Open data et création de valeur : cas d'usage d'une entreprise dans les cryptomonnaies

Izem El Mourabit

Mémoire de stage

Open data et création de valeur

Cas d'usage d'une entreprise dans les cryptomonnaies

Présenté par : EL MOURABIT Izem
Entreprise d'accueil : CoinShares

Date de stage : du 04/04/22 au 30/09/22

Tuteur entreprise : GOIS Julien
Tuteur universitaire : HOAREAU Emilie

Master 2 - Formation initiale
Master Systèmes d'information
Parcours Intelligence des données
2020 - 2022

CoinShares

GRENOBLE IAE
ECOLE PUBLIQUE DE MANAGEMENT
GRENOBLE INP - UGA

Avertissement :

Grenoble IAE, au sein de l'Université Grenoble Alpes, n'entend donner aucune approbation ni improbation aux opinions émises dans les mémoires des candidats aux masters en alternance : ces opinions doivent être considérées comme propres à leur auteur.

Tenant compte de la confidentialité des informations ayant trait à telle ou telle entreprise, une éventuelle diffusion relève de la seule responsabilité de l'auteur et ne peut être faite sans son accord.

RÉSUMÉ

Les données ouvertes (open data) sont considérées par beaucoup comme le nouvel or noir du 21e siècle. Elles sont une ressource, qui, une fois exploitée peut créer beaucoup de valeur pour une organisation. C'est pourquoi il est primordial pour les entreprises d'aujourd'hui de connaître les étapes à suivre pour exploiter au mieux cette ressource. Ce travail prend pour cadre théorique le modèle de la chaîne de création de valeur de l'open data (European Commission, 2015). Il détaillera et expliquera les différentes étapes du processus. Après avoir détaillé le fonctionnement du modèle, un cas d'usage sera présenté : nous verrons comment créer de la valeur à partir d'open data une entreprise dans le secteur des cryptomonnaies comme CoinShares. À travers ce cas d'usage, nous verrons que le modèle théorique a ses limites : une analyse et des critiques de ce modèle seront présentées. Le travail conclue avec des recommandations formulées pour CoinShares concernant l'attitude à avoir vis-à-vis de l'open data. (160 mots)

SUMMARY

Open data is considered by many as the new oil of the 21st century. It is a resource that, once exploited, can create a lot of value for an organization. This is why it is essential for today's companies to know the steps to follow in order to exploit this resource. This work takes the open data value chain model (European Commission, 2015) as a theoretical framework. It will detail and explain the different steps of the process. After detailing how the model works, a use case will be presented: we will see how to create value from open data in a company in the cryptocurrency ecosystem like CoinShares. Through this use case, we will see that the theoretical model has its limits: an analysis and criticisms of this model will be presented. The work concludes with recommendations for CoinShares regarding the attitude to have towards open data. (148 words)

MOTS CLÉS : données ouvertes, données, cryptomonnaie, crypto, création de valeur, API, application web, CoinShares

KEY-WORDS: open data, data, cryptocurrency, crypto, value creation, API, web application, CoinShares

REMERCIEMENTS

Premièrement, j'aimerais remercier Arnaud Dartois et Jean-Charles Dudek qui m'ont permis d'intégrer l'entreprise (Napoleon Group initialement, rachetée par CoinShares).

Ensuite, j'aimerais remercier Ghazi Mejaat qui avait initialement accepté d'être mon tuteur avant d'être licencié trois semaines avant mon arrivée.

Puis, j'aimerais remercier Jean-Marie Mognetti, CEO de CoinShares, d'avoir racheté Napoleon Group sans quoi, il n'est pas certain que Napoleon Group existerait encore ; et donc il n'est pas certain que j'aurais pu effectuer mon stage de fin d'études.

Enfin, j'aimerais me remercier moi-même pour avoir gardé la motivation et l'envie nécessaire de bien faire les choses et de rendre ce stage tout de même instructif malgré le fait que je n'avais plus de tuteur officiel ni de missions précises. J'ai su rendre mon stage productif en me rendant utile, en étant force de proposition en suggérant des projets à réaliser et j'ai ainsi développé des compétences (notamment en code) qui me seront fort utile plus tard en tant que data analyst.

TABLE DES MATIÈRES

“Open data is the new oil of the digital economy.”

INTRODUCTION

CONTEXTE DE L'ENTREPRISE

CoinShares est un pionnier européen de l'investissement de crypto-actifs avec plus de 2 milliards d'actifs crypto sous gestion. Un crypto-actif est « un actif numérique virtuel qui repose sur la technologie de la blockchain[1] à travers un registre décentralisé et un protocole informatique crypté. » (AMF, 2022). L'entreprise propose notamment aux particuliers et aux institutionnels d'investir sur des ETP (Exchange-Traded Product ou produit négocié en bourse en Français) basés sur des cryptomonnaies (cryptos) comme Bitcoin, Ethereum ou encore Litecoin. Depuis l'acquisition récente de Napoleon Group et de son produit phare 'Napbots', CoinShares propose dorénavant aux particuliers des stratégies d'investissement automatisées basées sur des cryptomonnaies dont le projet est solide. Les particuliers connectent leur plateforme de trading favorite à la plateforme napbots.com, ils choisissent une stratégie d'investissement et la cryptomonnaie qui leur convient le mieux (neuf cryptomonnaies sont aujourd'hui disponibles) et tout le reste est automatique : Napbots se charge d'envoyer les ordres d'achat / de vente à la plateforme de trading.

Le lecteur doit avoir conscience que l'écosystème de la cryptomonnaie est très communautaire. Chaque cryptomonnaie est liée à un projet technologique basé sur la blockchain et chaque projet crée plus ou moins d'engouement autour de lui. Par exemple, Ethereum (dont la cryptomonnaie native 'Ether' ou ETH est la deuxième cryptomonnaie derrière le Bitcoin), est une blockchain sur laquelle il est possible de développer des applications décentralisées (ou dApps). Un exemple de dApp est Gnosis : elle permet aux utilisateurs de faire des prédictions sur n'importe quel sujet allant de la météo aux prochaines élections. Un autre exemple est Etheria : c'est un jeu vidéo ressemblant à Minecraft dans lequel les joueurs peuvent, entre autres, acheter des parcelles de terre. Ces deux exemples de dApps fonctionnent sur la blockchain Ethereum. Ethereum est donc un projet majeur de l'écosystème crypto car il permet de construire tout type d'applications décentralisées et génère de ce fait, beaucoup d'engouement dans la communauté.

La clé pour CoinShares, qui propose à ses clients des services liés aux cryptomonnaies, est de pouvoir mesurer/détecter cet engouement. L'engouement autour d'une cryptomonnaie peut être mesuré notamment par l'activité sur les réseaux sociaux comme Twitter ou Reddit. Quand l'équipe du projet Ethereum fait une annonce officielle, on voit le nombre de publications Twitter avec le hashtag #ETH

[1] La blockchain (ou chaine de blocs en Français) est un registre numérique décentralisé qui recense l'historique de toutes les transactions effectuées en Bitcoin (ou dans une autre cryptomonnaie) et est maintenue par un réseau d'ordinateurs connectés les uns auxautres de pair à pair.

exploser. Mais l'engouement autour d'Ethereum et sa cryptomonnaie associée Ether est déjà connue et CoinShares propose déjà des stratégies d'investissement autour de cette crypto. Il est donc un enjeu important pour CoinShares de pouvoir détecter des cryptomonnaies qui connaîtront un engouement futur afin de les suivre et de potentiellement proposer des produits d'investissements basés sur ces dernières.

COINSHARES ET L'OPEN DATA

L'open data est d'une grande utilité pour CoinShares. L'open data (ou données ouvertes) sont des données publiées par une organisation publique ou privée et qui peuvent être utilisées par n'importe qui (plus d'explications dans la Partie I). Twitter, par exemple, est un fournisseur d'open data puisque l'entreprise propose l'accès à des données (via une API[2]) comme le nombre de Tweets avec la mention '#ETH' ou bien le nombre d'abonnés du compte Twitter d'Ethereum. Mais Twitter n'est pas un exemple isolé, aujourd'hui la plupart des entreprises technologiques (comme celles des réseaux sociaux) proposent ce genre de service. Elles mettent à disposition en accès libre certaines données qu'elles génèrent. CoinShares peut, dès lors, valoriser ces données ouvertes en les utilisant à son avantage.

ANNONCE DU PLAN

Mais alors, comment une entreprise privée peut-elle exploiter l'open data afin d'en créer de la valeur ? Ce mémoire présente un cas d'usage concret d'une application de suivi de tendances de cryptomonnaies. L'objectif de ce travail est de montrer comment une entreprise aujourd'hui, peut créer de la valeur en exploitant l'open data. Le cas d'usage utilisé est une application basée entièrement sur de l'open data pour suivre l'engouement autour de certaines cryptomonnaies.

Le mémoire se décompose en trois parties : la première partie aborde le processus de création de valeur de l'open data en se basant sur le modèle théorique de la chaîne de valeur de l'open data de la Commission Européenne (European Commission, 2015). La deuxième partie est une présentation de l'application de suivi d'engouement des cryptomonnaies qui servira comme cas d'usage. Enfin, dans la troisième partie, le modèle théorique évoqué en Partie I sera analysé et des améliorations basées sur le cas d'usage seront suggérées. Toujours dans cette dernière partie, des recommandations à CoinShares concernant l'attitude à avoir vis-à-vis de l'open data seront formulées.

PARTIE I. OPEN DATA ET CRÉATION DE VALEUR : THÉORIE

[2] API (ou Application Programming Interface) est un intermédiaire informatique qui permet à deux applications d'interagir ensemble.

QU'EST-CE QUE LA DONNÉE ?

En 2017, l'hebdomadaire The Economist affirmait que la donnée était le nouvel or noir (The Economist, 2017). A l'instar du pétrole au début de siècle dernier, la donnée serait, dans notre monde numérique d'aujourd'hui, une matière première exploitable et très prisée par les acteurs économiques. Il serait donc de l'intérêt des entreprises d'aujourd'hui de savoir exploiter cette nouvelle denrée afin de créer de la valeur et de rester compétitives. Mais la donnée brute seule n'a pas de valeur intrinsèque. A l'instar d'une matière première brute (comme le pétrole), la donnée a besoin d'être traitée et transformée afin de pouvoir exploiter tout son potentiel et d'en dégager des informations voire de la connaissance (à l'instar de l'énergie pour le pétrole). En effet, il est important de distinguer donnée, information et connaissance (Ackoff, 1989) :

1. Donnée brute : une donnée est un symbole représentant des propriétés d'objets, es évènements et leur environnement. Elle est le produit d'observation. Par exemple, une température mesurée en degré Celsius, à Grenoble, le 1 août (35°C) est une donnée brute. Elle est une représentation de la température extérieure à un moment donné.
2. Information : une information est une donnée brute à laquelle on a donné du sens (en la traitant par exemple, ou encore en l'agrégeant avec d'autres données). Une information répond à des questions du type « qui ? », « quoi ? », « où ?», « quand ?» et « combien ?». Par exemple, si on combine cette température de 35°C avec les températures (du jour et de la nuit) mesurées les deux derniers jours, on pourra calculer des températures moyennes. Or, à Grenoble, si la moyenne des températures dépasse 34°C le jour et 19°C la nuit, sur trois jours, on parle de canicule. Donc, en contextualisant une donnée brute avec d'autres, on peut obtenir du sens : on a dégagé une information, Grenoble connaît une période de canicule.
3. La connaissance : la connaissance c'est « savoir comment » ou « savoir pourquoi » un système fonctionne ainsi. C'est l'explication du système. Dans notre exemple de températures, une des causes de la canicule peut être le sirocco provenant du Sahara qui remonte jusqu'à l'hexagone ou bien un anticyclone, synonyme d'un temps chaud et sec, proche des îles britanniques qui expose la métropole. Bref, on essaie de donner du sens à l'information que l'on possède.

La donnée est donc la base de toute information et de toute connaissance. Il est donc primordial pour les entreprises de savoir l'exploiter. Cependant, contrairement aux géants des technologies de l'information et de la communication (TIC) comme Google ou Facebook, beaucoup d'entreprises ne génèrent pas autant de données. Il peut alors sembler difficile pour beaucoup d'entreprises de taille petite et moyenne de tirer profit d'une ressource à priori inexistante. Mais ceci est sans compter l'essor

de l'open data que l'on connait depuis la fin des années 2000 et qui a permis l'ouverture et la mise à disposition d'une immense quantité de données à tous.

QU'EST-CE QUE L'OPEN DATA ?

Pour que des données puissent être catégorisées comme ouvertes, il faut qu'elles puissent être librement utilisées, modifiées et partagées par n'importe qui et dans n'importe quel but (The Open Definition, 2022). Elles doivent être publiées dans un format facilement lisible par une machine (format CSV ou JSON par exemple). Les données ouvertes doivent être publiées sous une Open License (ou licence ouverte) qui donne la permission à quiconque d'accéder, de (ré-)utiliser et redistribuer les données avec peu ou pas de restrictions (voir la liste des licences qui respectent ces critères ici : https://opendefinition.org/licenses/). L'open data est donc une ressource qu'une entreprise peut librement utiliser ou modifier pour en créer de la valeur. Et le champ des possibles est grand car on connaît aujourd'hui un essor de l'open data permettant l'accès à de multiples sources et de types de données ouvertes.

L'ESSOR DE L'OPEN DATA

Depuis la fin des années 2000, le mouvement a sensiblement gagné en importance (Kitchin, 2014). En mai 2009, les Etats-Unis lançaient data.gov, un site web conçu pour donner accès à des ensembles de données non sensibles et historiques détenues par l'État américain et les agences fédérales. Sur ce site on peut retrouver des jeux de données portés sur l'agriculture, le climat ou encore l'énergie aux USA. En septembre de la même année, c'était au tour du Royaume-Uni de publier son site data.gov.uk qui comptait plus de 19000 jeux de données en 2015 et qui en compte plus de 47000 à l'écriture de ces lignes (re3data.org, 2014, p. 3). La France a elle aussi lancé sa plateforme ouverte des données publiques en décembre 2013 : data.gouv.fr. Le site « vise à centraliser et structurer les données ouvertes en France. Il favorise la transparence et l'efficacité de l'action publique tout en facilitant la création de nouveaux services. ». data.gouv.fr compte aujourd'hui plus de 40000 jeux de données et a été réutilisé officiellement plus de 3000 fois. Aujourd'hui La France est un des pays les plus avancé en matière d'open data puisqu'elle se hisse à la 4^{e} place (sur 30) du classement du site opendatabarometer.org de 2017 (site produit par la World Wide Web Foundation) et 2e du OURdata Index de l'OECD de 2019 (Perez et al., 2020). Parmi les réutilisations de ces jeux de données ouverts par la France, on peut trouver des visualisations sur les logements vacants en France ou encore des applications sur la qualité de vie des différents quartiers de France. Un effort a donc été mis sur l'accès et le partage des données non sensibles par les gouvernements car ces derniers ont compris les enjeux à la fois de transparence mais aussi de développement de nouveaux services que peut apporter cette

ouverture de données. Les acteurs privés l'ont d'ailleurs bien compris car le marché de l'open data a connu une forte croissance en quelques années.

LE MARCHÉ DE L'OPEN DATA

Alors qu'il ne valait que 27 milliards en 2005 (European Commission, 2015), le marché européen de l'open data[3] pesait 184 milliards d'euros en 2019 et il est estimé qu'il atteindra 199 à 334 milliards (pour les projections les plus optimistes) d'ici à 2025 (eurostat, 2022). À titre de comparaison le secteur de l'agriculture représentait seulement deux fois plus avec 414 milliards en 2020 (eurostat, 2022). Ce marché européen de l'open data représentait 1.09 million d'employés cette même année soit 12% du marché global de la donnée. Il est intéressant de noter que tous les secteurs économiques ne sont pas égaux vis-à-vis de l'open data. En prenant en compte la disponibilité, le nombre de téléchargements et l'applicabilité de l'open data, Buchholtz et al. (2014) ont montré que certains secteurs (comme les TIC, l'administration publique ou la santé) ont la capacité de jouir des bénéfices de l'open data plus que d'autres (le secteur du BTP ou de la justice par exemple) (Figure 1).

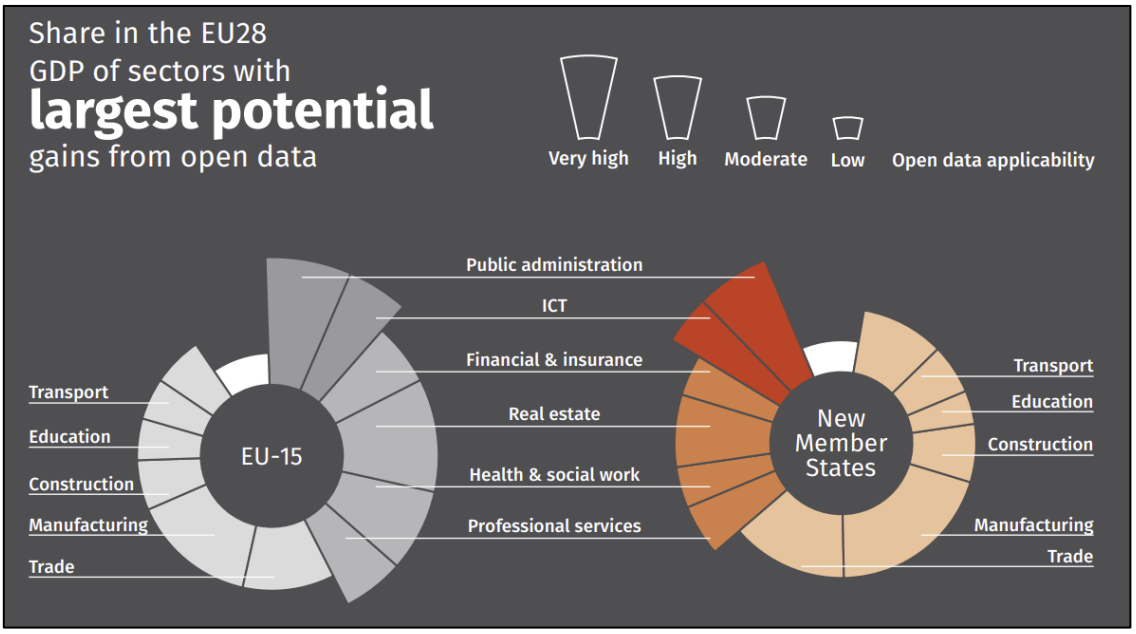

Figure 1. Secteurs économiques avec les plus grands gains potentiels venant de l'open data (Buchholtz et al., 2014). Les transports, l'administration publique et l'industrie sont les secteurs ayant le plus de potentiel de bénéficier des bienfaits de l'open data. À noter qu'il y a des disparités entre les membres historiques de l'UE et les nouveaux membres, notamment sur la part qu'occupe un secteur dans le produit intérieur brut total.

En effet, il sera plus facile pour une entreprise dans les TIC d'exploiter des données en lien avec son activité (qui sont très nombreuses) qu'un cabinet d'avocat pour qui il existe beaucoup moins de données disponibles en lien avec son activité (sans compter le manque de savoir-faire pour traiter la

[3] Marché de l'open data : marché des produits, services et contenus créés ou améliorés grâce à l'open data.

donnée ou même le manque d'éducation ou de culture liée à ses sujets). Tous les secteurs n'ont de fait, pas les mêmes chances au départ pour exploiter l'open data et créer de la valeur.

EXPLOITATION DE L'OPEN DATA ET CRÉATION DE VALEUR

Par valeur, il sera question dans ce travail de valeur d'utilité, c'est-à-dire « est-ce que le service/produit créé grâce à l'open data à une utilité pour son utilisateur ? ». Le pétrole, par exemple, à une forte valeur d'utilité car, une fois transformé en énergie par un transformateur, il permet de transporter des marchandises, aux gens de se déplacer ou encore de faire tourner nos machines. Dans la même idée, l'open data a de la valeur si et seulement si le résultat de son « extraction » et de sa « transformation » aura permis de créer (ou améliorer) un service/produit qui sera utile pour les utilisateurs (que ce soit pour les utilisateurs internes à l'entreprise ou bien les clients de l'entreprise).

Il y a différentes manières de créer de la valeur à partir de l'open data. Une entreprise peut, par exemple, s'aider de l'open data pour améliorer ses services et ses produits, améliorer ses prises de décisions et/ou son efficacité, mieux connaître les habitudes de ses clients ou encore créer des opportunités business. Google, par exemple, utilise des données ouvertes de transport pour enrichir son application Google Maps et permettre aux utilisateurs de planifier leurs voyages en utilisant les transports publics. Google a donc créé de la valeur en créant un service utile aux gens souhaitant se déplacer, et ce, sur la base de l'open data. Un autre exemple est Yelp qui est une application mettant en relation des personnes avec des restaurants et des entreprises. Elle exploite les données des inspections sanitaires pour informer les utilisateurs sur la qualité de l'hygiène des restaurants. Yelp a donc créé de la valeur en améliorant son service initial (avis utilisateurs de restaurants) en y ajoutant des informations d'hygiène.

MODÈLE THÉORIQUE : CHAÎNE DE VALEUR DE L'OPEN DATA

Le modèle théorique sur lequel se base ce travail est la chaîne de création de valeur de l'open data de la Commission Européenne (European Commission, 2015) (Figure 2). Ce modèle a pour but de décrire le processus de création de valeur de l'open data.

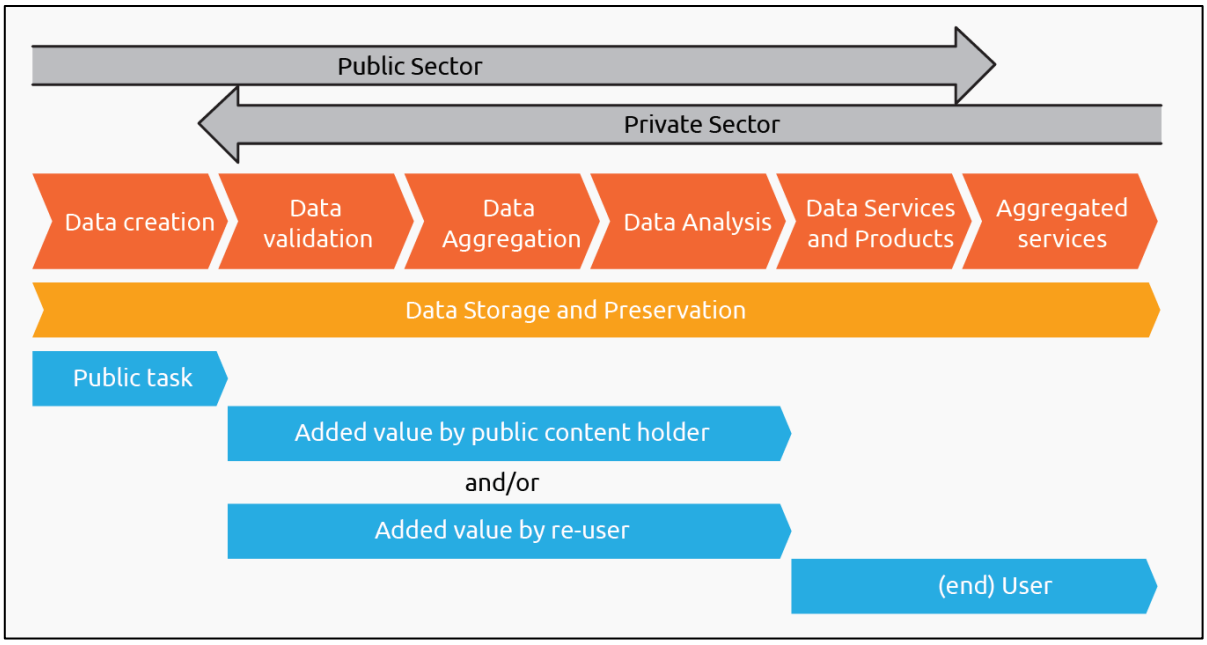

Figure 2. Chaîne de valeur de l'open data (European Commission, 2015). Le modèle comporte 6 étapes, de la création de la donnée à l'agrégation de services basés sur cette donnée. D'après ce modèle, le secteur public est au début de la chaîne (fournit la donnée) tandis que le secteur privé arrive en fin de chaîne (traite la donnée et créé des services). Durant ce processus, la donnée à besoin d'être stockée et préservée et des enjeux de sécurité peuvent apparaître. Enfin, de la valeur est créée soit par le secteur public publiant la donnée (selon son niveau de pré-traitement) et/ou par le secteur privé qui la réutilise et la traite pour en créer un service/produit ; l'utilisateur final jouit du service/produit créé en étape 5 et 6.

Ce processus comporte 6 étapes, de la création de la donnée jusqu'à l'agrégation de services basés sur cette donnée, en passant par l'analyse de cette donnée. Chacune de ces étapes sera détaillée dans les parties suivantes - voici un aperçu du processus : une fois la donnée créée, celle-ci a besoin d'être collectée et validée. Elle est ensuite analysée par l'entreprise utilisatrice de cette donnée : l'entreprise pourra combiner d'autres jeux de données pour ajouter du sens à cette donnée, créer des visualisations ou encore faire des tests statistiques pour en ressortir des résultats. Tout ceci permettra la création de services/produits et donc de valeur. Optionnellement, le service/produit créé pourra lui aussi être agrégé ou associé avec un autre service déjà existant au sein de l'entreprise pour créer encore une fois de la valeur. Durant tout ce processus, la donnée a besoin d'être stockée et préservée : le stockage de la donnée peut être effectué en interne ou dans le cloud, avec des enjeux de sécurité différents selon les cas (voir Deshpande et al. (2019) pour en apprendre plus sur les enjeux de sécurité de la donnée dans le cloud).

Le modèle indique que le secteur public (comme les gouvernements) est au début de la chaîne (c'est lui qui crée et rend accessible la donnée) tandis que le secteur privé arrive en fin de chaîne, et c'est à lui qu'incombe la tâche de traiter la donnée et de créer des services/produits. Le lecteur gardera en tête que c'est une vue simplifiée de la réalité car le secteur privé peut lui aussi créer et mettre à

disposition des données ouvertes (par exemple Twitter qui met à disposition des données gratuitement). Ce point sera discuté lors de l'analyse du modèle PARTIE III.

La valeur est créée lors des étapes 2, 3 et 4 de validation, agrégation, et analyse de la donnée ouverte. Elle peut donc être créée à la fois par l'entité qui récupère (et donc valide, agrège et analyse la donnée) ou par l'entité qui publie la donnée selon le niveau de pré-traitement qu'elle a effectué sur la donnée qu'elle publie. L'utilisateur (final) est celui qui profite de la valeur créée. Il peut s'agir de l'utilisateur d'une application créée et basée entièrement sur l'open data ou bien une entreprise qui a accumulé de l'information concernant son secteur et qui peut alors prendre des décisions plus éclairées.

ARCHÉTYPES DE LA CHAÎNE DE VALEUR DE L'OPEN DATA

Le modèle suggère que l'on peut distinguer quatre types d'acteurs intervenant tout au long du processus : les fournisseurs, les agrégateurs, les développeurs/les « enrichisseurs » et les facilitateurs (non présents sur le schéma mais mentionnés par les auteurs du modèle) (Figure 3). Chaque acteur intervient à un moment précis de la chaîne et chaque acteur exploite le travail de l'acteur qui le précède. En fin de chaîne, par la transformation de la donnée en information, de cette information en connaissance puis de cette connaissance en services, de la valeur est créée et proposée à l'utilisateur final.

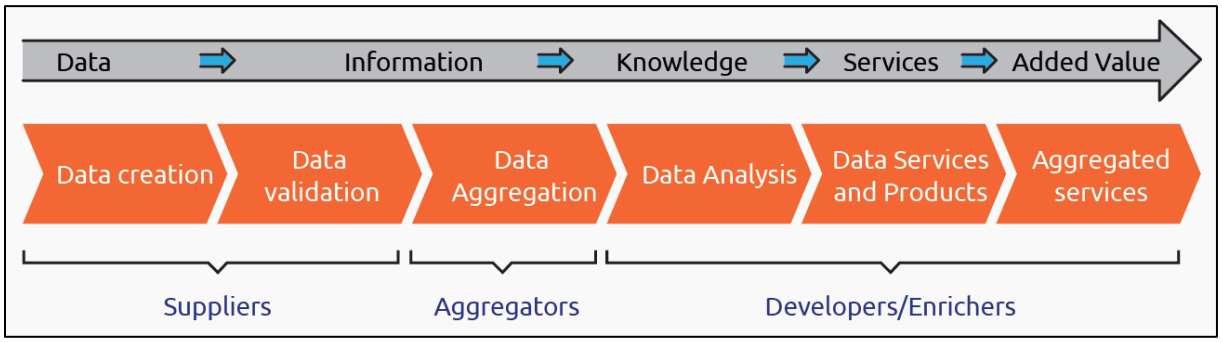

Figure 3. Archétypes de la chaîne de valeur de l'open data. Le modèle distingue trois acteurs principaux qui interviennent tout au long de la chaîne : les fournisseurs qui créent et valident la donnée ; les agrégateurs, qui agrègent les données provenant d'un ou plusieurs fournisseurs ; les développeurs/les « enrichisseurs » qui analysent les données et créent des services/produits basés sur cette dernière ; et les facilitateurs (même si ces derniers ne sont pas mentionnés dans le schéma, ils sont mentionnés par les auteurs du modèle) qui facilitent l'utilisation et l'exploitation de l'open data. Par la transformation de la donnée en information, de l'information en connaissance et de la connaissance en services, de la valeur est créée.

LES FOURNISSEURS

Les fournisseurs sont les organisations qui rendent leurs données ouvertes et accessibles à tous. Il s'agit non seulement d'organismes du secteur public mais aussi de certaines entreprises du secteur privé, comme la SNCF par exemple, qui met à disposition une API qui permet d'obtenir les itinéraires

et les horaires des trains SNCF. Il est vrai que les entreprises mettant à disposition librement leurs données ne génèrent pas de revenus directs, mais ces données peuvent augmenter le niveau d'engagement et de fidélité des clients, pouvant ainsi générer des revenus indirects. Il est également possible aux entreprises privées de mettre à disposition seulement certaines données gratuitement et proposer des données supplémentaires contre paiement ou abonnement. C'est d'ailleurs le cas de beaucoup de sites proposant des APIs. Comme nous le verrons dans la Partie II, la plupart des APIs privées utilisées dans l'application web présentée dans le cas d'usage, proposent une version gratuite et une version avec abonnement.

LES AGRÉGATEURS

Les agrégateurs sont les organismes qui collectent et agrègent l'open data mise à disposition par les fournisseurs. Les agrégateurs ont la responsabilité d'harmoniser les données malgré les différentes sources dont elles proviennent afin de les rendre facilement exploitables plus tard. Il y a souvent une composante géographique dans le travail des agrégateurs. Par exemple, data.gouv.fr, la plateforme ouverte des données publiques françaises, agrège une multitudes de sources de données en France : que ce soit des données de l'Institut National de la Statistique et des Etudes Economiques (Insee) sur les entreprises françaises[4], ou bien du Ministère de l'Économie, des Finances et de la Souveraineté industrielle et numérique sur le plan cadastral en France[5] ou encore venant du Ministère de l'Intérieur concernant les résultats définitifs du 1er tour de l'élection présidentielle[6]. Dans le secteur privé, on peut citer Transport API qui se veut être la première plateforme de données ouvertes concernant les données de transport du Royaume-Uni. Transport API gagne de l'argent en imposant un abonnement après un accès gratuit pendant 30 jours. Le rôle de agrégateurs est important car il facilite le travail des acteurs voulant exploiter l'open data.

LES DÉVELOPPEURS ET LES « ENRICHISSEURS »

Les développeurs sont des organisations ou des développeurs particuliers qui conçoivent, développent et vendent des applications (web ou mobiles) afin de fournir des services ou produits utiles (et donc créent de la valeur) aux usagers ou clients de ces-dites applications. Ces applications entrent parfois même en concurrence avec les applications officielles (on peut citer Citymapper qui concurrence directement l'application officielle RATP à Paris).

---

[4] Base Sirene des entreprises et de leurs établissements (SIREN, SIRET) https://www.data.gouv.fr/fr/datasets/base-sirene-des-entreprises-et-de-leurs-etablissements-siren-siret/

[5] Plan Cadastral Informatisé : https://www.data.gouv.fr/fr/datasets/plan-cadastral-informatise/

[6] Election présidentielle des 10 et 24 avril 2022 - Résultats définitifs du 1er tour : https://www.data.gouv.fr/fr/datasets/election-presidentielle-des-10-et-24-avril-2022-resultats-definitifs-du-1er-tour/

Dans la même lignée que les développeurs, les « enrichisseurs » sont des organisations qui utilisent l'open data pour s'informer ou trouver de nouvelles idées grâce auxquelles elles créent de nouveaux services ou produits qu'elles proposent à leurs clients. Ce sont souvent des produits qui auraient été impossibles à développer sans l'utilisation de l'open data. C'est le cas de Zillow, par exemple, une marketplace des biens immobiliers aux Etats-Unis permettant d'acheter, louer, vendre ou financer un bien de manière transparente. Les « enrichisseurs » sont sans doute le type d'organisations qui créent le plus de valeur ajoutée à l'open data et cela transparaît dans leur valeur marchande (Zillow est côté aux NASDAQ et est le site de biens immobiliers le plus visités des Etats-Unis).

LES FACILITATEURS

Les facilitateurs sont des organisations qui fournissent des plateformes et des services que les autres entreprises ou individus peuvent utiliser afin de faciliter l'exploitation de leurs données. Ils font partie à part entière de l'écosystème de l'open data. Par exemple, l'entreprise française OpenDataSoft a créé une plateforme open data qui permet aux entreprises, administrations ou autres collectivités territoriales de publier, visualiser ou de partager leurs données (la plateforme permet également de rendre leurs données disponibles via des APIs). Le rôle de ces organisations est donc de faciliter l'accès à l'open data pour tout type d'organisations (OpenDataSoft a notamment développé le portail open data de la SNCF).

LES ÉTAPES DE LA CHAÎNE DE VALEUR DE L'OPEN DATA

I. LA CRÉATION DE LA DONNÉE

Avant qu'elle puisse être exploitée, la donnée doit être créée (puis publiée). L'étape de création de la donnée est cruciale car elle est à la base de la chaîne de création de valeur. Comme nous l'avons vu précédemment, ce sont les fournisseurs qui sont à la manœuvre lors de cette étape.

D'après le modèle de la chaîne de valeur de l'open data (European Commission, 2015), les institutions publiques sont une source majeure d'open data, mettant à disposition des milliers de jeux de données. Reprenons l'exemple de Santé Publique France qui est l'agence nationale de santé publique en France ; elle publie et met à jour quotidiennement des données hospitalières relatives à l'épidémie de COVID-19 sur la plateforme data.gouv.fr. Elle joue donc le rôle de fournisseur de données de santé (ce sont en réalité les hôpitaux qui génèrent les données liées aux patients et Santé publique France joue le rôle d'intermédiaire, fournisseur de la donnée au public). Un autre exemple d'institution publique fournissant des données ouvertes est Météo France. Elle utilise des satellites, des stations de mesures au sol ou encore des capteurs embarqués sur des avions de lignes pour générer ces données.

L'institution met ensuite toutes ces données en accès libre sur son portail donneespubliques.meteofrance.fr. Météo France est donc un fournisseur de données météorologiques, qui peuvent ensuite être exploitées par les entreprises privées.

II. LA VALIDATION DE LA DONNÉE : QUALITÉ DE LA DONNÉE

Il est du rôle des fournisseurs de veiller à ce que la donnée proposée soit une donnée de qualité. Les développeurs et les « enrichisseurs » recueillant et utilisant ces données doivent également vérifier que les données recueillies sont de qualité. En effet, créer des services/produits ou baser ses décisions sur des données de mauvaise qualité peut mener à des services/produits non fiables ou de mauvais choix. Comme de tels problèmes devront de toute façon être corrigés plus tard, il est fortement conseillé d'éviter ces problèmes en début de chaine et de valider la qualité de la donnée en amont de l'exploitation.

De par la complexité et l'hétérogénéité des données du monde réel, les dimensions et les méthodes utilisées afin d'évaluer la qualité des données sont nombreuses. Cependant, on peut citer cinq propriétés qui reviennent souvent dans la littérature (Herzog et al., 2007) : la pertinence, l'exactitude, la fraîcheur, la comparabilité et l'intégrité.

A. Pertinence

Une donnée pertinente est une donnée appropriée pour l'usage que l'on veut en faire (ou lui donner). La donnée doit également pouvoir être utilisée dans d'autres scénarios non pensés initialement (ou du moins pouvoir être facilement traitée afin d'être utilisée dans différents cas d'usages). Par exemple, une donnée météo pertinente fournie par Météo France est la température relevée à une localisation L et à un instant T. La température est une donnée pertinente car elle va pouvoir être utilisée et exploitée pour différents usages. Au contraire, une donnée moins pertinente sera par exemple la concentration de gaz rares dans l'air. Cette donnée est moins pertinente dans un contexte de données météo car elle ne va pas pouvoir être exploitée aussi facilement et utilisée dans autant de contextes que la température. En somme, la pertinence d'une donnée dépend du contexte dans lequel elle est délivrée et de l'usage qui va en être fait.

B. L'exactitude

Une donnée exacte doit être « vraie », conforme à la réalité. Elle doit représenter le plus fidèlement la propriété du monde qu'elle est censée décrire. La température par exemple, doit représenter fidèlement la température de l'endroit auquel elle est relevée. Le fournisseur doit veiller à ce que son matériel de mesure soit fonctionnel. On pourrait également suggérer de fournir la température avec le nombre maximal de décimales donné par l'outil de mesure pour gagner en exactitude. De futurs

usages pourraient nécessiter une température d'une très grande précision, et pas seulement d'une température avec un nombre arrondi. Avoir des données exactes est la condition sine qua none d'une information et d'une connaissance fiable.

C. La fraîcheur

Une donnée « fraîche » est une donnée actuelle, à jour. Le fournisseur doit s'assurer qu'il y a le moins de temps possible entre le moment où la donnée est collectée et le moment où elle est publiée. Beaucoup d'utilisations peuvent nécessiter des données mise à jour tous les jours, toutes les heures, toutes les minutes voire toutes les secondes. Un exploitant agricole va par exemple avoir besoin de données de météo exactes et mises à jour toutes les heures afin de surveiller sa récolte le plus précisément possible. Ainsi la fraîcheur des données est une dimension indispensable pour avoir des données de qualité.

D. La comparabilité

Une donnée comparable est une donnée qui peut être comparée à une autre, en termes de valeur, de qualité, d'intensité etc. Il est important d'avoir des données comparables entre elles afin de faciliter leur utilisation future. Par exemple, les températures publiées par Météo France doivent conserver la même unité (Celsius), la même fréquence (horaire) et le même format de localisation (degrés décimaux ex. 48.864716, 2.349014). Si Météo France changeait d'unités et commençait à publier des données en Kelvin, toutes les 30 minutes, et avec un format de localisation en degrés, minutes et secondes (ex. 48°51'52.9776''N, 2°20'56.4504''E), il serait plus difficile de comparer les nouvelles données avec les anciennes données publiées dans des unités différentes. Il serait toujours possible de le faire en utilisant des conversions mais la comparabilité aurait grandement diminué. Il est donc important pour les fournisseurs d'assigner aux champs importants des jeux de données des clés d'identification. Dans notre exemple, Metéo France pourrait assigner un identifiant unique à chaque station de mesures au sol (qui est un champ important) afin de retrouver chaque station facilement même si le format de localisation était amené à changer dans le futur.

E. L'intégrité

Par « intégrité », il est fait référence à des données ne contenant aucun enregistrement ou champ manquant. Par exemple, pour des données météo horaires, une heure manquante est un enregistrement manquant. Un taux d'humidité manquant (parmi d'autres champs comme la température ou la pression atmosphérique) est un champ manquant. Des enregistrements manquants ou des champs manquants témoignent d'un manque de qualité. Dans beaucoup de bases de données (BDD) comme des BDD financières, des données manquantes peuvent avoir des conséquences

désastreuses. Si des problèmes de ce type surviennent, le processus de création de la donnée doit être examiné et corrigé par l'entreprise émettrice.

III. L'AGRÉGATION DE LA DONNÉE

L'agrégation de la donnée est la troisième étape de la chaîne de création de valeur de l'open data. Lors de cette étape, les données de différentes sources sont compilées dans le but de préparer le (pré-)traitement et l'analyse de données. Un jeu de données seul ne suffira souvent pas pour créer un service/produit innovant, il faudra donc l'agréger avec d'autres jeux de données venant de multiples sources, potentiellement hétérogènes. Comme vu plus haut, il existe des agrégateurs de données dont le rôle est précisément celui-ci : collecter et agréger différentes données provenant de différentes sources afin de les mettre à disposition (ex. Santé Publique France ou OpenDataSoft). Mais il n'est pas rare de ne pas trouver les jeux de données qui seront utiles à la création d'un service/produit sur les plateformes de ces agrégateurs. Ce sont alors aux entreprises voulant exploiter la donnée qu'incombe cette tâche. Prenons l'exemple fictif d'une entreprise voulant créer une application qui donne les endroits les plus souvent ombragés et frais à Paris. Elle aura par exemple besoin des données d'emplacement de tous les arbres plantés à Paris. De plus, elle aura besoin de la température précise de chaque segment de rue de Paris (si tant est que ces données existent). Il y a peu de chances qu'un agrégateur ait agrégé ces données. L'entreprise devra/pourra alors agréger et croiser ces données pour trouver les coins de Paris les plus ombragés et gardés au frais par l'ombre des arbres. Ces données proviendront sans doute de différentes sources, mettant à disposition leur données en différents formats. L'entreprise devra alors compiler et agréger ces données pour les rendre exploitables.

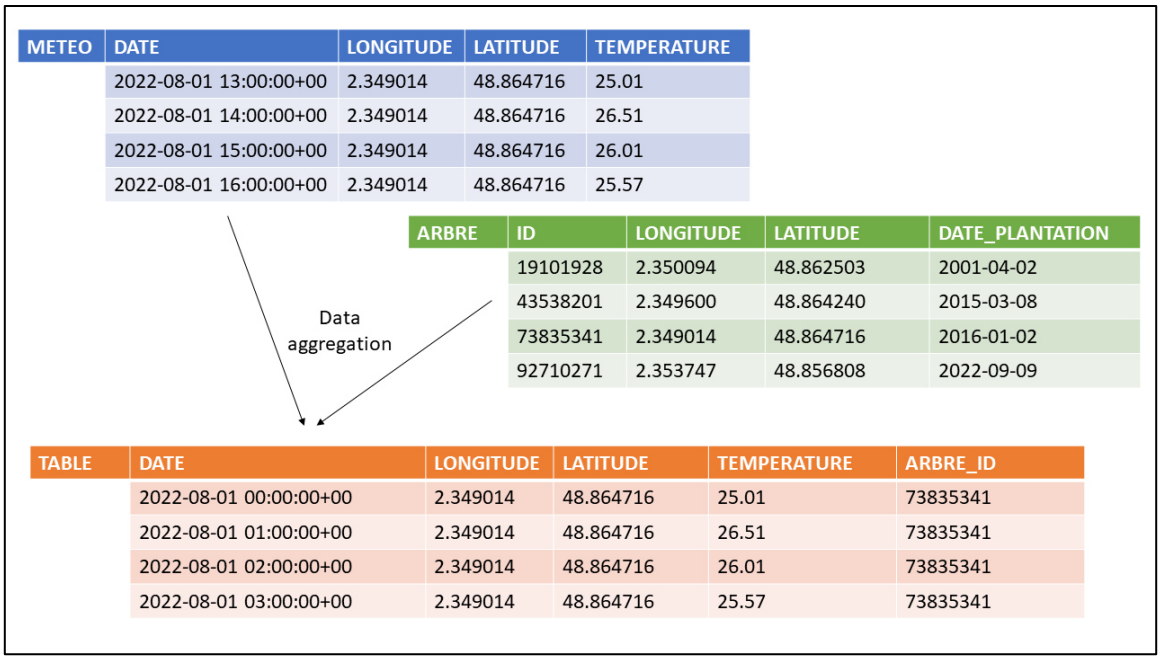

Figure 4. Exemple d'agrégation de données météo et de données d'emplacement d'arbres. Les deux jeux de données sont agrégés selon un identifiant unique (ici la longitude et la latitude des arbres et de la température) pour créer un troisième jeu de données contenant des informations plus riches. L'entreprise exploitant la donnée travaillera (analysera) ensuite sur ce troisième jeu de données.

C'est lors de cette agrégation de différents jeux de données que de la valeur est créée : les différentes données sont compilées entre elles et sont rendues exploitables, elles peuvent maintenant être analysées.

IV. L'ANALYSE DE DONNÉES

L'analyse de données est l'une des étapes les plus importantes de la chaîne de création de valeur de l'open data. Lors de cette étape, l'entreprise exploitant la donnée devra la nettoyer, la transformer, l'explorer, la visualiser et/ou la modéliser dans le but d'en sortir des informations qui aideront à la prise de décision ou bien qui aideront à créer un service/produit à forte valeur ajoutée.

A. (Pré-)Traitement des données

Afin de pouvoir être analysées, les données doivent auparavant être traitées : les données vont devoir subir un ensemble d'opérations afin d'être transformées pour faciliter les futures analyses. Par exemple, les données collectées peuvent ne pas toutes être dans le même format : certaines peuvent être en format CSV, d'autres en format JSON ou encore XML. Il sera du rôle du data analyst d'homogénéiser les formats de données. Les données tabulaires (comme les données en format CSV) étant les plus faciles à analyser, il est recommandé de transformer les données dans ce format (Figure 5).

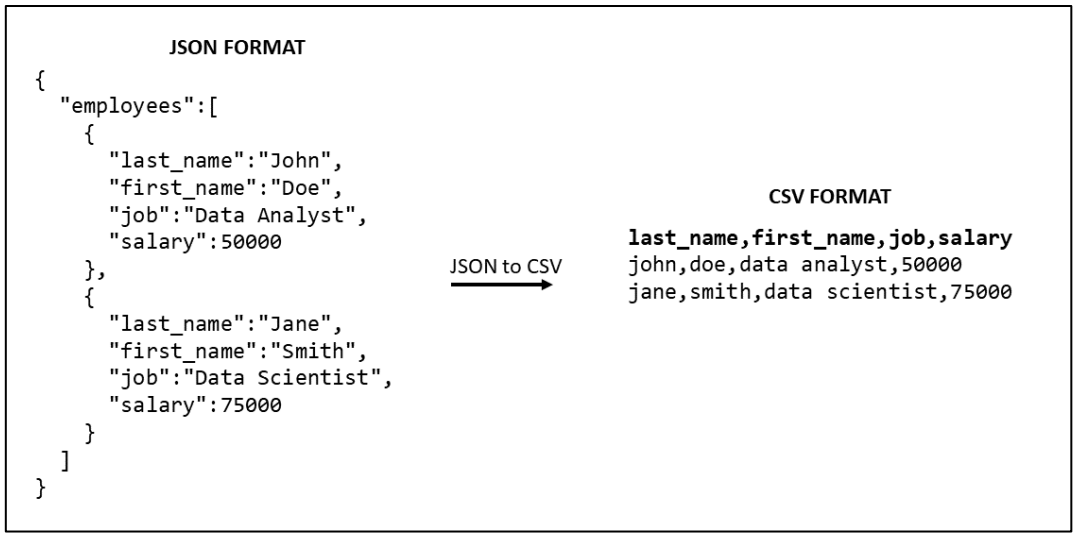

Figure 5. Exemple d'un traitement de données : conversion de données en format JSON en un format CSV.

Même une fois transformées en données tabulaires, les données peuvent avoir besoin d'une autre transformation. En effet, certaines données tabulaires sont en format « wide » quand d'autres sont en format « long ». Un format « wide » contient des valeurs qui ne se répètent pas dans la première colonne. Un format « long », au contraire, contiendra des valeurs qui se répètent dans la première colonne (Figure 6).

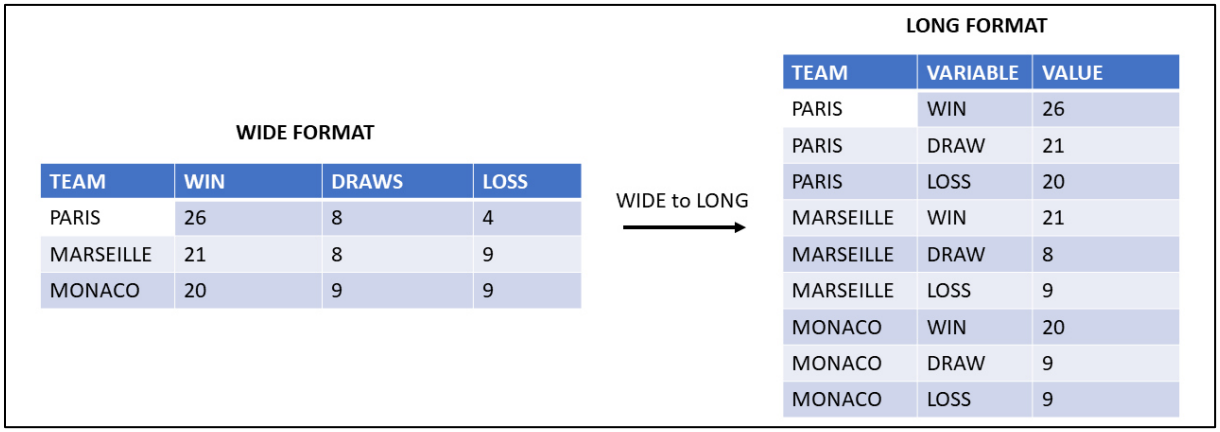

Figure 6. Transformation d'un jeu de données en format « wide » en un format « long ».

Il arrive souvent que les jeux de données récupérés dans le monde réel contiennent des données manquantes : c'est lorsqu'il n'y a pas de valeur stockée dans le champ d'une observation. Le data analyst devra décider comment traiter ces valeurs manquantes : supprimer l'observation entièrement, remplacer la valeur manquante par une moyenne/medianne ou même utiliser du machine learning pour estimer la valeur manquante.

L'énumération des opérations de pré-traitement faite ici n'est pas exhaustive. Il existe d'autres opérations non citées ici comme : supprimer les valeurs dupliquées, exclure les données aberrantes ou encore changer les types de variables (entier, chaîne de caractères, date etc.). Une fois le

prétraitement effectué (processus qui peut prendre beaucoup de temps), le data analyst peut passer à l'analyse de données à proprement parlé.

B. Analyse de données

L'analyse de données a différentes facettes et approches et a des applications très variées. De ce fait, il existe différents types d'analyses de données. On peut en citer quatre principales :

- L'analyse exploratoire (ou data mining) : elle met l'accent sur l'exploration et la découverte de patterns dans un jeu de données encore non connu par l'analyste. Il va permettre de gagner de la connaissance sur un sujet grâce aux informations « cachées » dans les données.
- L'analyse descriptive (liée à la business intelligence) : elle repose plutôt sur la visualisation et les statistiques descriptives (moyenne, médiane etc.) et porte souvent sur des données business. Par exemple, un histogramme présentant le chiffre d'affaires d'une entreprise sur une année, mois par mois.
- L'analyse confirmatoire : on va analyser les données pour confirmer ou non des hypothèses de départ. Par exemple, la campagne marketing a-t-elle généré plus de ventes ce semestre par rapport au semestre dernier.
- L'analyse prédictive (machine learning) : utilisation de modèles statistiques ou mathématiques afin de prédire une issue (ex. prédire les ventes d'un produit ou l'attrition) ou encore de faire de la classification (ex. bons clients, mauvais clients) ou encore de faire de la détection d'anomalies (ex. fraude).

Selon les besoins de l'entreprise et/ou le service/produit qu'elle veut proposer, elle n'utilisera pas la même approche, ni les mêmes ressources (elle aura besoin de data analysts pour faire les trois premiers types d'analyses de données et aura besoin de data scientists dans le cas d'analyse prédictive). Une fois que l'entreprise aura analysé les données à sa convenance et aura compris les possibilités qu'offrent les données collectées, elle aura la liberté de créer un service/produit basé sur ces dernières.

V. LES SERVICES ET PRODUITS BASÉS SUR LA DONNÉE OUVERTE

Un produit basé sur la donnée ouverte (PBD) peut se définir comme « le résultat d'un processus durant lequel des données sont transformées en informations accessibles via un service, une infrastructure, une analyse ou une combinaison des trois » (Arribas-Bel et al., 2021). La différence entre un PBD et de simples données ouvertes est la valeur ajoutée, qui étend l'accessibilité et l'utilisation des données ouvertes (sans quoi, ces données ne seraient pas exploitées et l'information qu'elles contiennent non

partagée). À noter qu'il est possible d'utiliser des données « fermées » en complément de données ouvertes collectées pour améliorer le service/produit proposé.

Afin de créer un PBD, il est primordial d'identifier un problème qui demande à être traité. Développer des PBD utiles (et donc à forte valeur ajoutée) requiert souvent de penser moins au « quoi » qu'au « qui » pourrait être intéressé de l'utiliser (Arribas-Bel et al., 2021). De ce fait, identifier les potentiels utilisateurs finaux, comprendre ce qui est faisable ou non avec la donnée disponible et mesurer les compétences et ressources disponibles aideront à maximiser la pertinence d'un PBD pour une entreprise. Un exemple récent et clair de l'importance d'identifier un problème qui a besoin de réponse se trouve dans la récente pandémie. Il était important de comprendre l'impact de la pandémie sur des populations hétérogènes à travers le monde. Or, il existait peu de jeux de données prêts à l'emploi pour comprendre les dynamiques de la pandémie ces populations hétérogènes. Pour combler ce manque, Dong et al. (2020) de l'université de Johns Hopkins ont fait un travail d'agrégation des données COVID publiques publiées par les états du monde entier et ont construit une application web interactive qui permet de suivre en temps réel l'évolution de la pandémie.

Ainsi, développer un PBD ne se résume pas à rendre des données brutes accessibles. Pour développer un tel produit, il est primordial de traiter, analyser et de construire sur les données originales. Cela renforce la valeur de l'information et augmente les possibilités d'insight. Améliorer l'utilisabilité de la donnée aide à augmenter l'accès à la donnée, notamment quand l'acquisition de la donnée est coûteuse. Mais le traitement et l'exploitation des données nécessite souvent des compétences de programmation qui ne sont pas à la portée du monde. Combler ce manque de compétences permet d'ouvrir la donnée et l'information qu'elle contient à un public beaucoup plus large. Cela est particulièrement pertinent pour les populations non initiées à l'analyse de données qui, si les PBD sont combinés avec des visualisations et des ressources interactives, peuvent interagir avec des données complexes d'une manière qui leur serait autrement inaccessible (comme dans le cas de la carte interactive de Dong et al. (2020)).

Maintenant que le processus de création de valeur de l'open data a été décrit, passons à un cas d'usage : une application web de suivi de tendances des cryptomonnaies. Cette application a été développée lors de mon stage de fin d'études chez CoinShares et a pour vocation d'aider l'entreprise à monitorer les cryptomonnaies connaissant un fort engouement.

PARTIE II. ÉTUDE DE CAS : APPLICATION WEB DE SUIVI DE TENDANCES DE CRYPTOMONNAIES

OBJECTIF DE L'APPLICATION

Comme évoqué en introduction, le cœur de métier de CoinShares est la proposition de services et de conseils en investissement de crypto-actifs. L'entreprise se doit donc d'être parfaitement informée des évolutions de ce marché, d'autant plus que l'écosystème crypto évolue très vite. On recensait environ 1300 cryptomonnaies en 2017, 6000 en 2021 et plus de 10000 en février 2022 (Statista, 2022). De plus toutes les cryptomonnaies ne connaissent pas le même engouement : Bitcoin connait un engouement croissant et assez stable depuis des années, Ethereum également depuis sa création en 2015 mais a récemment vu son engouement exploser courant 2022 depuis l'annonce de « The Merge » (passage de la blockchain du proof-of-work au proof-of-stake[7]) qui aura lieu courant septembre 2022. D'autres cryptos, au contraire, connaissent ou ont connu une perte d'intérêt comme le projet GetGems (cryptomonnaie GEMZ) qui était censée être une messagerie depuis laquelle les utilisateurs pouvaient envoyer et recevoir des Bitcoins. Après avoir levé près d'un million de dollars, l'application n'a jamais été délivrée, et GEMZ n'existe plus aujourd'hui. Le lecteur l'aura compris, l'écosystème crypto, car très jeune, est en perpétuelle évolution et de nouveaux projets sont créés chaque mois. Certains persistent dans le temps, d'autres disparaissent. Durant leur chemin, certains projets connaissent des hauts et des bas et peuvent attirer l'attention plus que d'autres. Le prix de la cryptomonnaie associée au projet fluctue tout autant. Il est donc primordial pour CoinShares d'être au plus près des évolutions d'un grand nombre de projets cryptos.

L'application web créée a pour vocation de répondre à ce besoin. Le processus de création de l'application est détaillé dans les parties suivantes mais voici brièvement comment fonctionne l'application : elle récupère des données rendues publiques par des entreprises privées comme Twitter ou CoinMarketCap, les agrègent et les stocke dans une base de données, puis requête cette même base pour calculer des indicateurs et pour visualiser les données dans une interface interactive (Figure 7).

[7] Le proof-of-work et le proof-of-stake sont des méthodes de validation des blocs d'une blockchain par les validateurs.

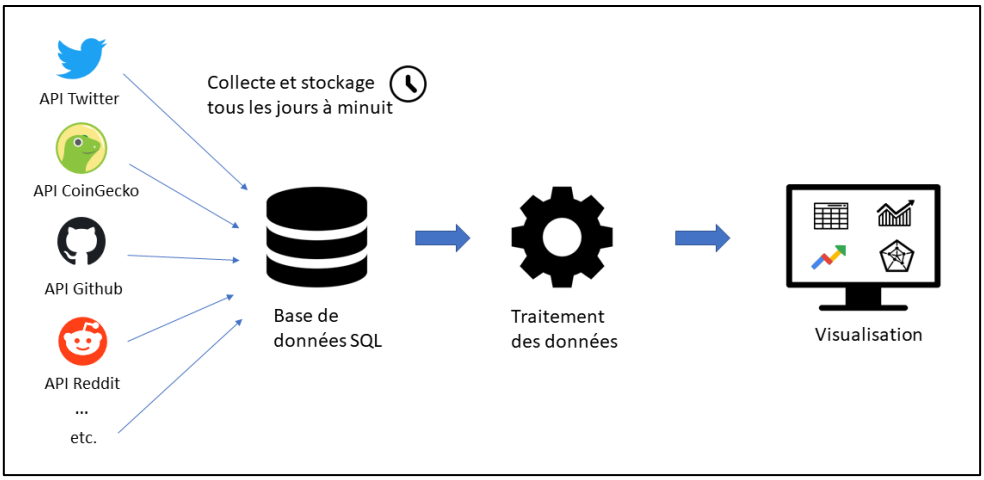

Figure 7. Fonctionnement de l'application web. Les données provenant de différentes APIs (Twitter, Reddit, Github etc.) sont collectées tous les jours à 00:00:00 UTC. Ces données sont ensuite stockées dans une base de données SQL. L'application requêtes ces données de la base, les traite (gére les données manquantes, calcul des métriques) et les affichent sous la forme d'un tableau de bord.

Ainsi, à très court-terme, l'application permettra aux utilisateurs (internes à l'entreprise pour le moment) de suivre les tendances des cryptomonnaies ce qui leur permettra de faire de meilleurs choix (par exemple le service marketing pourra communiquer sur cette cryptomonnaie et donner de l'information aux clients ou encore éviter de proposer des services autour d'une cryptomonnaie en perte de vitesse). Dans un futur proche, l'application permettra l'amélioration du produit phare Napbots car CoinShares pourra ajouter les cryptomonnaies qui connaissent un engouement (donc qui sont demandées par le public) au catalogue du produit. Ainsi cela pourra attirer de nouveaux clients qui seront séduits par la possibilité d'investir de manière professionnelle sur leur cryptomonnaie favorite.

ÉTAPES DU PROCESSUS DE CRÉATION DE VALEUR DE L'OPEN DATA APPLIQUÉ AU DÉVELOPPEMENT DE L'APPLICATION

I. LA CRÉATION DE LA DONNÉE : DES DONNÉES CRÉÉES ET PUBLIÉES PAR DES ENTREPRISES PRIVÉES

Contrairement au modèle théorique présenté Partie I, qui stipule que les données ouvertes sont fournies par le secteur public, l'ensemble des données récoltées qui serviront de matière première à l'application sont des données produites par des acteurs privés tels Twitter ou CoinMarketCap, et mises à disposition via des API. Il existe une multitude d'acteurs mettant à disposition une vaste panoplie de données qui permettront d'alimenter l'application.

A. Données de l'intérêt du public

Il est possible de récupérer des données traitant de la popularité publique autour de telle ou telle cryptomonnaie. Twitter par exemple, met à disposition le nombre de Tweets (horaires ou quotidiens) avec une mention particulière (nombre de Tweets avec la mention « #BTC » par exemple). Cette donnée peut être utile pour estimer l'intérêt d'une cryptomonnaie à travers le temps. Dans la même lignée, Google Trends permet de mesurer la popularité de mots clés recherchés dans le moteur de recherche Google. On peut ainsi avoir une estimation de la popularité de la recherche « Bitcoin » à travers le temps. Le lecteur peut sans doute voir se dessiner une méthodologie permettant d'arriver à notre but à savoir, mesurer et detecter l'engouement autour de telle ou telle cryptomonnaie. En outre, CoinMarketCap ou CoinGecko, des sites de recensement et de classement des cryptomonnaies, fournissent des données intéressantes pour le cas d'usage comme le nombre de personnes ayant ajouté une crypto à leurs favoris (CoinGecko), ou encore le nombre de fois qu'une crypto a été ajoutée à une « watchlist » (liste de suivi de cryptomonnaies). Toutes ces données permettront de mesurer un score de popularité.

B. Données communautaires

L'écosystème crypto étant très communautaire, il est également important de récupérer des données mesurant l'engagement de la communauté. Beaucoup de communautés cryptos se retrouvent sur Telegram ou Discord. Le canal Discord Uniswap (18^{e} crypto en termes de capitalisation) par exemple compte plus de 92000 membres à l'écriture de ces lignes. À titre de comparaison, Ethereum (qui est la 2^{e} cryptomonnaie en termes de capitalisation de marché) compte « seulement » 33000 membres sur Discord. Le nombre de membres Discord n'est donc pas proportionnel à la capitalisation des cryptomonnaies, il sera donc important de prendre en compte cette donnée dans le calcul du score d'engouement. Au même titre, l'activité du subreddit associé à la crypto (le nombre de membres, le nombre de publications ou même le nombre de commentaires) ainsi que le nombre d'abonnés sur le compte Twitter officiel du projet, sont des données disponibles librement et sont de bons indicateurs de l'engouement communautaire autour d'une crypto.

C. Données de marché

Les cryptomonnaies sont avant tout des actifs qui s'échangent entre des acheteurs et des vendeurs sur les marchés financiers. Lorsqu'il y a plus d'acheteurs que de vendeurs, le prix monte ; lorsque c'est le contraire, le prix baisse. Le prix est donc dans une certaine mesure, un indicateur de l'intérêt (ou du désintérêt) d'un crypto : quand les gens croient en la crypto et son projet, ils vont l'acheter et faire monter le prix, et au contraire, s'ils se méfient ou ne croient plus en le projet, ils vendront l'actif et feront baisser le prix.

Au prix, on peut ajouter le volume de transactions journalier : il s'agit de la valeur échangée (exprimée en $) entres les différents acteurs du marché . Si, au cours d'une journée, il y a 10 transactions d'un actif valant en moyenne 100$, le volume journalier sera de 1000$. C'est une donnée importante car un volume en constante augmentation signifie une entrée constante de nouveaux acteurs sur le marché (ou une augmentation de la valeur des transactions) et donc témoignent d'un intérêt pour la crypto. Ces données de marché sont facilement accessibles via l'API de CoinGecko (entre autres) qui fait un travail de recensement d'une multitude de données de marché des cryptomonnaies.

D. Données développeurs

Comme évoqué précédemment, les projets cryptos sont avant tout des projets technologiques développés sur une blockchain. Des développeurs vont coder le projet (que ce soit une application, un site web ou autre) et le maintenir. Personne d'autre n'est mieux placé pour connaître le fonctionnement du projet qu'eux. Les développeurs sont donc un maillon essentiel d'un projet crypto. Or, il existe un site dédié aux développeurs, leur permettant de poster leur code, de proposer des suggestions aux projets existants, ou encore de suivre des projets : Github. Github est le lieu privilégié des développeurs et il est possible de suivre l'activité de tel ou tel projet sur le site. Par exemple, le code source du projet Ethereum est posté sur Github[8] et l'on peut avoir accès à trois métriques importantes :

- « Watch » : nombre de personnes qui suivent le projet
- « Fork » : nombre de personnes ayant cloné le projet
- « Star » : nombre de personnes ayant ajouté le projet dans leurs favoris

Github possédant une API et mettant à disposition ces données via cette dernière, il sera possible de les intégrer à l'application.

Les acteurs privés mettent donc à disposition du grand public de nombreuses données très diverses. Un travail de sélection a été fait, avec toujours en vue concernant le but final de l'application : monitorer l'engouement des cryptomonnaies. L'application utilisera donc 14 données venant de 8 sources différentes (Tableau 1).

[8] Code source du projet Ethereum disponible à cette adresse : https://github.com/ethereum/go-ethereum

Source	Donnée	Catégorie
Twitter	Nombre d'abonnés	Community
	Nombre de Tweets	Public
Telegram	Nombre de membres du canal	Community
Discord	Nombre membres du serveur	Community
Reddit	Nombre de membres du subreddit associé au projet	Community
Google Trends	Score de la recherche « #SYMBOL » (SYMBOL étant le symbole de la cryptomonnaie, par exemple BTC pour Bitcoin)	Public
CoinGecko	Nombre de mentions j'aime	Public
	Prix, volume et capitalisation en $	Market
CoinMarketCap	Nombre d'ajouts à des watchlists	Public
Github	Nombre de 'watch'	Developer
	Nombre de 'fork'	Developer
	Nombre de 'star'	Developer

Tableau 1. Tableau récapitulatif des données sélectionnées pour alimenter l'application ainsi que leur source. On notera la diversité des sources utilisées. Un travail de validation et d'agrégation de la donnée sera essentiel.

Pour ajouter de la valeur à toutes ces données brutes, il convient de valider leur qualité.

II. LA VALIDATION DE LA DONNÉE

Pour pouvoir valider la donnée collectée, il faut s'assurer qu'elle respecte bien les cinq critères de pertinence cités en Partie I : l'exactitude, la fraicheur, la comparabilité et l'intégrité.

A. Pertinence

Les données collectées sont de facto pertinentes pour l'application puisqu'elles ont précisément été sélectionnées pour répondre à l'objectif initial, à savoir, détecter et comparer les engouements autour de certaines cryptomonnaies. Des données comme le nombre de « Star[9] » sur Github ont été préférées au nombre d'« Open Issues ». Les « Open Issues » sont des problèmes relevés par les utilisateurs/développeurs du projet sur Github qui n'ont pas encore été corrigés. Ces problèmes

[9] Sur Github, les « Star » représentent le nombre de fois qu'un projet à été ajouté aux favoris par les utilisateurs.

ouverts permettent de prévenir les développeurs cœurs du projet de corriger les éventuels bugs trouvés et d'ainsi résoudre le problème. Aussi intéressant que cette donnée soit, il a été estimé qu'elle n'était pas aussi pertinente que le nombre de « Star » du projet par exemple. En effet, un grand nombre d'open issues n'indique pas forcément un grand intérêt pour le projet crypto. Cela peut simplement vouloir dire que le projet comporte beaucoup de bugs. Ainsi la sélection faite en amont a permis de sélectionner les 14 données les plus pertinentes parmi toutes celles disponibles en open data sur internet.

B. Exactitude

Un travail de vérification de l'exactitude a été fait lors du développement de l'application afin de vérifier que les données récupérées via les APIs étaient bel et bien conformes à la réalité (quand c'était possible). C'est un travail critique car des données erronées peuvent induire en erreur et créer des fausses informations. Pour se faire, il a été vérifié manuellement (quand c'était possible) si la donnée renvoyée par l'API était correcte. Par exemple, il a été vérifié que le nombre de membres du subreddit d'une crypto en particulier retourné par l'API correspondait bien à ce qui était affiché sur le site web de Reddit ; ou encore que le prix d'une crypto à une date donnée renvoyé par l'API de CoinGecko corresponde bien au prix affiché par plusieurs plateforme de trading (comme Binance ou FTX) à cette même date. Malheureusement toutes les données ne sont pas vérifiables. Par exemple, le volume d'échange d'une cryptomonnaie sur une journée donné par l'API CoinGecko n'est pas vérifiable car CoinGecko additionne les volumes d'échange de plusieurs plateformes de trading sans préciser lesquelles. Dans ce cas, on est obligé de faire confiance à CoinGecko. Il s'agit d'une limite mais il n'y a pas vraiment d'alternative à ce problème. Cela vaut aussi pour la capitalisation ou encore le nombre de Tweets.

C. Fraîcheur

Les données récupérées jouissent d'une très bonne fraîcheur puisque toutes les données récupérées sont mises à jour quotidiennement par les entreprises émettrices. Et le fait d'utiliser des APIs (versus télécharger des jeux de données manuellement) pour récolter les données procure l'avantage de récupérer des données fraîches rapidement en format facilement traitable par un ordinateur. Pour notre cas d'usage, étant donné que l'on ne s'intéresse qu'à des données journalières, il n'est pas primordial de récolter les données le plus rapidement possible, mais l'utilisation d'API a tout de même cet avantage-là. Le travail de fraîcheur est donc fait en amont lors de la création et publication de la donnée de la part des entreprises émettrices.

D. Comparabilité

Les données récoltées jouissent (pour l'instant) d'une bonne comparabilité. En effet, toutes les données récupérées sont exclusivement des valeurs numériques, à savoir, des entiers ou des réels (ex. nombre d'abonnés Twitter, score Google Trends etc.). Il n'y a donc aucun souci de comparabilité pour l'instant. Cependant, il n'est pas exclu qu'un jour, l'une des entreprises émettrices décide de changer l'unité d'une des données publiées. Par exemple, Google Trends pourrait décider de changer l'échelle de son score de popularité (aujourd'hui de 0 à 100) en un score allant de -1 à 1. Cela impliquerait une modification du prétraitement (voir la partie sur L'analyse de données) afin de convertir les différentes échelles et de retrouver une unité commune. C'est pourquoi un travail de maintenance est indispensable sur ce genre d'application utilisant des APIs externes, car il n'est pas exclu de voir survenir des problèmes ou des surprises dans les données récupérées. C'est en partie le travail du data analyst au sein d'une entreprise. C'est pourquoi les professionnels de la donnée sont une ressource indispensable pour une entreprise voulant dégager de la valeur de l'open data.

E. Intégrité

Il est important de vérifier l'intégrité des données collectées car des valeurs manquantes peuvent exister. En effet, étant récupérées tous les jours via des requêtes API, il est possible que certaines requêtes n'aboutissent pas, le serveur interrogé pouvant renvoyé un message d'erreur HTTP du type *500 : Internal Server Error* ou *503 : Service Unavailable*. Il est également possible que ce soit le serveur sur lequel l'application est hebergée qui fasse défaut (même si c'est plus rare). Dans ce cas le programme assignera une valeur nulle au champ ou à l'enregistrement en question. Une valeur manquante peut aussi venir du simple fait que l'entreprise émettrice n'ait pas la donnée en question et renvoie elle-même une valeur nulle. Il faudra faire attention à ces valeurs nulles lors de l'analyse de données et décider de comment les traiter. Quoiqu'il en soit, lors du développement de l'application, il a été vérifié que les points API interrogés renvoient bien des valeurs non nulles. Mais des exceptions peuvent arriver, des gestionnaires d'exceptions ont donc été implémentés dans le cas où cela arriverait.

Il est également important de vérifier le type des données reçues. Un symbole (ex. ETH) doit être une chaîne de charactères, un nombre d'abonnés Twitter doit être un entier (positif) et une date doit être en format date. Un nombre d'abonnés Twitter égal à 0 ou -1 par exemple, signifiera une donnée manquante. Il est important de détecter et de corriger ces anomalies afin de les exclure des futures analyses. On peut les remplacer par des valeurs nulles ou bien les marquer comme valeurs aberrantes afin de les gérer plus tard lors de l'analyse de données.

En s'assurant que les données correspondent aux standards de qualité, on s'assure d'avoir une matière première exploitable et (presque) sans défaut pour notre application. Les données jouissant d'une

bonne qualité, il est maintenant possible de les agréger afin d'avoir suffisamment de matière première à exploiter.

III. L'AGRÉGATION ET LE STOCKAGE DE LA DONNÉE : RENDRE ACCESSIBLE LES DONNÉES COLLECTÉES

Une fois les valeurs manquantes et les types de données validés, les données sont agrégées et stockées dans une base de données SQL. Une base de données SQL est un type de base de données qualifiée de relationnelle. Cela signifie qu'il existe des relations entre les différentes données ; et c'est le cas pour nos données : chaque métrique est reliée à un symbole et une date. Le fait de stocker les données dans une BDD relationnelle les rend facilement accessibles par de simples requêtes SQL. Par exemple, si l'on veut récupérer le nombre d'abonnés Twitter d'Ethereum à la date du 20 août 2022, il suffira d'exécuter la requête suivante : *SELECT * FROM metrics WHERE symbol = 'ETH' AND date = '2022-08-20'*. Ainsi nos métriques sont facilement accessibles par un programme informatique ce qui facilitera le travail d'analyses plus tard. À titre d'exemple, une autre alternative de stockage aurait été de stocker les données dans des fichiers CSV en local sur le serveur herbgeur. Mais d'une part le requêtage des données aurait été plus compliqué : il aura fallu importer, lire les fichiers CSV et sélectionner la métrique désirée à chaque fois que l'on aurait eu besoin d'accéder à une donnée. D'autre part, il aurait été impossible d'accéder aux données via une plateforme d'administration de BDD comme pgAdmin.

Afin d'être réutilisables par d'autres personnes plus tard, il est important de documenter les tables de la base de données. Pour cela on établit un dictionnaire de données : un document répertoriant les différents champs d'une table, leur signification et leur domaine (ex. texte, date, entier etc.). C'est un document primordial car sans celui-ci il est parfois impossible de comprendre à quoi correspond tel ou tel champ d'une table dans une BDD. Le but de l'application étant de créer de la valeur à partir de la donnée, il est nécessaire de donner un minimum d'information (et donc du sens) à ces données stockées. En Annexe 1, le lecteur trouvera le dictionnaire de données de la table 'coins'. Cette table contient tous les attributs nécessaires à la récolte des données de chaque cryptomonnaie. Par exemple, le champ 'github' contient le chemin identifiant le dépôt du code source du projet en question. Par exemple, 'thesandboxgame/sandbox-smart-contracts' est le chemin que l'on va fournir en paramètres lors de l'appel API de Github. Github retournera les données concernant ce dépôt.

Le Tableau 2, quant à lui, est le dictionnaire de données de la table 'metrics' qui stocke toutes les métriques (nombre d'abonnés twitter, nombre de « Star » sur Github etc.) pour chaque crypto et pour chaque jour.

Attribut	Signification	Domaine
symbol	Symbole identifiant la cryptomonnaie	texte
date	Date de récupération des données	date
twitter_followers	Nombre d'abonnés Twitter	entier
n_tweets	Nombre de Tweets avec mention '#SYMBOL'	entier
telegram_followers	Nombre de membres du canal Telegram	entier
discord_followers	Nombre de membres du serveur Discord	entier
subreddit_followers	Nombre de membres du subreddit	entier
google_trends_score	Score Google Trends normalisé du terme '#CRYPTO' (ex. 'bitcoin')	réel
cmc_watchlists	Nombre d'ajouts de la crypto à des watchlists sur CoinMarketCap	entier
coingecko_likes	Nombre de mentions « j'aime » de la crypto sur CoinGecko	entier
github_watchs	Nombre de « Watch » du dépôt Github	entier
github_forks	Nombre de « Fork » du dépôt Github	entier
github_stars	Nombre de « Star » du dépôt Github	entier
price	Prix de la crypto en ouverture à 00:00:00 UTC (en USD)	réel
volume	Volume de transaction sur la journée entière (en USD)	réel
market_cap	Capitalisation (en USD)	réel

Tableau 2. Dictionnaire des données de la table 'metrics' de la base de données SQL. À noter qu'il y a d'autres attributs présents dans la table 'metrics', non utilisés par l'application et non cités ici.

Une fois les métriques stockées et rendues facilement accessibles, elles peuvent être analysées afin d'en tirer un maximum de valeur.

IV. L'ANALYSE DE DONNÉES : LE CŒUR DE LA CRÉATION DE VALEUR

L'analyse de données dans notre cas d'usage consistera essentiellement en une analyse de données descriptive avec quelques calculs de rendement, des normalisations et de la visualisation. Mais avant de calculer un quelconque indicateur, et de visualiser les données il faut les prétraiter.

A. Pré-traitement de la donnée

La qualité des données ayant été validée précédemment, il y a peu de prétraitement de données à effectuer. Toutes les données sont en format tabulaire (car stockées dans une BDD relationnelle), en format « long », il n'y a pas de données dupliquées car ne on récupère qu'une seule fois par jour les données nécessaires et les types de variables sont déjà corrects car définis dans la BDD. Il ne reste qu'à gérer les potentielles données aberrantes (comme un -1 ou 0 pour un nombre d'abonnés Twitter) : on remplace ces données par des valeurs nulles (NaN). En effet, il est inutile de supprimer tout l'enregistrement lorsqu'un seul champ contient une valeur aberrante. Le fait de remplacer ces valeurs par des valeurs nulles permettra d'ignorer ces valeurs de nos calculs.

Quelques transformations des métriques sont également nécessaires afin de calculer le score d'engouement. Il faut par exemple normaliser le score Google Trends ou encore lisser (en utilisant une moyenne mobile) le nombre de Tweets quotidien. On va également créer une métrique 'volume / market cap' qui est un indicateur du gain d'intérêt de la cryptomonnaie sur les marchés. Le Tableau 3 récapitule les transformations nécessaires afin de passer d'un champ de la BDD à une métrique exploitable.

Champ	**Métrique**	**Transformation**
twitter_followers	Nombre d'abonnés Twitter	-
n_tweets	Nombre de tweets sur 48h	Moyenne mobile (période=2 jours) du nombre de Tweets
telegram_followers	Nombre d'abonnés Telegram	-
discord_followers	Nombre d'abonnés Discord	-
subreddit_followers	Nombre d'abonnés du subreddit	-
google_trends_score	Score Google Trends	Produit cumulé de (1 + google_trends_score)
cmc_watchlists	Nombre d'ajouts à une watchlist	-
coingecko_likes	Nombre d'ajouts aux favoris	-
github_watchs	Nombre de 'Watch' sur Github	-
github_forks	Nombre de 'Fork' sur Github	-
github_stars	Nombre de 'Star' sur Github	-
price	Prix en dollars	-

volume	Volume en dollars	-
market_cap	Capitalisation en dollars	-
volume et market_cap	Engouement sur les marchés	Division du volume par la capitalisation

Tableau 3. Table de correspondance entre les champs de la table 'metrics' et les métriques calculées.

B. Méthodologie de calcul du score d'engouement

Afin de comparer les cryptomonnaies entre elles (en termes d'engouement), la méthodologie suivante a été appliquée :

Premièrement chaque métrique est regroupée dans sa catégorie respective (voir Tableau 1).

Ensuite, pour chaque métrique dans chaque catégorie (ex. nombre d'abonnés Twitter dans la catégorie 'Community'), la progression de chaque crypto est calculée sur une période donnée (ex. 1 mois). On obtient ainsi un taux de progression pour chaque cryptomonnaie.

Ensuite, une note est assignée à chaque cryptomonnaie en fonction du taux de progression. Pour cela, les taux de progression de chaque cryptomonnaie sont mis à l'échelle entre 0 et 1 en utilisant la formule utilisée par scikit-learn dans l'implémentation de leur MinMaxScaler[10]. Pour mieux comprendre, considérons 3 cryptomonnaies, A, B, C. La cryptomonnaie A obtient 15% de progression de ses abonnés Twitter sur un mois, B obtient 10% et C seulement 5%. Un score (de 0 à 1) est alors assigné à chaque crypto. La crypto A obtiendra 1/1, B obtiendra 0.5/1 et C obtiendra 0/1. Avec quatre cryptos obtenant 20%, 15%, 10% et 5% de progression, les scores obtenus seraient respectivement de 1, 0.66, 0.33, 0. Ce processus de notation est réitéré pour chaque autre métrique dans chaque catégorie. Ainsi chaque cryptomonnaie obtient un score pour chaque métrique (excepté pour les métriques qui manquent, par exemple certaines cryptos n'ont pas de métrique pour Github, elles n'auront donc pas de note 'Developer'.

Une fois les scores calculés pour chaque métrique de chaque catégorie, des poids sont appliqués sur chaque score selon le degré d'importance estimé (de manière arbitraire). Par exemple, dans la catégorie 'Public', le score Google Trends a plus d'importance que le nombre de mentions j'aime sur Coingecko (ceci est bien sûr arbitraire). Un poids plus important est donc assigné pour le score Google Trends que pour les mentions j'aime de Coingecko. Chaque métrique dans une catégorie a donc un poids différent. On calcule ensuite une moyenne pondérée des scores par catégorie (en tenant donc compte des poids assignés). On obtient ainsi quatre notes pour les quatre catégories.

[10] MinMaxScaler de scikit-learn : https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html

Enfin pour obtenir un score global d'engouement on calcule une moyenne pondérée des quatre scores (correspondants aux quatre catégories) en assignant un poids à chaque catégorie. Le Tableau 4 récapitule les poids assignés à chaque métrique et à chaque catégorie.

Catégorie	Métriques	Poids
Public		1
	Nombre d'abonnés Twitter	1
	Nombre d'abonnés Telegram	1
	Nombre d'abonnés Discord	1
	Nombre d'abonnés subreddit	1
Community		1
	Score Google Trends	1
	Nombre de Tweets (sur 48h)	1
	Nombre de watchlists CoinMarketCap	1
	Nombre de mentions j'aime CoinGecko	1
Market		0.25
	Prix	1
	Volume	1
	Capitalisation	1
	Volume / Capitalisation	1
Developer		0.5
	Nombre de Watch sur Github	1
	Nombre de Fork sur Github	1
	Nombre de Star sur Github	1

Tableau 4. Tableau récapitulatif des poids assignés aux différentes catégories et métriques. Le tableau se lit comme-suit : la catégorie 'Public' contient quatre métriques (Nombre d'abonnés Twitter, Nombre de membres Telegram, Nombre de membres Discord, Nombre de membres subreddit). Chaque métrique a un poids (ici elles ont toutes un poids de 1). Finalement, chacune des quatre catégories a elle aussi un poids (Public :1, Community : 1, Market : 0.25, Developer : 0.5).

Ainsi, et toujours dans un souci de maintenabilité de l'application, il sera aisé pour le développeur qui maintiendra l'application de changer le poids de quelconque métrique ou catégorie dans le calcul final du score de popularité.

V. VISUALISATION DE LA DONNÉE

Rappelons que l'un des objectifs de l'application est de permettre aux utilisateurs (collaborateurs de l'entreprise) d'avoir une idée rapide de l'engouement de telle ou telle cryptomonnaie. Même si présenter un score pour chaque crypto est possible, il est beaucoup plus attrayant et parlant de visualiser les données. La visualisation s'articule en 2 axes : classement des cryptomonnaies en termes d'engouement global et évolution des différentes métriques.

A. Classement global des cryptomonnaies

Afin d'avoir une idée rapide de quelles cryptomonnaies connaissent un engouement fort, le choix a été fait de présenter un classement des cryptos sous forme de tableau affichant avec le score d'engouement global ainsi que le détail des scores des différentes catégories (Figure 8).

name	commu	public	market	develop	overall
filter data...					
Evmos	0.42	0.65	0.47	0.59	0.54
Lido DAO	0.41	0.61	0.35	0.67	0.52
Chain	0.55	0.39	0.12	0.56	0.46
EOS	0.33	0.079	0.35	0.88	0.34
Radix	0.18	0.4	0.11	0.6	0.33
DeFiChain	0.47	0.12	0.089	0.54	0.32
Chiliz	0.2	0.31	0.82		0.32
Osmosis	0.2	0.13	0.3	0.9	0.31
Flow	0.25	0.2	0.24	0.58	0.29
Klaytn	0.34	0.11	0.12	0.61	0.29
Moonbeam	0.27	0.13	0.11	0.73	0.29
Ethereum Classic	0.21	0.24	0.27	0.53	0.28
PAX Gold	0.25	0.081	0.11	0.81	0.28
NEAR Protocol	0.24	0.081	0.17	0.79	0.28

Figure 8. Classement des cryptomonnaies en fonction de leur score d'engouement au 26 août 2022. Les cryptomonnaies sont ici classées de manière décroissante selon leur score d'engouement ('overall') (la première cryptomonnaie 'Evmos' a un score d'engouement plus élevé que 'Lido DAO' qui elle-même a un score d'engouement plus élevé que 'Chain'). Les quatre catégories utilisées pour calculer le score global sont aussi affichées afin de mieux comprendre quelle(s) catégorie(s) a/ont influencé le score global. Pour rappel, le score d'engouement est une moyenne pondérée des quatre catégories selon leur poids respectif.

D'un coup d'œil, l'utilisateur peut ainsi voir quelles cryptomonnaies connaissent un fort engouement (haut du tableau) versus lesquelles connaissent un faible engouement (bas du tableau). En ayant le détail des scores de chacune des quatre catégories, l'utilisateur peut aussi savoir quelles catégories influencent le plus le score global. Dans la Figure 8, on voit que Evmos a le score d'engouement le plus haut avec 0.54 (sur 1), et ceci est expliqué par des bonnes notes obtenues dans chacune des quatre catégories (0.42 pour 'Community', 0.65 pour 'Public', 0.47 pour 'Market', 0.59 pour 'Developeur'). Pour rappel le score d'engouement est calculé en pondérant le score de chacune des catégories ; on peut retrouver le score d'engouement d'Evmos via ce calcul : (0.42 [Communauté] * 1 [poids Community] + 0.65 [Public] * 1 [poids Public] + 0.47 [Market] * 0.25 [poids Market] + 0.59 [Developer] * 0.5 [poids Developer]) / (1 + 1 + 0.25 + 0.5) ~= 0.54. Le fait de connaître l'influence de chaque catégorie dans le score d'engouement final est important car avoir un score d'engouement élevé grâce à un score élevé de 'Community' n'a pas la même signification que d'avoir un score d'engouement élevé à grâce à un score élevé de 'Developer' et de 'Market'.

En plus du classement des cryptomonnaies sous forme de tableau, il peut être pertinent de présenter un graphique du type diagramme de Kiviat (ou radar chart) afin de comparer plus facilement les cryptomonnaies entre elles. Un diagramme de Kiviat est une manière de visualiser des données multivariées sous la forme d'un graphique à deux dimensions d'au moins trois variables quantitatives représentées sur des axes commençant au même point. Cette manière de représenter les données est ici appropriée car elle permet de représenter et comparer facilement les quatre catégories (qui font guise de variables quantitatives) pour chacune des cryptos. Ainsi, d'un coup d'œil, la Figure 9 permet de comprendre pourquoi Evmos a un meilleur score d'engouement qu'Ethereum Classic : Evmos a un meilleur score qu'Ethereum Classic dans les quatre catégories permettant de calculer le score global d'engouement.

Figure 9. Diagramme de Kiviat permettant de comparer deux cryptomonnaies (ou plus) entre elles selon les quatre catégories 'Community', 'Public', 'Market' et 'Developer'. Dans cet exemple, on compare les deux cryptos Evmos et Ethereum Classic. Grâce à la visualisation en diagramme de Kiviat, on comprend d'un coup d'œil pourquoi Evmos est mieux classée qu'Ethereum Classic dans le classement d'engouement : Evmos a un meilleur score qu'Ehtereum Classic dans les quatre catégories.

L'utilisateur sait maintenant quelles cryptomonnaies connaissent un fort engouement et quelles catégories permettent d'expliquer ce fort engouement. Mais l'utilisateur pourrait également vouloir savoir plus précisément quelle métrique a influencé telle ou telle catégorie (par exemple, est-ce l'évolution du nombre d'abonnés Twitter sur 48h ou le score Google Trends qui a le plus influencé le score de la catégorie 'Public').

B. Visualisation de l'évolution de chaque métrique

Le score global est calculé grâce à la moyenne pondérée des scores de quatre catégories, elles-mêmes composées de 15 métriques. Chacune des 15 métriques sera représentée via un graphique courbe (line plot) avec en abscisse, la date, et en ordonnée, l'évolution de la métrique en question (en %). Le choix a été fait de représenter le pourcentage d'évolution de la métrique et non la valeur absolue pour faciliter la comparaison entre cryptomonnaies. En effet, il existe de telles différences de valeur absolue des métriques entre chaque crypto qu'on aurait souvent un problème d'échelle. Par exemple, la différence d'abonnés Twitter entre Bitcoin et Evmos est trop importante pour pouvoir comparer leur évolution sur la même échelle : +5 millions d'abonnés pour Bitcoin versus 107K pour Evmos. On aurait

du mal à comparer les évolutions correctement entre les deux cryptomonnaie. Visualiser les % d'évolution résout ce problème (Figure 10).

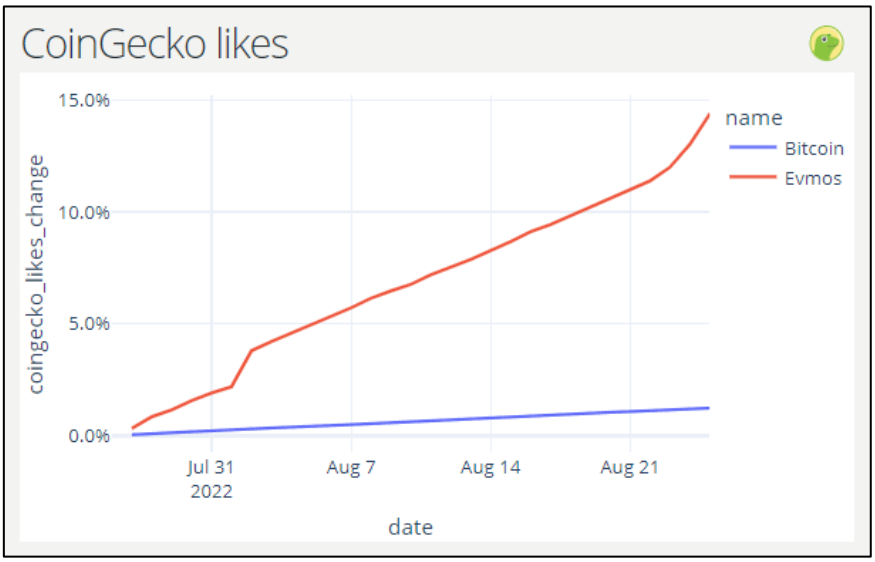

Figure 10. Visualisation de la métrique 'Mentions j'aime CoinGecko' pour les cryptomonnaies Bitcoin et Evmos. Le % d'évolution de cette métrique est représenté en ordonnée ce qui facilite la comparaison entre les deux cryptomonnaies (versus représenter des valeurs absolues).

La Figure 10 montre un exemple de visualisation de la métrique 'Mentions j'aime CoinGecko'. La comparaison des cryptomonnaies Bitcoin et Evmos est choisie pour l'exemple. On retrouve la date en abscisse et le % d'évolution de la métrique en ordonnée. Le % d'évolution étant une mesure normalisée de l'évolution du nombre de mentions j'aime Coingecko, on peut facilement comparer les deux cryptomonnaies (versus comparer l'évolution en termes de valeurs absolues) : on voit qu'Evmos a gagné 15% d'abonnés Twitter contre moins de 2% pour Bitcoin. De cette manière, on peut comparer très facilement les cryptomonnaies entre elles selon la métrique d'intérêt.

La Figure 11 montre un autre exemple de comparaison de trois cryptomonnaies pour la métrique 'Nombre d'abonnés Twitter'. Les cryptomonnaies Bitcoin, Ethereum et Evmos sont comparées entre elles et on peut facilement voir qu'Ethereum surperforme les deux autres en termes de % d'évolution du nombre d'abonnés Twitter.

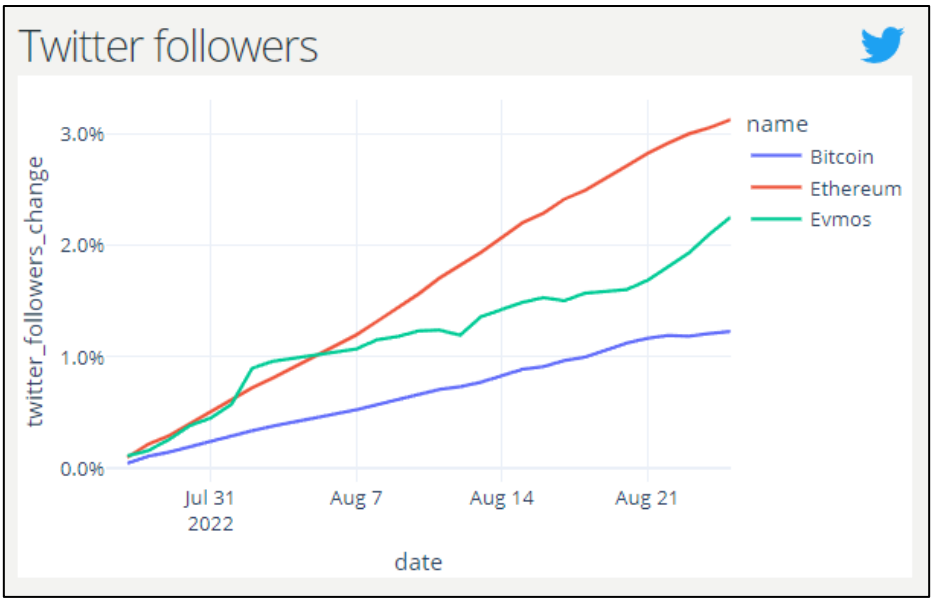

Figure 11. Visualisation de la métrique 'Nombre d'abonnés Telegram' pour les cryptomonnaies Bitcoin, Ethereum et Evmos. Cet exemple illustre l'intérêt de présenter un % dévolution plutôt qu'un nombre absolu pour la métrique donnée : la comparaison entre les différentes cryptomonnaies est plus aisée.

Le classement des cryptomonnaies, le diagramme de Kiviat et différents graphiques en courbe étant préparés, il suffit maintenant de les assembler dans une application web. Ces graphs seront également interactifs (pour pouvoir choisir les cryptomonnaies à comparer, passer sa souris sur les graphiques pour voir les valeurs numériques etc.) pour améliorer l'expérience utilisateur.

LA CRÉATION D'UN PRODUIT INTERNE : UN APPLICATION WEB INTERACTIVE

L'application créée est un tableau de bord interactif (Figure 12), déployé en ligne sur un serveur et accessibles par les collaborateurs de CoinShares via un VPN. Étant donné le nombre conséquent de graphiques à afficher, les différents graphiques sont répartis en cinq onglets :

1. Onglet 'Global' : présente le classement des cryptomonnaies sous forme de tableau ainsi que le diagramme de Kiviat permettant de comparer plusieurs cryptomonnaies selon les quatre catégories
2. Onglet 'Community' : présente les quatre graphiques courbe représentant les quatre métriques de la catégorie 'Community', à savoir, 'Nombre d'abonnés Twitter', 'Nombre d'abonnés Telegram, 'Nombre d'abonnés Discord' et 'Nombre d'abonnés subreddit'
3. Onglet 'Public' : présente les quatre graphiques courbe représentant les quatre métriques de la catégorie 'Public', à savoir, 'Score Google Trends score', 'Nombre de Tweets (48sh)', 'Nombre de mentions j'aime CoinGecko' et 'Nombre de watchlists CoinMarketCap'
4. Onglet 'Market' : présente les quatre graphiques courbe représentant les quatre métriques de la catégorie 'Market', à savoir, 'Prix', 'Volume', 'Capitalisation' et 'Volume/Capitalisation'.

5. Onglet ‘Developer’ : présente les quatre graphiques courbe représentant les quatre métriques de la catégorie ‘Developer’, à savoir, ‘Nombre de Watch’, ‘Nombre de Fork‘, ‘Nombre de Star’.

Ainsi chaque onglet ne présente que l’information d’une seule catégorie et évite de surcharger l’écran de graphiques.

Un bouton qui permet de choisir la période sur laquelle l’engouement est calculé a également été ajouté ; l’utilisateur peut ainsi choisir parmi les options suivantes : deux semaines (‘2W’), un mois (‘1M’), trois mois ('3M'), six mois (‘6M’) et un an (‘1Y’). Cela permet à l’application d’être plus flexible : possibilité à l’utilisateur de choisir la période sur laquelle l’engouement est calculé (court, moyen ou long terme).

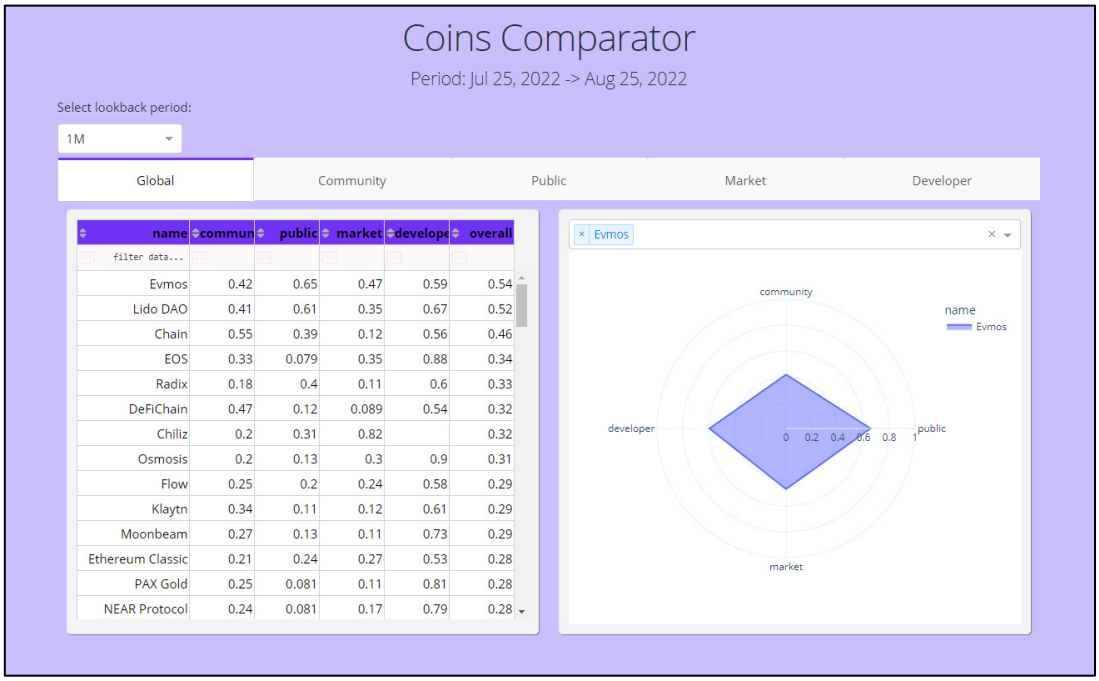

Figure 12. Landing page de l'application web. Une fois lancée, l’application affiche en premier lieu l’onglet ‘Global’ qui montre le classement des cryptomonnaies selon leur score d’engouement ainsi que le diagramme de Kiviat à droite. On peut faire défiler le tableau afin de voir le bas du classement ; le diagramme de Kiviat est interactif (le score s’affiche lorsque l’on passe la souris sur un angle). La période d’évolution est d’un mois par défaut (bouton en haut à gauche). Chacun des quatre autres onglets est accessible via un simple clic.

Si l’utilisateur veut comparer deux cryptomonnaies, il peut facilement le faire via le menu de sélection au-dessus de diagramme Kiviat (Figure 13).

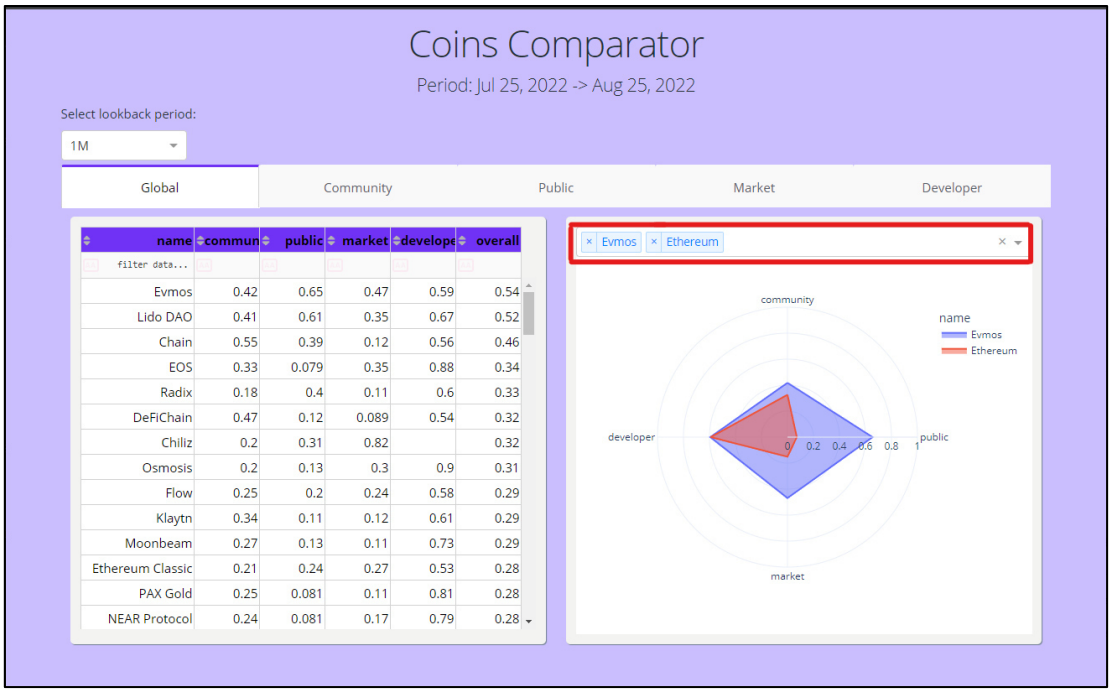

Figure 13. Sélection de deux cryptomonnaies sur le diagramme de Kiviat (à droite). L'utilisateur peut comparer autant de cryptomonnaies qu'il le souhaite en les sélectionnant via l'onglet au-dessus du graphique (encadré rouge).

Une fois que les cryptomonnaies d'intérêt ont été comparées grâce au diagramme Kiviat, l'utilisateur peut maintenant affiner sa compréhension de l'engouement de ces dernières en explorant les évolutions de chaque métrique à travers le temps. La Figure 13 est la vue de l'utilisateur lorsqu'il clique sur l'onglet 'Community ', sélectionne les cryptomonnaies Bitcoin et Evmos et choisit '2W' (soit deux semaines) comme période d'évolution.

Figure 14. Affichage de l'onglet 'Community' et de ses quatre graphiques courbe. L'onglet 'Community' a été sélectionné par l'utilisateur et les cryptomonnaies Bitcoin et Evmos sont comparées sur deux semaines. On observe qu'Evmos surperforme Bitcoin sur 3 métriques 'Nombre d'abonnés Twitter', 'Nombre d'abonnés Telegram' et 'Nombre d'abonnés subreddit'. On remarquera que Bitcoin n'a pas de données Discord car il n'existe pas de serveur Discord officiel Bitcoin.

On observe que sur les deux dernières semaines, Evmos surperforme Bitcoin sur toutes les métriques composant la catégorie 'Community'. On voit par exemple que son nombre d'abonnés subreddit a augmenté de près de 8% contre moins d'1% pour le subreddit Bitcoin. Le lecteur notera qu'il n'y a pas de données Discord disponibles pour Bitcoin, ce dernier ne possédant pas de serveur officiel. Ainsi l'utilisateur a une compréhension bien plus fine de pourquoi telle ou telle cryptomonnaie connaît un fort engouement et surtout, il peut comparer plusieurs cryptomonnaies entre elles.

AMÉLIORATION D'UN PRODUIT EXISTANT : AJOUT DE CRYPTOMONNAIES AU CATALOGUE NAPBOTS

Pour rappel Napbots est l'un des produits phare de CoinShares acquis lors de l'acquisition de Napoleon Group en décembre 2021. Napbots est un service de trading automatisé qui offre aux particuliers la possibilité de jouir de stratégies de trading développées par des professionnels. Le fonctionnement est le suivant : le client connecte sa plateforme de trading préférée (Binance, FTX etc.) au service, il choisit sa stratégie de trading parmi une sélection (Wise : ETH, Wise : BTC, Pulse ETH etc.), et enfin surveille les performances du bot via un tableau de bord sur le site web. Aujourd'hui seules neuf cryptomonnaies sont listées au catalogue. Il s'agit des cryptomonnaies ayant les plus grosses capitalisations : Bitcoin, Ethereum, Binance Coin, Ripple, Cardano, Solana, Doge Coin, et Polygon.

Actuellement, les clients n'ont donc pas d'autre choix que de sélectionner une stratégie contenant l'une de ces cryptomonnaies. Or, même si les cryptomonnaies listées sur le catalogue ont l'avantage d'être les plus grosses cryptomonnaies en termes de capitalisation, il n'est pas certain que celles-ci soient les neufs cryptomonnaies les plus en vogue ces derniers mois. Et même si c'était le cas, il n'est pas certain que ce soit le cas les prochains mois. L'idée serait alors d'utiliser l'application web pour détecter des cryptomonnaies connaissant un fort engouement et décider de les ajouter sur Napbots.

D'ailleurs aujourd'hui, d'après l'application, aucune des neufs cryptomonnaies présentement listées sur Napbots n'apparaissent dans le top du classement. Ethereum est la mieux positionnée parmi les neufs cryptomonnaies listées sur Napbots mais se place seulement 19/106 ; Polygon la deuxième mieux positionnée se place seulement 27/106. Au contraire, Evmos, Lido DAO ou encore Chain (qui ne sont pas listées sur Napbots) sont dans le top 3 du classement (Figure 12). Il serait donc intéressant pour CoinShares de réfléchir à lister l'une d'entre elles sur le catalogue de Napbots afin de répondre à l'engouement autour d'elles (et donc de répondre à une potentielle demande des clients/prospects). Aussi, le fait de pouvoir facilement changer les poids de chaque métrique dans le calcul du score d'engouement permettra d'accorder plus d'importance à telle ou telle métrique. Si, par exemple, CoinShares décide qu'un nombre d'abonnés Twitter est la métrique la plus importante pour détecter un engouement autour d'une cryptomonnaie, il sera facile de le faire grâce au design de l'application.

LIMITES ET AMÉLIORATIONS FUTURES DE L'APPLICATION

I. LIMITES DE L'APPLICATION

Il ne sera pas aussi trivial pour CoinShares de choisir la première cryptomonnaie du classement et de l'ajouter au catalogue.

En effet, certaines cryptomonnaies listées dans le classement ne sont pas assez liquides c'est à dire qu'elles ne jouissent pas d'un assez grand volume d'échange (il y a peu d'acheteurs et de vendeurs dans le marché). Evmos par exemple, a un volume d'échange d'environ 2 millions de dollars ces dernières 24h (contre 24 milliards pour Bitcoin ou 16 milliards pour Ethereum). Cela est un problème car Napbots passe des ordres au marché pour un nombre conséquent de clients au même moment. Or, passer le même ordre au marché plusieurs fois à un même moment sur un marché non liquide peut poser des problèmes de spread et de volatilité[11] (certains ordres pourraient ne pas être exécutés au prix escompté et/ou les ordres passés pourraient faire monter/descendre significativement le cours

[11] Plus d'information sur les marchés non liquides ici : https://www.investopedia.com/terms/i/illiquid.asp

de l'actif). CoinShares ne peut donc pas se permettre de lister une cryptomonnaie non liquide dans son catalogue.

Ensuite, il y a un problème de stabilité de l'engouement dans le temps. Une fois listée car temporairement populaire, la cryptomonnaie pourrait connaître un déclin de popularité quelques mois plus tard. Que faire dans ce cas ? Faut-il délister la cryptomonnaie au risque de décevoir des clients ayant choisi d'investir dans cette dernière ou bien faut-il garder cette cryptomonnaie au catalogue ce qui peut demander des ressources supplémentaires. Car plus de cryptomonnaies listées signifie plus de développement et de maintenance pour le service informatique. C'est donc un enjeu crucial pour CoinShares de choisir méticuleusement les cryptos à lister dans son catalogue.

Une autre limite de l'application est qu'il ne sera pas possible d'ajouter et de suivre toutes les cryptomonnaies possibles. Et ce pour deux raisons :

Premièrement, de par la méthode de calcul de l'engouement, afin d'obtenir un score d'engouement robuste et fiable, il est nécessaire que cette dernière possède le maximum de métriques (notamment les métriques des catégories 'Community' ou 'Public'). Certaines cryptos actuellement suivies ne possèdent pas de Github ce qui empêche le calcul de la catégorie 'Developper' et diminue ainsi la robustesse du score d'engouement de la crypto et diminue aussi la pertinence de la comparaison avec d'autres cryptos. Comme la catégorie 'Developer' n'a pas le plus gros poids dans le calcul du score d'engouement, ce n'est pas un problème majeur. Le problème serait plus gênant si c'était la catégorie 'Community' ou 'Public' qui était impactée : si la cryptomonnaie ne possédait par exemple qu'un compte Twitter et pas de Telegram, de Discord ni de subreddit.

Deuxièmement, les appels API quotidiens permettant de récupérer les données sont soumis à des 'rate limit' : il y a une limite au nombre de requêtes que l'on peut effectuer dans une période donnée. L'API de Twitter par exemple a un 'rate limit' de 900 appels toutes les 15 minutes. Ce qui signifie dans notre cas, que l'on peut récupérer le nombre d'abonnés de 900 comptes Twitter maximum en 15 minutes. Si un jour l'application monitorait 1000 cryptos, cela poserait problème car le nombre d'abonnés de tous les comptes ne serait pas récupéré au même moment. L'application récupérant les données à 00:00:00 UTC, le nombre d'abonnés des 900 premiers comptes serait récupéré à la bonne heure tandis que les derniers comptes seraient récupérés en retard de plusieurs minutes. Cela poserait un problème d'exactitude des données. Le nombre d'abonnés Twitter (ou Discord ou Telegram) ne variant pas énormément en quelques minutes, le problème est cependant à relativiser mais à garder à l'esprit pour la personne qui maintiendra l'application.

II. POSSIBLES AMÉLIORATIONS DE L'APPLICATION

Pour résoudre le problème de liquidité évoqué plus haut, il pourrait être pertinent d'ajouter la possibilité de filtrer les cryptomonnaies selon leur volume d'échange. Dans le classement de l'onglet 'Global', on pourrait ajouter une colonne 'trading volume' et associer à chaque crypto son volume d'échange sur 24h. Ainsi l'utilisateur (CoinShares) pourrait facilement trier les cryptomonnaies qu'il estime ne pas être assez liquides. Une autre possibilité plus radicale serait de monitorer seulement les cryptomonnaies les plus liquides. CoinShares fixerait un seuil de volume d'échange minimum à atteindre pour suivre une cryptomonnaie. Cela aurait pour avantage d'éliminer toutes les cryptomonnaies qui, de toute façon, ne sont pas assez liquide pour être implémentées dans Napbots. Par contre, cela aurait le désavantage d'arrêter la collecte de données de ces cryptomonnaies. Or, une petite cryptomonnaie ne restera pas petite ad vitam aeternam ; il est donc primordial de pouvoir monitorer les petites cryptos et de pouvoir détecter de potentielles cryptos en plein essor (et donc ce, avant qu'elles deviennent de grosses cryptomonnaies). De plus, avoir l'historique de données comme le nombre d'abonnés Twitter ou le nombre de Tweets pourrait se révéler très utile pour une autre application dans le futur. CoinShares devrait donc simplement ajouter un filtre 'volume d'échange' dans le tableau classant les cryptomonnaies pour résoudre le problème de liquidité.

Actuellement, il y a un nombre fixe de cryptomonnaies (une centaine) qui sont suivies dans l'application et dont les données sont récoltées quotidiennement. Il pourrait être intéressant de savoir quand des cryptomonnaies se hissent dans le haut du classement global des cryptomonnaies ou encore quand de nouvelles cryptos sortent et deviennent disponibles sur le marché. Par exemple Audius – une plateforme musicale de streaming décentralisée – est aujourd'hui seulement 126^{e} des cryptomonnaies en termes de capitalisation[12] et n'est pas monitorée par l'application. Or, il n'est pas impossible qu'Audius grossisse et se hisse bientôt dans le top 100 des cryptomonnaies les plus capitalisées. Il serait fort dommageable de ne pas suivre l'engouement que connaîtrait cette cryptomonnaie. Ou encore, StarkWare, un projet qui connait déjà un fort engouement dans le monde de la finance décentralisée alors même que l'équipe n'a pas encore déployé sa cryptomonnaie. Pour pallier à ces problèmes, une amélioration possible serait de notifier à l'utilisateur (soit par email soit dans directement dans l'application) des « petites » cryptomonnaies qui montent dans le classement global ou des nouvelles cryptomonnaies qui viennent d'être listées sur les plateformes d'échange afin que l'utilisateur puisse ajouter ces dernières à l'application.

Les poids associés à chaque métrique et à chaque catégorie dans le calcul du score d'engouement ont été déterminé de manière arbitraire, il sera toujours temps d'améliorer le calcul de ce score en

[12] D'après CoinMarketCap, le 25 août 2022 : https://coinmarketcap.com/currencies/audius/

modifiant les poids. En effet, on ne pouvait pas avoir assez de recul sur la potentielle influence de telle ou telle métrique dans l'engouement autour d'une cryptomonnaie lors de la conception de l'application. CoinShares, avec le temps, devrait pouvoir améliorer le calcul du score de l'engouement. D'ici quelques mois, peut-être qu'une cryptomonnaie encore méconnue aujourd'hui connaîtra un engouement fort et fera l'unanimité dans le monde de la crypto. L'application ayant potentiellement monitoré la crypto depuis ce temps, CoinShares pourra regarder quels indicateurs ont donné le plus d'indices avant-coureurs sur l'engouement et l'explosion de cette crypto. Un engouement est-il mieux détecté par une explosion du nombre d'abonnés Twitter ou bien par une augmentation soutenue de mentions j'aime CoinGecko ou encore par une augmentation soudaine de la métrique 'volume / market cap'. Une fois que CoinShares saura quels indicateurs ont le plus d'importance dans la détection d'engouement, elle pourra assigner des poids plus forts sur tel ou tel indicateur. On notera encore l'importance d'ajouter le plus possible de cryptomonnaies à l'application afin d'avoir le plus de chances d'avoir des données historiques sur des cryptos qui connaîtront un essor futur.

Il pourrait être pertinent d'ajouter à la catégorie 'Public' une métrique sur l'actualité des cryptos dans la presse. Pour rappel, actuellement cette catégorie contient les métriques 'score Google Trends', 'nombre de Tweets', 'nombre de mentions j'aime CoinGecko' et 'nombre d'ajouts à une watchlist CoinMarketCap' ; ajouter une dimension d'actualité permettrait d'étoffer cette catégorie et de rendre son score plus robuste. Une idée serait de récupérer le nombre d'articles quotidien citant le nom de la cryptomonnaie en question. Pour aller plus loin, on pourrait également faire de l'analyse de sentiments sur les titres des articles de presse citant la cryptomonnaie afin de déterminer le climat émotionnel autour de ladite crypto. L'application devait d'ailleurs initialement récupérer ces données via l'API de newsapi.org mais celle-ci a un 'rate limit' de 100 requêtes par 24h (50 requêtes toutes les 12 heures) pour son plan gratuit. Il aurait fallu débourser 449$ par mois pour passer au plan 'Business' et avoir accès à un meilleur rate limit, l'idée a donc été abandonnée. C'est l'une des limites de l'utilisation d'open data venant d'entreprises privées : on a accès à la donnée selon les conditions (plus ou moins contraignantes) fixées par l'entreprise fournisseuse.

PARTIE III. ANALYSES ET RECOMMANDATIONS

CRITIQUES DU MODÈLE DE LA CHAÎNE DE VALEUR DE L'OPEN DATA (EUROPEAN COMMISSION, 2015)

Cette application est donc un cas d'usage de création de valeur à partir d'open data au sein d'une entreprise dans les cryptomonnaies. Lors du développement de cette application, j'ai toutefois

observé des divergences entre le modèle théorique de la Commission Européenne (European Commission, 2015) présenté Partie I et la réalité du terrain.

Premièrement le modèle stipule que la création de l'open data est effectuée par le secteur public. Or, même si cela est en partie vrai, nous avons vu lors du cas d'usage que le secteur privé est également un fournisseur essentiel d'open data. Toutes les données collectées pour la création de l'application web sont issues de sources privées. Dans notre cas, seules les sources privées fournissaient des données pertinentes (cf. qualité des données) pour les besoins de l'application. En revanche, il serait faux de penser que le secteur public ne fournit pas de données pertinentes et exploitables, seulement, cela va varier selon le cas d'usage. En effet, par définition le secteur public fournit des données publiques. L'INSEE, par exemple, va pouvoir fournir seulement des données publiques qu'elle collecte comme la liste des personnes décédées, des données d'activité selon l'âge et le sexe, ou encore les prénoms donnés à la naissance. Ces données peuvent avoir un intérêt dans d'autres contextes mais elles nous seront inutiles dans notre cas. Et après quelques recherches, aucun jeu de données intéressant pour notre cas n'était fourni par le secteur public. C'est pourquoi, je me suis tourné vers le secteur privé.

En lien avec ce premier point, le modèle explique que le secteur privé est en début de chaîne de création de valeur tandis que le secteur privé arrive en fin de chaîne. Ici encore ce point est à nuancer car les deux acteurs privés et publics peuvent intervenir à n'importe quel moment de la chaîne. Par exemple, l'administration française a mis en place une plateforme visant à mettre en valeur l'activité des députés de l'Assemblée Nationale : nosdeputes.fr. Le gouvernement exploite donc ses propres données publiques (la liste des parlementaires et des indicateurs de l'activité parlementaires) pour mettre à disposition aux citoyens français un site web qui leur donne les outils pour « comprendre et analyser le travail de leurs représentants », comme le précise le site. Le début de la chaîne de valeur de l'open data n'est donc pas réservé exclusivement au secteur public tout comme la fin de cette chaîne n'est pas exclusivement réservée au secteur privé. Pour aller plus loin, Deloitte (2012) indique même qu'il est possible de considérer un troisième acteur : les particuliers . En effet, un particulier peut très bien, à condition qu'il possède les compétences nécessaires, exploiter un jeu de données ouvertes à des fins purement personnelles. Il créerait donc aussi de la valeur avec l'open data.

Ensuite, la deuxième partie du modèle indique que la création de valeur arrive en fin de chaîne (donnée > information > connaissance > services > valeur ajoutée). Ce serait qu'une fois que la donnée a été créée, validée, agrégée, analysée et utilisée pour créer un service/produit que de la valeur serait créée. Or, avec l'expérience acquise durant le développement du cas d'usage, je peux témoigner que de la valeur est créée à chaque étape du processus : lorsque l'on valide la donnée avant de l'agréger, on rend plus exploitable la donnée pour dans le futur. Lorsqu'on agrège la donnée, on crée des jeux de

données plus riches qui seront également plus exploitables par la suite et donneront plus d'informations. Enfin, le traitement et l'analyse de données est, par expérience, l'étape la plus créatrice de valeur. C'est lors de cette étape que l'on va traiter la donnée, la transformer, calculer des indicateurs, trouver des corrélations, visualiser les données etc., et c'est à ce moment que la donnée est mise à rude épreuve. Si les données collectées possèdent peu de valeur intrinsèque (à l'instar d'un pétrole ayant une faible densité), l'analyse de données permettra de voir s'il sera possible de créer quelconque valeur à partir d'elles. Si au contraire, les données ont un bon potentiel et renferme des informations pertinentes, c'est lors de l'analyse de données que leur plein potentiel pourra être exploité. Ce sera tout le rôle du data analyst de trouver les meilleurs types d'analyses afin de créer le maximum de valeur à partir de ces données. Le service/produit créé ensuite ajoutera une couche supplémentaire de valeur mais sans cette étape d'analyse de donnée, le produit ne pourra ajouter que peu de plus-value.

Ainsi, au vu des divergences entre le modèle initial et la réalité du terrain, je propose une version adaptée du modèle qui prend en compte les différentes critiques évoquées à l'instant (Figure 15).

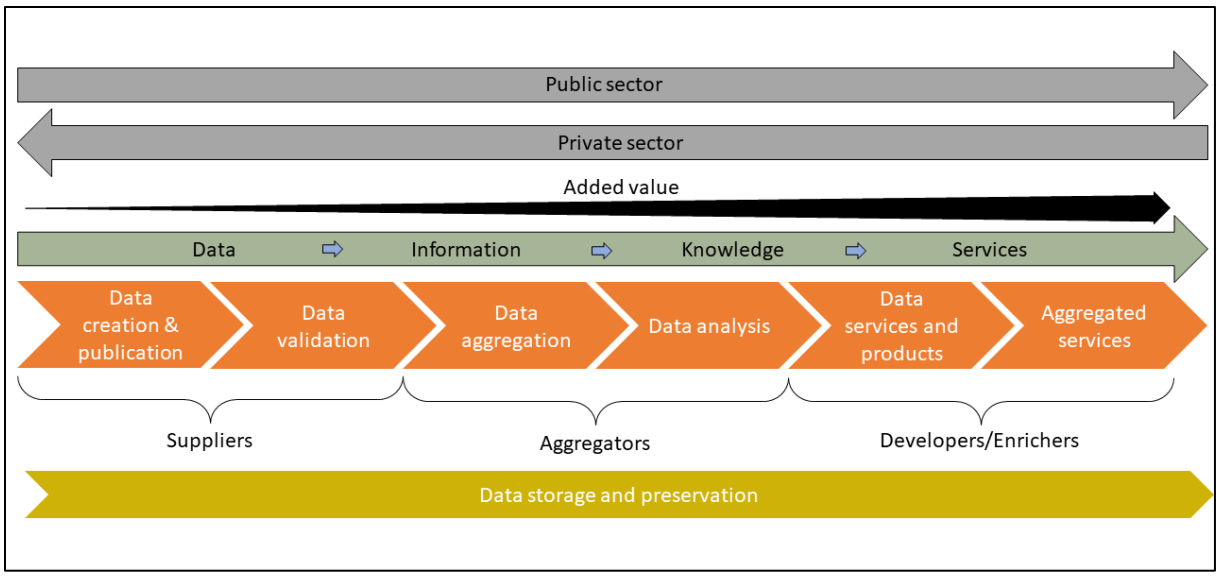

Figure 15. Modèle de la chaîne de valeur de l'open data adapté. Ce modèle théorique reprend le modèle de la Commission Européenne **(European Commission, 2015)** en l'adaptant selon l'expérience vécue lors du développement de l'application web. D'après ce modèle adapté, le secteur privé comme le secteur public créent et mettent à disposition des données ouvertes ; le secteur public peut lui aussi exploiter de l'open data pour créer de la valeur, cette tâche n'est pas réservée qu'au secteur privée. Aussi, la valeur est créée au cours de toutes les étapes (et non pas qu'en fin de processus) et augmente significativement à partir de l'étape d'analyse des données, étape clé dans le processus.

Ce modèle prend donc en compte les critiques faites précédemment au modèle initial et le rend plus cohérent avec l'expérience vécue lors du développement de l'application web. Premièrement, le modèle adapté indique que le secteur privé comme le secteur public peut jouer le rôle de fournisseur

en créant et publiant des données ouvertes (les deux flèches grise en haut). Ensuite, la création de services/produits grâce à l'open data n'est pas réservée aux acteurs privés, les institutions publiques peuvent également exploiter des données publiques pour créer leur propre service/produit (ex. du gouvernement Français et son site nosdeputes.fr). Enfin, de la valeur est créée à chaque étape de la chaîne, cette valeur croissant tout au long du processus. Chaque étape ajoute un peu plus de valeur que l'étape précédente. L'étape d'analyse de données est l'étape clé du processus : c'est lors de cette étape que le data analyst exploitera toute la valeur que renferme la donnée en trouvant les bons types d'analyses, calculs d'indicateurs, tests statistiques ou encore visualisations à effectuer. L'étape de création de services/produits basés sur ces analyses ajoute encore une couche de valeur aux analyses faites à l'étape d'avant. Optionnellement l'entreprise exploitant la donnée utilisera ce service/produit pour améliorer un ou plusieurs produits existants en les agrégeant entre eux.

RECOMMANDATIONS FAITES À COINSHARES CONCERNANT L'ATTITUDE À AVOIR VIS-À-VIS DE L'OPEN DATA

Comme le développement de cette application web l'a montré, CoinShares peut tirer beaucoup de valeur de l'exploitation de l'open data. Je conseille à l'entreprise d'intégrer plus fréquemment des données ouvertes dans ses produits ou dans sa création de connaissance de l'écosystème crypto. Elle peut commencer par utiliser les données récoltées mais non utilisées (pour l'instant) par l'application web (nombre de commentaires quotidiens sur le subreddit d'une crypto, le nombre d'open issues Github, le taux de votes positifs/négatifs des utilisateurs CoinGecko etc.). Aussi, je conseille à l'entreprise de faire de la veille sur les sources (publiques ou privées) existantes qui pourraient fournir des données (ouvertes ou non) intéressantes pour elle. Elle peut commencer par monitorer les APIs déjà utilisées dans l'application afin d'être alertée lorsque de nouvelles métriques seront ajoutées par les fournisseurs. Elle peut également chercher de nouvelles sources ou rester au fait de nouvelles APIs qui pourraient être publiées. Les nouvelles données pourront être intégrées à l'application pour étoffer le calcul du score d'engouement ou bien réutilisées dans d'autres contextes (comme une autre application) ou même donner des idées de nouveaux services/produits.

Afin de pouvoir exploiter ces données (ouvertes), CoinShares aura besoin de ressources (humaines) compétentes dans ce domaine : des data analysts/scientists. Ces professionnels de la donnée auront les compétences pour comprendre le besoin, collecter, analyser et construire des produits ou des insights basés sur la donnée. Il va sans dire que les data analysts/scientists pourront également être utilisés pour exploiter et valoriser les données internes à l'entreprise. Les entreprises produisant de plus en plus de données, il devient indispensable pour elles d'exploiter cette ressource et de ne pas passer à côté de la valeur qu'elles peuvent en tirer. D'ailleurs le métier de data scientist arrive dans le

top 3 des meilleurs métiers (en termes de salaire, satisfaction et positions ouvertes) selon le classement Glassdoor de 2022[13], et ce, depuis cinq années consécutives, ce qui montre l'attrait et l'intérêt pour ce métier d'avenir.

En plus de pouvoir exploiter techniquement les données ouvertes ou internes à une entreprise, ces professionnels de la donnée transmettront la culture data au sein de l'organisation. Cela pourra inciter les collaborateurs à penser à utiliser plus souvent la donnée pour résoudre des problèmes ou même pour gagner en insight. L'application web développée durant ce travail est d'ailleurs une extension d'un travail qui était fait auparavant, manuellement, chaque semaine. Chaque semaine, un stagiaire récoltait des métriques de popularité/engouement d'une quinzaine de cryptomonnaies, calculait un score de popularité/engouement et pouvait ensuite classer les cryptos. Chaque semaine, le classement était comparé avec celui de la semaine précédente afin de voir si des cryptomonnaies gagnaient en popularité. L'application web développée ici apporte donc une amélioration significative du travail précédent. Elle ajoute en effet l'automatisation de la récolte des données et la fréquence de la récolte (quotidienne) ; elle améliore aussi le nombre des métriques récoltées (le nombre de Tweets quotidien ne peut pas être calculé manuellement par exemple), le nombre de cryptos suivies (plus d'une centaine) et permet le partage de l'information à tous les collaborateurs via un tableau de bord interactif rendant facile et intuitive la compréhension des données.

CONCLUSION

Les données ouvertes sont donc une ressource primordiale pour les entreprises aujourd'hui. Si exploitées correctement par des professionnels de la donnée, elles peuvent permettre de créer beaucoup de valeur pour une entreprise. À travers le cas d'usage décrit dans ce travail, on a vu comment les données ouvertes liées à des cryptomonnaies pouvait créer de la valeur pour CoinShares en, d'une part, donnant des insights au personnel sur l'engouement autour des cryptomonnaies et, d'autre part, en permettant d'améliorer leur produit phare existant : Napbots. Dans le futur, il sera important pour CoinShares de multiplier les cas d'usages de ce type pour exploiter une ressource qui n'a jamais été aussi abondante et indispensable qu'aujourd'hui.

[13] 50 Best Jobs in America for 2022: https://www.glassdoor.com/List/Best-Jobs-in-America-LST_KQ0,20.htm

BIBLIOGRAPHIE

Ackoff, R. L. (1989). From Data to Wisdom. *Journal of Applied Systems Analysis*, *16*, 3–9.

AMF. (2022). *Qu'est-ce qu'une « cryptomonnaie » ?* AMF. https://www.amf-france.org/fr/quest-ce-quune-cryptomonnaie

Arribas-Bel, D., Green, M., Rowe, F., & Singleton, A. (2021). Open data products-A framework for creating valuable analysis ready data. *Journal of Geographical Systems*, *23*(4), 497–514. https://doi.org/10.1007/s10109-021-00363-5

Buchholtz, S., Bukowski, M., & Sniegocki, A. (2014). *Big and open data in Europe. A growth engine or a missed opportunity?*

Deloitte. (2012). *Open data Driving growth, ingenuity and innovation.*

Deshpande, P. S., Sharma, S. C., & Peddoju, S. K. (2019). *Security and Data Storage Aspect in Cloud Computing*. Springer.

Dong, E., Du, H., & Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases*, *20*(5), 533–534. https://doi.org/10.1016/S1473-3099(20)30120-1

European Commission. Directorate General for the Information Society and Media., Capgemini Consulting., Intrasoft International., Fraunhofer Fokus., con.terra., Sogeti., Open Data Institute., Time.lex., & University of Southampton. (2015). *Creating value through open data: Study on the impact of re use of public data resources*. Publications Office. https://data.europa.eu/doi/10.2759/328101

eurostat. (2022). *Performance of the agricultural sector*. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Performance_of_the_agricultural_sector

Herzog, T. N., Scheuren, F., & Winkler, W. E. (2007). *Data quality and record linkage techniques*. Springer.

Kitchin, R. (2014). *The Data Revolution: Big Data, Open Data, Data Infrastructures & Their Consequences*. SAGE Publications Ltd. https://doi.org/10.4135/9781473909472

Perez, J., Emilsson, C., & Ubaldi, B. (2020). *OECD Open, Useful and Re-usable data (OURdata) Index: 2019* (No. 1). https://www.oecd.org/governance/digital-government/ourdata-index-policy-paper-2020.pdf

re3data.org. (2014). *DATA.GOV.UK*. 47.248 datasets. https://doi.org/10.17616/R3Q89G

Statista. (2022). *Number of crypto coins 2013-2022*. Statista. https://www.statista.com/statistics/863917/number-crypto-coins-tokens/

The Economist. (2017). The world's most valuable resource is no longer oil, but data. *The Economist*. https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data

The Open Definition. (2022). https://opendefinition.org/

TABLES DES FIGURES

ANNEXE 1 : DICTIONNAIRE DE DONNÉES DE LA TABLE 'COINS'

Attribut	Signification	Domaine
symbol	Symbole identifiant la cryptomonnaie	texte
cg_id	Identifiant de la cryptomonnaie de l'API CoinGecko	texte
cg_name	Nom donné à la cryptomonnaie par CoinGecko (ex. The Sandbox)	texte
cg_url	Nom donné à la cryptomonnaie dans l'URL CoinGecko ex. 'the-sandbox' pour https://www.coingecko.com/en/coins/the-sandbox	texte
cmc_url	Nom donné à la cryptomonnaie dans l'URL CoinMarketCap ex. 'the-sandbox' pour https://coinmarketcap.com/currencies/the-sandbox/	texte
discord	Identifiant d'invitation du canal discord (ex. 'sandboxgame' pour le lien d'invitation de The Sandbox https://discord.com/invite/sandboxgame)	texte
github	Chemin du dépôt Github du projet crypto (ex. 'thesandboxgame/sandbox-smart-contracts' pour The Sandbox https://github.com/thesandboxgame/sandbox-smart-contracts)	texte
subreddit	Nom du subreddit officiel de la crypto	texte
telegram	Identifiant du groupe Telegram dans le lien d'invitation (ex. 'sandboxgame' pour le lien d'invitation de The Sandbox https://t.me/ sandboxgame)	texte
twitter	Nom du compte twitter du projet crypto	texte

Annexe 1. Dictionnaire de données de la table 'coins'. Cette table recense toutes les informations qui permettent de récupérer les données d'une crypto via des appels API. Par exemple le chemin de dépôt Github sert à récupérer les Watch, Fork et Star du dépôt en question via l'API Github.