HETEROGENEITE ET TRAITEMENT DES DONNEES

Alain van Cuyck

HETEROGENEITE ET TRAITEMENT DES DONNEES

Alain van Cuyck

Introduction : le contexte général de l'ouverture des données ouvertes.

L'ouverture des données publiques en France et dans le monde a généré un important mouvement de mise à disposition auprès des citoyens de nombreuses données détenues par les pouvoirs publics. En France, c'est à la suite du décret n° 2011-194 du 21 février 2011 portant création d'une mission « Etalab **»** chargée de la mise au point d'un portail unique interministériel des données publiques que le site « *data.gouv.fr* » a été mis en place par décret au JO du 22 février 2011.

Au 11 juillet 2014 ce site recensait 354 894 jeux de données publiques**,** un forum et une communauté. En juillet 2014, 42 pays dans le monde se sont également lancés dans des initiatives similaires[i].

Ainsi le site « data.gouv.uk », l'équivalent anglais du site français « data.gouv.fr », répertoriait 37 599 jeux de données portant sur des secteurs aussi variés que les dépenses du gouvernement, l'économie, la démographie, la santé et la prévention, le marché du travail, l'environnement, les opérations gouvernementales, la population et la santé, l'éducation, la justice, les finances, le monde des affaires, le géo spatial, les transports, le logement, l'agriculture, le planning, les conditions sociales, l'art le sport et la culture, l'énergie…

Il y a donc autant de domaines et d'activités que de types de traitement de l'information, ce qui nous a conduit à mieux comprendre et à mieux saisir la diversité et l'hétérogénéité des données dans les processus de traitement des open data, sachant que tous les acteurs ne sont pas sur les mêmes avancées en matière de traitement des données. Bien souvent l'existence des open data repose sur un maillage relativement fin et complexe des environnements concernés avec pour chaque domaine une « écologie propre », relevant non seulement de plusieurs facteurs comme le niveau des savoir-faire techniques et informatiques ou le degré de mise en place de traitements de l'information complexe, mais aussi des possibilités d'agrégation et de traitement entre les données entre elles.

Ce mouvement plus profond s'inscrit également dans l'évolution profonde du web des données à partir des années 2007qui a donné lieu à l'émergence de ce que Tim Berners Lee a appelé le web sémantique, ayant pour objectif de rendre interopérable les données entre elles. La problématique de fond de ce mouvement sur l'open data était en effet de rendre compatible les données pour les rendre accessibles à tous les types de machines. On parle alors de « données bridées » lorsqu'elles ne respectent pas les standards du Web et que leur utilisation est amoindrie, voire impossible.

Une donnée ouverte selon la définition de Wikipédia (Wikipédia, 2015) est « une donnée numérique d'origine publique ou privée. Elle peut être notamment produite par une collectivité, un service public (éventuellement délégué) ou une entreprise. Elle est diffusée de manière structurée selon une méthode et une licence ouverte garantissant son libre accès et sa réutilisation par tous, sans restriction technique, juridique ou financière ».

1. <u>Organisation des données ouvertes</u>

1.1. Les licences

En 2008 le standard SPARQL est devenu une recommandation du W3C. Ce langage de requête permet aux développeurs de tester leurs requêtes directement depuis leurs navigateurs Web sur les données ouvertes en ligne puis de développer leur propre programme pour analyser ces données. On peut ainsi consommer les données à distance sans avoir à les transformer ou à les déplacer. Ainsi, les gouvernements au Royaume-Uni et aux États-Unis ont commencé à basculer leurs données dans le Web des données ouvert (en anglais, *Linked Open Data* ou LOD) en respectant les standards du W3C et en offrant un point d'accès SPARQL pour les développeurs.

En France, depuis 2013, l'harmonisation des pratiques a conduit à l'usage de deux licences types utilisées dans des proportions équivalentes: la Licence Ouverte et l'Open Database Licence.

1.1.1 La Licence Ouverte

Cette licence créée par ETALAB est destinée à être utilisée notamment sur data.gouv.fr. afin d'encadrer l'ouverture des données de l'État français.

1.1.2 L'Open Database License

La licence ODbL a été traduite en français par la ville de Paris afin de l'adapter à un usage national. Une variété de projets utilisent cette licence, depuis Open Street Map aux collectivités locales (Paris, Nantes, Toulouse…).

1.2. Quelques caractéristiques des données ouvertes

Comme le rappelle l'article relatif aux données ouvertes de wikipédia,

> Les principaux problèmes de l'exploitation des données ouvertes sont d'ordre technique, les données en masse ne peuvent pas être traitées humainement. Le concept de Web des données appliqué aux données ouvertes met en œuvre 3 mécanismes :
>
> 1. Permettre l'existence de la donnée sur le réseau à travers une URI unique (cela inclut les URL).
> 2. Diminuer le coût de transformation de la donnée en apportant des formats standards lisibles par les machines (comme avec RDF, RDFa ou les microdonnées dans le HTML5).
> 3. Améliorer la qualité de la donnée pour éviter qu'un traitement de mise à disposition ne puisse les altérer. Un entrepôt de données même avec des erreurs est préférable qu'un entrepôt biaisé. Ainsi, des mécanismes pour la fréquence et l'automatisation des mises à jour de la donnée par les producteurs des données est possible avec un service SPARQL sur ces données.
>
> Les données ouvertes ne sont contrôlables par leurs producteurs (contrôle des mises à jours) et réellement exploitables par d'autres qu'à la condition d'utiliser ces 3 mécanismes.

1.3. Les principaux critères pour la définition d'une donnée ouverte

En 2010, la Sunlight Fundation a établi une liste de 10 critères[ii] caractérisant une donnée ouverte. Pour qu'une donnée soit dite ouverte, elle doit être : Complète, Primaire, Opportune, Accessible, Exploitable, Non-Discriminatoire, Non-Propriétaire, Libre de droits, Permanente, Gratuite.

Tim Berners-Lee, toujours en 2010, proposait une échelle de qualité des données ouvertes de 1 à 5 étoiles[iii]

★	Données non filtrées (éventuellement dégradées) par exemple mises en ligne avec n'importe quel format[41]
★ ★	Données disponibles de manière structurée (ex : données tabulaires en CSV, XML, Excel, RDF)[42]
★ ★ ★	Données librement exploitables - juridiquement (Cf. licences), - techniquement (dans des formats non-propriétaires, pas sous Excel notamment)
★ ★ ★ ★	Données identifiées par des URL (avec date de mise à jour) afin que l'on puisse « pointer » un lien vers elles (et la retrouver éventuellement mise à jour)
★ ★ ★ ★ ★	Données liées à d'autres données pour les contextualiser et les enrichir

Cette échelle concerne les données numériques de base, mais ses critères peuvent être adaptés à des données agrégées ou des informations publiques plus complexes (photos, vidéo, rapports, études, etc. qui devront aussi être mis en ligne, idéalement avec une métadonnée de qualité et pouvant aussi intégrer de l'hypertexte dans le cas des rapports et études, voire de certaines vidéos).

Une métadonnée selon la définition de Wikipédia[iv] est « une donnée servant à définir ou décrire une autre donnée quel que soit son support (papier ou électronique) ».

Un exemple type est d'associer à une donnée la date à laquelle elle a été produite ou enregistrée, ou à une photo les coordonnées GPS du lieu où elle a été prise.

Les métadonnées sont à la base des techniques du Web sémantique. Elles sont définies dans le cadre du modèle « Resource Description Framework (RDF) ».

On peut également rajouter à ces deux échelles de critères un troisième modèle emprunté au big data et qui apparait pertinent par rapport à l'open data. Enoncé initialement par le cabinet Mc Kinsey en 2011, ce modèle s'appuie sur la règle des 3V : « un grand Volume de données, une importante Variété de ces mêmes données et une Vitesse de traitement s'apparentant parfois à du temps réel. Ces technologies étaient censées répondre à l'explosion des données dans le paysage numérique (le « data déluge »). Puis, ces qualificatifs ont évolué, avec une vision davantage économique portée par le 4ème V la Valeur, et une notion qualitative véhiculée par le 5ème, celui de Véracité des données (disposer de données fiables pour le traitement). ».[v]

La dimension économique de la valeur des données est au cœur des logiques de l'open data et la fiabilité des données est un critère extrêmement crucial, notamment quand à leur validité temporelle le plus souvent très volatile à l'instar de nombreux secteurs soumis à la grande variabilité des données (secteur du transport, de la météorologie, de la santé, des finances, de l'économie etc….)

1.4. Bases et plateformes de données

Selon la définition de Wikipédia (Wikipédia, 2005) :

> La base de données est au centre des dispositifs informatiques de collecte, mise en forme, stockage, et utilisation d'informations. Le dispositif comporte un système de gestion de base de données (abr. SGBD) : un logiciel moteur qui manipule la base de données et dirige l'accès à son contenu. De tels dispositifs — souvent appelés base de données — comportent également des logiciels applicatifs, et un ensemble de règles relatives à l'accès et l'utilisation des informations.
>
> Lorsque plusieurs choses appelées bases de données sont constituées sous forme de collection, on parle alors d'une banque de données (en anglais : *data bank*).
>
> « Les bases (de données) sont souvent des « entrepôts » de données[2] (anglais *datawarehouse*), utilisées pour collecter d'énormes quantités d'éléments historiques de manière quotidienne depuis une base opérationnelle. Dans de telles applications l'accent est mis sur la capacité d'effectuer des traitements très complexes et le logiciel moteur (le SGBD) est essentiellement un moteur *d'analyse* ».
>
> Les bases de données sont également classifiées selon les caractéristiques du contenu : des données *non structurées* sont stockées à l'état brut, et nécessitent d'être retraitées en vue de produire de l'information - de la connaissance. Inversement des données *structurées* sont formatées en fonction de l'usage qui va en être fait, en vue de faciliter le stockage, l'utilisation et la production d'information finie. Il est alors possible d'effectuer des calculs.[vi]

La plupart du temps les données mises à la disposition des usagers de l'open data sont des données brutes ne permettant pas un usage direct pour effectuer des traitements et des calculs. le retraitement de l'information est donc nécessaire et les usagers non qualifiés dans le domaine de l'informatique ne peuvent en avoir qu'une utilisation restreinte. Toutefois, ces données peuvent servir de base pour la mise en œuvre de diverses applications. On donnera par exemple la possibilité d'avoir les coordonnées géographiques de toutes les gares SNCF, mais cela demandera des traitements spécifiques si l'on veut par exemple organiser l'ensemble des gares par département.

1.5. Organisation des plateformes de données

Lorsqu'elles sont mises en place à un niveau national, les jeux de données sont généralement centralisées au sein d'une plateforme de données les rendant accessibles au grand public (en France data.gouv.fr, en Angleterre data.gov.uk etc… par l'intermédiaire d'un service chargé de centraliser les données issues des administrations publiques (ex : Datalab, France)..

Bien entendu, les choses sont relativement compliquées et complexes dans la réalité, car les informations provenant de différentes administrations concernent différents types de données,

avec des temporalités et des actualisations différentes. Ce manque d'homogénéité appelle donc une véritable recomposition de l'ensemble pour permettre une meilleure accessibilité aux différents publics.

Si les Etats sont de fait les plus importants fournisseurs de données publiques, les collectivités locales territoriales sont également appelés à se mobiliser autour de politique d'open data à leurs échelles respectives. Ce mouvement s'est réalisée en France à partir d'innovations portées essentiellement par des villes, des départements et des régions.

Ainsi au 9 mai 2014, plus de 40 collectivités locales s'étaient engagées dans une politique d'ouverture publique des données. Evidemment, l'importance et la taille des collectivités jouent un grand rôle sur le nombre de jeux de données offerts à la consultation, mais ce n'est toutefois pas systématique. Ex. : 94 jeux de données pour Paris et 244 pour La Rochelle…. Le nombre de jeux de données n'est pas par ailleurs un indicateur absolu car il peut varier en fonction des choix de présentation qui sont fait : ainsi telle ville découpera ses données par années ce qui lui permettra d'afficher 10 jeux de données, alors que telle autre les synthétisera sur une même liste regroupant les 10 années.

Sans citer l'ensemble des jeux de données répertoriées, nous indiquons ici les collectivités affichant le nombre le plus important de jeux de données

Parmi les villes répertoriées, la ville de Nantes affiche 485 jeux de données, le Grand Lyon 422, La Rochelle 244, Rennes 195, Montpellier 119, Bordeaux 106, Paris 94, Nice 94

Pour les départements et régions, la région Paca répertoriait 448 jeux de données, le conseil régional d'Ile de France 409, la région Aquitaine210, la région Corse 112, la région Nord Pas de Calais 61, le conseil régional d'Auvergne 12… cette liste n'étant ni exhaustive ni finie, il faut savoir que le le processus de mise en place d'ouverture des données publiques ne fait que commencer et la plupart des plateformes ne sont apparues que depuis 2011.

La plupart de ces données sont retraitées pour être interopérables, mais il s'agit en fait essentiellement de données brutes. La ville de Bordeaux mentionne également des services offrant des applications concernant surtout 6 domaines : accessibilité, stationnement payant, parcs et jardins, pistes cyclables, monuments, vins. On peut déceler ici une politique d'open data à deux niveaux : au premier niveau, celle d'une simple mise à disposition de données de base – certes nettoyées, actualisées, mises aux normes et nettoyées et au second niveau une véritable offre d'applications diverses faite aux usagers, service d'un type plus raffiné, correspondant à la mise en place d'un ensemble de dispositifs de traitement des données. Il semble logique que l'open data évoluera au fil des années vers ce second niveau.

2. Traitement et catégorisations

2.1. Les thématiques importantes traitées par les collectivités.

Si l'on s'en tient aux 422 jeux de données du grand Lyon (au 10 mai 2014), celles-ci concernent essentiellement les transports (28), l'imagerie (197), la citoyenneté (12), les services (15), la localisation (178), les limites administratives (13), l'environnement (21), l'occupation du sol (139), l'urbanisme (59), les équipements (23), l'accessibilité (2)….sachant que ces ensembles de données peuvent être classées dans plusieurs thèmes à la fois.

Claude HANCLOT, dans son mémoire consacré au développement et contraintes de l'open data dans les collectivités territoriales reprend la liste élaborée en 2011 par Simon Chignard[vii] en y ajoutant la rubrique éducation sports, loisirs.

Ces 10 catégories sont les suivantes : vie démocratique et données financières, état-civil, démographie, population, économie, arts, culture, patrimoine et tourisme, sports, loisirs, éducation, transports et déplacements, localisation et information géographiques, environnement, urbanismes et habitat, équipements et service d'intérêts publics.

Les données concernant le transport et les données géographiques semblent être les plus importantes, d'une part en fonction d'un usage déjà très développé de l'information en matière de transports et d'autre part en raison de l'importance des données mises en œuvre notamment les données géographiques fournies avec la collaboration du RGU – répertoire Géographique urbain. Claude HANCLOT citant un entretien avec Gregory BLANC BERNARD, chef de projet du grand Lyon en septembre 2013, fait remarquer que les données géographiques au Grand Lyon représentent 95 % des données mises en ligne.[viii]

2.2. Les étapes de la publication des données : collecte, traitement et diffusion des données

Selon le schéma de Libertic[ix] la démarche pour la mise en œuvre d'une plateforme d'open data serait structuré en 4 grandes étapes : découvrir et répertorier les données, formater et nettoyer les données, valider et publier en recommandant les développements et l'enrichissement des métadonnées.

La première étape concernerait l'établissement d'un répertoire interne des données de façon à les répertorier pour créer ensuite un catalogue de données avec un descriptif des formats des sources et des droits et permettant aussi d'identifier les données déjà sollicitées. La seconde étape consisterait à nettoyer et organiser les tableaux de données, les convertir en format ouvert, et les anonymiser le cas échéant (protection des données personnelles). La troisième étape est une étape de validation destinée à s'assurer des droits de propriétés sur les données et à vérifier la conformité des publications aux recommandations de la CNIL et à la loi de 1978. Enfin la dernière étape, concernant la publication, consistera à documenter les données, à rendre opératoire leur interface de programmation (souvent dénommé API pour application programming interface) et à enrichir les métadonnées.

Le département de Loire Atlantique propose une approche assez analogue en différenciant 6 étapes au total : la sélection, l'extraction, le nettoyage, la transformation, la publication et la réutilisation des données.[x]

Le schéma général de stratégie d'ouverture des données publiques Libertic[xi] caractérise les trois étapes importantes de la chaîne des open data : catalogage des données (linked data), enrichissement sémantique par l'adjonction de métadonnées (interopérabilité), et mise en diffusion sur l'espace public (open data). Il faut également veiller au respect de la confidentialité des données lorsqu'elles ne sont pas anonymes dans l'objectif de les rendre compatibles avec les recommandations de la CNIL.

Toutes ces étapes sont nécessaires pour garantir un bon usage des données qui seront ensuite publiées et mises à la disposition des publics afin de pouvoir les réutiliser.

2.3. Catégories et complexité des données : l'exemple du transport et des déplacements

Même si les 10 catégories descriptives citées ci-dessus permettent de mettre un peu d'ordre dans ce foisonnement des données, il y a également à l'intérieur de chacune d'entre elles beaucoup de variété et de diversité. Si l'on prend comme exemple la catégorie transports et déplacements, on peut décliner comme le fait Claude Hanclot[xii] les sous-rubriques suivantes :

> Accidents, tracés d'autoroutes, accessibilité des trottoirs, pistes cyclables, stationnements payants, places réservés, infractions, emplacement handicapés, stations de vélos, voies et adresses, parc de stationnement vélos, stations d'auto-partage, schémas directeurs accessibilité, zones apaisées, disponibilité des parcs relais, accessibilité des équipements, liste des places de livraisons, mobiliers de stationnement et de transports en commun, panneaux indicateurs de signalisation, nomenclature des voies, zones piétonnes, localisation des surfaces podotactiles, escaliers droits, aires de stationnement, localisation des changements de pentes sur les trottoirs, horaires de passages des transports en commun, horodateurs, comptage du trafic, disponibilité des parking, localisation sur voie, lignes de bus, stationnement règlementaire, administration portuaire, code de la route, réseaux de transports, infrastructures, courbe de circulation en temps réel, état du trafic, informations sur les lignes commerciales, offre de bus et de tramway, tables de relation, réseaux hiérarchique de voierie…

Et la liste n'est certainement pas exhaustive. On pourrait en effet rajouter notamment le critère « pollution » comme dans le projet MoCoPo de la région de Grenoble[xiii] où les mesures du trafic routier sont associées et corrélées aux mesures de la pollution sonore et aérienne.

Voilà donc beaucoup de données à agréger, traiter, analyser, mettre en forme pour les rendre visibles et accessibles, ce qui complexifie à la fois les tâches et les traitements.

Les formes de traitements, de visualisations et d'applications varient ainsi fortement d'une collectivité locale à l'autre, que l'on fasse une comparaison internationale comme l'a fait Catherine Bouteiller entre plusieurs grandes villes dans le monde pour le domaine du transport[xiv] ou une comparaison nationale, selon l'exemple de l'étude du CEREMA de 2014.

Ainsi si l'on compare les applications open data pour smartphone de Nantes et Rennes, les données concernant les transports en commun varient de 11 % pour Nantes à 41 % pour Rennes, le vélo 7 % à Nantes, 11 % à Rennes, 36 % pour les parking à Nantes contre 4 % à Rennes, le transport multimodal 11 % à Nantes et 22 % à Rennes, enfin les données concernant le hors transport sont de 36 % à Nantes et 15 % à Rennes.

Selon les auteurs du CEREMA (2014) :

> Ces différences s'expliquent en partie par les jeux de données mis à disposition. Ainsi, la mise à disposition de jeux de données sur les parkings publics de Nantes est a l'origine de nombreuses applications. En revanche, malgré des données semblables, la proportion des applications concernant les transports collectifs (TC) est bien plus faible à Nantes qu'à Rennes : à Rennes, on recense 19 applications basées sur les données d'horaires de transports collectifs, tandis qu'on n'en compte que 6 à Nantes.[xv]

Enfin les données peuvent être orientées en fonction des objectifs de traitement que l'on se donne, ce qui conditionne bien entendu les usages qui en seront fait.
Les données sont toujours liées à des contextes spécifiques – maturité du projet, compétences significatives, existence de données déjà existantes, méthodologies envisagées, solutions

logicielles disponibles, budget, politique générale etc. Tous ces facteurs constituent ainsi un très grand registre de variabilité dans la logique de la production et de la mise en œuvre des projets qui conduisent à une très grande variabilité de l'offre en fonction des spécificités de chaque contexte local.

2.4. Complexité des relations et partenariats

Enfin, pour compléter cette vision « écologique » des caractéristiques hétérogènes des produits de l'open data, il est également à noter que la partie visible de la production de données se situe au sein d'un réseau complexes d'acteurs et d'interactions. L'existence de plateformes mutualisés supposent en effet la coopération de multiples acteurs et services et nécessitent le management de projet multi-partenaires complexes. Ainsi, comme le fait remarquer le rapport du Cerema, (CEREMA, 2014) :

> Dans la plupart des démarches, des partenariats ont été noués afin d'enrichir les jeux de données diffusés sur la plate-forme Open Data. Ces partenariats peuvent concerner d'autres collectivités (Région PACA...) ou des acteurs privés comme les exploitants des réseaux de transport en commun (Rennes Métropole, Toulouse Métropole...), les exploitants des services de vélo en libre-service (Toulouse Métropole...), ou encore les exploitants des parkings. Certains de ces partenaires peuvent avoir un rôle moteur dans la démarche de la collectivité, comme l'exploitant du réseau de transports urbains de Rennes. Toutefois, dans une majorité de cas, les collectivités qui portent le projet Open Data doivent engager un travail pédagogique auprès de leurs partenaires pour les convaincre de l'intérêt d'ouvrir leurs données et de les rendre accessibles sur la même plateforme[xvi].

D'où également la nécessité de travailler en « groupes-projet », pour fédérer les différents acteurs autour d'un objectif commun, le tout avec une volonté politique forte.

3. Complexité du traitement des données

3.1. Complexité des logiciels et des formats

Enfin pour rajouter à la complexité ambiante, les formats des données ne sont pas toujours les mêmes et elles peuvent varier selon les types de données. Ainsi selon le rapport CEREMA, les données ouvertes sur les arrêts de bus (étude comparative sur les sites des Villes de Bordeaux, de Montpellier, Grand Lyon, Toulouse Métropole, Montpellier, Agglomération Nantes Métropole, Gironde, Loire-Atlantique) ont révélé l'emploi des formats suivants :
SHP (11), CSV (10), GTFS(5), KML / KMZ (4), XLS (3),XML, XML Neptune, Mapinfo, JSON, DXF.
Concernant l'emplacement des arceaux publics de stationnement des vélos (y compris en gare SNCF), la comparaison entre les sites de de Bordeaux, Marseille, Paris Pays de la Loire, PACA, Communauté urbaine de Bordeaux) a permis de recenser les formats suivants : CSV (5), JSON (3), XLS (3),KML / KMZ (2), SHP (2), XML,ATOM, WMS, WFS, DWG, ODS.

De la même façon l'étude constatait une grande diversité et une large variété de formats pour les critères de tarification, de localisation des emplacements de vélos en libre service, d'arrêts de lignes, de qualification de l'offre et services des transports, des finances, des tarifications etc..[xvii].

Claude Hanclot, cite dans ce sens Pierre-Paul Pénillard, chef de projet du département de Saône-et-Loire (Hanclot, 2014) :

> La question de l'homogénéité des données est un leurre. Au conseil général de Saône-et-Loire il y a différentes missions. Il y a 100 progiciels. Il n'y a pas un système d'information unique capable de gérer tout ça. C'est un leurre de vouloir tout homogénéiser. Nous avons un projet autour de l'eau, mais nous n'avons pas une vision complète. Il faut avoir suffisamment de données pour pouvoir y travailler dessus. Concernant le tourisme au niveau départemental nous avons 60 % des données.[xviii]

La question d'un ensemble cohérent de données est donc au centre de toutes les préoccupations, notamment au niveau d'Etalab. On se retrouve ainsi sur la question de l'interopérabilité des données, de leurs utilisations et de leurs enjeux.

3 .2. Importance des Métadonnées et traitement des données

Selon l'article de Wikipédia (Wikipédia, 2015) :

> Les métadonnées correspondent à des marqueurs que l'on introduit dans les fichiers ou dans des langages de programmation appropriés, les langages de marquage XML. Les marqueurs ont pour effet d'améliorer l'efficacité des recherches d'information par rapport aux recherches plein texte. Le format RDF (Resource Description Framework) crée les conditions d'interopérabilité, avec des réseaux de métadonnées, et l'utilisation du langage XML. Les ressources numériques balisées transportent avec elles leurs propres métadonnées lorsqu'elles sont téléchargées, copiées, répliquées ou transmises par des messageries électroniques. Ceci s'applique à tous les types de ressources numériques (texte, son, image, multimédia).

Les métadonnées constituent ainsi l'un des principaux éléments de ce que l'on a appelé le web sémantique et elles permettent de croiser différentes sortes de données cohérentes en reliant par exemple la position spatiale et la temporalité.

> Le potentiel des métadonnées est beaucoup plus important, car elles peuvent faire inter opérer les ressources informatiques, dans la mesure où elles ont été paramétrées et structurées dans des dictionnaires de données (ou registres de métadonnées). On peut alors faire communiquer les bases de données classiques, utilisées dans les progiciels de gestion intégrés) et les données non structurées (documents, images, manipulés en gestion des connaissances…).

Le passage de données brutes au moyen de l'adjonction de métadonnées servant à leur interopérabilité est donc la plupart du temps un passage obligé permettant de traiter et de combiner des données de types différentes (par exemple arrêts, lignes de bus, horaires..). Cela demande bien entendu d'avoir à disposition des logiciels spécifiques pouvant traiter ces données - on emploie alors le terme d'analyse de données qui renvoie à des algorithmes très spécialisés et précis rendant possible l'agrégation des données mais impliquant une grande maîtrise des mathématiques combinatoires.

3.3 Approche statistique du traitement des données

Si le traitement des données s'est très tôt intéressé aux traitements statistiques cela ne relève pas du hasard mais de la nécessité. En effet les statistiques reposant sur la loi des grands nombres - (de données), elles s'intéressent plus particulièrement aux phénomènes de variabilité et de leurs mesures en terme de probabilité. La statistique fournit des outils pour

étudier, entre autres, des phénomènes aux données variables; au cœur de la statistique se trouve alors le concept de variabilité et de corrélation. De plus de nombreuses méthodes statistiques permettent de visualiser les résultats sous forme de graphique les résultats, pour mieux visualiser et analyser de nombreux éléments en les regroupant autour de schémas permettant de mieux évaluer leur répartition statistiques. Parmi les méthodes les plus utilisées et les plus connues on peut citer la droite de régression linéaire, les méthodes vectorielles (analyse en composante principale, analyse factorielle de correspondance, analyse de correspondance multiple, analyse factorielle de données brutes, analyse factorielle multiple (ACP, AFC, ACM, AFDM, AFM).

Ces méthodes permettent à la fois de travailler sur un grand nombre de données, d'étudier les corrélations possibles entre deux ou plusieurs variables et de mieux comprendre de façon visuelle la façon dont les variables peuvent (ou non) interagir entre elles. Les données pourront être cartographiées dans un espace mathématique à deux ou trois dimensions (2D et 3D) et constitueront la base d'une représentation dynamique de leur ensemble , notamment dans le cas de phénomènes soumis à des fluctuations dynamiques dans le temps (ex : circulation automobile, météorologie, connexions sur intranet etc..).

La droite de régression linéaire permet d'une part de déterminer s'il y a une forte corrélation entre deux variables et d'autre part de déduire l'équation moyenne de la droite. Ainsi, plus les points seront alignés autour de la droite, plus la corrélation sera parfaite, mais au contraire, plus les points seront dispersés, plus la corrélation deviendra faible ou nulle. Si la droite monte il y a corrélation positive (effet du plus qui mène au plus), si la droite descend, nous sommes en présence d'une corrélation négative (effet du moins qui mène au plus). La fluctuation des corrélations entre deux variables peut donc être observée de cette manière.

Les méthodes qui dérivent de l'analyse vectorielle, (ACP, AFC, ACM, AFDM, AFM) organisent les données statistiques sous la forme de vecteurs dans l'espace. Cela permet de représenter l'ensemble des données sous la forme de nuages de points vectorisés dans l'espace et de visualiser s'il y a corrélation positive, négative ou neutre. Les représentations peuvent varier au moyen d'un graphique organisé selon deux axes ou constituant un cercle . L'intérêt principal de ces méthodes réside dans le fait qu'elles permettent à la fois de visualiser les données sous formes d'espace et de relation mathématique et d'en figurer leur corrélation. Ainsi les variables appartenant à des régions d'un même espace sont soit corrélées de façon positives (attraction, dans la plupart des cas), soit de façon négative (opposition de leur position dans l'espace) soit ne sont pas corrélées (ni attraction, ni répulsion, lorsqu'elles forment un angle droit). L'analyse en termes de corrélations des variables dans l'espace statistique prend alors une forme visuelle et peut considérablement faciliter l'interprétation, une fois ces a-priori intégré. Nous renvoyons le lecteur à la définition et aux illustrations de chacune de ces catégories que l'on peut trouver sur Wikipédia[xix] mais également dans de nombreux dicta ciels statistiques présents sur le net

Toutefois si ces différentes catégories sont les plus connues il en existe d'autres, moins connues du grand public mais citées sur Wikipédia :

- L'analyse en composantes indépendantes ; les cartes auto-adaptatives (SOM, *self organizing maps* en anglais), appelées aussi cartes de Kohonen employées notamment en cartographie spatiale, l'analyse en composantes curvilignes, la compression par ondelettes, l'analyse de séries dynamiques d'images
- A noter : l'ACP, désignée en général dans le milieu du traitement du signal et de l'analyse d'images plutôt sous son nom de Transformée de Karhunen-Loève (TKL)

> est utilisée pour analyser les séries dynamiques d'images, c'est-à-dire une succession d'images représentant la cartographie d'une grandeur physique.

Le traitement des données est donc profondément tributaire de modèles mathématiques spécifiques et les processus multiples de visualisation de ces données sont ~~bien sur~~ toujours associés à des modèles mathématiques extrêmement élaborés et spécifiques à un champ donné, à l'exemple de la TKL mentionné par Wikipédia dans le domaine de la topographie (Wikipédia, 2015).

> De la même manière, la TKL permet de mettre en évidence des cinétiques différentes lors de l'analyse topographique dynamique, c'est-à-dire l'analyse de l'évolution du relief au cours du temps. Elle permet alors de déceler des phénomènes invisibles par simple observation visuelle, mais se distinguant par une cinétique légèrement différente (par exemple pollution d'une surface rugueuse par un dépôt).

3.5. L'algorithmie

Selon Wikipédia (2015) :

> « Un algorithme est un processus systématique de résolution, par le calcul, d'un problème permettant de présenter les étapes vers le résultat à une autre personne physique (un autre humain) ou virtuelle (un calculateur).
>
> En d'autres termes, un algorithme est un énoncé d'une suite finie et non-ambiguë d'opérations permettant de donner la réponse à un problème. Il décrit formellement une procédure concrète. »

Le portail consacré aux algorithmes[xx] sur Wikipédia en définit plusieurs grands groupes, mais de manière non exhaustive : algorithmes de structures de données, algorithmes de tris, algorithmes de la théorie des graphes, algorithmes géométriques, algorithmes mathématiques, Union-find, algorithmes de balayage,

En matière de gestion de données, les algorithmes de tri sont très utilisés et sont définis comme suit par Wikipédia (2015) :

> « Un algorithme de tri est, en informatique ou en mathématiques, un algorithme qui permet d'organiser une collection d'objets selon un ordre déterminé. Les objets à trier font donc partie d'un ensemble muni d'une relation d'ordre (de manière générale un ordre total). Les ordres les plus utilisés sont l'ordre numérique et l'ordre lexicographique (dictionnaire). »

Les métadonnées descriptives dont on a évoqué plus haut l'importance pour les questions d'interopérabilté et de web sémantique constituent des variables spécifiques sur lequel l'ordinateur et les programmes pourront ensuite agir en appliquant des algorithmes bien spécifiques pour pouvoir les traiter. D'une certaine façon ces métadonnées fonctionnent comme un registre d'adresse, dans lequel on pourra chercher pour identifier la donnée. Ainsi l'adresse URL par exemple doit être considérée comme une métadonnée qui sera (ensuite) traitée de façon algorithmique afin d'identifier la page pour pouvoir la localiser et l'ouvrir. Les algorithmes sont donc à la du processus de traitement des données et constituent ainsi les clés logiques du traitement des données . Ils constituent aussi le socle de la démarche de ~~la~~ recherche d'informations.

Enfin de nombreux algorithmes sont liés à la théorie des graphes et interviennent plus spécifiquement dans la cartographie des réseaux et des relations et bien sûr dans le traitement mathématique et statistique de l'ensemble. Nous renvoyons le lecteur au portail consacré aux algorithmes pour approfondir le sujet, portail dans lequel sont abordés entre autres de nombreuses catégories : algorithme récursif, réparti, émergent, adaptatif, métaheuristique, d'approximation, évolutionniste, génétique, mémétique, de tri, de recherche de chemin, algorithme de la théorie des graphes, de recherche de chemin….. qui pourront être utilisés pour des besoins très spécifiques de traitement des données, en fonction de la nature des problématiques considérées.

On pourra à titre d'exemple consulter l'algorithme A de recherche de chemin : comment aller dans le moins d'un point A à un point B. Les algorithmes de chemin (en anglais pathfinding) sont notamment très utilisés pour établir un trajet de bus ou métro le plus court et le plus économique possible.

(Source https://fr.wikipedia.org/wiki/Recherche_de_chemin)

3.6. La fouille de données - ou le data mining

Comme le rappelle le site du FREMIT, (Structure Fédérative de Recherche en Mathématiques et en Informatique de Toulouse)[xxi] (FREMIT, 2015) :

> « L'extraction de connaissances par l'exploration ou la fouille de données (data mining) est, depuis son émergence dans les années 90, un domaine dans lequel l'interaction entre mathématiciens (en particulier statisticiens) et informaticiens (en particulier spécialistes des bases de données) est forte et primordiale […./....] La fouille de grandes masses de données fait face à plusieurs verrous majeurs : l'intégration des données, leur dimensionnalité, leur hétérogénéité, la gestion de la temporalité et des contextes. L'intégration de données fait tout d'abord référence à l'hétérogénéité des données mais également à leur distribution. L'homogénéisation nécessite alors des mécanismes sophistiqués pour rechercher l'information (indexation contextuelle, text mining, extraction des informations utiles), les structurer (élicitation de structure), les lier entre elles et de les stocker de façon pertinente (datamart, XML). La dimensionnalité des données fait non seulement référence au problème de volume mais également à l'aspect multi-points de vue ou multi-facettes des données. Les visualisations pour l'interprétation associées restent mal adaptées aux volumes (quelques dizaines d'éléments au maximum). La temporalité des données fait référence d'une part au fait que certaines données ont des durées de validité variables et d'autre part que les données sont dynamiques. Cette temporalité doit être prise en compte à la fois au niveau des représentations des données, au niveau des algorithmes adaptatifs mais également au niveau des représentations visuelles associées. »

Wikipédia -décidément notre encyclopédie universelle - définit ainsi la fouille de données (Wikipédia, 2015) :

> « L'exploration de données, connue aussi sous l'expression de fouille de données, forage de données, prospection de données, *data mining*, ou encore extraction de connaissances à partir de données, a pour objet l'extraction d'un savoir ou d'une connaissance à partir de grandes quantités de données, par des méthodes automatiques ou semi-automatiques.
>
> Elle se propose d'utiliser un ensemble d'algorithmes issus de disciplines scientifiques diverses telles que les statistiques, l'intelligence artificielle ou l'informatique, pour construire des modèles à partir des données, c'est-à-dire trouver des structures

intéressantes ou des motifs selon des critères fixés au préalable, et d'en extraire un maximum de connaissances utiles à l'entreprise.[…/…] C'est aussi le mode de travail du journalisme de données[1]. »

L'article distingue cinq grands types de fouilles (Wikipédia, 2015) :

- La fouille de flots de données (*data stream mining*) qui consiste à explorer les données qui arrivent en un flot continu, […/….],
- La fouille audio, technique récente, parfois apparentée à la fouille de données, permet de reconnaître des sons dans un flux audio. Elle sert principalement dans le domaine de la reconnaissance vocale et/ou s'appuie sur elle,
- La fouille d'images est quant à elle la technique qui s'intéresse au contenu de l'image. Elle extrait des caractéristiques dans un ensemble d'images, par exemple du web, pour les classer, les regrouper par type ou bien pour reconnaître des formes dans une image dans le but de chercher des copies de cette image ou de détecter un objet particulier, par exemple.
- La fouille de textes ou l'exploration des textes en vue d'en extraire une connaissance de haute qualité : cette technique est souvent désignée sous l'anglicisme *text mining*. C'est un ensemble de traitements informatiques consistant à extraire des connaissances selon un critère de nouveauté ou de similarité, dans des textes produits par des humains pour des humains. Dans la pratique, cela revient à mettre en algorithmes un modèle simplifié des théories linguistiques dans des systèmes informatiques d'apprentissage et de statistiques. Les disciplines impliquées sont donc la linguistique calculatoire, l'ingénierie du langage, l'apprentissage artificiel, les statistiques et l'informatique.[xxii]

Nous sommes, avec les techniques de « data minding », au cœur des logiques d'exploration des moteurs de recherche et de leurs capacités à retrouver de l'information à partir d'immenses quantité de données. Tout le développement de Google s'est construit sur cette expertise.

A titre d'exemple le lecteur pourra se reporter au graphique des termes des discours inauguraux des présidents des Etats Unis par Santiago Ortiz en 2011. Source : http://intuitionanalytics.com/other/inauguralSpeeches/

3.7. La visualisation des données ou Dataviz

3.7.1 La data visualisation, une étape du processus de production

Pour beaucoup de chercheurs aujourd'hui, la « data visualisation » fait partie du processus de la science des données, en tant que dimension finale, liée aux autres étapes de construction, d'exploration et d'analyse. Nous avons vu préalablement comment les modèles mathématiques et statistiques basés sur les algorithmes de traitement de données enrichissait et synthétisait les données. L'objectif de la data visualisation sera donc de fournir des clés de lecture accessibles à la compréhension humaine, différentes d'une longue suite de codes de 0 et de 1 caractéristique du langage informatique, et évitant les longues données de nombres, de chiffres et de mesures, inintéressantes et fastidieuses si aucun travail d'organisation statistiques, heuristiques ou algorithmiques n'y étaient appliqué.

Dans le schéma consacré à la data science process, la data visualisation fait partie intégrante de la chaîne de production des données. Les étapes pour la mise en place d'un processus scientifique de mise en place des données sont en effet les suivantes : collecte des données brutes, implémentation en métadonnées, nettoyage et vérification des données, mise en place de modèles mathématiques et d'algorithmes pour le traitement des données, exploration et

analyse des données (data mining), production des données (publication) et data visualisation afin de permettre l'analyse et la prise de décision[xxiii].

Comme le rappelle Marina Palakartcheva (2015) qui s'est spécialisée sur les questions de visualisation coopérative :

> « Le projet de l'intelligence artificielle qui visait à substituer l'humain par la machine dans ses fonctions cognitives n'a pas donné les résultats escomptés. Cela nous ramène au paradigme de l'augmentation de la cognition humaine dans le sens d'Engelbart (1962). C'est dans ce contexte que la visualisation de l'information, qui vise à en optimiser l'appréhension et la compréhension, en sollicitant les compétences usuelles de la perception spatiale, a été développée. Il s'agit de fournir des représentations visuelles et interactives de données abstraites pour amplifier la cognition et faciliter la prise de décision[xxiv] ».

Pour cet auteur, la visualisation désigne la construction de modèles mentaux qui amplifient la cognition humaine afin de rendre d'une part l'information abstraite compréhensible et de permettre d'autre part de synthétiser l'information collectée à partir de multiples sources de données hétérogènes.

Outre l'aspect statistique des représentations des données déclinées plus haut, de nouveaux enjeux sont liés à leurs présentations, notamment en 3D. Comme le fait remarquer M.Palakartcheva, plus la complexité augmente, plus les besoins de présentation en trois dimensions sont élevés. La 3D permet en effet de visualiser la complexité et la spécificité des données car elle autorise leur répartition et leur différenciation dans l'espace.

Les adresses suivantes ouvrent l'accès à de nombreux exemples de visualisation en 3D de la dispersion de quatre variables :

https://en.wikipedia.org/wiki/Glyph_(data_visualization)#/media/File:Scatter_plot.jpg ou encore la visualisation en 3D d'une carte géographique (Perspective Landsat TM RGB visualization) à la page suivante : grass.osgeo.org

3.7.2 L'Analyse des réseaux

La visualisation est aussi particulièrement adaptée dans le cas de l'analyse des réseaux et des graphes, qu'il s'agisse de nuages de mots et de fouilles textuelles vus ci-dessus, ou bien de la visualisation de flux et de connections.

Pour mémoire, voici quelques adresses à consulter pour la visualisation de l'analyse des réseaux :

- Cartographie du réseau NICHE en avril 2001, 340 suiveurs sur le réseau social twitter par le chercheur William J. Turkel. La data visualisation et la méthode des graphes est particulièrement pertinente pour l'analyse des réseaux de communication. Source : http://williamjturkel.net/2011/08/02/social-network-analysis-and-visualization/

- Aperçu du projet « Metropolitain », un projet de visualisation de données urbaines développé par les équipes de Dataveyes. Un projet centré sur la réutilisation des données (temps de transport) par exemple) et des données ouvertes par la RATP, pour l'affluence de voyageurs. Source http://frenchweb.fr/datavisualisation-zoom-sur-5-start-upsfrancaises/116061#umRPT4dBGt2z1Hl7.99

- Cartographie des nationalités de membres ayant inscrit le terme « python » sur leur profil twitter. Source : http://giladlotan.com/wp-content/uploads/2012/11/Screen-Shot-2012-12-06-at-1.19.15-PM.png

Pour montrer tout l'intérêt que permet la data visualisation dès que l'on a à faire à des données susceptibles d'être représenté sous la forme de graphes de connexions (réseaux).

3.7.3 La visualisation des flux dynamique de données

Enfin, les techniques de visualisation permettent de plus en plus de traiter les flux dynamiques de données en "temps réel" en actualisant sans cesse les représentations. Quantité de données, flux, vitesse, actualisation deviennent ainsi les nouvelles contraintes de la data visualisation.

Ainsi, pour exemple, on pourra consulter le site « meteorage.fr » pour suivre en direct la carte dynamique des orages en temps réel sur le territoire français, ainsi qu'une animation sur les derniers 24 h. Bien entendu, dès lors que l'on s'oriente vers ce type de traitement dynamique cela présuppose également tout un dispositif de captation permanent de données (vidéos, capteurs atmosphériques, capteur de pollution, de trafic etc.) préfigurant déjà ce que pourra devenir le web des objets connectés et leurs applications dans le domaine du transport de la santé, de l'économie, de la domotique, de la surveillance de la pollution, des flux etc…

La visualisation permet de jouer également sur le niveau de détail recherché (exemple google map qui permet de zoomer plus ou moins sur une carte et de focaliser ainsi la perception de l'échelle). Elle rend aussi possible la démultiplication du réarrangement de l'écran en ayant accès à d'autres liens ou schéma (navigation). Enfin, elle autorise la mise en place des systèmes interactifs par l'utilisateur.

La question principale devient alors (Palakartcheva, 2015) :

> « Comment adapter les applications de visualisation de l'information au nouveau contexte des environnements collaboratifs des médias numériques et ce dans un contexte de développement extrêmement rapide des technologies de l'information et d'une quantité de données à la disposition des utilisateurs s'accroissant de façon exponentielle ».

La collaboration constitue de ce fait un des grands défis pour la visualisation de l'information et l'analytique visuelle et s'ouvre sur un nouveau champ d'applications logicielles permettant la coopération tels que (Palakartcheva, 2015) :

> Vibrela, OpenQuaq, OpenSpace3D, MindJet, iMindMap, FreeMind, Creately, InVisionApp, Deekit, ReviewStudio, ActiveColab, TeamLab Office, TeamWorkLive, Tableau, IBM Cognos, MicroStrategy, TIBCO Spotfire, VizQL - Natively visual, naturally faster, Live Query Engine - Data in your hands, In-memory Data Engine - Blazing speed qui permet l'analyse rapide de grandes bases de données, Tableau Public - MindManager etc…

Ainsi nous pouvons dire que nous entrons, avec toutes ces interfaces visuelles ouvrant sur l'exploration et le partage d'immenses quantités de données, dans l'ère de l'immersivité, pour reprendre ce terme cité à la fois par M. Palakartcheva et le site disko.fr centré sur la thématique de la data visualisation.

En guise de Conclusion provisoire : vers l'émergence d'une data science

Nous ne sommes qu'au tout début d'une mise à disposition massifiée de données ouvertes dans tout les domaines, y compris dans le registre particulier de la data visualisation, domaine destiné à rechercher l'appréhension visuelle et compréhensive d'immenses flux de données. Cela peut être considéré comme l'une des caractéristiques essentielles de l'émergence d'une ère nouvelle dans la connaissance de l'humanité que certains ont appelés société de l'information et de la connaissance.

Au-delà d'une simple mise à disposition des données publiques, l'enjeu est celui d'une profonde révolution de la connaissance par l'irruption massive de quantité de données qu'il va falloir traiter, analyser, comparer et faire circuler mais surtout explorer.

Nous avons essayé de montrer dans cet article comment, bien loin de l'idée centrée sur la dimension d'une transparence absolue des données, celles-ci constituent en fait le maillage cognitif plus poussé de nos perceptions, ainsi que de l'analyse et de la compréhension de nos environnements. Le processus de l'analyse et du traitement des données est essentiellement un dispositif de construction du réel par l'augmentation de notre perception et de notre connaissance, grâce aux dispositifs informatiques permettant de traiter, arranger, et combiner ces données, au même titre que le microscope ou le télescope à leurs époques.

Aujourd'hui avec l'arrivée de l'open data, c'est à une autre dimension à laquelle nous nous voyons confronté : celle de l'infiniment complexe et de l'infiniment nombreux, celle de l'hyper explosion des données à appréhender. Un univers nouveau de perception s'ouvre ainsi à nous, regroupant l'univers de la data et celui de l'information. Un univers qu'il va aussi falloir apprendre à explorer, décrypter, lire mais aussi écrire et produire. Aussi dans ce nouvel univers aurons-nous besoin de nouvelles boussoles pour nous orienter et nous donner le cap. Cependant, la motivation profonde restera l'exploration, dans le sens où elle reste liée à la connaissance. Mais l'exploration et la connaissance ne sont-ils pas des moteurs fondamentaux et éternels de la psyché humaine et dans ce sens le phénomène historique et social de l'open data constituerait un continuum somme toute logique..

Enfin, le dernier point important à souligner à l'issue de ce travail, c'est l'incontestable émergence d'une « data science » *inscrivant l'open data en tant qu'étape primordiale de son évolution.*

Ce n'est pas un hasard si Wikipédia cite cette nouvelle science selon ces termes (Wikipédia, 2015) :

> « La science des données (en anglais *data science*[3]) est une nouvelle discipline qui s'appuie sur des outils mathématiques, de statistiques, d'informatique (cette science est principalement une *« science des données numériques »* et de visualisation des données. Elle est en plein développement, dans le monde universitaire ainsi que dans le secteur privé et le secteur public. »

Et si le même article fait référence à la mission Etalab pour évoquer la présence des premiers « data scientist » de l'histoire (Wikipédia, 2015) :

> « Cette science s'inscrit dans les efforts d'accompagnement du numérique, en lien depuis qu'elle existe avec la mission Etalab (dont le directeur, Henri Verdier, est aussi « administrateur général des données de l'État », assisté par des data-scientists recrutés pour pour « accélérer la possibilité de politiques publiques « augmentées » par les données et leur analyse».

Les futurs développements de l'open data se situent clairement à l'intersection de toutes ces logiques que nous n'avons cessé de croiser pour sa compréhension même et dont la logique centrale est bien celle d'une véritable science de l'administration publique et de la gouvernance des données, de leur élaboration et de leur diffusion dans un espace public plus

en plus numérique. La « data science » peut donc être considérée comme une science complexe, située à l'intersection de nombreux domaines : ingénierie de la donnée, méthode scientifique, mathématiques, statistiques, informatique, visualisation, expertise mais aussi gouvernance, management de projet, systèmes coopératifs et politiques publiques[xxv].

"L'avenir est ouvert" disait Popper et il semble bien qu'il soit devenu ouvert pour l'open data…

[i] Source www.data .gov consulté en juillet 2014
[ii] Article Wikipédia sur les données ouvertes
[iii] Ibid. note III
[iv] Source : article Wikipédia sur les métadonnées https://fr.wikipedia.org/wiki/M%C3%A9tadonn%C3%A9e
[v] Source : Guide du big data 2013/2014, l'annuaire de référence à destination des utilisateurs, p.5
[vi] Source : article Wikipédia base de données - https://fr.wikipedia.org/wiki/Base_de_donn%C3%A9es
[vii] Simon Chignard, open data, ed.Fyp, 2012, pp. 60-64
[viii] Claude HANCLOT, Développement et contraintes de l'open data dans les collectivités territoriales, Mémoire de Master 2 VAE Mention infirmation et documentation, université Jean Moulin Lyon 32014 pp. 61
[ix] https://libertic.wordpress.com/
[x] Etude Delloite, Open data 44, évaluation de la démarche, Juin 2014, p 10.disponible à l'adresse suivante : http://www.loire-atlantique.fr/upload/docs/application/pdf/2014-07/deloitte-opendata44-rapport-vfinal.pdf
[xi] Source : https://libertic.wordpress.com/
[xii] Claude Hanclot, ibid., P. 60
[xiii] Les conférences de l'open data en Rhône Alpes, http://lesmatineesdelopendata.fr
[xiv] Catherine Bouteiller, Laboratoire d'Economie des transports Lyon 2, conférence donnée le 20 décembre 2013 dans le cadre des conférences MODRA - http://lesmatineesdelopendata.fr
[xv] CEREMA, l'open data en collectivité à la lumière des données de mobilité, collection connaissance, mars 2015, p. 20
[xvi] CEREMA, l'open data en collectivité à la lumière des données de mobilité, collection connaissance, mars 2015, p. 36
[xvii] Cf. p. 47 et 48 du rapport CEREMA
[xviii] Entretien du 26 août 2013, cité par Claude HANCLOT, Développement et contraintes de l'open data dans les collectivités territoriales, Mémoire de Master 2 VAE Mention information et documentation, université Jean Moulin Lyon 3, 2014 p.65.
[xix] https://fr.wikipedia.org
[xx] https://fr.wikipedia.org
[xxi] Source : http://www.irit.fr/FREMIT/Federation/fouille.html
[xxii] Pour approfondir la fouille de texte voir le diaporama de Philippe Gambette à ce sujet http://fr.slideshare.net/PhilippeGambette/analyse-de-textes-avec-treecloud-et-lexico3
[xxii] Place de la data visualisation dans le data science progress. Source : https://commons.wikimedia.org/wiki/File:Data_visualization_process_v1.png
[xxii] cf. Le schéma de la data science. Source : https://upload.wikimedia.org/wikipedia/commons/4/44/DataScienceDisciplines.png

Bibliographie et webographie

CEREMA, l'open data en collectivité à la lumière des données de mobilité, collection connaissance, mars 2015.
CHIGNARD Simon (2012), l'open data, comprendre l'ouvertur des données publiques, Paris, Fyp éditions.
DENIS Jérôme , GOETA Samuel .(2014), Exploration, Extraction and 'Rawification'. The Shaping of Transparency in The Back Rooms of Open Data *After The Reveal. Open Questions on Closed Systems - communication au congrès Neil Postman Graduate Conference*, Feb 2014, New York, United States, https://hal.archives-ouvertes.fr/halshs-00954302v1
E- goverrnemnt survey, (1992) , rapport des nations unies 1992.
http://www.unpan.org/egovkb/global_reports/08report.htm
Etude Delloite, Open data 44, évaluation de la démarche, Juin 2014, p 10.disponible à l'adresse suivante : http://www.loire-atlantique.fr/upload/docs/application/pdf/2014-07/deloitte-opendata44-rapport-vfinal.pdf

HANCLOT Claude, Développement et contraintes de l'open data dans les collectivités territoriales, Mémoire de Master 2 VAE Mention infirmation et documentation, université Jean Moulin Lyon 3, 2014
Livre blanc sur l'open data (2015) - http://bluenove.com/livres-blancs/open-data/.
Guide du big data 2013/2014, l'annuaire de référence à destination des utilisateurs
PALAKARTCHEVA Marina (2015), Les outils de visualisation de l'information : évolution des enjeux et des problématiques, communication au colloque "nouveaux enjeux de l'ère numérique", Université Lyon 2, 17 et 18 juin 2015.
VAN CUYCK Alain, (2013). L''open data comme nouvelle forme de gouvernance numérique : enjeux, marchés, modèles, idéologies, communication XIX° colloque international franco-roumain " Culture et Responsabilité sociale dans la communication des organisations " https://hal.archives-ouvertes.fr/hal-00963337

Wikipédia (2015). Données ouvertes, https://fr.wikipedia.org/wiki/Open_data, consulté le 20 novembre 2015.
Wikipédia (2015). Exploration de données, https://fr.wikipedia.org/wiki/Exploration_de_donn%C3%A9es, consulté le 20 novembre 2015.
Wikipédia, (2015). Science des données, https://fr.wikipedia.org/wiki/Science_des_données, consulté le 20 novembre 2015.
Wikipédia, (2015). Algorithmes**,** https://fr.wikipedia.org/wiki/Algorithme, consulté le 20 novembre 2015.
Wikipédia, (2015). Métadonnées, https://fr.wikipedia.org/wiki/Métadonnée , consulté le 20 novembre 2015.
Wikipédia, (2015). Data visualization, https://en.wikipedia.org/wiki/Data_visualization, consulté le 20 novembre 2015.

Plateformes nationales de données Open Data : France : data.gouv.org, USA : data.gov, Angleterre data.gov.uk, Australie : data.gov.au, Nouvelle Zélande : cat-open.org.nz
Site de la mission ETALAB : https://www.etalab.gouv.fr/
Plateformes régionales : liste des principales plate formes : https://opendata.hauts-de-seine.net/comprendre/demarches/le-point-sur-les-initiatives-liste-des-plateformes-des-collectivites-francaises
https://libertic.wordpress.com, site de l'association pour la liberté des données
www.regardscitoyens.org/open-data-en-France - association pour le partage de l'information politique
Site de l'association des collectivités territoriales engagées dans le mouvement open data
http://www.opendatafrance.net/
Les conférences de l'open data en Rhône Alpes, http://lesmatineesdelopendata.fr
www.datavisualization.fr/ site consacré à la visualisation des données en France
http://www.disko.fr site d'une agence consacrée à la data visualisation
Site de William J. Turkel sur la data visualisation des réseaux de communication
http://williamjturkel.net/2011/08/02/social-network-analysis-and-visualization/
Blog consacré à la data visualisation, l'infographie et les statistiques (en anglais) : FlowingData
Site communautaire et d'information autour de la data visualisation : visualizing.orgtutoriel sur les algorithmes : http://tcuvelier.developpez.com/tutoriels/algo/introduction-algorithmes-structures-donnees/#LI-B
Diaporama sur la fouille de données : http://fr.slideshare.net/PhilippeGambette/analyse-de-textes-avec-treecloud-et-lexico3
Site du FREMIT sur la fouille de données : http://www.irit.fr/FREMIT/Federation/fouille.html
Suristat**,** portail des enquêtes et de l'analyse des données : http://www.suristat.org
Site sur le journalisme des données de l'université de standford : http://datajournalism.stanford.edu/
Place de la datavisualisation dans le data science progress. Source :
https://commons.wikimedia.org/wiki/File:Data_visualization_process_v1.png