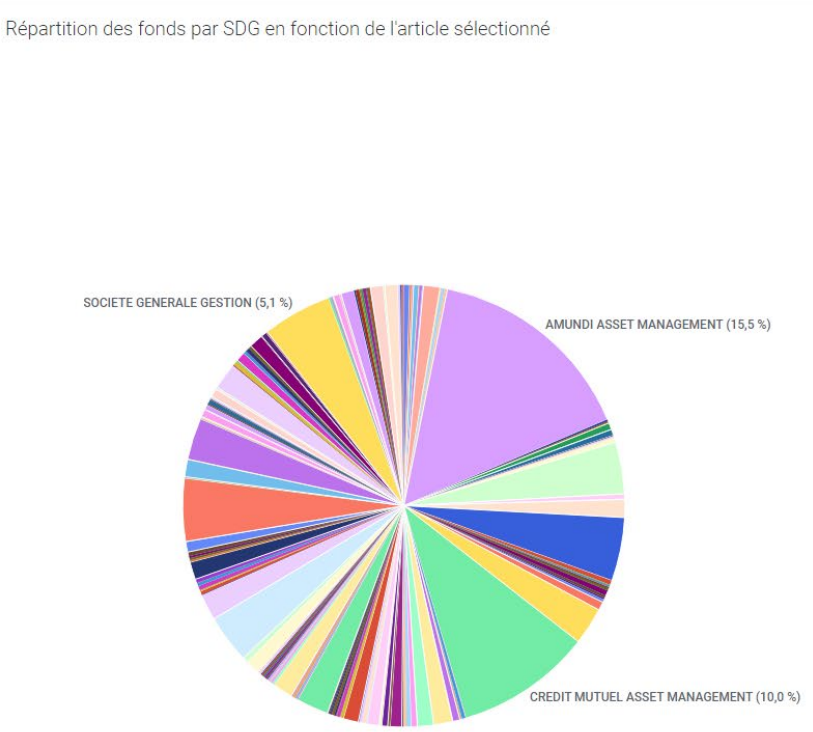
›

AUTORITÉ
DES MARCHÉS FINANCIERS

AMF

ANNEXE 2 : INTERFACE DE L'OUTIL DEPLOYE PERMETTANT L'EXPLORATION DES ANNEXES SFDR





AUTORITÉ
DES MARCHÉS FINANCIERS

AMF

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

ANNEXE 3 : PROBLÈMES DE FORMATS ET ENCODAGE

L'aspect visuel, qui est directement compréhensible pour un être humain, peut être encodé de diverses manière selon le format choisi (Word, PDF, XHTML, …). Ces différences s'observent aussi pour un format donné, tel que le XHTML, pour lequel la méthode de construction choisie par un rédacteur peut différer de celle d'un autre[36]. L'illustration 3 montre comment une partie du tableau du document de l'illustration 1 est présenté sous forme XHTML. L'encodage de la table située au bas de l'illustration 3 n'utilise pas les balises adéquates[37] permettant de savoir que le contenu se situe dans un tableau avec des lignes et des colonnes. L'ordre dans lequel les éléments apparaissent n'est pas clair et ne correspond pas à l'ordre visuel, le mot « Taxonomie » y est découpé en 4 parties séparées en plusieurs balises (« T », « a », « xonom », et « ie »).[38]

Afin que les informations contenues dans cette table soit considérées comme permettant l'extraction de données, la table aurait du être contenue dans des balises « table », les en-têtes dans des balises « thead », chaque ligne dans une balise « tbody » et enfin chaque colonne appartenant à une ligne dans des balises « tr ». Rendre ces éléments *lisible par la machine* aurait nécessité la création d'une taxonomie XBRL propre au règlement Taxonomie afin d'assigner une étiquette à chaque valeur présente dans le tableau.

Illustration 3 : partie du contenu du tableau précédent en xhtml

Les illustrations 4 et 5 ci-après montrent un exemple de document XHTML correctement construit : on y trouve un tableau des CAPEX Taxonomie publié en 2023 par un émetteur (illustration 4), et une partie de son code XHTML (illustration 5). *A contrario* de l'illustration 3 ci-dessus, la table de l'illustration 4 est correctement construite en utilisant les balises « sémantiques », comme spécifiées dans les normes du W3C. Celle-ci est donc automatiquement détectable par une machine et transformable en une table de données exploitable par tout type d'outil de traitement de données structurées.

[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 du 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 lisible par la machine à certaines totalement illisibles.

Par exemple, il est instantanément possible d'isoler la ligne « TOTAL A.1 + A.2 » (carré rouge) et la colonne « % de CAPEX » (carré vert) pour extraire la valeur de l'ICP, soit « 19,10 % » (croisement entre les deux).

Illustration 4 : Exemple de tableau normé relatif aux CAPEX et publié en 2023

Activités économiques	Code	CAPEX Absolu Euros	% de CAPEX %
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%

Illustration 5 : Code XHTML relatif au tableau de l'illustration 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_bis">
        <p class="eolng_tab_total_resserre"> == $0
          <span class="eolng_approche-25">Total  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_bis">…</td>
      <td class="border_rule_column border_rule_column_5 border_rule_column_end_5 eolng_base_c2_resserre_bis">
        <p class="eolng_tab_total_r_resserre">19,10%</p>
      </td>
```

Enfin pour les PDF, et en particulier les annexes SFDR, l'extractabilité des données n'est pas nécessairement possible à moindre coût et sans erreur.

L'illustration 2.a montre un formulaire pour lequel il a fallu identifier les cases cochées et les pourcentages automatiquement. L'approche la plus efficiente et la plus utilisée afin de traiter des documents PDF est de les convertir dans un format texte[39]. Or la façon dont est rempli le formulaire peut être différente selon la société de gestion ayant édité l'annexe, jusqu'à empêcher le convertisseur d'extraire l'ensemble des informations.

L'illustration 6 ci-après illustre les problèmes issus de l'hétérogénéité des méthodes de complétion du formulaire. Elle montre qu'après la conversion de l'annexe SFDR dont est issue l'illustration 3, les cases cochées et non-cochées ont disparu, impliquant que celles-ci étaient présentes sous forme d'image ou de dessin[40], et pas en caractère « ☑ » ou « ☐ ». L'AMF a aussi remarqué que certaines sociétés de gestion utilisaient le caractère « X », ou n'importe quel caractère qu'une police personnalisée transforme en case cochée ou en croix.

Illustration 6 : exemple de conversion en texte d'une partie de l'illustration 2.a

Ce produit financier a-t-il un objectif d'investissement durable ?
Oui
Non
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 : ___%

[39] Pour cela des outils de conversion en source ouverte existent et offrent des performances relativement satisfaisantes. L'AMF a par exemple utilisé PyMuPDF et Unstructured.
[40] Les dessins sont des objets spécifiques aux documents PDF que les rédacteurs peuvent ajouter afin de créer des formes telles que des cercles et des rectangles.

ANNEXE 4 : TRAITEMENTS AUTOMATIQUES

TAXONOMIE

L'outil d'extraction des informations relatives à la Taxonomie verte européenne dans les rapports 2022 des entreprises non-financières est basé sur une succession de modules comprenant des SAI. Cette séquence de briques permet de structurer l'information dans les DEU, identifier la section pertinente, extraire les indicateurs clés de performance présents dans les paragraphes ou dans les tableaux, d'agréger les résultats et enfin de procéder à des vérifications et des inférences.

Afin d'aider l'AMF à gagner du temps dans l'extraction des informations relatives à la Taxonomie, l'outil doit être en mesure :

- ☐ d'extraire les parts d'éligibilité des CAPEX, OPEX et CA dans chacun des documents ;
- ☐ d'identifier une éventuelle référence à une clause d'exemption de matérialité ou à la non-éligibilité pour un, ou plusieurs de ces ICP ; et,
- ☐ de référencer les paragraphes et/ou les tableaux relatifs aux informations liées à la Taxonomie.

Les DEU sont des documents très denses (plusieurs centaines de pages) contenant diverses sections dont celle(s) sur les informations relatives à la taxonomie. Les documents traités dans le cadre de l'étude sont publiés en XHTML, un format qui aurait dû permettre leur exploitation automatique grâce à l'usage d'un système de balises pour notamment définir la structure du contenu (titre, section, sous-section, etc.) mais également le référencement de certaines informations spécifiques. Toutefois, compte tenu de l'absence de spécifications sur l'usage des balises, de nombreux développements supplémentaires ont dû être menés. Ces développements ont dû être réalisés avant le développement des modules spécifiques à l'extraction de contenu en lien avec la thématique taxonomie, et ont donné lieu au développement de briques techniques (réutilisables au-delà de cette étude) permettant à la machine de naviguer dans les DEU[41]. Il est à noter que certains de ces développements n'auraient pas été nécessaires si les documents avaient encore été publiés en PDF, un format *pourtant* moins favorable à l'extraction de données.

Une fois que l'outil est en mesure d'isoler les portions d'un document relatives au rapport Taxonomie, il doit ensuite extraire les ICP (et narratifs associés) des contenus de ces parties. Deux situations sont alors à distinguer : soit l'émetteur a exclusivement publié ses ICP dans des paragraphes texte, auquel cas l'outil « se contente » d'utiliser des techniques du traitement de texte, soit tout ou partie des informations recherchées sont présentées dans un tableau. Dans ce deuxième cas de figure, l'étude a montré que le traitement d'image était l'approche donnant les meilleurs résultats pour extraire les ICP des tableaux. Par ailleurs, il est à noter que dans la majorité des documents, les deux configurations sont combinées. Aussi, l'outil doit pour la plupart des DEU à la fois traiter le texte et le tableau, avant de proposer une consolidation finale des résultats, notamment pour gérer les cas où le niveau de granularité des ICP[42] ne serait pas le même entre les deux.

L'approche développée pour le traitement du texte, dans le cas de l'extraction des informations relatives à la Taxonomie, repose sur un ensemble de techniques s'appuyant entre autres sur de l'apprentissage non-supervisé et supervisé[43]:

[41] L'ensemble des briques développées dans le cadre de l'étude sont présentées dans les annexes II.
[42] Alors que les ICP doivent être présentés au niveau du groupe ceux-ci peuvent être décomposés par entité ou par activité / groupe d'activités.
[43] Pour rappel, l'apprentissage non supervisé est une branche du *machine learning*, caractérisée par l'analyse et le regroupement de données non-étiquetées alors que l'apprentissage supervisé utilise des données étiquetées afin d'apprendre à prédire ces étiquettes.

- la reconnaissance d'entités nommées, qui a pour tâche d'identifier (et d'extraire) les mentions faites aux ICP dans le texte (tels que « CAPEX », « dépenses d'investissement » ou « revenu ») ainsi que leurs valeurs quantitatives (comme par exemple « 35% », « 249 millions » ou encore « nulle »), aux différentes activités (« activités d'extraction » ou « activités de production d'énergie » pour n'en citer que deux), ou, encore aux organisations (tels que « le Groupe » ou « ses filiales ») ;
- la résolution d'entités, qui vise à déduire pour chaque mention d'une entité dans un texte (telles que « dépenses d'investissement» ou « dépenses opérationnelles») à quel indicateur l'émetteur fait très exactement référence[44]. Autrement dit, l'algorithme apprend à différencier les subtilités dans le texte, comme à faire la distinction entre une référence à l'OPEX tel que défini par la Taxonomie ou à l'OPEX tel que défini par les normes IFRS. Cette étape permet également à l'outil de faire la différence entre le sens des mentions « éligibilité » et « alignement » des activités, ou encore, à identifier si le texte fait référence aux activités d'une filiale en particulier ou de l'ensemble du groupe de l'émetteur.
- Par exemple, dans l'extrait suivant:

« *Le montant des OPEX au sens du Règlement Taxonomie représente moins de 3% du total des dépenses d'exploitation du Groupe sur l'exercice 2021 et n'est pas considéré comme significatif.* »

Les simples mentions « OPEX » (1) et « dépenses d'exploitation » (2) ne permettent pas de conclure sur le type d'OPEX auquel l'émetteur fait référence, mais la résolution d'entités permet ici à l'outil de comprendre automatiquement que :

- (1) est l'OPEX tel que défini par la Taxonomie (l'ICP que l'outil cherche à extraire) ;
- (2) est l'OPEX IFRS du groupe (sur lequel la part de l'éligibilité est calculée).

Grâce à la perception de ces subtilités, le système est en mesure de déduire que le narratif de l'émetteur justifiant que ses OPEX Taxonomie ne sont pas significatifs est correct. Il est aussi possible de déduire que ces narratifs sont appliqués à l'ensemble du groupe et pas uniquement à une filiale.

- la détection d'attributs, qui vise à identifier des caractéristiques spécifiques à certains types d'entités, par exemple si un ICP est caractérisé comme « Non calculé », « Non significatif », « Non matériel » ou « Non éligible » par l'émetteur.
- l'extraction de relations qui permet de lier les diverses mentions des entités entre elles et, par exemple, faire correspondre un ICP avec le bon montant ou pourcentage correspondant dans le paragraphe, l'activité avec le montant, le pourcentage, ou l'organisation (émetteur ou filiale) qui lui correspond.

L'approche développée pour le traitement des tableaux est composée de trois grandes étapes[45] :

- l'outil doit d'abord être capable d'identifier la présence éventuelle de tableaux. Pour cela, le système transforme en image la section « Taxonomie » extraite précédemment, et utilise un algorithme spécifique de détection de tableaux, déjà pré-entraîné[46]. Une nouvelle image est produite pour chaque tableau identifié ;
- dans un second temps, l'outil doit appréhender la structure de chaque tableau pour détecter où les informations à extraire se trouve. Autrement dit, à partir de l'image de chaque tableau identifié à l'étape précédente, un ensemble d'algorithmes et de règles sont utilisés pour détecter les contours des cellules ;

[44] Ces entités réelles sont, dans le cadre de l'expérimentation, limitées aux types suivants : « CAPEX éligible », « CAPEX tel que défini par les normes IFRS », « OPEX éligible », « OPEX tel que défini par la Taxonomie », « OPEX tel que défini par les normes IFRS », « Revenu éligible », « Revenu tel que défini par les normes IFRS », « Activités spécifiques » et « Organisation ».

[45] Les quelques rares DEU ayant correctement défini leurs tableaux via l'usage des balises spécifiques du format XHTML n'ont pas nécessité de traitement d'image.

[46] La détection se base sur le *framework* Paddle (lien) – le modèle est décrit dans ce papier : lien

- la dernière étape consiste à extraire les informations. Une fois les contours de la structure bien identifiés, le tableau est converti en une table de données structurées[47]. Puis, un ensemble de règles identifient la cellule exacte contenant la valeur de chacun des ICP à partir des noms de colonnes, de lignes, du format du tableau et du contenu de la cellule[48].

SFDR

L'outil d'aide à la supervision des annexes SFDR dans les prospectus de fonds consiste en une succession de modules comprenant des SIA. Cette séquence de briques permet de structurer l'information dans les annexes SFDR : localiser les annexes et les couples question-réponse s'il y a, identifier l'article auquel est soumise l'annexe, extraire certaines informations et procéder à un ensemble de vérifications automatiques.

Afin d'aider l'AMF à gagner du temps dans la supervision des annexes SFDR, l'outil doit être en mesure :

- de déduire l'article correspondant au prospectus (article 6, 8 ou 9) ;
- de localiser, référencer les couples question-réponse des annexes ;
- d'extraire les informations présentes dans le formulaire d'objectifs et les graphiques d'allocation des actifs et d'alignement à la Taxonomie verte européenne.

Les *Prospectus* (Règlement (UE) 2017/1129) sont des documents longs et denses (plusieurs dizaines de pages) contenant diverses sections et annexes, dont celle relative aux informations extra-financières requises par SFDR. Les annexes traitées dans le cadre de l'étude sont publiées en format PDF relativement standardisé (cf. : Annexes II et III du règlement (UE) 2019/2088).

La standardisation du document limite les libertés des rédacteurs en contraignant la forme et le contenu ce qui simplifie la recomposition de la structure du document et le référencement d'informations spécifiques. Cependant le format PDF est un format facilitant la lisibilité humaine plus que l'exploitabilité machine, par exemple : lors de l'exploitation du document par une machine la plupart de la structure du document est perdue. Cette perte de structure impose des développements spécifiques pour prétraiter ces annexes afin qu'elles soient de qualité suffisante pour être traitées par une machine. Les développements nécessaires au prétraitement incluent la conversion du PDF en données interprétables par une machine, le nettoyage de ces données et le référencement de l'ensemble des questions et réponses du document.

L'approche développée pour le prétraitement des annexes SFDR repose sur une séquence de tâches successives qui reposent majoritairement sur des outils en source ouverte, des implémentations d'algorithmes et de l'ingénierie humaine :

- la conversion des fichiers PDF, qui a pour objectif de transformer le contenu lisible par un humain en en données exploitables par une machine. Cette conversion est effectuée par le biais de l'outil PyMuPDF, et permet à la machine d'interagir avec le contenu du PDF ;
- le nettoyage des données, qui permet d'améliorer sensiblement la qualité des données extraites en corrigeant les erreurs de conversion inhérents à ce format de données ;
- le référencement de l'ensemble des questions, qui a pour objectif d'indexer l'ensemble des portions du document. L'identification des questions se fait par à l'aide de la librairie fuzzysearch, dont un

[47] La conversion se base sur le *framework* PaddlOCR - Lien
[48] La recherche académique a récemment développé des approches poussées basées sur des réseaux de neurones, mais celles-ci n'étaient pas à date disponibles en langue française. Pour référence, voir : « TAPAS: Weakly Supervised Table Parsing via Pre-training », J. Herzig et al. - Lien

algorithme implémenté permet pour chaque question telles que définies dans les textes et pour chaque article, de rechercher si et où celles-ci sont présentes dans le document ;

- la déduction de l'ensemble des réponses apportées par la société de gestion aux questions présentes dans le document. L'identification se fait de manière mécanique, en déduisant les coordonnées de la réponse dans le document à partir de celles de la question en cours et de la question suivante dans le document.

Une fois le document prétraité, le processus continue avec quatre briques : un module d'identification de l'article correspondant et trois modules d'extraction des informations contenues dans les questions relatives aux objectifs, à l'allocation des actifs et enfin à l'alignement à la taxonomie. Ces modules exploitent le texte mais aussi la représentation visuelle du document, par exemple en utilisant les coordonnées des caractères, mots et paragraphes ainsi que les images qu'il contient.

Les processus de traitement développés afin d'extraire les informations demandées et qui permettront d'aider à la supervision des annexes SFDR sont décomposés ci-dessous :

- la classification du fonds, qui a pour tâche d'identifier le numéro d'article (6, 8 ou 9) auquel est soumis le prospectus. Celui-ci considère le fonds comme article 6 lorsqu'il n'identifie pas d'annexe SFDR, et article 8 ou 9 selon que le système identifie majoritairement des questions propres aux annexes article 8 ou article 9 ;
- l'analyse des objectifs d'investissement durable, qui a pour but d'extraire les informations des formulaires à cases présentés en toute première page des annexes que celles-ci soient montrées par une case avec une croix ou non, ou par un pourcentage renseigné. Comme nous l'avons montré plusieurs fois dans cette note, une partie du contenu est perdu ou « obfusqué » lors de la conversion du PDF en document exploitable par une machiné. Ce module est composé de plusieurs traitements déterministes : identification des cases, identifications des narratifs, liage des cases aux narratifs correspondants, classification case vide / cochée et enfin l'extraction des pourcentages s'ils sont présents ;
- l'extraction des allocations d'actifs, qui a pour objectif d'extraire les pourcentages d'allocation d'actifs dans chacune des composantes (« #1 Alignés sur les caractéristiques E et S », « #2 Autres », « #1A Durables », …) requises par la classification du fonds, que ceux-ci soient dans le graphique ou dans le texte. L'extraction à partir du graphique se fonde sur des expressions régulières, celle à partir du texte utilise SpaCy pour décomposer la réponse en phrases, et identifier les composantes avec leur pourcentage associé ;
- l'analyse de l'alignement des investissements à la taxonomie de l'UE, qui vise à extraire les pourcentages d'alignement présents dans les graphiques en camembert. Ce module utilise les informations visuelles présentes dans le PDF (coordonnées, couleurs, proximités, …) et le texte afin d'identifier les pourcentages, trouver à quelle légende de chaque graphique ceux-là correspondent et associer le pourcentage à son étiquette (aligné, non-aligné, obligation souveraine ou non, …).

Enfin, le dernier traitement vise à détecter un certain nombre d'incohérences, ou de remonter certaines alertes quant à la qualité du *reporting* des documents :

- Ce module regarde un certain nombre de choses telles que : la langue du document, la cohérence entre les questions présentées et l'article indiqué dans la question sur les objectifs d'investissement durable, l'absence de réponse à certaines questions, l'exploitabilité du document par une machine, et quelques contrôles de conformité selon la question étudiée.