Gestion des Données Ouvertes dans L'Agriculture et la Nutrition

Ce cours en ligne est le fruit d'une collaboration entre les partenaires de GODAN Action, y compris Wageningen Environnemental Research (WUR), Agro Know, AidData, l'Organisation des Nations Unies pour l'Alimentation et l'Agriculture (FAO), le Forum Mondial sur la Recherche Agricole (GFAR), l'Institut des Etudes du Développement (IDS), le Land Portal, l'Open Data Institute (ODI) et le Centre Technique de Coopération Agricole et Rurale (CTA).

GODAN Action est un projet de trois ans du Département pour le Développement International du Royaume-Uni pour permettre aux utilisateurs, producteurs et intermédiaires de données de s'engager efficacement avec les données ouvertes et maximiser leur potentiel d'impact dans les secteurs de l'agriculture et de l'alimentation. Nous travaillons en particulier à renforcer les capacités, à promouvoir des normes communes et les meilleures pratiques et à améliorer la manière dont nous mesurons l'impact. [www.godan.info]

MODULE 2: UTILISATION DES DONNÉES OUVERTES

Leçon 2.1 A La découverte des données ouvertes

Photo par Neil Palmer (CIAT) Sous licence CC BY -SA 2.0

Objectifs et résultats d'apprentissage

Cette leçon vise à é une base pour découvrir et accéder aux données disponibles sur le Web. Des téléchargements de données et des fournisseurs de services de données aux éditeurs de données ouvertes liées, cette leçon couvrira l'ensemble des outils nécessaires pour trouver plus rapidement les données recherchées sur le Web.

A la fin de cette leçon, vous devrez être en mesure :

- *De répertorier les différents types de services qui permettent d'accéder aux données ouvertes*
- *De répertorier différentes méthodes permettant d'accéder aux données à partir de ces services.*
- *D'expliquer la différence entre ces types de services*
- *D'utiliser les différents types de services pour accéder aux données ouvertes*
- *De découvrir des données téléchargeables*
- *De découvrir des données cachées*
- *De déterminer si une source de données est ouverte pour ces types de services*
- *De décrire les avantages et les inconvénients de ces types de services.*

Sommaire

Illustrations

Liste des tableaux

1. Introduction

Au fur et à mesure que le Web évolue, il continue d'être utilisé pour partager des ressources et des données de plus en plus complexes, mais il remet en question les paradigmes existants. Le World Wide Web constituait un élément central des données d'information et est considéré comme un espace d'information donnant accès à des documents et à d'autres ressources du Web (World Wide Web).

Une telle vision du Web a conduit à la création d'un réseau de documents, ou pages Web, plus connus, conçus pour permettre aux humains de lire et d'accéder à environ 4,75 milliards d'entre eux [http://www.worldwidewebsize.com]. De nombreux documents sont liés entre eux et ces connexions ajoutent de la valeur (lien hypertexte). Dans un blogue, un article de journal ou un article universitaire, nous pouvons utiliser de tels liens pour nous appuyer sur des discussions antérieures ou pointer vers des sources factuelles. Ils nous aident à explorer le réseau de documents.

2. Les générations du web.[1]

Il y avait beaucoup moins de documents dans les premiers jours du Web, mais les gens devaient quand même être en mesure de s'y retrouver. Les premiers efforts pour maintenir manuellement un index[2] ont été faits par Sir Tim Berners-Lee, Président et Co-Fondateur de l'ODI[3]. Les gens pouvaient alors consulter la liste et ensuite sauter aux pages qui leur semblaient intéressantes.

Nous avons ensuite créé des portails, tels que DMOZ[4] et early Yahoo![5] Il s'agissait de listes de sites Web et de pages organisées par thèmes particuliers. Avec l'expansion du Web, les portails n'étaient plus viables et les gens sont passés aux moteurs de recherche de métadonnées, tels que Altavista[6] et Lycos[7], qui utilisaient des métadonnées définies manuellement dans la page Web et fournissant des informations sur le document.

La recherche était plus évolutive car les pages étaient découvertes automatiquement, mais les résultats étaient peu fiables et facilement manipulables. La génération suivante des engins de découverte Web est arrivée avec la recherche PageRank-style[8], comme Google, qui a utilisé

[1] https://theodi.org/blog/we-need-to-learn-how-to-search-the-web-of-data
[2] https://www.w3.org/History/19921103 hypertext/hypertext/DataSources/WWW/Servers.html
[3] https://theodi.org/team/timbl
[4] https://www.dmoz.org
[5] https://en.wikipedia.org/wiki/Yahoo!_Directory
[6] https://en.wikipedia.org/wiki/AltaVista
[7] https://en.wikipedia.org/wiki/Lycos
[8] https://en.wikipedia.org/wiki/PageRank

Beaucoup plus d'indices pour la recherche, y compris une compréhension du contenu, l'utilisation et la liaison. Cette troisième génération a permis de scruter directement le web pour découvrir à quel point chaque document serait pertinent pour les utilisateurs.

Il était essentiel que nous apprenions à construire ces différents types de recherche. Le web des documents ne pouvait pas évoluer jusqu'à ce que les engins de recherche[9] deviennent le principal moyen de découverte. Toutes ces méthodes existent toujours et répondent à des besoins différents, l'un étant de répondre aux exigences de mise en ligne des données sur le Web.

3. Les générations de données ouvertes sur le web

Les premières techniques de publication de données ouvertes reflètent les mêmes générations que celles du web de documents. Nous avons d'abord créé des portails, tels que data.gov.uk, opendata.go.tz et data.sncf.com. Ce sont des catalogues structurés de jeux de données organisés par thèmes ou structures organisationnelles particuliers.

À mesure que la quantité de données ouvertes commence à croître, des agrégateurs de données commencent à apparaître qui fournissent des services pour faciliter la découverte de données liées à des sujets ou des régions particuliers. Les exemples comprennent enigma.io[10], Européen data portal[11] et transport API[12]. Ces services reposent sur la disponibilité de métadonnées provenant d'autres portails, sites Web et prestataires de services pour fournir des informations et un accès aux données.

La troisième génération de moteurs de recherche de données est encore très axée sur le web étant un espace d'information, nous devons donc compter sur les mêmes moteurs de recherche pour trouver des données ainsi que des informations. Le développement des moteurs de recherche pour les données est encore à ses balbutiements et ceci est fortement lié aux méthodes utilisées pour publier des données sur (ou dans) le web.

[9] http://onlinelibrary.wiley.com/doi/10.1002/1097-4571(2000)9999:9999%3C::AID-ASI1607%3E3.0.CO%3B2-F/full
[10] https://www.enigma.com
[11] https://www.europeandataportal.eu
[12] https://www.transportapi.com

4. Les données sont juste une autre ressource sur le web

À mesure que le Web évoluait, il a commencé à devenir un lieu de partage de ressources multimédias. L'inclusion des images, de l'audio et de la vidéo a ouvert de nouveaux potentiels pour fournir des services tels que les services de streaming. L'audio était le pionnier ici, où les premiers sites comme last.fm utilisaient des métadonnées sur les pièces de musique pour créer des recommandations personnalisées pour les usagers. Cette technologie précède et constitue la base de nombreux systèmes conseils utilisés aujourd'hui. Last.fm a fourni des recommandations mais n'a pas permis aux gens d'écouter la musique elle-même. Cette fonctionnalité n'est apparue que trois ans plus tard en 2005 avec le lancement de Pandora, une application de radio personnalisée. Trois années plus tard, Spotify est lancé, premier service de streaming utilisant les technologies Web pour fournir une plate-forme audio dédiée en dehors du navigateur Web.

Toujours plus en avant, web et internet sont devenus une plate-forme d'accès à d'énormes quantités de ressources différentes. En s'insérant dans le navigateur Web, les moteurs de recherche comme Google ont ajouté une capacité de recherche multimédia pour trouver des images et des vidéos, tandis que des portails spécifiques comme YouTube offrent désormais un accès Web via leur site Web ainsi que via des applications et des téléviseurs connectés.

Cela n'est pas tout à fait de même quand il s'agit des données. Les moteurs de recherche n'ont toujours pas de recherche spécifique de données, peut-être qu'ils ne le feront jamais. Mais trouver des données sur le web demeure un défi, qui commence par la définition même des données.

Qu'est-ce qu'une donnée ?

Une image est une représentation visuelle de quelque chose. Un fichier audio émet un son lorsqu'il est joué. Une vidéo combine plusieurs images avec de l'audio pour créer une image animée. Les données sont ... difficiles à définir et demeurent toujours un objet de recherche difficile à saisir.

Les données représentent le niveau d'abstraction le plus bas à partir duquel l'information et les connaissances sont dérivées. Ce sont des termes abstraits et, par conséquent, les données peuvent être une image, une feuille de calcul ou un fichier audio. De plus, si le web est un espace d'information, alors les données sont quelque chose qui est d'un niveau inférieur ?

Traditionnellement, les données sont considérées comme une feuille de calcul ou un ensemble de chiffres qui peuvent être analysés de diverses façons. Sur le Web, ces données sont souvent partagées via des portails de données simplement sous la forme d'un fichier pouvant être téléchargé. Certains portails offrent des fonctionnalités semblables à YouTube où les données peuvent être explorées sans téléchargement, mais les données elles-mêmes sont toujours une ressource statique, chargées par quelqu'un, et prêtes à être téléchargées par quelqu'un d'autre.

Alors que les données restent une ressource statique, les services Web de deuxième et troisième génération sont parfaitement adaptés pour collecter des métadonnées sur ces ressources statiques et fournir des entrées dans les résultats de recherche. Les ressources statiques peuvent être directement liées et ainsi des algorithmes comme PageRank restent pertinents.

Cette approche convient à l'approche du Web, où les métadonnées constituent toujours la clé de la recherche des données. Cela signifie que les moteurs de recherche existants peuvent être utilisés pour trouver des données si vous pouvez faire la bonne requête.

Cependant, toutes les données ne sont pas statiques.

Lorsque vous visitez un site Web de magasinage, un site Web de voyage ou un site Web météorologique. Le contenu est susceptible d'être différent à chaque fois en fonction des données brutes. Sur un site d'achat, certains produits peuvent varier en termes de prix et de niveau de stock. Sur un site Web de voyage, les options varieront dans la disponibilité et le prix selon les critères de recherche pendant que la météo change constamment.

Les données qui alimentent ces sites sont vastes et cachées sur le Web. Les données lisibles par machine créent un nouveau web de données qui, selon les experts, a le potentiel de déclencher l'ère des données. Peut-être le web se transforme t'il alors d'un espace d'information en un espace de données et d'informations.

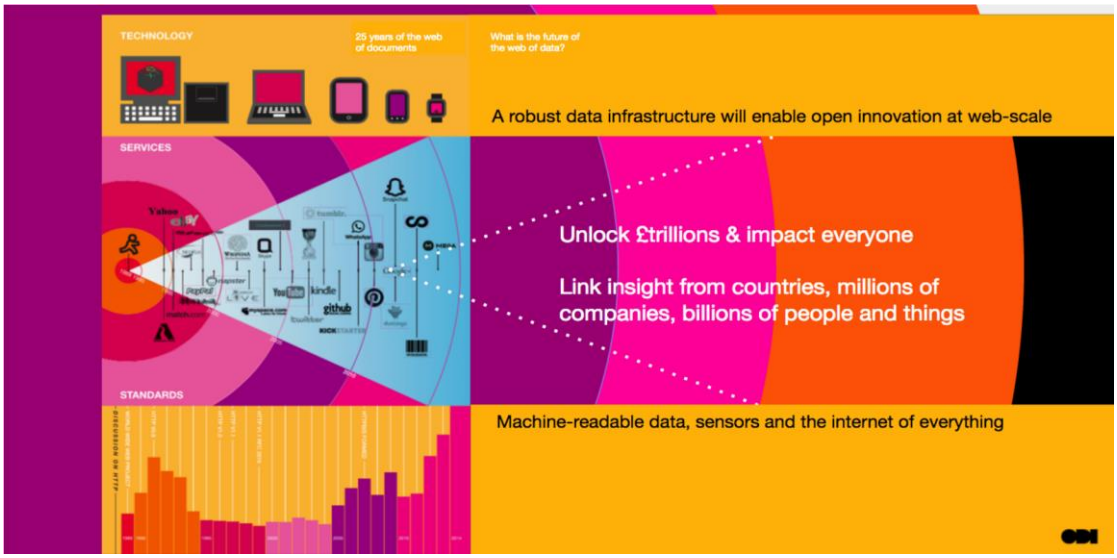

Illustration 1 L'étendue croissante du web

Les applications qui utilisent de telles données sont déjà répandues, depuis les planificateurs de voyages, les applications de météo et d'achats jusqu'aux ensembles. Ces applications échangent des données pour les aider à fonctionner, mais ces données sont souvent cachées, ce qui rend leur accès et leur utilisation difficiles pour les autres.

Le reste de cette leçon explore les deux approches des données sur le web de documents et comment les rechercher, avant d'explorer le réseau de données et comment commencer à libérer son potentiel.

5. Portails gouvernementaux

L'histoire des données ouvertes est étroitement liée aux lois qui régissent l'accès à l'information publique. Ces lois garantissent que le public a le droit d'accéder à l'information de ceux qui fournissent des services publics. Le mouvement de données ouvertes tente une inversion complète de cette logique. Plutôt que d'avoir à demander des données, ces données devraient déjà être « ouvertes par défaut »; fermer un ensemble de données devrait exiger une bonne raison plutôt que le contraire. Les gouvernements et les fournisseurs de services publics devraient travailler de manière proactive et ouverte.

Pour aller plus loin, les gouvernements de nombreux pays ont signé le Partenariat pour un gouvernement ouvert (Open Government Partnership - OGP). OGP a été lancé en 2011 pour fournir une plate-forme internationale aux réformateurs nationaux qui s'engagent à rendre leurs gouvernements plus ouverts, plus responsables et plus réceptifs aux citoyens. Un élément clé de l'OGP est l'engagement d'être ouvert par défaut et d'ouvrir les données. Cela a mené au lancement de nombreux portails de données ouvertes du gouvernement qui ont été construits pour contenir leurs données et les rendre facilement accessible au public.

Au fil des ans, les activités de données ouvertes ont évolué dans la mesure où les gouvernements sont maintenant évalués sur la qualité de leur activité et sur leur durabilité.

Le Baromètre des données ouvertes [Open Data Baromètre (ODB)] de la World Wide Web Fondation (Fondation pour le Web) note les gouvernements sous trois aspects : la préparation, la mise en œuvre et l'impact. Le résultat de mise en œuvre est mesuré en recherchant des ensembles clés de données ouvertes, présents, accessibles et à la hauteur des données. Dans le baromètre 2016, la note de mise en œuvre recherche la disponibilité de 15 types de données.

1. Données cartographiques
2. Données sur la propriété foncière
3. Données détaillées du recensement
4. Budget détaillé du gouvernement
5. Dépenses gouvernementales détaillées
6. Registre des entreprises
7. Législation
8. Horaires des transports en commun
9. Données sur le commerce international
10. Performance du secteur de la santé
11. Données de performance de l'éducation primaire et / ou secondaire
12. Statistiques de la criminalité
13. Statistiques nationales de l'environnement
14. Résultats des élections nationales
15. Contrats publics

Les ensembles de données tels que les résultats des élections, les dépenses gouvernementales et la performance en matière d'éducation sont des documents d'importance historique. De tels ensembles de données sont de nature très statique et peuvent être facilement rendus disponibles dans une feuille de calcul pour être téléchargés via un portail. En même temps, ce sont ces documents qui sont davantage liés aux lois sur l'accès à l'information et ont un potentiel économique moindre pour une large réutilisation.

À l'inverse, les ensembles de données tels que les cartes, les entreprises et les données commerciales sont beaucoup plus dynamiques et, en tant que tels, conviennent mieux à un niveau de service différent à partir d'un simple téléchargement de fichiers. Cela est d'autant plus vrai que les données cartographiques prennent la forme de nombreux formats complexes et que les données commerciales peuvent dépasser les tailles de fichiers habituelles pour l'accès par téléchargement.

L'agriculture est un secteur qui recoupe un grand éventail de domaines, avec de nombreux ensembles de données différents disponibles dans un grand nombre de portails. Par exemple, les données de cartographie sont essentielles pour donner des informations sur l'agriculture, bien au-delà des aspects de propriété foncière tels que le captage et le ruissellement, l'utilisation des terres ainsi que les zones protégées / défendues peuvent tous contribuer à donner des informations sur l'utilisation des terres.

De nombreux gouvernements utilisent maintenant un portail data.gov.XX (ou des variantes dans la langue locale, par exemple datos.gob.mx) qui fournissent un catalogue central pour les données gouvernementales. De nombreux portails ont des sections dédiées à l'agriculture, souvent étroitement liées au département gouvernemental concerné, qui offrent de nombreux liens et un accès à des données téléchargeables.

Comme mentionné précédemment, trouver des données dans ces portails repose fortement sur la recherche de métadonnées. Pour ce faire, le titre et la description de chaque ensemble de données sont essentiels pour en permettre la découverte. Cependant, étant donné que le portail sera souvent organisé par département ou activité, la recherche d'un ensemble de données nécessite soit une connaissance du domaine spécialisé, soit une connaissance de la manière dont le gouvernement local organise et décrit ses données dans le portail. Un bon point de départ est de parcourir le portail par vous-même pour obtenir des indices sur la manière dont il est organisé et découvrir comment les données sont décrites avant de cibler votre recherche en utilisant ces nouvelles connaissances.

6. Obtenir des données de 'sur le web'

Comme mentionné précédemment, la plupart des données ouvertes disponibles ne sont disponibles que sur le Web ; soit via un bouton de téléchargement ou contenues dans les pages Web elles-mêmes. Cette section examine les techniques pour commencer à découvrir et déverrouiller ces données prêtes à être utilisées.

6.1. Comment trouver les données téléchargeables

De nombreux fournisseurs sont utiles et vous donnent des liens de téléchargement lisibles par l'homme afin d'obtenir leurs données. La plupart des ensembles de données sur les portails de données gouvernementaux fonctionnent de cette façon, explorez les liens ci-dessous pour en découvrir plus :

- Statistiques nationales du Royaume-Uni - Dernières statistiques sur les stocks de céréales
- Tanzania Agricultural Census (Recensement agricole de la Tanzanie)[13]

De nombreux moteurs de recherche, y compris Google, vous permettent d'utiliser des recherches et des préfixes avancés pour extraire des données de sources telles que celles mentionnées ci-dessus. Les recherches avancées utilisent des filtres et des préfixes afin de limiter le type de recherche et les résultats. Une liste de préfixes et d'exemples liés peut être vue dans le tableau1.

Tableau 1 Préfixes pour la recherche avancée

Préfixe	**Description**	**Exemple de recherche**
Type de fichier:	Rechercher uniquement les types de fichiers spécifiques	Type de fichier : stocks de céréales xls
site:	Rechercher avec un domaine ou un site spécifique uniquement	site : opendata.go.tz agriculture
lié:	Rechercher du contenu lié à une page connue	related: https://www.gov.uk/government/statistics/cereal-stocks
lien:	Lister seulement les pages qui pointent vers celle donnée	link: https://www.gov.uk/government/statistics/cereal-stocks

Chacun d'entre eux peut aider à affiner votre recherche de données. Tandis que les deux premiers aident à affiner votre recherche, les deux derniers l'élargissent à nouveau une fois que vous avez trouvé une ressource pertinente. Tandis que « lié » peut vous aider à trouver d'autres contenus pertinents, ceux qui sont liés à votre ensemble de données peuvent avoir utilisé les données et aider à fournir un contexte par rapport à l'utilisation existante. Étant donné que la majorité des ensembles de données sous licence ouverte nécessitent une attribution, la recherche spécifique 'lien :' devrait renvoyer au moins un nombre de résultats si l'ensemble de données a été utilisé.

[13] http://opendata.go.tz/dataset/tanzania-utafiti-wa-sampluli-sensa-ya-kilimo-2007

6.2. Agrégateurs de données

L'un des principaux défis auxquels sont confrontées les données « sur le Web » est le manque de capacité de recherche à l'intérieur des jeux de données mêmes. Les moteurs de recherche existants n'autorisent que la recherche de métadonnées, quel que soit le format du fichier (cela est tout aussi vrai pour les audio et les vidéos, mais Google vous permet maintenant d'effectuer une recherche en utilisant une image existante).

De nombreux portails de données agissent comme agrégateurs et permettent une certaine exploration des données, à partir d'un seul fournisseur de données ou d'un ensemble. En ce qui concerne l'agriculture, la Banque mondiale regroupe des données statistiques clés provenant de nombreux pays et en permet l'exploration et le téléchargement.

Un exemple fourni par la Banque mondiale est la liste d'Indicateurs de l'Agriculture et du Développement rural (rural Agriculture and Rural développement indicateurs)[14]. L'agrégation des données permet d'explorer des indicateurs tels que la superficie des terres agricoles par rapport à la population rurale. Cet exemple propose aussi une la comparaison entre un certain nombre de pays d'Afrique de l'Est et de l'Ouest. Le service Data Bank de la Banque mondiale offre un accès facile pour explorer et télécharger ces données prêtes à l'emploi.

De tels agrégateurs peuvent être une source cruciale de données ouvertes lorsque les portails de pays ne sont pas à jour, ou ne sont tout simplement pas disponibles.

Enigma.io, vainqueur de Techcrunch Disrupt en 2013[15], rassemble des données provenant d'une multitude de sources de données ouvertes et permet une recherche fine dans les données elles-mêmes. C'est en fait une recherche inversée sur les données ; plutôt que de rechercher les métadonnées pour trouver les données, enigma.io effectue une recherche dans les données et vous indique les ensembles de données dans lesquels se trouve votre terme de recherche. Par exemple, une Recherche sur 'Monsanto'(searching for 'Monsanto')[16] produit toutes sortes d'ensembles de données intéressantes, allant des contributions de la campagne fédérale en 2016 jusqu'aux bulletins météorologiques de 1949.

[14] https://data.worldbank.org/topic/agriculture-and-ruraldevelopment?locations=KE-TZ-RW-GH-NG-ML-BF
[15] http://www.businessinsider.com/techcrunch-disrupt-winner-enigma-2013-5?IR=T
[16] https://public.enigma.com/search/Monsanto

6.3. Extracteurs de données (Scrapers)

Parfois, les données ne pourront pas être téléchargées dans un format utilisable. Parfois, les données seront disponibles uniquement à partir de la page Web sous forme de tableau ou de liste. Dans d'autres cas, les données peuvent être disponibles dans un format de document (tel que PDF) plutôt que dans un format de données. Dans les deux cas, l'utilisation de scrapers de données peut aider à extraire ces données visibles.

6.4. Scrapers de données Web

Les scrapers Web permettent l'extraction automatique de données structurées à partir d'une page Web. Des outils comme grepsr[17] permettent l'extraction automatique des données de sites Web structurés en quelques secondes, y compris la capacité de gérer la pagination et les résultats de défilement infinis. Actuellement, ces outils impliquent généralement un coût par enregistrement pour l'extraction avec un nombre limité de crédits gratuits par mois.

6.5. Scraper de données PDF

Un autre endroit où les données utiles sont souvent intégrées est dans les rapports PDF produits par les agences de statistiques. Ceux-ci contiennent souvent de longues annexes de données tabulaires qui peuvent être extraites à l'aide d'outils comme PDFTables.com. Essayez-le vous-même avec statistiques agricoles [18] de la Tanzanie. Les données commencent à la page 124 de ce rapport et il est conseillé de réduire le PDF aux pages exactes qui contiennent les données requises avant de télécharger vers un extracteur PDF comme PDFTables.com.

7. Obtention de données depuis "sur le web"

L'évolution du Web a conduit à la nécessité de séparer l'infrastructure principale et les données de la couche de présentation telles que les sites Web et les applications mobiles qui utilisent toutes ces mêmes données. Les applications de magasinage, de météo et de voyage offrent toutes différentes options permettant aux utilisateurs d'interagir essentiellement avec les mêmes données.

Ces applications utilisent tous les services de données dédiés pour accéder aux données et les interroger. Beaucoup de ces services de données sont documentés pour quiconque à utiliser alors que beaucoup restent cachés pour diverses raisons commerciales ou budgétaires. Cette section se penche

[17] https://www.grepsr.com
[18] http://harvestchoice.org/sites/default/files/downloads/publications/Tanzania_2007-8_Vol_5g.pdf

sur les différentes techniques qui peuvent être essayées pour accéder aux données « sur le Web», ce qui aide à servir de telles applications.

7.1. Extensions de type de fichier

Certains sites Web ont été conçus pour offrir un moyen d'extraire des données en ajoutant une extension de fichier à l'URL de la page Web consultée. Pour ces sites Web, généralement gérés par des organisations qui publient également des données ouvertes téléchargeables, l'ajout de l'extension correcte entraîne le téléchargement de cette page dans un format de données, par opposition à un format de document.

Un bon exemple de cela est le site Web du gouvernement britannique (www.gov.uk), qui fournit n'importe quelle page dans un format de données simplement en ajoutant l'extension appropriée comme '.json', par exemple www.gov.uk/browse/business.json. Pour afficher les données sous une forme plus facilement lisible, copiez-les dans jsonlint.io.

Le Tarif commercial du Royaume-Uni [19] a également la même fonctionnalité et contient des détails sur les codes du commerce international qui peuvent être liés aux données commerciales disponibles auprès de Revenu et Douanes [20].

Malheureusement, peu de sites Web indiquent clairement aux humains que des formats alternatifs (tels que JSON) sont disponibles. Un bon indicateur est de trouver des sites web modernes où les pages contiennent clairement des données, telles que des entreprises individuelles [21], où de telles extensions peuvent être essayées. Le Tableau 2.1.2 répertorie les formats de données communs disponibles pour les données « sur le Web».

Tableau 2 *Formats communs pour les données 'sur le web'*

Extension	**Description**
.csv	Valeurs séparées par des virgules. Format de données tabulaire comme Excel mais dépouillé juste pour contenir des données dans une structure simple.
.json	Notation d'objet JavaScript. Un format de données hiérarchique natif du langage JavaScript qui est largement utilisé sur le Web car il fait partie de la spécification HTML5.

[19] https://www.trade-tariff.service.gov.uk/trade-tariff/sections
[20] https://www.uktradeinfo.com/Statistics/BuildYourOwnTables/Pages/Home.aspx
[21] https://opencorporates.com/companies/ch/471356

.xml	extensible Mark up Langage. Une spécification de balisage qui a un large éventail d'utilisations. A été critiqué pour sa complexité et sa verbosité par rapport à JSON.
.rdf	Bien que RDF ne devrait pas être un format de données (non couvert ici). RDF définit une structure de données formelle qui peut être appliquée aux formats xml, json et csv. L'utilisation de l'extension implique que la structure est utilisée et le plus souvent les données elles-mêmes sont au format XML.
.rss	Une autre structure XML spécifique souvent utilisée pour les flux de données qui sont régulièrement mis à jour, tels que les actualités et la météo.

7.2. Interfaces de programmation d'application (APIs)

Les API sont l'un des meilleurs moyens d'accéder aux données. Les API sont un service décrit comme une « promesse» par un système de fournir constamment et constamment un service à un autre qui permet aux deux d'interagir. Pour cette raison, les API présentent de nombreux avantages par rapport à toute autre forme d'accès aux données, comme indiqué ci-dessous.

1. Contrats de service. Comme une API est un service, cela garantit l'accès aux données et peut souvent s'accompagner d'accords de niveau de service pour ceux qui souhaitent les utiliser.

2. Accès direct. Les API fournissent un mécanisme grâce auquel les données peuvent être incluses en direct dans une application. L'exemple le plus courant d'une API de données est le temps de transport en direct. Au dos d'une seule API, plusieurs centaines d'applications peuvent être créées.

3. Conçu pour les données. Le plus grand avantage d'une API réside peut-être dans le fait qu'elle est conçue pour les données et les machines plutôt que pour les humains. Cela signifie que la disponibilité des données n'est plus limitée par les paradigmes de la façon dont les humains utilisent le Web, mais cela crée des défis lors de la recherche de données qui pourraient être dans une API.

Le principal inconvénient des API est que les données ne sont pas aussi facilement accessibles au téléchargement et à l'utilisation immédiate. Certaines applications tierces, comme enigma.io, utilisent déjà des API pour accéder aux données d'autres services afin de faciliter l'accès, alors que d'autres comme OpenCorporates autorisent les téléchargements par extension de fichier dans le cadre de leur API.

Des exemples de services comportant des API incluent: OpenCorporates[22], OpenSteetMap[23], Twitter[24], Flickr[25], et LinkedIn[26]. Ces API fournissent un accès direct aux données brutes et permettent de larges requêtes pour une recherche à multiples facettes.

La plupart des plates-formes de données ouvertes fournissent des API pour accéder aux données, y compris Socrata et OpenDataSoft. Ces plates-formes sont utilisées par un certain nombre de gouvernements et de départements, Socrata principalement aux Etats-Unis, et OpenDataSoft dans toute l'Europe. CKAN, une alternative open source, possède également une API bien que dans de nombreux cas, cette API ne donne accès qu'aux enregistrements de métadonnées.

Le tableau 3 contient quelques exemples de chaque plate-forme et certains des ensembles de données agricoles disponibles, dont certains ont été mentionnés précédemment.

Tableau 3 *Exemples de plates-formes de données ouvertes et de jeux de données agricoles disponibles*

CKAN	page web: https://data.gov.uk/dataset/cereal_stocks_england_and_wales API: https://data.gov.uk/api/3/action/package_show?id=cereal_stocks_england_and_wales
Socrata	Page Web: https://data.code4sa.org/dataset/List-of-RegisteredDams-2014/iety-gmha API: https://data.code4sa.org/resource/cig6-sz38.csv
OpenDataSoft	*Aucun exemple agricole trouvé.*

7.3. Utilisation des API

De nombreuses API Web prennent la forme d'API REST. Représentationnel State Transfer (REST) est une API spécialement conçue pour le Web. Il a un ensemble spécifique de directives et de règles qui déterminent si quelque chose est une API RESTFUL.

De manière générale, une API REST nécessite l'utilisation d'identificateurs de ressources requis pour télécharger / télécharger la ressource requise. Dans le

[22] https://api.opencorporates.com
[23] http://wiki.openstreetmap.org/wiki/API
[24] https://developer.twitter.com/en/docs
[25] https://www.flickr.com/services/api/
[26] https://developer.linkedin.com/docs/rest-api

cas de l'exemple Socrata de la section précédente, l'adresse de l'API est l'identificateur basé sur le Web de la ressource (https://data.code4sa.org/resource/cig6sz38). Cliquez sur cette ressource pour vous rediriger vers la page Web, non parce que c'est ce que vous cliquez sur ce lien, mais parce que c'est ce qui a été demandé lorsque vous avez cliqué sur le lien dans un navigateur Web. Le navigateur Web va à une demande GET pour une représentation de la page Web (texte / html) de cette ressource comme indiqué ci-dessous.

GET /resource/cig6-sz38 HTTP/1.1
HOST: data.code4sa.org
ACCEPT: text/html

Une API REST spécifie qu'une machine devrait être capable de changer la requête afin de demander différentes représentations de la même ressource. C'est un peu comme ajouter des extensions de fichiers, sauf lorsque la ressource demandée ne change aucune partie de son emplacement sur le web (l'ajout de '.csv' change effectivement l'URL). L'exemple ci-dessous montre deux exemples de requêtes pour une version JSON et CSV de la même ressource à l'aide d'une API REST.

JSON	CSV
GET /resource/cig6-sz38 HTTP/1.1 HOST: data.code4sa.org **ACCEPT: text/csv**	GET /resource/cig6-sz38 HTTP/1.1 HOST: data.code4sa.org **ACCEPT: application/json**

Les API REST sont simplement des extensions du protocole de transfert hypertexte existant sur les sites Web, sauf pour les données. Ainsi, il est possible de changer le type de requête d'un GET à un PUT puis d'envoyer des données structurées au serveur pour remplacer les données existantes par de nouvelles données (en utilisant évidemment l'authentification). La ville de Chicago utilise la méthode POST pour envoyer la mise à jour statistiques de criminalité[27] à leur portail de données et le font quotidiennement depuis 2001.

Les API permettent non seulement aux utilisateurs d'accéder aux données, mais elles constituent également un élément clé de l'infrastructure de données du fournisseur, ce qui permet de gérer et de maintenir les données à jour.

7.4. API cachées

[27] https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2/data

Tous les sites Web qui chargent dynamiquement des données ne rendent pas leur API publique, même si elle existe. Cependant il est possible de les découvrir. Cela nécessite une bonne quantité de connaissances techniques ; néanmoins, une bonne recherche sur Google fait souvent apparaître des communautés de personnes qui ont déjà créé quelque chose pour le service particulier dont vous voulez extraire des données.

Comme de nombreuses API sont basées sur la conception de l'API REST, dans de nombreux cas, il est assez simple pour quelqu'un qui connaît les API REST de trouver rapidement si un service en a un et comment cela fonctionne. Cela peut être fait en essayant certaines demandes REST avec des extensions de navigateur comme Postman for Google Chrome[28].

L'outil Extracteur expérimental de Donnée Cachées[29] a été conçu par l'ODI pour rechercher automatiquement les API REST qui échangent des données JSON lorsqu'une page Web est chargée.

Vérification de vos droits d'utilisation des données :

Il existe plusieurs façons d'obtenir des données sur le Web, qu'elles soient clairement visibles via un bouton de téléchargement ou disponibles via une API publique ou cachée. Quelle que soit la méthode d'acquisition des données, il est essentiel de vérifier vos droits à la fois pour utiliser cette méthode et pour utiliser les données qui en résultent.

Tout comme les données elles-mêmes, certaines déclarations légales seront seulement lisibles par l'homme, certaines seulement lisibles par machine et d'autres une combinaison des deux. Toutefois, les fournisseurs de services disposeront généralement d'une version lisible par l'homme de leurs conditions d'utilisation et / ou de leur licence de données qui couvrira à la fois les conditions d'utilisation du service et les droits d'utilisation des données une fois acquises.

De nombreux portails de données gouvernementaux verront la licence de données répertoriée comme une métadonnée liée à l'enregistrement consulté. Par exemple, dans data.gov.uk, toutes les licences sont répertoriées directement sous le titre de l'ensemble de données en tant que lien cliquable. La plate-forme CKAN (dont data.gov.uk est une version) est particulièrement efficace pour exposer les déclarations de droits. Cela aide les utilisateurs à s'assurer que les données qu'ils consultent sont des données ouvertes.

Des services comme Flickr ont également des licences contre chaque photo. Chaque utilisateur de Flickr est capable de spécifier des licences pour ses

[28] https://chrome.google.com/webstore/detail/postman/fhbjgbiflinjbdggehcddcbncdddomop?hl=en

[29] http://odinprac.theodi.org/hidden-data-extractor/

propres photos. Flickr fournit même une recherche qui permet aux autres de trouver des photos avec des licences spécifiques. Si les licences sont également lisibles par machine (comme c'est le cas avec CKAN et Flickr), les moteurs de recherche peuvent les utiliser comme métadonnées, ce qui signifie que les résultats de recherche peuvent être filtrés instantanément pour ne contenir que du contenu sous licence libre (essayez Google recherche avancée [30]).

Si vous utilisez une API REST, l'instruction de droits peut être renvoyée dans le cadre d'un en-tête Link. Ceci sépare l'énoncé de droits du contenu, ce qui permet à la réponse d'être toujours la donnée pure, par ex. Fichier CSV.

Si aucune de ces options n'existe, il peut être nécessaire de lire les termes et conditions des fournisseurs pour s'assurer que votre méthode d'accès et vos droits d'utilisation des données sont autorisés. Ce n'est pas parce que quelque chose est accessible sur le web que tout le monde a le droit de l'utiliser.

Résumé

Cette leçon a introduit plusieurs des méthodes et des lacunes reliées à la découverte de données ouvertes sur le Web. Il est encore tôt dans l'évolution d'un « âge des données » à la suite de «l'ère de l'information» et les services spécialisés dans l'accès rapide aux données évoluent.

Dans le même temps, le nombre de services fournissant des données augmente également, reflétant les premiers jours du web. Il reste encore des leçons à tirer, mais les méthodes d'accès aux données commencent à se stabiliser avec l'émergence d'API communes telles que REST.

Les formats de données ont également évolué et ont donc aussi des méthodes pour découvrir et accéder aux données. Les moteurs de recherche deviennent beaucoup plus intelligents et peuvent être personnalisés pour effectuer des requêtes très ciblées. En même temps, les outils pour extraire et travailler avec les données ont évolué de telle sorte qu'il est très facile de commencer à travailler avec des données quel qu'en soit le format.

L'évolution des applications mobiles qui exigent un accès instantané aux données a également augmenté le nombre d'API disponibles, même si certaines d'entre elles restent cachées. Il est clair que nous vivons à l'ère des données, mais nous devons faire attention à nos droits d'utilisation de ces données. Avoir des licences de données ouvertes claires est essentiel pour l'avenir de notre infrastructure de données.

[30] https://www.google.co.uk/advanced_search