L'optimisation de la stratégie commerciale des organisations grâce à l'analyse des données ouvertes

Pablo Perez

le cnam
intd

L'optimisation de la stratégie commerciale des organisations grâce à l'analyse de données ouvertes.

Mémoire pour l'obtention du Master Sciences humaines et sociales mention humanités numériques

Parcours Mégadonnées et analyse sociale (MEDAS)

Pablo PEREZ

Date et lieu de la soutenance

- 07/09/2021
- À distance

Membres du jury

- Claudia Marinica, tutrice CNAM
- Xavier Aimé, tuteur CNAM
- Maxime Morinière, tuteur entreprise
- Ghislaine Chatron

Promotion (2019-2021)

Entreprise OCTOPUSMIND

Expert en analyse de données économiques.

PEREZ Pablo

Le modèle social actuel est en train de changer. Nous passons d'une société industrielle à une société de la connaissance, une société dans laquelle les données sont une matière première. Les organisations et les gouvernements évoluent. Ces acteurs génèrent et essaient de traiter de plus en plus de données, avec l'objectif de créer de la valeur. Néanmoins il existe différents éléments et enjeux à prendre en compte afin de constituer une stratégie viable d'exploitation des données ouvertes (Open Data).

Ce mémoire a une approche sur les avantages économiques et sociaux de l'exploitation des données ouvertes, mais également sur les principales problématiques liées à la récupération et le traitement de ces informations. Différents exemples de projets innovants liés à l'exploitation des données complémentent cet ouvrage avec la finalité de guider et encourager les acteurs sociaux à entreprendre des projets d'analyse et de valorisation de l'Open Data.

Descripteurs

Open Data

Mégadonnées

Smart Data

Data mining

Les enjeux de l'Open Data

Stockage des données

Data wrangling

Python

The current social model is changing. We are moving from an industrial society to a knowledge society, a society in which data is a raw material. Organizations and governments are changing. These actors generate more and more data. However, there are various elements and issues to know before constituting a strategy to exploit Open data.

This thesis has an approach based on the economic and social benefits of the exploitation of Open Data, but also on the main problems related to the retrieval and processing of this information. Various examples of innovative projects related to these exploitation of data complete this work with the aim of guiding and encouraging social actors to undertake similar initiatives.

Keywords

Open Data

Big data

Smart Data

Data mining

Challenges of Open Data

Data storage

Data wrangling

Python

Para Guillermo Cintora,

Jorge Alberto Gutiérrez Limón,

y Walter Fernández Baca

Profesores de vida.

Remerciements

Je souhaite avant tout remercier ma tutrice du mémoire, Claudia MARINICA, pour le temps qu'elle a consacré à m'apporter les outils méthodologiques indispensables à la conduite de ce travail. Son accompagnement m'a grandement stimulé.

Un grand merci également à Xavier AIMÉ pour m'avoir accompagné dans l'élaboration de ce mémoire et pour avoir eu la patience de répondre à toutes mes questions.

Je remercie en particulier Frédéric OLIVEAU et Maxime MORINIÈRE, pour tous vos conseils et également de m'avoir donné l'occasion extraordinaire de réaliser mon travail de terrain.

L'enseignement de qualité dispensé par le Master Sciences humaines et sociales mention humanités numériques (MEDAS) a également su nourrir mes réflexions et a représenté une profonde satisfaction intellectuelle, merci donc aux enseignants-intervenants.

J'aimerais exprimer ma gratitude à tous les chercheurs et spécialistes, trop nombreux pour les citer, qui ont pris le temps de discuter avec moi sur mon sujet. Chacun de ces échanges m'a aidé à faire avancer mon analyse.

Table des matières

Figures

Introduction

Nous vivons dans “l'ère des données”, dans une époque marquée par le mouvement d'ouverture des données[1].

Les données ouvertes sont des ensembles d'informations disponibles dans des formats numériques standards et accessibles de façon gratuite. Elles sont publiées sur le web afin que tout le monde puisse y accéder facilement. Ces données suivent une philosophie qui tend à la liberté et l'ouverture sans restriction du droit d'auteur, des brevets ou d'autres mécanismes de contrôle ou limitation; car les motivations du public ou des utilisateurs de ces informations sont diverses : par exemple s’informer ou développer de nouveaux services ou produits en augmentant la valeur sociale et commerciale des données.

Comme a indiqué la CNIL dans son glossaire[2] “*l’Open data désigne un mouvement d’ouverture et de mise à disposition des données produites et collectées par les services publics (administrations, collectivités locales...)*”.
L'ouverture des données est donc le processus qui met des données à disposition de la société, dans des formats numériques standardisés et ouverts en suivant une structure claire permettant leur compréhension et leur utilisation. Ces données peuvent, par conséquent, être réutilisées et redistribuées ou partagées librement [9].

Avant ce partage en masse d’information, les entités publiques hébergeaient et protégeaient de grandes quantités de données concernant l'activité du gouvernement et de la société, par exemple les statistiques de de circulation routière, des consommations des ménages (issus des analyses fiscales), le répertoire des logements locatifs des bailleurs sociaux, etc... Néanmoins, l’ouverture de ces informations, ces dernières années, a aidé à stimuler la croissance économique, à soutenir la bonne gouvernance et à impulser l'innovation sociale. Motivés par ces avantages, plusieurs gouvernements dans le monde ont commencé à inclure des initiatives d’ouverture des données dans leurs stratégies.

En France, depuis 2018, toutes les collectivités locales de plus de 3500 habitants sont obligées de publier leurs données selon la loi République Numérique[3]. D'autres États dans le monde prennent des initiatives comme celle-ci, toujours avec des objectifs similaires liés à trois aspects principaux: la croissance économique (innovation commerciale, création d'entreprises et d'emplois), la participation citoyenne et l'amélioration de l'efficacité des opérations et des services (meilleure prise de décision grâce à l'accès aux données d'autres organisations).

Réutilisatrices de données ouvertes, certaines entreprises développent de nouvelles applications pour leurs clients et créent de nouveaux marchés, de nouvelles entreprises et des modèles commerciaux innovants [2].

[1] www.francearchives.fr/fr/open_data
[2] www.cnil.fr/fr/definition/open-data
[3] https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000033202746/

Aujourd'hui, le concept Open Data évoque généralement une allusion à la modernité, le chemin qui connecte les futures Smart cities, des plateformes des données innovantes et ergonomiques, faciles d'utilisation. Néanmoins la réalité est souvent différente, car cette jeune et accélérée tendance d'ouverture d'information a entraîné certains problèmes liés aux producteurs de données qui rendent difficile le travail des acteurs économiques intéressés par leur exportation.

Ces problématiques favorisent la résilience à l'utilisation des données. Par exemple, les données ouvertes sont malheureusement publiées, souvent, sous des formats très hétérogènes sans aucune stratégie de réutilisation. De plus, certaines questions se posent sur leur qualité, pertinence et fiabilité, et même si pour faire face à ces contraintes, des alternatives techniques et des outils de traitement existent, une nouvelle problématique apparaît : où chercher la data ? La communication sur les sources qui partagent ces informations et leur niveau de confiance sont des informations encore difficiles d'accès.

Les quatre principaux obstacles à l'utilisation des informations sont donc :

- les obstacles ou les risques juridiques liés à l'extraction et exploitation des données ;
- la disponibilité des données et le manque d'information qui les décrivent et qui facilitent leur exploitation (problèmes de publication) ;
- le manque de connaissances techniques et d'infrastructures nécessaires limitant la capacité à collecter, sauvegarder, gérer et utiliser ces données (problèmes techniques) [4] ;
- la question sur la qualité des informations, par exemple l'abondance des données incomplètes, doublées, obsolètes ou non pertinentes. Ces éléments alourdissent le travail de préparation et de nettoyage des données, une étape fondamentale pour la réussite des résultats fiables issus des analyses.

Dans ce contexte, est-il encore possible de tirer du bénéfice de l'exploitation des données ouvertes? est-ce qu'il existe des techniques et outils qui peuvent aider à réduire le risque et d'augmenter le retour sur investissement (ROI); plus concrètement comment est-ce que les entreprises et les entrepreneurs intéressés par l'exploitation des données publiques peuvent trouver, analyser et extraire de la valeur afin d'améliorer leur stratégie commerciale, prendre de meilleures décisions et créer de la valeur avec le développement des nouveaux produits et services ?

En plus, de répondre à ces questions, ce travail présente plusieurs projets et initiatives en cours liées à l'analyse de l'Open Data et notamment le cas d'étude de la plateforme Databiz (plateforme interactive d'analyse économique de données mondiales), dans le but d'inviter le lecteur à découvrir les avantages et les opportunités qui se cachent derrière l'exploitation des données ouvertes et de comprendre la valeur qu'elle peut avoir.

I. Les données ouvertes et leur potentiel économique

I.A Les données ouvertes

I.A.1 Les producteurs des données ouvertes

I.A.1.1 Ouverture des données publiques

Comme l'a indiqué Marc Garriga-Portolà, militant des processus d'ouverture des données publiques en Espagne, et ex-membre du groupe de pilotage des premiers projets de données ouvertes de la mairie de Barcelone (2011), dans un article [15], "*tout ce qui a été généré avec l'argent public doit être accessible à toute la société sans discrimination, y compris les données*".

L'ouverture des données est donc à la fois un une philosophie d'accès à l'information, une politique publique et une pratique de publication de données librement accessibles et exploitables basée sur l'idée que ces données sont un bien commun.

Depuis sa création, le mouvement des données ouvertes s'est concentré sur l'ouverture des données du secteur public. Cependant, cette perception a évolué et il est aujourd'hui plus courant de parler d'Open Data d'une façon plus générale qui englobe tous les autres secteurs producteurs, comme le montre par exemple la déclaration des principes de l'Open Data Charter[4], qui est parfaitement applicable à n'importe quel secteur.

Néanmoins, les gouvernements restent les principaux producteurs de données ouvertes [2]. Généralement, ces informations ont été collectées par l'administration publique ou par des organisations privées, dans différents domaines, tels que la santé, la cartographie, la météo, l'éducation, les marchés publics, la législation, etc. [5]. Ces données publiques, détaillées dans la figure 2, incluent un large répertoire de données collectées ou financées par les gouvernements nationaux, régionaux et locaux et les agences publiques ou institutions qui font partie de la fonction gouvernementale (par exemple, les archives judiciaires).

[4] www.opendatacharter.net/principles

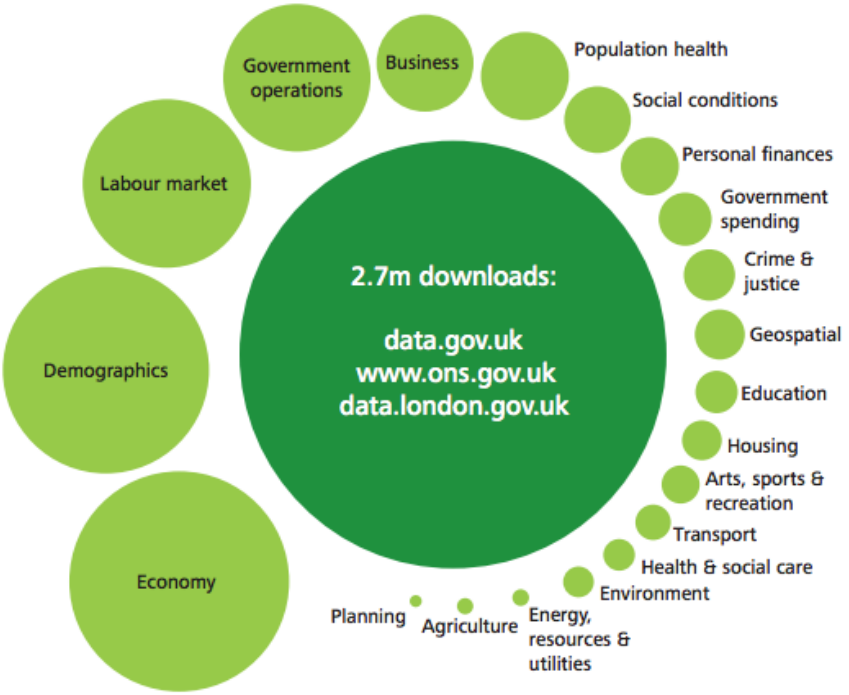


Figure 1 : Exemple de données ouvertes publiées par le gouvernement britannique[5].

Un exemple de la mise en pratique de cette philosophie est la signature, par l'ex-président des Etats-Unis, Barack Obama, lors de son premier jour en tant que président, d'un mémorandum pour toutes les agences gouvernementales dans le but de promouvoir la transparence, la participation et la collaboration entre le gouvernement et les citoyens. Ce document est appelé Open government directive[6] et il définit les lignes directrices que les agences gouvernementales américaines doivent suivre pour parvenir à plus de transparence, de participation et de collaboration. Parmi d'autres objectifs, il force l'ouverture des données publiques et leur publication sur le portail Data.gov[7], créé pour favoriser l'ouverture de l'information publique.

Tout cela était (et continue d'être) d'une grande importance politique et sociale, car d'un point de vue interne aux États-Unis, la mise à disposition des données à la société lui permet de valider et de contrôler leur qualité, entraînant une amélioration de l'efficacité du secteur public. Aujourd'hui, les gouvernements qui divulguent des informations sur le secteur public sont appelés Open Governments ou Gouvernements ouverts [7].

Les avantages économiques et sociaux de l'ouverture des données sont nombreux, mais en général, la mise en œuvre de technologies de cette nature contribue à la modernisation de l'État et des organisations, facilite le contrôle interne et externe, contribue à la transparence, réduit la distance entre les acteurs économiques, publics et les citoyens et enfin favorise la participation sociale aux différents processus décisionnels.

[5] www.koombea.com/blog/why-the-future-of-your-business-could-rely-on-open-data
[6] www.digital.gov/open-government-directive
[7] www.data.gov

Un processus d'ouverture des données permet donc aux citoyens d'en savoir plus sur les organisations publiques, de mieux percevoir leurs limites, ce qui se traduit, à long terme, par une meilleure perception des services publics.

Cela devient tellement important que les gouvernements tentent d'impulser cette révolution numérique en soutenant collectifs tel que Open Knowledge Foundation, et en allant encore plus loin, le G8 soutient même le concept d'ouverture par défaut (open by default) [1], reconnaissant le rôle central des données ouvertes dans la transparence de l'action et gestion gouvernementale.

L'Open Government Working Group[8] a été lancé au début de 2010 dans les États Unis en tant que forum pour les professionnels des Gouvernement ouverts afin de partager les meilleures pratiques et de promouvoir la transparence, la participation et la collaboration. Les huit principes[9] des données publiques qui ont étés définis dans le forum sont :

1) Ouverture de toutes les données publiques.

2) Publier les données directement depuis la source, avec un niveau de granularité le plus fin possible.

3) Mise à jour dans des délais appropriés pour préserver leur valeur. Par exemple, il ne sert à rien d'ouvrir des informations du trafic routier si elles ne sont pas mises à jour très souvent.

4) Faciliter son utilisation en fournissant des outils d'accessibilité, de tri, de recherche et de contrôle de la qualité de l'information pour l'ensemble de la société.

5) Faciliter la réutilisation automatique en utilisant des formats de fichiers appropriés à cette fin.

6) Les distribuer, sans aucune discrimination ni limitation, car *si ces données ne peuvent pas être indexées, elles n'existent pas*. [19]

7) Utilisez des *formats* sans restrictions afin que personne n'ait le contrôle exclusif.

8) Utiliser une licence gratuite qui encourage la réutilisation, sans objet de droit d'auteur qui la limite (sauf dans les cas où des restrictions sont autorisées pour des raisons de sécurité, de confidentialité ou qui sont régies par une loi spécifique ou par une procédure administrative).

Ces principes ont un seul objectif : ouvrir les données pour qu'elles puissent être utilisées. Par exemple, une réalité sur Internet est que si les moteurs de recherche ne peuvent pas trouver d'informations, elles n'existent pas (principe 6 - indexation). Les 8 axiomes de l'ouverture des données publiques, parlent également des formats qui facilitent leur réutilisation, par exemple des formats standards (shp, csv,) qui permettent la réutilisation d'applications, langages et logiciels faits pour des données ouvertes. Ces formats dépendent des caractéristiques de chaque donnée.

Les types de format sont détaillés dans la section Annexes - *Formats de données ouvertes*.

[8] https://obamawhitehouse.archives.gov/open/about/working-group
[9] https://opengovdata.org/

Enfin, un cadre juridique [10] est nécessaire afin de partager ce qui a été créé pour motiver la société à offrir de nouveaux services avec des informations ouvertes ou simplement montrer un fait intéressant qui est déduit de ces informations. En d'autres termes, l'Open Data doit être proposé avec des conditions d'utilisation qui encouragent au maximum sa réutilisation, même à des fins commerciales.

I.A.1.2 Ouverture des données privées

Aujourd'hui les progrès des technologies numériques ont permis à tous les types d'organisation de générer de manière permanente et exponentielle de grandes quantités de données. Il s'agit d'informations provenant à la fois des systèmes internes de l'organisation et de son interaction avec les fournisseurs et les clients.

Il existe un grand volume de données du secteur privé qui pourraient être ouvertes au grand public et aux autres organisations. Par exemple, les informations de suivi des véhicules pour la gestion du trafic et le développement d'infrastructures, ou les données de ventes qui pourraient aider à l'estimation de l'indice des prix à la consommation.

Un autre exemple de partage et d'ouverture de données privées pourrait être un portail d'information pour les producteurs ruraux, permettant la mise en place de coopératives de données.

La disponibilité de ces données pourrait diminuer l'indépendance des producteurs aux grands brokers mondiaux et faciliter ainsi le développement et l'augmentation de leur compétitivité. [6]

La collecte et l'analyse des données en interne impliquent des avantages importants pour les organisations, tels que l'optimisation des processus et l'amélioration de la prise de décision. Néanmoins, l'ouverture de ces données offre également des opportunités de croissance économique, de compétitivité et d'innovation. [16]

Les principaux avantages qu'apporte l'ouverture de données, soit dans des environnements B2B (avec d'autres entreprises) soit B2G (avec des administrations publiques et des universités ou des organismes de recherche), peuvent être par exemple la monétisation des données comme moyen de générer des bénéfices supplémentaires, la possibilité de collaborations avec d'autres organisations, le positionnement stratégique de l'entreprise et/ou le soutien à l'innovation.

La figure 1 présente la chaîne de valeur autour des données ouvertes. Elle commence premièrement avec la source de données, celles-ci peuvent provenir d'organisations publiques ou privées.

Après il y a les mécanismes et techniques qui permettent la publication et la mise à disposition des données (des fois ces services sont offerts pour des infomédiaires). Enfin les

[10] Ce sujet est traité dans la section *II.A La législation.*

utilisateurs, qui peuvent être des citoyens ou des entreprises, créent de la valeur en combinant et en utilisant efficacement ces ensembles de données.

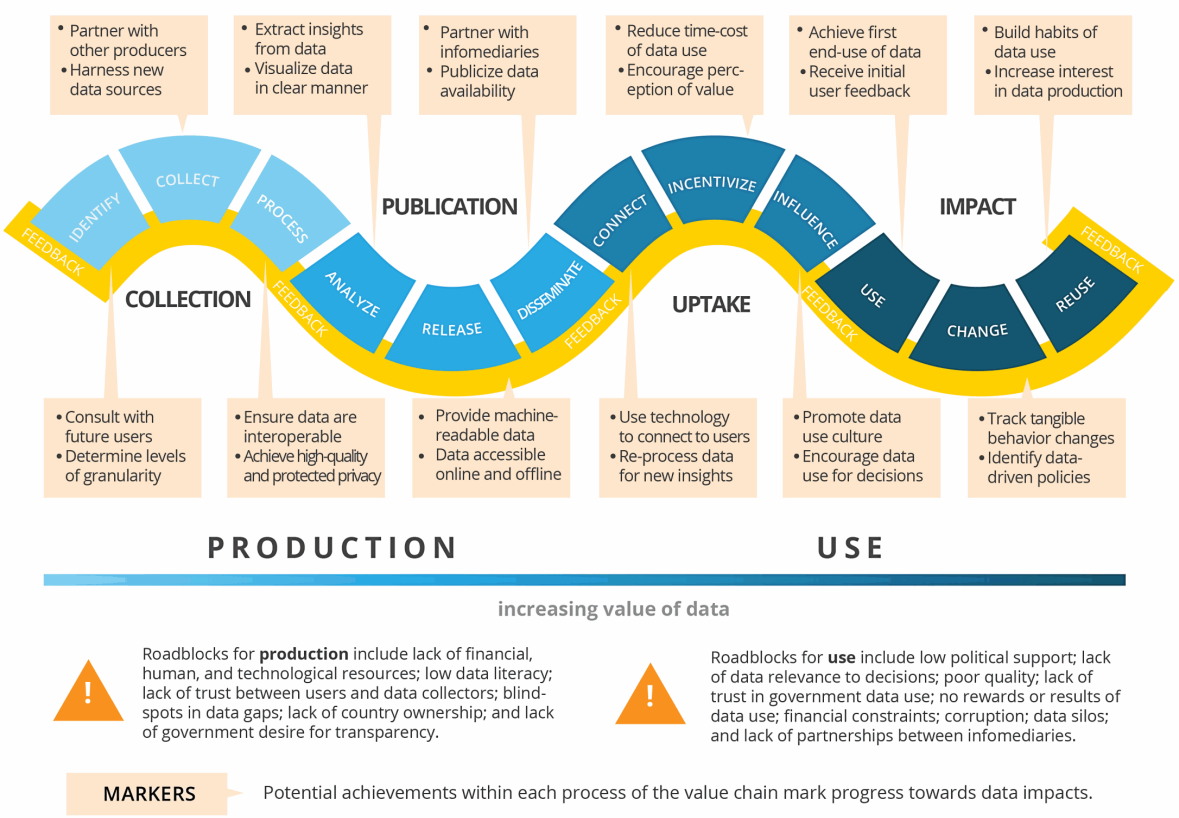


Figure 2 : Chaîne de valeur autour des données ouvertes (Open Data Watch Organization)[11].

Il est important de mentionner que le partage de données n'implique pas nécessairement que l'entreprise offre l'accès à tous ses ensembles de données. Elle doit plutôt décider quelles sont les données à rendre publiques, en fonction de sa stratégie commerciale.

Un exemple de réussite est le service BBVA Valora. Le projet de cette banque internationale a été de créer un nouveau service ouvert à tout public, basé sur ses données. Cette plateforme offre des informations utiles avant d'acheter ou de louer une maison ou bien une voiture, et vous permet de gérer toutes les informations liées au crédit et aux dépenses du ménage (en cas de l'achat d'un bien immobilier) en un seul endroit.[12]
Grâce à cette initiative, la banque est devenue plus compétitive, elle a renforcé son image et sa capacité d'innovation, ainsi que son portefeuille clients. Également, cette plateforme lui a permis d'élargir sa base de données en collectant de nouvelles informations.

Cependant, l'ouverture ou partage de données implique un changement au niveau des infrastructures techniques, ainsi que dans les processus organisationnels et dans la culture de l'entreprise.
Par conséquent, si l'organisation n'a pas de connaissances internes en analyse de données, la chose la plus appropriée est d'avoir le soutien d'un cabinet ou d'un personnel externe qui

[11] www.opendatawatch.com/publications/the-data-value-chain-moving-from-production-to-impact
[12] www.bbva.es/personas/experiencias/bbva-valora.html

va partager ses connaissances et qui donnera un conseil sur la valeur et le potentiel des ces données.

En résumé, afin qu'une initiative de données partagées ou ouvertes soit efficace, il faut disposer d'un personnel qualifié pour la gestion et le partage des données, ainsi qu'une stratégie :

- soutenue par un modèle commercial viable.
- avec un modèle de financement associé qui soit rentable (qui dépasse le seuil de rentabilité de l'investissement).
- qui respecte le cadre juridique existant.
- qui favorise un changement de la culture organisationnelle. [17]

I.A.1.3 Données ouvertes dans différents domaines

Ces dernières années, de nouveaux secteurs producteurs de données sont apparus. Ces données sont complémentaires aux centaines de catalogues Open Data des gouvernements et du secteur privé [14]. Elles proviennent surtout du :

- Secteur universitaire et de l'innovation
- Troisième secteur (les associations non lucratives)
- Des citoyens

On peut citer, par exemple, le projet *Save the Rain*[13]. Cette application, développée aux Etats-Unis, montre aux utilisateurs comment ils peuvent aider à réduire l'impact de la réduction alarmante du niveau annuel des précipitations. Grâce à des cartes, les utilisateurs calculent la quantité d'eau tombée qu'ils pourraient réutiliser chaque année.

La World Bank présente beaucoup d'autres exemples d'Open Data provenant de secteurs et de sujets très variés[14]. La figure 3 présente plusieurs exemples de ces projets ainsi que d'autres sites web et plateformes de données ouvertes classés par secteur d'activité.

Secteur	Exemples de portails des données ouvertes
Agriculture	Guide national des marchés des agriculteurs des États-Unis https://search.ams.usda.gov/farmersmarkets/
Finances et macroéconomie	Catalogue des données de la Banque mondiale https://datacatalog.worldbank.org/

[13] www.savetherain.org
[14] www.opendatatoolkit.worldbank.org/es/essentials.html

Education	UNESCO - Organisation des Nations Unies pour l'Éducation http://data.uis.unesco.org/
Environnement	Portail de données sur le climat, alimenté par le think-tank The Shift Project https://www.theshiftdataportal.org/
Information géospatiale	OpenStreetMap www.openstreetmap.org
Santé	IPUMS et un projet consacré à la collecte et à la distribution des données de recensement du monde entier https://international.ipums.org/international/
Transport et mobilité	Rassemblement des données de toute l'offre de mobilité à travers la France https://transport.data.gouv.fr/
Eau potable et assainissement	Base de données mondiale sur l'eau http://map.mwater.co/
Développement durable	Objectifs de développement durable de l'ONU https://unstats.un.org/sdgs/indicators/database
Développement social	Socialprogress est un site web qui évalue le développement social et environnemental de différents pays https://www.socialprogress.org/
Société et culture	Worldometer est une plateforme qui fournit des estimations et des statistiques en temps réel https://www.worldometers.info/

Figure 3 : Exemples de sites web et plateformes de données ouvertes classés par secteur d'activité.

Grâce à cette variété de producteurs de données, les possibilités d'enrichissement de l'information sont augmentées, et en même temps un nouveau besoin se crée pour améliorer la coordination et l'intégration des données ouvertes. De là sont nés de nombreux projets innovants basés sur l'Open Data.

Par exemple, après les Rencontres Nationales du Transport Public[15] du 2015, Keolis et OpenDataSoft [16] ont lancé la solution Open Data Mobilité (data.explore.star.fr), plateforme qui permet au Réseau STAR de publier et de gérer ses jeux de données. Ces données ont impulsé la création de l'application mobile START[17] permettant aux citoyens de se déplacer dans Rennes en transport en commun.

Il est intéressant de remarquer l'importance sur les données économiques ouvertes, car ces informations peuvent être utiles pour une institution gouvernementale ou éducative, une organisation non lucrative ou une entreprise traditionnelle. La recherche des données économiques et financières, joue un rôle central dans l'élaboration des politiques et des initiatives dans de nombreux domaines importants de nos économies et de nos sociétés, ainsi que dans la création de services et de produits innovants.

Il est donc possible de les croiser de différentes informations avec des données économiques afin d'obtenir des indicateurs innovants. Par exemple, le fait de savoir si l'éducation, le développement humain (HDI), le changement climatique ou le sentiment de sécurité de la population est corrélé à la demande de certains produits ou services [42, 43].

La plateforme Social progress (www.socialprogress.org) a créé un modèle business avec cette base de connaissances. Leur site internet fournit aux décideurs et aux citoyens les meilleures données sur la santé sociale et environnementale afin de les aider à établir des priorités.

Ce sujet est traité dans la section suivante.

I.A.1.3.1 Les données économiques ouvertes

Les données économiques prennent de nombreuses formes, telles que celles issues de la bourse et les marchés financiers, des enquêtes de consommation et de production industrielle, du commerce de détail, du chiffre d'affaires des entreprises cotées en bourse, des prix des logements, etc.

Avant, seuls les courtiers institutionnels et les puissants investisseurs avaient accès à ce type de données. Mais, aujourd'hui dans de nombreux pays, ces informations sont ouvertes à tout public, accessibles gratuitement et publiées dans plusieurs sites dédiés [11].

Cependant, l'ouverture des données économiques aurait de nombreux avantages, comme le précise l'Open Economics Working Group, fondée par l'Open Knowledge Foundation en association avec le Center for Intellectual and Property Law (CIPIL) de l'Université de Cambridge, dans ses Principes d'économie ouverte.

Outre des points déjà relevés plus hauts tels que le sujet de la connaissance en tant que bien public ou encore l'amélioration de l'efficacité, ce groupe de travail voit aussi des points positifs pour les chercheurs du monde entier (accès facilité aux données, reproductibilité des travaux de recherche).

---

[15] www.rencontres-transport-public.fr
[16] www.keolis.com , www.opendatasoft.com
[17] www.play.google.com/store/apps/details?id=com.keolisrennes.starmobilites

Voici une liste des principales sources de données économiques ouvertes.

Description	Source	Granularité géographique
Le portail data.economie.gouv.fr a pour objectif de valoriser les données ouvertes produites par les Ministères économiques et financiers et de faciliter leurs réutilisations par le plus grand nombre afin d'améliorer le service rendu aux usagers et citoyens.	www.data.economie.gouv.fr	France
Les tableaux de l'économie française de l'INSEE s'adressent à tous ceux qui souhaitent disposer d'un aperçu rapide et actuel sur la situation économique, démographique et sociale de la France.	www.insee.fr/fr/statistiques	France
Publications, dossiers, actualités, textes officiels, statistiques, informations pratiques et boursières de la banque centrale française.	www.banque-france.fr/statistiques	France et l'etranger
Opendatasoft est une société française qui propose des logiciels et conseils pour le partage et l'ouverture de données. Elle offre aussi des jeux de données dans sa Data Network.	www.public.opendatasoft.com/explore	France et l'etranger
L'Eurostat est la direction générale de la Commission européenne chargée de l'information statistique à l'échelle communautaire.	www.ec.europa.eu/eurostat/fr/web/main/data/database	Europe
Le site DataBank est un outil d'analyse et de visualisation qui contient des séries de données chronologiques sur des sujets divers.	www.databank.worldbank.org/home.aspx	Monde entier
L'OCDE, offre une publication statistique mensuelle présentant des données pour un large éventail d'indicateurs pour les pays membres ainsi que pour une sélection d'économies non membres.	www.data.oecd.org	Monde entier
BDnomics est une plateforme gratuite d'agrégation de données économiques ouvertes. Le site a été conçu comme un outil pour améliorer le processus de traitement des données en économie.	www.db.nomics.world	Monde entier
Our World in Data est un projet du Global Change Data Lab, une organisation à but non lucratif basée au Royaume-Uni. Cette initiative est le fruit d'une collaboration entre des chercheurs de l'Université d'Oxford et l'organisation à but non lucratif Global Change Data Lab, qui possède, publie et entretient la plateforme et les outils de données.	www.ourworldindata.org	Monde entier
Site des statistiques officielles du commerce international des Nations Unies	www.comtrade.un.org/data	Monde entier
L'Organisation mondiale du commerce (WTO) fournit des informations quantitatives sur les questions de politique économique et	www.data.wto.org	Monde entier

commerciale. Ses bases de données donnent accès à des données sur les flux commerciaux, les tarifs, les mesures non tarifaires (NTLs) et le commerce en valeur ajoutée.		
UNdata est un service de données basé sur Internet qui amène la Division de statistique des Nations Unies qui s'engage à faire progresser le système statistique mondial.	www.data.un.org , www.population.un.org/wpp	Monde entier
Statistiques financières internationales de Fonds Monétaire International (FMI).	www.data.imf.org	Monde entier
Le Global Housing Watch suit l'évolution des marchés du logement à travers le monde sur une base trimestrielle. Il fournit des données actuelles sur les prix des logements ainsi que des paramètres utilisés pour évaluer l'évaluation sur les marchés du logement, tels que les ratios prix des logements par rapport au loyer et prix des logements par rapport au revenu.	www.imf.org/external/research/housing	Monde entier
IndexMundi contient des statistiques détaillées sur les pays, des graphiques et des cartes compilées à partir de sources multiples. Par exemple, la les prix des produits de base au bourse comme les metaux.	www.indexmundi.com/commodities	Monde entier
Ce site propose plus de 20 millions d'indicateurs économiques, des données historiques, des graphiques, des nouvelles et des prévisions de 196 pays.	www.tradingeconomics.com/indicators	Monde entier
Worldometer est un site web qui fournit des estimations et des statistiques en temps réel pour divers sujets sur la base de différents algorithmes. Il est détenu et géré par la société de données Dadax.	www.worldometers.info	Monde entier
Yahoo ! Finance est un service de Yahoo qui fournit des informations, des données et même des commentaires financières et boursières du monde entier.	www.finance.yahoo.com/	Monde entier
Le SME Finance Forum gère un réseau mondial de plus de 200 membres qui rassemble des institutions financières, des entreprises technologiques et des institutions de financement du développement pour partager des connaissances, stimuler l'innovation et promouvoir la croissance des PME.	www.smefinanceforum.org/data-sites	Monde entier
Gapminder est une fondation suédoise indépendante. Cette initiative identifie les idées fausses systématiquement sur les tendances économiques mondiales et utilise des données fiables pour développer du matériel pédagogique facile à comprendre.	www.gapminder.org/data	Monde entier

I.A.1.3.1.1 Classification des données économiques ouvertes

Dans un sens large, il est possible de classifier les données économiques selon leur champ d'étude en données macro et micro économiques.

La microéconomie se concentre sur l'observation et l'analyse des interactions à une petite échelle (offre et demande, détermination des prix, etc.). Inversement, la macroéconomie étudie l'économie au niveau national ou international. Ces deux types de données sont détaillés dans les sections suivantes.

I.A.1.3.1.1.1 Données macroéconomiques

Aujourd'hui, les données macroéconomiques sont, parmi les données économiques ouvertes, les plus abondantes [34]. L'ouverture des données du secteur privé, c'est-à-dire les données issus de l'activité des entreprises, des associations, des ONG, etc. est en effet encore précaire (comme évoqué dans la section *I.A.1.2 Ouverture des données privées)*.

La macroéconomie est la science qui étudie les phénomènes économiques à l'échelle globale (nationale, internationale, collective) et leur répartition dans un secteur donné. Par exemple: la consommation, la production, l'emploi, le revenu, l'investissement, l'inflation, le taux de chômage, etc. [35] .

Cet approche théorique a pour objectif l'étude globale de l'économie à partir des agrégats (grandeur synthétique mesurant le résultat de l'activité économique de même nature en valeur ou en volume[18]). Les principaux agrégats sont : le produit intérieur brut (PIB), le rapport de la balance commerciale, l'indice des prix à la consommation (IPC), le taux de chômage, le taux d'intérêt et le taux d'inflation.

I.A.1.3.1.1.2 Données microéconomiques

Comme l'indique le Centre National de Ressources Textuelles et Lexicales (CNRTL)[19] , la microéconomie est la science qui étudie les phénomènes économiques restreints: individuels, unités économiques réduites. Cette discipline a pour objectif l'étude des agents économiques : le consommateur, les entreprises, les associations, etc.

Les données microéconomiques ont une origine dans l'achat de biens et services. Par exemple :

- les acquisitions réalisées par une entreprise dans un secteur donné (par exemple celles faites par Nestlé dans le secteur de l'eau).
- l'estimation des dépenses annuelles dans l'alimentation par ménage.
- la répartition des achats publics de biens et de services par rapport aux marchés publics conclus.

[18] www.insee.fr/fr/metadonnees/definition/c1672
[19] www.cnrtl.fr/definition

L'accès à ce genre des données est généralement plus compliqué que pour les données macroéconomiques. Par exemple, les bases de données microéconomiques d'Eurostat sont gratuites, mais ne sont pas accessibles librement. Elles nécessitent la présentation d'un projet de recherche justifiant l'ouverture de ces données anonymisées [20].

Cependant, il existe des sources comme l'INSEE (Institut national de la statistique et des études économiques) où il est possible de télécharger plusieurs jeux de données, des études et des enquêtes sur des sujets liés aux sentiments et consommations des ménages en France ou sur la conjoncture des différents secteurs économiques [21] .

D'autre part, les détails sur l'achat des biens et de services de l'État peuvent être analysés aussi à travers l'étude des appels d'offres [36, 71]. Ce type d'information publique est disponible sur différents sites, par exemple le site du Bulletin officiel des annonces des marchés publics (boamp.fr) ou le site Tenders Electronic Daily (ted.europa.eu).

Ces types d'études font partie de la théorie microéconomique des contrats, un approche qui conçoit les organisations et les individus comme des entités (des nœuds dans le jargon économique) qui interagissent par le biais de contrats. Une entreprise est, par exemple, un nœud composé de contrats de travail, liant l'entreprise à ses salariés ou ou des contrats avec ses clients (par exemple un contrat de prestation de services), liant celle-ci à eux en tant que fournisseur.

Les marchés publics sont un autre cas particulier de tels nœuds de contrats, ici des contrats d'échange. Les États, au sens des organisations politiques gérant des espaces géographiques déterminés, ont des contrats d'achat qui les lient aux organisations et aux individus qu'ils gouvernent et vice versa [37].

[20] www.sciencespo.libguides.com/economie/etudes-statistiques
[21] www.insee.fr/fr/statistiques/4983805

I.A.2 Ensembles de données massives, ou mégadonnées

Le Big Data (les données massives ou mégadonnées) n'est pas seulement un terme marketing, c'est une réalité avec une histoire solide et un chemin évolutif. Le terme Big Data désigne des ensembles de données si volumineux et complexes qu'ils nécessitent des technologies et des infrastructures non traditionnelles pour les traiter correctement [23].

Des exemples de ce type de données sont : les messages, photos et vidéos des réseaux sociaux, les données issus des capteurs qui collectent des informations du climat, du trafic routier, les images satellites numériques, les coordonnées GPS des téléphones mobiles, etc.

Dans la figure 4, il est possible de se persuader que cette explosion des données ne cesse pas de s'agrandir.

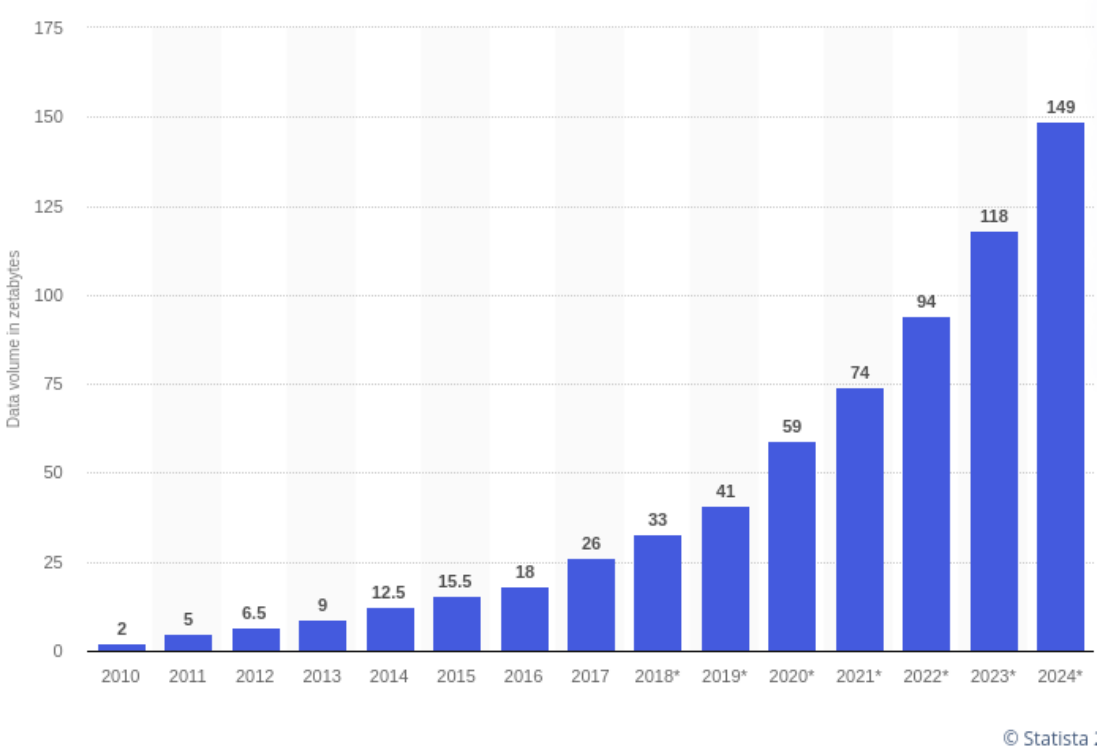


Figure 4 : Volume de données/informations créées, capturées, copiées et consommées dans le monde entier de 2010 à 2024 , Statista 2021 [22].

Avec le développement continu des technologies liées aux mégadonnées, comme le cloud computing (informatique dans les nuages) , le calcul parallèle, les bases de données NoSQL, l'apprentissage automatique, la visualisation des données, etc.), cette masse extraordinaire de données est de plus en plus exploitée [23, 24].

Néanmoins, l'adoption d'une stratégie big data pour une entreprise n'est pas encore simple.

[22] www.statista.com/statistics/871513/worldwide-data-created

Les coûts de l'infrastructure nécessaire pour stocker et traiter un très grand volume de données (comme l'achat des ordinateurs et serveurs dédiés), l'investissement en capital humain et dans l'acquisition des connaissances nécessaires pour obtenir des résultats satisfaisants et efficaces, ainsi que la difficulté de trouver des données fiables et de qualité.sont des enjeux complexes qui peuvent empêcher l'implémentation des business plan (plan d'affaires) basés sur l'exploitations de données massives.

Ces nombreux défis imposés par le Big Data sont caractérisés par ses 5 élément principaux, modélisées avec des V [25, 38] :

- Volume : par rapport à l'ampleur exceptionnelle du volume des informations.
- Vitesse : les traitements des mégadonnées doivent êtres rapides et en temps réel.
- Variété : cette caractéristique symbolise les différents formats et la nature des informations. Par exemple, une base des données peut être composée des images, du texte, d'enregistrements audio, etc. [23]
- Véracité : afin d'aboutir à des résultats fiables il est nécessaire d'avoir dans la mesure du possible des informations pertinentes et réelles.
- Valeur : Il est primordial de bien sélectionner les données à analyser, en fonction de l'activité ou secteur d'action et de les objectifs recherchés, afin de réussir à créer de la valeur issus des données massives.

Ces éléments sont représentées dans la *Figure 5 :*

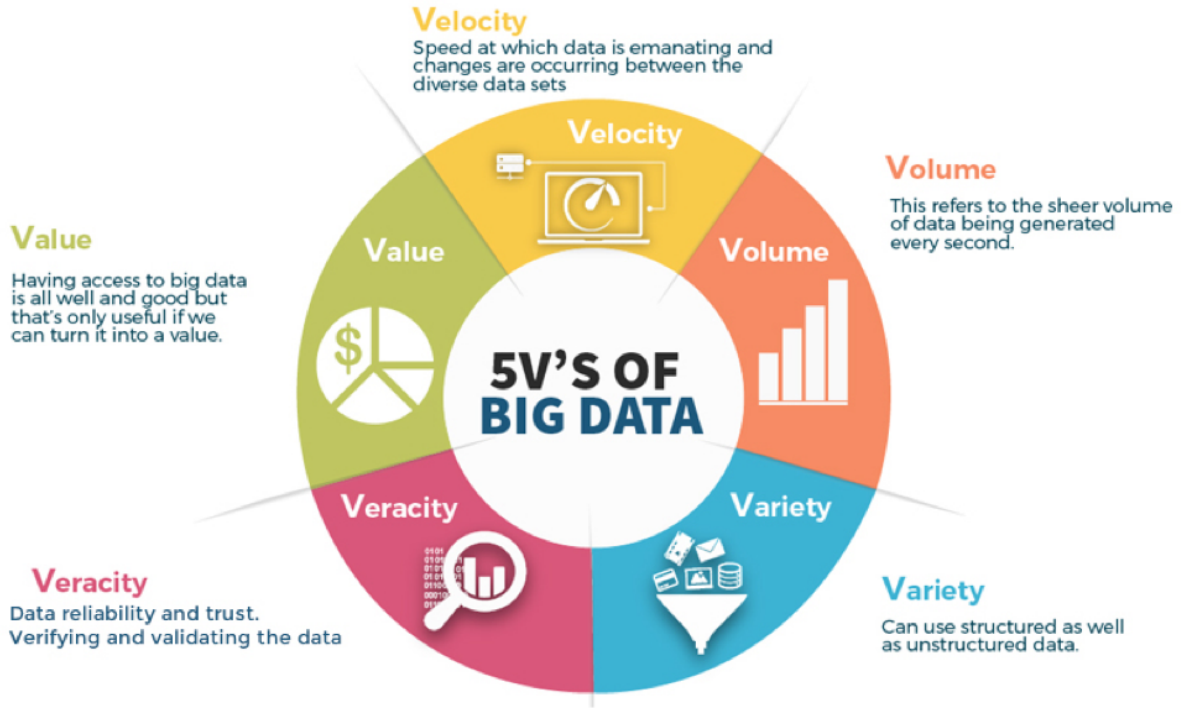


Figure 5 : Les cinq V des données massives [24].

[23] Les principaux formats des données sont présentés dans la section Annexes - Formats des données ouvertes.
[24] www.techentice.com/the-data-veracity-big-data

I.A.2.1 La méthode des données intelligentes et les données massives

Le concept de Smart Data (données intelligentes) apparaît après la naissance du Big Data. Le Smart Data fait référence à la *présélection et à la transformation* de longues listes de chiffres et de données en informations précieuses, réactives et utiles. Donc, par définition, les données massives s'opposent au Smart data qui représente une solution simple, exploitable et facile de la data. Par définition, le Smart Data se différencie donc du concept du Big data par rapport à la simplicité des analyses et les économies dans les infrastructures qu'il offre [40].

La proposition ou la solution suggérée pour la méthode des données intelligentes*,* est de continuer à profiter de l'abondance des données disponibles, mais en le faisant de façon plus concentrée ou focalisée et en définissant un objectif clair et spécifique [39].

Avoir trop d'informations ne conduit pas nécessairement à de meilleures décisions, plus démocratiques ou plus rationnelles. Plus d'informations peuvent entraîner moins de compréhension, plus de confusion et moins de confiance, et un surcoût beaucoup plus élevé [23]. Surtout si on considère tous les problèmes potentiels liés à la qualité des données ouvertes (cet élément a été mentionné dans la section *II.Principaux problèmes lors de l'exploitation de données ouvertes et comment les confronte*r).
Une autre solution ou avantage du Smart Data est que les informations collectées peuvent être stockées (sauvegardées) sur un fichier csv, contrairement aux données massives, qui est le résultat d'une situation informatique complexe, qui nécessite un espace de sauvegarde et visualisation dans un serveur dédié [40].

I.B Création de valeur liée à l'exploitation des données ouvertes et les principales techniques pour y arriver

I.B.1 Les principaux types de création de valeur

Dans le cadre de la mondialisation et du passage d'une logique de production à une logique d'innovation, la puissance d'une entreprise ou d'un État est liée à la maîtrise de toute la chaîne d'informations produites (information interne) ainsi que de la connaissance de son environnement (information externe).

Le cabinet international de consulting Capgemini[25] estime que la contribution directe des données ouvertes à l'économie de l'Union européenne atteindra 199,51 milliards d'euros en 2025 dans un scénario de croissance prudent, et dans le scénario optimiste, elle est de 334,20 milliards d'euros en 2025.
Une croissance est également attendue du nombre de personnes employées grâce aux données ouvertes. Il est prévu que le nombre d'employés directs et indirects des données ouvertes pourrait atteindre 1,97 million en 2025, dans un scénario optimiste [3].

Selon une étude menée en 2020 par l'European Data Portal, les domaines des données les plus réutilisées semblent être des données statistiques (27,3 %), suivies des données géospatiales (25,8 %), des entreprises (19,5 %) et des transports et infrastructures (18,8 %) [80], comme le montre la figure suivante.

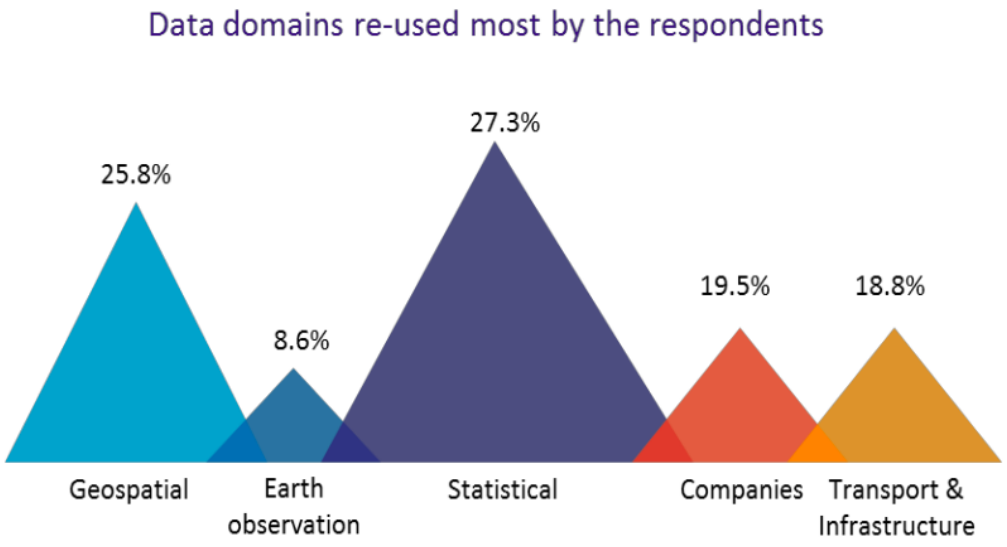


Figure 6: domaines de données les plus réutilisés pour la création des nouveaux produits et services.

[25] Capgemini est un leader mondial du conseil, de la transformation numérique, des services technologiques et d'ingénierie.

Mais pourquoi travailler avec des données ouvertes ? Travailler avec des données ouvertes apporte des avantages multiples et synergiques aux organisations [11, 17, 80]. Certains de ces avantages pour les entreprises travaillant avec des données ouvertes sont :

- L'innovation, principalement le cas dans lequel les données utilisées étaient auparavant fermées et sont maintenant utilisées pour développer de toutes nouvelles entreprises commerciales ou ou des solutions à des problèmes qui n'avaient pas été envisagés auparavant, comme par exemple pour lutter contre l'épidémie de la COVID-19.
  Grâce aux données ouvertes de plusieurs institutions médicales, universités et gouvernements, des chercheurs ont utilisé des méthodes dérivées des réseaux neuronaux qui semblent être beaucoup plus efficaces que les stratégies classiques des contentions des épidémies [83, 84].
- L'amélioration des modèles commerciaux grâce à l'utilisation des données ouvertes.
- Les données ouvertes peuvent également aider à gagner du temps car les données ouvertes peuvent être utilisées pour réduire la charge administrative d'une organisation ou éliminer le temps passé à demander des données.
- Fiabilité accrue de l'entreprise puisque les données ouvertes utilisées sont des informations officielles d'une administration publique qui créent la confiance et la fiabilité de ceux qui utilisent les services fournis par l'entreprise. Des exemples sont des applications qui utilisent les données d'arrivée ou de départ en temps réel des services de transport public ou des applications qui indiquent où trouver des places de stationnement disponibles ou des stations de vélos en libre-service dans une ville.

Ces éléments donnent lieu à la création de nouvelles sources de revenus innovantes et très rentables. Par exemple, le fait même de collaborer dans des projets avec un organisme public qui utilise des données raffinées basées sur ses propres données ouvertes [11, 80].

Les principales sources de revenus issus de l'exploitation des données ouvertes sont en majorité celles qui viennent de la création directe des services ou des produits ou une combinaison des deux. Les organisations qui travaillent avec ce type d'information comme matière première sont encore majoritairement regroupées dans la sphère des TIC [80] comme il est possible de constater dans la figure suivante.

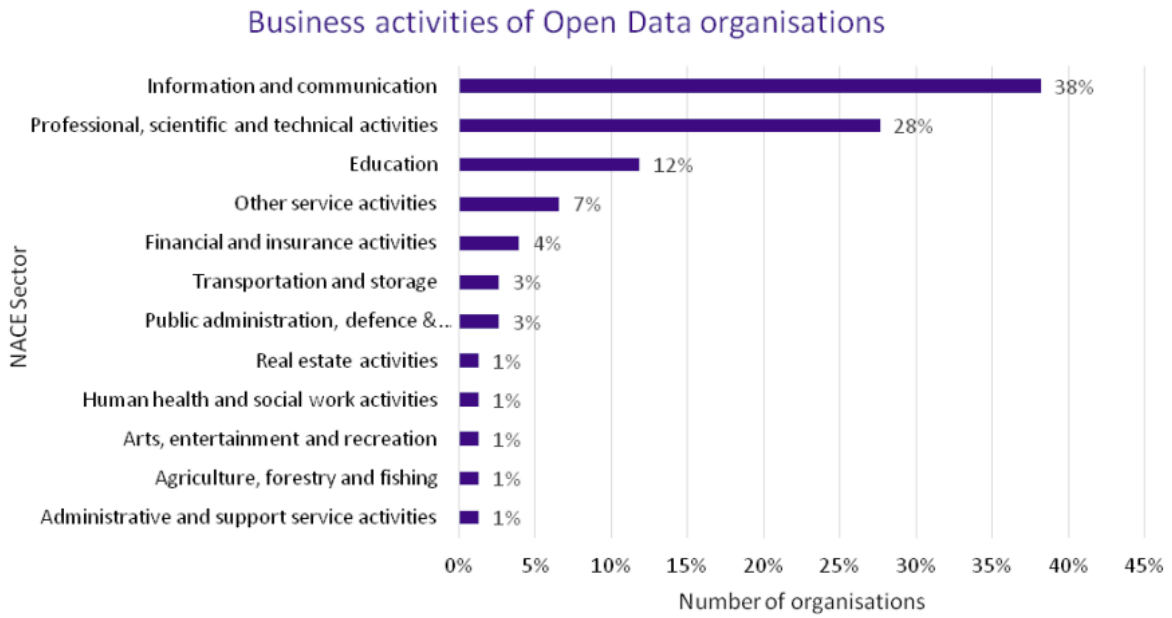


Figure 7: Activités commerciales des organisations de données ouvertes.

Il existe de nombreux autres exemples d'initiatives liées à l'exploitation des données ouvertes qui conduisent les acteurs économiques à stimuler leur capacité d'innovation, à améliorer leur stratégie commerciale, à prendre des meilleures décisions et à augmenter leurs bénéfices économiques ainsi que leurs avantages concurrentiels (certains d'entre eux seront présentés au cours de ce travail).

Les expériences d'ouverture de données d'entreprises sont, malgré cela, encore rares et les bénéfices de l'utilisation des données publiques externes restent encore inconnus pour la plupart des acteurs du secteur privé [2, 80].

Les départements de BI (Business Intelligence), par exemple, se sont traditionnellement concentrés sur l'exploitation des données spécifiques aux produits, services et processus de l'entreprise, c'est-à-dire des données internes. Une donnée interne est donc une donnée qui est générée par un des composants du système d'information interne de l'entreprise, ou qui est saisie en interne.

Néanmoins, le fait d'enrichir ces informations avec des données ouvertes ou des données externes peut fournir une perspective plus large et plus précise des clients et des marchés.
L'une de ces sources peut être l'utilisation des données des médias sociaux. Twitter propose ses données via une API publique qui permet le traitement de "tweets" pour faire de la fouille d'opinion: pour savoir en temps réel quelle perception les gens ont sur un sujet précis.

Afin d'essayer d'encourager les entreprises et les entrepreneurs intéressés à ce sujet, sont présentés ci-dessous les 4 principaux types de création de valeur liés à l'exploitation des données économiques ouvertes.

I.B.1.1 Promouvoir l'expérimentation, découvrir de nouveaux besoins et améliorer les performances

Les technologies de l'information permettent aux organisations de créer des processus et d'établir des expériences contrôlées. L'utilisation de données comme source de ces analyses peut aider à comprendre les phénomènes externes (analyse SWOT), découvrir de nouvelles tendances du marché, pondérer les données internes et réaliser des études comparatives du secteur. Tout cela afin d'améliorer la performance globale de l'organisation.

Il est possible par exemple de créer de nouveaux services par rapport aux besoins clients plutôt que par rapport à la concurrence. Par exemple, à l'aide des techniques du traitement du langage naturel (NLP)[26], il est possible d'analyser une extraction de quelques mois des messages clients et de scruter les messages afin d'identifier de façon générale les besoins clients, c'est-à-dire essayer de répondre aux questions : que recherchent-ils ? Quels sont leurs principaux problèmes ou demandes ? [78].

Les algorithmes de fouille des données textuelles (text mining) peuvent aussi faciliter la recherche des mots-clés précis pour savoir combien de clients ont parlé d'un sujet en particulier, par exemple sur le fait d'avoir accès à une API, à un dashboard, etc. En plus, la réalisation d'une séance brainstorming avec les gens qui répondent à ces messages afin de valider, peut servir à approfondir ou élargir les résultats ou propositions obtenus.

Il est important de mentionner que le fait d'améliorer les performances peut impliquer des avantages au-delà de l'aspect économique. Il y a, par exemple, plusieurs entrepreneurs de l'Open Data qui ont décrit l'impact de leur travail d'un point de vue sociétal, comme l'identification d'économies publiques potentielles, la contribution à l'ensemble des connaissances publiques ou l'aide au suivi et à la prévention de la propagation des maladies [80].

I.B.1.2 Segmentation des populations pour personnaliser les actions

Les données ouvertes permettent aux organisations d'adapter leurs produits et services aux besoins de chaque segment de marché. Par exemple, de nombreuses entreprises commencent à appliquer la micro-segmentation des clients en temps réel pour guider les promotions personnalisées [27].

Il est possible par exemple, d'appliquer l'apprentissage automatique non supervisé[27] pour mieux segmenter la population d'un pays en particulier à partir des données démographiques accessibles à tout public issus des recensements et des autres études sociales (en France ce types des données sont disponibles au site de l'Institut national de la statistique et des études économiques). Le clusters ou groupes résultants peuvent être utilisés pour planifier une campagne électorale, identifier des besoins en matière de santé publique et de services sociaux ou pour créer des campagnes marketing ciblées [85, 92].

[26] Ce sujet est traité dans la section suivante *I.B.2 Les différents types d'analyses*.
[27] Ce sujet est abordé dans la section *I.B.2.1.2 Machine Learning et Deep Learning*.

I.B.1.3 Algorithmes d'aide à la décision

La data science et l'intelligence artificielle appliquées aux données ouvertes peuvent considérablement améliorer la prise de décision, minimiser les risques et découvrir des informations précieuses qui, autrement, resteraient inconnues. Ces analyses peuvent être appliquées, par exemple, pour affiner les entrepôts et les prix en réponse aux ventes en temps réel en magasin et en ligne [24].

Par exemple, DuPont propose le logiciel Pioneer Field360 Select, un outil en ligne qui combine des données ouvertes avec des informations agronomiques et météorologiques en temps réel pour aider les cultivateurs à prendre des décisions de gestion éclairées.
Ce logiciel s'appuie sur des données relatives au sol, à la météo et aux précipitations couvrant plusieurs décennies. Il s'appuie également sur le croisement de données de Pioneer Agronomy Sciences pour déterminer les stades de croissance des hybrides de maïs [77].

I.B.1.4 Nouveaux modèles, produits et services innovants

L'exploitation des données ouvertes permet aux entreprises de créer de nouveaux produits et services, d'améliorer ceux déjà existants et d'inventer de nouveaux modèles commerciaux [2, 11, 25, 81].
Par exemple, le site Doctrine (www.doctrine.fr) est un outil qui aide les avocats à construire des stratégies juridiques pour leurs clients. La plateforme agrège toute des données juridiques ouvertes disponibles, les enrichis avec la création de liens entre différentes informations (entre la chronologie des affaires, les décisions similaires, les tous les cas d'une entreprise ou d'un avocat, etc.) et les met à jour quotidiennement afin d'offrir aux utilisateurs une vu global et contextualisée dans une seul. Cette proposition de valeur se distingue d'une recherche juridique classique qui peut prendre beaucoup plus de temps et qui nécessite de consulter plusieurs sources [80].

Dans le domaine de la cartographie des données, il existe aussi une infinité de possibilités de création de business models[28] innovants basés sur l'utilisation d'informations géographiques et cadastrales.
La carte des ressources touristiques d'Euskadi Espagne[29] est un service à travers lequel il est possible de consulter de manière géolocalisée l'offre touristique de la ville, c'est-à-dire, les hébergements, restaurants, aéroports, gares, offices du tourisme, musées et monuments, installations sportives, tourisme thermal, etc. de façon gratuite.

L'organisation anglaise Open Opps (openopps.com) utilise les données ouvertes pour donner un aperçu des marchés publics dans le monde entier. Il publie des appels d'offres du monde entier de manière ouverte, afin que les clients puissent accéder à un monde d'opportunités.

[28] Le business model (ou modèle d'affaires) désigne généralement la façon dont un projet ou une activité doit générer des revenus.
[29] www.turismo.euskadi.eus/aa30-15524/es

I.B.2 Les différents types d'analyses

La sélection des analyses à appliquer aux données ouvertes récupérées sont fondamentales pour arriver à extraire de la valeur dans ces gigantesques ensembles d'informations. Il est possible de faire l'analogie entre une mine où se cache de l'or : pour le trouver et l'exploiter, il faut établir une stratégie bien précise, ainsi que choisir ou acheter les machines nécessaires pour y arriver.

I.B.2.1 Data mining (exploration de données)

L'exploration de données peut être définie comme l'application des analyses ou de processus d'extraction de connaissances à partir de grandes quantités de données. Ces techniques sont issues de différents domaines dépendant principalement de la statistique, mathématiques et de l'intelligence artificielle [57, 59].

Les principaux objectifs du data mining sont de trouver des structures originales et des corrélations informelles entre les données. Cela permet de mieux comprendre les liens entre des phénomènes en apparence distincts et d'anticiper des tendances encore peu discernables [25].

Avec la croissance des données disponibles, le data mining est né de la nécessité de Comprendre les informations utiles des bases de données ou Data warehouse (entrepôt de données)[30]. Dans le domaine des sciences et de l'ingénierie, il existe une large gamme de domaines d'application, par exemple en marketing, commerce, santé, sport, transport, météorologie [57].

Le processus du data mining se décompose en quatre étapes principales [59, 91] :

1. Collecte et consolidation : tout d'abord, les organisations collectent des données et les sauvegardent.
2. Filtrage et pré-traitement : les analystes commerciaux, les équipes de gestion et les professionnels techniques accèdent aux données et déterminent comment ils veulent les organiser. Ensuite, les spécialistes des données seront chargés de vérifier la pertinence des formats et de nettoyer ces informations.
3. Analyse ou fouille de données : cette étape est caractérisée pour la modélisation des algorithmes qui ont la finalité de générer de nouvelles connaissances et de la valeur issus des informations.
4. Visualisation et interprétation : les données sont distribuées dans un format facile à partager, et les résultats sont présentés dans graphiques ou dans un tableau.

[30] Cet élément est traité dans la section *II.C.3.1.2.3 Entrepôt de données*.

La figure suivante présente, à titre d'exemple, le processus général d'exploration des données appliqué aux séries chronologiques (des ensemble des observations d'une variable statistique ou un événement faites à intervalles réguliers, par exemple à l'année, au trimestre, par mois, par jour, comme les données météorologiques ou les fluctuations de prix des actions sur le marché boursier).

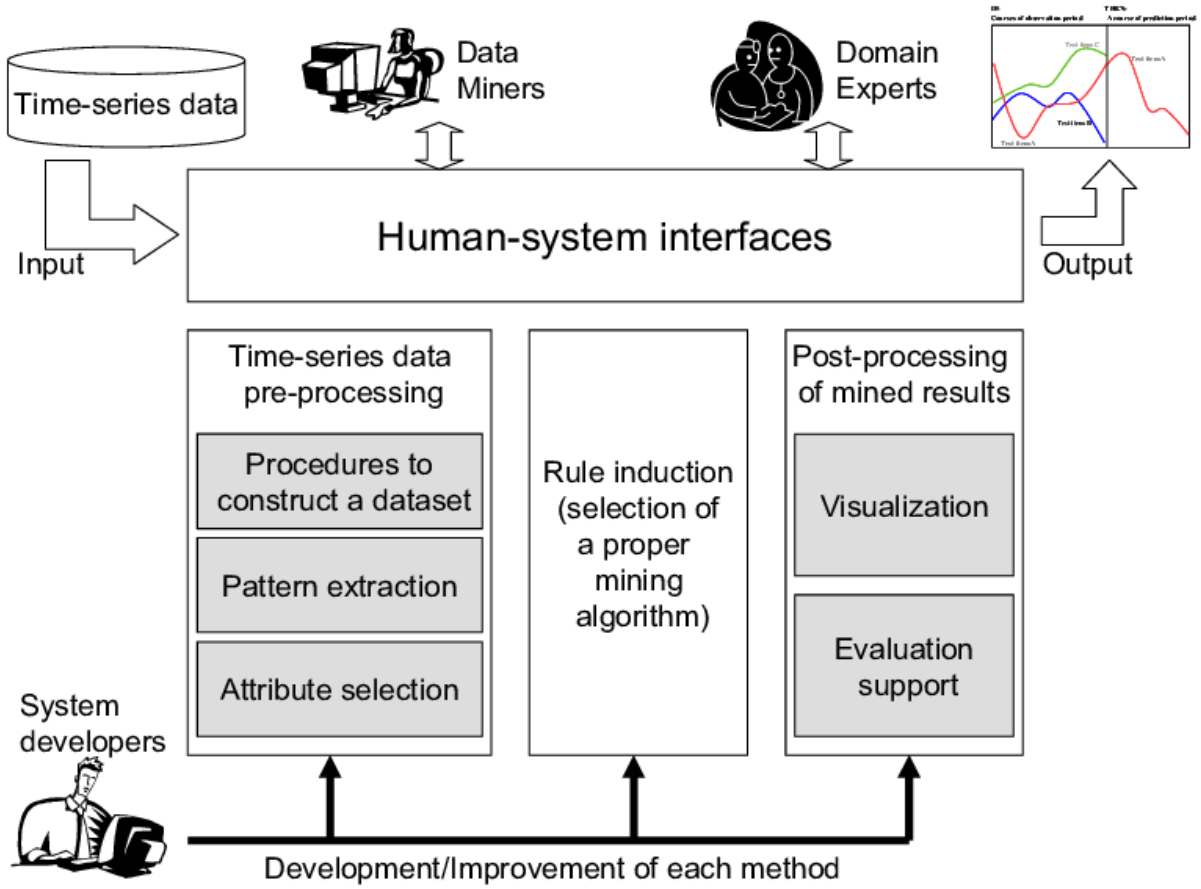


Figure 8 : Processus du data mining appliqué aux séries temporelles [86].

Pour la sélection d'une type analyse d'extraction des connaissances (valeur recherchée) appropriée, les alternatives disponibles sont nombreuses, néanmoins les deux plus répandus actuellement sont l'analyse relationnelle et l'apprentissage automatique [23, 25, 87]. Ces éléments sont détaillés ci-dessous.

I.B.2.1.1 Analyse relationnelle

Les requêtes relationnelles font référence à l'extraction de connaissances à partir d'un stockage relationnel. Ces requêtes prennent en compte les modèles relationnels, dans lesquels il existe une structure logique avec des relations ou des tables (ce sujet est traité à détail dans la section *II.C.3.1.1.1 Système de Gestion de Base de Données Relationnelles*.
Les relations ont un nom et sont formées par des attributs (colonnes) de données qui présentent un type particulier (entiers, flottants, caractères, etc.).

Pour la génération de requêtes relationnelles, le langage de préférence est SQL. Les opérations fondamentales de l'algèbre relationnelle sont : sélection, projection, produit cartésien, union et différence d'ensembles. Il existe également des opérations d'union, d'intersection et de division, qui sont exprimées en termes d'opérations de base suivantes : compte, somme, moyenne, min et max. Ces opérations sont connues sous le nom d'opérations de regroupement ou agrégation.

I.B.2.1.2 Machine Learning et Deep Learning

L'apprentissage automatique (machine Learning) est un domaine de recherche bien connu en informatique, qui consiste principalement à découvrir des modèles, des schémas et des régularités dans les données. L'apprentissage automatique peut être envisagé selon deux approches : le symbolique et le statistique.

L'approche symbolique, travaille sur l'apprentissage inductif de descriptions symboliques (par exemple, les arbres de décision), tandis que les seconds se concentrent sur les méthodes de reconnaissance des formes ou les statistiques.
Les deux approches peuvent être utilisées pour analyser des données quantitatives ainsi que des données textuelles. Ce dernier domaine d'étude est appelé Text Mining (fouille de textes). Le Text Mining consiste à transformer un texte non structuré[31] en données vectorielles ou structurées pour ensuite procéder à l'analyse. Cette pratique repose sur la technologie du Natural Language Processing (traitement naturel du langage), permettant aux machines de comprendre et de traiter le langage humain automatiquement.

Les algorithmes d'apprentissage automatique sont classés en deux catégories : supervisés et non supervisés [53, 60].

Le Machine Learning supervisé est un ensemble d'algorithmes qui permettent à l'ordinateur d'apprendre à prédire un résultat à partir d'un ensemble de données historiques. Le jeu de données doit inclure une variable dépendante aussi appelée variable Y. Il s'agit de la variable que l'ordinateur devra apprendre à prédire. Les autres variables sont les "prédicteurs", également appelés variables X et utilisées par l'ordinateur pour construire des modèles permettant de prédire Y. Quand Y est une variable qualitative (par exemple, gagnant ou perdant), il s'agit d'une problématique de classification. Quand Y est quantitative, il s'agit d'une problématique de régression [61].

À la différence de l'apprentissage supervisé, l'apprentissage non supervisé est celui où l'algorithme doit opérer à partir d'exemples non annotés. Les algorithmes non supervisés comprennent des méthodes permettant de cartographier des consommateurs, des clients, des patients, des échantillons environnementaux ou tout autre ensemble d'objets, ou de les classer en groupes d'objets similaires (partitionnement de données ou data clustering en anglais) [53]. Il n'y a pas de variable à prédire. Les outils courants de Machine Learning non supervisé comprennent notamment la classification k-means, la Classification Ascendante Hiérarchique, l'Analyse en Composantes Principales, etc. [53, 61].

[31] Voir section Annexes - Structure des données ouvertes.

Ces dernières années, l'utilisation du Machine Learning s'est rapidement répandue, avec des applications, par exemple, dans des domaines tels que la détection de la fraude, la détection du spam, les prédictions financières[32], entre autres [27, 53].
L'apprentissage profond (deep learning) est un terme général utilisé pour désigner une série d'architectures multicouches. L'une des principales caractéristiques est la modélisation d'abstractions de haut niveau par des méthodes non supervisées, dans le but d'obtenir une nouvelle représentation des données qui contribue à la tâche de prédiction [25, 26].

Cette approche a été appliquée avec succès dans des domaines tels que la réalité augmentée, la reconnaissance d'images, le traitement du langage naturel et la bioinformatique.
L'apprentissage profond est prometteur pour la modélisation et prévisions des données de séries chronologiques grâce à des techniques telles que les machines de Boltzmann restreintes (RBM), les RBM conditionnelles, les réseaux neuronaux récurrents, les réseaux neuronaux convolutifs, etc. [25].

Compte tenu de la grande variété des possibilités présentées de cet ensemble d'algorithmes, la difficulté de leur mise en œuvre réside dans la complexité des données étudiées. Des explorations complexes sont nécessaires pour comprendre les relations et les corrélations des phénomènes et pour trouver les techniques qui correspondent le mieux au domaine et permettent d'extraire des informations pertinentes.
En outre, l'application de chaque analyse nécessite la manipulation d'ensembles de données très particuliers et avec un périmètre bien travaillé (nettoyage des données[33]) [25, 26].

Dans le secteur de l'agriculture, par exemple, une plateforme de gestion agricole intelligente (smart agri-management platform[34]) a été développée par l'Institut taïwanais d'information industrielle. Cette solution offre un service de prévision des prix des cultures à partir de l'analyse des données ouvertes sur les prix historiques. Les algorithmes mis en place anticipent les prix et créent des tendances à partir de l'étude de ces séries chronologiques [76].

[32] Un exemple de ce type de prédiction appliqué aux séries chronologiques est présenté dans la section *III. Développement du produit prototype - Les prédictions.*
[33] Ce sujet a été détaillé dans la section *I.B.1 Nettoyage et préparation des données.*
[34] www.intelligentagri.com.tw/en

I.C Comprendre et prendre en compte le contexte socio-économique actuel à l'aide des données ouvertes

I.C.1 Analyse fondamentale et réseaux sociaux

L'analyse fondamentale est une méthode d'évaluation de produits financiers qui cherche à estimer le mieux possible la valeur intrinsèque d'une action, c'est-à-dire sa valeur réelle ou patrimoniale, en ne s'appuyant pas seulement sur l'analyse graphique de leur comportement. Ainsi, l'analyse fondamentale est la méthode qui a pour but de déterminer la valeur d'un titre financier en se focalisant sur les facteurs sous-jacents et "externes" qui influent l'activité présente et future de l'entreprise. Au sens contraire, l'analyse technique établit qu'il suffit de considérer le marché en tant que tel pour prévoir l'évolution des marchés.

Ce type d'analyse, qui cherche à comprendre la situation dans son ensemble, peut être utilisée dans l'analyse des données, et surtout dans le calcul des prédictions. Par exemple, aucun modèle de prévision et aucune IA n'avait prévu cette crise sanitaire et son impact dans l'économie mondiale. Donc, il est possible d'affirmer que la plupart des prévisions économiques de 2020, notamment les prévisions de croissance, se sont trompées, non pas à cause de leurs bases statistiques ou mathématiques, mais à cause d'un événement imprévu non considéré dans le modèle. "Le désordre est un ordre auquel on ne s'attend pas" disait Henri Bergson[35].

Cependant, il existe aujourd'hui des algorithmes complexes basés sur des exploitations similaires à l'analyse fondamentale qui tentent d'établir des corrélations entre des informations publiques provenant de centaines de sources différentes afin de prédire l'avenir à court terme [90]. Par exemple, au début de la crise, si un algorithme avait suivi le prix du pétrole et avait complété ses prévisions par une analyse des sentiments en faisant du scraping sur le réseau social Twitter [72], il était relativement facile de prédire que le prix allait baisser à court terme en raison du confinement. Un algorithme aurait pu prédire également la crise du secteur aérien et touristique, etc...

Il est donc possible d'imaginer des modèles et des systèmes de prévision plus généralisés capables de détecter, récupérer et interpréter des informations sur le web (journaux virtuels, réseaux sociaux) et sur la base d'analyses, issues du text mining par exemple et de réussir à établir des estimations pertinentes du futur.

L'application des méthodes d'analyse mentionnées auparavant est un élément fondamental pour la création de valeur économique et sociale issue de l'exploitation des données ouvertes. Néanmoins, il est nécessaire de prendre connaissance des principaux enjeux et défis de l'Open Data et les techniques pour les surmonter. Cela sera expliqué dans le chapitre suivant.

[35] www.wikipedia.org/wiki/Henri_Bergson

II. Les principaux problèmes lors de l'exploitation de données ouvertes et comment y faire face

L'exploitation des données ouvertes est un secteur qui peut faire évoluer la société, impulser l'innovation et inciter les entreprises et les entrepreneurs à créer des produits et services révolutionnaires. Néanmoins, il n'y a pas que des avantages dans le monde de l'Open Data. Si cette matière première semble être en principe gratuite, les ressources et connaissances nécessaires pour leur exploitation sont loin d'être négligeables.

Le chemin qui mène de la création de la stratégie d'exploitation et de la recherche de données à la création de produits prototypes peut être long et accidenté, rempli de défis et d'obstacles à franchir.
C'est pourquoi il est important de connaître les principaux problèmes potentiels et les solutions existantes à mettre en œuvre. De plus, ces solutions doivent être adaptées à nos ressources disponibles et à notre business plan (ce qu'on cherche à accomplir), afin de réduire les coûts, les efforts humains et le temps qu'implique la mise en œuvre de ces solutions.

Les principaux problèmes qui empêchent l'utilisation et exploitation des données ouvertes peuvent être classés comme suit :

- Problèmes de législation

  L'équilibre entre les réglementations juridiques, qui bloquent l'accessibilité et l'utilisation potentielle des données ouvertes, et l'absence de lois, qui ne protègent pas les producteurs et les utilisateurs de ces données, n'est pas toujours clair. C'est pourquoi une législation inadaptée constitue un des plus grands problèmes de l'exploitation des données ouvertes.

- Problèmes de publication

  Ces types de problèmes sont liés à l'absence d'une stratégie de publication qui devrait favoriser l'accès et le traitement des données ouvertes. Citons, par exemple : la difficulté à trouver l'information voulue, la fragmentation des publications et l'existence des jeux de données dupliquées, ainsi que le manque d'informations sur les données publiées (les métadonnées).

- Problèmes techniques

  Le fait d'obtenir des avantages économiques ou informationnels issus de l'analyse des données ouvertes, actuellement se limite aux connaissances techniques requises pour extraire, lire et traiter de grands volumes de données, ainsi qu'aux coûts liés à l'infrastructure nécessaire pour le faire.

- Problèmes de qualité

  La qualité de données est un grand domaine de recherche et un sujet qui intéresse les spécialistes des données, car c'est le dernier obstacle à franchir mais c'est loin d'être le plus facile. Le nettoyage et l'ensemble du processus de préparation des données ou data wrangling [36] est parfois la partie la plus longue (et la plus coûteuse en temps et en efforts) du processus de traitement des données ouvertes [75].

Ces problèmes sont détaillés ci-dessous, auxquels s'ajoutent quelques solutions ou techniques possibles qui peuvent aider à les franchir.

[36] Le Data Wrangling, est le processus qui permet à partir des données brutes de les découvrir, structurer, nettoyer, enrichir, valider et de publier les résultats dans un format adapté à l'analyse des données.

II.A La législation

Dans de nombreux pays, il existe déjà des initiatives de données ouvertes, mais dans la plupart des cas, elles n'ont pas un cadre juridique établi pour réglementer leur utilisation. Cela peut conduire à une diminution de la légitimité des origines des données, notamment en cas où ces informations aient été prises sans le consentement des personnes concernées.

Une réglementation faible peut représenter des menaces à la sécurité des utilisateurs. Par exemple, le risque potentiel de contamination du système informatique avec des malwares incrustés à la data.

De plus, les données à traiter peuvent contenir des informations qui violent la vie privée et le secret statistique[37]. Par conséquent, les utilisateurs, étant légalement responsables de l'exploitation, peuvent être exposés à d'éventuelles poursuites judiciaires. Par exemple l'ouverture des données de santé pose des questions liées au cadre juridique impliqué, car l'essor sans précédent de l'ouverture des nouvelles données implique aussi des possibilités croissantes de réidentification des personnes initialement concernées par le croisement de ces informations [88].

Par rapport à la confiance juridique et politique qu'il doit y avoir dans tout le cycle de vie des données et métadonnées ouvertes[38], Evelyn Ruppert, sociologue des données (data sociologist) et maître de conférences au Goldsmiths College de l'université de Londres, exprime l'opinion suivante [73] ; "Au-delà des questions politiques et éthiques sur le respect de la vie privée, la confidentialité et la protection des données, l'ouverture des données implique de repenser les relations avec le public dans la production de données statistiques si l'on veut que les citoyens leur fassent confiance. Nous esquissons une approche qui implique la co-production de données, avec des citoyens comme partenaires de la production statistique, de la conception de plateformes de production de données à leur interprétation et leur l'analyse. Si des questions sur la qualité et la fiabilité des données méritent d'être posées, nous estimons que la co-production a le potentiel d'atténuer les problèmes associés à la ré-utilisation de données massives à des fins statistiques. Dans ce contexte, nous estimons que l'avenir des statistiques publiques repose non seulement sur des données et des méthodes inédites, mais aussi sur la mobilisation des possibilités offertes par les technologies numériques pour établir de nouvelles relations avec le public."

En somme, puisqu'il s'agit d'information publique, les barrières juridiques devraient se concentrer sur la protection des données privées (RGPD), la propriété intellectuelle, la sécurité et le secret statistique [21], afin de ne pas obstruer la diffusion et la réutilisation de ces informations, mais en assurant la sécurité des utilisateurs et des producteurs de ces données.

---

[37] Le secret statistique est une forme particulière du secret professionnel qui s'applique aux organismes qui relèvent de la statistique publique. La finalité est de ne pas divulguer de l'information qui peut permettre d'identifier une personne ou un foyer en particulier, donc envahir à la vie personnelle et familiale. Pour en savoir plus voir la section *Glossaire.*

[38] Ce sujet est traité dans la section *II.C.3.1 Gestion du cycle de vie des données*.

II.A.1 Licence et question sur l' ouverture gratuite des données

Comme l'indique le huitième principe des données publiques, défini par l'Open Government Working Group [39] :

"Il est nécessaire de promouvoir l'utilisation d'une licence gratuite qui encourage la réutilisation, sans objet de droit d'auteur qui la limite (sauf dans les cas où des restrictions sont autorisées pour des raisons de sécurité, de confidentialité ou qui sont régies par une loi spécifique ou par une procédure administrative)."

Néanmoins, il est essentiel que toute donnée partagée ou ouverte comporte des informations explicites sur les conditions d'utilisation afin d'établir quelles sont les conditions de réutilisation de cette donnée et dans quel contexte elle peut être réutilisée [81].

Afin de trouver un équilibre, les entités publiques peuvent publier leurs données avec des licences simples. Dans ce cas, les données seront entièrement disponibles pour une réutilisation sous un certain nombre de conditions minimales, par exemple :

- citation obligatoire de la source des données.
- la date de la dernière mise à jour doit être mentionnée.
- il ne peut être fait allusion au fait que les propriétaires des données participent à leur réutilisation, la parrainent ou la soutiennent.
- les métadonnées doivent être maintenues et non modifiées, ainsi que les conditions de réutilisation applicables.

D'un autre côté, les licences payantes sont généralement établies pour les entreprises ou institutions privées qui cherchent à monétiser leur travail et faire de l'ouverture de leurs données une source de revenus.

Par rapport aux modèles de tarification, les données peuvent être vendues selon :

- leur volume. Par exemple, un accès gratuit jusqu'à un nombre de requêtes par mois avec l'utilisation d'une API[40] publique et un système premium offrant un accès privilégié.
- les prix peuvent varier en fonction des types de données (granularité, fraîcheur).
- selon un système hybride volume/type de données et même des solutions d'abonnements [2, 11].

Une autre option est la vente de bases de données en particulier dans le domaine marketing et le secteur commercial, ainsi que dans le marché des informations financières [11].

[39] Ce sujet a été traité dans la section *I.A.1.1 Ouverture des données publiques*.
[40] Cet élément est traité dans la section *II.B.5.2.2 API.*

Est-ce que la gratuité des données justifie une moindre qualité par rapport aux données payantes ? Il pourrait sembler logique de déduire que les données payantes, (qui sont la base des revenus pour de nombreuses entreprises) soient de meilleure qualité que les données ouvertes, ou de justifier que les institutions publiques qui ouvrent leurs données doivent s'ajuster à un budget limité, et comme il n'y pas un ROI (retour sur l'investissement) direct, elles sont autorisées à rendre ces données publiques avec une qualité inférieure.

Cette question est difficile à répondre. Un oui ou un non ne suffira sûrement pas, car il doit y avoir des institutions publiques qui offrent des données d'une qualité enviable pour plusieurs motifs (une longue histoire d'ouverture, la nature et la façon de recueillir l'information, etc.). Néanmoins, selon certains exploitants de données, il semble beaucoup plus intéressant d'avoir des informations disponibles, même avec certains problèmes, que de ne pas les avoir ou de les avoir trop en retard (données obsolètes) [2, 11, 80,81].

Les entreprises, entrepreneurs, journalistes et des associations aujourd'hui développent des méthodes, connaissances et des stratégies techniques et statistiques de plus en plus performantes afin de pouvoir exploiter le volume le plus grand possible des données dans leur état brut (et avec les défauts que cela peut impliquer).

Ce sujet est traité dans la section *II.D Qualité des données ouvertes.*

Cependant, afin de diminuer les coûts et le temps dédié au data munging (processus de traitement de la donnée qui écarte les erreurs et rend les informations pleinement exploitables), il est nécessaire d'exiger une amélioration continue de la qualité de la part de toutes les institutions productrices de données publiques, qu'elles soient publiques ou privées [82].

II.B Problèmes de publication

II.B.1 Inadéquation entre la demande et l'offre des données ouvertes

La figure suivante montre la comparaison des catégories de données les plus populaires avec les ensembles de données disponibles sur le Portail européen de données (data.europa.eu).

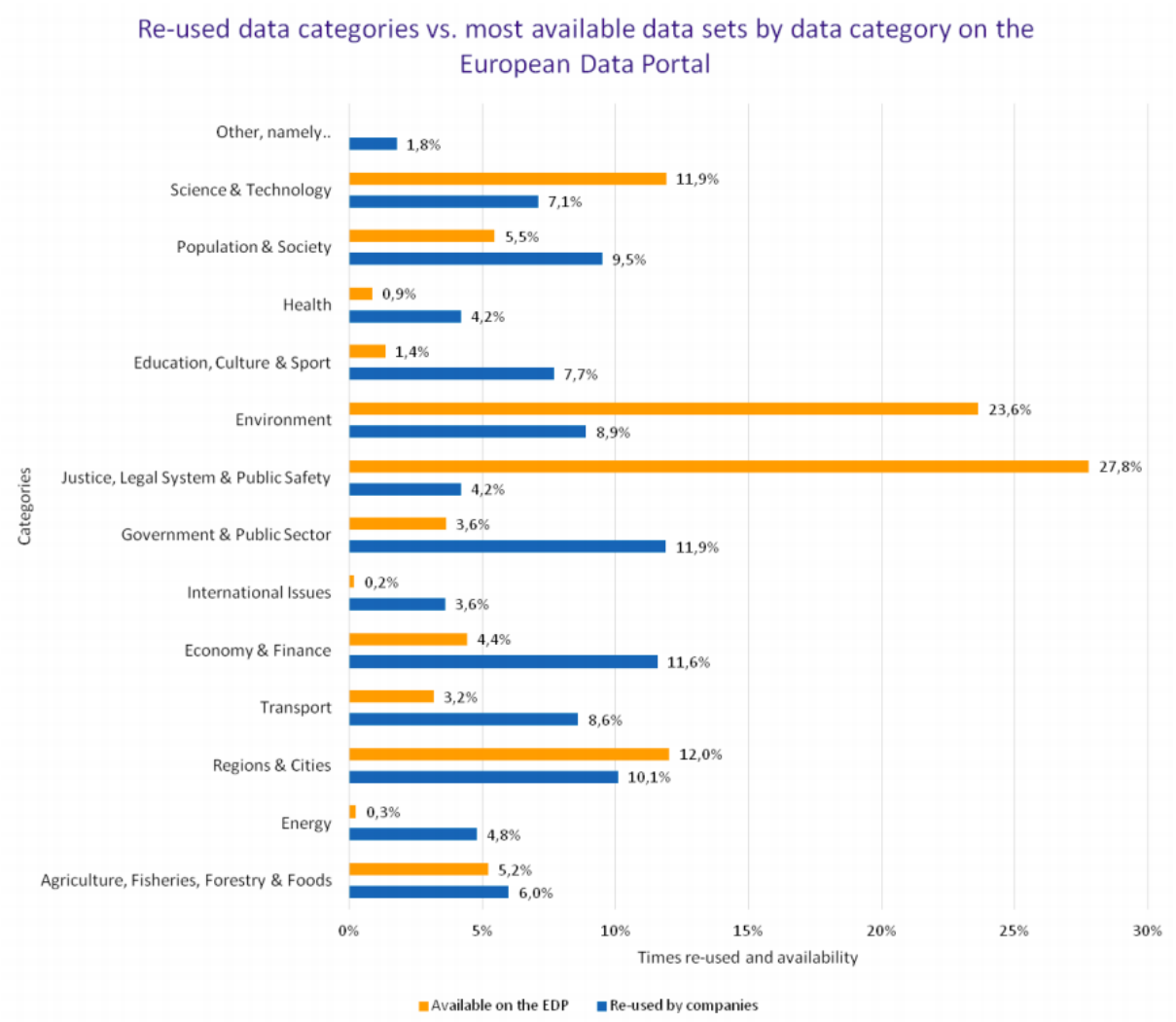


Figure 9 : Catégories de données réutilisées par rapport aux ensembles de données les plus disponibles sur le portail européen de données.

Les catégories de données les plus utilisées pour créer de nouvelles sources de revenus, par rapport aux 13 catégories de données ouvertes définies par Eurostat, sont: gouvernement et secteur public (11,9%), économie et finances (11,6%), régions et villes (10,1%), population et société (9,5%), et environnement (8,9%). Ces cinq catégories de données représentent 52% de la réutilisation totale des données ouvertes par les répondants.
Inversement, les catégories de données les moins utilisées sont les questions internationales (3,6%), la santé (4,2%) et la justice, le système juridique et la sécurité publique (4,2%) mais c'est le domaine où l'on dispose de plus de données.

Une solution à l'inadéquation entre les ensembles de données disponibles et les données utilisées pourrait être que les administrations publiques développent une stratégie de publication qui soit plus en phase avec les besoins des utilisateurs de données [80].

Une étude des besoins secteur par secteur doit donc être menée, car différents ensembles de données sont utiles pour différents secteurs sociaux. Autrement dit, les gouvernements et les acteurs économiques devraient travailler ensemble pour identifier les priorités sociales les plus urgentes et les besoins en données qui sont liés en termes de priorités.

II.B.2 Fragmentation des publications de données

En ce qui concerne la publication des données, l'un des principaux problèmes c'est la fragmentation des publications de données. c'est-à-dire que des données qui appartiennent à la même classe ou sujet sont souvent publiées dans plusieurs sources différentes. A cause de cela, les utilisateurs ont des difficultés à trouver les données dans leur ensemble.

Par exemple, dans le cas de la publication d'appels d'offres publics, bien qu'il existe des sites dédiés à la centralisation de ces informations, comme la plateforme BOAMP (Bulletin officiel des annonces de marchés publics[41]) pour la France, de nombreuses communes publient des annonces "exclusives" sur leurs sites.
Par conséquent, pour de nombreuses entreprises qui n'ont pas une équipe dédiée uniquement à la veille des appels d'offres, il est facile de passer à côté d'opportunités potentielles s'ils regardent uniquement le BOAMP.

D'un autre côté, il arrive aussi que plusieurs jeux de données ouvertes soient dupliquées, c'est-à-dire présentes dans différentes sources ou sites, ou dans la même source mais avec un autre nom [20].

Enfin, un autre grand défi est le simple fait de trouver les informations voulues, car elles ne sont pas faciles à trouver dans de nombreux cas.

II.B.3 Manque de métadonnées

Comme indiqué par le portail des données ouvertes du Québec : "les métadonnées sont les données qui accompagnent les jeux de données afin de décrire leur contenu et leur contexte. Elles sont un complément essentiel aux jeux de données ouvertes, car elles permettent de structurer la définition de chaque variable des jeux de données." [42]

Un catalogue de données est un emplacement centralisé où sont regroupées les jeux de données d'un sujet en particulier et les informations qui les décrivent en détail : les métadonnées.

Malheureusement, il est souvent difficile de trouver ces informations relatives aux données publiées. Par conséquent, en l'absence de métadonnées, les utilisateurs ont du mal à les interpréter, et cela rend difficile leur compréhension car ils ne connaissent pas les définitions et les catégories utilisées, ce qui complique leur comparaison, leur liaison et la réutilisation des informations.

[41] www.boamp.fr
[42] www.donneesquebec.ca/lignes-directrices-sur-la-diffusion-de-donnees-ouvertes

II.B.4 Le format des données et leur potentiel d'exploitation

Sur de nombreux portails des données ouvertes, il existe des informations statistiques d'intérêt ou des données pertinentes disponibles dans divers formats selon leur nature. Ces données, dans le meilleur des cas, sont dynamiques, générées en temps réel. Une équipe de personnes peut aussi les générer et les publier périodiquement.

Ces données peuvent être présentées dans différents formats, ce qui permettra ou non de les réutiliser plus facilement pour construire des produits et services. Plus les données sont structurées et enrichies, plus il sera facile de les réutiliser et de créer des applications qui les gèrent automatiquement.

Parmi les formats de données les plus appropriés pour la réutilisation, il existe : XML, JSON, Excel (XLS), CSV, ainsi que le modèle RDF (qui permet la requête de données en utilisant le langage SPARQL). Ils permettent de créer des services à haute valeur ajoutée [10].

Ces éléments sont détaillés dans la section *Annexes - Formats des données ouvertes.*

Tim Berners-Lee, l'inventeur du World Wide Web (WWW), du HTML et qui en 2012 était chargé de s'assurer de la disponibilité sans discrimination des données publiques ouvertes au sein de l'Open Data Institute (ODI) à Londres, a classé les formats des ensembles de données et leur donne une note avec un nombre d'étoiles compris entre 1 et 5. [11]

Stars	Characterization	Example format
*	Data online under an open license	PDF
**	As * and structured data	Excel
***	As ** and nonproprietary formats	CSV
****	As ***, using URIs as identifiers	RDF
*****	As ****, linked to other data	RDF

Figure 10 : Échelle de qualité des données ouvertes de Tim Berners-Lee.

La figure 10 montre une échelle qui va d'un format de mauvaise qualité en référence au potentiel d'exploitation de l'information (une étoile), et qui remonte jusqu'à 5 (cela représente une excellente qualité pour la réutilisation).

Au premier niveau il y a des ensembles de données non structurés[43] pouvant être consultés sur le Web, mais non traités automatiquement. Par exemple, une image au format JPG ou PNG, ou un document numérisé au format PDF.

Le niveau augmente progressivement et avec 4 étoiles, il y a des données pouvant être référencées par des adresses Web ou des URI (Uniform Resource Identifier).
Les formats utilisés sont de type standard et ouverts du W3C (RDF et SPARQL) , afin de décrire sémantiquement les informations.

[43] Voir section Annexes - Structure des données ouvertes.

Par exemple, une représentation RDF des bâtiments de l'État, avec leurs informations de contact et leur localisation. Toutes ces données doivent être accessibles via des adresses Web (URI).

Dernièrement, il existe les données liées à d'autres ensembles des informations externes qui fournissent un contexte à la donnée. Pour faire cela, des relations sémantiques sont établies entre les informations liées, permettant une réutilisation ultérieure beaucoup plus efficace.

Un exemple de ceci serait la description de l'emplacement des bâtiments publics enrichie de liens vers Geonames[44], qui est une grande base de données spécialisée dans les emplacements géographiques. Grâce à ces liens, il serait possible d'inclure une description détaillée des localités, régions ou pays et ainsi avoir un accès direct aux informations socio-économiques de ces lieux.

En résumé, l'essence des données ou des informations partagées, c'est qu'elles peuvent être mélangées ou croisées avec une autre source. Cette interopérabilité est absolument fondamentale pour tirer profit de l'ouverture des données : l'augmentation de la capacité à combiner différentes bases de données ou de jeux de données, ainsi que le développement de produits et services plus nombreux et de meilleure qualité.

II.B.5 Stratégies de gestion des problèmes de publication des données ouvertes

II.B.5.1 Veille

La veille peut être définie comme l'ensemble des actions de recherche, d'analyse, de diffusion, d'utilisation et de protection des données utiles aux différents acteurs économiques (entreprises, entrepreneurs, associations, etc.) [47]. Ces actions permettent à une entreprise ou à une organisation de disposer de la bonne information au bon moment pour être en mesure de prendre des décisions stratégiques en connaissance de l'environnement interne et surtout externe.
C'est pour cela que la veille stratégique (la recherche des données qui peuvent aider à la prise des décisions stratégique) devrait être intégrée dans le fonctionnement normal des organisations, et elle devrait être systématique, régulière et méthodique [46].

L'activité la plus connue de ce processus est la recherche d'informations, mais si les besoins ou les objectifs ne sont pas clairement définis en amont, les recherches apporteront de la matière inutilisable, sans valeur ajoutée. Pour cela, avant de commencer à collecter des données, il est indispensable de :

- spécifier les objectifs de l'utilisation des données.
- définir une stratégie à suivre : basée soit sur le big data, soit sur le smart data (ce sujet était traité dans la section *I.A.2 Ensembles données massives, ou mégadonnées)*.

[44] geonames.org

- délimiter les sujets liés à nos problématiques ou objectifs, afin d'affiner notre recherche.
- cibler des potentielles sources à exploiter [46, 48].

Ces aspects sont la base du cycle de la veille [48]. Ce cycle comprend les étapes de surveillance de l'environnement selon des objectifs définis, mais aussi l'intégration des données collectées pour les transformer en véritables informations stratégiques et opérationnelles, comme le montre la figure suivante :

Figure 11 : Le cycle de la veille.

Chacune des étapes doivent être adaptées à notre stratégie de veille économique (parfois appelée intelligence économique) [46, 48]. Le tableau suivant présente les étapes de veille économique en détail.

Etape	Description
Expression du besoin	La détection des besoins est basée sur l'étude des projets en cours, des analyses benchmarking, de la surveillance de l'environnement concurrentiel, de l'anticipation de risques stratégiques, etc.
Recherche et collecte de l'information	Mise en œuvre de techniques et outils de veille pour collecter l'information pertinente. Choix des sources d'information (réseaux sociaux, presse en ligne, appels d'offres, etc.). Cette étape comprend aussi la compréhension des risques et des menaces (comme le non-respect des droits d'auteurs ou le piratage, l'espionnage, etc.) et le plan d'action pour les éviter.
Validation et traitement de l'information	Cette étape consiste à valider la véracité des données et à les transformer en informations utiles. Elle comprend une analyse, un enrichissement et une mise en forme des données pour pouvoir les exploiter (ce sujet sera traité dans la section *II.D.1 Stratégies de gestion des problèmes de qualité des données ouvertes*).

Diffusion aux acteurs économiques	Choix des outils et méthodes pour diffuser l'information collectée aux acteurs concernés.
Processus de décision organisationnel	Prise de décisions managériales et de décisions stratégiques, changement de la stratégie d'influence [45], etc.
Mise à jour du système ou de l'information	Répétition des premières 3 étapes du cycle afin de faire la mise à jour des besoins, des nouvelles sources et des renseignements, ainsi que la coordination des acteurs et des activités liées à la stratégie de veille.

Figure 12 : Les étapes de la veille économique.

Le veille stratégique peut contribuer donc à simplifier et standardiser la recherche des données ouvertes, une tâche qui est généralement complexe à cause de :

- la disponibilité limitée de certaines informations, par exemple les données de santé.
- ou inversement, à cause de la masse croissante des informations disponibles. Par exemple dans la recherche d'un sujet très précis du domaine juridique (le domaine plus abondant des données ouvertes).

II.B.5.2 Collecte des données

On peut globalement distinguer deux types de collecte des données : la collecte externe et la collecte interne.

Les exemples de collecte interne des données au secteur privé (généralement non ouverts au public) peuvent être une base de données clients ou les données générées par des automates[46] [46]. Ces données peuvent être quantitatives, comme les capacités du parc machines, la situation financière et la trésorerie, le carnet de commandes, etc. ou qualitatives comme la situation de la ressource humaine, les relations avec les clients, etc.

Les collectes externes, pour leur part, portent sur les données ouvertes qui peuvent être d'intérêt pour les entreprises, par exemple les données ouvertes financières des organisations concurrentes ou les données juridiques sur une réglementation en particulier. Cette collecte comprend la récupération des informations :

- d'origine gouvernementale (généralement des données macroéconomiques, par exemple l'indice de la production industrielle[47]).
- issues du secteur éducatif et de la recherche.
- des autres entreprises.

[45] Il s'agit de la capacité à maîtriser son environnement : relationnel, lobbying, etc.
[46] Dispositif électronique programmable, destiné à la commande de processus industriels par un traitement séquentiel.
[47] www.insee.fr/fr/metadonnees/source/indicateur/p1646/description

Le secteur public et des institutions éducatives ou de recherche, peuvent à leur tour enrichir leurs bases des connaissances avec l'exploitation des données externes issus du secteur privé, comme par exemple les données sur les ventes des produits et services par secteur économique.

Comme indiqué plus haut, peu d'entreprises ouvrent réellement leurs données, alors où est-il possible de trouver ces informations ? L'État a notamment accès à tous les bilans financiers des entreprises, mais le grand public peut consulter aussi des données sur les sociétés cotées en bourse qui peuvent être très intéressantes. Il est important de mentionner que ces sociétés sont obligées de publier tous les informations annuelles et semestrielles significatives ou susceptibles d'avoir un impact sur le cours de bourse[48].

Il est très important de bien choisir les sources à partir desquelles les informations vont être collectées, car c'est la base de tout le processus de la gestion des connaissances. Puis, décider quelles sources d'information sont les plus utiles et les plus significatives, ainsi que valider leur fiabilité, par exemple en termes d'actualité ou d'autorité. Cela peut contribuer à réduire le périmètre de recherche, les efforts dédiés à la recherche et les coûts impliqués.

La collecte d'information peut être réalisée en visitant directement le site de la source, et en téléchargeant les informations souhaitées. Par exemple, le site des statistiques officielles du commerce international des Nations Unies (Comtrade), permet de télécharger différentes ensembles de données au format csv en un clic (figure 13).

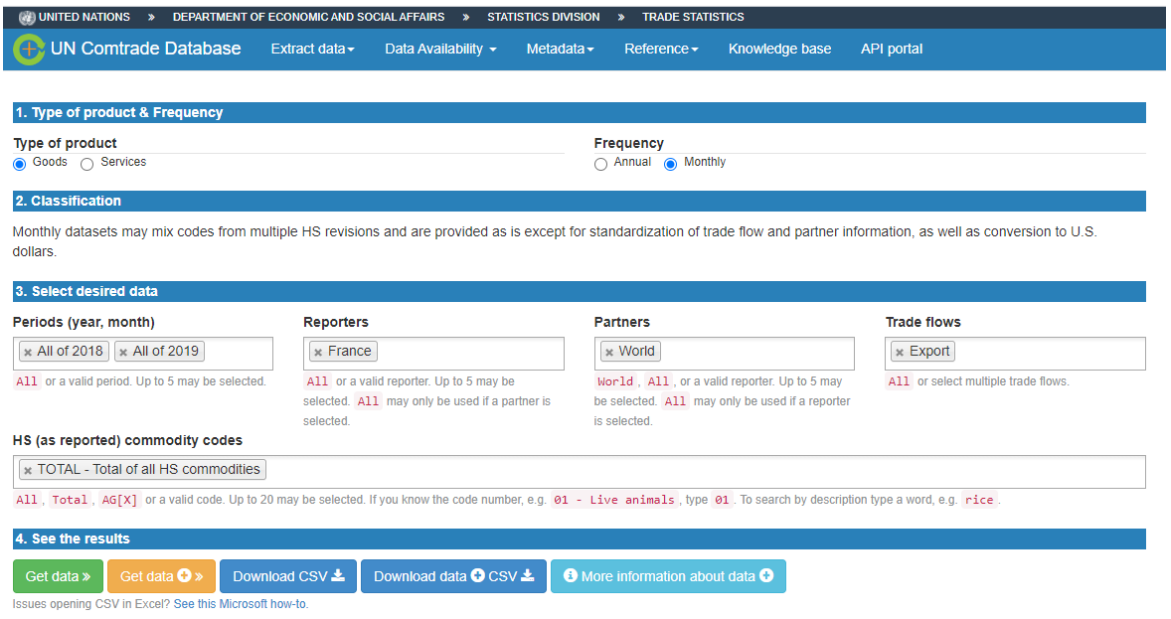


Figure 13 : Téléchargement manuel des données.

II.B.5.2.1 Collecte via URL

Il est possible de collecter des données ouvertes en reliant directement le système d'exploitation de données aux sources ciblées avec les URL de téléchargement de chaque jeux de données.

[48] www.amf-france.org/fr/actualites-publications/dossiers-thematiques/info-periodique-et-permanente

Par exemple, une société dédiée à la gestion des déchets agricoles peut récupérer directement l'historique de la production mensuelle de biométhane dans le site *data.gouv.fr* (le site de données publiques ouvertes en France) dans le but de valider une stratégie de diversification ou incursion au marché du biogaz (en donnant une nouvelle valeur à certains déchets).

Un autre exemple serait le cas d'un investisseur qui veut développer une startup de puces géolocalisables pour aider les personnes à trouver les objets qui se perdent le plus souvent dans la vie quotidienne.
Il pourra consulter la base des objets trouvés de la SNCF (Société nationale des chemins de fer français), en faisant une mise à jour quotidiennement avec l'aide de la bibliothèque Python *Threading*[49] ou en utilisant le programme cron[50]. Un programme qui utilise la bibliothèque Threading est présenté dans la section A*nnexes - Codes.*

II.B.5.2.2 API

Les interfaces de programmation d'applications (API) sont l'un des services d'échange d'informations et d'accès aux données les plus courants aujourd'hui. Dans le contexte de l'Open Data, le terme fait généralement référence aux API Web, appelées dans certains domaines "Web API", qui sont un moyen courant de tenir l'échange d'informations. Ce type d'API offre un ensemble de fonctionnalités sur un serveur Web qui peuvent être utilisées par des applications clientes grâce à l'utilisation de procédures standard [24].

Même s'il existe différentes alternatives dans les modèles architecturaux qui inspirent la conception des API, le modèle d'API REST sur HTTP est le type d'API le plus populaire et le plus répandu [49].

Une caractéristique de ce modèle est l'utilisation de normes ouvertes telles que HTTP/HTTPS, qui ne lient pas les implémentations d'API sur le serveur ou les applications client à une implémentation spécifique, c'est-à-dire que les deux composants peuvent être implémentés à l'aide de différents langages de programmation, tant qu'ils peuvent formuler des demandes et comprendre les réponses à l'aide du protocole HTTP [49].

Les API REST sont conçues pour exposer et interagir avec des ressources qui sont des objets, des données ou des services auxquels un client peut accéder. Les APIs peuvent augmenter la disponibilité des données en donnant accès à fichiers téléchargeables et sont un moyen essentiel pour l'accès et l'exploitation des données. Il s'agit d'un mécanisme d'accès idéal pour publier des données avec une fréquence de mise à jour élevée, telles que des données en temps réel ou dynamiques [24, 49].

[49] https://docs.python.org/3/library/threading.html
[50] https://doc.ubuntu-fr.org/cron

Comme il était mentionné dans l'exemple de la section *II.B.5.2 Collecte des données*, le site Comtrade offre aussi une API gratuite[51] pour faciliter l'accès aux données. Le URL de l'API est :

http://comtrade.un.org/api/refs/da/view?<<parameters>>

Pour l'utiliser, il est nécessaire d'ajouter certains paramètres. Par exemple, pour spécifier le type de marchandise cherchée il est nécessaire d'ajouter à l'URL le symbole "&" suivi, soit d'un C si nous cherchons de l'informations des transactions des produits commercialisés (commodities), soit d'un S pour des services. Il faut aussi délimiter la granularité temporelle (mois ou l'année) et la date des enregistrements composée du nombre de l'année et du mois. S'il est souhaité de consulter les données des produits commercialisées dans le monde en octobre du 2019 il est nécessaire de saisir :

https://comtrade.un.org/api/refs/da/view?type=C&freq=M&ps=201910

Comme pour de nombreuses API publiques, il existe des limites ou des restrictions. Il s'agit d'éviter les abus, afin de ne pas surcharger le système de transfert des données. Pour le cas de l'API de Comtrade par exemple, les limites sont :

- Limite de débit : 1 demande par seconde (par adresse IP ou utilisateur authentifié).
- Limite d'utilisation : 100 demandes par heure (par adresse IP ou utilisateur authentifié).
- Limite de combinaison des paramètres : par exemple le choix de pays, années ou le code des marchandises et limite à 5 et un seul de ces éléments peut être choisi dans sa totalité. Par exemple, il est possible de consulter tout l'historique des enregistrements des transactions commerciales de la France (exportations et importations de tout type de marchandises) avec le Mexique, la Colombie et l'Espagne.

Si la limite d'utilisation est atteinte , une erreur 409 (conflit) est renvoyée avec un message précisant pourquoi la demande a été bloquée et quand les demandes peuvent reprendre.

[51] www.comtrade.un.org/Data/doc/api

II.B.5.2.3 Web scraping

Le scraping ou l'extraction automatique de données, est une technique permettant d'extraire du contenu (des informations) d'un ou de plusieurs sites web de manière totalement automatique, afin de les enregistrer et les analyser. Ce sont des scripts ou des programmes informatiques qui sont chargés de collecter ces informations [50].

Le web scraping permet en effet de créer de grands ensembles de données massives avec des dizaines de milliers de variables, mais il peut aussi être utilisé pour créer des ensembles de données de taille modeste, plus faciles à gérer de la même façon qu'une API.
La différence entre ces deux techniques de collecte des données est que le scraping peut permettre de collecter des données ouvertes de façon plus dynamique et en réduisant des efforts, du temps et des coûts aux agences économiques [51].

Dans la section *Annexes - Codes - Spider DILA*, un programme Python démontre comment les API Webs et le scrapping peuvent être utilisés ensemble pour collecter des données issus de la Direction de l'information légale et administrative - DILA (echanges.dila.gouv.fr/OPENDATA) qui assure la publication des lois et décrets au Journal Officiel. Le site DILA garantit également la transparence économique et financière par la publication, au niveau national, de l'ensemble des informations légales, économiques et financières relatives à la vie des entreprises et au milieu associatif. Cet algorithme basé sur le framework Scrapy[52] récupère toutes les pièces jointes 'pdf' de chaque dossier, c'est-à-dire la description de chaque base de données associée à DILA, par exemple le BODACC (Bulletin officiel des annonces civiles et commerciales).
Cela peut être utile si un acteur économique souhaite trouver des sources d'appels d'offres. Pour le faire, tous les fichiers pdf peuvent être convertis en texte, puis des recherches traditionnelles par mots-clés peuvent être effectuées afin d'identifier les sources comprenant, par exemple, le mot "marchés publics".

Il est important de mentionner qu'il est fortement suggéré de ne pas négliger la légalité et l'éthique de l'utilisation de ces outils. Par exemple, certains sites stipulent explicitement qu'il est interdit de collecter leurs données. [50].

[52] www.docs.scrapy.org/en/latest/intro/install.html

II.C Problèmes techniques

II.C.1 Fracture numérique

Dans le domaine technique, il existe des barrières comme le manque de soutien aux utilisateurs inexpérimentés, ainsi que le manque d'infrastructure technologique pour l'hébergement, le téléchargement, la disposition et l'utilisation des données.
Par conséquent, puisque tous les utilisateurs ne sont pas en mesure d'utiliser ces données, il existe un risque de se limiter à certains groupes et cela contribuerait à la fracture numérique et aux inégalités sociales.

Il faut donc prendre en compte toutes les capacités et les niveaux de connaissances des utilisateurs qui voudraient exploiter des données complexes et sophistiquées. En effet, le fait de publier des ensembles de données ne sert à rien si les utilisateurs qui pourraient bénéficier de leur réutilisation ne disposent pas des outils pour en profiter.

Pour faire face à cette barrière, des formations et des canaux d'assistance doivent être fournis pour la société civile, pour les entreprises et pour les organisations. [20]

II.C.2 Coûts élevés liés aux infrastructures technologiques

En premier lieu, l'organisation des données est généralement une tâche complexe et coûteuse à réaliser, qui peut consommer une grande partie des ressources informatiques et humaines. Cette organisation des données est définie comme le fait de :

- délimiter les informations à héberger avec précision : prévoir l'archivage des données brutes et des données traitées.
- prévoir la conservation des données : périodes d'observations, interruptions, etc.
- organiser a priori, et non a posteriori, la constitution informatique du catalogue des données (ou un inventaire des données) et faciliter l'utilisation ou l'intégration des nouvelles données en introduisant des champs supplémentaires pour sélectionner et identifier ces observations.
- établir et tenir à jour des grilles des données pour retrouver les données rapidement.

Même avant de créer une stratégie d'organisation des données, il faut structurer l'architecture des données. L'architecture de données est le processus qui permet de standardiser la façon dont les acteurs économiques collectent, stockent, transforment, distribuent et utilisent les données.
Le chargé de ce processus doit définir la vision des données en traduisant les exigences business en exigences techniques (structures), ainsi que les normes pour les modèles de données, les métadonnées et les flux de données qui déterminent qui dans l'organisation génère des données, qui les utilise et comment les flux de données sont gérés [79].

Ainsi, même si les données sont gratuites, le coût du stockage et du traitement peut devenir hors de portée pour de nombreux entrepreneurs et de petites entreprises.

Néanmoins, ce coût élevé, lié aux infrastructures et au personnel spécialisé dans l'architecture des données, doit être considéré dans son intégralité, c'est-à-dire après avoir effectué l'analyse de l'impact positif dans l'organisation ou par un retour sur investissement.

II.C.3 Stratégies de gestion des problèmes techniques lors de l'exploitation des données ouvertes

II.C.3.1 Gestion du cycle de vie des données

Le terme de gestion du cycle de vie des données ou DLM (Data life cycle management) désigne la gestion du flux de données d'un système informatique tout au long de son cycle de vie, de la création des données à leur suppression [45, 74]. Ce cycle est montré dans la figure suivante *Cycle de vie des données*.

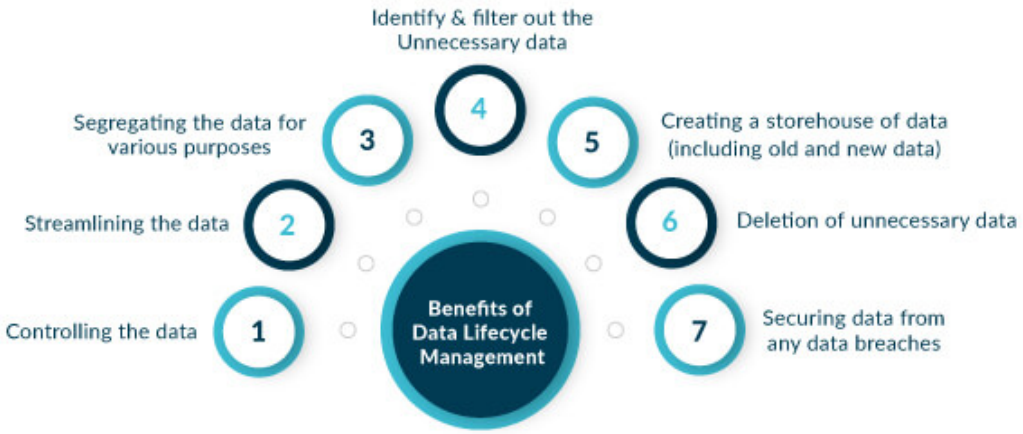


Figure 14 : Avantages d'un correct management du cycle de vie des données [53].

Un cycle de vie des données efficace permet d'améliorer les performances des systèmes informatiques et de réduire les coûts de stockage[74]. Par exemple, les données plus récentes et les données auxquelles il est nécessaire d'accéder fréquemment sont sauvegardées sur des supports de stockage plus rapides, tandis que les données plus anciennes sont hébergées sur des supports moins performants (et moins chers).

Comme suggéré par la CNIL (Commission Nationale de l'Informatique et des Libertés)[54], une stratégie de gestion du cycle de vie des données, appliqué aux données en général, peut être le fait d'utiliser 3 niveaux de stockage afin d'assurer la base des données et diminuer les coûts :

- Conservation en base active : il s'agit de sauvegarder les données pendant la durée nécessaire à la réalisation de nos analyses. En pratique, les données seront alors facilement accessibles dans l'environnement de travail immédiat pour les services opérationnels qui sont chargés de ce traitement.

[53] https://www.audienceplay.com/blog/what-is-data-lifecycle/
[54] https://www.cnil.fr/fr/les-durees-de-conservation-des-donnees

- Archivage intermédiaire : les données ne sont plus utilisées pour atteindre l'objectif , mais présentent encore un intérêt administratif pour l'organisme ou doivent être conservées pour répondre à une obligation légale (par exemple, les données de facturation doivent être conservées dix ans en application du Code de commerce, même si la personne concernée n'est plus cliente). Les données peuvent alors être consultées de manière ponctuelle et motivée par des personnes spécifiquement habilitées.
- Archivage définitif : en raison de leur valeur et intérêt, certaines informations sont archivées de manière définitive et pérenne.

II.C.3.1.1 Stockage de données

L'explosion des données ces dernières années pose d'énormes défis en matière de collecte et de stockage, ce qui a un impact sur la gestion et l'analyse des données. Toutefois il existe différentes technologies pour héberger de grands volumes de données, qui ont leurs avantages et leurs limites, allant des modèles les plus traditionnels (relationnels) aux nouvelles tendances qui incluent la possibilité de sauvegarder des données non structurées ou semi-structurées.

Ces éléments sont présentés dans la section Annexes - Structure des données ouvertes.

II.C.3.1.1.1 Système de Gestion de Base de Données Relationnelles

Un système de gestion de base de données relationnel (SGBDR) est un logiciel permettant de gérer une base de données relationnelle. Par exemple : Oracle, MySQL, Neo4j [56]

Ce type de bases de données structurées permet des connexions ou des relations entre les enregistrements qui sont contenus dans des tables. Une table représente l'espace de stockage d'une collection de mêmes objets.

Une table est constituée d'attributs. Ces attributs sont déterminés par un type de données (chaîne de caractère, numérique, date, etc.) qui dépend de la base de données [45]. Ces tables présentent un certain nombre de caractéristiques particulières, parmi lesquelles :

- Il n'y a pas deux tables avec le même nom.
- Une clef primaire désigne une colonne (ou un ensemble de colonnes) dont les valeurs sont uniques au sein de la table.
- Une clef étrangère désigne une colonne dont les valeurs font référence à la clef primaire d'une autre table.

Pour manipuler les données contenues dans ces tables, l'algèbre relationnelle et le calcul relationnel sont utilisés. Également, le SQL (Structured Query Language) est un langage commun d'interrogation des bases de données relationnelles.

Dans un SGBDR, le terme "transaction" désigne les opérations apportant des modifications aux données.

Les propriétés d’un SGBDR visent à garantir la validité du système, même en cas d’erreur ou de panne informatique. Andreas Reuter et Theo Härder ont inventé l'acronyme ACID (atomicité, cohérence, l’isolation et durabilité) en 1983, pour aider à expliquer ces propriétés.

- Atomicité : une transaction se fait au complet ou pas du tout. Si une partie d'une transaction ne peut être faite, il faut effacer toute trace de la transaction et remettre les données dans l'état où elles étaient avant la transaction.

- Cohérence : la propriété de cohérence assure que chaque transaction amènera le système d'un état valide à un autre état valide. Tout changement à la base de données doit être valide selon toutes les règles définies, incluant mais non limitées aux contraintes d'intégrité, aux rollbacks en cascade, aux déclencheurs de base de données, et à toutes combinaisons d'événements.

- Isolation : toute transaction doit s'exécuter comme si elle était la seule sur le système. Aucune dépendance possible entre les transactions. La propriété d'isolation assure que l'exécution simultanée de transactions produit le même état que celui qui serait obtenu par l'exécution en série des transactions. Chaque transaction doit s'exécuter en isolation totale.

- Durabilité : la propriété de durabilité assure que lorsqu'une transaction a été confirmée, elle demeure enregistrée même à la suite d'une panne d'électricité, d'une panne de l'ordinateur ou d'un autre problème. Par exemple, dans une base de données relationnelle, lorsqu'un groupe d'énoncés SQL a été exécuté, les résultats doivent être enregistrés de façon permanente, même dans le cas d'une panne immédiatement après l'exécution des énoncés.

II.C.3.1.1.2 NoSQL

Une alternative qui a émergé pour le stockage de grands volumes de données et qui envisage des données non structurées ou semi-structurées est le mouvement NoSQL (Not Only SQL).

Ce mouvement propose l'utilisation de systèmes qui permettent de gérer ces grands volumes de données de manière efficace et économique, en respectant des caractéristiques telles que :

- la possibilité d'évoluer horizontalement et de répliquer et distribuer les données sur de nombreux serveurs
- l'utilisation efficace des index distribués et de la RAM pour le stockage des données
- donner la possibilité d'ajouter dynamiquement de nouveaux attributs aux enregistrements de données [56]

Par rapport aux systèmes de bases de données relationnelles traditionnels, les systèmes NoSQL sont recommandés lorsqu'il est nécessaire de servir des millions d'utilisateurs sans perdre en performance, comme dans le cas des réseaux sociaux. Ces systèmes offrent une solution qui est généralement facile à utiliser, pas trop chère et évolutive. Cependant, parmi ses inconvénients, il n'y a pas de modèle de données unique au niveau du système, en termes d'architecture, il y a peu de normalisation des interfaces pour les services, et il n'y a pas non plus de sémantique standard, ce qui entraîne des problèmes d'interopérabilité [24].

Les quatre modèles de stockage NoSQL les plus couramment utilisés [56] sont présentés dans la figure 13 modèles de données pour les bases de données NoSQL :

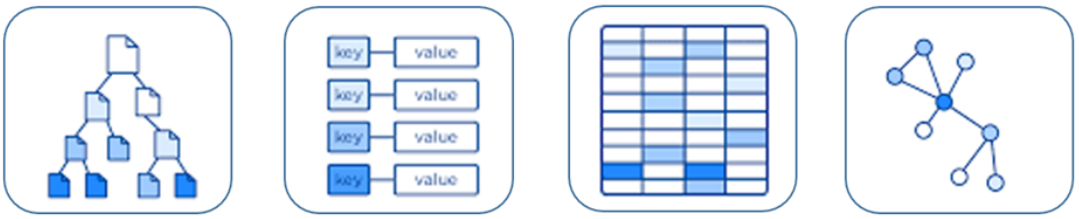


Figure 15 : modèles de données pour les bases de données NoSQL [55]

[55] www.docs.microsoft.com/fr-fr/dotnet/architecture/cloud-native/relational-vs-nosql-data

II.C.3.1.1.3 Entrepôt de données

Un entrepôt de données (data warehouse) est un dépôt d'informations historiques recueillies auprès de sources multiples des données structurées, unifiées sous un même schéma et généralement situées en un seul endroit. Les entrepôts de données sont construits par un processus de nettoyage, d'intégration, de transformation, de chargement et de mise à jour périodique des données.

La caractérisation d'un entrepôt de données comporte plusieurs aspects fondamentaux, tels que l'intégration, la non-volatilité et la variation dans le temps. Le fait qu'un entrepôt de données soit intégré implique qu'il sera alimenté par des données provenant de différentes sources, qui doivent être nettoyées et structurées selon un schéma. Un entrepôt de données doit être non volatil, cela implique que, contrairement à ce qui se passe dans les systèmes transactionnels traditionnels (où les informations sont constamment insérées et modifiées), dans un entrepôt de données, les données sont généralement chargées et consultées massivement sans être modifiées.

Le data warehouse peut être conçu selon des modèles multidimensionnels, dans lesquels il y a la présence d'une table de faits qui est entourée de dimensions. Les principales architectures (ou modèles) utilisées sont la structure en étoile et la structure en flocon de neige. Un schéma de ces structures est présenté à la figure 16 : *Modélisation en étoile et en flocon.*

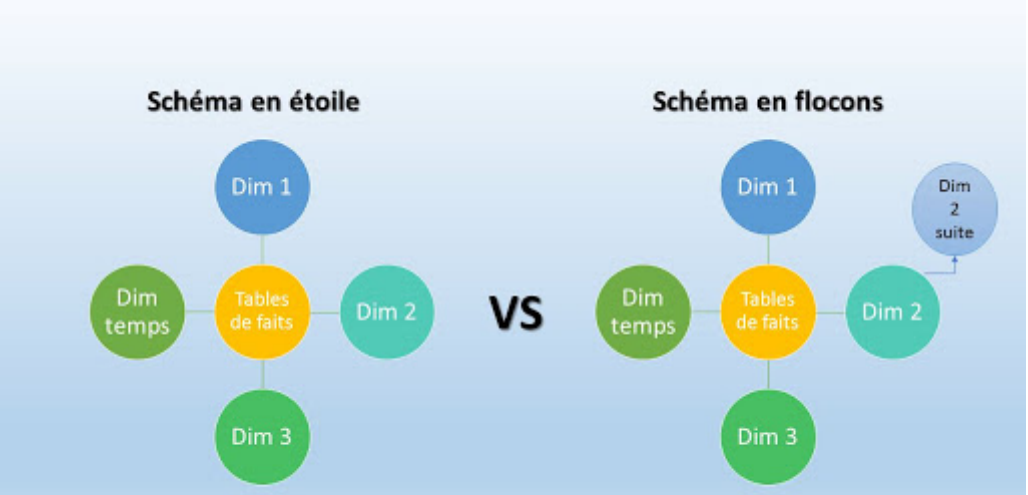


Figure 16 : Modélisation en étoile et en flocon

La structure en étoile est le modèle multidimensionnel classique, avec une seule table de faits entourée de deux tables de dimensions ou plus. Le flocon de neige est une variante comportant plusieurs tables de faits qui partagent certaines tables de dimensions entre elles.

Dans le tableau suivant, montre les principales caractéristiques des modélisations des bases de données en forme d'étoile et en forme de flocon.

	Schéma en étoile	Schéma en flocon
Structure du schéma	Contient des tables de faits et de dimensions.	Contient des tables de sous-dimension comprenant des tables de faits et de dimensions.
Complexité de requête	Faible	Haute
Jointures utilisées	Moins	Plus
Utilisation de l'espace	Plus	Moins
Temps d'exécution de la requête	Moins	Plus (en raison de la plus grande utilisation de jointures).

Figure 17 : Différences entre le schéma en étoile et en flocon.

L'ensemble des données collectées dans le cadre d'un projet de données ouvertes peut nécessiter un stockage en masse et à l'état brut. Ceci signifie un coût d'hébergement un peu plus bas que pour le cas d'un data warehouse, contrairement au coût du traitement par la suite pour pouvoir exploiter la donnée.

Cette forme de stockage à l'état brut est appelée lac de données (data lake) et s'oppose au data warehouse qui stocke les données dans un format plus structuré [40], comme s'explique dans la figure suivante *Data warehouse vs Data lake.*

DATA WAREHOUSE	vs.	DATA LAKE
structured, processed	DATA	structured / semi-structured / unstructured, raw
schema-on-write	PROCESSING	schema-on-read
expensive for large data volumes	STORAGE	designed for low-cost storage
less agile, fixed configuration	AGILITY	highly agile, configure and reconfigure as needed
mature	SECURITY	maturing
business professionals	USERS	data scientists et. al.

Figure 18 : Data warehouse vs Data lake[56].

[56] www.panoply.io/data-warehouse-guide/data-warehouse-vs-data-lake

II.C.3.1.2 L'enjeu écologique de la gestion des données

En novembre 2014, un groupe d'experts indépendants nommé par les Nations Unies a publié un rapport proposant des recommandations sur la manière dont la révolution des données peut être mobilisée pour le développement durable (Nations unies (IEAG) Groupe consultatif d'experts indépendants sur la révolution des données pour le développement durable 2014).

Ce document montre comment le volume croissant de données, permet aux décideurs de suivre et d'atteindre les objectifs de développement durable[57] qui ont été annoncés dans l'Agenda 2030 (Nations unies 2015b). Selon ce document, l'objectif "santé et bien-être" peut être facilité par la collecte de données sur les mouvements des utilisateurs de téléphones mobiles afin de prédire la propagation des maladies infectieuses et l'objectif "énergie propre et abordable" peut être atteint en réduisant les déchets grâce aux données de consommation d'électricité, de gaz et d'eau collectées par des capteurs intelligents.

Le rapport souligne également plusieurs défis pour le développement durable et équitable soulevés par la masse gigantesque de données actuelles, ou mégadonnées. Le groupe d'experts encourage les mesures visant à élaborer des normes de qualité des données dans l'ensemble de la communauté scientifique [70].

Paradoxalement, de nombreux autres scientifiques expriment leurs inquiétudes par rapport à la révolution des données qui menace le développement durable en raison de son empreinte environnementale. Il est largement reconnu dans le contexte de la recherche environnementale que les technologies de l'information et de la communication (TIC), en général, et les data centers et le cloud des données, sont des éléments de l'épine dorsale du Big Data (les données massives) [70]. Ces façons de stocker des données ont une lourde empreinte caractérisée par une forte consommation d'énergie non renouvelable, la production de déchets et des émissions de CO_2 (Pohl et al. 2019 ; Whitehead et al. 2014 ; Williams 2011).

Malgré cette relation ambiguë entre le Big Data et la durabilité, les questions environnementales sont étonnamment absentes, non seulement dans les initiatives politiques soutenant le Big Data, mais aussi dans la littérature récente sur l'éthique du Big Data [70].

C'est pourquoi l'impact écologique de l'adoption d'une stratégie ou d'un projet d'exploitation des données ouvertes doit être évalué dans sa globalité. C'est surtout vrai pour les organisations ou les acteurs économiques qui souhaitent améliorer leur image corporative, ainsi que leur responsabilité sociale (RS) et écologique.

[57] www.un.org/sustainabledevelopment/fr/objectifs-de-developpement-durable

II.D Qualité des données ouvertes

Dans l'idéal, les données ouvertes sont prêtes pour leur exploitation, pour permettre de se concentrer immédiatement sur l'analyse et la compréhension, tout en évitant les retards répétés, les coûts et les risques liés à la recherche et à la préparation des données.
Néanmoins, en ce qui concerne la qualité des informations, beaucoup de portails rendent disponibles des ensembles de données obsolètes, ou avec peu d'informations pertinentes et, dans de nombreux cas, les données ne sont pas bien travaillées.

Nettoyer et traiter les ensembles de données avant de pouvoir les utiliser induit ainsi une charge de travail importante, augmentant ainsi le coût et la difficulté d'exploitation.

Les principaux problèmes de qualité des données ouvertes sont présentés dans la figure suivante et détaillés plus bas :

Figure 19 : Les principaux problèmes de qualité des données ouvertes.

Comme montré dans la figure 19, un des principaux problèmes de qualité des données ouvertes est qu'elles peuvent être incomplètes, c'est-à-dire que les informations ne sont pas publiées dans leur intégralité pour faciliter leur téléchargement et leur utilisation. Les données peuvent ainsi être fragmentées et dispersées dans plusieurs sections d'un même site, voire même être publiées sur différents sites web, ce qui les rend très difficiles à localiser [82].

Comme indiqué dans la section *II.B.3 Manque des métadonnées* les métadonnées sont des champs ou des informations qui décrivent les données. Ils fournissent à l'utilisateur suffisamment d'éléments pour traiter et comprendre les données. Par conséquent, s'il y a des erreurs ou des contradictions présentes aux métadonnées, cela peut entraîner une perte de temps, un effort plus important dans le traitement et même des erreurs dans les résultats de l'analyse.

Le point 3 fait référence à un travail de coordination commun qui est nécessaire pour conduire et établir les formats et la structuration de l'information de manière standardisée au niveau national et même international (au niveau européen, des pays qui font partie de l'Organisation de Coopération et de Développement Économiques, etc.). Par exemple, dans la commande publique, plus particulièrement pour le cas des appels d'offres, la standardisation ou la publication de ces informations avec une même structure au niveau international pourrait entraîner la création d'une Linked Data (réseau des données liées) des marchés publics, ce qui faciliterait la consultation, l'exploitation et permettrait de réduire le coût de stockage de données car plusieurs gouvernements pourraient stocker ces informations sur une base partagée [18].
Cela est le cas de la plateforme TED (Tenders Electronic Daily)[58] . Cette plateforme est la version en ligne du supplément au Journal officiel de l'Union européenne, et elle est consacrée aux marchés publics européens. Le site était fondé sur des directives de l'Union européenne, sur laquelle les pouvoirs adjudicateurs et les opérateurs économiques peuvent mener leurs activités quotidiennes en matière de marchés publics. Les transactions commerciales sont traitées de manière efficace dans un environnement sécurisé et tous les services proposés étant gratuits.

Les données périmées ou obsolètes sont les données qui ne sont pas mises à jour à la fréquence qui serait appropriée en fonction de leur nature. Aussi, il n'est pas toujours indiqué à quelle fréquence les données sont mises à jour. Elles perdent alors beaucoup de leur intérêt pour les réutilisateurs potentiels.

Enfin, la pertinence des données définit à quel point les données apportent une valeur ajoutée dans leur utilisation. Des informations ou colonnes "complémentaires" peuvent par exemple être ajoutées sans une vraie stratégie d'utilisation, ce qui alourdit leur volume et augmente les coûts de stockage et de traitement.

II.D.1 Stratégies de gestion des problèmes de qualité des données ouvertes

Des données propres sont considérées comme des données cohérentes et uniformes, sans doublons et prêtes à être consommées par des machines.

Les données propres facilitent la combinaison de différentes séries de données et permettent une compréhension plus approfondie, et assurent un point de départ fiable à partir duquel vous pouvez travailler les données et générer de la nouvelle valeur.

[58] https://ted.europa.eu/

II.D.1.1 Nettoyage et préparation des données

Le travail du nettoyage des données consiste en principe à :

- Traiter les données manquantes (N/A) : la plupart des bases de données comprennent très souvent des données manquantes. Nombreuses sont les raisons derrière cette problématique : erreurs de typographie, capteur endommagé, absence initiale de données, par exemple dans le cas des questionnaires ou des formulaires où il y des questions non obligatoires, etc.

- Diminuer, dans la mesure du possible, les erreurs de données, en corrigeant les données incohérentes ou aberrantes.

- Résoudre les redondances causées par l'intégration des données, comme les enregistrements en double (les doublons) ou les données incomplètes [40].

II.D.1.1.1 Traitement des données manquantes (N/A)

Il existe différentes options pour le traitement des données manquantes, par exemple :

- Ignorer les observations contenant des valeurs manquantes.
- Remplir automatiquement les valeurs manquantes avec des constantes globales, par exemple par la médiane, la moyenne de l'attribut ou la valeur la plus probable (basée sur une inférence telle que la méthode bayésienne ou l'arbre de décision).
- Définir le format des valeurs manquantes. Par exemple, en python, la bibliothèque numpy met à disposition numpy.nan. Cela permet de continuer l'analyse sans messages d'erreur[59].

L'algorithme K-Means peut également être utilisé pour combler les données manquantes. K-Means est une méthode de regroupement par voisinage dans laquelle un nombre donné de prototypes est utilisé et un ensemble d'exemples à regrouper. K-Means est l'un des algorithmes de clustering les plus fréquemment utilisés [52, 53].
Le "K" fait référence au fait que l'algorithme fonctionne pour un nombre fixe de clusters, qui sont définis en termes de proximité entre les points de données.

II.D.1.1.2 Doublons

Il arrive aussi qu'un niveau élevé de redondance dans les ensembles de données ouvertes existe. La réduction de la redondance, c'est-à-dire, des données ou des observations doublons, sont efficaces pour diminuer le coût indirect du traitement et de l'hébergement des données.

La méthode *duplicated* de la bibliothèque Pandas de Python[60], permet de savoir par exemple s'il y a des doublons dans le jeux des données (ou appelé aussi *dataframe*).

[59] www.numpy.org/doc/stable/user/misc.html
[60] Cet élément est traité dans la section *II.E.1 Python.*

La méthode *sum* appliquée à la méthode duplicate permet de connaître le nombre de doublons : df.duplicated().sum() . La méthode *drop_duplicates* permet de supprimer les doublons au sein du jeux de données, il est possible de spécifier une colonne entre guillemets ou une partie des colonnes entre crochets : df.drop_duplicates().

II.D.1.1.3 Évaluer la cohérence et délimiter le périmètre des analyses

Une valeur aberrante correspond à une valeur éloignée de la distribution de la variable. Cela pourra être dû à une erreur de typographie ou à une erreur de mesure mais cela pourra également être une valeur extrême. Un valeur extrême est une valeur non erronée qui s'éloigne néanmoins fortement du reste des valeurs de la variable.

Une façon assez simple de détecter ces valeurs est d'utiliser la règle de l'intervalle interquartile, en suivant ces étapes :

- Calculez l'intervalle interquartile pour les données, défini par la différence entre les troisième et premier quartiles : IQR = Q3 - Q1. L'intervalle interquartile montre comment les données sont réparties autour de la médiane.
- Multipliez l'intervalle interquartile (IQR) par 1,5 (une constante utilisée pour discerner les valeurs aberrantes).
- Ajoutez 1,5 x (IQR) au troisième quartile. Tout nombre supérieur à cette valeur est présumé aberrant.
- Soustrayez 1,5 x (IQR) du premier quartile. Tout nombre inférieur à cette valeur est présumé aberrant.

Néanmoins, toute valeur extrême obtenue par la méthode interquartile doit être examinée dans le contexte de l'ensemble complet de données, avant de le supprimer. En effet, il faut remarquer que des observations extrêmes ne sont pas forcément aberrantes.

Les valeurs extrêmes sont des valeurs réelles mais "rares" et les données aberrantes sont considérées comme erreurs. Par exemple, dans un cours universitaire, il peut y avoir des élèves de 70 ans, mais s'il y a un élève qui est enregistré avec un âge de 150 ans ce sera très probablement un erreur.
Ces deux types de données peuvent tromper les modèles statistiques mais sont traités de différentes façons.

Généralement les données incohérentes ou aberrantes sont éliminées ou remplacées par une valeur précédemment stipulée, par exemple, dans le cas de données quantitatives, il est possible de remplacer une valeur aberrante par la moyenne de l'échantillon ou la moyenne[61].

Les valeurs extrêmes peuvent être traitées ou étudiées séparément afin d'essayer de trouver des explications sur leur existence, surtout les valeurs extrêmes qui font partie des variables significatives ou définies comme importantes, car des petits changements dans la position de ces observations peuvent causer des modifications majeures dans les analyses.

[61] Ce sujet est traité dans la section *II.D.1.1 Nettoyage et préparation des données.*

Il y a différentes possibilités pour identifier ces observations et mesurer l'effet de leur suppression (par exemple, la distance de Cook[62]) [54, 55].

Plus concrètement, dans le cadre d'un projet de Machine Learning (apprentissage automatique), par exemple, il est nécessaire de faire le choix de supprimer une valeur aberrante, afin d'obtenir une meilleure qualité de prédiction, car un modèle pourra être très sensible aux données extrêmes ce qui va biaiser les prédictions.

Le sujet Machine Learning est traité dans la section *I.B.2.3 Machine Learning et Deep Learning*.

[62] https://fr.wikipedia.org/wiki/Distance_de_Cook

II.E Outils de traitement de données

Choisir parmi les nombreuses technologies est un challenge pour les entreprises. Tous les outils que nous présentons ici peuvent être utilisés pour une approche stratégique des mégadonnées ou smart data (données intelligentes).

II.E.1 Python

Python est un langage de programmation Open Source (source ouverte ou gratuite) orienté objet interprété. Guido van Rossum a commencé à travailler sur Python à la fin des années 1980 et l'a publié pour la première fois en 1991 sous le nom de Python 0.9.0[63].

La polyvalence de Python pour le développement d'applications multiples est ce qui a fait que son utilisation a dépassé le cadre des développeurs pour atteindre des groupes de recherche de différentes universités dans le monde entier qui ont développé des bibliothèques pour toutes sortes de domaines tels que la biologie, la physique, les mathématiques et l'ingénierie, entre autres [62]. Et l'utilisation n'arrête pas de s'étendre, comme montre le graphique suivant :

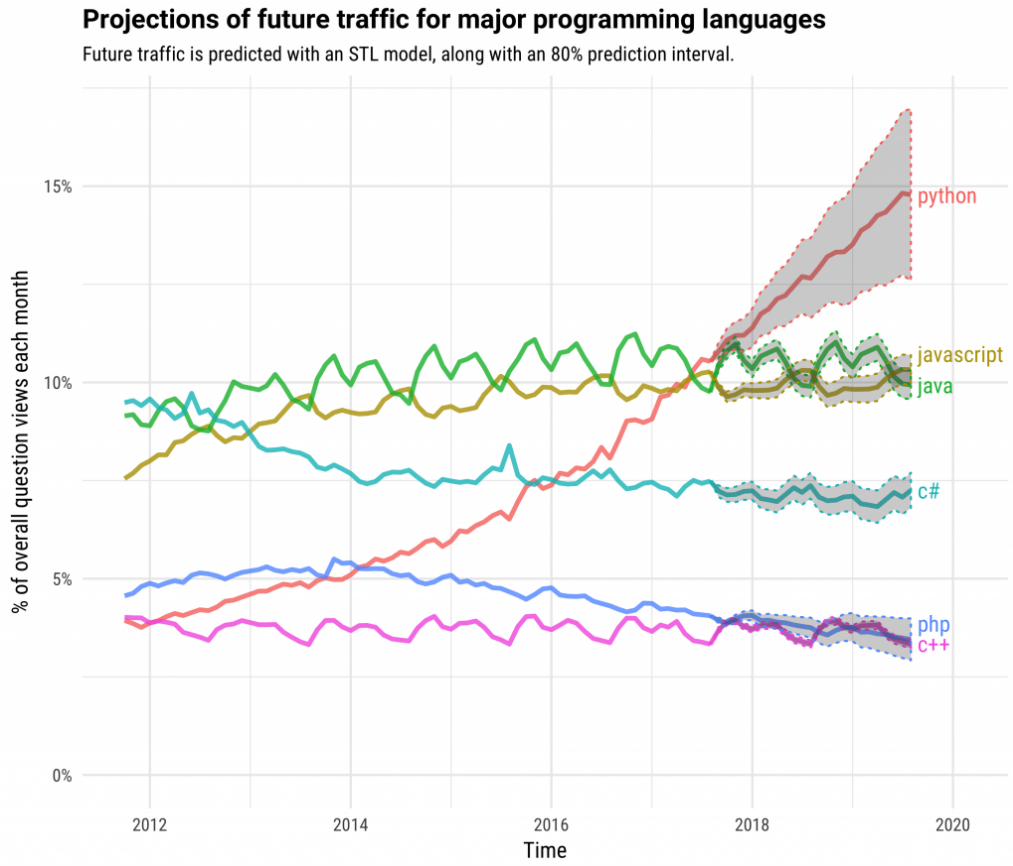


Figure 20 : Projections de l'utilisation du langage Python dans le monde, comparé aux autres langages de programmation les plus populaires[64].

[63] www.python.org/about

[64] https://stackoverflow.blog/2017/09/06/incredible-growth-python/

Ces bibliothèques sont des paquets contenant un grand nombre de fonctions, d'outils et d'algorithmes programmés. Elles sont disponibles gratuitement sur l'entrepôt public PyPI[65]. Leur objectif est de faire gagner beaucoup de temps de programmation aux utilisateurs.

Comme indiqué dans la section *I.B.2.1 Data mining (exploration de données)*, dans le développement d'un projet Data, il y a en effet différentes étapes à développer, certaines d'entre elles sont :

- Extraction des données
- Traitement des données
- Application d'algorithmes, qui peuvent être d'apprentissage automatique, d'apprentissage profond ou de traitement du langage naturel (NLP).
- L'évaluation des résultats

Avec Python, ces tâches peuvent être réalisées sans nécessiter de connaissances avancées en programmation, avec peu de lignes de code et dans des environnements de programmation faciles à prendre en main, qui facilitent la programmation et la visualisation des résultats.

Pandas[66], dont il a déjà été question plus haut, est un exemple de ces bibliothèques Open Source Python. Elle est maintenue par la communauté PyData et est principalement utilisée pour l'analyse et le traitement des données.

Le tableau suivant présente quelques bibliothèques qui font partie des outils les plus populaires pour le traitement des données à l'heure actuelle [62, 89].

Bibliothèque	**Fonctionnalité ou application principal**
Tensorflow	Creation des reseaux de neurones
Seaborn, Altair, Matplotlib	Visualisation des données
Pandas	Gestion et traitement des jeux de données
NLTK	Résoudre des problèmes dans le domaine du traitement du langage naturel.
NumPy	Calcul de données sous forme de matrices multidimensionnelles.
Scikit-learn	Algorithmes de traitement des données et d'apprentissage automatique.

Figure 21 : Bibliothèques Python populaires pour l'exploitation des données.

[65] https://pypi.org/
[66] www.pandas.pydata.org

II.E.2 R

R est un logiciel de traitement statistique des données. Il fonctionne sous la forme d'un interpréteur de commandes. Il dispose d'une bibliothèque très large de fonctions statistiques, d'autant plus large qu'il est possible d'en intégrer de nouvelles par le système des paquets [64].

R est aujourd'hui l'un des écosystèmes les plus riches pour effectuer des analyses de données. Il y a environ 12000 paquets disponibles dans CRAN (dépôt open-source). Il est possible de trouver une bibliothèque pour n'importe quelle analyse que vous souhaitez effectuer. La grande variété de bibliothèques fait de R le premier choix pour l'analyse statistique, en particulier pour les travaux analytiques spécialisés. R propose également une palette étendue de fonctionnalités graphiques [63, 65].

La différence entre R et les autres logiciels statistiques est la sortie. R dispose d'outils très performants pour communiquer les résultats [63]. Rstudio est l'interface utilisateur de ce langage.

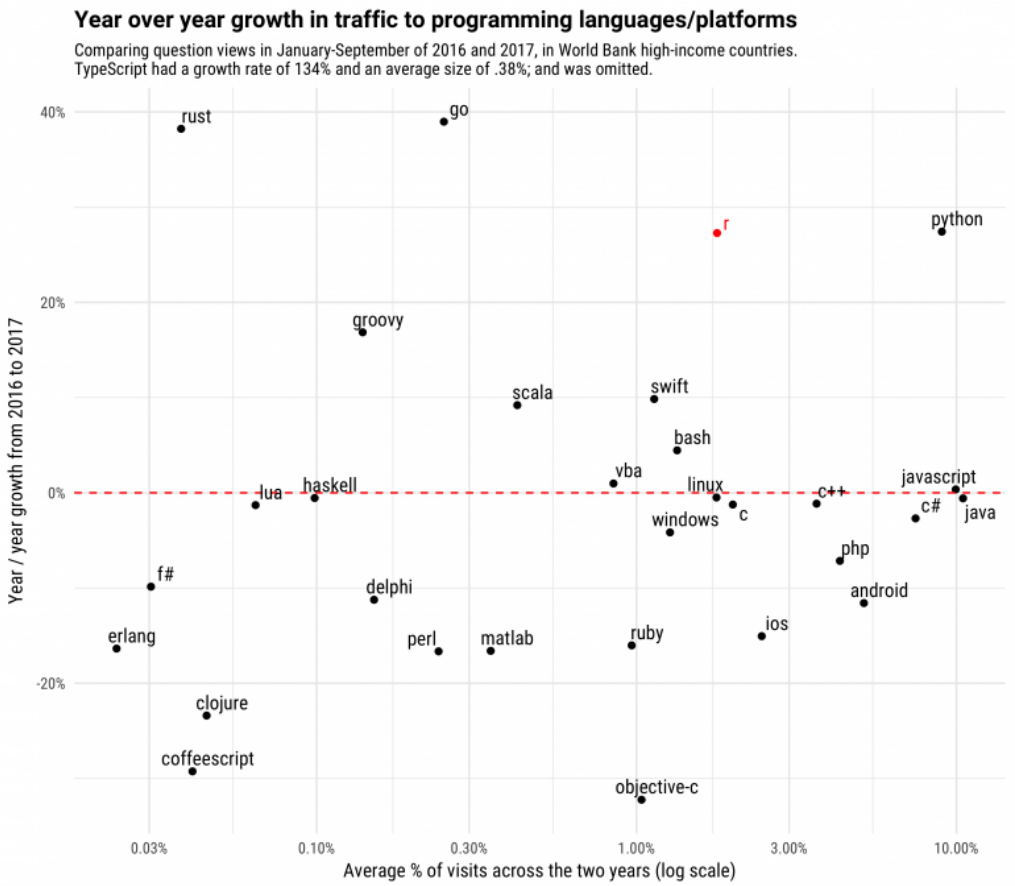


Figure 22 : Pourcentage de la croissance annuelle de l'utilisation des langages informatiques comparé à l'augmentation des consultations liées à ces langages sur le site Stackoverflow.[67]

[67] www.stackoverflow.blog/2017/10/10/impressive-growth-r

Comme l'indique la figure 22, le langage de programmation R a connu une croissance remarquable au cours des dernières années. En fait, R croît à un rythme similaire à celui de Python en termes de pourcentage d'une année sur l'autre [65].

II.E.3 L'écosystème Hadoop

Hadoop est une bibliothèque Apache définie comme un cadre permettant le traitement *distribué* de données sur de grands volumes de données. Il est conçu pour fournir une puissance de mise à l'échelle allant de quelques serveurs à des centaines de machines ou nœuds, qui gèrent le stockage et le traitement de ces informations.

Une base de données distribuée est une base qui n'est pas limitée à un seul système d'exploitation, elle est donc répartie sur différents sites, c'est-à-dire sur plusieurs ordinateurs ou sur un réseau d'ordinateurs. Cela peut être nécessaire lorsqu'une base de données particulière doit être consultée par différents utilisateurs à l'échelle mondiale[68] . Ainsi, pour aborder des projets des données massives, d'un point de vue strictement architectural, la stratégie de distribution du stockage de données est une bien meilleure stratégie que la stratégie d'architecture client/serveur classique [23].

Hadoop est devenu très populaire ces dernières années, devenant l'une des solutions les plus utilisées au niveau organisationnel pour traiter les méga données. Hadoop a deux composants essentiels, un système de fichiers distribué (HDFS) et MapReduce.

Le système de fichiers distribué (HDFS) possède les principales caractéristiques suivantes :

- Tolérance aux pannes
- Accès aux données en continu
- Facilité d'utilisation
- Modèle de cohérence simple
- Portabilité de la coexistence

Hadoop MapReduce est un Framework fait pour écrire des applications pouvant dialoguer sur un cluster Hadoop. Il permet le calcul sur des environnements distribués. En effet, chaque requête est fragmentée en deux phases principales :

- une phase "map" pour filtrer et transformer les données d'entrée, afin qu'elles soient exploitables par le « Reducer »
- une phase "reduce" pour agréger les données (simple calcul statistique ou traitement plus complexe)

Les problématiques des données massives sont segmentées d'un point de vue fonctionnel et pour chaque segment, des technologies qui s'appuient sur Hadoop ont été développées pour répondre à ses challenges. L'ensemble de ces outils forment ce qui s'appelle l'écosystème Hadoop.

[68] www.oracle.com/fr/cloud/definition-base-donnee-distribuee

L'écosystème Hadoop le rend capable de résoudre une grande variété de problématiques métiers. A ce jour, cet écosystème est composé d'une centaines de technologies qui sont regroupés dans en 14 catégories selon leur segment de problématique : les langages d'abstraction, le SQL sur Hadoop (Hive, Pig), les modèles de calcul (MapReduce, Tez), les outils de traitement temps réel (Storm, Spark Streaming), les Bases de données (HBase, Cassandra), les outils d'ingestion streaming (Kafka, Flume), les outils d'intégration des données, (Sqoop, Talend), les outils de coordination de Workflow (Oozie, Control M for Hadoop), les outils de coordination de services distribués (Zookeeper), les outils d'administration de cluster (Ranger, Sentry), les outils d'interface utilisateur (Hue, Jupyter), les outils d'indexation de contenu (ElasticSearch, Splunk), les systèmes de fichier distribués (HDFS), et les gestionnaires de ressources (YARN et MESOS) [45].

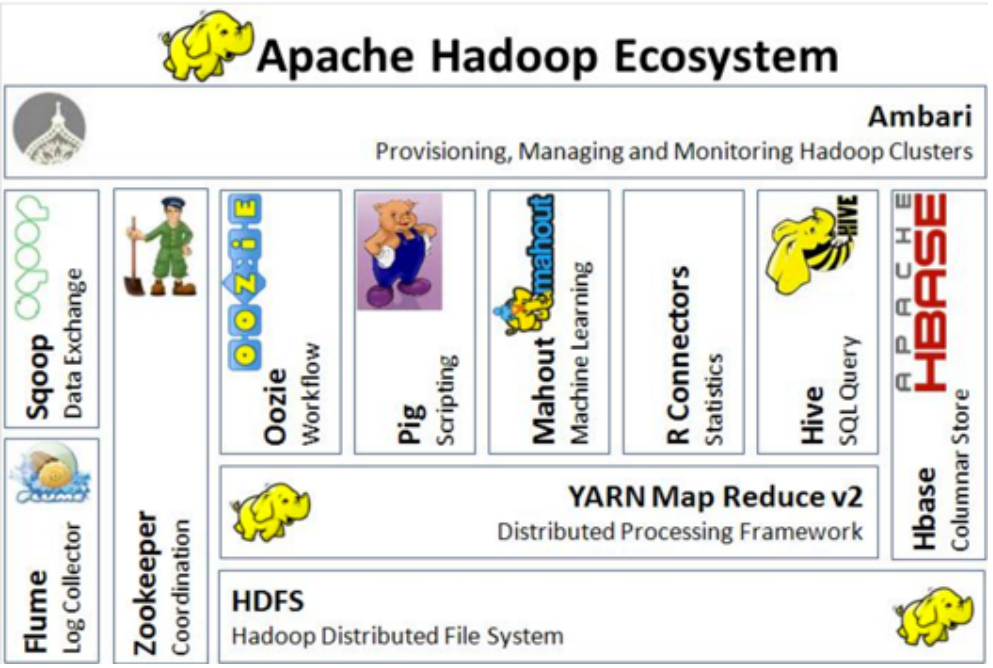


Figure 23 : Environment Apache Hadoop.

II.E.3.1 Elasticsearch et Solr

Dans l'écosystème Hadoop, les technologies spécialisées dans la recherche d'information s'appellent les *moteurs d'indexation de contenu*. Ces moteurs sont des types particuliers de SGBD auxquels s'intègrent des fonctionnalités d'indexation et de recherche de contenu.

Apache Solr[69] et ElasticSearch sont deux de ces types de moteur. Apache Solr et ElasticSearch sont des moteurs NoSQL d'indexation de contenu scalables, qui s'appuient sur Apache Lucene, une bibliothèque de d'indexation de contenu, pour fournir des fonctionnalités d'indexation et de recherche de contenu. Apache Lucene ne gère pas le stockage des documents, ces deux moteurs fournissent le support de stockage des données de sorte que l'indexation et la recherche puisse se faire directement dans le moteur [66].

[69] www.solr.apache.org

Certaines de ses caractéristiques sont décrits au tableau suivant :

SGBDR	ElasticSearch
Base de données	Index ElasticSearch
Table de la base de données	Indice de l'index ElasticSearch
Colonne de la table	Type de l'indice ou propriétés des documents JSON
Ligne de la table	Document de l'indice

Figure 24 : Comparaison d'ElasticSearch avec un SGBDR.

ElasticSearch est donc un moteur distribué de stockage, de recherche et d'analyse de contenu, Open Source qui permettent d'effectuer les recherches de contenu très rapidement. En tant que moteur de stockage, il sauvegarde les données en format JSON. En tant que moteur d'analyse de contenu, il s'appuie sur Logstash, un logiciel de gestion de logs et Kibana, une plateforme d'exploration et de visualisation des données, pour effectuer des analyses sur les données qu'il stocke [67].

Figure 25 : ElasticSearch Ecosystem [70].

Sa particularité vient du fait qu'il s'interroge à partir d'une API REST, est accessible à partir du protocole HTTP et utilise le format JSON, aussi bien pour le stockage des données que pour le renvoi des réponses de requêtes. Les standards HTML, REST et JSON le rendent facile à intégrer avec d'autres applications, et il est simple à utiliser pour les utilisateurs.
De plus, le fait qu'il utilise le JSON pour le stockage des données rend l'exploitation des données possible à partir de n'importe quel langage de programmation possédant des API permettant de lire du JSON.

Le chapitre suivant présente un cas d'étude d'un business model inspiré par l'exploitation de données ouvertes. L'objectif est de donner un exemple de la mise en pratique et de l'utilisation de certains des outils décrits ci-dessus, ainsi que d'illustrer des stratégies de résolution des enjeux liés aux données ouvertes.

[70] www.wilsonmar.github.io/elastic-ecosystem

III. Cas d’étude Databiz

III.A OctopusMind

OctopusMind est une entreprise experte en collecte, traitement, analyse et visualisation de données économiques qui développe des solutions innovantes de prospection commerciale et de CivicTech, comme l'application J360 (j360.info) et Cityzenmap (cityzenmap.com).

Afin de pouvoir analyser des importants volumes des données, la société réalise systématiquement des investissements dans la recherche et le développement (R&D), notamment en matière de data mining. Le savoir-faire caractéristique est la création de solutions de veille et d’analyse de données économiques en combinant les intelligences artificielle et humaine.

III.B Plateforme J360

J360 est une plateforme de détection d'opportunités d'affaires (appels d'offres publics et privés) dans plus de 180 pays qui permet à tous les professionnels de détecter des opportunités commerciales, trouver des partenaires pour y répondre et publier gratuitement des intentions d'achats.

Cette application fonctionne grâce à des algorithmes intelligents de matching de compétences métiers que suggèrent des marchés (appels d’offres) pertinents en fonction du profil de l'utilisateur (recommandations en fonction de ses compétences enregistrés, selon la zone géographique, la langue, etc.). L’utilisateur bénéficie aussi des suggestions de marchés similaires, d’organisations potentiellement concurrentes et des estimations du montant ou du prix moyen attendu de l’annonce.

La caractéristique innovante du service J360 repose sur la pertinence des informations proposées en fonction du profil utilisateur grâce à des techniques de machine learning (apprentissage automatique) et l'apprentissage profond ou deep learning (classification des données en fonction des compétences métiers et suggestions personnalisées multilingues) et le fait de pouvoir consulter, sur un seul site, des informations propres compilées à partir de plusieurs centaines de sources nationales et étrangères.

Contrairement à certains concurrents du secteur, cette solution propose aussi aux utilisateurs de diffuser leurs intentions d’achats sur la plateforme.

III.C Databiz

Databiz est une plateforme itérative d'analyse économique mondiale et de stratégie commerciale.

Avec l'essor de l'ouverture des données, ce projet a pour ambition de révolutionner l'analyse de données économiques multi-sources. Cette idée est née du constat que de nombreuses données disparates, contenues dans plusieurs sources peuvent être de formidables aides à la décision une fois combinées.

L'enjeu de DataBiz est de construire, de façon itérative, un outil d'analyse, avec une approche Smart Data[71], à partir de 3 cas d'usages clients :

- Analyse de marché : étude de l'évolution d'un secteur d'activité.
- Développement commercial : recommandations de produits ou services en croissance proche de l'activité de l'entreprise afin de diversifier son activité ou accéder à de nouveaux marchés.
- Prospection : recommandations d'acteurs économiques pour générer des leads qualifiés, c'est-à-dire des clients ou partenaires potentiels du secteur.

Les 3 axes d'analyses de données proposés par DataBiz sont issus de besoins exprimés par des entreprises, banques et cabinets conseil, d'après une analyse des besoins des clients. La stratégie de ce développement est de proposer une version initiale (produit prototype) aux clients potentiels pour identifier les usages ou les besoins et mieux cerner la valeur ajoutée perçue du produit.

Un des défis du déploiement est de traiter automatiquement les données pertinentes et de les agréger de manière intelligente pour obtenir des indicateurs qui auraient pris beaucoup de temps à calculer par une étude menée de manière classique.

La différenciation par rapport aux outils actuels concurrentiels est de construire une analyse automatisée, non pas à partir de données macroéconomiques, mais à partir de données microéconomiques, issues des appels d'offres.
Néanmoins, même si la source initiale de données est constituée de la base d'appels d'offres du site J360.info, d'autres sources de données ouvertes sont intégrées (par exemple la base SIRENE, qui identifie et caractérise les acteurs économiques, ou encore la base des enregistrements du commerce extérieur *Comtrade*).

Un des enjeux du projet est de trouver ces données externes pertinentes pour leur traitement et le croisement ou la pondération statistique avec la base des appels d'offres. Certaines de ces sources offrent une interface simple de collecte (API, fichiers xml, etc.) mais d'autres demandent d'utiliser du web scraping (exploration web automatisée). Il est donc nécessaire de gérer la mise à jour des informations, leur suppression ainsi que la qualité des données (doublons, données incomplètes).

[71] Ce sujet a été traité dans la section *I.A.2.1 La méthode des données intelligentes et les données massives.*

III.C.1 Développement du produit prototype

Le produit prototype de Databiz est focalisé sur le scénario d'analyse de marché (l'étude de l'évolution et le comportement économique d'un secteur d'activité). Les informations présentées en forme d'indicateurs donnent des perspectives pour valider un business plan ou un investissement.

Le produit est composé de 8,867,356 documents (annonces des appels d'offres) de la base J360 vers l'espace dédié au développement du produit (la "pré-production"). Ces données concernent des appels d'offres récupérés de 2017 à 2019, comme décrit dans la figure 27 Tableau de métadonnées - source J360.

Tableau de métadonnées	
Source	J360.info
Périmètre géographique	Pays de la zone euro
Granularité géographique	Par pays et pour le cas de la France, par département
Langues	Multiple
Périmètre temporel	2017-2019
Granularité temporelle	Secondes
Sujets	Appels d'offres
Granularité [72]	Appels d'offres (AO), Avis d'attribution (AA)
Fréquence de mise à jour	Une fois au début du projet[73]

Figure 26 : Tableau de métadonnées - source www.J360.info.

L'extraction, le traitement et le chargement des données pour la construction des visualisations (processus ETL) s'effectue avec un approche scriptée ou à la main, c'est-à-dire à partir du développement complet du processus en utilisant principalement les outils Python comme le framework (cadre de travail web)[74] Django, la bibliothèque de traitement des données Pandas ou le paquet de visualisation Altair, ainsi que Elasticsearch.

Ce sujet est détaillé dans les sections suivantes.

[72] La notion de granularité définit la taille du plus petit d'un élément d'un système. Pour en savoir plus : https://fr.wikipedia.org/wiki/Granularit%C3%A9

[73] Il est important de signaler que l'objectif à court terme est de faire les mises à jour en temps réel, ou au moins de manière journalière.

[74] L'objectif d'un framework est généralement de simplifier le travail des développeurs informatiques, en leur offrant une architecture pré-construite qui leur permette de ne pas repartir de zéro à chaque nouveau projet.

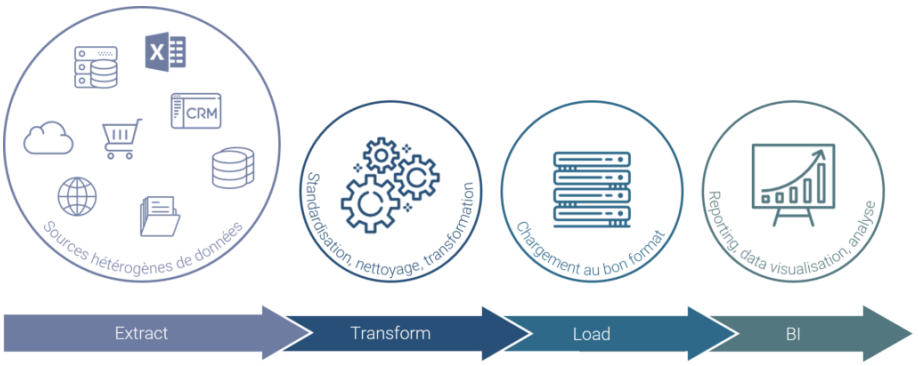


Figure 27 : Processus ETL [75]

III.C.1.1 Récupération et stockage des données

Comme mentionné auparavant, la plateforme Databiz est composée des données ouvertes macroéconomiques issues des appels d'offres[76] présentes dans le site www.J360.info.

Le processus de récupération de ces données est le suivant :

- D'abord l'équipe de veille repère de nouvelles sources d'appels d'offres, en accord aux besoins identifiées des utilisateurs.

  Par exemple, si un nouveau client souhaite avoir des opportunités d'affaires liées aux services informatiques au Brésil, un pays dont (imaginons-nous) nous n'avons pas assez d'informations. L'équipe de veille sera donc chargée de rechercher de nouvelles sources des marchés publics où figurent des acheteurs potentiels.

- Les sources ciblées seront communiqués à l'équipe IT afin de procéder à la collecte des informations. L'extraction et l'intégration de ces données s'effectuent par l'utilisation de techniques de web scraping (technique qui a été détaillée dans la section *II.B.5.2 Collecte des données).*

Cette approche de ciblage et de collecte d'informations est conçue avec une vision de Smart Data (données intelligentes). L'objectif est de récupérer uniquement des données exploitables afin de :

- Économiser de l'espace de stockage.
- Réduire l'impact écologique de l'hébergement des données.
- Faciliter l'intégration et la modélisation des sources.
- Simplifier la recherche d'informations précises.

[75] www.axysweb.com/processus-etl-talend

[76] Ce sujet était traité dans la section *I.A.1.3.1.1.2 Données microéconomiques*.

III.C.1.2 Indexation et exploration des données

III.C.1.2.1 Emploi d'Elasticsearch

Les documents ou les annonces d'appels d'offres sélectionnés pour le développement de la version prototype du site Databiz, ont été indexés dans Elasticsearch (ES).
Même si l'origine de ces informations est la base relationnelle du site www.J360.info, pour cette phase initiale du projet, nous avons choisi d'utiliser le moteur d'indexation d'Elasticsearch comme une base des données afin de rendre plus agil l'interrogation et l'intégration des données. D'autres raisons de ce choix sont les performances de stockage et de recherche d'informations dans des champs qui contiennent une quantité de texte importante. L'objectif final est de vérifier si c'est viable et pertinent de continuer avec cette méthode d'indexation à long terme ou de visualiser et de passer à une base de données relationnelle.

De plus, Elasticsearch offre la possibilité de visualiser les données via l'interface utilisateur Kibana. L'application Kibana nous a permis de commencer l'exploration des données et de tester la construction des indicateurs avec l'objectif de valider la pertinence du mapping établi pour l'index. Le mapping est similaire à la définition de schéma dans une base de données relationnelle. Il permet de définir la structure des types.
Cela a été très important car, si le développement de la plateforme de Databiz continue avec l'indexation ou la synchronisation de toute la base de J360; le changement du mapping et la ré-indexation subjacente des informations représenteraient un travail long, difficile et coûteux.

III.C.1.2.2 Jupyter Notebook

Après avoir indexé et exploré les données, l'étape suivante consistait à développer les indicateurs prévus sur Jupyter Notebook [77]. Cet outil est un bloc des notes de calcul (computational notebook) gratuit et interactif qui prend en charge plus d'une quarantaine de langages de programmation.
Certains de ses avantages sont qu'il est possible de visualiser le code, de l'exécuter depuis une interface utilisateur, d'apporter des changements, d'intégrer des données et de vérifier les résultats de ces modifications instantanément.

La principale raison pour laquelle nous avons utilisé cet outil était de développer les indicateurs en simulant de manière la plus proche possible la face de production.
Le service Jupyter nous a permis donc de récupérer et traiter les données, ainsi que de construire les indicateurs en forme de graphiques dynamiques tout en utilisant les mêmes technologies que dans l'état de production, c'est-à-dire : DSL d'Elasticsearch ainsi que les bibliothèques Python de Pandas et Altair.

[77] https://jupyter.org/

III.C.1.3 Traitement et visualisation des données

Le traitement des données démarre dans le premier onglet “Écran de paramétrage” (*figure 28*). D'abord, l’utilisateur doit saisir les mots-clés qui caractérisent son activité, et ensuite il doit choisir :

- un des métiers proposés (liste de clusters suggérés et liés à son activité commerciale).
- des autres mot-clés associées à son activité commerciale.
- au moins un pays de la zone euro (il a été choisi de délimiter le prototype aux pays de la zone euro afin de ne pas avoir besoin de convertir les devises).
- ainsi qu'une période de temps à analyser (des dates allant de début 2017 à fin 2019).

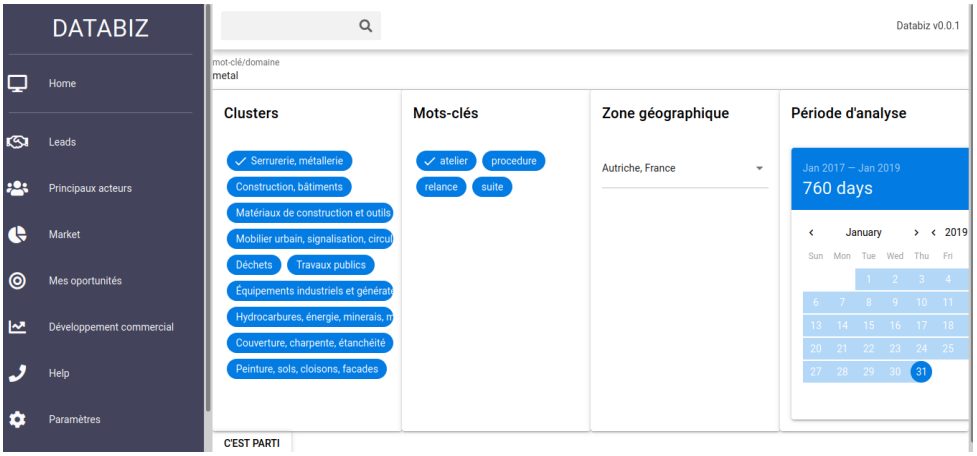


Figure 28 : Landing page (écran de paramétrage) du prototype Databiz.

La finalité de ce paramétrage est de délimiter le périmètre de l’étude (filtrer notre base des appels d’offres) ou d’affiner la recherche afin de personnaliser les indicateurs attendus.
Ce processus de triage d’information est géré automatiquement dans la plateforme

Ensuite les données sont traitées ou préparées afin d’établir une base “propre” pour les visualisations. Cette étape de préparation et visualisation des données été réalisé dans Jupyter Notebook où nous avons expérimenté et finalement déterminé :
:

- les techniques les plus efficaces pour nettoyer et préparer les données.
- la façon la plus performante et précise de construire les indicateurs.

Enfin, les programmes résultats ont été adaptés et implémentés directement dans le produit (la pré-production).

Ces éléments seront détaillés dans la section suivante.

III.C.1.2.1 Les indicateurs clés

Le scénario analyse de marché est composé de 4 indicateurs en forme de visualisations qui partent de deux grandes familles des données : soit des appels d'offres ou marchés publiés (AO), soit des avis d'attribution ou résultats de marchés (AA)[78].

Indicateurs basés sur des appels d'offres	**Indicateurs basées sur des avis d'attribution**
Évolution historique du secteur.	Chiffre d'affaires du marché (valeur commerciale du secteur).
Analyse géographique (carte du volume des marchés).	Principaux acheteurs du secteur.
Principaux concurrents du secteur.	

[78] Voir granularité des données Figure 27 Tableau de métadonnées - source J360

III.C.1.2.1.1 Évolution historique du secteur

Le graphique de l'évolution historique d'un secteur d'activité, montre le nombre d'appels d'offres trouvés pendant la période du temps et la zone géographique choisie. Les données utilisées sont décrites dans la figure 29 - *Dictionnaire des données de l'indicateur Évolution du volume des marchés publiés*.

Dictionnaire des données					
Indicateur : Evolution du volume des marchés publiés					
Description	**Champ**	**Format**	**Exemple**	**Source**	**Granularité**
Date de récupération de l'annonce.	created	datetime64	2017-01-01	J360	AO
Données calculées					
Description	**Champ**	**Format**	**Exemple**	**Librairies utilisées**	**Calcul**
Marchés comptés par période.	doc_count	int64	100	Elasticsearch DSL	Agrégation des annonces par mois.
Ligne de tendance LOESS.	loess	int64	100	Altair	Méthode de régression non paramétrique[79] qui permet de produire des courbes lissées, ajustées à un nuage de point [28].

Figure 29: Dictionnaire des données de l'indicateur de l'évolution du volume des marchés publiés.

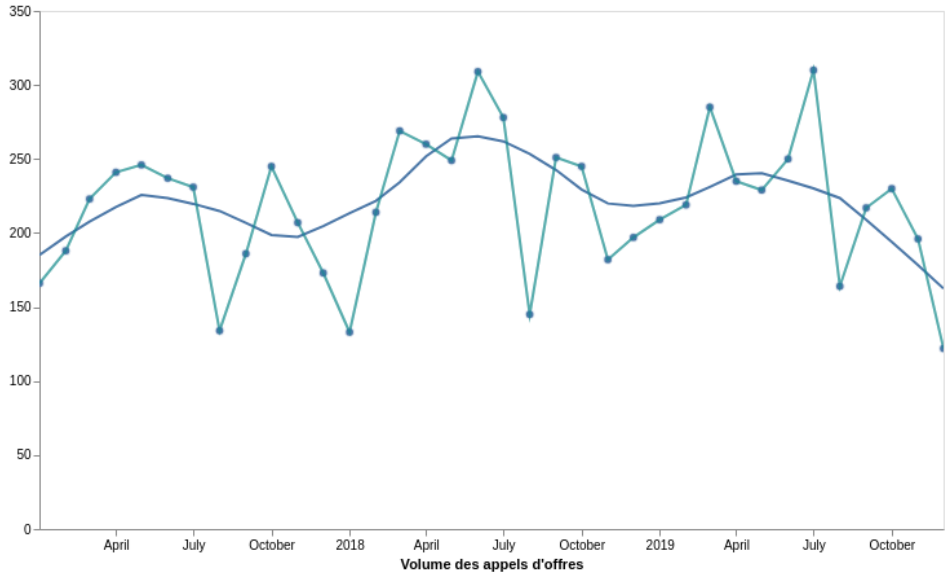


Figure 30: Indicateur de l'évolution historique du secteur.

[79] C'est-à-dire que la méthode n'est pas associée à une équation, comme par exemple une régression linéaire ou polynomiale classique.

Avec quelques paramètres à fixer, la bibliothèque des visualisations *Altair*[80] nous permet d'avoir un histogramme interactif (zoomable), avec une répartition de l'axe 'X' (dates) qui s'auto génère, ainsi qu'une ligne de lissage (tendance) calculée aussi de façon automatique.
Cette courbe de lissage utilise la méthode LOWESS - locally-weighted scatterplot smoother[81] ou nuage de points localement pondéré lissé, qui est connue aussi comme régression mobile [28].
Cette méthode de régression est définie comme "non paramétrique", c'est-à-dire qui n'est pas associée à une équation, comme par exemple une régression linéaire ou polynomiale classique. Cela est l'un de ses principaux avantages car le fait de ne pas avoir besoin de définir une formule adaptée aux types de données rend son application très large et flexible.

III.C.1.2.1.2 Analyse géographique - carte de densité des marchés

L'utilisateur de Databiz aura aussi accès à une carte du placement géographique des marchés, avec des couleurs qui augmentent leur tonalité selon la densité des annonces comptées (voir figure 31 - Dictionnaire des données de la carte de densité des marchés).

Dictionnaire des données					
Indicateur : Carte de densité des marchés (analyse géographique)					
Description	**Champ**	**Format**	**Exemple**	**Source**	**Granularité**
Date de récupération de l'annonce.	created	datetime64	2017-01-01	J360	AO
Coordonnées géographiques (latitude et longitude).	localisation	object	{'lon': 6.3355935, 'lat': 48.1446427}	J360	AO
Données calculées					
Description	**Champ**	**Format**	**Exemple**	**Librairies utilisées**	**Calcul**
Géocodage des coordonnées géographiques (latitude et longitude) en Geohash.	geohash_grid	object	gbwfm5vg6yc	Elasticsearch DSL	Agrégation des coordonnées, type geohash_grid avec 11 niveaux de précision
Marchés comptés par zone géographique.	doc_count	int64	100	Elasticsearch DSL	Agrégation des coordonnées
Conversion de geohash en latitude / longitude.	geo	object	(48.2292, -1.53007)	Pygeohash	Décodage des coordonnées

[80] www.altair-viz.github.io/gallery
[81] La régression LOWESS est une généralisation de la moyenne mobile. Comme base de son application, nous considérons par défaut que nos données présentent une distribution gaussienne. Pour en savoir plus : www.altair-viz.github.io/user_guide/transform/loess.html

Polygones pour le template de la carte du monde (base géographique).	gdataframe	geometry	POLYGON ((61.20961 35.64925, 62.23202 35.27011...	Vega_datasets [82], Geopandas [83]	Importation des données via url, et décodage des polygones au format géographique.

Figure 31 : Dictionnaire des données de la carte de densité des marchés.

Exemple du jeu de données utilisé pour la carte :

geohash_grid	doc_count	lat	long
gbwfm5vg6yc	126	48.2292	-1.53007
gbqsywty9de	125	47.27805	-1.81576
ezzxju8hhxx	76	44.84967	-0.45024
gbt4w840rjk	65	48.25202	-3.93005

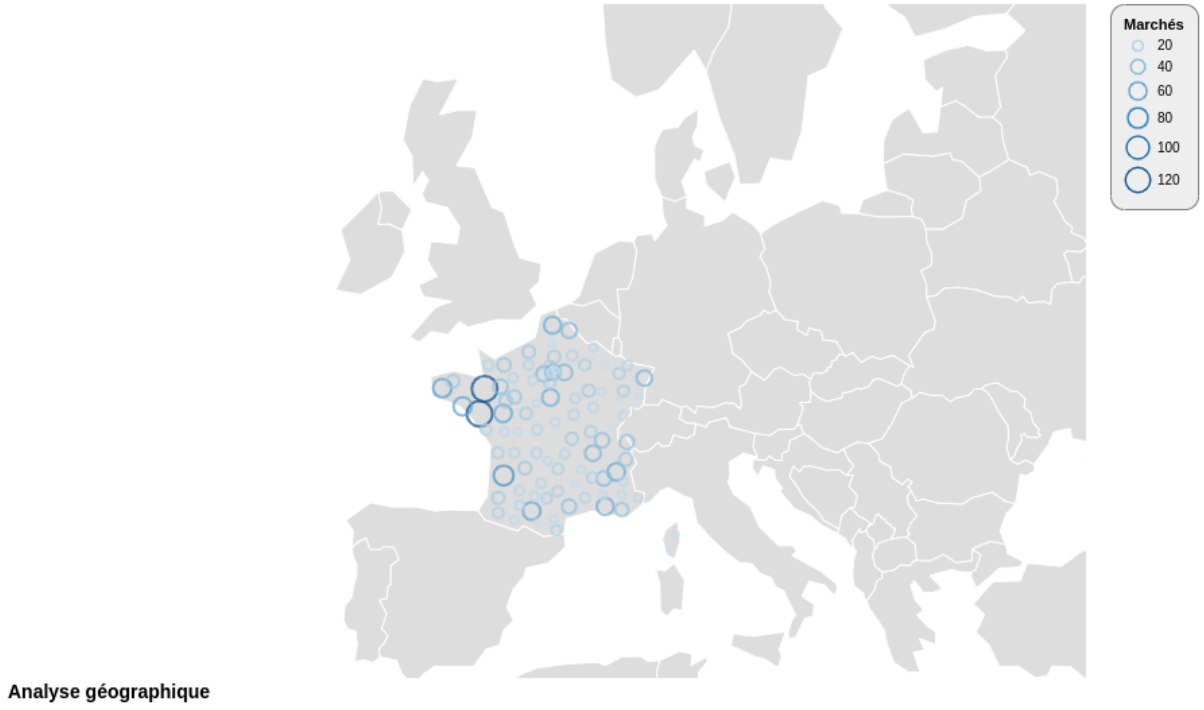


Figure 32 : Carte de densité des marchés.

[82] www.github.com/altair-viz/vega_datasets
[83] www.geopandas.org/io.html

III.C.1.2.1.3 Chiffre d'affaires du marché

L'indicateur de la valeur commerciale du secteur est présenté comme l'évolution historique du chiffre d'affaires (valeur en euros) d'un secteur d'activité. Les données traitées sont indiquées dans la figure suivante *Dictionnaire des données de valeur commerciale du secteur.*

Indicateur : Chiffre d'affaires du secteur					
Description	**Champ**	**Format**	**Exemple**	**Source**	**Granularité**
Date de récupération de l'annonce.	created	datetime64	2017-01-01 0:00:00	J360	AA
Information sur le montant de l'avis d'attribution (montant, devise, gagnant..) .	budgets	object (nested field)	[{'montant': 96643.25, 'type': 'total', 'devise': 'EUR'}]	J360	AA
Information concernant chaque lot présenté dans l'avis d'attribution (montant du lot, gagnant..).	lot	object (nested field)	[{'budgets': [{'montant': 75000.0, 'type': 'total', 'devise': 'EUR'}], [{'titulaires': [{'id': 2716996, 'cp': '54700', 'nom': 'HOLLINGER'	J360	AA

Données calculées					
Description	**Champ**	**Format**	**Exemple**	**Librairies utilisées**	**Calcul**
Total des montants attribués.	total	float64	10000.25	Elasticsearch DSL, Pandas	Valeur totale du montant attribué enregistré de chaque annonce, plus la valeur totale des lots des annonces qui n'ont pas un montant total général enregistré (cela est fait avec deux requêtes différentes).
Ligne de tendance LOESS.	transform_loess	int64	100	Altair	Méthode de régression qui permet de produire des courbes lissées, ajustées à un nuage de point [28].
Moyen du total de montants remportés par mois	mean_total	float64	10000.25	Altair	Moyenne de l'agrégation (somme) des valeurs gagnées par mois.

Figure 33: Dictionnaire des données de valeur commerciale du secteur.

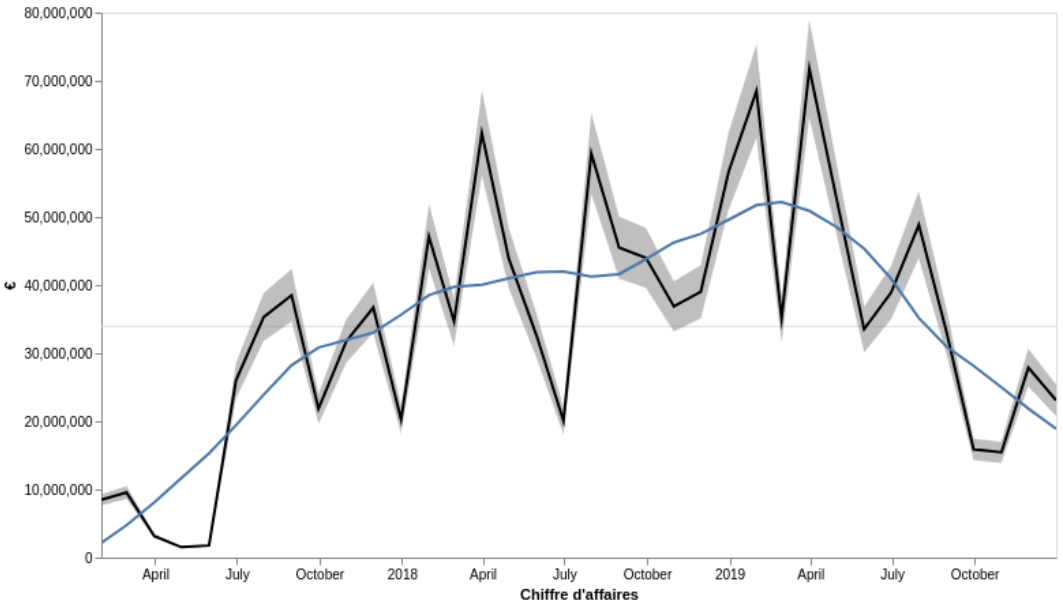


Figure 34 : Chiffre d'affaires du secteur

Pour ce graphique il était nécessaire de supprimer toutes les valeurs extrêmes, afin d'avoir comme résultat un graphique qui soit lisible et pertinent. Ces données ont étés repérées grâce à l'utilisation de la règle l'intervalle interquartile[84]. Cette méthode indique qu'une valeur est considérée comme extrême si la valeur de l'écart avec le premier ou troisième quartile est supérieure à plus de 1,5 x l'écart interquartile [33] .

L'intérêt de ce graphique tient aux faits que :

- Le volume des opportunités d'affaires (marchés) peut être élevé, mais la valeur du secteur peut être beaucoup moins importante.
- Le volume des appels d'offres peut être en augmentation mais le chiffre d'affaires du secteur être en décroissance. Par exemple, l'Etat va acheter un service ou produit de façon fractionnée, en moindre quantité qu'avant, ou à un prix inférieur.
- Le prix moyen du marché peut ne pas être assez intéressant par rapport au coût de la préparation pour répondre aux appels d'offres.
- La valeur économique du secteur peut donner une idée de la part de marché qu'a l'entreprise dans son secteur (chiffre d'affaires de l'entreprise ou le chiffre d'affaires de l'ensemble des entreprises du secteur) .
- Le chiffre d'affaires peut aider à calculer la valeur ajoutée du secteur (valeur ou richesse créée par une entreprise, un secteur d'activité ou un agent économique au cours d'une période donnée).

[84] Cet élément a été traité dans la section *II.D.1.1.3 Évaluer la cohérence et délimiter le périmètre des analyses*.

III.C.1.2.1.4 Principaux acheteurs et concurrents du secteur

Finalement l'utilisateur du produit prototype de Databiz pourra connaître les principales organisations acheteurs et ses principaux concurrents du secteur, c'est-à-dire :

- les principaux acheteurs du secteur (par rapport aux marchés achetés).
- les principaux acheteurs du secteur (par rapport aux montants (€) achetés).
- les principaux concurrents du secteur (par rapport aux marchés gagnés).
- les principaux concurrents du secteur (par rapport au CA remporté).

Les données utilisés pour ce graphique imbriqué sont indiquées au tableau suivant *Figure 34 : Dictionnaire des données de l'indicateur Principaux acheteurs et concurrents du secteur*

Dictionnaire des données					
Indicateur : Principaux acteurs du secteur : top 10 des acheteurs et des titulaires					
Description	**Champ**	**Format**	**Exemple**	**Source**	**Granularité**
Information sur le montant de l'avis d'attribution (montant, devise, gagnant..) .	budgets	object (nested field)	[{'montant': 96643.25, 'type': 'total', 'devise': 'EUR'}]	J360	AA
Information réfèrent à chaque lot présent dans l'avis d'attribution (montant du lot, gagnant..).	lot	object (nested field)	[{'budgets': [{'montant': 75000.0, 'type': 'total', 'devise': 'EUR'}], [{'titulaires': [{'id': 2716996, 'cp': '54700', 'nom': 'HOLLINGER'	J360	AA
Nom des acheteurs normalisés	acheteur_normalise	object (texte)	mairie de lagny sur marne	J360	AO
Nom des entreprises ou professionnels qui ont remporté l'avis d'attribution.	titulaires	object (texte)	DEOBAT à SENONES	J360	AA

Données calculées					
Description	**Champ**	**Format**	**Exemple**	**Librairies utilisées**	**Calcul**
Total des montants attribués.	total	float64	10000.25	Elasticsearch DSL, Pandas	Valeur totale du montant attribué enregistré de chaque annonce, plus la valeur totale des lots des annonces qui n'ont pas un montant total général enregistré.
Marchés comptés par acteur (acheteur / titulaire)	doc_count	int64	100	Elasticsearch DSL	Agrégation des annonces par acheteur / titulaire
Nom des titulaires normalisé.	titulaire_normalise	object (texte)	deobat senones	Spacy	Normalisation du texte (traitement des espaces et des accents, suppression des caractères, etc.) [85].
Liste des 10 meilleurs acheteurs et 10 gagnants du secteur en fonction du nombre des marchés achetés / gagnés et du montant total acheté / gagné.	data_best_players	object (list)	[{ highest_budgets_awardes : [{ titulaire : spie ouest centre, total : 5492923.97, ...}...]	Pandas	Somme des montants achetés ou gagnés par participant, ainsi que l'agrégation des marchés achetés ou gagnés par chaque participant.

Figure 35 : Dictionnaire des données de l'indicateur Principaux acheteurs et concurrents du secteur.

[85] Voir Annexes , section Codes.

Exemple de l'ensemble de données utilisé pour le graphique :

total	created	acheteur_normalise	titulaires
167250.0	2019-09-18 16:47:07.143482	centre hospitalier reg marseille	abiomed france
200000.0	2019-09-25 10:51:43.561980	chu brest	meditor
90000.0	2019-09-09 09:30:46.708238	centre hospitalier reg marseille	orthopedie biomeca locomotion

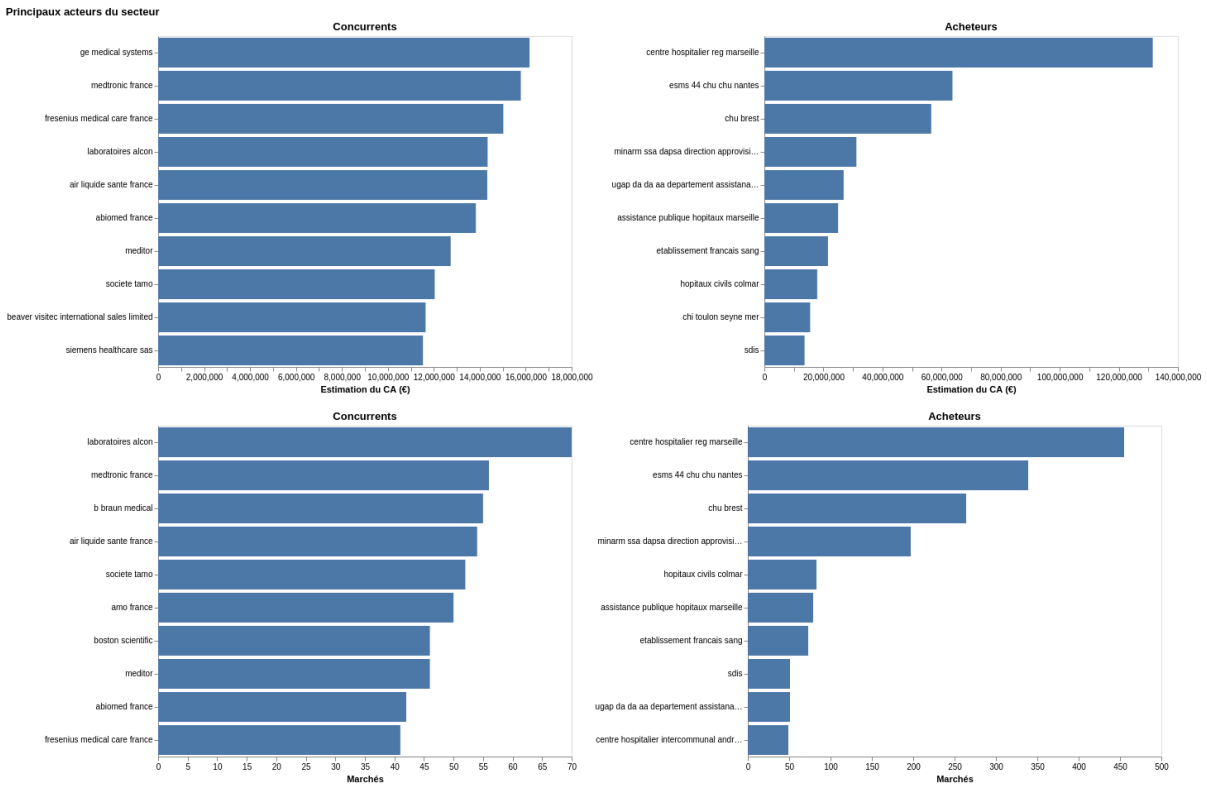


Figure 36 : Principaux acteurs du secteur.

Avant de grouper les annonces par acheteur ou titulaire, il est nécessaire de *normaliser* chacun des noms des acteurs. Cela consiste en :

- enlever les accents et des caractères spéciaux.
- convertir les majuscules en minuscules.

Une fois les noms standardisés, nous éliminons les mots considérés comme ‘stop-words’. Les mots vides sont les mots sans signification tels que les articles, les pronoms, les

prépositions, etc. Cela permet de réaliser un calcul plus fiable de similarité textuel avec l'aide de l'algorithme *Jaro winkler*[86] .
Une fois que la similarité entre les noms des acheteurs et titulaires est calculé, il est possible de clusteriser les noms qui sont assez similaires en utilisant l'algorithme *Agglomerative Clustering* de Sklearn[87].
L'implémentation de cette méthode non supervisée de machine learning a pour objectif d'éviter d'avoir plusieurs fois des entreprises ou acteurs qui sont les mêmes, par exemple "esms 06 chu nice" et "chu nice" doivent un seul acheteur "chu nice".

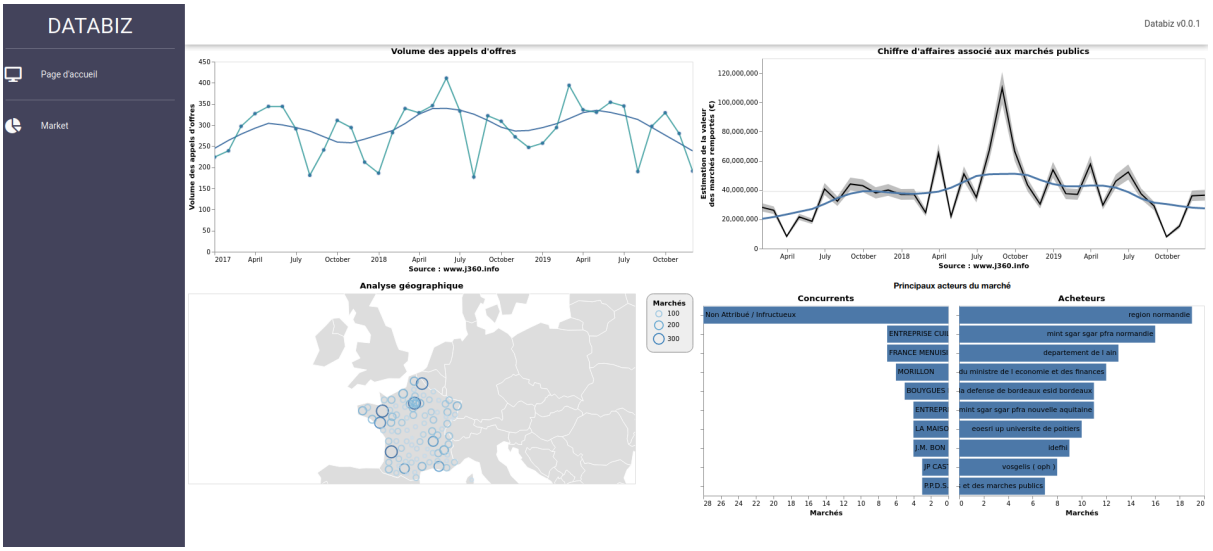


Figure 37 : Prototype de Databiz.

Néanmoins, la solution idéale à moyen terme est de tout normaliser en amont, en effet des nouvelles bases acheteurs et titulaires normalisés font partie des étapes suivantes du développement du produit.

[86] https://pypi.org/project/textdistance/
[87] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html

III.C.1.2.2 Intégration multi-sources et croisement des données externes

Une prochaine étape à développer, suite à l'achèvement du produit prototype sera l'intégration des sources externes des données.

Comtrade, le plus grand dépositaire de statistiques officielles du commerce international des Nations Unies, est la première source externe qui est prévue d'être intégrée dans le projet Databiz. Dans le tableau suivant, nous pouvons voir les principales caractéristiques des données qui la composent.

Tableau de métadonnées	
Source	comtrade.un.org
Périmètre géographique	Pays de la zone euro
Granularité géographique	Pays
Languages	EN
Périmètre temporel	2016-2018
Granularité temporel	Année
Topics	Exportations et importations
Granularité	HS2 - HS6 du système harmonisé [88]
Fréquence de mise à jour	Une fois au début du projet

Figure 38: Tableau de métadonnées - source Comtrade.

Ces données sont disponibles sous forme d'API, et elles ont étaient extraites à travers de algorithmes d'indexations (web crawler ou scrapping) basé sur le framework de Scrapy (cet élément a été détaillé dans la section *II.B.5.2.3 Web scraping*).

[88] Ce concept est défini dans le Glossaire, dans la section Annexes.

Le schéma des données complet pour Databiz est présenté ci-dessous :

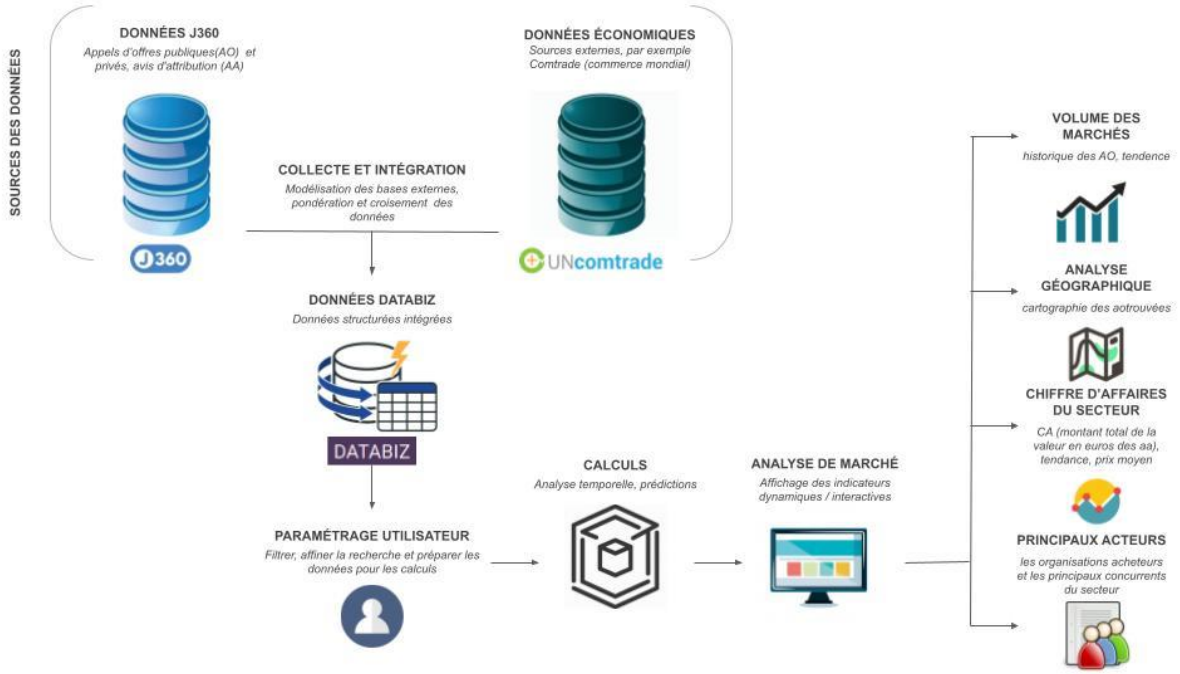


Figure 39 : Futur schéma des données - Databiz.

III.C.1.2.2.1 Application des modèles de prédiction

Cette étape de la construction du produit prototype est en cours de développement. Les modèles de prédiction seront basés sur l'intégration et le croisement des données externes, afin d'offrir aux utilisateurs de Databiz des prédictions statistiques enrichies avec ces données macro-économiques externes, ainsi que de diminuer la probabilité d'un biais potentiel issu de la complexité du secteur, notamment parce qu'il n'est pas possible de récupérer 100% des appels d'offres en France ou en Europe. De plus, les marchés publics ne représentent qu'une petite partie de tout ce qui se passe dans l'économie actuelle.

Ces prévisions seront d'abord ajoutées aux indicateurs ou histogrammes de l'*Évolution historique des marchés* et du *Chiffre d'affaires du secteur (section III.C.1.2.1.3 Chiffre d'affaires du marché).* Les données et leur traitement sont décrits dans la figure 39 : Dictionnaire des données des analyses des prédictions.

Indicateur : Prédictions					
Description	**Champ**	**Format**	**Exemple**	**Source**	**Granularité**
Date de l'enregistrement ou transaction commerciale (soit par année, soit par mois).	period	object (texte)	2016, 201601 (janvier 2016)	Comtrade	HS-2
Montant en dollars de la valeur commerciale de la transaction.	trade_value	float64	10000.25	Comtrade	HS-2
Type de flux commercial.	trade_flow	int64	1 - Imports, 2 - Exports	Comtrade	HS-2
Date de récupération de l'annonce.	created	datetime64	2017-01-01 0:00:00	J360	AA
Information sur le montant de l'avis d'attribution (montant, devise, gagnant..).	budgets	object (nested field)	[{'montant': 96643.25, 'type': 'total', 'devise': 'EUR'}]	J360	AA
Information réfèrent à chaque lot présenté dans l'avis d'attribution (montant du lot, gagnant..).	lot	object (nested field)	[{'budgets': [{'montant': 75000.0, 'type': 'total', 'devise': 'EUR'}], [{'titulaires': [{'id': 2716996, 'cp': '54700', 'nom': 'HOLLINGER'	J360	AA

Données calculées					
Description	**Nom du champ**	**Format**	**Exemple**	**Librairies utilisées**	**Calcul**
Valeur totale du montant attribué de chaque avis d'attribution, plus la valeur totale des lots des annonces qui n'ont pas un montant total général enregistré.	total	float64	10000.25	Elasticsearch DSL, Pandas	Agrégation ES des annonces par mois, avec une addition des montants (budgets) récupérés.
Marchés comptés par période.	doc_count	int64	100	Elasticsearch DSL	Agrégation des annonces par mois.
Normalisation des données.	df_normal	float64	[{'appel d'offres : 0.4665, 'ca_exportations': '0.8884', ..}]	Pandas	Les données quantitatives sont mises à l'échelle dans une petite gamme de valeurs, par exemple entre -1 et 1 ou entre 0 et 1 [52.]
Coefficients de régression linéaire multiple.	coef_	float64	0.0078	Sklearn	Ces coefficients représentent la force et le type de relation entre les variables explicatives et la variable à prévoir[89] [29].
Montants ajustés.	total_pred	float64	10000.25	Sklearn	Montants multipliés par le coefficient calculé..
Nombre de marchés comptés ajusté.	doc_count_pred	int64	100	Sklearn	Total des marchés comptés par an multiplié par le coefficient calculé.
Prédictions des valeurs ajustées pour les 6 périodes (mois) suivants.	forecast	float64	10000.25	Statsmodels	Modèle ARIMA [30, 31]

Figure 40 : Dictionnaire des données des analyses des prédictions.

[89] Ce sujet est détaillé à la fin de cette section.

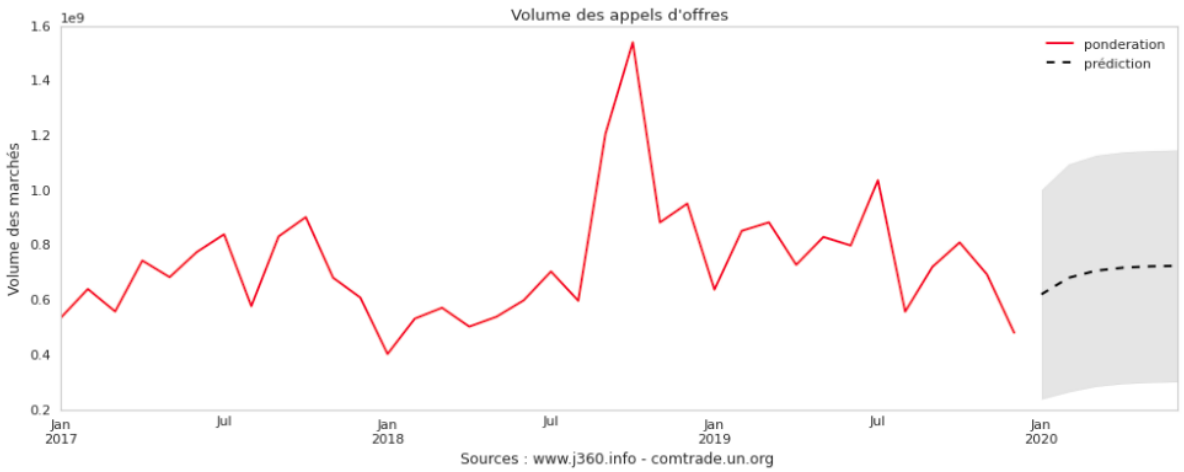


Figure 41 : Volume des marchés attendus.

Avant de calculer une projection estimée (c'est-à-dire d'appliquer un modèle de prédiction), il est nécessaire d'établir un pré-traitement des données et des tests d'acceptation, c'est-à-dire, délimiter des conditions que les données doivent suivre pour que notre modèle de prédiction soit pertinent. Ce test est composé de :

- la détection et la suppression des valeurs aberrantes[90].
- la normalisation des données et le calcul de corrélation.

Le calcul de corrélation (corr)[91] entre les importations et les exportations des produits liés aux métiers des utilisateurs de Databiz, indique s'il existe une correspondance entre ces variables et la variable à prédire (soit le volume des marchés, soit le chiffre d'affaires pour ce cas). Par exemple, un coefficient positif indique qu'à mesure que la valeur des exportations augmente, la moyenne de la variable du volume des marchés a également tendance à augmenter [32].

Si la corrélation dépasse le seuil d'acceptation [63], qui pour notre analyse est de 0.2, le test peut continuer. Ce seuil est relativement bas car il a été estimé préalablement qu'il existe une corrélation entre la commande publique et le commerce extérieur. Par exemple, l'augmentation des importations des masques chirurgicaux en France est corrélée à une augmentation des appels d'offres d'achat de ce produit.

Une fois ces hypothèses validées[92], les prédictions sont calculées, cela est illustré dans la Figure suivante - Flux des données pour le calcul des prédictions.

[90] Ce sujet a été traité dans la section *III.C.1.2.1.3 Chiffre d'affaires du marché*.
[91] www.pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.corr.html
[92] Note : la "linéarité" fait partie des hypothèses de la régression multiple. Nous considérons cette "linéarité" comme valide par défaut, du fait de la nature des données et des observations des visualisations étudiées précédemment.

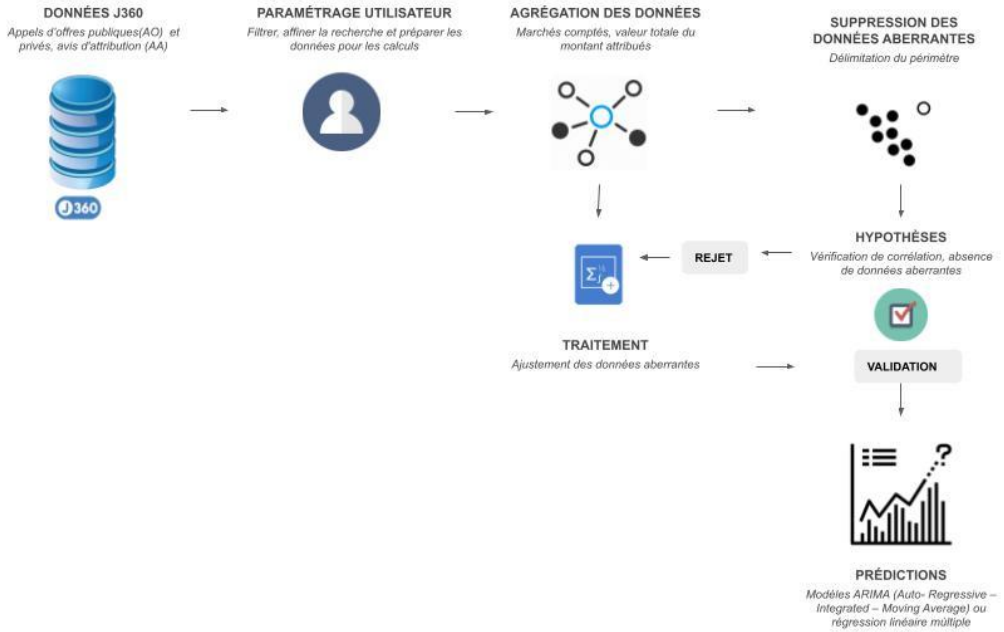


Figure 42 : Flux des données du calcul des prédictions.

Pour le calcul des prédictions, deux modèles ou techniques sont proposés. Il sera possible d'estimer le volume des marchés ou le chiffre d'affaires lorsque des prévisions des variables explicatives sont connues à l'avance (dans le cas de la source Comtrade, pour la valeur des exportations et importations).
Par exemple, avec la crise économique et sanitaire actuelle, si c'est prévu à ce que les importations diminuent de 30% l'année prochaine, il sera possible d'ajuster les données actuelles et de prédire les offres ou la valeur du secteur estimées.

Une alternative (en l'absence de données préalables suffisantes relatives aux variables explicatives) est la prédiction basée sur une stratégie autorégressive comme le modèle ARIMA (autoregressive integrated moving average). Cette technique est implémentée notamment pour les prévisions des séries temporelles car des éléments tels qu'une possible saisonnalité y sont pris en compte, et le résultat est une extrapolation de l'avenir en fonction des éléments passés [30, 31].

III.C.1.2.2.2 Préconisation des changements du marché à l'aide des réseaux sociaux.

Pour le futur l'entreprise, nous souhaitons intégrer les données des réseaux sociaux afin d'enrichir et d'affiner nos modèles de prévision[93].

La volumétrie des marchés, le chiffre d'affaires généré ou estimé et leurs prévisions peuvent être pondérés à l'aide de données économiques issues de sources de données ouvertes et des évènements détectés via les réseaux sociaux.

La détection et la prise en compte des événements au sein des réseaux sociaux liés aux différentes activités économiques des utilisateurs, peuvent nous aider à préconiser des comportements inattendus du marché. Ces évènements peuvent biaiser les prédictions statistiques, au sens négatif comme positif.

Par exemple, nous pouvons prédire une augmentation du volume des appels d'offres liés au recrutement de personnel l'année prochaine. Néanmoins, si nous ajoutons des données issues des réseaux sociaux qui parlent de la fermeture, le rachat ou la restructuration d'entreprises (ce qui peut causer des vagues du chômage, des baisses d'investissement temporels et créer de l'incertitude dans le secteur); elles pourront rajouter de la pertinence dans les prédictions.
Une future étape du projet Databiz pourrait donc être l'ajout d'un indicateur au tableau de bord de l'analyse du marché en indiquant le risque de volatilité[94] du secteur. Ce nouvel indicateur permettrait uniquement de donner une estimation du risque lié à l'instabilité du secteur.

Le flux des données et de leur traitement prenant en compte l'intégration des réseaux sociaux dans le projet de Databiz est illustré dans la figure suivante :

[93] Ce sujet a été traité dans la section *I.C.1 Analyse fondamentale et réseaux sociaux.*
[94] Le concept de volatilité est couramment utilisé en finance, et il qualifie l'instabilité d'un titre, qui fluctue dans une large fourchette, et dont les variations semblent incohérentes. La valeur du titre peut fortement baisser ou monter soudainement.

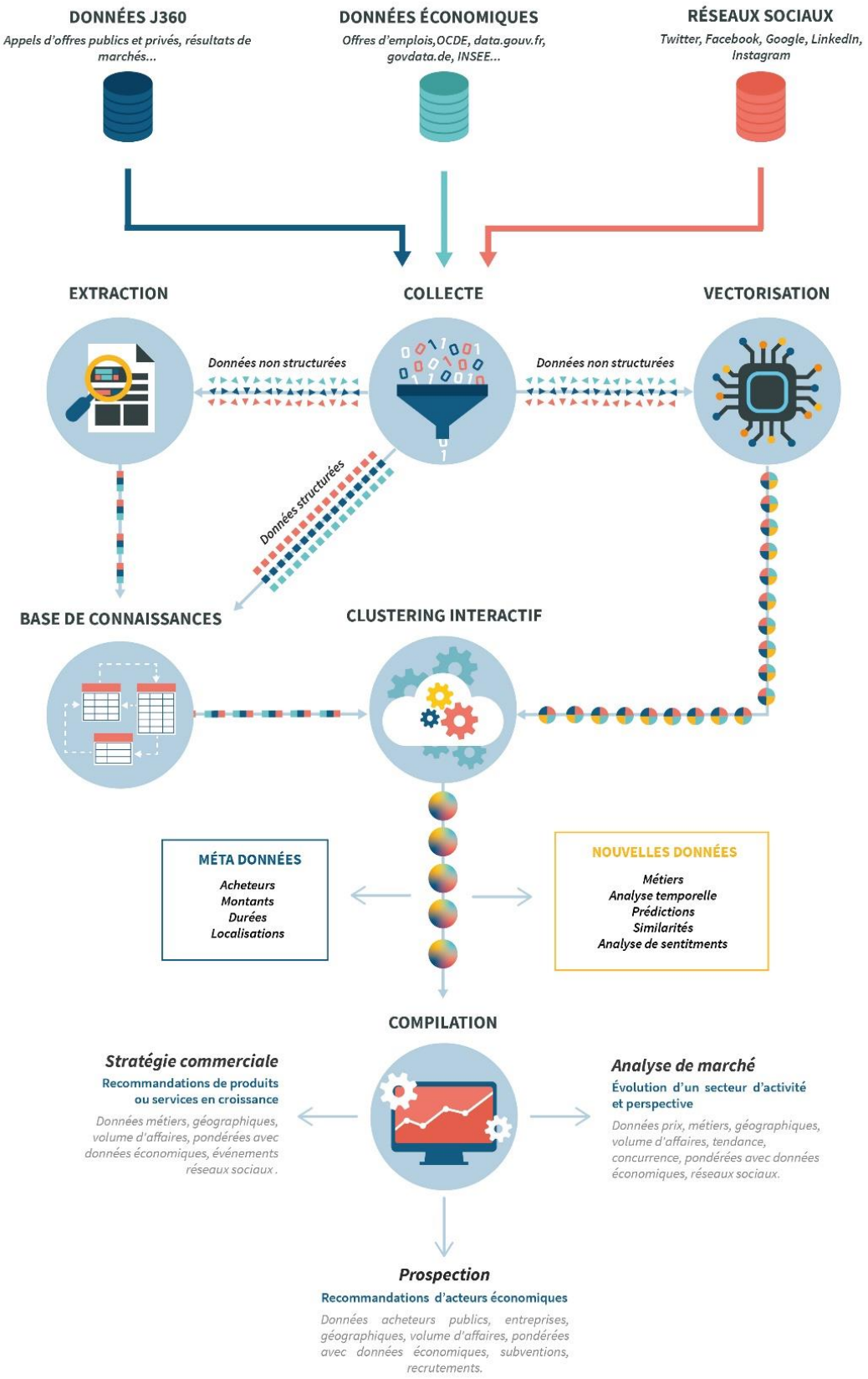


Figure 43 : Intégration des données issus des réseaux sociaux au projet Databiz[95].

[95] Graphique interne à l'entreprise OctupusMind.

CONCLUSION

L'exploitation de données ouvertes deviendra probablement une activité courante au sein des entreprises. Les potentiels bénéfices économiques que cela représente sont de plus en plus connus. En effet, tout au long de cet ouvrage ont été présentées plusieurs idées de projets commerciaux, de business models innovants, ainsi que des initiatives sociales liées à l'analyse et la valorisation de ces données.

De plus, les barrières qui découragent les entreprises et les entrepreneurs commencent à disparaître puisque d'une part, les États sont en train d'améliorer la qualité de leurs données, ainsi que d'adapter et d'agrandir leur offre des informations publiques. Et d'autre part, le développement de nouvelles technologies et techniques aident à diminuer les coûts de la collecte, du stockage, du traitement ainsi que du croisement de ces informations.

La portée de ce mouvement d'ouverture et d'exploitation des données est assez large. D'un point de vue économique, les sociétés et les entrepreneurs créent de nouvelles sources de revenus et d'emplois et en conséquence les indicateurs macroéconomiques et sociaux s'améliorent. De plus, l'ouverture des données publiques contribue à la transparence en réduisant la distance entre les acteurs économiques, les institutions et les citoyens et favorise la participation sociale aux différents processus décisionnels. Un exemple qui l'illustre est la recherche collaborative de solutions aux problèmes de santé publique. Les répercussions économiques de l'exploitation des données ouvertes entraînent donc des innovations sociales pouvant permettre d'améliorer considérablement la société.

Cependant, il est important de souligner qu'actuellement un des enjeux sur l'ouverture et la gestion de ces grands volumes de données ouvertes est l'impact environnemental que cela implique. Les besoins croissants de data centers et de leur refroidissement ont une lourde empreinte caractérisée par une forte consommation d'énergie non renouvelable, une production de déchets importante et ainsi que des émissions de CO_2.

Il est donc nécessaire de chercher des alternatives de stockage qui permettent d'atténuer cet impact écologique indésirable et de continuer ainsi à démocratiser et encourager l'ouverture, le partage, le traitement et l'exploitation des données ouvertes.
Par exemple, plusieurs entreprises leaders dans le domaine du stockage de données, travaillent sur un projet commun qui cherche à mettre en place un écosystème de Bio-stockage économiquement abordable et apte à répondre à l'explosion du volume de données numériques. Grâce à cette initiative, il a été possible de stocker 1Go de données dans un ADN (acide désoxyribonucléique) synthétisé et de récupérer ces données par la suite. Cette expérience a permis donc de prouver qu'il est possible de concevoir un système entièrement automatisé pour le stockage et la récupération de données dans une molécule d'ADN [93].
Cet exemple de solution qui nécessite, en principe, beaucoup moins d'espace et d'énergie pour fonctionner, en plus d'éviter l'extraction de métaux (rares et traditionnels), semble une solution viable à moyen terme pour un stockage plus durable des données.

Dès qu'une solution viable sera adoptée et démocratisée, il sera possible d'exploiter les avantages que l'analyse des données ouvertes peut apporter à la lutte contre le changement climatique. Ce problème prend de plus en plus d'importance parmi les problèmes "sociaux" les plus urgents à résoudre.

Comment les données ouvertes peuvent-elles aider l'écologie? L'étude des données sur la qualité de l'air, les niveaux de contamination de l'eau, etc... ainsi que le suivi des données issues des images satellitales sont des pistes à exploiter pour anticiper et trouver des solutions face au changement climatique.
Si nous souhaitons aller plus loin : est-ce qu'il serait pertinent d'étudier la perception de l'Homme sur le changement climatique? Autrement dit, est-il viable de créer un forum mondial ouvert où les participants témoignent des principaux changements et problèmes écologiques qu'ils ont perçus ? Ce forum pourrait être étudié avec des techniques de fouilles des données textuelles[96] afin de les classifier et d'évaluer si les problèmes "perçus" sont les mêmes que les problèmes prédits par les publications scientifiques. Peut-être l'enjeu est-il plus important que ne l'indiquent les modèles de prédictions, ou peut-être existe-t-il des domaines d'action prioritaire qui n'ont pas été visualisés.

Le potentiel économique, social et environnemental de l'exploitation des données ouvertes est donc considérable dans l'amélioration future de l'état de l'humanité.

[96] Ce sujet a été traité dans la section *I.B.2.1.2 Machine Learning et Deep Learning.*

BIBLIOGRAPHIE

[1] OCDE (2015), « Open government data », dans Government at a Glance 2015, Éditions OCDE, Paris, https://doi.org/10.1787/gov_glance-2015-48-en.

[2] Groupe SNCF, La Poste, SUEZ ENVIRONNEMENT, Groupe POULT sur la base de l'étude bluenove – BVA (Ed.). (2011, November). *Open Data : quels enjeux et opportunités pour l'entreprise ?* BVA Group. http://mobile.bva.fr/data/actualite/actualite_fiche/329/fichier_download22386.pdf

[3] Moore Duhen, M. M. D. (2020, February). The Economic Impact of Open Data: Latest study shows that open data is an enabler for the economy. Capgemini Invent. https://www.capgemini.com/fr-fr/news/limpact-economique-de-lopen-data-selon-la-derniere-etude-de-capgemini-invent-lopen-data-est-un-catalyseur-pour-leconomie/

[4] P. J. Stephenson, "Unblocking the flow of biodiversity data for decision-making in Africa," Biol. Conserv., 2016 https://www.researchgate.net/publication/308301822_Unblocking_the_flow_of_biodiversity_data_for_decision-making_in_Africa

[5] Y. M. González, "Política europea de reutilización de la información del sector público. De la norma Jurídica al portal de datos abiertos.," Rev. UnIII. Eur., vol. 19, pp. 113–134, 2013. https://dialnet.unirioja.es/servlet/articulo?codigo=4682830

[6] Sánchez-Padial, Antonio Jesús, 2019. "Portal de datos abiertos de origen privado para el sector privado," AgriXiv wpva8, Center for Open Science. https://www.researchgate.net/publication/336820844_Portal_de_datos_abiertos_de_origen_privado_para_el_sector_privado

[7] C. S. Albano, "Open government data : a value chain model proposal," in Proceedings of the 14th Annual International Conference on Digital Government Research, 2013, pp. 285–286. https://www.researchgate.net/publication/262398531_Open_government_data_a_value_chain_model_proposal

[8] Ortiz, Alberto. "El nudo gordiano de la apertura de datos públicos". Administraciones en Red, 31 marzo 2011. http://eadminblog.net/post/2011/03/31/el-nudo-gordiano-de-la-apertura-de-datos-publicos

9] Open Knowledge Foundation, "Manual de los Datos Abiertos," 2012.

[10] European Commission under SMART 2012/0107. (2014). Introduction to RDF & SPARQL [Slides]. Europeandataportal.Eu. https://www.europeandataportal.eu/sites/default/files/d2.1.2_training_module_1.3_introduction_to_rdf_sparql_en_edp.pdf

[11] R. A. M. Calenti and F. S. Lorenzo, Open Data y RISP generando valor social y económico ciudadanos gobierno abierto servicios. 2011. https://www.cpeig.gal/sites/default/files/libros/Open_Data_%2By_RISP.pdf

[12] ESRI (July 1998). "ESRI Shapefile Technical Description" (PDF). Retrieved 2007-07-04. https://www.esri.com/content/dam/esrisites/sitecore-archive/Files/Pdfs/library/whitepapers/pdfs/shapefile.pdf

[13] Geolittoral. (2017, September). Qu'est qu'un service WMS / WFS ? Comment utiliser celui de Géolittoral ? Note explicative. Services web d'intéropérabilité. http://www.geolittoral.developpement-durable.gouv.fr/services-web-d-interoperabilite-a803.html

[14] https://datos.gob.es/. (2020, February). *Datos abiertos más allá del sector público PUBLICADORES, MOTIVACIONES Y MODELOS DE COLABORACIÓN*. Gobierno de España. https://datos.gob.es/sites/default/files/doc/file/toc_informe_1_-_datos_abiertos_mas_alla_de_los_gobiernos_final_0.pdf

[15] Garriga-Portolà, Marc. "¿Datos abiertos? Sí, pero de forma sostenible". El profesional de la información, 2011, mayo junio, v. 20, n. 3, pp. 298-303. DOI: 10.3145/epi.2011.may.08 http://profesionaldelainformacion.com/contenidos/2011/mayo/08.pdf

[16] P. Craig Boardman, Branco L. Ponomariov, University researchers working with private companies, Technovation, Volume 29, Issue 2, 2009, Pages 142-153, ISSN 0166-4972, https://dl.acm.org/doi/abs/10.1145/2757401.2757411

[17] Anneke Zuiderwijk, Marijn Janssen, Kostas Poulis, and Geerten van de Kaa. 2015. Open data for competitive advantage: insights from open data use by companies. In Proceedings of the 16th Annual International Conference on Digital Government Research (dg.o '15). Association for Computing Machinery, New York, NY, USA, 79–88. DOI:https://dl.acm.org/doi/abs/10.1145/2757401.2757411

[18] Linked Data: Evolving the Web into a Global Data Space
Tom Heath and Christian Bizer. Synthesis Lectures on the Semantic Web: Theory and Technology, February 2011, Vol. 1, No. 1 , Pages 1-136

(https://doi.org/10.2200/S00334ED1V01Y201102WBE001)

[19] Eaves, D. E. (2019, January 4). The first decade of open data has been a win — but not for the reasons you think. Apolitical. https://apolitical.co/en/solution_article/the-first-decade-of-open-data-has-been-a-win-but-not-for-the-reasons-you-think

[20] Zuiderwijk, Anneke & Janssen, Marijn & Choenni, Sunil & Meijer, Ronald & Sheikh_Alibaks, Roexsana. (2012). *Socio-Technical Impediments of Open Data. Electronic Journal of eGovernment*. 10. 156 - 172. https://academic-publishing.org/index.php/ejeg/article/view/571

[21] Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). *Benefits, Adoption Barriers and Myths of Open Data and Open Government. Information Systems Management*, 29(4), 258–268. doi:10.1080/10580530.2012.716740 https://scinapse.io/papers/2171527181

[22] César, C y Lorenzo, S. (2010.). *Open Government: Gobierno abierto*. Jaén, España : Algón Editrores MMX, 2010. https://dspace-libros.metabiblioteca.com.co/handle/001/163

[23] P. Chen and C.-Y. Zhang, *Data-intensive applications, challenges, techniques and technologies: A survey on Big Data*, Inf. Sci. (Ny)., vol. 142 275, no. 10, pp. 314–347, 2014. https://www.sciencedirect.com/science/article/abs/pii/S0020025514000346

[24] N. Khan, I. Yaqoob, I. Abaker, T. Hashem, Z. Inayat, W. Kamaleldin, M. Ali, M. Alam, M. Shiraz, and A. Gani, *Big Data : Survey, Technologies, Opportunities, and Challenges*," Sci. World J., 2014. http://romisatriawahono.net/lecture/rm/survey/machine%20learning/Chen%20-%20Big%20Data%20Challenges%20and%20Techniques%20-%202014.pdf

[25] X. Chen, S. Member, and X. Lin, *Big Data Deep Learning : Challenges and Perspectives*, vol. 2, 2014. https://ieeexplore.ieee.org/abstract/document/6817512

[26] Yang, Mei & Nazir, Shah & Xu, Qingshan & Ali, Shaukat. (2020). *Deep Learning Algorithms and Multicriteria Decision-Making Used in Big Data: A Systematic Literature Review*. Complexity. 2020. https://www.hindawi.com/journals/complexity/2020/2836064/

[27] Dullaghan, C., & Rozaki, E. (2017). *Integration of Machine Learning Techniques to Evaluate Dynamic Customer Segmentation Analysis for Mobile Customers*. ArXiv, abs/1702.02215. https://arxIII.org/abs/1702.02215

[28] Cleveland, W. S.. "*Robust Locally Weighted Regression and Smoothing Scatterplots*." Journal of the American Statistical Association 74 (1979): 829-836. http://home.eng.iastate.edu/~shermanp/STAT447/Lectures/Cleveland%20paper.pdf

[29] Winship, Christopher and Larry M. Radbill. "*Sampling Weights and Regression Analysis.*" Sociological Methods & Research 23 (1994): 230 - 257. https://www.semanticscholar.org/paper/Sampling-Weights-and-Regression-Analysis-Winship-Radbill/d28c3fc36349a3b8a4c80f2c3bf10a1dad2cb618

[30] Didier Delignières, D. D. (2000, mars). *Séries temporelles – Modèles ARIMA*. Séminaire EA « Sport – Performance – Santé ». https://didierdelignieresblog.files.wordpress.com/2016/03/arimacomplet.pdf

[31] Cayir, Aykut & Kozan, Ozan & Dağ, Tuğçe & Yenidoğan, Işıl & Arslan, Çiğdem. (2018). *Bitcoin Forecasting Using ARIMA and PROPHET*. 10.1109/UBMK.2018.8566476. https://www.researchgate.net/publication/329400738_Bitcoin_Forecasting_Using_ARIMA_and_PROPHET

[32] Tranmer, M., Murphy, J., Elliot, M., and Pampaka, M. (2020) Multiple Linear Regression (2nd Edition); Cathie Marsh Institute Working Paper 2020-01. http://hummedia.manchester.ac.uk/institutes/cmist/archive-publications/working-papers/2020/multiple-linear-regression.pdf

[33] Erwan, E. J. (2020). *USID17 STATISTIQUES 2 - Régression* [Diapositives]. Conservatoire National Des Arts Et Métiers. https://lecnam.net

[34] López, R. (2005). Trade and Growth: Reconciling the Macroeconomic and Microeconomic Evidence. Wiley-Blackwell: Journal of Economic Surveys. https://www.semanticscholar.org/paper/Trade-and-Growth%3A-Reconciling-the-Macroeconomic-and-L%C3%B3pez/512ab58f1d90ed97a7a7291835b66893869ebe3d

[35] Félix Jiménez, 2001. "Macroeconomía: enfoques y modelos. Tomo I," Libros PUCP / PUCP Books, Fondo Editorial - Pontificia Universidad Católica del Perú, edition 1, number lde-2001-01, January.

[36] Mougeot, M. M. (1988). *Analyse micro-économique du Code des marchés publics*. Revue économique. https://www.persee.fr/doc/reco_0035-2764_1988_num_39_4_409095

[37] Patrick Bolton & Mathias Dewatripont, 2005. "Contract Theory," MIT Press Books, The MIT Press, edition 1, volume 1, number 0262025760, September. http://www.gbv.de/dms/hebis-darmstadt/toc/125875967.pdf

[38] *Dejemos de usar el término « Big Data »*. (2015, 26 août). Deloitte Colombia. https://www2.deloitte.com/co/es/pages/technology/articles/stopusingthetermbigdata.html

[39] García-Gil, D., Luengo, J., García, S., & Herrera, F. (2019). Enabling Smart Data: Noise filtering in Big Data classification. Inf. Sci., 479, 135-152. https://arxIII.org/pdf/1704.01770.pdf

[40] Rissoan, R., & Jouin, R. (2018). La boîte à outils de la Stratégie big data (BàO La Boîte à Outils) (French Edition). DUNOD.

[41] Adam, A. R. (2019, 29 janvier). The 80% Blind Spot : Are You Ignoring Unstructured Organizational Data ? *Forbes*. https://www.forbes.com/sites/forbestechcouncil/2019/01/29/the-80-blind-spot-are-you-ignoring-unstructured-organizational-data/?sh=196893f7211c

[42] Sanchez-Garcia, D. S. G. (2019). Towards the quantification of energy demand and consumption through the adaptive comfort approach in mixed mode office buildings considering climate change. *Energy and Buildings*, *187*, 173-185. https://www.sciencedirect.com/science/article/abs/pii/S0378778818330883 https://www.mdpi.com/2076-3417/10/4/1513/pdf

[43] Ritchie, H. R. (2021, 19 mars). What are the carbon opportunity costs of our food ? *Our World in Data org*. https://ourworldindata.org/carbon-opportunity-costs-food

[44] Harry Pommier, H. P. (2020). Manipulation des données structurées. GitHub. https://cours-harrypommier.gitbook.io/cnam-m1-medas-manipulation-donnees-structures/-MF5jr7uBM91OIBAPxZu/

[45] Vincent LE ROUX, V. L. R. (2020). *Etat de l'art de la BI, du Big Data et de l'IA* [Diapositives]. Le Cnam - Conservatoire National Des Arts Et Métiers. https://lecnam.net/

[46] Gómez Obando, C. (2015). INTELIGENCIA ECONÓMICA Y COMPETITIVIDAD. http://hdl.handle.net/10251/55356.

[47] Hernández Gómez, J. R. (2011). Inteligencia económica. LOGOS CIENCIA & TECNOLOGÍA, 3(1), 1-19. https://dialnet.unirioja.es/servlet/articulo?codigo=4166171

[48] Équipe de Manager GO. (2020, 4 septembre). Intelligence économique : le cycle du renseignement. Manager GO. https://www.manager-go.com/intelligence-economique/

[49] Iniciativa de datos abiertos del Gobierno de España. (2019). Guía práctica para la publicación de Datos Abiertos usando APIs. datos.gob.es. https://datos.gob.es/sites/default/files/doc/file/guia_publicacion_apis.pdf

[50] Krotov, Vlad & Silva, Leiser. (2018). Legality and Ethics of Web Scraping. https://www.researchgate.net/profile/Vlad-Krotov/publication/324907302_Legality_and_Ethics_of_Web_Scraping/links/5aea622345851588dd8287dc/Legality-and-Ethics-of-Web-Scraping.pdf

[51] Landers, Richard & Brusso, Robbie & Cavanaugh, Katelyn & Collmus, Andrew. (2016). A Primer on Theory-Driven Web Scraping: Automatic Extraction of Big Data From the Internet for Use in Psychological Research. Psychological Methods. 21. 10.1037/met0000081.https://www.researchgate.net/publication/303502399_A_Primer_on_Theory-Driven_Web_Scraping_Automatic_Extraction_of_Big_Data_From_the_Internet_for_Use_in_Psychological_Research

[52] Hernández, C. L. (2008, mars). *Preprocesamiento de datos estructurados - Structured Data Preprocessing*. Universidad Distrital Francisco José de Caldas. https://revistas.udistrital.edu.co/index.php/vinculos/article/view/4123/5790

[53] Claudia Marinica (2020). *Fouille de données* [Diapositives]. Conservatoire National Des Arts Et Métiers. https://lecnam.net

[54] Planchon, Viviane. (2005). Traitement des valeurs aberrantes : concepts actuels et tendances générales. Biotechnologie, Agronomie, Société et Environnement. 9. https://www.researchgate.net/publication/26406403_Traitement_des_valeurs_aberrantes_concepts_actuels_et_tendances_generales/fulltext/0e6053d3f0c46d4f0ab02728/Traitement-des-valeurs-aberrantes-concepts-actuels-et-tendances-generales.pdf

[55] Erwan, E. J. (2020). *Analyse des données* [Diapositives]. Conservatoire National Des Arts Et Métiers. https://lecnam.net

[56] Dupuy, L. D. (2020). Bases de données et modélisations [Diapositives]. Conservatoire National des Arts et Métiers. https://www.cnam.fr/

[57] Besse, Philippe & Gall, Caroline & Raimbault, Nathalie & Sarpy, Sophie & Toulouse, Motorola & Eads, Airbus & Toulouse, — & Informatique, Carso & Balma, —. (2001). Data Mining et Statistique. https://www.math.univ-toulouse.fr/~besse/pub/sfdsmin.pdf

[58] Magdalena Auvinet, M. A., & Lucie Cellier, L. C. (2014). DETECTION ET TRAITEMENT DES VALEURS EXTREMES ET INFLUENTES DANS LA MESURE D'AUDIENCE

INTERNET. 8ème Colloque Francophone sur les Sondages.
http://paperssondages14.sfds.asso.fr/submission_77.pdf

[59] Principles of Data Mining, David Hand, Heikki Mannila and Padhraic Smyth ISBN: 026208290. The MIT Press © 2001 (546 pages)
http://box.cs.istu.ru/public/docs/other/_New/Books/Misc/Principles%20of%20Data%20Mining.pdf

[60] Oussama Ahmia, O. A. (2021). *Analyse des données Formation Machine Learning Chapter 1: The basics* [Diapositives]. OctopusMind. https://www.octopusmind.info/

[61] Alexander Y Sun and Bridget R Scanlon 2019 Environ. Res. Lett. 14 073001,
https://iopscience.iop.org/article/10.1088/1748-9326/ab1b7d/meta

[62] Lt Col Rahul Dutt Sharma, R. D. S. (2019). Python Tools for Big Data Analytics. *International Journal of Science and Research (IJSR)*, 597-601.
https://www.ijsr.net/archive/v9i5/SR20507222308.pdf

[63] Guay, J. H. (2014). *Statistiques en sciences humaines avec R*. De Boeck.

[64] DataCamp Team. (2020). *Choosing Python or R for Data Analysis ? An Infographic*. Datacamp.
https://www.datacamp.com/community/tutorials/r-or-python-for-data-analysis

[65] Lord, Laurel & Sell, John & Bagirov, Feyzi & Newman, Mark. (2018). Survival Analysis within Stack Overflow: Python and R. 10.1109/Innovate-Data.2018.00015.
https://www.researchgate.net/publication/327202224_Survival_Analysis_within_Stack_Overflow_Python_and_R

[66] Luburic, N., & Ivanovic, D. (2016). Comparing Apache Solr and Elasticsearch search servers.
http://www.eventiotic.com/eventiotic/files/Papers/URL/icist2016_54.pdf

[67] Laure Bourgois. (2019). ELASTIC STACK. [Diapositives] . CODASCHOOL.www.codataschool.com

[68] Nextdecision. (2020). Outil ETL ou Script ? www.next-decision.fr.
https://www.next-decision.fr/wiki/outil-etl-script

[69] Nassim KHOULALEN. (2020). Formation TALEND. [Diapositives]. Business&Decision. https://lecnam.net

[70] Lucivero, Federica. (2020). Big Data, Big Waste? A Reflection on the Environmental Sustainability of Big Data Initiatives. Science and Engineering Ethics. 26. 10.1007/s11948-019-00171-7.
https://link.springer.com/article/10.1007/s11948-019-00171-7

[71] Mamavi, Olivier & Morin, Stephane. (2014). Quelle intelligence peut-on trouver dans les données massives ? La cas des marchés publics français. Revue internationale d'intelligence économique. 6. 131-142. 10.3166/r2ie.6.131-142.
https://www.researchgate.net/publication/272491702_Quelle_intelligence_peut-on_trouver_dans_les_donnees_massives_La_cas_des_marches_publics_francais

[72] Sarah Djemili, Julien Longhi, Claudia Marinica, Dimitris Kotzinos, Georges-Elia Sarfati. What does Twitter have to say about ideology?. NLP 4 CMC: Natural Language Processing for Computer-Mediated Communication / Social Media - Pre-conference workshop at Konvens 2014, Oct 2014, Hildesheim, Germany. http://www.uni-hildesheim.de/konvens2014/data/konvens2014-workshop-proceedings.pdf: p.16-25. ⟨halshs-01058867v2⟩
https://hal.sorbonne-universite.fr/ETIS-MIDI/halshs-01058867v2

[73] Ruppert, E., Grommé, F., Ustek-Spilda, F. & Cakici, B. (2018). Citizen Data and Trust in Official Statistics. Economie et Statistique / Economics and Statistics, 505-506, 171–184.
https://doi.org/10.24187/ecostat.2018.505d.1971

[74] Open Data Support by the European Commission. (2013). *Le cycle de vie des données et métadonnées publiques liées ouvertes* [Diapositives]. https://data.europa.eu/.
https://data.europa.eu/sites/default/files/d2.1.2_training_module_2.1_the_linked_open_government_data_lifecycle_fr_edp.pdf

[75] Furche, T., Gottlob, G., Libkin, L., Orsi, G., & Paton, N. W. (2016). Data Wrangling for Big Data: Challenges and Opportunities. In Advances in Database Technology — EDBT 2016: Proceedings of the 19th International Conference on Extending Database Technology (pp. 473-478). (Advances in Database Technology).
https://doi.org/10.5441/002/edbt.2016.44

[76] Y. Peng, C. Hsu, and P. Huang, “Developing Crop Price Forecasting Service Using Open Data from Taiwan Markets,” in Technologies and Applications of Artificial Intelligence (TAAI), 2015, pp. 172–175.
https://ieeexplore.ieee.org/document/7407108

[77] Pioneer. (2014). *Pioneer Field360 Tools App for Crop Management Decisions*. DuPont Pioneer Agronomy Sciences.

https://www.pioneer.com/CMRoot/Pioneer/Canada_en/Programs_Services/mobile/Pioneer_Field360_Tools_App.pdf

[78] Pabón, Oswaldo Solarte; Torres, José Heriberto; Bucheli, Víctor Andrés.Revista Ibérica de Sistemas e Tecnologias de Informação; Lousada N° 39, (Oct 2020): 52-66. DOI:10.17013/risti.39.52-66
https://www.proquest.com/openview/c18095a019385a0fd2598cb9d0ff5dde

[79] K. Bakshi, "Considerations for big data: Architecture and approach," 2012 IEEE Aerospace Conference, 2012, pp. 1-7, doi: 10.1109/AERO.2012.6187357.
https://ieeexplore.ieee.org/abstract/document/6187357

[80] European Data Portal. (2020, juillet). Re-using Open Data - A study on companies transforming Open Data into economic & societal value (OA-03-20-042-EN-N). Luxembourg : Publications Office of the European Union. https://doi.org/10.2830/876679
https://data.europa.eu/sites/default/files/re-using_open_data.pdf

[81] COTEC. (2019). Guía para la apertura y compartición de datos en el entorno empresarial. Fundación COTEC para la innovación.
https://cotec.es/proyecto/el-potencial-del-open-data-en-el-sector-privado/975e45f8-ce1e-494f-b477-397358cc1765

[82] Barrau, Delphine & Barthélémy, Nathalie & Kedad, Zoubida & Laboisse, Brigitte & Nugier, Sylvaine & Thion, Virginie. (2016). Gestion de la qualité des données ouvertes liées - État des lieux et perspectives.
https://www.researchgate.net/publication/304262365_Gestion_de_la_qualite_des_donnees_ouvertes_liees_-_Etat_des_lieux_et_perspectives

[83] Innovations in Smart Cities Applications Volume 4. 2021; 183: 1282–1296. Published online 2020 Dec 10. doi: 10.1007/978-3-030-66840-2_98
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7972548/

[84] University of Gothenburg, & Laura Natali, L. N. (2021, avril). *Machine learning can help slow down future pandemics*. The Faculty of Science - University of Gothenburg.
https://www.gu.se/en/news/machine-learning-can-help-slow-down-future-pandemics

[85] Amazon, & Han Man, H. M. (2018, juin). *Analyze US census data for population segmentation using Amazon SageMaker*. AWS Machine Learning Blog.
https://aws.amazon.com/fr/blogs/machine-learning/analyze-us-census-data-for-population-segmentation-using-amazon-sagemaker/

[86] Abe, Hidenao & Ohsaki, Miho & Yokoi, Hideto & Yamaguchi, Takahira. (2005). Evaluating an integrated time-series data mining environment–a case study on a chronic hepatitis data mining.
[87] Yiyo Kuo, Taho Yang, Guan-Wei Huang, The use of grey relational analysis in solving multiple attribute decision-making problems, Computers & Industrial Engineering, Volume 55, Issue 1, 2008, Pages 80-93, ISSN 0360-8352, https://doi.org/10.1016/j.cie.2007.12.002.
https://www.sciencedirect.com/science/article/abs/pii/S0360835207002732

[88] La Commission Nationale de l'Informatique et des Libertés. (2017, août). *Open data : la protection des données comme vecteur de confiance*. CNIL.
https://www.cnil.fr/fr/open-data-la-protection-des-donnees-comme-vecteur-de-confiance

[89] Towards data science, & Rashi Desai, R. D. (2019, December). Top 10 Python Libraries for Data Science. Towardsdatascience.com.
https://towardsdatascience.com/top-10-python-libraries-for-data-science-cd82294ec266

[90] Escande, P. (2017, 30 mars). BlackRock : « Aladdin » et l'investissement merveilleux. Le Monde.fr.
https://www.lemonde.fr/economie/article/2017/03/30/blackrock-aladdin-et-l-investissement-merveilleux_5103222_3234.html

[91] Lajnef, Mohamed-Ali & Ben Ayed, Mounir & Kolski, Christophe. (2005). Convergence possible des processus du data mining et de conception-évaluation d'IHM: adaptation du modèle en U. 243-246. 10.1145/1148550.1148587.
https://www.researchgate.net/publication/220745687_Convergence_possible_des_processus_du_data_mining_et_de_conception-evaluation_d'IHM_adaptation_du_modele_en_U/citation/download

[92] Open Data Camden. (2015). *Population segmentation*. Islington Council.
https://data.gov.uk/dataset/41f3e46c-e039-433d-af58-6d3624fe6f3e/camden-demographics-population-segmentation-supplementary-analysis-2015

[93] Vitak, S. (2021). Technology alliance boosts efforts to store data in DNA. Nature. Published. https://doi.org/10.1038/d41586-021-00534-w

ANNEXES

Formats des données ouvertes

JSON

Le format JSON ou JavaScript Object Notation est un format léger d'échange de données basé sur la notation des objets JavaScript, sa syntaxe est très simple, ce qui le rend facile à lire par n'importe quel langage de programmation, cela signifie que les ordinateurs Ils peuvent le traiter plus facilement vers d'autres formats tels que XML (eXtensible Markup Language). Ce dernier est largement utilisé dans l'échange de données, car il offre de grandes opportunités pour maintenir la structure des données et la façon dont elles sont construites, en plus de permettre aux développeurs d'écrire des parties de la documentation sans interférer avec sa lecture.

Un exemple de ce format de données peut être l'extrait suivant du jeux de données du *Référentiel des communes 2015 en Pays de la Loire* [97] qui décrit les caractéristiques ou informations géographiques stockées de la commune Aigrefeuille-sur-Maine.


```
[{
	"datasetid": "communes-2015",
	"recordid": "91bd87028eca10143b0a523799c705cddf463e07",
	"fields": {
		"code_postal": "44140",
		"objectid": 2,
		"st_length_shape": 20281.5609802155,
		"insee_comm": 44002,
		"st_area_shape": 14700903.0,
		"gml_id": "sig.referentiels.pdl_communes_2015.2",
		"libelle_commune_actuel": "Aigrefeuille-sur-Maine",
		"geom": {
			"type": "Polygon",
			"coordinates": [
				[
					[-1.4417655427, 47.0521620443],
					[-1.4353111425, 47.0465651485]

				]
			]
		},
		"insee_commune_actuel": "44002",
		"centroid": [47.071441629, -1.415354875],
		"nom_commun": "Aigrefeuille-sur-Maine"
	},
	"geometry": {
		"type": "Point",
		"coordinates": [-1.415354875, 47.071441629]
	},
	"record_timestamp": "2021-04-08T04:01:11.243+02:00"
}]
```


[97] data.paysdelaloire.fr/explore/dataset/communes-2015

XML

Le format XML a été développé par le World Wide Web Consortium (W3C), un consortium international renommé où les organisations membres, le personnel à plein temps et le grand public travaillent ensemble pour développer des normes Web. Ce format XML est une norme de création post-SGML.

SGML est l'acronyme de Standard Generalized Markup Language, qui consiste en un système d'organisation et d'étiquetage des documents, normalisé par l'Organisation internationale de normalisation (ISO). Il est utilisé pour spécifier les règles de balisage des documents et n'impose lui-même aucun jeu de balises spécial (par exemple, le langage HTML est défini en termes de SGML).

Mais XML est plus facile à implémenter car il évite certaines fonctionnalités SGML avancées. Grâce à cela, est proposé comme standard pour l'échange d'informations structurées entre différentes plates-formes, car il peut être utilisé dans des bases de données, des éditeurs de texte, des feuilles de calcul, etc.

Le flux RSS (Really Symple Sindication) est un type de format XML pour la distribution de contenu de page Web. Il facilite la publication d'informations mises à jour, généralement en temps réel, sans utiliser de navigateur, en utilisant un logiciel spécialisé dans ce format.

Par exemple, le site Open Data des Pays de la Loire (data.paysdelaloire.fr) permet de télécharger le catalogue des données du trafic des Transports en commun - TAN en temps réel, en utilisant le flux RSS suivant :

```
<rss version="2.0" hv="a3">

    <channel>

      <title>Datasets</title>

      <link>http://data.paysdelaloire.fr/explore/</link>

      <description>Datasets for the domain 'paysdelaloire'</description>

      <item>

        <title>Fluidité des axes routiers de Nantes Métropole</title>

        <link>http://data.paysdelaloire.fr/explore/dataset/244400404_fluidi
        te-axes-routiers-nantes-metropole@nantesmetropole/</link>

        <description><p> Indicateurs de fluidité du trafic sur les tronçons
        routiers de Nantes Métropole. </p><hr/>Plus d'informations sont
        disponibles sur le site <a
        href="https://metropole.nantes.fr/infocirculation"
        target="_blank">metropole.nantes.fr</a>. <hr
        id="horizontalrule"/></description>

        <pubDate>Thu, 08 Apr 2021 09:12:07 GMT</pubDate>
```

```
      </item>
  </channel>
</rss>
```

CSV

Un autre type de format sont les fichiers CSV ou de valeurs séparées par des virgules, qui représentent de manière simple, des données au format tableau, en séparant les colonnes par des virgules (ou des points-virgules) et les lignes par des sauts de ligne. [9]

Pour revenir à notre exemple des informations géographiques stockées de la commune Aigrefeuille-sur-Maine sur le site data.paysdelaloire.fr, les mêmes données seraient présentés dans ce type de format (csv) de la façon suivante :

```
gml_id;objectid;insee_comm_2015;nom_commun_2015;st_area_shape_;st_length_sh
ape_;geom_2015;centroid_2015;libelle_commune_actuel;insee_commune_actuel;Co
de_postal_actuel
sig.referentiels.pdl_communes_2015.2;2;44002;Aigrefeuille-sur-Maine;1470090
3.0;20281.5609802155;"{""type"": ""Polygon"", ""coordinates"":
[[-1.4417655427, 47.0521620443], [-1.4353111425,
47.0465651485]]}";47.071441629,-1.415354875;Aigrefeuille-sur-Maine;44140
```

Formats géospatiaux

En ce qui concerne les données géospatiales, les formats les plus utilisés sont GeoJSON, Shapefile et KML, ainsi que les flux de données en temps réel qui sont de plus en plus utilisés, par exemple WFS, WMS, etc.

SHP Shapefile est un format de “données spatiales propriétaires standard”, développé par la société ESRI, qui stocke à la fois des informations géométriques et alphanumériques et peut en général être réutilisable par les systèmes d'information géographique (SIG). Ce format est composé de 4 fichiers toujours joints. Le fichier .shp contient la géométrie des entités, le .dbf a les attributs au format dBase, le .shx l’index et le fichier ayant l’extension .prj contient les informations sur le système de coordonnées. Il peut y avoir encore d’autres fichiers associés aux données shapefile. [12]

Les flux wms (Web Map Service) produisent des cartes à partir d'informations géographiques vectorielles présentant les informations sous forme d'images numériques pouvant être visualisées à l'écran aux formats PNG (Portable Network Graphics), GIF ou JPEG et parfois, ils sont représentés sous forme d'informations vectorielles au format SVG (Scalable Vector Graphics), optimal pour les diagrammes complexes nécessitant un zoom ou une visualisation par couches.

Les flux WFS permettent d'afficher des couches vecteur, non directement modifiables mais qui peuvent être téléchargées au format shapefile. Les cartes affichées peuvent se superposer, tant que les paramètres géographiques et la taille de sortie sont identiques. [13]

Plus bas, un exemple de ce type de données est le flux WMS du site *vuduciel.loire-atlantique.fr* qui permet de visualiser les photographies aériennes de la Loire-Atlantique :

```
<ows:ExceptionReport xmlns:xs="http://www.w3.org/2001/XMLSchema"
xmlns:ows="http://www.opengis.net/ows"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" version="1.0.0"
xsi:schemaLocation="http://www.opengis.net/ows
https://wms-vuduciel2.makina-corpus.net/geoserver/schemas/ows/1.0.0/owsExce
ptionReport.xsd">
<ows:Exception exceptionCode="InvalidParameterValue" locator="service">
<ows:ExceptionText>No service: ( ows )</ows:ExceptionText>
</ows:Exception>
</ows:ExceptionReport>
```

PDF

Les PDF ou Portable Document Format sont un autre type de spécification ouverte et multiplateforme (elle peut être présentée dans les principaux systèmes d'exploitation: Windows, Mac ou Unix / Linux). Il a été développé par Adobe Systems et permet de stocker des documents de type composé (image vectorielle, bitmap et texte), c'est-à-dire qu'il a été conçu pour les documents pouvant être imprimés, car il spécifie toutes les informations nécessaires à la présentation finale du document, déterminant tous les détails de son apparence, ne nécessitant pas de processus d'ajustement ou de mise en page préalable.

Il convient de se rappeler que l'idéal est de proposer au public des données et non des documents. Cela implique qu'il n'y a aucun sens d'ouvrir des fichiers au format pdf, Word ou similaires. Un service Open Data n'est pas un service ”Open Acces”, c’est plutôt un site où les données brutes (raw data) sont proposées pour être réutilisées et créer d'autres services publics. Pour cette raison, les documents ne peuvent pas être proposés comme informations dans un portail de données ouvertes. [11]

Structure des données ouvertes

Données structurées

Les données structurées sont des données dont le format est bien défini. Il s'agit généralement de dates, de nombres (données quantitatives) et de chaînes de caractères. Ils sont généralement stockés dans des tables. Dans les entreprises, ces données se trouvent dans les informations obtenues à partir du CRM (Customer Relationship Management), de l'ERP (Enterprise Resource Planning), SRM (Supplier Relationship Management) ainsi que issus de la récupération de données externes (données exogènes) [24, 44, 45].

Ces données sont stockées dans un entrepôt de données (data warehouse)[98] si le volume est considérablement grand, mais en cas contraire, si l'entreprise ou la société ne génère pas une grande quantité de données, ces informations peuvent être stockées dans une base de données relationnelle. Pour interroger ces données, il est utilisé généralement des requêtes SQL [23].

Les données structurées sont donc, l'épine dorsale des bases de données relationnelles. Dans les modèles relationnels, toutes les informations sont stockées dans un schéma de tables et ces tables auront des champs définis et des relations entre elles (cet élément est développé en détail dans la section *II.C.3.1.1.1 Système de Gestion de Base de Données Relationnelles*).

Les données structurées peuvent être :

- Données issues de capteurs : comme celles des GPS, des compteurs électriques, des équipements médicaux, etc.
- Web Log Data : les serveurs, les réseaux, les applications, etc. génèrent de grandes quantités de données structurées.
- Données provenant des points de vente : par exemple les enregistrements des passages des codes-barres dans un lecteur.
- Données financières : de nombreuses opérations bancaires et boursières sont des données structurées qui sont générées automatiquement.

Données non structurées

Contrairement aux données structurées, ce type de données n'ont pas un modèle prédéfini. Elles sont le plus souvent classées comme des données qualitatives et ne peuvent pas être traitées et analysées à l'aide d'outils et de méthodes classiques [23, 24].

[98] Ce sujet est traité dans la section *II.C.3.1.1.3 Entrepôt de données.*

Les origines de ce type de données sont :

A. Données générées par des machines et des ordinateurs.
- Des images satellites.
- Données scientifiques : graphiques sismiques, graphiques atmosphériques, etc...
- Photographie et vidéo : par exemple, les caméras de surveillance.
- Données collectées à partir des sonars et des radars (positionnement par smarthphone, etc.).

B. Données générées par des personnes :

- Textes inclus dans les systèmes d'information internes des organisations : par exemple: présentations, courriels, etc.
- Données provenant des réseaux sociaux.
- Données provenant de nos appareils mobiles : par exemple, les messages texto.
- Contenu des sites web : par exemple, des vidéos YouTube.

Presque 80% des données d'une entreprise ne sont pas structurées (*Adam Rogers, conseiller technologique de Forbes, 2019*). Dans la plupart des organisations, ces données sont “ignorées”, mais ces informations peuvent être particulièrement précieuses en raison de leur capacité à fournir une vue riche, détaillée et qualitative de ce qui se passe réellement au sein d’une organisation, c'est-à-dire que ces informations peuvent aider à résoudre la question, non seulement de ce qui se passe, mais aussi pourquoi cela est en train de se produit. Un exemple des utilisations possibles des données non structurées est l'analyse des sentiments des employés, au sein du département de RH, à l'aide de modèles de traitement du langage naturel (PNL) [41].

Données semi-structurées

Les données semi-structurées sont un mélange de données structurées et non structurées, c'est-à-dire que ces données suivent une sorte de structure implicite, mais pas assez régulière pour pouvoir être gérées et automatisées comme des informations structurées [24].

Ces données ont la particularité que leur format évolue vers un protocole, avec une série de caractéristiques “déterminantes”. Il est possible de dire que les données semi-structurées ont leurs propres "métadonnées semi-structurées", qui décrivent les objets, donc les relations entre eux peuvent être déduites [24].

Les avis de décès, les demandes d'emploi, les listes de biens immobiliers, les avis juridiques ou les noms de comptes bancaires sont des exemples de ces données.

Principaux avantages de l'ouverture des données économiques

1. *Reproductibilité*. Pour que la recherche économique soit fiable et digne de confiance, il devrait être possible d'examiner et de reproduire les résultats. Cela est difficile, voire impossible, si les données et les analyses ne sont pas disponibles. Rendre le matériel ouvertement disponible réduit au minimum les obstacles à la réalisation de recherches reproductibles.

2. *La connaissance en tant que bien public*. Les données doivent être considérées comme un bien public. La recherche financée par des fonds publics et effectuée dans l'intérêt public, devrait être aussi librement accessible au public.

3. *Stabilité et efficacité des marchés*. Des informations transparentes et disponibles peuvent être essentielles au bon fonctionnement des marchés. La meilleure façon de garantir la transparence et de garantir que les informations sont disponibles pour toutes les parties concernées, y compris les régulateurs et les chercheurs, est de rendre les données ouvertes.

4. *Engagement et confiance du public*. L'économie, et en particulier les données et les analyses économiques, jouent un rôle important dans de nombreux domaines de l'élaboration des politiques qui affectent directement tous les membres de nos sociétés. À ce titre, l'engagement et la confiance du public sont importants et l'ouverture est essentielle pour gagner et conserver cette confiance, ainsi que d'accroître l'engagement social.

5. *Nouvelles utilisations potentielles des données*. Dans de nombreux cas, la meilleure utilisation des données peut finalement être trouvée en dehors de leur utilisation immédiate et la mise à disposition des données peut générer de nouvelles recherches et créer de nouvelles connaissances.

6. *Accès équitable*. Les chercheurs et les instituts de recherche du monde entier, peuvent accéder à la recherche, aux données et aux analyses économiques sans discrimination quant à leur affiliation, leur objectif de recherche ou leur capacité à payer pour l'accès.

7. *Un impact plus élevé de la recherche*. Rendre la recherche et les données économiques librement accessibles permet une meilleure diffusion des résultats de la recherche et améliore la visibilité et l'impact de la recherche.

8. *Démocratisation de la recherche économique*. Une grande partie de la recherche économique est effectuée dans le but d'améliorer l'économie, les politiques et les institutions. La recherche économique ouverte mènera à un engagement plus élevé des citoyens menant à de meilleures politiques et à une vie meilleure.

9. *De meilleures ressources pour l'éducation et la formation*. L'ouverture de la recherche économique aide à former une nouvelle génération d'économistes et de spécialistes des sciences sociales qui sera en mesure de produire des recherches de haute qualité.

10. *Meilleure prestation de services et nouveaux modèles d'entreprise*. Les données ouvertes peuvent améliorer la qualité et la cohérence des services publics en révélant les inefficacités et la corruption et en proposant de nouvelles idées sur l'utilisation efficace des ressources publiques. Cela peut également entraîner une meilleure intégration des chaînes d'approvisionnement, exploiter l'innovation, révolutionner les modèles commerciaux et stimuler l'esprit d'entreprise, générant des externalités de connaissances dans l'économie.

Modélisation des bases des données relationnelles

La modélisation des données consiste à analyser puis concevoir une représentation des informations à stocker.
L'analyse peut prendre deux formes, soit l'analyse des données existantes pour réutiliser, mutualiser et faire évoluer des modèles existants, soit l'analyse des nouveaux besoins en données. [56].

MERISE est une méthode d'analyse et de conception des bases des données qui est utilisée dans les projets complexes et de grandes ampleurs, ainsi que pour modéliser les SGBD relationnelles. La méthode MERISE permet une analyse systémique entre les données, leurs traitements et leurs flux [45, 56] .

On retrouve trois type de schémas dans le modèle MERISE de données :

- MCD (abstrait) : Modèle Conceptuel de Données. Représentation des concepts et des relations. Le but est ici de définir le Quoi.
- MLD : Modèle Logique de Données. Transcription du niveau conceptuel en pseudo langage informatique. Avec quoi est ici la question que propose de résoudre le niveau logique.
- MPD : Modèle Physique de Données : ensemble de script SQL qui automatise la génération de la base de données [56].

Dans les trois cas, les principales étapes de la modélisation des données sont :

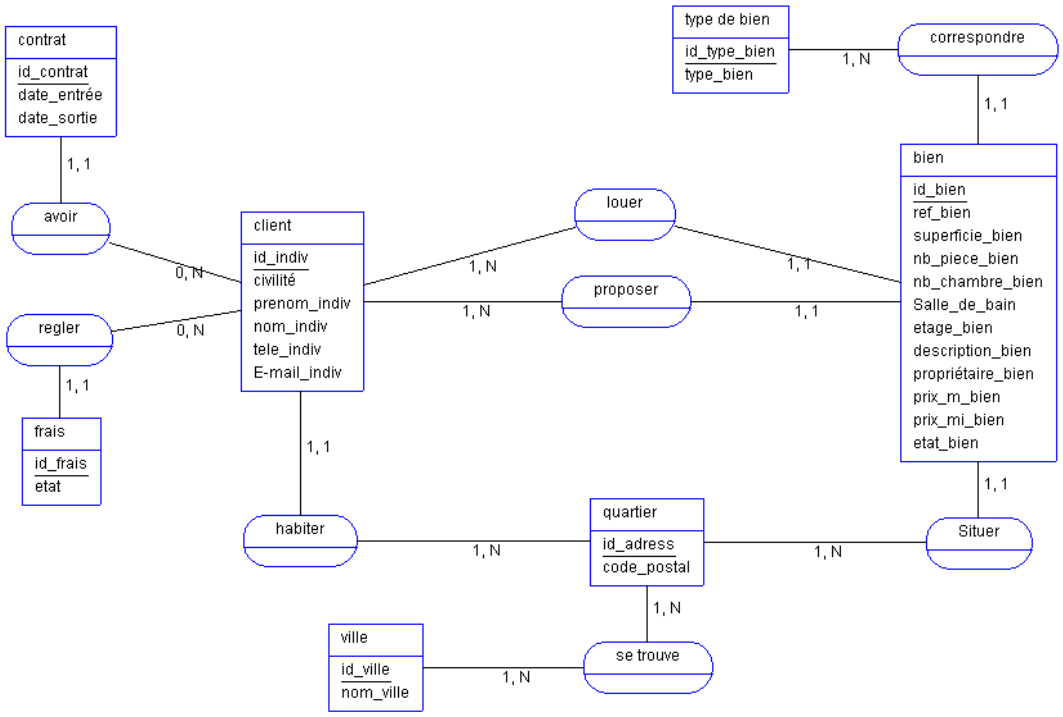
- Réaliser un dictionnaire de données : le dictionnaire décrit des données telles que les clients, les nomenclatures, les produits, les services, les annuaires, etc. Il précise la signification des données, leurs relations entre elles, leur structure, leur format ainsi que des détails utiles pour l'utiliser [56].

- Réaliser une matrice des dépendances fonctionnelles : matrice avec le nombre de colonnes et lignes correspond au nombre de données à analyser et leur granularité temporelle. Cet outil met en évidence des liens existants entre certaines données et déterminer les entités et leurs propriétés [45]. La figure suivante montre un exemple d'une matrice de dépendances d'une petite base des données d'une agence de gestion locative.

		TEMPS	BIEN	LOCATAIRES	PROPRIETAIRES
		année	code_bien	code_locat	code_prop
		mois	adresse	nom	nom
		semaine	ville	prénom	prénom
			code postal	adresse	adresse
			département	code postal	code postal
			région	ville	ville
			surface	département	département
			type bien	région	région
				âge	âge
Fait	**Indicateur / Attribut**	**TEMPS**	**BIEN**	**LOCATAIRES**	**PROPRIETAIRES**
BAIL_Marketing	Charges	mois	X	X	
	Montant contrat demandé	mois	X	X	
BAIL_CDG	Numéro de contrat	semaine	X	X	X
	Montant reversé	semaine	X	X	X
	Montant contrat demandé	semaine	X	X	X
	Charges	semaine	X	X	X
	Date début contrat	semaine	X	X	X
	Date fin contrat	semaine	X	X	X
BAIL_Paiment	Montant réellement perçu	semaine	X	X	
	Date réception montant	semaine	X	X	

Exemple d'une matrice des dépendances fonctionnelles

- Déduire la modélisation : d'une modélisation juste découle une bonne compréhension des flux d'information et de la logique métier. Les outils peuvent évoluer : si l'analyse est bonne, les données et les bases de données qui permettent d'y accéder serviront de longues années. À l'inverse, lorsque la modélisation est mauvaise ou bien lorsque la base de données est construite de façon empirique, les différentes briques ne s'accordent pas entre elles car il n'y a pas de logique commune.



Modélisation de la gestion des locations [99]

99

www.developpez.net/forums/d810641/general-developpement/alm/modelisation/schema/gestion-locations-d-agence-immobiliere

Cette figure présente l'exemple de la modélisation des données issus de la gestion des locations d'une agence immobilière.

Les nombres et lettres de chaque node représentent la *cardinalité* d'une relation est le nombre d'occurrence d'une entités qui participent à la relation, où :

- 0,1 → 0 ou 1
- 1,1 → 1 et 1 seul
- 0,n → 0 ou plus
- 1,n → Au moins 1

C'est-à-dire, par exemple, qu'un client peut louer ou proposer au moins un bien jusqu'à une infinité d'entre eux (1,N). Inversement, un bien est toujours lié à un seul client (1,1).

Transformation des données

La différence principale entre le nettoyage des données et la transformation des données est que le nettoyage des données est le processus de correction ou assouplissement des erreurs présentes d'un jeu ou d'une base de données (valeurs aberrantes, doublons, etc.), tandis que la transformation de données est le processus de conversion de données d'un format à un autre afin de faciliter leur analyse.

La transformation des données recouvre les opérations suivantes :

- Normalisation : les attributs sont mis à l'échelle dans une petite gamme de valeurs, par exemple entre -1 et 1 ou entre 0 et 1 [100].

- Agrégation : des opérations de compression ou d'agrégation sont appliquées aux données. Par exemple, les ventes quotidiennes peuvent être agrégées en ventes mensuelles ou annuelles.

- Généralisation : les données de bas niveau ou primitives sont remplacées par des concepts de plus haut niveau, en faisant appel au concept de hiérarchie. Par exemple, pour l'âge des individus, une correspondance peut être établie avec des concepts de niveau supérieur tels que jeune, adulte, personne âgée [52].

[100] Cet élément est détaillé dans la section *Annexes - Principales techniques de normalisation*.

Principales techniques de normalisation des données

Normalisation Min-Max

Elle exécute une transformation linéaire des données d'origine. Sur la base des valeurs minimale et maximale d'un attribut, une valeur de normalisation zi est calculée sur la base de la valeur xi [52] selon l'expression suivante :

$$zi=xi-min(x)max(x)-min(x)$$

Cette méthode préserve les relations entre les données d'origine. Exemple : en supposant que les valeurs minimale et maximale d'un attribut sont respectivement de 12 $ et 98 $, il est nécessaire de faire correspondre les valeurs dans une plage comprise entre 0 et 1. En prenant une valeur de 73, le résultat de la normalisation min-max est :

$$\frac{73-12}{98-12}(1-0)+0=\frac{61}{86}=0,7093$$

Normalisation z-core

Les valeurs d'un attribut sont normalisées sur la base de la moyenne et de l'écart-type [52] . Une valeur x est normalisée en calculant l'expression suivante :

$$v' = x-\mu / \eth$$

Cette méthode est utilisée lorsque le maximum et le minimum de l'attribut A sont inconnus ou lorsqu'il existe des valeurs aberrantes qui prédominent lors de l'utilisation de la normalisation min-max. Exemple : supposons que la moyenne et l'écart-type d'un attribut soient respectivement de 54 et 16 dollars. Avec la normalisation du z-core, une valeur de 73 deviendrait :

$$v'=\frac{73-54}{16}=\frac{19}{16}=1,1875$$

Normalisation de l'échelle décimale

Dans cette technique, nous déplaçons le point décimal des valeurs de l'attribut. Le nombre de points décimaux déplacés dépend de la valeur absolue maximale de A[101]. Une valeur v de A est donc normalisée à v' en calculant l'expression suivante :

$$v' = v / 10^j$$

Où j est le plus petit entier de Max(|v'|) < 1. Par exemple, supposons que l'intervalle de valeurs des enregistrements de l'attribut A soit de -986 à 917. La valeur absolue maximale de A est 986. Pour normaliser à l'échelle décimale, il faut diviser chaque valeur par 1000 (j=3), puis -986 est normalisé comme -0,986. Il faut noter que la normalisation peut modifier un peu les données originales, en particulier les deux dernières méthodes mentionnées[52].
Il est également nécessaire d'enregistrer des paramètres tels que la moyenne ou l'écart-type pour une utilisation ultérieure, afin de pouvoir les normaliser de manière uniforme.

Ces méthodes sont utilisées dans différents algorithmes de machine learning avec l'objectif de transformer toutes les données à la même échelle, car si les échelles des différentes observations sont très différentes, cela peut avoir un effet indésirable sur l'entraînement ou dans l'apprentissage (en fonction des méthodes).[102]

[101] https://uomustansiriyah.edu.iq/media/lectures/6/6_2021_06_10!10_05_09_PM.pdf
[102] Cet élément a été traité plus en détail dans la section *II.D.1.1.3 Évaluer la cohérence et délimiter le périmètre des analyses.*

Le processus ETL (Extraire - Transformer - Charger)

Un ETL ou Extract Transform Load (Extraire Transformer Charger) est un middleware (intergiciel ou logiciel tiers qui crée un réseau d'échange d'informations entre différentes applications informatiques) ou un groupement des programmes permettant d'effectuer des synchronisations de données entre différents systèmes [45].

Les processus ETL fournissent donc des données de qualité et cohérentes qui peuvent être stockées et traitées dans une phase d'analyse ultérieure. Il est important de souligner la complexité des tâches ETL, cette complexité est représentée à la fois par le coût, le temps et la consommation de ressources. Pour cela, le processus ETL doit être conçu avec soin et il faut choisir entre l'utilisation d'un outil dédié (Talend[103] par exemple) ou une approche scriptée.

L'approche scriptée, ou "hand made" , consiste à développer le ou les outils nécessaires à la mise en place d'un traitement ETL, à l'aide d'un ou plusieurs langages de programmation (Python, R, etc.). Cette méthode consiste, généralement, en une combinaison de procédures sauvegardées au niveau des bases de données ainsi que de différents scripts nécessaires au transport des données et aux transformations/ traitements complexes [68].

Les avantages de l'approche scriptée sont :

- Elle permet une homogénéité technologique avec les solutions déjà en place au niveau du SI.
- Elle permet d'utiliser les langages que les équipes maîtrisent déjà, sans apprentissage et médiation d'un outil tiers.
- Elle peut, dans certains cas, simplifier le traitement des fichiers plats.
- Elle permet une intégration poussée avec des outils de gestion de versions.

Dans un autre côté, il se trouvent certains inconvénients liés à cet approche, en effet :

- Il nécessite de tout développer.
- Il impose de nombreuses modifications lors d'une modification au niveau de la source, du contenu ou de la destination.
- Il est facteur d'un grand nombre de bugs et de régressions à chaque modification.
- Il nécessite l'utilisation d'un planificateur de tâches externe.
- Il nécessite de tout documenter, au risque de se perdre dans le code produit.
- Il peut être éparpillé sur plusieurs serveurs.

[103]Talend est un logiciel spécialisé dans l'intégration de données. Pour en savoir plus: www.talend.com

Glossaire

Base de données

Fichier ou ensemble de fichiers permettant le stockage, la sauvegarde et l'accès à des informations structurées (ou par abus de langage, le contenu de la base).

Ensembles de données ou datasets

Le terme fait référence à la catégorisation ou à la classification des données publiques dans des catalogues de données, pour être facilement indexées et localisées. Pour cela, sont utilisés des champs qui définissent le groupe de données, tels que: description, fréquence de mise à jour, format, licence d'utilisation, entre autres.

Inventaire des données

Un inventaire des données disponibles permet d'y voir clair, de repérer les dossiers, les textes, les chiffres, et de s'assurer de leurs utilités pour un projet. Cet inventaire des données comporte le maximum d'informations sur l'observation, tout ce qui est nécessaire au traitement ainsi que les références aux résultats éventuellement déjà obtenus.

Marchés publiques

Les marchés publics sont des contrats entre un acheteur public (pouvoir adjudicateur) et un (ou des) opérateur(s) économique(s). Ces contrats répondent aux besoins en matière de travaux, de fournitures ou de services du gouvernement. Les marchés publics doivent respecter les principes de liberté d'accès à la commande publique, d'égalité de traitement des candidats et de transparence des procédures. Ces principes permettent d'assurer l'efficacité de la commande publique et la bonne utilisation des deniers publics.

Métadonnées

Les métadonnées sont des champs ou des informations qui décrivent les données. Ils fournissent à l'utilisateur suffisamment d'éléments pour traiter et comprendre les données. Ces champs peuvent varier en détail, depuis des descriptions très basiques telles que l'explication du sujet général de la base de données, jusqu'à la fourniture de détails sémantiques pour permettre un degré élevé de lisibilité par machine; cela augmente le degré d'ouverture et l'utilité des données.

Modèle ARIMA

Modèle de prédiction des séries temporelles ARIMA (moyenne mobile intégrée auto-régressive) ou méthodologie de prévision de Box-Jenkins.

Ce modèle offre une approche complémentaire de la prévision de séries chronologiques. c'est-à-dire que l'analyse ARIMA peut supprimer la composante de tendance et de la saisonnalité dans les données, visant à décrire les autocorrélations dans les données où les valeurs passées ont un effet sur les valeurs actuelles, afin de prédire avec précision les valeurs futures.

Mouvement d'ouverture des données

Ce concept peut être expliqué en développant chacun de ses principaux éléments [9] :

Disponibilité et accès : les informations doivent être disponibles dans leur ensemble et à un coût de reproduction raisonnable, de préférence en les téléchargeant sur Internet. De plus, les informations doivent être disponibles sous une forme pratique et modifiable.

Réutilisation et redistribution : les données doivent être fournies sous des termes qui leur permettent d'être réutilisées et redistribuées, voire de les intégrer à d'autres ensembles de données.

Participation universelle : chacun doit pouvoir utiliser, réutiliser et redistribuer les informations. Il ne devrait y avoir aucune discrimination en termes d'efforts, de personnes ou de groupes. Les restrictions qui empêcheraient l'utilisation commerciale des données ou les restrictions d'utilisation à certaines fins (par exemple uniquement pour l'éducation) devraient être limitées au maximum.

Pondération des données

Pondérer un indicateur ou une information consiste à donner aux valeurs qui les composent un poids différent (un coefficient de pondération) en fonction des divers critères qui rendent compte de l'importance relative de chacun des éléments. La finalité est de corriger la représentativité de l'échantillon en fonction de certaines variables clés, afin d'être en mesure d'extrapoler les résultats à la population.

Secret statistique

Le secret statistique est défini par la loi n° 51-711 du 7 juin 1951 modifiée sur l'obligation, la coordination et le secret en matière de statistiques. Il interdit, pendant une durée de soixante-quinze ans, toute communication de données ayant trait à la vie personnelle et familiale, et plus généralement, aux faits et comportements d'ordre privé recueillis au moyen d'une enquête statistique. Des dérogations peuvent néanmoins être accordées, sur avis du Comité du secret statistique, exclusivement en réponse à des besoins dont la finalité relève de la statistique publique ou de la recherche scientifique ou historique [104].

Sources d'information statistique

Ces sources d'information peuvent être classifiées comme "primaires" ou "secondaires". Les sources primaires sont celles qui contiennent de l'information de première main, sans avoir subi d'altérations ou de modifications, par exemple des résultats statistiques. Les sources

[104] Extrait de insee.fr/fr/information/1300624

secondaires fournissent des informations modifiées, résumées ou représentées par des tiers.

Système harmonisé

Le Système harmonisé est un code de classement des marchandises international à six chiffres. Les deux premiers chiffres (HS-2) identifient le chapitre dans lequel les marchandises sont classées, par exemple 09 = café, thé, maté et épices. Les deux chiffres suivants (HS-4) identifient les groupes au sein de ce chapitre, par exemple 09.02 = Thé, même aromatisé. Les deux chiffres suivants (HS-6) sont encore plus spécifiques, par exemple 09.02.10 Thé vert (non fermenté).

Traitement automatique de données

Ensemble des opérations réalisées par des moyens automatiques, relatif à la collecte, l'enregistrement, l'élaboration, la modification, la conservation, la destruction, l'édition de données et, d'une façon générale, leur exploitation.

Codes

Text normalisation

```
import re
import unicodedata

# our own list of stop_words
from index.search_stopwords import JM_STOP_WORDS


def strip_accents(text):
    text = unicodedata.normalize("NFD", text).encode("ascii",
"ignore").decode("utf-8")
    return str(text)


def _normalize_whitespace(text):
    """
    This function normalizes whitespaces, removing duplicates.
    """
    corrected = str(text)
    corrected = re.sub(r"//t", r"\t", corrected)
    corrected = re.sub(r"( )\1+", r"\1", corrected)
    corrected = re.sub(r"(\n)\1+", r"\1", corrected)
    corrected = re.sub(r"(\r)\1+", r"\1", corrected)
    corrected = re.sub(r"(\t)\1+", r"\1", corrected)
    return corrected.strip(" ")


def normalize(text):
    """ Keep only lower-cased text and numbers"""
    text = _normalize_whitespace(text)
    removed_accents = " ".join(
        strip_accents(word) for word in text.lower().split() if
word not in JM_STOP_WORDS
    )
    return re.sub("[^a-z0-9]+", " ", removed_accents).strip()
```

Clustering of buyers and competitors

```
from sklearn.cluster import AgglomerativeClustering

# text-clustering
def group_texts(texts, threshold= 0.5):
  """ Replace each text with the representative of its cluster"""
  normalized_texts = np.array([normalize(text) for text in texts])
  distances = 1 - np.array([
      [textdistance.jaro_winkler(one, another) for one in normalized_texts]
      for another in normalized_texts
  ])
  clustering = AgglomerativeClustering( #these parameters needs to be tuned
carefully
    distance_threshold = threshold,
    affinity ="precomputed",
    linkage = "complete",
    n_clusters = None,
  ).fit(distances)
  centers = dict()
  for cluster_id in set(clustering.labels_):
    index = clustering.labels_ == cluster_id
    centrality = distances[:, index][index].sum(axis=1)
    centers[cluster_id] = normalized_texts[index][centrality.argmin()]
  return [centers[i] for i in clustering.labels_]
```

Remove the outliers

```
def remove_outliers(self, dataframe):
        """
        Remove the outliers identified via the "1.5 x interquartile
range" rule.
        """
        column_name = "total"
        q1 = dataframe[column_name].quantile(0.25)  # first quartile
        q3 = dataframe[column_name].quantile(0.75)  # third quartile
        interquartile_range = q3 - q1
        lower_bound = q1 - 1.5 * interquartile_range
        higher_bound = q3 + 1.5 * interquartile_range
        return dataframe.loc[
            (dataframe[column_name] >= lower_bound) &
(dataframe[column_name] <= higher_bound)
        ]
```

Launching periodic tasks with Threading

```
import pandas as pd
import threading

def sncf_lost_properties():
    lost_properties =
pd.read_json('https://ressources.data.sncf.com/explore/dataset/objets-trouves-restitution/download/?format=json&timezone=Europe/Berlin&lang=fr')
    lost_properties.to_json('lost_properties.json') #save file
    threading.Timer(86400, sncf_lost_properties).start() #86400 secondes per day

sncf_lost_properties()
```

Spider DILA

```
from urllib.parse import urlparse
from urllib.parse import urljoin

import scrapy
from scrapy import Selector
from scrapy.http import Request
from scrapy.linkextractors import LinkExtractor
from scrapy.spiders import CrawlSpider, Rule, Spider

from scrapy.loader import ItemLoader
import sys
from items import ExtractionItem

class Dila(scrapy.Spider):
    name = "dila"

    start_urls = ["https://echanges.dila.gouv.fr/OPENDATA/"] # API

    def parse(self, response):

        rows = response.xpath("//body//pre//a[not(contains(@href,'?'))]")
        for i, row in enumerate(rows):
            if i == 0:
                continue  # skip table header

            loader = ItemLoader(item=ExtractionItem(), selector=row)
            relative_path = row.xpath(".//@href").get("")
            absolute_path = urljoin(self.start_urls[0], relative_path)
            if ".pdf" in absolute_path :
                loader.add_value('file_urls', absolute_path)
                yield loader.load_item()
            else :
                yield Request(
                url=absolute_path,
                callback=self.parse_detail
                )


    def parse_detail(self, response):
```

```
        """
        Parses the detail page
        """
        rows = response.xpath("//body//pre//a[not(contains(@href,'?'))]")
        for i, row in enumerate(rows):
            if i == 0:
                continue  # skip table header

            loader = ItemLoader(item=ExtractionItem(), selector=row)
            relative_path = row.xpath(".//@href").get("")
            absolute_path = urljoin(self.start_urls[0], relative_path)
            loader.add_value('file_urls', absolute_path)
            yield loader.load_item()
```