Didak'TIC

Convergence des cultures de l'information

n°1

2017

BIG DATA

Une revue proposée par le Master 2 Management et Valorisation de l'Information Numérique (MAVINUM) Promotion 2016-17

71ème section CNU : sciences de l'information et de la communication (SIC)

information — ITIC — & documentation

MAVINUM

MASTER

MASTER2

SCIENCES HUMAINES & SOCIALES

SOMMAIRE

ÉDITORIAL 1

ACTUALITÉS 4

LE DOSSIER BIG DATA (PRÉSENTATION) 5

PRÉLIMINAIRES : UNE JOURNÉE AVEC LES BIG DATA ! 6

Technologie :
VEILLE 11
DONNÉES DE LA RECHERCHE 14
ARCHIVAGE DU WEB 19
BUSINESS INTELLIGENCE 23
OURAGAN ANNONCÉ DANS LES SI 27
DATASTORELLING 31
Juridique :
GOUVERNANCE DES DONNÉES 36
Économie :
CROWDFUNDING ET RÔLE DES BIG DATA 41
DATA MARKETING 45
Social :
INTELLIGENCE COLLECTIVE 49
EN MATIÈRE D'ÉLECTION, TOUT EST PERMIS... OU PRESQUE 52
Environnement :
BIG DATA ET ENVIRONNEMENT 56

EN CONCLUSION, LES AMBIGUÏTÉS DU BIG DATA 60

RENCONTRE AVEC DES PRO... ILS VOUS EN DISENT PLUS !
VINCENT MORENO 67
THIBAUT BRITZ 70
LUCIE ALBARET 73
SANDRA BRINGAY 75
OLIVIER RAJZMAN 78
...un juriste anonyme 80
PIERRE DELORT 82
PIERRE SAVARY DE BEAUREGARD 85

FICHE PRATIQUE 88
BIBLIOGRAPHIE 90
GLOSSAIRE 91

ÉDITO

Au delà des frontières temporelles, territoriales, culturelles et professionnelles : la construction de connaissances !

Cher monde de l'information,

Comment introduire un premier numéro ?... Nous réfléchissons en séance de visio, nous, les copilotes du projet, autour de cet exercice qui pourrait paraître anodin après toutes les étapes franchies : écrire un éditorial. Et là, la lumière : quel a été notre postulat de base à l'initiative de ce magazine ? Le Big Data... Ces deux mots que beaucoup peinent à comprendre. Ces deux mots qui effraient certains. Ces deux mots que nous-mêmes avions du mal à expliquer SIMPLEMENT. Nombre de personnes clament d'emblée : "je ne comprends pas, c'est de l'informatique", "je ne comprends pas, c'est trop technique" "Big Data : kesako ?". Réaction légitime et humaine face à cette tempête qui engendre de vrais changements.

Mais de quoi s'agit il exactement ? Est-ce un effet de mode ? Un nouveau mode de calcul de probabilité ? Pourquoi s'y intéresser : quels enjeux, quelles opportunités ? Quels impacts dans notre quotidien, dans les entreprises et les organisations ? Quels nouveaux concepts accompagnent ce paradigme ? Datastorelling, crowdfunding, machine learning : tant de terminologies "barbares" qui rebutent le tout un chacun.

Dans ce monde complexe, la variété et la quantité d'informations, trop souvent mises à la disposition du public sous le seul angle technique, nous avons décidé d'amorcer un dispositif pouvant combler ce manque d'accompagnement au savoir. Bien des enjeux nous concernent tous pourtant et certains sont très accessibles !

A travers ce magazine, nous souhaitons, par conviction, apporter des réponses claires, simples et concises aux sujets complexes qui nous entourent.
C'est aussi, par passion de transmettre ce que nous avons appris tout au long des enseignements de master.
C'est également par obligation de réaliser un projet universitaire et, soyons honnêtes, viser une bonne évaluation ! :-)

Quoi qu'il en soit (ou quelles que soient nos motivations profondes), la mise en œuvre de ce projet a mis en évidence qu'un professionnel de l'information ne connaît plus de frontière, ni temporelle ni géographique ! De tout âge, de sensibilité différente quant à la fonction et aux métiers de l'information, des quatre coins de la planète, telle fut notre équipe.

Ce projet a su surmonter de nombreux challenges, bravant ainsi décalages horaires, vies professionnelles et familiales en parallèle. Il est une Superbe illustration de l'évolution de notre époque en terme de culture numérique sur les modes de travail et de gestion de projet 2.0.

Mais le challenge ne résidait pas seulement là. Il fallait également concrétiser un projet qui nous rassemble et nous ressemble. Alors quoi de mieux que de mettre en place un magazine et un site Internet à notre image et nous permettre pour une fois de laisser une trace concrète de notre collaboration !
Ce projet nous a apporté l'interdisciplinarité qui caractérise si bien les Sciences de l'Information et de la Communication. Cessons de voir le monde en silos et partageons nos connaissances !
Ce projet nous a fédéré, ce projet nous a rassemblé, ce projet nous a professionnalisé.

Alors partagez avec nous, ce fabuleux voyage ! Le premier d'une série que feront vivre les futures promotions des MAVINUM qui vous proposeront d'autres études de fond sur des sujets autour du Management et valorisation de l'information numérique.

■ Alexandre HUBERT, Béatrice TOFILI et Valérie GRIMAN

ILS ONT PARTICIPÉ

Firouze Betou
Paris, France

Hélène Kaizer
Thionville, France

Amélie Gallo
Lyon, France

Aurélie Quinidoz,
Grenoble, France

Valérie Griman
Dax, France

Isabelle Rouquette
Fort de France, La Martinique, France

Alexandre Hubert
Avignon, France

Camille Subtil
Colombo, Sri Lanka

Céline Jourdren
Paris, France

Béatrice Tofili
Païta, Nouvelle-Calédonie

Le magazine Didak'TIC
ITIC Bâtiment Marc Bloch (Bât. E) Université Paul-Valéry Montpellier 3 - Route de Mende 34199 Montpellier Cedex 5 - www.didaktik.fr
DÉPÔT LÉGAL : à parution.

Didak'TIC

DIRECTRICE DE LA PUBLICATION : Lise VERLAET

COORDINATION
Copilotes du magazine : Valérie GRIMAN, Alexandre HUBERT et Béatrice TOFILI.

RÉDACTION
Directeurs de la rédaction : Lise VERLAET et Hans DILLAERTS.
Rédacteurs : Firouze BETOU, Adèle DURET, Amélie GALLO, Valérie GRIMAN, Céline HOARAU-LEGOUX, Alexandre HUBERT, Céline JOURDREN, Hélène KAIZER, Aurélie QUINODOZ et Béatrice TOFILI.

CONCEPTION TECHNIQUE
Pour la partie PRINT : Chef technique : Jean-Michel DUBOIS. Maquettistes : Aurélie QUINODOZ, Amélie GALLO et Béatrice TOFILI.
Iconographes : Amélie GALLO, Laure RECHNER, Isabelle ROUQUETTE et Camille SUBTIL.
Pour la partie WEB : Chef technique : Mireille BACHELOT. Webistes : Firouze BETOU, Céline JOURDREN, Alexandre HUBERT et Béatrice TOFILI.

DIFFUSION Imprimé en France

Didak'TIC

Convergence des cultures de l'information

Sur le web : www.didaktic.fr

LES ACTUALITÉS DU BIG DATA

RÉGLEMENTATION

Les apports successifs de la Loi pour une République Numérique (LRN) et le Règlement Général européen sur la Protection des données (RGPD) qui sera applicable en mai 2018, imposent aux entreprises des projets conséquents en matière de conservation des données. Aujourd'hui seulement 10% des entreprises implantées dans l'Union Européenne seraient en mesure de faire face à ces mesures.

CHIFFRE

Tous les deux jours, l'humanité produit autant d'information que ce qu'elle a généré depuis l'aube de la civilisation jusqu'en 2003.

ALLIANCES

De nombreux médias français ont décidé de faire des alliances afin de mettre en commun leurs données pour mieux les commercialiser sur un marché publicitaire où le numérique a redistribué les cartes.

La première alliance « Gravity », annoncée le 4 juillet, rassemble 18 entreprises du monde de la presse et de l'audiovisuel figurent L'Equipe, SFR, NextRadioTV, FNAC-Darty, Marie-Claire, Sciences et Avenir, Condé Nast, Sud-Ouest, M6, Solocal, Prisma Média ou encore Challenges. Deux jours plus tard, les groupes Le Monde et Le Figaro ont annoncé un rapprochement stratégique à travers « Skyline ».
Cette logique d'alliance pour faire contrepoids aux GAFA s'observe aussi dans d'autres secteurs comme dans les transports où La SNCF, la RATP, Transdev et Blablacar ont décidé de mettre leurs informations de transport en commun.

NOUVEAUTÉ

L'ANSSI, l'agence nationale en charge de la cybersécurité, propose un MOOC accessible à tous dédié à la sécurité du numérique. Une excellente façon de s'initier ou d'en apprendre plus sur la cybersécurité afin d'agir sur la protection de nos outils numériques.

ÉVÈNEMENT

L'exposition « Terra Data, nos vies à l'ère du numérique », présentée du 4 avril 2017 au 7 janvier 2018 à la Cité des sciences et de l'industrie explique ce que sont les données dites *big data*. Par quels traitements peuvent-elles produire de la connaissance et comment elles agissent sur nos existences ?

Big Data, qui es-tu ?

Concept, technologie, artifice[1], le Big Data est de plus en plus présent dans notre quotidien même là où on s'y attend le moins.
Il impacte à la fois au niveau technologique, juridique, économique, social et en matière d'environnement.
Mais qu'est-ce que c'est ? A-t-il une histoire ?
Se cache-t-il pour autant partout ?
Et quel impact a-t-il dans notre usage quotidien ?
Tant de questions auxquelles les articles qui vont suivre essaieront d'apporter des éclairages.

[1] Brun, C. (2016).Big data et autre articife d'intelligence. *Les Echos.fr*, 19 avril 2016 , [en ligne] : https://www.lesechos.fr/idees-debats/cercle/cercle-156278-big-data-et-autre-artifice-dintelligence-1214881.php

Big Data et gouvernance des données : approche simple et non technique

■ Valérie Griman

Vous lisez le titre et vous froncez les sourcils ? Vous ne comprenez aucun des concepts énoncés ? Alors cet article est pour vous. Vous êtes un expert ? Gardez l'article sous le coude pour être capable d'illustrer simplement ce sujet à vos collègues ou autres. Essayons ensemble de rapprocher ces notions de notre « connu », des situations du quotidien, de VOTRE quotidien, quel que soit votre profil.

Le Big Data : c'est quoi alors exactement ?

Ce concept, apparu en 1997, signifie « Gros volume de données » ou « Big bang de l'information stockée »[1]. Il est souvent lié aux solutions informatiques de traitement et de stockage de données, d'où ce ressenti de « trop technique pour moi ! ».

Mais ce que tout le monde n'a pas intégré, c'est que notre vie de tous les jours est centrée et est guidée par ces mêmes données. Et oui ! Aujourd'hui « notre vie se transforme en données »[2]. Nous sommes les émetteurs des données Big Data

Prenons l'exemple d'une journée classique

Vous vous levez le matin et en prenant votre petit déjeuner vous consultez vos sms et mails. Tiens ! une alerte ! c'est l'anniversaire d'un de vos amis aujourd'hui. Vous vous notez de l'appeler dans la journée.

Vous consultez vos mails et vous pestez contre toutes les publicités reçues qui inondent votre messagerie. Vous faites le vide. Et, ah non ! tiens ! sympa cette photo des dernières Mizuno. Vous cliquez sur le lien pour aller voir les détails de l'offre.

Vous partez au bureau et vous connectez à nouveau à votre messagerie professionnelle. Le même exercice se répète. Vous avez reçu des offres de formation ou de prestation conseil pour vous accompagner dans vos métiers, noyées dans votre flux de veille et les mails de vos collègues.

Vous partez courir à votre pause déjeuner avec un de vos collègues. Vous activez votre montre cardio qui vous indique votre rythme cardiaque, votre temps chrono, votre distance parcourue et les calories dépensées, informations qui s'enregistrent dans votre historique des sorties.

L'après-midi vous partez en déplacement et activez votre GPS qui

vous conseille le meilleur itinéraire en fonction du trafic du jour. Vous prenez l'autoroute et vous voyez votre plaque d'immatriculation s'afficher sur un écran, vous indiquant de ralentir car vous êtes au-dessus des vitesses autorisées.

Le soir, vous vous connectez sur votre application musicale pour vous détendre. Cette dernière vous propose de nouvelles sorties d'albums, qui sont dans vos goûts habituels.

Vous vous connectez sur votre page Facebook et le site vous propose des lectures dans vos goûts, de nouveaux jeux à découvrir. Il vous incite à découvrir de nouvelles pages thématiques.

Mais comment est-ce possible ?

Il suffit d'entrer dans les coulisses de tous ces objets du quotidien dont nous sommes devenus dépendants : les smartphones, les ordinateurs, les tablettes, les montres connectées, les GPS, etc..., et toutes leurs applications.

Le Big Data « *nous observe, essaie de comprendre notre façon de vivre, nos habitudes*. »[3].

Il étudie tous nos actes « connectés » du quotidien, quels que soient les supports : nos navigations, nos comportements sur les réseaux sociaux, nos lectures pour anticiper nos comportements futurs. Nos objets de tous les jours deviennent une extension de notre vie physique, d'où la notion d'IoT : Internet of Things, l'Internet des Objets[5].

Toutes les informations, les recherches, les achats en ligne, les jeux, les pages consultées, nos profils sociaux complétés, les échanges sur nos « murs » se transforment en datas et viennent s'agglomérer à notre fameux « Big Data ».

Extrait de la vidéo « Big data : données, données, donnez-moi ! »[4]

En effet, toutes nos données générées sont regroupées, entassées, stockées dans des bases de données, des datacenters « dans les nuages ». Le stockage dans le cloud nous permet, de plus, un accès à nos informations quel que soit le support.

Ce flux d'informations lié à un individu, cumulé à celui de tout internaute, utilisateur connecté, organisations, entreprises, génèrent, tous les deux ans, dans le monde, autant de données que depuis le début de l'humanité. En 2020, seront décomptés « 50 milliards d'objets connectés capables de générer et d'échanger des données » (source data-business.fr). En 2014, ce nombre était de 12 milliards.

Le Big Data : quelles Opportunités et pour qui ?

Nous venons de voir que tous nos flux de navigations et toutes nos traces numériques, qu'elles soient personnelles ou professionnelles, sont stockées. Mais qui les exploitent ? Et à quelle fin ?

Cette marée noire informationnelle est valorisée et qualifiée de « nouvel or noir »[7] dans de nombreux secteurs d'activités qui ont vite compris les opportunités liées à cette masse d'informations : marketing, énergies, recherche et développement, formation, politique, sites de

Extraits de la vidéo « Big Data c'est quoi ? »[6]

rencontres, etc... d'autant plus que nous libérons toutes ces données GRATUITEMENT.

Ces données sont exploitées dans des « calculateurs », qui, par des algorithmes informatiques, font parler nos traces numériques. Ces machines analysent la répétition de certaines recherches, les mots récurrents dans nos écrits, nos « likes » sur les réseaux sociaux personnels et professionnels, les blogs et définissent notre profil. Ce profilage nourrit les industries ci-dessous.

Le Marketing

L'analyse de nos informations personnelles stockées dans les bases de données : nos achats en magasin, en ligne, nos navigations, nos « like » sur les réseaux sociaux donne aux industries Marketing une connaissance plus approfondie du client.

Le profil déterminé par le calculateur va permettre aux industries e-commerce de cibler les annonces et les mails que nous recevrons à notre prochaine connexion et tenter de nous influencer pour un prochain achat.

Les assurances

Les contrats d'assurance automobile : l'analyse de nos usages « transport » génère des profils type et donc détermine des contrats type avec des systèmes de bonus-malus plus appropriés à chaque assuré et ses bonnes ou mauvaises pratiques.
La maison intelligente : l'analyse de nos pratiques au quotidien donne aux assureurs l'opportunité de mieux cerner nos comportements et de créer différents profils de contrats d'assurances. Les capteurs positionnés dans la maison leur indiquent nos habitudes (bonnes ou mauvaises) quotidiennes, ce qui leur permet de faire des préconisations et d'ajuster leur offre aux pratiques.

Santé : certains assureurs pensent même inclure des capteurs de type bracelets connectés pour inciter à une meilleure hygiène de vie et tenter d'offrir « une planche de salut pour notre système de santé à bout de souffle »[1]. Les données collectées permettraient aux organismes de santé d'identifier des risques potentiels par rapport à nos routines alimentaires et loisirs, dans un objectif de prévention.

La Politique

Dans ce secteur, par exemple, l'analyse des comportements des indécis en situation d'élections est essentielle aux candidats et leurs équipes pour identifier les freins et déterminer et affiner les arguments pour faire pencher la balance.

Cette liste n'est bien évidemment pas exhaustive mais chacun se retrouvera dans les exemples cités pour une meilleure compréhension du sujet !

Extrait de la vidéo « Big data : données, données, donnez-moi ! »[4]

Les enjeux pour les entreprises : la gouvernance des données

Face à cette infobésité, cette multiplication des sources et des typologies des données internes et externes, les entreprises doivent s'organiser. Comme illustré ci-dessus, cette marée noire impacte les relations BtoB, BtoC et CtoC.

Le nouvel or noir doit être intégré dans la stratégie d'entreprise et doit devenir un outil de création de valeur. Il donne une connaissance plus affinée de l'environnement dans lequel tout organisme évolue et doit devenir un outil d'aide à la décision.

Extrait de la présentation HUPI CCI Bayonne

L'organisme ne doit garder de cette pluie torrentielle de données que l'eau vitale à son fonctionnement, son évolution et sa pérennité. Elle gagne ainsi en agilité afin de mieux s'adapter aux attentes de ses parties prenantes mais détecte, également, les « *signaux faibles* » pour innover et identifier de nouvelles opportunités.

L'entreprise doit chercher, filtrer, traiter les informations de façon à les transformer en connaissance stratégique selon un but : meilleure approche des marchés, évaluation de sa réputation... L'information doit être vulgarisée, simplifiée, résumée « *pour*

la rendre accessible et interprétable à toute personne qui en a besoin pour décider »[2].

Mais comment faire ?

La gouvernance des données

Il ne s'agit pas pour les entreprises de collecter des données à tout va mais de déterminer au préalable une stratégie, avec des objectifs, des méthodes, des règles d'usage des données, des outils et des responsables nommés. Ces personnes doivent avoir la compétence pour juger de la qualité, de la sûreté et de la fiabilité des informations collectées dans leur contexte, pour y apporter une valeur ajoutée, leur donner du sens. Les informations se managent.

Une approche systémique de l'entreprise et de ses métiers est essentielle. Identifier ses parties prenantes, ses risques et opportunités permet une définition plus précise du besoin des utilisateurs de l'information en interne.

Gouverner les données commence donc par la question du pourquoi et à quelle fin, dans un cadre précis. En effet, collecter et analyser les datas répondent à un besoin bien défini, comme vu dans les exemples traités précédemment en Marketing ou Politique.

Les organisations ont tendance à souvent faire l'impasse sur cette phase de réflexion. L'amalgame est fait entre Big Data et solutions informatiques. Mais comme pour tout choix d'outil, ce dernier doit avant tout répondre à un besoin, lui-même intégré dans la stratégie globale de l'entité. Ce n'est qu'après avoir défini sa stratégie que l'on peut réfléchir au comment et donc à l'outil informatique le plus adapté.

Cette solution doit prendre en compte le caractère hétéroclite des data, à structurer et à indexer en temps réel, grâce aux référentiels métiers et aux ontologies[10] et en tout respect de la propriété intellectuelle et de la vie privée.

Pourquoi hétéroclite ?

Le flux d'informations centralise des données de tout type dans une organisation :
- Des données non structurées telles que des documents Word, des fichiers .pdf,
- Des mails
- Des éléments de bases de données externes : Hadoop, MySQL
- Des ressources web : pages web, google analytics, portails publics institutionnels, journaux...
- Les flux de commentaires sur les blogs, les réseaux sociaux
- Des données centralisées sur des bases de données internes sur Excel, ACCESS ou outils spécifiques de GED (Gestion Electronique du Document) ou Intranet, données dites structurées par des métadonnées.

Autant de domaines et de sources à prendre en compte que de nouvelles compétences à intégrer dans les ressources d'un organisme.

Il s'agit là d'une véritable transformation du modèle économique habituel. Ce qui explique une majorité des freins identifiés actuellement dans les entreprises : nouvelles stratégies, évolutions organisationnelles et technologiques.

La gouvernance des données est une affaire de technologie, certes, mais surtout d'actions sur les hommes et les processus. Pour manager ce flux d'informations externes, l'entreprise doit d'abord cadrer son système d'informations interne. Elle doit sortir de sa vision informatique du sujet pour tendre vers une gouvernance de son patrimoine informationnel.

Le bémol

L'exploitation des données Big Data est régulièrement un sujet sensible car il touche à la vie privée, à l'intimité des utilisateurs.

Les bases de données qui stockent nos données sont comparées à des « *Systèmes totalitaires capables de nous contrôler, nous surveiller ou nous influencer* »[11].
Le parallèle est régulièrement fait avec l'ouvrage 1984, de George Orwell, publié en 1949.
L'aspect « *Big Brother is watching you* » pose la question de la limite entre vie privée et vie publique.

Les utilisateurs connectés ne donnent pas leur accord explicite sur cet usage gratuit de leurs données.

Des sites comme Facebook ont intégré dans leurs informations une page expliquant à l'utilisateur les « coulisses » des annonceurs et proposent d'affiner le paramétrage du profil.

Encore faut-il être conscient de ces usages et de prendre le temps de chercher et lire les informations. En effet, il est souvent reprocher l'absence d'informations ou le manque de visibilité sur l'exploitation des données collectées et stockées par les sites.

Les institutions souhaitent mettre en place des autorités de contrôle pour une éthique Big Data. Des textes de lois sont parus pour légiférer la protection des données personnelles. Le General Data Protection Regulation EU 2016/679 a été voté le 27 avril dernier avec pour objectif :
- « *consolider les droits des personnes physiques,*
- *renforcer les pouvoirs des autorités européennes et*
- *responsabiliser les entreprises qui traitent les données à caractère*

personnel. » qui renforce le « droit à l'oubli, ainsi que le droit à la portabilité des données et à l'opposition de toute opération marketing. »[12].
Le Big Data est nourri de nos pratiques numériques quotidiennes, qu'elles soient personnelles ou professionnelles et offre ainsi de belles opportunités de sources informationnelles à certaines industries.

Le cadre juridique posé nous permettra de rester maître, autant que faire se peut, de nos informations. Nous aurons le choix de contribuer à cette transformation digitale de notre système économique.

Mais ce qu'il ressort de cette étude est que ce soit à titre individuel ou dans un contexte d'entreprise, il est essentiel aujourd'hui de cadrer et structurer notre patrimoine d'informations afin de le maîtriser et de pouvoir le faire évoluer.

---

[1] C'est quoi le Big Data ? [En ligne] : https://www.youtube.com/watch?v=UwAENZ3-up8

[2] http://bfmbusiness.bfmtv.com/01-business-forum/le-big-data-pour-les-nuls-621028.html

[3] http://bfmbusiness.bfmtv.com/01-business-forum/le-big-data-pour-les-nuls-621028.html

[4] Big data : données, données, donnez-moi ! [En ligne] : https://www.youtube.com/watch?v=5otaBKsz7k4

[5] Internet of Things : extension d'Internet à des objets ou à des lieux (montres connectées, Transport, GPS, etc)

[6] Big Data c'est quoi ? [En ligne] : https://www.youtube.com/watch?v=a5KXurVhu4Y

[7] Conférence : « Le BIG DATA pour les nuls », HUPI - CCI Bayonne - 26 Octobre 2016. [En ligne] : http://www.bayonne.cci.fr/js/libs/tiny_mce/plugins/filemanager/files/Presentation-Hupi-Bigdata-26-oct-2016.pdf

[8] Pourquoi les objets connectés font rêver les compagnies d'assurances. Les Echos, le 3 mars 2015. [En ligne] : https://www.lesechos.fr/03/03/2015/LesEchos/21889-042-ECH_pourquoi-les-objets-connectes-font-rever-les-compagnies-d-assurances.htm#vYIwiY2qcS9OerVX.99

[9] Big Data : face à l'abondance, l'entreprise doit gérer la rareté. Les Echos, 23 février 2012. [En ligne] : http://archives.lesechos.fr/archives/cercle/2012/02/23/cercle_43767.htm#f6CMH4OREqjt5ICu.99http://archives.lesechos.fr/archives/cercle/2012/02/23/cercle_43767.htm#f6CMH4OREqjt5ICu.99

[10] Définition Wikipédia : ensemble structuré des termes et concepts représentant le sens d'un champ d'informations, que ce soit par les métadonnées d'un espace de noms, ou les éléments d'un domaine de connaissances.

[11] Big Data et données personnelles: vers une gouvernance éthique des algorithmes, Jérôme Béranger, 22 décembre 2014, http://parisinnovationreview.com/2014/12/22/big-data-ethique/

[12] Pourchet, D. Protection des données personnelles : qu'est-ce que la nouvelle règlementation va changer ?, Les Echos, le 07 mars 2017. [En ligne] : https://www.lesechos.fr/idees-debats/cercle/cercle-167196-protection-des-donnees-personnelles-quest-ce-que-la-nouvelle-reglementation-va-changer-2070321.php#WkCUwYAhpJiWIDsA.99

[13] Le Big Data vous observe mais n'ayez pas peur ! et gardez le contrôle...

La veille à l'ère du Big data

■ Hélène Kaizer

La veille (CC0)

La veille, selon le vocabulaire de la documentation, est définie comme « *un dispositif organisé, intégré et finalisé de collecte, traitement, diffusion et exploitation de l'information qui vise à rendre une entreprise, une organisation, quelle qu'elle soit, capable de réagir, à moyen et long termes, face à des évolutions ou des menaces de son environnement, que celles-ci soient technologiques, concurrentielles, sociales, etc...* ». A l'ère du web 2.0 où les blogs, les avis de consommateurs et les flux RSS se sont très largement développés, les dispositifs de veille évoluent au rythme du cycle de l'information. De la collecte à la diffusion des données, les veilleurs sont confrontés à l'infobésité, à la variété des sources et à la volatilité des informations.

La veille est à envisager pour appréhender la gestion du risque, elle permet d'anticiper et de réduire les zones d'incertitudes. Traditionnellement, la veille est associée à des compétences métier (veille juridique, veille technologique, veille réglementaire, etc.[1]). Elle s'attache à la maîtrise d'une compétence spécifique, d'une compétence clé. La veille aide à la maîtrise de la communication et de l'information, elle permet de s'assurer du lien entre l'organisation et son environnement en plus de la maîtrise des règles organisationnelles. Jusqu'à l'avènement du web 2.0, la veille se limitait à la surveillance de sources accréditées, bien souvent de nature textuelle. Depuis lors, il a fallu qu'elle évolue au gré des innovations.

Comment la veille s'adapte-t-elle aux spécificités du Big data qui peuvent être résumées en trois mots : volume, variété et vélocité des données ?

Les défis du veilleur 2.0

Le volume des données disponibles est colossal. La mise en ligne du service public de la donnée[2], courant avril 2017, laisse présager un accroissement constant des ressources en open data. Le volume des données du web croît sans cesse et le développement de l'usage des objets connectés vient confirmer cette tendance. Le veilleur n'a plus la possibilité de ne compter que sur la collecte manuelle, celle-ci doit être automatisée. Désormais, elle s'effectue aussi sur des données collectées par des robots.

Cette masse de données variées pose la question de l'analyse. Pour le veilleur, la difficulté consiste à dégager une information qualitative à partir des données quantitatives issues de la collecte. Il convient de dévoiler la valeur de l'information transmise par les données : est-ce positif, négatif ou neutre ? Il est impossible pour un humain d'en extraire la portée sans l'aide de machines.

Il s'agit en effet d'appréhender une information issue d'une masse de données non assimilable sans assistance technologique. La restitution de l'information est souvent associée à des outils de datavisualisation et d'infographie. La datavisualisation prend sa source dans la restitution de données brutes quand l'infographie met en lumière une information grâce à une représentation graphique qui peut être de l'ordre de la relation entre des données, de la modélisation d'un processus ou bien élaborer une vue textuelle comme un nuage de tags.

Les compétences du veilleur doivent permettre d'anticiper les tendances, de détecter les opportunités de développement et procurer un avantage concurrentiel à l'organisation. Il doit capter, analyser et diffuser les informations stratégiques aux décideurs en s'adaptant à un environnement aujourd'hui centré sur la donnée. Il convient de maîtriser les outils informatiques et d'en comprendre les mécanismes. Par extension, la maîtrise de la langue anglaise lui offre de sérieux atouts.

Les spécificités du web nouvelle génération auxquelles la veille doit faire face...

L'une des spécificités essentielles du web est caractérisée par « l'information multiforme », c'est à dire

que les informations sont de nature multimédia. Les contenus peuvent prendre la forme de vidéo, d'image fixe, de texte, d'enregistrement sonore, etc. Une partie de l'information est produite par les internautes (on qualifie cela de « User Generated Content »), presque dans l'immédiateté, au sens où celle-ci est rendue disponible extrêmement rapidement voire instantanément. La multiplication des canaux de diffusion rend la veille complexe.

En effet, le web se compose d'une multitude de réseaux. Il peut s'agir de réseaux sociaux qui concernent la sphère privée tels Facebook et Google+, ou bien la sphère professionnelle tel LinkedIn. Il est composé de communautés virtuelles d'internautes qui partagent des informations sur des blogs ou sur des chaînes Youtube. Ainsi il est recommandé de veiller sur les réseaux sociaux, les blogs, les plateformes de partage vidéo pour ne pas manquer une information stratégique et faire valoir sa réactivité.

Une autre spécificité du web réside dans la viralité. L'information se propage sur les réseaux à la manière du bouche-à-oreille. Un internaute va apporter du crédit ou du discrédit à une information en la commentant, en la partageant. Il se crée alors de nouvelles informations qu'il convient de surveiller de près, notamment lorsqu'elles concernent l'image de marque de l'organisation. L'analyse de la viralité vient s'ajouter aux défis rencontrés par les dispositifs de veille et ouvre les portes de l'analyse de l'impact stratégique. La possibilité de suivre la propagation d'une campagne marketing, par exemple, donne accès à de précieuses informations sur la réception de cette même campagne.

Les nouveaux axes de veille

La veille sur l'e-réputation consiste à suivre les flux d'informations qui circulent sur l'organisation et son image de marque. Les informations peuvent émaner de l'organisation (site institutionnel), mais aussi des clients, des partenaires, des collaborateurs. Ce type de veille est très intéressant dans la gestion du risque car elle permet d'être alerté rapidement et de réagir en cas d'usurpation d'identité ou de fuites d'informations confidentielles par exemple. Elle anticipe la veille de l'information sensible où l'accent est mis sur la surveillance des plateformes de partage pour vérifier qu'il n'y ait pas de fuites de données.

La veille sur les réseaux sociaux (qualifiée aussi par l'expression « social media listening ») est importante à cause de l'information produite par les utilisateurs (user generated content) qui sont très nombreux. Il devient possible de saisir l'influence des campagnes émises par l'organisation grâce à une offre en social media listening dynamique et innovante. L'image de marque de l'organisation est soumise à l'analyse de l'humeur et du comportement des internautes. L'exploitation des métadonnées issues des réseaux sociaux sert de capteur d'opinion et favorise le dégagement de tendances.

La veille « temps réel » intervient au moment où il faut gérer une crise. L'information recueillie sert à prendre des décisions d'urgence. L'information est collectée via des requêtes sur les moteurs de recherche, sur les sites de presse et d'actualités et fournit des informations de publication récente.

Les innovations des logiciels dédiés à la veille sur le web

Les outils de veille répondent aux problématiques posées par les spécificités du web en mettant à disposition des interfaces « social media listening » via des connecteurs intégrés sur les API des détenteurs d'information (Twitter, Instagram, Facebook et Youtube notamment). Les données collectées sont conservées dans le cloud. Même si cela peut sembler poser des problèmes de confidentialité, il n'en est rien car il s'agit pour l'essentiel de données brutes dont on garantit la protection

Social media listening
(source : http://www.activitycloud.com/)

mais qui ne représentent pas de connaissance particulière.

Les outils de veille de nouvelle génération permettent de produire des tableaux de bord, de formuler des hypothèses d'interprétation, de faire des recommandations. Les outils et les résultats de la veille sont partagés par des équipes et peuvent être utilisés en mobilité, être commentés et enrichis. Le volume des données est rendu intelligible par des interfaces de datavisualisation. Il devient possible de visualiser les tendances, les relations, les sujets en régression ou en progression d'un seul coup d'oeil.

Certains acteurs du domaine proposent également des outils prédictifs, par exemple Palantir[3], spécialisés dans l'intégration de jeux de données en temps réel. Ces acteurs se spécialisent dans le Big data, leurs produits permettent de dévoiler les tendances qui se dégagent d'un volume de données massif collecté et exploité en temps réel.

Nous pouvons noter un changement dans les pratiques de veille à l'ère du web 2.0 : « il faut substituer à la validation de la source la validation de l'information qu'elle produit[4] ». Il ne suffit plus de suivre des sources d'autorités identifiées, il faut dénicher l'information qui fera la différence parmi le volume informationnel du web. Les technologies Big data permettent l'analyse d'opinion, de comportement de masse ce qui était impossible avec une approche de veille traditionnelle.

Avec le web 2.0, la veille devient un outil collaboratif au service d'une intelligence collective avec un bémol cependant : les profils utilisateurs du web ne sont pas forcément représentatifs de la société actuelle ce qui signifie que l'interprétation peut s'éloigner de la réalité... Toutefois, les générations nées avec le numérique viendront bientôt rétablir la confiance dans la représentativité de la population.

[1] http://www.adbs.fr/veille-19022.htm?-RH=OUTILS_VOC

[2] http://www.data.gouv.fr/fr/reference

[3] https://www.palantir.com/

[4] « La veille dans un environnement numérique mouvant » de Christophe Deschamps, dans Guide Pratique Archimag n°47 « Outils et efficacité d'un système de veille »

La recherche d'informations (CC0)

Voici un panel d'outils de veille, classés alphabétiquement, compatible avec les spécificités du Big data :

- ***4Dconcept : http://www.4dconcept.fr/fr/veille-et-analyse-concurrentielle***
- ***Bolero live : http://www.bolero.fr/prestations/bolero-live/***
- ***Brandwatch : https://www.brandwatch.com/fr/***
- ***GraphyStories : http://www.graphystories.com/fr/***
- ***Oktopost : https://www.oktopost.com/***
- ***Palantir : https://www.palantir.com/***
- ***Talkwalker : https://www.talkwalker.com/fr***

BIBLIOGRAPHIE

« La veille multimédia ou la kaléidoscopie de l'information » / Antoine Raulin, I2D – Information, données & documents 2016/4 (Volume 53), p. 4-6

Archimag Guide pratique n°47 «Outils et efficacité d'un système de veille»

Archimag n°301 de février 2017 «Réinventer la veille»

https://www.1min30.com/dictionnaire-du-web/social-media-monitoring-veille (consulté le 10/04/2017)

L'intelligence économique au service de l'innovation / Christian Coutenceau, François Barbara, Véronique Chapuis-Thuault... [et al.]. Editions Eyrolles, 2014.

Nouveaux usages de la veille : 5 pratiques en émergence / Groupe de travail intelligence économique et économie de la connaissance - Groupement Français de l'Industrie de l'Information, 2012.

L'ouverture des données de la recherche. Nouvel Eldorado du Big Data ?

■ Aurélie Quinodoz

Avec son volume vertigineux de données produites, le Big Data s'est largement imposé dans le débat public. Exploitation commerciale des données personnelles, nouveau gisement d'emplois, création de valeurs… Effet de mode ou nouvelle révolution ?

Caractérisé par la règle dite des «5V», pour volume (extension indéfinie des données à traiter et à stocker), vélocité (rapidité de production de ces données), variété (multiplicité des formats de ces données, structurées ou non), vitesse (actualisation des données en quasi temps réel) et valeur (pour les usages qu'elles produisent), le Big Data ouvre de nombreuses perspectives. Toutefois, les risques sont tout aussi nombreux, dans tous les domaines d'activités.

Les données de la recherche (Source : http://www.helsinki.fi/kirjasto/en/get-help/management-research-data/)

Dans le secteur public et notamment dans l'enseignement supérieur et la recherche, l'idée d'une ouverture des données s'est progressivement imposée avec les mouvements de l'Open Access et de l'Open Data. Après la valorisation des résultats de la recherche[1], l'ouverture des données de la recherche, données utilisées ou produites lors du processus de recherche même, est une nouvelle étape. Récemment, la loi française «Pour une République numérique[2]» entérine et encourage également le partage et l'exploitation des savoirs produits par la recherche, en accordant le libre accès aux données de recherche publique et l'accès sécurisé des chercheurs aux grandes bases de données publiques.

Quelles que soient les technologies développées par le Big Data pour l'utilisation des données, les questions se multiplient. Quelles données transmettre? Comment les rendre accessibles mais aussi les protéger ? Comment faciliter le travail des chercheurs, aussi bien en tant qu'utilisateurs ou que producteurs ?

Pourquoi s'intéresser aux données de la recherche ?

Les données de la recherche constituent un matériau essentiel à la recherche, d'autant plus précieuses qu'elles sont produites dans le cadre d'expériences complexes et coûteuses (voir encadré 1). Invisibles pourtant, elles restent la part immergée de la recherche.

De nombreux enjeux

Un enjeu épistémologique

Garantir la mise à disposition massive de jeux de données pouvant être exploités scientifiquement est un enjeu épistémologique pour les chercheurs. Le partage des données ouvre les portes du processus de recherche, assure l'exactitude des données, leur authenticité, leur fiabilité et la reproductibilité des résultats de la recherche. La mise en oeuvre d'une protection et d'une diffusion des données de la recherche « *joue un rôle dans le processus scientifique, en particulier pour la validation des hypothèses et résultats* »[3].

Un enjeu patrimonial

Conserver et transmettre les données de la recherche permet de réduire le risque de perte grâce à des plateformes d'hébergement correctement sécurisées. Elles permettent de construire une mémoire des savoirs acquis et des savoirs faires et

améliorent la visibilité des chercheurs et de leurs résultats (augmentation du nombre de citations).

Un enjeu économique

Communiquer les données de la recherche permet d'éviter les doubles emplois en reproduisant certaines données existantes mais non exploitables (confidentialité, problème de format). Produites grâce à des financements publics, elles appellent un retour sur investissement pour accroître l'efficacité de la recherche. Les financeurs sont très attentifs aux potentiels de cette ouverture. Au niveau européen, le programme Horizon 2020[4] mène une opération pilote en la matière, Open Research Data. Tout projet Horizon 2020 faisant partie du pilote doit mettre en oeuvre le dépôt des données dans un référentiel de données. Enfin, la connaissance des travaux menés par les laboratoires concurrents d'un même secteur permet d'écarter des pistes de recherche déjà engagées et de mesurer les intérêts portés sur certains secteurs d'étude. Elle induit un esprit de prospective et une meilleure compréhension des études en cours.

Un enjeu sociétal

Améliorer la confiance dans la science et la recherche. Pour répondre à ces enjeux comme au volume exponentiel des données, le Big Data apporte des outils nouveaux.

Quels outils pour le Big Data ?

De nouvelles méthodes de gestion et d'analyse d'importants volumes de données voient le jour. Calculées par des algorithmes, les données produisent des indices de réputation (on peut penser aux like de Facebook ou encore les followers sur Twitter), d'autorité (le fameux PageRank de Google) ou encore de popularité pour mesurer l'audience des sites. De nouveaux outils s'écartent des calculs statistiques pour produire une « prédictivité » : c'est le rôle des Machines Learning, ces machines apprenantes.

Capables d'appréhender de nouvelles données et de tirer profit des résultats antérieurs, elles permettent la création de valeurs. La détection de l'information utile est facilitée grâce à la fouille de texte ou encore Data mining (voir encadré 2), qui prend en charge la variété des données, structurées ou non (textes, images, contenus multimédia...)

Afin d'offrir de nouveaux services, d'anticiper des comportements et des besoins, d'améliorer la prise de décision ou encore de réaliser des économies, les applications se développent du côté des données publiques ouvertes. Les promesses sont-elles aussi grandes avec l'ouverture des données de la recherche ? Malgré la multiplication des incitations, des freins persistent. Ils concernent en premier lieu la protection des données et la crainte du mésusage de ces données. En effet, certaines données sont dites « sensibles » pour des raisons de sécurité ou parce qu'elles font l'objet d'accords commerciaux. Les coûts humains (en temps et en compétences) et les coûts techniques (infrastructures matérielles lourdes) sont aussi régulièrement cités par les chercheurs. Enfin, la diffusion de données personnelles attise les craintes.

Dans ce contexte, il convient de faciliter l'ouverture aux données de la recherche en construisant un écosystème qui lève ces difficultés. Cela suppose de mettre en oeuvre une stratégie pour maximiser les investissements humains et financiers et accroître le potentiel de la recherche.

Diffuser et protéger les données de la recherche ?

Nous l'avons vu. Le Big Data comme l'ouverture des données de la recherche soulèvent de nouvelles opportunités, mais aussi de nouvelles questions. En réponse, les actions gouvernementales se multiplient. Certains pays comme les États-Unis ou le Royaume Uni sont déjà très avancés dans la mise en oeuvre de plans nationaux pour organiser la diffusion des données. Plus récemment, les institutions de recherche françaises multiplient les initiatives.

Ces plans ont pour but d'optimiser la valeur de la recherche et de répondre aux risques associés à une absence de gestion des données de la recherche (perte de temps, d'argent, absence de réemploi). La mise en œuvre d'un plan de gestion (DMP) conjuguée à une évolution du droit impulse une politique de protection et de diffusion des données.

DONNÉES DE LA RECHERCHE ?

« Enregistrements factuels (textes, chiffres, images, sons) utilisés comme sources principales pour la recherche scientifique et reconnus par la communauté scientifique comme nécessaires pour valider les résultats de la recherche » (Déf. de l'OCDE – 2007)»

- **Données brutes (raw data)** : données d'observation, données expérimentales, données de simulation

- **Données dérivées (derived data)** : obtenues par compilation ou traitement des données brutes

- **Jeux de données (dataset)** : agrégation de données brutes ou dérivées rassemblées dans un tout cohérent

La réutilisation des données de la recherche est libre si les données sont issues d'une recherche financée pour moitié par des fonds publics ou si les données ne sont pas protégées par un droit spécifique.

- les données ont été rendues publiques par le chercheur ou un établissement

ZOOM SUR LA FOUILLE DE DONNÉES / DATA MINING

La fouille de textes et de données (text and data mining – TDM) consiste à explorer, via des techniques de fouille, des corpus immenses – composés notamment d'articles scientifiques ou de données expérimentales – afin d'en déduire des nouvelles connaissances. Imaginez qu'un chercheur en épidémiologie puisse analyser automatiquement les 2 millions d'articles scientifiques traitant du cancer, afin de déterminer s'il existe une corrélation inconnue à ce jour entre une pratique et l'apparition de la maladie : voilà ce que permet le TDM. Si cela vaut pour l'ensemble des disciplines, cette technique prend tout son sens lorsqu'il s'agit d'objets complexes, qui nécessitent un traitement pluridisciplinaire : les épidémies donc, mais aussi les objets climatiques, les phénomènes politiques, etc.

http://www.cpu.fr/actualite/recherche-la-fouille-de-donnees-desormais-accessible/

Qu'est que le DMP ?

Data Management Plan est un outil conçu pour garantir la collecte, la protection et la diffusion des données de la recherche. Il permet de disposer, d'accéder et d'exploiter de données de qualité, accompagnées de leurs métadonnées. Il répond en outre aux objectifs des organismes de financement de la recherche publique. Il s'agit d'un document rédigé au démarrage d'un projet de recherche. Il décrit le cycle de gestion de l'ensemble des données qui seront collectées ou produites, les conditions de conservation et les modalités de traitement, de préservation et de réutilisation. Un coup d'oeil sur le site écossais Digital Curation Centre permet de prendre en compte les questions essentielles avant le démarrage du DMP.

Processus de planification, il implique la mise en place de bonnes pratiques tout au long du cycle de vie des données. Il permet de contrôler la qualité des données. Sa mise à jour régulière en garantit la valeur.

Le DMP est un outil de pilotage qui implique l'ensemble des acteurs. Le chercheur détermine et prend en charge le cycle de vie des données. Un document formalisé précise le choix des données à ouvrir, du lieu du dépôt et des conditions d'usage. Il revient à l'établissement de décider de la protection et de l'exploitation des données[5].

Quelques points de vigilance essentiels :

■ Description des données : par une documentation et avec des métadonnées partagées par une large communauté, lisibles par l'humain et validées par des standards. Comment les rédiger correctement ? Consulter le guide Réaliser un plan de gestion des données de l'Université Paris-Diderot[7].

■ Stockage et la conservation des données : stockage adapté au niveau de confidentialité des données, durée de conservation, choix du format de fichier.

■ Création d'identifiants pérennes et uniques : pour faciliter l'accès et le partage.

■ Partage et publication des données : par le dépôt dans des entrepôts de données de recherche, par la publication de Data papers.

De nombreux guides disponibles en ligne aident à la rédaction...Mais rien ne remplace l'accompagnement des chercheurs par les établissements pour faciliter l'accélération de ce mouvement d'ouverture.

Protéger juridiquement les données et les chercheurs

Pour amplifier ce mouvement tout en protégeant les données et les informations personnelles, directives européennes et loi française encadrent la communication des documents et des données publiques[6]. Elles limitent et spécifient les usages possibles.

Côté communication, elles prévoient des restrictions (secret défense, secret professionnel, informations personnelles ou présentant des risques pour la sécurité publique, dépôt de brevet...). De plus, la collecte de données à caractère personnel doit faire l'objet d'un respect préalable des formalités prévues par la CNIL.

Côté protection, l'usage des données est soumis au Code de la propriété intellectuelle. Toutefois, la nouvelle loi « Pour une République numérique » créée un statut de « bien commun » pour les données. La lecture détaillée du texte par Lionel Maurel sur le site S.I.Lex montre une évolution majeure. En effet, «une fois les données de la recherche publiées, leur usage devient pleinement libre et donc « commun à tous » (sous réserve de protéger les données personnelles et les droits des tiers[7]) ».

Photo by Drew Hays on Unsplash

Pour préciser, l'article 30 de la loi introduit « un droit d'exploitation secondaire des publications scientifiques au profit des chercheurs[8] » pour limiter le phénomène de captation des données par les grands éditeurs scientifiques qui demandent à présent, outre les textes, les jeux de données associés.

Par ailleurs, sans véritablement parvenir à un statut juridique spécifique pour les données de la recherche, la loi prévoit aussi leur protection grâce à la préconisation de licence de diffusion. Les licences Creative Commons (CC), Open Data base Licence (ODbL), Etalab (initiative gouvernementale) protègent les fournisseurs de données et les chercheurs (respect de l'intégrité des données et mention des sources).

Attention toutefois !
Les licences ne proposent pas toutes les mêmes niveaux de protection.

Comment protéger les données et favoriser les échanges ? Un exemple

Le CADS (Centre d'accès aux données sécurisées) est développé avec le soutien de la France. Il déploie un ambitieux projet à l'attention de la communauté scientifique française et étrangère. Le CASD a développé un pilote de centre d'accès sécurisé européen entre l'Italie, l'Allemagne et la France. « L'objectif est de construire à terme un réseau européen à l'échelle d'une vingtaine de pays. En 2014 aussi, les premières entreprises privées testent la technologie pour donner accès à leurs propres données[9]».

Tout est-il résolu ?

Pas vraiment !

Un récent rapport du Parlement Européen sur l'initiative européenne sur l'informatique en nuage[10] pointe la nécessité d'une harmonisation commune à l'ensemble des membres de l'Union européenne. Les risques liés à l'ouverture des données doivent donner lieu à la mise en place d'opérations de pseudonymisation, d'anonymisation ou encore de cryptage. Disposer d'infrastructures sûres et sécurisées est un autre point souligné par le rapport.

En conclusion, les auteurs du rapport invitent la Commission européenne à mettre en oeuvre une stratégie en matière de cybersécurité pour protéger les travaux de la communauté scientifique européenne.

De plus, le tout récent document de travail rédigé par le CNRS et le projet ISTEX appellent eux aussi à des ajustements pour sécuriser les pratiques des chercheurs (réalisation de guide d'application, charte, sensibilisation). Il demande une clarification des termes et des notions (données de la recherche, écrits scientifiques...) afin de définir les droits afférents.

En conclusion, quels risques pour les données de la recherche ?

S'attacher à comprendre les risques permet d'apporter des solutions techniques, humaines et juridiques. Toutes les données sont concernées et les données de la recherche, précieuses et coûteuses, doivent faire l'objet d'une attention particulière. L'adoption de la loi « Pour une République numérique » ouvre le chemin en préconisant des outils de protection, en favorisant leur diffusion et l'utilisation des outils du Big Data. Ces avancées ont d'ailleurs été saluées par la conférence des présidents d'université, notamment par l'autorisation de la fouille de textes et de données alors que jusqu'ici, « au nom de la protection du droit d'auteur, la loi interdisait aux chercheurs de réutiliser des articles scientifiques et/ou des données publiés, alors même que ces résultats étaient le fruit de leur travail et financées par de l'argent public [11]».

Pour autant, si la loi encourage et rend possible le recours à de nouvelles applications informatiques pour l'analyse des données massives, ce sont bien les hommes qui seuls, peuvent donnee du sens aux données. L'ouverture des données de la recherche ne pourra dès lors réaliser ses promesses qu'avec l'adhésion des chercheurs, en tant que producteurs et utilisateurs.

---

[1] Article L112-1 du Code de la Recherche

[2] http://www.gouvernement.fr/action/pour-une-republique-numerique

[3] Joachim Schöpfel, Eric Kergosien, Hélène Prost. « Pour commencer, pourriez-vous définir 'données de la recherche' ? » Une tentative de réponse. Atelier VADOR : Valorisation et Analyse des Données de la Recherche, INFORSID 2017, May 2017, Toulouse, France. 2017. <hal-01530937>

[4] La liste est disponible sur OpenAire : https://www.openaire.eu/opendatapilot

[5] Les modalités sont précisées sur le site d'Horizon2020 : http://www.horizon2020.gouv.fr/cid87457/publication-guide-d-aide-a-la-redaction-d-un-plan-de-gestion-de-donnees.html

[6] Directive européenne 2013/37/UE dite «PSI » pour «Public Sector Information » et, en France, par la loi dite « CADA» du 17 juillet 1978 codifiée depuis le 19 mars 2016 dans le Code des relations entre le public et l'administration.

[7] Maurel, L. (2016). Quel statut pour les données de la recherche après la loi numérique ? <https://scinfolex.com/2016/11/03/quel-statut-pour-les-donnees-de-la-recherche-apres-la-loi-numerique/>

[8] Le travail de la science et le numérique – Données, publications, plateformes. Une analyse systémique de la loi pour une République numérique. (Consulté le 1 mars 2017) <http://www.donneesdelarecherche.fr/spip.php?article997>

[9] Antoine Frachot. L'accès aux données sécurisées : première conférence du CASD. Disponible sur : https://casd.eu/documents/dossier-de-synthese-des-interventions-conference-casd-04-2016.pdf

[10] Parlement Européen. Rapport sur l'initiative européenne sur l'informatique en nuage. (Consulté le 28 février 2017) <http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//NONSGML+REPORT+A8-2017-0006+0+DOC+PDF+V0//FR>

[11] http://www.cpu.fr/actualite/recherche-la-fouille-de-donnees-desormais-accessible/

BIBLIOGRAPHIE SÉLECTIVE

- Cardon D. (2015). *À quoi rêvent les algorithmes ? Nos vies à l'heure des big data.* Paris : Editions du Seuil.

- Cartier, A., Moysan, M. et Reymonet N. *Construire des outils pour la gestion des données de la recherche dans une communauté d'universités.* Récupéré du site HAL : https://hal-descartes.archives-ouvertes.fr/hal-01138663/document

- Digital Curation Centre. http://www.dcc.ac.uk/

- La directive INSPIRE pour les nuls (3ème édition). Récupéré le 15 février 2017 du site Géopal : http://www.geopal.org/upload/gedit/1/file/Inspire_pour_les_neophytes_V3.pdf

- Lmberger, P., Batty M., Morel M. et al. (2015). *Big Data et Machine Learning : Manuel du data scientist.* Paris : Dunod.

- Maurel, L. *Quel statut pour les données de la recherche après la loi numérique ?* Récupéré le 17 février 2017 du site S.I.Lex : https://scinfolex.com/2016/11/03/quel-statut-pour-les-donnees-de-la-recherche-apres-la-loi-numerique/

- MIT Libraries, Cambridge. *Data Management.* Récupéré le 10 février 2017 du site MIT Libraries : https://libraries.mit.edu/data-management/plan/write/

- Ouverture des données de recherche : Guide d'analyse du cadre juridique en France. Récupéré le 10 février 2017 du site Biblithèque scientifique numérique : http://www.bibliothequescientifiquenumerique.fr/wp-content/uploads/2017/01/Guide_analyse_Cadre_Juridique_Ouverture_donness_Recherche_V1.pdf

- Réaliser un plan de gestion des données. Récupéré le 10 février 2017du site Université Paris Diderot : http://www.univ-paris-diderot.fr/DocumentsFCK/recherche/Realiser_un_DMP_V1.pdf

- data.gouv.fr. Récupéré le 10 février 2017 du site Data.gouv. : http://www.data.gouv.fr/fr/

- Université de Montréal. *Gestion des données de recherche.* Récupéré le 6 février 2017 : http://guides.bib.umontreal.ca/disciplines/500-Gestion-des-donnees-de-recherche?tab=2304

- Urfist, Strasbourg. *Gestion et diffusion des données de la recherche*. Récupéré le 12 février 2017 : http://www.inist.fr/IMG/pdf/support_formation_urfist_strasbourg_20161125.pdf.

Research data management
(Source : https://www.elsevier.com/connect/managing-your-research-data-to-make-it-reusable)

Archivage du Web

■ Hélène Kaizer

Depuis l'avènement d'Internet, le nombre de données créées ne cesse de croître. Chaque internaute est producteur de données, chaque donnée fournit de l'information qu'il convient de conserver. Comment est né l'archivage du web français ? Qui en a la charge ? Comment les archiveurs font-ils face à ce tsunami d'informations ?

L'archivage du web mondial et le dépôt légal du web français ont fêté les 22 et 23 novembre dernier leur anniversaire. Alors qu'Internet Archive, fondation américaine pionnière en la matière, fête ses 20 ans, le dépôt légal du web français a eu 10 ans. Plusieurs initiatives ont eu lieu au cours de ces deux décennies dans différents pays notamment en Suède, en Australie et au Portugal [1] .

Pour harmoniser les pratiques et unir les efforts face à un défi sans frontières, l'organisation International Internet Preservation Protocol (IIPC) a été fondée en juillet 2003 à l'initiative de la Bibliothèque Nationale de France. Son rôle est de favoriser la coopération entre les acteurs internationaux qui s'intéressent à la préservation du web. C'est en 2006 que la loi relative au droit d'auteur et aux droits voisins dans la société de l'information (dite loi DAVDSI) a rendu obligatoire le dépôt légal du web en France. Les missions, allant de la collecte à la sauvegarde, ont donc été confiées aux institutions déjà en charge du dépôt légal des autres médias, respectivement la Bibliothèque Nationale de France (BNF) et l'Institut National de l'Audiovisuel (INA).

Cependant, la méthodologie de collecte du dépôt légal diffère de la méthodologie classique car il n'y a pas d'obligation de dépôt pour les sites web. Il est question d'un moissonnage de l'Internet où le patrimoine est volumineux, volatile et mouvant.

L'archivage du web sous le prisme des 5 V

V comme Volume

La fondation Internet Archive conserve plus de 25 Pétabytes de données (10^{15}), ce qui équivaut à 510 milliards de pages ! Fin 2016, la BNF conservait 29 milliards de fichiers, pour un volume de 668 Téraoctets d'archives web avec un accroissement de 120 Téraoctets chaque année.

Près de 4,5 millions de pages web sont collectées tous les ans ! Tandis que l'INA conservait 52,3 milliards de ressources web et avait capté 389 millions de tweets qui occupent un espace de 3,95 Pétabytes ! En comparaison, la moitié du catalogue analogique de la Bibliothèque Nationale de France (ca. 6 millions de livres) équivaut à 1 Téraoctet[2] !

V comme Variété

La variété des données n'est pas en reste ! Le web est un média qui propose de multiples ressources : du texte, de l'image fixe, de l'image animée, du son, des créations telles que les webdocumentaires, les infographies, les e-books, etc. Les données sont complexes et hétérogènes. L'archivage du web doit permettre de reconstituer les éléments collectés, quelles que soient les ressources.

V comme Vélocité (rapidité, flux)

En plus du volume et de la variété, les sites média, c'est-à- dire les sites qui procurent du contenu sous forme visuelle, textuelle et sonore, proposent des données volatiles et réactives. Les sites web sont des sources primaires, ce qui signifie que les producteurs d'informations les diffusent en premier lieu sur Internet.

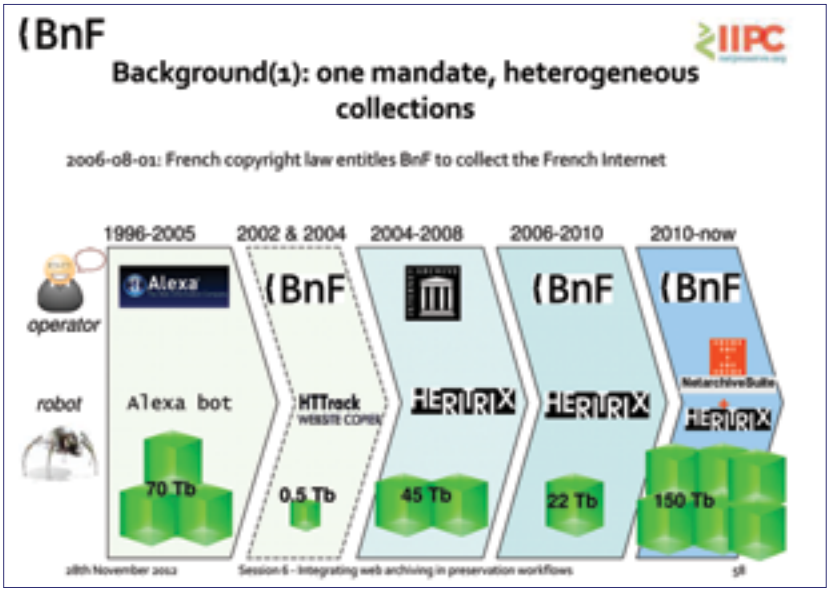
La caractéristique première d'un site web est d'être un élément dynamique, où le contenu et la mise en forme sont régulièrement mis à jour et modifiés. « Un site Web change ou disparaît tous les quatre-vingts jours en moyenne. Et avec, c'est notre culture, notre histoire, la trace de nos vies qui disparaissent. Le Web a besoin d'une mémoire », expliquait Brewster Khale dans un article du Monde intitulé « Vingt ans d'archivage du web : les coulisses d'un projet titanesque » [3] .

V comme Valeur et Véracité

Les 3 V précédemment cités s'accompagnent des enjeux liés à la valeur ainsi qu'à la véracité des sites

collectés. Chaque site est porteur d'informations véhiculant des valeurs, qu'elles soient financières,qualitatives ou bien symboliques. Prenons l'exemple du site de la Maison Blanche aux États-Unis qui s'est vu dévaloriser de son accès en langue espagnole dès l'installation de Donald Trump dans le bureau ovale [4]. Pour poursuivre avec l'exemple américain, il n'est plus nécessaire de rappeler à quel point la véracité des informations peut être mise en doute et donc à quel point il est primordial de la préserver.

Tous ces éléments (Volume, Variété, Vélocité, Valeur et Véracité) permettent de rapprocher l'archivage du web des projets de Big data.



Historique des outils de collecte utilisés à la BNF depuis le début du projet d'archivage du web (Source : http://netpreserve.org/)

Quelle gouvernance des données pour archiver le web ?

La gouvernance des données consiste en la bonne gestion du cycle de vie des données. Les organisations documentaires sont expertes en la matière mais les spécificités du numérique leur ont demandé quelques ajustements et adaptations. Passons en revue les différentes étapes de la gouvernance avec l'objectif d'archiver le web, un patrimoine évolutif et dynamique, qui est soumis au dépôt légal mais qui, nous le rappelons, n'est pas soumis à une obligation de dépôt par les producteurs de données.

Première étape : la collecte

En France, la collecte du dépôt légal du web est assurée par la BNF et l'INA depuis la promulgation de la loi DAVDSI en 2006 et son décret d'application en 2011. Ce cadre réglementaire définit le périmètre de collecte attribué à chacune des deux institutions. Bien que tous les sites ne puissent pas être archivés, la BNF a pour périmètre de collecte le web français en .fr ou toutes autres extensions liées au territoire national par exemple : .re (La Réunion), .nc (Nouvelle-Calédonie), etc. mais également d'autres sites grâce à un élargissement du périmètre défini selon des thématiques ciblées.

L'objectif est de constituer un échantillon représentatif du web à un moment donné mais aussi de recueillir les données associées à des événements ou à des thématiques ciblées. Pour cela, l'archivage du web est effectué via des collectes « larges » et des collectes « ciblées » dont la périodicité de moissonnage et la profondeur de recueil sont décidées par les différents départements. Une procédure de collecte d'urgence vise aussi à couvrir la collecte des sites éphémères (sites web de festival par exemple). Les outils utilisés, depuis mars 2017, sont la dernière version de NetArchive Suite[5] basé sur le crawler web Heritrix 3. Ces outils open-source améliorent la qualité et la couverture des collectes.

À partir de 2009, l'INA a pris en charge la collecte du web français du secteur de la communication audiovisuelle. Celui-ci se limite à 14000 sites mais leur moissonnage est quotidien.
Il y a 4 grands axes : « Les sites web des chaînes de télé et radio, les Web Tv et les Web Radios, les contenus édités par les acteurs de l'audiovisuel (accès depuis les plateformes de partage audiovisuelles), et les sites liés aux programmes diffusés par les chaînes (émissions, séries, sites de fans...) ». Les outils de collecte sont développés en interne. La documentation concernant ces outils est très peu disponible sur le web.

Deuxième étape : Classement / Traitement / Analyse

Le traitement des données collectées intervient ensuite. L'objectif visé à cette étape est de pouvoir utiliser les données via des applications qui permettent de les modifier tout en conservant les données originelles. On procède alors à un encapsulage des données [6].

A la BNF, depuis 2014, on a fait le choix du format WARC. Il signifie Web ARChive et correspond aujourd'hui à la norme ISO 28500 : 2009 pour la préservation du web. Il s'agit de l'héritier du format ARC développé par Internet Archive. Le format WARC est un format conteneur qui permet à un fichier de contenir plusieurs objets pour le stockage, la gestion et l'échange. Le WARC rend possibles « notamment l'enregistrement des entêtes de requête HTTP, l'enregistrement des métadonnées,

l'attribution d'un identifiant pour chaque fichier conteneur, la gestion des doublons et des dossiers ayant migrés, la segmentation des documents. Les fichiers au format WARC sont destinés à stocker tout type de contenu numérique, qu'il soit récupéré par HTTP ou via un autre protocole[7] ».

L'INA utilise un format développé en interne, le format DAFF (Digital Archive File Format) pour l'encapsulage des données après la collecte. Comme précédemment, il y a très peu d'informations disponibles sur ce format. S'agit-il d'un format ouvert ? Il est difficile de répondre à cette question.

Troisième étape : Conservation / stockage / préservation

La BNF a créé le système SPAR pour la conservation mais surtout pour la préservation des données. Comme indiqué sur son site, « *SPAR (Système de Préservation et d'Archivage Réparti) cherche à garantir la préservation des documents numériques : à faire en sorte que l'information reste lisible, compréhensible et réutilisable sur le long terme, même si l'environnement technique et humain dans lequel ces documents ont été produits change.*
Comme il est souvent difficile de réparer des documents numériques après dégradation ou de les relire après disparition des matériels et logiciels d'époque, l'effort porte sur les opérations préventives de veille et de migration ».
SPAR est construit avec des briques logicielles ouvertes et est évolutif. L'objectif est de pouvoir adapter le système aux évolutions technologiques en facilitant la restitution et le traitement.

Actuellement, l'INA s'engage dans une démarche de stockage des données orientée Big data. En effet, il est question de fondre tous les systèmes documentaires au sein d'un lac de données dans le but de rassembler toutes les données que l'institution conserve (ce qui inclue les « métadonnées documentaires, commerciales, juridiques et d'usage » d'après Gautier Poupeau interviewé par Bruno Texier d'Archimag [8]). Un lac de données (Data Lake en anglais) est défini par l'absorption de flux de données bruts rendus utilisables pour analyse. Des données disparates sont collectées puis stockées en continu dans un espace que l'on pourrait qualifier de « réservoir ».

Quatrième étape : Communication / restitution

Les archives du web ne sont pas en accès libre en France. Toutefois, la BNF, l'INA et certains pôles associés permettent un accès aux archives dans leurs locaux. La Bibliothèque Nationale Universitaire de Strasbourg, sous l'égide de la BNF, a mis en ligne un tutoriel [9] qui permet de comprendre ce qu'est le dépôt légal du web, ce qu'on y trouve et comment le consulter. L'INA donne accès à la consultation des archives du web dans 27 lieux répartis sur le territoire français [10] via une interface de restitution des sites web développée en interne. Cette interface met à disposition un accès temporel aux sites archivés tout en restituant le contenu dynamique. D'autres outils complémentaires viennent enrichir ces interfaces tels des outils d'analyse qui permettent d'élargir la recherche aux sites, aux URLs, aux dates. La BNF utilise une interface de restitution similaire à la Wayback Machine [11] d'Internet Archive mais ne peut la proposer en dehors de ses locaux pour des raisons de respect des droits d'auteur.

Challenges et enjeux

Le premier enjeu de l'archivage du web est de préserver la mémoire des 20ème et 21ème siècles dans un environnement de grands changements technologiques. Il faut assurer la pérennité de l'archivage tout en trouvant des solutions concernant les difficultés de collecte notamment la collecte des sites sur abonnement (par exemple sites de journaux ou de vidéo à la demande), les applications de musique en ligne ou les sites proposant des livres numériques avec dispositif anti-copie.

Le second enjeu est la restitution des éléments collectés ainsi que leur ouverture à la recherche. Il convient de réfléchir aux usages possibles. Comme le signale Claude Mussou[12] « *l'archive web se construit "au fil de l'eau", par la collecte automatique de contenus qui n'ont, pour partie, pas perdu leur valeur d'usage et, à la différence de la radio ou de la télévision, ne proviennent pas d'un flux temporel linéaire. C'est pourquoi, coller au plus près à la fréquence de mise à jour des pages, ajuster à la taille des sites les profondeurs de collecte par les robots, assurer une veille permanente pour suivre l'apparition des contenus, leur évolution, ou même anticiper leur disparition, et travailler en relation étroite avec la communauté académique sont des priorités pour la constitution de collections qui, à défaut d'être exhaustives, s'efforcent de répondre le mieux possible aux exigences de recherche.* »

La BNF a lancé un projet : Préfiguration d'un service de fourniture de corpus numériques à destination de la recherche qui couvre la période 2016-2019. Ce projet « *doit permettre aux chercheurs d'utiliser les outils d'analyse qui permettent d'aller au-delà de ce que l'humain peut lire, pour établir de nouvelles connaissances. La fouille de textes et de données (ou TDM pour text & data mining) offre justement l'opportunité de voir émerger de nouvelles pistes de recherche : il ne s'agit plus de consulter un document pour son seul contenu, mais d'interroger la place qu'il occupe dans un corpus, questionner ses métadonnées, ses occurrences internes et externes.A l'heure du "big data", de nouveaux cas d'utilisation émergent dans le monde de la recherche et plus spécifiquement des humanités numériques.* »[13]. Les technologies Big data font leur entrée à la BNF pour exploiter plus encore

les données issues de l'archivage du web !

D'après le rapport de Gail Truman[14], daté de mars 2016, 23 institutions de par le monde ont un programme d'archivage du web. Celles analysées sont des institutions publiques qui ont pour mission de rendre accessibles les ressources aux chercheurs et aux historiens (bibliothèques, archives et musées). Les organisations faisant un usage commercial des données archivées ont été écartées de ce rapport. Les archives du web sont dotées d'une valeur patrimoniale et scientifique.

Les technologies du Big data viennent apporter un nouvel éclairage sur ces corpus encore très peu exploités du fait de leur volumétrie, de leur variété et de leurs aspects dynamiques.

POUR ALLER PLUS LOIN

Workflow de l'archivage du web à la BNF :

http://netpreserve.org/sites/default/files/resources/Putting%20it%20all%20together.pdf

Exemples de valorisation des archives du web :

- À la BNF : http://www.bnf.fr/fr/collections_et_services/anx_pres/a.parcours_guides_arch_internet.html

- Sur Internet Archive : https://archive.org/web/

---

[1] Initiatives d'archivage du web. En Suède : http://www.kb.se/english/find/internet/websites/ . En Australie : http://pandora.nla.gov.au/ . Au Portugal : http://arquivo.pt/?l=en

[2] Siméon, G. Données le vertige. [en ligne] http://www.liberation.fr/futurs/2012/12/03/donnees-le- vertige_864585

[3] Tual, M. Vingt ans d'archivage du web : les coulisses d'un projet titanesque. [en ligne] http://www.lemonde.fr/pixels/article/2016/10/26/vingt-ans- d-archivage- du-web- un-projet- titanesque_5020433_4408996.html

[4] Rebourg, A. Donald Trump Président : le site de la Maison Blanche n'est plus disponible en espagnol. [en ligne] http://www.lci.fr/international/donald-trump- president-le- site-internet- de-la- maison-blanche- n-est- plus-disponible- en-espagnol-2023074.html

[5] https://sbforge.org/display/NASDOC52/NetarchiveSuite+Overview

[6] Voir définition dans le glossaire.

[7] Source : http://europa-eu- audience.typepad.com/fr/2009/06/format-warc- publi%C3%A9e-en- tant-que- norme-iso- pour-la-pr%C3%A9servation-du- web.html

[8] Texier, B. Big data : l'Ina se plonge dans un lac de données. [en ligne] http://www.archimag.com/univers-data/2015/11/26/big-data- ina-plonge- lac-donn%C3%A9es

[9] Voir tutoriel : http://dai.ly/x5bmq3j

[10] Lieux de consultation des archives web collectées par l'INA : http://www.inatheque.fr/consultation.html

[11] Wayback Machine d'Internet Archive : https://archive.org/web/

[12] Mussou, C. « Et le Web devint archive : enjeux et défis », Le Temps des médias 2012/2 (n° 19), p. 259-266. DOI: 10.3917/tdm.019.0259

[13] Bibliothèque Nationale de France http://actions-recherche.bnf.fr/BnF/anirw3.nsf/IX01/A2016000032_prefiguration- d-un- service-de-fourniture- de-corpus- numeriques-a- destination-de- la-recherche

[14] Truman, G. Web Archiving Environmental Scan. Harvard Library Report. 2016 https://dash.harvard.edu/handle/1/25658314

Business Intelligence & Big Data

■ Alexandre Hubert

Digital meeting – Business Intelligence (source : https://www.novaloca.com)

Trop d'organisations commettent encore l'erreur suivante : trop d'états statistiques, utiles lors de leur mise en production, continuent d'être exploités par routine mais parasitent, au fil du temps, la visibilité, la lecture de la performance. Ce surplus d'information tue la décision[1].

Le fait de définir des indicateurs est un investissement, en temps et en coût de mise en oeuvre notamment : mobilisation des experts métiers (définir, valider), achat de solutions logicielles et de prestations techniques etc. Partir de ce que l'on souhaite mesurer avec fiabilité pour construire l'environnement technique correspondant. Ceci est rentable si les parties se focalisent tout d'abord sur l'essentiel en cherchant à bâtir un système d'indicateurs cohérent et pertinent tout en s'assurant que l'élaboration de chaque indicateur en « vaut le coût[2] ». Des projets à ne pas prendre à la légère donc.

Piloter l'activité

Construire des tableaux de bord synthétiques

La description en français « non technique », en « langage métier », minimise les quiproquos, évite à ces indicateurs de subir des aléas indésirables lors de l'évolution des systèmes d'information. Ils pourront constituer un référentiel pour l'entreprise et permettront ainsi, pour commencer, une comparaison des performances dans le temps. L'idée est aussi de constituer des tableaux de bord qui permettront en outre d'avoir une vision du lien opérationnel/décisionnel[3].

Définir des indicateurs, c'est aussi faire une synthèse des informations de pilotage, c'est-à-dire perdre de l'information mais au profit de plus sens.

De fait, comment si savoir un sportif s'améliore en dehors de son propre ressenti ? Pour déterminer s'il y a amélioration ou non, il faut disposer d'indicateurs de performance et être au clair avec ce dont on parle précisément.

Des principes très anciens !

« *If you can't measure it, you can't improve it* », « *Administrer, c'est d'abord définir* », « *Le début de la sagesse est la définition des termes* » s'exprimaient déjà respectivement Lord Kelvin, Confucius et Socrate[4].

Etablir un dictionnaire des indicateurs permet de se fixer les objectifs suivants[5] :

- « Disposer d'un référentiel d'indicateurs cohérents et pertinents par domaine.

- Surveiller l'usure et garantir l'efficacité du pilotage sur la durée.

- Responsabiliser les personnes émettrices et destinataires de chaque indicateur ».

Système d'information et aide à la décision

Système d'information

Le Dictionnaire de l'Information définit[6] le système d'information comme un *« ensemble cohérent constitué par l'identification et la description des processus liés à une tâche, un ensemble de tâches ou un métier et des produits et services d'information nécessaires au fonctionnement de ces processus. Par extension l'expression « système d'information » désigne les technologies*

et les moyens informatiques (matériels et logiciels) qui assurent le stockage, le traitement et la diffusion des données sous forme électronique dans une organisation. »

Business Intelligence

La *business Intelligence*[7] – *intelligence d'affaires* ou *informatique décisionnelle* dans la langue de Molière – « *englobe les solutions IT apportant une aide à la décision aux professionnels avec, en bout de chaîne, des rapports et tableaux de bord de suivi des activités de l'entreprise à la fois analytiques et prospectifs* ».

La *business intelligence* vise à construire des environnements - des *univers* - dans lesquels des données sont architecturées techniquement pour faire l'objet d'analyses multidimensionnelles. Elle permet de modéliser des données selon plusieurs axes (chiffre d'affaires / zone géographique, catégorie de client, de produit...). La finalité est d'offrir à certains utilisateurs prédéterminés (management intermédiaire, analystes etc.) la possibilité de manipuler aisément des objets en utilisant la modélisation précitée sur des interfaces conviviales. En effet, les solutions modernes de business Intelligence ont pour objectif de créer un système dans lequel on peut sélectionner, croiser, utiliser et mettre en exergue de façon ergonomique les indicateurs souhaités de façon récurrente ou ponctuelle. Cela, sans pour autant passer de temps pour les obtenir et se focaliser ainsi sur la construction, l'exploitation des tableaux de bord et l'analyse métier.

Photo by William Iven on Unsplash

Rien de tel qu'un exemple. Un centre de relation clients de taille significative cherche à clarifier, définir ce qu'il entend par *nombre de ventes par mois.*

« Ah, bon, n'est-ce pas simple à obtenir une stat. aussi commune ? ».

Sans entrer dans une explication jargonneuse, il est probable que les services *Marketing*, *Finances*, *Fidélisation* ou *Réclamation* ne perçoivent pas une vente de la même façon, sous le même angle. En d'autres termes, une vente n'aura pas la même signification pour eux car les règles de gestion pour les comptabiliser, ou du moins les quantifier pour commencer, varient d'un service à l'autre : sa valeur à l'instant *t* pourra ne pas être la même. Il s'agit donc de clarifier, voire d'ajouter un niveau de granularité supplémentaire à ce concept pour qu'il corresponde à un élément fiable ayant fait consensus.

Les différents cas de figure viendront alimenter la réflexion des analystes. Les équipes pluridisciplinaires pourront ensuite construire, avec les experts techniques, ces *objets* en tenant compte dans cet exemple de la mécanique et des spécificités métiers (cf. gestion de la relation client : promesse, vente, vente contestée, annulation etc.) dans le contexte de l'organisation. Ces objets pourront être croisés avec une période de temps (jour, mois, année) pour arriver à l'analyse multidimensionnelle précitée et donc l'indicateur *nombre de ventes par mois.* Les spécifications, formalisées dans des documents de référence, contiendront précisément ce que l'on entend par tel ou tel indicateur.

Pour mener à cet objectif, il convient de bien étudier *ab initio*, avec les experts métiers, les indicateurs que l'on souhaite avoir à disposition en vue de construire les entrepôts de données et l'ensemble des mécanismes et environnements techniques intermédiaires. Pour être au clair avec ce que l'on évoque ici, tout en restant dans la vulgarisation, il faut avoir à l'esprit l'idée suivante : les systèmes techniques sont construits pour récupérer et répliquer des données émanant des systèmes informatiques de production et approvisionner un entrepôt de données. Ces mécanismes, appelés *Extract, Transform & Load* (ETL)[8], récupèrent les données.

On part tout d'abord d'une réplication plus ou moins complète des bases de données des systèmes d'information de l'entreprise vers un entrepôt de données (*datawarehouse*).
Pourquoi ? D'une manière générale, faire des requêtes en bases de données en temps réel sur un environnement de production nuit à la stabilité des systèmes d'information en production et n'est pas conseillé. Nous n'entrerons pas dans le détail à ce niveau car ce n'est pas l'objectif.

Les mécanismes ETL agissent comme des bras qui vont sélectionner les données utiles. En fin de chaîne, on arrive à des bases de données (*datamart*), plus petites, spécifiquement taillées pour que les logiciels de *business intelligence* fournissent les *objets* et *dimensions* dont les analystes ont besoin. On pourra ainsi avoir un *datamart* spécifique pour tel ou tel service. La technique s'adapte au besoin fonctionnel d'où l'importance de co-construire tout cela.

Big Data et nouvelle donne

Le site *bigdata.fr* explicite le concept éponyme de la manière suivante[9] :

« Littéralement, ces termes signifient mégadonnées, grosses données ou encore données massives. Ils désignent un ensemble très volumineux de données qu'aucun outil classique de gestion de base de données ou de gestion de l'information ne peut vraiment travailler. […]. Cependant, aucune définition précise ou universelle ne peut être donnée au Big Data. Etant un objet complexe polymorphe, sa définition varie selon les communautés qui s'y intéressent en tant qu'usager ou fournisseur de services ».

Le Big Data peut être vu comme un phénomène majeur marquant une nouvelle ère : les sociétés modernes doivent désormais composer avec la collecte et la profusion de données mais surtout avec l'ensemble des enjeux qui en découlent : impacts économiques et sociétaux. Tout cela est encore très souvent mal compris. Les quantités de données considérables découlent de trois phénomènes comme l'explique Jean-Daniel Fekete[10] :

« 1. la capacité de stockage a augmenté de façon exponentielle et nous pouvons trouver des disques durs de plusieurs téraoctets (milliers de milliards d'octets) pour moins de 100 euros,

2. nous capturons des données à une vitesse jamais égalée avec nos satellites, nos capteurs de pollution, nos dossiers médicaux informatisés, nos enregistrements de transactions informatisées, les articles mis en ligne sur le Web, etc.

3. les archives de données collectées depuis la nuit des temps sont maintenant mises à disposition sur Internet, DVD ou serveurs de données visibles en ligne ».

On utilise souvent les 3 V : Volume, Vélocité, Variété pour caractériser le big data mais le considérer uniquement comme un « seul » gros volume de données et le traduire ainsi est réducteur : ce n'est pas vraiment la quantité de données qui fait la nouveauté mais plutôt la combinaison[11] des 3 V.

De nombreux experts et observateurs considèrent que le Big Data est aussi déterminant que la révolution industrielle tant il ouvre nombre de perspectives dans tous les secteurs de l'économie. Il fédère en outre une multitude de communautés d'intérêt, pluridisciplinaires.

Business Intelligence et le Big Data

« Et le Big Data, alors ? On parlait d'entrepôt de données précédemment ! Le Big data, entre-t-il dans la Business intelligence ? Je suis un peu perdu là »

Pas de panique ! Il est toujours délicat de mettre en rapport des termes de ce type car ce sont ce que l'on appelle des *buzzwords*. Nous entendons précisément par *buzzword*[12] un *« terme ou une expression de jargon qui est utilisée, pendant une certaine période, comme slogan pour désigner une nouveauté (technologie, produit, concept, etc.) et ainsi attirer l'attention sur cette nouveauté. Son utilisation donne l'impression qu'il s'agit de quelque chose d'important et à la mode ainsi qu'une impression de compétence auprès du public ».*

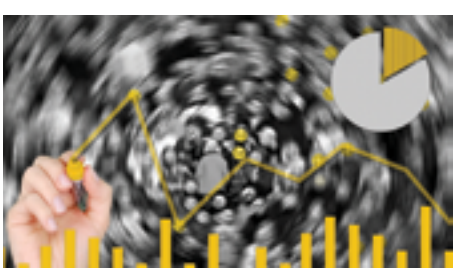
La *business intelligence*, dans l'acception proposée précédemment, a globalement 25-30 ans, voire davantage selon les degrés de maturité, et selon ce que l'on met derrière ces domaines et termes. Professionnels et experts ne sont d'ailleurs pas toujours d'accord.

Elle vise à mettre en oeuvre des indicateurs basés sur les systèmes d'information « interne », c'est-à-dire les applications de l'entreprise, pour constituer des tableaux en bord en vue d' « aider à la décision ».

Le problème avec les buzzwords, c'est aussi qu'ils représentent une innovation « nommée » dans leur contexte bien précis. C'est comme imaginer de faire allusion à des « nouvelles » technologies de l'information et de la communication (NTIC) en 2050. Elles sont nouvelles, celles-là, fin des années 1990, début des années 2000 même si, bien sûr, elles continuent de progresser. L'appellation « art contemporain » peut être un autre exemple illustratif dans un registre totalement différent.

Il faut en outre bien avoir à l'esprit que le terme *intelligence* est très connoté *renseignement* chez les anglo-saxons. Buzzwords et décalage linguistique (voire culturel) n'aident pas !

Le big data ne recouvre pas forcément les mêmes choses selon les agents économiques qui prononcent cette expression. Il devient aussi un écosystème économique[13]. Précédemment, il a été explicité comme un phénomène mais les marketeurs de solutions logicielles aiment à semer une certaine confusion en mélangeant le phénomène et les technologies pour l'exploiter. Ils ajoutent parfois des V comme « Vérité » ou « Valeur » !



CCO

Différence avec l'informatique décisionnelle

La principale différence concerne les données et leur utilisation[14].

D'un côté, l'informatique décisionnelle porte sur l'utilisation de statistiques descriptives et sur des données à forte densité en information afin de mesurer des phénomènes, détecter

des tendances..., pour mémoire, la statistique descriptive[15] *« vise à décrire de façon synthétique et parlante des données observées pour mieux les analyser ».*

De l'autre, comme le résume bien Wikipédia, le Big data mobilise des compétences pour utiliser des *« statistiques inférentielles, sur des données à faible densité en information dont le grand volume permet d'inférer des lois [...] donnant dès lors [...] au big data des capacités prédictives. [...]. L'informatique traditionnelle, informatique décisionnelle comprise, est basée sur un modèle du monde ; le big data vise à ce que les mathématiques trouvent un modèle dans les données ».*

Alors que la complexité va continuer d'augmenter, la principale valeur ajoutée des Big Data réside dans l'analyse[17] :

- Prédiction : qu'est-ce qui est le plus susceptible de se produire ?
- Estimation : quelle est la valeur future d'une variable ?
- Description : quelles relations existent dans les données ?
- Simulation : que pourrait-il arriver ?
- Prescription : que doit-on faire[18] ?

Les finalités ne sont pas donc les mêmes. Le principal changement résulte non pas des technologies mais surtout du changement d'architecture qu'implique le management du Big Data pour des questions stratégiques.

---

[1] Foucard, J.-L. (2014). La boîte à outils du pilote des systèmes d'information. Paris : Dunod.

[2] *Ibidem*.

[3] *Ibidem*.

[4] Bazin, H. Recueil de pensées sélectionnées par Hubert BAZIN. [En ligne] http://bazin-conseil.fr/pensees.html.

[5] Foucard, J.-L. (2014). op. cit.

[6] Cacaly, S., Le Coadic, Y.-F., Pomart, P.-D. & [et al.] (2008). Dictionnaire de l'information. Paris : A. Colin.

[7] Journal du Net. Business Intelligence (informatique décisionnelle). [En ligne] http://www.journaldunet.com/solutions/intranet-extranet/business-intelligence/.

[8] « Extraire, Transformer & Charger ».

[9] Définition : Qu'est-ce que le Big Data ? [En ligne] http://www.lebigdata.fr/definition-big-data.

[10] Fekete, J.-D. La visualisation analytique, pour comprendre des données complexes. [En ligne] https://interstices.info/vismaster.

[11] Fouchard, A.-C. Perdu(e) dans la définition du "Big Data"? [En ligne] http://www.theaccessgroup.fr/ressources/blogs/perdu-e-dans-la-d%C3%A9finition-du-big-data/.

[12] Wikipedia. Buzzword. [En ligne] https://fr.wikipedia.org/w/index.

[13] Cartier, M. Le 21e siècle : le Big Data. [En ligne] http://www.21siecle.quebec/table-des-matieres-2/le-big-data/.

[14] Wikipedia. Big data.
[En ligne] https://fr.wikipedia.org/w/index.php?title=Big_data&oldid=138130937.

[15] Ayache, A. & et Hamonier, J. Cours de Statistique Descriptive. [En ligne] http://math.univ-lille1.fr/~ayache/cours_SD.pdf.

[16] Wikipedia. op. cit.

[17] Madsen, M. Bi isn't big data and big data isn't BI », 05:55:14 UTC. [En ligne] https://www.slideshare.net/mrm0/bi-isnt-big-data-and-big-data-isnt-bi.

[18] *Ibidem*.

BIBLIOGRAPHIE

Ayache, A. et Hamonier, J. « Cours de Statistique Descriptive », [En ligne: http://math.univ-lille1.fr/~ayache/cours_SD.pdf].

Bazin, H., « Recueil de pensées sélectionnées par Hubert BAZIN », [En ligne: http://bazin-conseil.fr/pensees.html].

Cacaly, S. et Le Coadic, Y-F. et Pomart P.-D. et al. (2008). *Dictionnaire de l'information*, Paris : A. Colin

Cartier, M. « Le 21e siècle: le Big Data », [En ligne: http://www.21siecle.quebec/table-des-matieres-2/le-big-data/].

« Définition: Qu'est-ce que le Big Data? » [En ligne: http://www.lebigdata.fr/definition-big-data].

Fekete, J.-D., « La visualisation analytique, pour comprendre des données complexes », [En ligne: https://interstices.info/vismaster]. .

Foucard, J.-L. (2014). *La boîte à outils du pilote des systèmes d'information.* Paris : Dunod.

Fouchard, A.-C. *Perdu(e) dans la définition du "Big Data"?* [En ligne] http://www.theaccessgroup.fr/ressources/blogs/perdu-e-dans-la-d%C3%A9finition-du-big-data/.
Journal du Net. « Business Intelligence (informatique décisionnelle) », [En ligne: http://www.journaldunet.com/solutions/intranet-extranet/business-intelligence/].

Madsen, M. *Bi isn't big data and big data isn't BI », 05:55:14 UTC. [En ligne] https://www.slideshare.net/mrm0/bi-isnt-big-data-and-big-data-isnt-bi.*

Wikipedia. (2017, juin 8). Buzzword. Consulté 1 juin 2017, à l'adresse https://fr.wikipedia.org/w/index.php?title=Buzzword&oldid=138016653

OURAGAN ANNONCÉ DANS LES SI

■ Béatrice Tifoli

L'intégration du Big Data dans les systèmes d'information n'est pas récente. Toutefois, lors de sa mise en oeuvre, la protection de l'information doit être garantie sur quatre critères : l'authenticité, la sécurité, l'intégrité et la pérennité.

Le terme Big Data est apparu récemment permettant de traduire le concept ou la méthodologie d'un système ayant la capacité à traiter en temps quasi réel d'immenses gisements d'informations, de grandes masses de données structurées, semi-structurées, voire non structurées et disparates pour extraire de leur recoupement de nouvelles valeurs demeurées jusqu'alors inconnues (Zolunski, 2015).

Toutefois, son intégration dans les systèmes d'information n'est pas récente. Comment concilier les anciennes technologies avec le Big Data afin de garantir une protection de l'information ?

Explications

Une grande partie des cas d'usage du Big Data existait déjà avant son émergence, déclarait Elisabeth Gabbay[1], consultante confirmée en technologies et service de l'information décisionnelle. Il remonterait d'ailleurs depuis les années 1944[2]! En effet, des technologies ont été acquises par l'entreprise pour permettre des traitements de l'information. Face aux enjeux du marché concurrentiel, les entreprises ont dû adapter leurs méthodes et outils afin de pressentir la naissance d'une tendance, de prédire l'évolution d'un phénomène ou encore de procéder à des contrôles pour découvrir des fraudes. D'où l'émergence du marketing. Toute l'approche systémique (c'est-à-dire la visualisation globale du système de l'entreprise) était donc basée sur l'amélioration du fonctionnement externe en assurant que le système d'information garantissait la qualité, la conception d'un produit orienté client et la conception orientée processus et centrée métiers.

Toutefois, face à un marché concurrentiel de plus en plus rude où le client possède de nouvelles exigences, et dans un contexte de profusion de l'information, de multiplication des données et de production massive des documents, l'entreprise se retrouve à acquérir de nouvelles technologies permettant le traitement rapide, sécurisé et pertinent de l'information. Certains logiciels détenus dès lors sont poussés vers la sortie. L'approche engagée par l'entreprise évolue donc vers une approche systémique contrôlée interne et externe. Données (statistiques) et informations (rapports par exemple) issues de ces contrôles deviennent la valeur première du système de l'entreprise.

En effet, l'enjeu majeur d'aujourd'hui n'est plus de trouver l'information mais de trouver immédiatement l'information ayant de la valeur par le biais d'indicateurs de mesure amenant ainsi vers une stratégie : tel est le rôle des solutions Big Data. L'une des technologies ayant évolué avec l'arrivée du Big Data est la Gestion Electronique du Document (GED) (du

moins ceux qui ont fait l'objet d'une évolution[3]).

En effet, les solutions GED sont déjà capables de traiter des gros volumes, des dizaines de millions de données et de documents complexes pour les analyser et en tirer de la valeur (Guerrier 2013). Pour rappel, sa mission première a été de mettre en place au sein de l'entreprise une gestion centralisée de l'information et des documents afin de satisfaire le critère de sécurité et de protection de ces derniers. Ensuite, elle a évolué en GED collaborative en proposant des workflows pour finir par son intégration dans le système d'information de l'entreprise permettant ainsi de proposer des indicateurs nécessaires à la stratégie de celle-ci.

Qu'est ce que le système d'informations (SI) d'une entreprise ?

Le système d'information (SI) représente l'ensemble des éléments (personnel, matériel, logiciel, ...) participant à la gestion, au traitement, au transport et à la diffusion de l'information au sein de l'organisation. Depuis les années 1990, les entreprises ont conçu leur SI orienté qualité, satisfaction client, processus métiers afin d'obtenir un fonctionnement plus efficace (voire un meilleur rendement). Les résultats étaient souvent perçus financièrement peu de temps après sa mise en place.

Des acronymes comme ERP[4], ECM[5], CMS[6] ont donc fait leur apparition dans le vocabulaire entrepreneurial. Afin de clarifier la situation, on définira simplement que l'ERP (Enterprise Ressource Planning) correspond au progiciel de gestion intégré (PGI), l'ECM (Enterprise Content Management) correspondant à la gestion des contenus de l'entreprise peut être assimilé à la GED et le CMS (Content Management system) au programme informatique utilisant une base de données et permettant de gérer de A à Z l'apparence et le contenu d'un site Web. Ces acronymes font partie du SI.

Dans une organisation, le SI intervient entre le système opérant (système de production) et le système de pilotage (système de décision). Il apporte les informations opérationnelles nécessaires aux activités et les informations indispensables à la gestion. Ainsi, le SI peut être désigné SI pour le management ou pour la gestion (cela dépend de l'angle d'approche de l'utilisateur du SI). Ci-après une représentation sommaire du système d'information au sein d'une entité.

Pourquoi faire évoluer le SI d'une entreprise ?

À l'ère du numérique où les nouvelles technologies produisent de plus en plus de données et d'informations, et dans un contexte de concurrence marketing, il est nécessaire pour les entreprises de faire évoluer leur SI. En effet, la multiplicité de l'information est constatée à l'intérieur de l'entreprise comme aux portes d'entrée et de sortie de celle-ci. Le SI doit permettre la survie de l'entreprise c'est-à-dire qu'elle doit connaitre le présent, anticiper le futur, comprendre les problèmes récurrents d'un produit ou service afin de prendre des décisions rapidement se basant sur des informations fiables et pertinentes. Les solutions mises en place dès lors trouvent leur limite face à la diversification et la volumétrie des données à traiter rapidement.

C'est là que les solutions Big Data font leur apparition.
Selon Gartner, elles visent à tirer un avantage concurrentiel des méthodes de collecte, d'analyse et d'exploitation des données qu'on ne pouvait utiliser jusqu'à présent du fait des contraintes économiques, fonctionnelles et techniques liées à la volumétrie, à la vitesse de traitement et à la variété des données à considérer. (Business&Decision).
En effet, elles arrivent à prendre en compte une variété de données soit dans leur typologie (structurées, non structurées, semi-structurées), soit dans le format de documents dont les données sont issues (pdf, txt, graph, audio, vidéo...), soit de leur source (Internet, réseaux sociaux, Open Data...). Elles peuvent également traiter une volumétrie de données s'élevant en téraoctets en toute vélocité. Enfin, le dernier avantage de ces solutions est qu'elles permettent la véracité de l'information afin d'en tirer des informations à valeur ajoutée pour l'entreprise.

Le premier domaine impacté par le Big Data est le marketing où l'analyse des réseaux sociaux devient la matière première pour connaitre la tendance du consommateur, l'avis d'un produit afin de l'adapter, ... Qu'advient-il alors des sondages d'opinion des clients ? En reprenant l'exemple de la GED, d'une armoire de stockage de fichier informatique, celle-ci a évolué en proposant des workflows (travail collaboratif) au sein de l'entreprise en dématérialisant les procédures existantes. Elle a ensuite pu prendre en compte différents formats comme l'email afin de répondre aux nouveaux besoins de l'entreprise.

En somme, les solutions Big Data sont une opportunité pour les entreprises dans un marché fort concurrentiel. Toutefois, l'entreprise doit rester prudente quant à leurs utilisations. En effet, à ce stade, les solutions Big Data n'ont pas vocation à gérer toutes les données mais du moins celles les plus pertinentes pour l'entreprise. Une phase de transition vers le Big Data est nécessaire.

Comment les solutions Big Data s'intègrent dans le SI ?

Les solutions Big Data intègrent le SI comme tout projet d'intégration de nouveau logiciel. Toutefois, un important travail en amont est nécessaire afin que l'entreprise obtienne les résultats escomptés. Trois questions simples doivent être posées.

De quoi dispose-t-on ?

Il s'agit bien sûr de dresser un état des lieux des sources et ressources détenues par l'entreprise tant au niveau de son architecture informatique (réseaux, matériel, logiciel...) que son architecture informationnelle (procédures décisionnelles, financières ou autres, stockage physique des documents, référentiels...). Il est important de dresser l'inventaire des référentiels et des typologies des données existants au sein de l'entreprise afin que la solution Big Data puisse « traduire » les nouvelles données dans l'objectif qu'elles y trouvent leur place (ou pas).

Que veut-on ?

Il s'agit de poser clairement les attentes que devra satisfaire la solution Big Data. Néanmoins, la décision ne devrait pas être issue de la hiérarchie mais d'un travail de collaboration avec tous les acteurs impactés par ce nouveau logiciel.

Quelles sont les solutions existantes ?

Il s'agit de sonder les solutions Big Data existantes sur le marché. Ce recensement permettra, en s'appuyant sur l'inventaire dressé ainsi que les besoins de l'entreprise, d'identifier la solution Big Data adéquate à l'entreprise. La réussite de l'intégration d'un logiciel relève de deux ressources « binômes » : l'informaticien et le professionnel de l'information. Suite à la mise en place d'un tel projet, l'expert en interprétation des données ou « Data scientistes » est nécessaire. Enfin, le cas échéant, l'entreprise sera confrontée à une mise à jour ou une évaluation de ses référentiels et processus.

Où se situe la protection de l'information dans le SI ?

Le SI a pour objectif d'identifier, de collecter, de traiter et de diffuser l'information ou la donnée à la bonne personne attributaire sous une forme utilisable et compréhensible afin de faciliter sa prise de décision dans son action ou sa stratégie. Afin d'atteindre ses objectifs, il est nécessaire que l'information soit protégée contre tout risque de destruction, de détérioration, de piratage ou de divulgation. La protection de l'information est donc assurée par le dispositif technique, l'application et les usagers du SI.

Les mesures de protection de l'information de l'infrastructure

Le système informatique est protégé par les anti-virus, les anti-spams, les anti spyware /adware et est renforcé par :
- Les contrôles d'accès : les authentifications des usagers à l'entrée du SI,
- Les sécurités réseaux : l'utilisation de pare-feu, de routeur filtrant permet de contrôler les échanges entrant et sortant du SI,
- Les solutions de virtualisation : réseau local virtuel (ou VLAN) et un réseau virtuel privé (VPN),
- Les solutions d'externalisation : les serveurs dupliqués ou hébergés hors site de l'entreprise.

Les mesures de protection de l'information de l'application

Les applications disponibles sur le marché doivent satisfaire la réglementation en vigueur (loi Informatiques et libertés, code des postes et télécommunications, horodatage, gestion des droits, RGS, RGI, RGDP...) et notamment certaines normes. Par exemple, dans le cas d'une GED où le workflow est permis, la sécurité de l'information sera assurée par sa mise en conformité à la norme AFNOR 15489-2. Certaines applications proposent également les contrôles d'accès et les solutions d'externalisation pour permettre une protection de l'information suite à un incident (incendie, dégât des eaux...). Les données pertinentes pour l'entreprise contenues dans l'application sont stockées en silo et ne peuvent être vues par une tierce personne non autorisée.

Les mesures de protection de l'information du personnel

La charte informatique est un document technique définissant les règles de bonne utilisation du SI mis à la disposition des usagers. L'utilisation de l'intranet, d'internet et de la messagerie y est souvent définie. Enfin, la sensibilisation est l'une des mesures d'accompagnement à privilégier auprès des utilisateurs afin qu'ils connaissent leurs droits et obligations et qu'ils adoptent les bons réflexes lors de son utilisation. Un climat de confiance doit être instauré.

Les solutions Big Data peuvent-elles passer outre ces mesures de protection de l'information ?

La loi métropolitaine « Informatique et Libertés[7]» garantit au citoyen qu'en cas de collecte de données à caractère personnel, celui-ci doit être consentant et qu'il doit savoir qui collecte les informations et pour quel usage.

Toutefois, les solutions Big Data remettent en question cette garantie.
En effet, dans le cadre de l'intégration de solution Big Data, la récolte d'informations (essentiellement issues de profil) permettant le croisement des données de différentes sources et donc de différentes bases de données s'avère essentiel pour obtenir des informations à valeur ajoutée pour l'entreprise. La garantie définie par la loi de 1978 est renforcée par l'adoption en 2016 par le Parlement Européen du RGPD[8]. À noter que les états membres ont jusqu'en 2018 pour promulguer et donc appliquer ledit règlement sur leur territoire respectif. Ainsi, les entreprises seront contraintes d'informer dans un langage simple et clair lors de récolte de données personnelles et

de concevoir des fonctionnalités par défaut et des produits de sorte de collecter et traiter le moins possible des données à caractère personnel.

Il est tout de même rappelé que le transfert de données à caractère personnel vers un pays tiers à l'Union européenne est, en principe, interdit, sauf si le pays de destination assure un niveau de protection suffisant des données personnelles.

La mise en œuvre par les entreprises de tout ce dispositif sera latente par rapport à l'entrée en vigueur de normes juridiques et donc par effet de ricochet par rapport à l'évolution technologique. La capacité analytique trouvera certainement ses limites (elle peut être augmentée jusqu'à un certain moment d'après le principe d'Heisenberg et de Godel) entraînant de fait que la pertinence des informations stockées et traitées baisse. Alors qu'adviendra–t–il de toutes ces normes juridiques lors du revirement de la technologie vers une analyse synthétique ?

[1] http://www.redsen-consulting.com/fr/inspired/Dataanalyse/cas-dutilisation-Big-Data

[2] http://www.forbes.com/sites/gilpress/2013/05/09/a-veryshort-history-of-Big-Data

[3] http://www.archimag.com/demat-cloud/2017/04/11/ged-fini

[4] http://www.entreprise-erp.com/articles/definition-erp.html

[5] https://www.ged.fr/ecm/

[6] https://www.medialibs.com/cms/definition-cms/

[7] Loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés

[8] Règlement général sur la protection des données

BIBLIOGRAPHIE/WEBOGRAPHIE

Assouline, M (2016). Le « BIG DATA » : Quelles précautions juridiques prendre en matière de données de protection des données personnelles ? », Solutions numériques. [En ligne] http://www.solutionsnumeriques.com/le-Big-Data-quelles-precautions-juridiquesprendre-en-matiere-de-protection-des-donnees-personnelles/.

Gabbay, E. (2013). Cas d'utilisation du Big Data », REDSEN Consulting. [En ligne] http://www.redsenconsulting.com/fr/inspired/Data-analyse/cas-dutilisation-Big-Data.

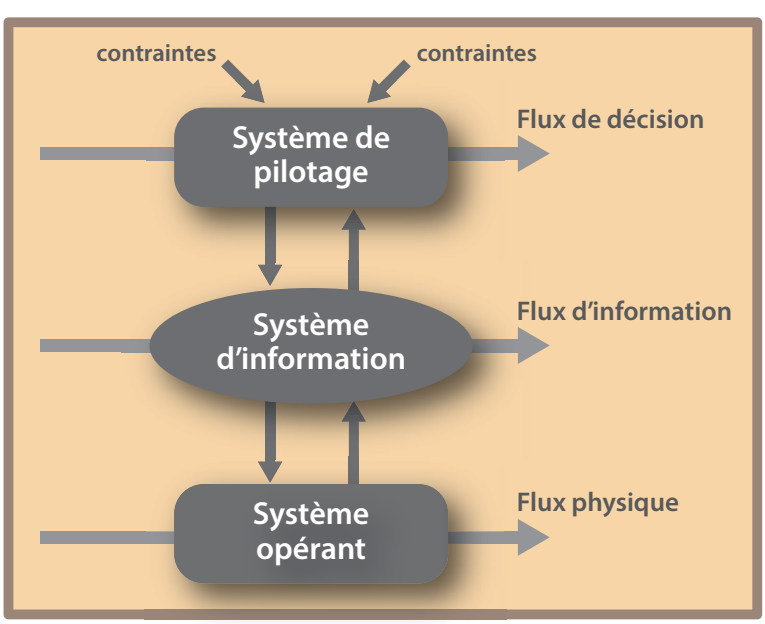
Fernadez, A. « Le Big Data et la conception du système d'information », Piloter.org. [En ligne] http://www.piloter.org/business-intelligence/Big-Data-si.htm.

Guerrier, P. (2013). Fabrice Lacrois – Antidot : « Le Big Data résume notre message : les données ont de la valeur », ITexpresso, 2013, [En ligne] : http://www.itespresso.fr/fabricelacroix-antidot-Big-Data-resume-message-donnees-valeur-63685.html.

Hubert, C. (2016). « La GED, une réelle opportunité pour l'analyse Big Data », GPO Mag.fr. [En ligne] http://www.gpomag.fr/web/technologies/la-ged-une-reelleopportunite-pour-l-analyse-Big-Data.

« Chapitre 1 : Système d'information de l'entreprise », [En ligne] http://profs.vincimelun.org/profs/adehors/CoursWeb2/Cours/Ch1/Ch1.php.

Zolynski, C. (2015). « Big Data : pour une éthique des données », La revue I2D – Information, données & documents, pp. 24-26, n°2.



L'intéraction des systemes dans une entreprise (Sources http://profs.vincimelun.org/)

Conter les données ? Datadesign et datastorytelling : l'univers visuel des Big Data

■ Aurélie Quinodoz

« Un dessin vaut mille mots » Jacques Bertin

Aussi monumentales soient-elles, les données massives (Big Data) restent muettes sans traitement et assistance pour les comprendre.
Les représentations graphiques pour la visualisation des données (data driven) sont ainsi devenues des instruments couramment utilisés pour rendre intelligibles les données. La visualisation constitue en outre une aide à la décision (insight). En effet, se limiter aux outils statistiques courants ne permet plus de saisir la richesse et la structure des données.
Pour élargir les types d'explorations possibles, les savoir-faire du design renouvellent les approches de la visualisation. Plus récemment encore, le récit via le data storytelling contribue à la compréhension des données massives.

La data visualisation est l'étude, la science ou l'art de représenter des données de façon visuelle. Graphiques, camemberts, diagrammes, cartographies, chronologies, infographies, ces représentations visuelles rendent les données parlantes, sans effort excessif. Les volumes des Big Data compliquent néanmoins l'accès aux informations. Le design, parce qu'il « *recherche de nouvelles manières d'apparaître (faire-être) en vue de rendre possibles d'autres expériences du monde (faire-faire*[3]*) », répond aux besoins de comprendre et de faire agir. Dans l'univers numérique, le design a recours à «une matière calculée comme matière à modeler en elle-même et pour elle-même, [...] en donnant forme avant tout à des matériaux informatisés*[4] ».

Couplée avec le data storytelling, la visualisation des données accroît encore la capacité informationnelle grâce au renouvellement de l'expérience visuelle. Puissant vecteur de communication, le data storytelling est la mise en récit des données des données. « Processus qui consiste à traduire les analyses de données en langage courant dans le but d'orienter une décision ou une action[5] », le data storytelling utilise la narration pour simplifier l'interprétation des données.

À partir d'exemples de réalisations concrètes, cet article est une contribution à la compréhension des apports du design et du storytelling pour la visualisation des données

La visualisation des données n'est pas nouvelle !

Depuis toujours, la représentation visuelle est un outil au service de la réflexion. Au premier chef, la cartographie a permis d'élaborer des outils communs pour appréhender l'espace et les ressources. D'abord descriptive, la cartographie a considérablement évolué pour représenter des données de plus en plus abstraites. Mais c'est grâce aux travaux scientifiques du cartographe Jacques Bertin (Bertin, 2013), que les bases de la sémiologie graphique

sont jetées. Bien que sa recherche se développe sur des cartes statiques, la sémiologie graphique demeure l'outil principal pour l'étude et la réalisation de représentations graphiques et plus largement pour la visualisation des données.

Les apports de la sémiologie graphique

Étude de la signification, la sémiologie graphique est une méthode pour faire passer un message visuel. À partir de l'analyse du message à transmettre, elle conditionne le choix des supports, le choix des icônes et des légendes et la structuration de chartes graphiques. L'exemple le plus connu est celui du Code de la route mais les applications sont bien plus nombreuses.

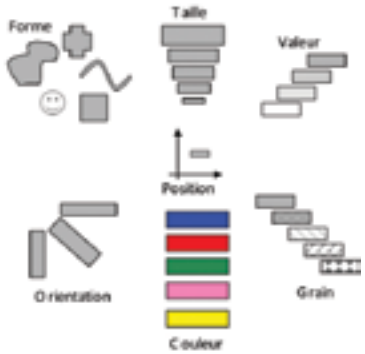
S'appuyant sur l'analyse d'une multitude d'images, Jacques Bertin[6] a dégagé les principes essentiels à la transmission d'un message. Véhiculé par un ensemble de structures visuelles (formes, pictogrammes...), le message devient une image rationnelle dont l'information est immédiatement perceptible.

Cette construction d'images s'appuie sur les lois de la perception visuelle. Se distinguant de l'image abstraite sujette à des interprétations contradictoires, le graphique réduit la confusion par des éléments définis au préalable (système monosémique = une seule signification) appelés variables.

Par exemple, la variation d'une couleur traduit une relation d'ordre et de différences. Ce sont les relations mises en oeuvre à travers les variables qui rendent compréhensible l'information contenue dans le graphique.

Et dans l'univers du Big Data ?

Les règles de la sémiologie graphique, écrites en 1965, restent valables dans l'univers du Big Data et de l'informatique. Le nombre d'informations susceptibles d'être comprises par le cerveau humain demeure «autour de trois et ne



Variables visuelles définies par Jacques Bertin

dépasse pratiquement jamais sept[7] ». Avec le nombre croissant de données à traiter, l'assimilation des informations se complexifie.

Le design d'information ou encore Data design, riche des études antérieures et de ses méthodes propres, renouvelle dès lors les capacités de visualisation des données.

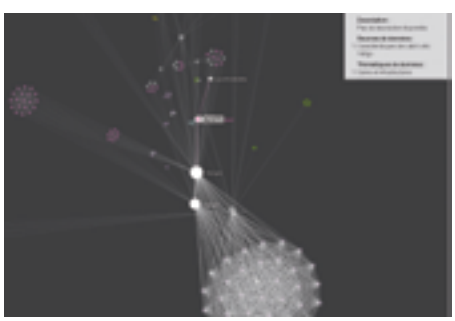
Le design dans l'univers du Big Data

Les évolutions des métaphores visuelles sont parallèles à l'évolution du monde, de plus en plus abstrait et complexe à appréhender. Mais la nécessité de façonner un langage commun demeure pour donner du sens aux données et les rendre pertinentes.

L'exemple de la SCNF présenté par Caroline Guérin[8], designer et chef de projet – SNCF Innovation & Recherche, montre que le design, au-delà de l'aspect graphique, produit des nouvelles pistes pour explorer les données. Afin d'appréhender et anticiper les besoins de mobilité, la SNCF dispose de très grands volumes de données. Pour leur traitement, elle dispose d'une cellule spécifique composée de statisticiens, de mathématiciens et d'une designer.

L'approche du design vient enrichir la conception des visualisations de données en amont de la démarche. Puisant dans la théorie de Jacques Bertin, chaque acteur participe à la création progressive d'un langage commun fait de variables visuelles assimilables et reconnaissables par tous. Il s'agit, comme le définit Stéphane Vial, de considérer que « l'exercice cognitif qui consiste à structurer les données se confond avec l'exercice perceptif qui consiste à afficher la donnée ».

Le design d'information n'est pas la dernière touche aux traitements statistiques des données. Il réduit l'ambiguïté pour favoriser le jugement humain. Il est la composante qui réunit les compétences pour identifier ce que l'on souhaite voir, comment le faire à voir et même, comme le précise Caroline Guérin, « représenter les données pour qu'elles racontent des histoires ». Le design permet de co-construire des algorithmes de représentation au service des algorithmes statistiques.



Cartographie des données : donner à voir l'open data SNCF par PELLOUX Sonia Sciences du Design, 2/2016 (n° 4), p. 11-15

Un outil privilégié : le graphe

Dépassant les outils traditionnels de la statistique (tableau, listes...) au service de la visualisation, le graphe est la forme privilégiée de la représentation des données. Outil mathématique, le graphe est une représentation simplifiée constituée d'un ensemble de nœuds et de liaisons formant un réseau.

Pour expliquer les liaisons possibles, Isabel Meirelles utilise l'exemple du réseau Internet où « les noeuds sont les ordinateurs ou routeurs, les liaisons étant les câbles ou les connexions sans fil[9]». Les liaisons

permettent d'exprimer les interactions entre les noeuds. Changement majeur dans l'appréhension des données, celles-ci construisent des chemins et montrent les relations plus qu'elles ne mesurent les causes. En outre, la prolifération d'informations à visualiser avec les Big Data nuit à la lisibilité par la démultiplication des points et lignes. « *l'exercice cognitif qui consiste à structurer les données se confond avec l'exercice perceptif qui consiste à afficher la donnée* ».

Le design d'information n'est pas la dernière touche aux traitements statistiques des données. Il réduit l'ambiguïté pour favoriser le jugement humain. Il est la composante qui réunit les compétences pour identifier ce que l'on souhaite voir, comment le faire à voir et même, comme le précise Caroline Guérin, « représenter les données pour qu'elles racontent des histoires ». Le design permet de co-construire des algorithmes de représentation au service des algorithmes statistiques.

Quelques exemples :

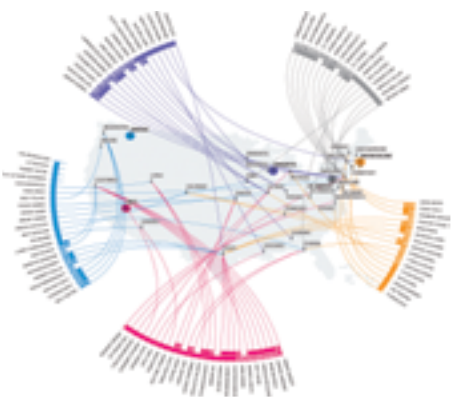
- Been there / une réalisation du studio Catalogtree installé à Arnhem (NL) et Berlin (DE)
Réseau de déplacements des coaches du basket américain pour découvrir de nouveaux champions. Projet commandité par la chaîne ESPN.

- Data visualisation of a social network / une réalisation de Felix Heinen

Data visualisation of a social network / Felix Heinen

Par l'utilisation des codes postaux, un fournisseur peut vérifier la pénétration de son réseau dans différentes agglomérations. Deux grandes affiches (200 x 90 cm - 80 x 36 pouces) montrent la variété et l'attitude des membres d'une communauté Internet comme MySpace.
Ici sur la première affiche, les fonctions et les informations complémentaires (âge, formation, situation familiale, sexe, fréquence de connexion) fournissent une information précieuse pour définir une stratégie commerciale pour une clientèle ciblée.

- Marking Europe High Speed / une réalisation de Lust
Carte des lignes ferroviaires à grande vitesse en Europe. Les croisements des lignes et les points de passage sont les marqueurs des forces économiques et touristiques de l'Europe.



Been there, réalisation du studio Catalogtree installé à Arnhem (NL) et Berlin (DE). Source : Catalogtree.net

Les exemples de visualisations réalisées par des designers graphiques se multiplient. La nécessité de présenter les données pour offrir une compréhension de l'information est largement acquise. Les graphiques interactifs étendent encore la popularité du design d'information.

Dans la foulée de la data visualisation, le data storytelling se développe rapidement et fournit un autre puissant vecteur d'information. Nous passons alors d'une visualisation centrée sur les données, pas toujours compréhensibles immédiatement, à une mise en récit qui ajoute une couche informationnelle.

Tirer une histoire des données

Le récit est depuis toujours un outil de transmission de valeurs et un outil d'édification. Parce qu'il mobilise les émotions, le récit permet de retenir l'attention pour mieux appréhender les messages. Les contes pour enfants en sont l'exemple majeur.

Le data storytelling associe cette technique simple de transmission aux données les plus complexes pour les transformer en histoires. Le monde de l'entreprise et plus particulièrement la Business Intelligence[10] regarde de près la mise en scène des données et ses possibilités. Sur son blog, Brent Dykes, directeur de la stratégie pour la société Domo[11], explique pourquoi le data storytelling est essentiel pour le succès des entreprises. Alliant aspects visuels des données et histoires construites, les présentations deviennent dynamiques et modifiables en jouant sur les variables et les données. Elles sont construites autour de récits prospectifs en s'appuyant sur les visualisations. Elles plongent l'auditeur dans une immersion totale pour lui faire toucher du doigt les futurs possibles, pour le pire comme le

Les mots du Big Data

Data driven : Data driven est une expression anglaise qui signifie littéralement « conduit ou tiré par les données[1] ». Le data-driven est souvent associé à une dimension marketing. Il s'agit de faciliter la prise de décision grâce au sens produit par les données

Insight : Notion encore plus floue ! L'insight vient du moyen anglais « inner sight » autrement dit le regard intérieur, la sagesse. Pour les psychologues[2], l'insight est le processus qui consiste à résoudre un problème soudainement. Les idées de créativité et de solution sont au centre de l'insight. Dans le contexte du Big Data, les données doivent être mises en forme pour produire des solutions.

meilleur. Les données ne viennent plus seulement valider des hypothèses de comportement ou d'actions... elles les font apparaître !

Pour bien comprendre le fonctionnement du data storytelling, il utilise un visuel qui illustre efficacement le processus à l'oeuvre.

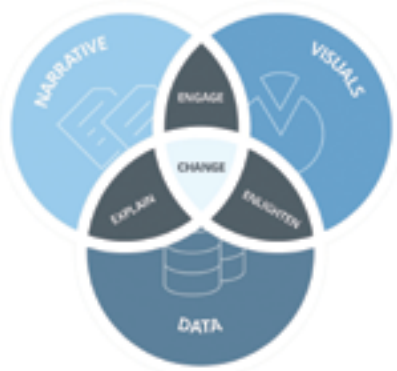


Schéma extrait de l'article Data storytelling pt. 2: Why telling stories with data is essential to success.

Domo et d'autres entreprises telles que Qlik, le leader du marché d'outils d'analyses visuelles, développement des applications analytiques pour l'aide à la décision. Aux classiques tableaux de bord sont inclues des fonctionnalités permettant de raconter une histoire et de naviguer parmi des données interactives. Au sein d'une même interface, les données sont reliées, contextualisées et structurées en multipliant les sources validées. À titre d'exemple, la société Idacity[12] propose un outil complet pour traiter les données et organiser des story-boards en fonction du récit.

Les outils ne doivent cependant pas faire oublier que le meilleur ambassadeur des données reste le conteur des données, une fonction spécifique à créer pour piloter les outils et les données, au service d'une stratégie.

Pour conclure provisoirement

Au-delà de rendre compte des multiples facettes des données, le design s'adapte et révèle les besoins pour construire des outils qui fondent un équilibre entre complexité et simplicité. À la complexité des attentes et des besoins, le designer construit de la « simplexité ». Alain Berthoz l'a définie comme «l'ensemble des solutions trouvées par les organismes vivants pour que, malgré la complexité des processus naturels, le cerveau puisse préparer l'acte et en projeter les conséquences[13] ». De cette façon, les designers aident à poser les problèmes pour donner forme aux données.

Les grandes entreprises, et plus discrètement les collectivités publiques, sont les premières intéressées par le design d'information et le data storytelling. Mais le monde universitaire commence aussi à leur porter attention. Le design comme la narration sont interrogés en vue d'apporter un éclairage sur les nouvelles pratiques et sur les nouveaux usages qu'ils engendrent.

Dans la présentation du dossier spécial intitulé Design et Transmedia de la Revue française des sciences de l'information et de la communication, Stéphanie Cardoso et Mélanie Bourdaa affirment vouloir « saisir la valeur de la rencontre entre deux disciplines de SHS, toutes deux au coeur de problématiques sociétales, le design et le Transmedia Storytelling, dans une perspective communicationnelle[14]».

De son côté, le sociologue Dominique Cardon constate que l'évolution des techniques de visualisation est « le témoin d'une transformation des manières de construire et de représenter la société dans le travail des sciences humaines[15] ». Il montre l'opération de renversement des régimes d'interprétation des données par les Big Data. Avec la prolifération des données enregistrées quotidiennement, le temps de la publication, qui venait traditionnellement après le travail de filtrage, est inversé. Avec la généralisation des graphes, il conclut que « la visualisation précède l'interprétation[16]».

Enfin, le sociologue Antoni Casilli appelle à « introduire de la transparence[17] ». Cela veut dire que si les graphiques d'information apportent des explications à des problèmes complexes, le risque d'une manipulation des données ne doit pas être perdu de vue. Dans le domaine du data storytelling, de mêmes données peuvent produire des récits totalement différents, adaptés à l'audience visée.

Savoir qui contrôle la construction des données et qui détermine ce qui est donné à voir ou à entendre s'impose pour veiller à la qualité de l'information.

Ces nouvelles façons de représenter les données transforment en profondeur le monde professionnel comme les approches en sciences humaines et sociales. Outre les espoirs et les craintes suscités, le phénomène du Big Data crée de nouveaux types d'informations dont le traitement et l'impact sur la société commencent seulement à émerger.

---

[1] Définitions marketing. Définition : data driven. En ligne : http://www.definitions-marketing.com/definition/data-driven/

[2] Papanicola, M. Qu'est ce qu'un insight ?. En ligne : https://fr.slideshare.net/marcpapanicola/quest-ce-quun-insight

[3] Vial, S. (2013). L'être et l'écran. Paris : PUF, 2013. p.262.

[4] Ibid. p. 264

[5] LeMagIT. Data Storytelling. En ligne : http://www.lemagit.fr/definition/Data-Storytelling

[6] Bertin, J. (2013. Sémiologie graphique : Les diagrammes, les réseaux, les cartes. Paris : Editions de l'Ecole des Hautes Etudes en Sciences Sociales.

[7] Ibid. p.14

[8] Pelloux, S. (2016). Cartographie des données : donner à voir l'open data SNCF », Sciences du Design, 4(2). p.12

[9] Meirelles, I. (2014). Design de l'information : Représenter visuellement les informations. Parramon.

[10] Nous retiendrons la définition simple du site Définitions Marketing. « La business intelligence est l'ensemble des techniques et procédures de traitement de l'information à des fins de prises de décisions managériales». En ligne : https://www.definitions-marketing.com/definition/business-intelligence/

[11] Dykes, B. Data storytelling pt. 2: Why telling stories with data is essential to success. En ligne : https://www.domo.com/blog/data-storytelling-success/

[12] Idaciti. En ligne :https://hello.idaciti.com/

[13] Berthoz, A. (2009). La simplexité. Paris : O. Jacob.

[14] Cardoso, S. & Bourdaa M. (2017). Design et Transmedia : projet, expérience usager, worldbuilding au coeur des disciplines SHS. Revue française *des sciences de l'information et de la communication*, 10. En ligne : http://rfsic.revues.org/2558

[15] Cardon, D. (2012). Regarder les données. Multitudes, 49(2). En ligne : http://www.cairn.info/revue-multitudes-2012-2-page- 138.htm

[16] Ibid.

[17] Casilli, A. (2015). Voir pour comprendre et comprendre pour voir. I2D – Information, données & documents, 52(2). p. 41.

Atelier HSL : Making Europe High speed, par Lust (2003 - 2004)
Source : https://lust.nl/#projects-34

POUR ALLER PLUS LOIN

Avenati, O. et Chardel, P.-A. (2016). *Datalogie : Formes et imaginaires du numérique*. Paris : Loco.

Bertin, J. (2013). *Sémiologie graphique : Les diagrammes, les réseaux, les cartes.* Paris : Editions de l'Ecole des Hautes Etudes en Sciences Sociales.

Bourquin, N. et Klanten, R. (2009). *Data flow, design graphique et visualisation d'information*. Paris : Thames & Hudson.

I2D – Information, données & documents 2015/2 (Volume 52). 88 pages.

Lloveria, V. (2014). Data design-moi un mouton. De la data visualisation au data storytelling chez Michael Paukner. *Communication & Organisation, 46*, p. 99-112.

Meirelles, I. (2014). *Design de l'information : Représenter visuellement les informations*. Parramon.

Pelloux, S. (2016). Cartographie des données : donner à voir l'open data SNCF. *Sciences du Design*, *4*(2), p. 11-15.

Vial, S. (2013). *L'être et l'écran*. Paris : PUF.

Quelques sites

Catalogtree.net. https://fr.pinterest.com/source/catalogtree.net

INTACTILE. *Visalisation et Big Data*. [En ligne] : http://intactile.com/index.php/etudes-de-cas/visu-info-big-data

Cartographie sémantique. [En ligne] : http://www.cartographie-semantique.fr/

Gouvernance des données à caractère personnel et sensible dans le secteur juridique

LES PERSPECTIVES DE L'OPEN DATA JURIDIQUE

■ Adèle Duret

« Ouvrir les données juridiques, et les données de jurisprudence en particulier, c'est ouvrir un immense terrain de jeu pour l'application des techniques les plus avancées de traitement des données. Pour l'intérêt final du justiciable et de la justice elle-même : espérance de gain de cause et de dédommagement, pilotage statistique d'activité de tribunaux ou de résultats. Et par intérêt commercial. Car l'open data c'est l'opportunité de développer des produits et services commerciaux innovants et disruptifs. »

Comme le souligne Guiraude Lame, l'enjeu de la mise en place d'un open data des décisions de justice est double. Il répond à la fois à la volonté de transparence de la vie publique et à un enjeu économique et social lié à la valorisation des données[1].

Jusqu'alors, le principe était l'accès aux décisions de justice. Cet « open access » (mise gratuitement à la disposition du public) est prévu depuis le décret du 7 août 2002 et prévoit « de faciliter l'accès du public aux textes en vigueur ainsi qu'à la jurisprudence [2]». Pour ce faire, le site Légifrance met en ligne une sélection d'arrêts.

La loi de la République numérique, dite loi Lemaire, va plus loin : « *la totalité des décisions de premier et second degré seront mises à disposition en open data* » (articles 20 et 21). Il est également question, dans l'optique d'une mise à disposition des données de décisions en open data, d'utiliser un standard ouvert et réutilisable.

Comme on le voit, le secteur juridique n'échappe pas à l'ouverture des données détenues par les acteurs publiques en open data. Toutefois, la composition des décisions de justice pose question. En effet, elles sont composées à la fois de données publiques et de données personnelles.

D'ailleurs, la loi Lemaire stipule d'ailleurs que la mise à disposition des jurisprudences ne pourra être effectuée que « *sous réserve du principe du respect de la vie privée* » c'est-à-dire sous condition de protéger les données à caractère personnel ou sensible. On entend par données personnelles Article 2 de la loi Informatique et Liberté : « *Constitue une donnée à caractère personnel toute information relative à une personne physique identifiée ou qui peut être identifiée directement ou indirectement, par référence à un numéro d'identification ou à un ou plusieurs éléments qui lui sont propres* [3]».

Dès lors, la question se pose : quelle gouvernance des données à caractère personnel et sensible est à mettre en place ?

Pour ce faire, nous verrons quelles sont les méthodes actuelles de protection des données personnelles et sensibles (I) puis en étudierons les limites (II-2.1) et les solutions envisagées pour y faire face (II-2.2).

Les méthodes de protection des données personnelles et sensibles

L'utilisation des décisions de justice contenant des données à caractère personnel ou sensible ne peut être possible que dans un des cas suivants :

1 - Consentement des personnes concernées,
2 - Anonymisation ou pseudoimisation préalable,
3 - Dispositif légal le permettant.

Dans le cas des décisions de justices, il est fort compréhensible que la pseudonymisation ou l'anonymisation est la solution la plus fréquemment envisagée.

Pseudonymisation

Ce terme porte sur la protection apportée directement sur un texte, et en aucun cas sur tous les autres éléments extérieurs qui pourraient permettre la réidentification des personnes physiques. Autrement dit la pseudonymisation s'entend, hors de tout croisement de jeux de données, comme étant un moyen efficace de rendre anonyme une personne physique ou sein d'un ensemble d'informations[4].

La pseudonymisation est une mesure technique qui permet de compliquer l'identification d'une personne. Autrement dit, il s'agit d'un procédé par lequel les données à caractère personnel ou sensible sont rendues inidentifiables. Par exemple, on peut remplacer un nom de famille par un autre nom de famille ou par une chaine de caractères aléatoires.

Il est question dans ce procédé des données contenues uniquement dans le corps de la décision. En d'autres termes, la pseudonymisation s'intéresse exclusivement aux données intrinsèques aux jurisprudences.

A ce jour, cette pseudonymisation des données à caractère personnel et sensible peut intervenir aux trois moments suivants[5]:

- L'écriture en « privacy by design » : prise en compte du risque de ré-identification en amont, dès l'écriture de la jurisprudence. Cette méthode peut nécessiter l'aide d'un outil d'aide à la rédaction.

- La pré-pseudonymisation :
la pseudonymisation advient une fois la décision de justice signée. Ce sont les greffiers qui en sont chargés.

- La pseudonymisation centralisée : elle s'appuie sur un acteur tiers qui peut être par exemple un diffuseur public ou privé (comme un éditeur juridique).

Elle intervient dans deux cas :

- Editeurs juridiques classiques : récupèrent les jurisprudences directement auprès des autorités judiciaires. En général, ils « pseudonymisent » les décisions une fois reçues. Ce processus n'est pas géré en interne par les éditeurs juridiques. Si l'on prend le cas de l'éditeur juridique Lexbase, cette dernière fait appel à un prestataire extérieur appelé NUMELEX.

- Sociétés de diffusions du droit plus nouvelles (appelées « legal tech ») : collectent les décisions de justice via les données mises à disposition en open data. La pseudonymisation n'est pas directement traitée par ces dernières, elle intervient en amont[6].

Actuellement, c'est la technique de la pré-pseudonymisaion qui est envisagée. La Cour de Cassation s'inscrivant dans le mouvement de l'open data, teste un logiciel d'aide à la pseudonymisation qui sera en soutien au contrôle humain. Ce logiciel substituera le nom de famille des parties ou témoins par une lettre de l'alphabet et gommera l'adresse. D'autres éléments d'identification seront aussi effacés[7].

On l'a vu, la pseudonymisation ne s'intéresse qu'aux données incluses dans la décision de justice. Elle assure la non-identification au sein de la jurisprudence.

Une des limites majeures de cette méthode est qu'elle ne garantit pas la non ré-identification des personnes qui interviendrait par un croisement avec d'autres données. En d'autres mots, elle rend l'identification plus complexe mais pas impossible.

Une des solutions serait d'avoir recourt à l'anonymisation qui permet d'empêcher la ré-identification d'une donnée personnelle ou sensible car le recoupement des informations est rendu impossible.

Selon le rapport d'information rédigé par le Sénat de 2014, « *l'anonymisation peut être définie comme l'opération de suppression de l'ensemble des informations permettant d'identifier directement ou indirectement un individu, contenues dans un document ou une base de données* [8]».

L'anonymisation peut être définie comme l'opération de suppression de l'ensemble des informations permettant d'identifier directement ou indirectement un individu, contenues dans un document ou une base de données[9].

Selon l'Avis 05/2014 sur les techniques d'anonymisation adopté le 10 avril 2014 par groupe de travail « article 29 » sur la protection des données[10], trois critères sont nécessaires pour assurer l'anonymisation complète :

- L'individualisation : possibilité d'isoler des données spécifiques à une personne d'un ensemble de données,

- La corrélation : revient à l'éventualité de relier des données pour en détacher des données sur un même individu,

- L'inférence : consiste à déduire des informations concernant un individu.

Ce même groupe, composé de chaque autorité indépendante de protection des données personnelles dans l'ensemble de l'Union européenne (en France, cette autorité correspond à la Commission Nationale de l'Informatique et des Libertés, la CNIL), examine quelles sont les méthodes possibles pour la mise en oeuvre d'une solution d'anonymisation. En définitive, les techniques d'anonymisation se composent de deux modèles :

- La *randomization* est une altération de la véracité des données personnelles

afin d'atténuer le lien entre les données et la personne concernée. Cela signifie que les données seront assez incertaines pour ne pas pouvoir être reliées à un individu précis. Par exemple, une de ces méthodes s'appelle « l'ajout de bruit » qui conduit à la distension du lien entre les données et la personne réelle. Un autre de ces moyens, « à permutation » consiste à mélanger les valeurs liées à une personne de telle manière que certaines données sont artificiellement rattachées à des personnes concernées différentes.

- La *généralisation* consiste à généraliser ou diluer les données personnelles de sorte à ce qu'elles perdent en précision. Ces données ne sont plus ciblées et ne caractérisent plus une personne avec exactitude. Elles ne pointent pas avec précision une personne mais peuvent être partagées par plusieurs personnes. Ainsi, l'identification devient moins aisée. Ainsi, on affectera à un individu une région plutôt qu'une ville ou une semaine plutôt qu'un jour. En pratique, une anonymisation aujourd'hui partielle est seulement possible. En effet, de nombreux obstacles tant techniques que financiers se dressent devant la mise en oeuvre de cette pratique.

Le procédé n'est pas encore abouti puisque l'anonymisation complète n'est pas encore garantie. Pour y parvenir, de nombreuses recherches et études devraient être envisagées générant un coût non négligeable.

De surcroît, le contexte de l'ouverture des données complexifie encore plus ce travail. Les jeux de données de décisions de justices qui seront mises à disposition dans la mouvance de l'open data juridique sont énormes. Il s'agit alors de trouver des techniques pour éviter de ré-identifier un individu par le croisement de ces mêmes données avec le big data alors même que les processus de ce dernier sont en constante amélioration.

Des techniques d'anonymisation complète ou partielle existent déjà. C'est le cas de l'assistant d'anonymisation NOME[11], solution logicielle développée au Québec. Entre autres, NOME permet d'automatiser le remplacement des noms mentionnés dans un document par leurs initiales ou par d'autres signes. Toutefois, afin d'assurer cette anonymisation, une intervention humaine demeure indispensable. L'on peut prendre l'exemple d'un justiciable s'appelant M. Pierre. Dans le corps de la jurisprudence, une coquille s'est glissée le dénommant alors M. Piere. Si un des noms d'un des participants au système judiciaire est mal orthographié, une vérification humaine est indispensable, la technologie NOME ne permettant de relever et corriger cette erreur.

Afin de palier à ce manque, nous pourrions imaginer une solution basée sur le « machine learning ». On pourrait alors concevoir un système d'apprentissage automatique qui apprendrait à reconnaitre les erreurs dans les chaînes de caractères, ou mieux, à repérer elle-même les données à caractère personnel ou sensibles.

Toutefois, cette perspective, bien que très prometteuse, n'est pas encore disponible !

Quelle que soit la solution envisagée – pseudonymisation ou anonymisation – un contrôle humain est nécessaire. Pourtant, comme le souligne Axelle Lemaire, ancienne Secrétaire d'État chargée du numérique :

« *Imposer une vérification et une analyse des risques à chaque fois qu'il est question de publier le jugement d'un tribunal de première instance reviendrait en pratique à empêcher la mise en place de l'open data, puisque les ressources des tribunaux ne le permettraient pas*[12]».

Aujourd'hui, les moyens techniques disponibles rendent chimériques la garantie de la non ré-identification tout comme la nature irrémédiable de l'anonymisation.

Les limites des méthodes actuelles

Risques de ré-identification

L'efficience des méthodes d'anonymisation n'est pas garantie. Selon la nature des données ayant trait aux caractéristiques personnelles ou sensibles d'un individu pris dans le processus « raté » d'anonymisation, le risque de ré-identification est réel.

Cette non maitrise de ce risque pourrait donner lieu à des diffusions fortuites d'informations personnelles.
A ce propos, dans son article Ouverture des bases de données de jurisprudence et protection des données sensibles[13] Nathalie Metallinos donne un exemple éclairant.
Il en est ainsi d'un arrêt rendu par la cour d'appel de Nîmes publie sur Legifrance concernant un jeune homme âgé de 19 ans, qui a été grièvement blessé alors qu'il disputait un match de rugby et qui a subi un traumatisme crânien ayant occasionné un grave handicap.

Si son nom patronymique a bien été occulté, il n'en reste pas moins d'autres informations éminemment sensibles: date et lieu de naissance, numéro de sécurité sociale, commune d'habitation, date de l'accident, nom des clubs sportifs.

Cette affaire a bénéficié d'une forte couverture médiatique régionale et a été relayée en tant que « faits divers » dans les journaux sportifs.

Sans bénéficier d'une forte expertise en informatique, il suffit d'une simple recherche sur un moteur de recherche pour relier cette jurisprudence à l'évènement concerné et retrouver ainsi l'identité de la victime.

Si la ré-identification est aisée via des méthodes de recherche d'informations classiques et manuelles via un moteur de recherche quelconque, les technologies développées par le big data intensifient le risque de ré-identifier un individu.

Cyber sécurité - Haker

Dans le contexte d'ouverture des données publiques via l'open data et l'émergence voire l'explosion des techniques issues du big data, l'on se rend bien compte que la ré-identification devient plus facile pour des données à caractère personnel ou confidentielles même plus « protégées ».

Par le croisement de données issues de différentes sources d'informations, la déduction d'une information initialement manquante ou parcellaire est rendue possible.

La réutilisation des données est un des principes de base de l'open data. D'ailleurs, le W3C a formalisé une démarche afin d'évaluer la valeur du jeu de données mis à disposition. Plus le nombre d'étoiles attribué sera grand, plus la valeur de réutilisation sera grande. Les données de jurisprudences « 5 étoiles » lorsqu'elles sont publiées en « Linked Open Data » facilitent le croisement des données.

Par la suite, l'on pourrait imaginer faire corréler des données entre elles.
Par exemple, le nom d'un magistrat pourrait être relié à son style de rédaction. Grâce à des méthodes spécifiques, il est possible de rattacher un texte court à un style de rédaction spécifique d'un auteur donné.[14] Il serait alors possible de rattacher et ainsi démasquer d'éventuelles « tromperies » afin de vérifier si l'auteur d'une décision est vraiment le juge qui l'a « signé ».[15]

Solutions envisagées

Afin de contourner les risques manifestes et inévitables de ré-identification des individus par les techniques non-abouties de l'anonymisation et de la pseudonymisation, une gouvernance des données à caractère personnel et confidentiel doit être mise en oeuvre.

Anticiper pour réduire le risque

Il s'agirait ici selon le rapport du Sénat[16] d'anticiper les risques en prévoyant dès sa perspective de mise en « open data » les modalités de ré-identifications éventuelles ainsi que les mesures à suivre sur le risque de ré-identification.

Il serait donc souhaitable de prévoir le plus en amont possible les risques de réutilisations « malveillantes » de données à caractère personnel ou sensible possible. Ainsi, il serait plus aisé de procéder aux ajustements nécessaires.

En cas de risques concernant les données à caractère personnel ou sensible, les administrations pourraient avoir la possibilité de ne pas mettre à disposition le jeu de données concernées. L'on pourrait aussi envisager une mise à disposition partielle revenant ainsi à l'administration judiciaire de procéder à la sélection des jurisprudences publiables sans risques sur les données permettant la ré-identification.

Le bienfondé et les finalités d'une telle démarche semblent évidents. Il s'agit notamment pour l'administration, de se prémunir du risque de « procès » des personnes ainsi ré-identifiées. L'on pourrait imaginer dans ce contexte, l'administration rendue coupable de la non-protection des données à caractère personnel et sensible.

Toutefois, cela pose inévitablement la question même de l'open data. En effet, peut-on encore qualifier d'open data des données mises à disposition uniquement sur sélection et non réutilisables ? Dans un tel cadre, les caractéristiques de l'open data des jurisprudences voulues par le législateur ne semblent plus garanties.

Rappelons ici que le critère principal est la mise à disposition totale permettant une réutilisation. L'on reviendrait ainsi au schéma connu de la diffusion des jurisprudences qui a cours avant l'émergence de l'open data, à savoir des jurisprudences sélectionnées et diffusées garantissant seulement « l'open access ».

Adopter une gouvernance des données à caractère personnel et confidentiel efficiente

Il est prévu également de permettre à l'administration de procéder à des contrôles sur l'usage qu'il est fait des données. En effet, avec l'évolution technologique toujours croissante, il est envisageable qu'une technologie de demain rende possible une ré-identification, là où la technologie en cours à l'instant de la mise en ligne ne le permettait pas.

Une veille active et probante devra être mise en place afin de répondre à ce risque. Il s'agirait donc de surveiller la réutilisation des données qui serait effectuée. On peut être en droit de se demander ici dans quelle mesure et par quelle méthode une telle veille serait effectuée. Compte-tenu de la lourdeur d'un tel processus, s'il est réalisé entièrement par des humains, il semble difficilement réalisable.

Qui plus est, une des failles identifiées dans la mise en oeuvre de l'open data juridique est le manque d'accompagnement des administrations concernées par son déploiement.

Une des recommandations formulées dans le rapport du Sénat est de « mutualiser les bonnes pratiques et les diffuser (mise en ligne) : l'expérience en matière d'anonymisation des données personnelles pourrait être non seulement mutualisée, mais également mise à disposition de tous afin de sécuriser la diffusion de données publiques issues de données personnelles ».

En somme, la pseudonymisation est la solution pour le moment retenue par l'administration, l'anonymisation restant une exception tant les solutions techniques ne semblent pas encore abouties.

Les limites de ces deux méthodes conduisent à un risque accru des ré-identifications des individus. Il s'agirait donc de contourner ces méthodes en introduisant dans l'administration judiciaire une gouvernance des données opérante et efficiente pour y faire face.

C'est un préalable indispensable à la mise à disposition des décisions de justices en open data. Si son objectif affiché est d'accroitre la transparence de la vie publique et d'améliorer les rapports entre les administrations et les citoyens, nous ne saurions alors imaginer une justice dans laquelle les justiciables hésitent à saisir la justice par peur de voir leurs données personnelles ou sensibles dévoilées.

---

[1] Open Case Law. « Livre Blanc Open Data Jurisprudentiel ». p.9 http://openlaw.fr/, 1 janvier 2017. http://www.calameo.com/read/005075651e5c75eb4e579

[2] Décret n° 2002-1064 du 7 août 2002 relatif au service public de la diffusion du droit par l'internet. https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000000413818&categorieLien=id [consulté le 18/04/2017]

[3] Loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés [consulté le 18/04/2017]

[4] Cassar, Bertrand. « De la pseudonymisation et l'anonymisation des données à caractère jurisprudentiel ». Mémoire d'université, Université Paris 1 - Panthéon - Sorbonne, 2015. http://openlaw.fr/images/7/7d/Anom_Pseudo_Donn%C3%A9es_Justice_(CASSAR_B).pdf.

[5] Open Case Law. « Livre Blanc Open Data Jurisprudentiel ». http://openlaw.fr/, 1 janvier 2017., p.51 et 52 http://www.calameo.com/read/005075651e5c75eb4e579

[6] Cassar, Bertrand. « De la pseudonymisation et l'anonymisation des données à caractère jurisprudentiel ». Mémoire d'université, Université Paris 1 - Panthéon - Sorbonne, 2015. http://openlaw.fr/images/7/7d/Anom_Pseudo_Donn%C3%A9es_Justice_(CASSAR_B).pdf.

[7] Cour de cassation. « Anonymisation des décisions : la Cour teste un logiciel | Cour de cassation ». Cour de Cassation, 12 septembre 2016. https://www.courdecassation.fr/cour_cassation_1/open_data_7821/cour_teste_35045.html.

[8] Gorce, Gaëtan, et François Pillet. « La protection des données personnelles dans l'open data : une exigence et une opportunité ». Rapport d'information du Sénat. Sénat, 16 avril 2014. http://www.senat.fr/rap/r13-469/r13-4697.html.

[9] *Ibidem*.

[10] Le G29 publie un avis sur les techniques d'anonymisation. CNIL [en ligne]. avril 2014. https://www.cnil.fr/fr/le-g29-publie-un-avis-sur-les-techniques-danonymisation-0

[11] Pelletier, Frédéric, Luc Plamondon, et Guy Lapalme. « L'assistant d'anonymisation NOME », 2004. http://www.frlii.org/IMG/pdf/Pelletier_Paris2004_NOME.pdf.

[12] Séance du 27 avril 2016 (compte rendu intégral des débats). Sénat [en ligne]. [Consulté le 11 mai 2016]. http://www.senat.fr/seances/s201604/s20160427/s20160427_mono.html

[13] Cour de Cassation. « La jurisprudence dans le mouvement de l'open data : dossier documentaire / bibliographie », 20, 2016. https://www.courdecassation.fr/venements_23/colloques_4/2016_7423/open_data_35038.html

[14] Jardino, Michèle. Identification des auteurs de textes courts avec des n-grammes de caractères. Université Paris 3 - Sorbonne Nouvelle [en ligne]. http://lexicometrica.univ-paris3.fr/jadt/jadt2006/PDF/II-048.pdf

[15] Cassar, Bertrand. « De la pseudonymisation et l'anonymisation des données à caractère jurisprudentiel ». Mémoire d'université, Université Paris 1 - Panthéon - Sorbonne, 2015. http://openlaw.fr/images/7/7d/Anom_Pseudo_Donn%C3%A9es_Justice_(CASSAR_B).pdf

[16] Gorce, Gaëtan, et François Pillet. « La protection des données personnelles dans l'open data : une exigence et une opportunité ». Rapport d'information du Sénat. Sénat, 16 avril 2014. http://www.senat.fr/rap/r13-469/r13-4697.html.

L'exploitation des big data pour le Crowdfunding

■ Amélie Gallo

« Crowdfunding », la plupart des gens connaissent le principe sans forcément connaître le mot. Le crowdfunding, c'est le financement participatif, c'est simplement un outil de financement alternatif.
Mais quel rapport peut avoir ce système de financement et le big data ? C'est un sujet de réflexion qui, depuis quelques années intéresse, à tel point que la PSB (Paris Buisness School) a organisé le 23 mars dernier, pour la deuxième année consécutive une journée échange-atelier de réflexion sur cette thématique : « Big data et crowdfunding ».

Le crowdfunding pas si nouveau

Revenons tout d'abord sur le Crowdfunding. Le financement participatif en français est un moyen de financer un projet, quels que soient son ampleur et son secteur, par des financements privés, c'est-à-dire en évitant les acteurs traditionnels : les banques.

Ce mot anglais nouvellement employé en France désigne cependant un vieux procédé. Un procédé vieux comme le monde ? On ne peut pas le déclarer, cependant, cela existait en 1875 avec la statue de la liberté qui a été réalisée grâce à une levée de fonds privés et individuels. Le film « Shadows » de John Cassavetes, également. En 1958, le réalisateur lança un appel à la radio demandant à chaque auditeur d'envoyer un billet de 1 dollar. Plus récemment, on peut parler de la rénovation du prieuré du Mont Saint-Michel ou de l'appel en 2016 des Musées Royaux de Belgique pour la restauration du « Portrait de Suzanne Bambridge » de Paul Gauguin.

Dans un tout autre registre, on peut citer le financement de la campagne électorale de Barack Obama ou celle de « Mouvement en marche ». Bien sûr, ce sont les petits projets qui sont les plus nombreux : projets musicaux, associatifs, caritatifs, audiovisuels, création de start-up etc.

À chaque époque ses moyens de communication. Pour la statue de la liberté, l'information est passée essentiellement par les journaux, de nos jours, ce sont les réseaux sociaux qui diffusent les projets auprès d'un large public.

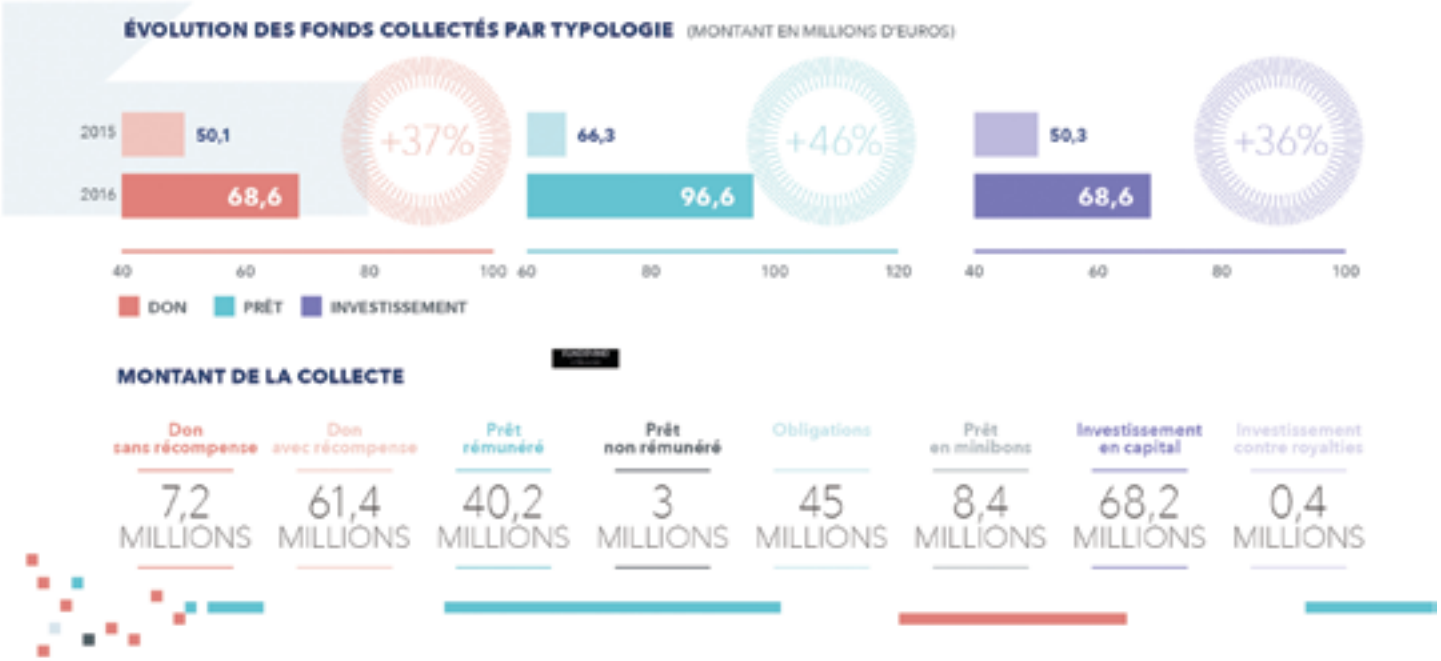
(source http://ict.io/)

Le fonctionnement du crowdfunding

On trouve plusieurs types de financements participatifs. On peut diviser d'un côté le don qui prend deux formes, on peut appeler ça du Love Money[1] :

- Le pur don (Crowdgifting ou crowdgiving) : on donne généreusement de l'argent pour une cause ou un projet qui nous tient à coeur sans la moindre attente en retour. Ce sont surtout les entreprises sociales et solidaires, certaines startups et quelques projets très innovants qui peuvent mener une campagne de crowdfunding en dons.

- La contrepartie (reward-based crowdfunding) : on donne de l'argent en échange de quelque chose de matériel. C'est la forme la plus connue de crowdfunding, auprès des



Le financement participatif en France : évolution des financement entre 2015 et 2016 (Source : FPF http://financeparticipative.org)

particuliers. Bien souvent, il s'agira de recevoir l'objet que l'on a contribué à créer : Cela peut être un CD du groupe dont nous aidons à sortir son premier album, un pass VIP pour une exposition ou l'inscription de son nom sur une plaquette à l'entrée d'un bâtiment dont nous avons contribué à sa rénovation, des photos de la distribution de nourriture que nous avons aidé à acheter etc.

De l'autre côté, on trouve le financement intéressé, c'est-à-dire le prêt d'argent avec retour sur investissement. La motivation du financeur, en plus de soutenir un projet qui lui tient à coeur, est donc la perspective de réaliser un gain financier.

- Le prêt (crowdlending): comme on aime être généreux mais sans vraiment donner, on prête simplement l'argent. Il s'agit là de l'une des formes les plus récentes du crowdfunding. Si le projet rapporte suffisamment, l'argent est rendu. Evidemment il y a un système d'intérêts avec un taux pré défini en fonction de la durée du prêt. Ce type de fonctionnement est plutôt récent. En effet, c'est à la suite des Assises du financement participatif, que l'ordonnance du 30 mai 2014 est mise en vigueur. Elle supprime le monopole des banques pour le prêt d'argent et permet donc la mise en place de prêts sur les plateformes de crowdfunding.

- L'equity-crowdfunding : ou « Crowdinvesting » ou encore « Crowdequity ». L'Equity consiste, pour un entrepreneur, à lever des fonds auprès des particuliers qui investissent de l'argent en échange d'actions dans la future PME ou start-up créée. La société Wiseed est la première plateforme de crowdfunding à s'être lancée dans l'investissement en 2008.

L'état du crowdfunding en France

En France, il est important de noter la création en août 2012, de l'association professionnelle des acteurs du crowdfunding « Financement Participatif France ». FPF est une association loi 1901 ayant pour objectif la représentation collective, la promotion et la défense des droits et intérêts des acteurs de la finance participative notamment auprès desautorités réglementaires pour faire progresser le financement de projets en France. Cette structure promeut activement ce mode de financement, permet aux acteurs de rester informés des décrets et législations et cadre le crowdfunding.

Chaque année, FPF, fait une analyse du crowdfunding en France et propose un baromètre à partir de l'analyse des données des plateformes adhérentes à l'association.

Pour l'année 2016 les fonds récoltés par le crowdfunding ont progressé de 40 % par rapport à 2015. Ceci est expliqué entre autres par la multiplication du nombre de plateformes et l'émergence de plateformes spécialisées. Cette hausse des montants collectés touche tous types de crowdfunding.

Il est intéressant de noter la différence des montants des collectes entre les dons ou prêts sans intérêt et toutes les autres formes de crowdfunding avec des récompenses ou avec investissements. On remarque que la croissance du secteur des entreprises a été essentiellement tirée par les levées en obligations (+86%, à 45 millions d'euros), qui concernent majoritairement les opérations de promotions immobilières. L'apparition

des «minibons» (destinés à faciliter et sécuriser les prêts) a drainé 8,4 millions d'euros, un bon score pour ce nouveau produit mis en place courant 2016.

Et le Big Data dans tout ça ?

Le crowdfunding utilise le pouvoir des big data et surtout leur analyse pour faire des prédictions et rendre les campagnes plus efficaces.
Le Big data est un outil qui est utilisé à deux niveaux dans le milieu du crowdfunding. Premièrement, du point de vue des investisseurs, il permet de prédire quel projet pourra rapporter de l'argent, de préférence, le plus d'argent possible. Deuxièmement, du côté des leveurs de fonds, l'analyse des big data permet, entre autres, de choisir les meilleurs moyens de communication autour du projet ou simplement de trouver les moyens de réussite de son projet.

Les lois du succès d'une campagne d'une levée de fonds reposent sur trois aspects :

- Evidemment, le projet doit tout d'abord être « impressionnant ». Plus un projet est innovant et plus il sort de l'ordinaire, plus il attirera du monde.

- Le projet doit être décrit sur une plateforme spécialisée et mis en avant. Il ne faut jamais hésiter à faire de la publicité pour un projet.

- Il est important de faire vivre la campagne de crowdfunding, jusqu'à la fin de la levée de fonds. Il convient de mettre à jour le projet le plus régulièrement possible, simplement pour signifier aux investisseurs que le projet est vivant et donc bien réel.

La prédiction

Le Graal pour les analystes des big data est de trouver un modèle d'analyse prédictif qui permet de dire rapidement si une campagne va réussir ou non. Ce modèle doit pouvoir donner les clés de la réussite d'une campagne de levée de fonds, c'est-à-dire, que le modèle doit pouvoir trier les projets en fonction de l'intérêt de l'investissement, pour que les funders puissent réussir leur investissement.

À cet égard, des chercheurs de l'Ecole d'informatique et communication et de l'Ecole Polytechnique Fédérale de Lausanne (EPFL) ont mis au point une méthode exclusive pour déterminer si les campagnes de crowdfunding réussiront dans les quatre heures suivant leur lancement (campagne de Kickstarter). Le taux de précision des prédictions est de 76%.

Les big data permettent d'analyser, de prédire et de réagir en conséquence. En comprenant très tôt si une campagne sera réussie ou non, cela permet aux gestionnaires de projets de faire des ajustements pour améliorer leurs chances d'être pleinement financés.

Mais l'analyse permet de comprendre les risques avant même le lancement de la campagne, par exemple :

- Quels sont actuellement les types de projets les plus populaires ?

- Quels sont les statuts de projets similaires, et ont-ils réussi ?

- S'ils réussissaient, alors, quelle méthode de crowdfunding utilisaient-ils ?

- Quelle était la taille de leur base de supporters ?

Ainsi, les big data aident pour les actions de marketing d'une campagne avec un ciblage de la communauté la plus à même d'investir.

En résumé, l'analyse des big data permet une meilleure gestion des risques du crowdfunding, que ce soit pour les investisseurs ou pour les porteurs de projet : risque de ne pas retrouver son argent en cas de prêt ou d'achat d'action, ciblage d'une mauvaise communauté et donc d'une mauvaise communication autour du projet, risque d'échec total du projet.

Mais pour comprendre des comportements, on utilise de plus en plus l'intelligence des données. Dans ce cas, on ne parle plus vraiment de big data, mais de smart data[4]. Ce qui est important est non plus la quantité d'informations recueillies, mais bien la qualité de l'information qui peut être traitée et permet alors de positionner son projet dans le bon sens ou bien d'investir sur le bon projet.

L'analyse sémantique pour éviter les faux projets

Sur la toile, on trouve des milliers (voire plus) de nouveaux projets au quotidien.
Certaines plateformes ne sont pas modérées, et on voit apparaître des faux projets, des SPAM[5]. En France, une liste officielle des plateformes a été établie avec la création d'un label pour éviter la multiplication de faux projets. Mais on peut également compter sur les big data pour éviter d'investir dans un projet qui n'est pas réel. Pour cela, des systèmes d'analyse sémantique ont été mis en place pour le repérage de comportements frauduleux. Il s'agit de comparer, pour des projets ayant les mêmes thématiques, les descriptions, les sources, les mots-clés etc. D'où l'importance de la personnalisation et la description lors de la mise en ligne de son projet.

Le rôle des médias sociaux

Le rôle des réseaux sociaux est devenu tellement important ces dernières années tant sur le plan de la communication et de la diffusion d'informations que sur le plan de la collecte d'informations, de données. Ils jouent un rôle essentiel dans le succès des campagnes individuelles de crowdfunding. Des chercheurs de l'Université de l'Illinois ont vérifié cette hypothèse en identifiant un lien tangible entre les activités des médias sociaux et la réussite du financement. En particulier, ils ont constaté que Twitter et Facebook sont les platesformes les plus importantes. En effet, les réseaux sociaux sont comme une caisse de résonance des projets de crowdfunding et permettent

de rassembler très rapidement une communauté. Un projet très présent sur les médias sociaux partira avec des milliers de personnes qui le soutiennent.

L'utilisation de données non financières pour créer de l'argent, c'est la nouveauté dans l'utilisation des big data. Des entreprises ont bien compris le principe, et se sont spécialisées dans la vente et l'analyse des data, mais aussi d'actions d'automatisation pour le crowdfunding.

La startup française installée à Lyon depuis 2015, Crowdybox, vend des packages clé en main, d'analyse des big data aux futurs porteurs de projet. Le mot d'ordre de la startup est : « Inspirez-vous des success stories pour réussir votre propre campagne ». Crowdybox ne s'adresse qu'aux leveurs de fonds dans la réussite de leur projet (lire encadré : Rencontre avec German Bruno créateur de Crowdybox).

La société travaille avec des outils d'analyse des réseaux sociaux et des plateformes de crowdfunding afin de définir les modes de communication les plus adaptés à un projet et propose à ses clients de déployer une stratégie de campagne, quelle que soit la taille du projet. La société vend également à ses clients des listes de personnes influentes (22.000 personnes) tels blogueurs, journalistes, animateurs radio qui peuvent relayer l'information. Une fois le projet lancé, la société, propose de mesurer en temps réel la notoriété sur Facebook et Twitter.

Aux États-Unis, la société Krowdster, fondée en 2014, fonctionne sur le même principe en analysant 500.000 anciens projets. Cependant, contrairement à la startup lyonnaise, cette société californienne a comme clients des porteurs de projets mais aussi des investisseurs en quête de projet à haut potentiel.

En soit, à l'heure actuelle, le pouvoir des distributeurs de données (revendeurs de données ou analystes) est plus important que les producteurs de données. Ce sont eux qui transforment les big data en smart data et rendent la masse de données intelligible et utilisable.

---

[1] Love money : traduction argent de l'amour et non pas amour de l'argent. C'est l'argent donné généreusement, souvent à un proche (ami, famille).

[2] KPMG France. La finance alternative en France, Communiqué de presse de FPF, février 2017.

[3] *Ibidem.*

[4] Smart data : Donnée numérique pertinente

[5] SPAM : courriel indésirable

BIBLIOGRAPHIE

KPMG France. La finance alternative en France, Communiqué de presse de FPF, février 2017. En ligne hier: http://financeparticipative.org/wp-content/uploads/2017/02/CP-Barom%C3%A8tre-annuel-du-crowdfunding-KPMG-Financement-Participatif-France.pdf

BOLAND Joe, Big Data, Crowdfunding and the Fundraising Technology Craze. nonprofitpro.com, aôut 2014 (consulté le 30 mars 2017). [En ligne] http://www.nonprofitpro.com/article/big-data-crowdfunding-fundraising-technology-craze/all/

PAL Kaushik, How big data is influencing crowdfunding. Technopedia, aout 2016 (consulté le 28 mars 2017). [En ligne] https://www.techopedia.com/2/31895/trends/big-data/how-big-data-is-influencing-crowdfunding

National crowdfunding association of Canada (NCFA), How big data is transforming the world of crowdfunding. Ncfacanada.org janvier 2017 (consulté le 30 mars 2017). [En ligne] http://ncfacanada.org/how-big-data-is-transforming-the-world-of-crowdfunding/

EPITA, Big, Fast, Open data : décrire, décrypter et prédire le monde, l'avènement des données. LEJEUNE Yannick dir., FYP éditions : 2014.

GUITIERREZ Daniel, Krowdster: Big Data Analytics Meets Crowdfunding. Insidebigdata.com, février 2015 (consulté le 28 mars 2017). [En ligne] http://insidebigdata.com/2015/02/03/krowdster-big-data-analytics-meets-crowdfunding/

MINISTERE DE L'ECONOMIE ET DES FINANCES, Crowdfunding : tout savoir pour se lancer !. economie.gouv.fr, aout 2016 (consulté le 28 mars 2017). [En ligne] http://www.economie.gouv.fr/entreprises/crowdfunding-comment-se-lancer

CROWDYBOX (consulté le 15 mars 2017). [En ligne] http://www.crowdybox.com/

BERTRAND Cyril, JAKUBOWSKI, Le fric, c'est chic : panorama du crowdfunding en 2016. Annales des Mines - Réalités industrielles 2016/1 (Février 2016), p. 38-43. Disponible en ligne au format PDF : http://www.cairn.info/revue-realites-industrielles-2016-1-page-38.htm

Data marketing : notions clés, technicités et enjeux pour l'Entreprise

■ Firouze Betou

Le big data est un rouage de l'architecture globale de la nouvelle stratégie marketing des entreprises. L'innovation technique permet de croiser de grands volumes de données pour offrir une publicité taillée sur mesure et plus efficace. Le progrès apporte une capacité de ciblage et une connaissance du client jusque là inégalée avec des outils classiques de gestion de base de données. Le big data amorce la fin de la publicité de masse, l'individualisation de la stratégie marketing, soit une logique de l'offre à un client. Le domaine de la gestion des données devient un enjeu pour le développement des firmes dans un marché connecté. La personnalisation de l'offre est donc un avantage compétitif d'autant que les technologies du big data ont un coût, ce qui en limite l'accès à certains acteurs.

Le data marketing : le nouvel exercice du marketing

L'application des technologies du big data au domaine du marketing reste récente. L'usage du terme « data marketing » est né en 2000. Le data marketing permet, par exemple, de croiser des données sociales, commerciales et météorologiques pour fournir une image plus ciblée d'un consommateur. Le data marketing est une approche globale de la gestion des données clients.

L'analyse prédictive des données permet une segmentation stratégique fine du marché. Le big data permet un meilleur contrôle de l'offre en termes de message et d'image. L'objet de l'utilisation des technologies du big data est d'analyser rapidement de grands volumes de données pour la prise de décision stratégique. Les algorithmes prédictifs sont d'autant plus importants qu'ils donnent du sens à l'information collectée. L'offre est sélectionnée par un algorithme en fonction de l'analyse des données sur le comportement passé du client.

Cette offre « sur mesure » est envoyée au consommateur pour motiver l'achat. L'entreprise réduit ses frais publicitaires et la méfiance des clients vis-à-vis de sa marque.

La connaissance client : la matière première des actions marketing

Le marketing s'adresse à des marchés. Il s'agit d'influencer le comportement d'un ensemble de personnes qui sont

capables de procéder à l'échange pour satisfaire un désir. Le management de l'échange est le moteur de l'entreprise puisqu'il crée de la valeur économique.

L'enjeu de l'optique marketing est donc dans la notion de valeur. L'économie attribue à une marchandise une double nature : l'utilité sociale et l'utilité économique. Le prix qu'un consommateur est prêt à payer pour acquérir un bien dépend de son utilité sociale. La consommation est donc un acte économique influencé par des déterminants sociaux.

Un ensemble de normes sociales, démographiques et communautaires influencent les comportements de consommation et donc d'achat. Le marketing traditionnel incite la décision du consommateur pendant la phase de recherche d'information. La bonne connaissance du profil des clients et de ses réseaux d'influence sont des atouts compétitifs pour la marque.

Cette connaissance est la matière première des actions marketing et de fidélisation. La donnée client est un élément essentiel dans les pratiques marketing. Le digital permet de connaître les comportements des clients de manière plus directe pour un ciblage plus efficace. Les traces en ligne ou hors ligne laissées par le prospect se multiplient. Face à ce déluge informationnel, un changement d'échelle s'impose : celui de la donnée au big data.

Le digital : un nouveau paradigme de consommation

Les usages du Web font naître de nouveaux comportements de consommation spécifiques aux objets ou *« devices »*. Aujourd'hui, le consommateur est sursollicité. Les offres promotionnelles foisonnent dans un univers marchand vaste et volumineux. Aussi, le consommateur est-il devenu méfiant : il compare, demande des conseils. La démocratisation d'Internet facilite l'accès à l'information et le client se connecte simplement via sa tablette ou son smartphone. La marque et le consommateur sont donc dans un espace d'interactions permanentes et multicanales. Boutiques en ligne, Plateformes Web, magasins physiques : le parcours client se densifie. Désormais, le succès d'une marque dépend de la connexion qu'elle réussit à installer avec ses clients ou prospects. La marque doit créer son propre univers et sortir du carcan traditionnel de l'offre généralisée. La digitalisation de notre mode de vie produit un volume de données qui alimente en continu le big data. Ce sillage de données associé à une personne trahit son mode de consommation. Désormais, le volume d'information marketing disponible est sans limite. La publicité digitale est donc une réponse à la mobilité du client dans un espace informationnel étendu.

La publicité digitale : une conséquence de la « datification »

La convergence des médias remet en cause la distinction entre publicité et marketing direct. Selon, le dictionnaire du marketing « Le marketing direct est une approche classique de la relation client. Il s'agit d'un ensemble d'opération pour adresser directement au consommateur une offre générale ».

Les outils du marketing direct sont le téléphone, le courrier, l'e-mail, le SMS, les prospectus, etc. Le marketing digital coordonne des actions menées sur le Web. Par exemple : l'enregistrement des informations de navigation grâce aux cookies d'un site web marchand, la publication de contenu via les médias sociaux, l'échange avec des consommateurs via des blogs. Le marketing digital permet une communication publicitaire interactive plus large et à un meilleur coût, mais l'information client se disperse sur les canaux. Le big data est donc une réponse aux problématiques du marketing digital. L'entreprise a une meilleure communication avec le consommateur. Les solutions logicielles transforment en valeur l'abondance d'information produite via les médias sociaux, les achats en ligne ou hors ligne. Les systèmes de calcul big data agrègent des millions de traces numériques. Leur traitement en masse et en vitesse permet de tester la corrélation entre des suites de données multidimensionnelles. Ainsi, le big data est un levier de la stratégie marketing cross canal[1]. La marque peut pratiquer un placement publicitaire réactif et moins fragmenté sur l'ensemble des points de contact.

Les technicités du Big Data : technicités NoSQL et écosystème Hadoop

Les bases de données traditionnelles évoluent vers des solutions dédiées à l'exploitation de données massives.

En effet, le silo traditionnel de données marketing ne permet pas une gestion globale du cycle de vie du client. Ce fichier client ne contient que des données simples. Or, l'univers numérique foisonne de données non structurées et porteuses de sens pour l'Entreprise.

La technologie NoSQL[2] est favorable au stockage de données aux formats variés. Des solutions logicielles, nativement NoSQL automatisent la collecte à grande échelle de données dispersées dans les différents systèmes d'information de l'Entreprise et les canaux de connexion entre la marque et le client (Facebook, sites Web, centre d'appels, etc.). L'enthousiasme pour le « mouvement NoSQL » tient donc en partie aux limitations du modèle relationnel. Les utilisateurs et les entreprises recherchent une meilleure solution de stockage, mais également de restitution de l'information à l'échelle du Web.

Ainsi, le choix de la technique NoSQL répond aux critères du « big [3]» connus sous l'acronyme « 3V ». Le volume, la variété et la vélocité du big data imposent un traitement en

Logo Hadoop et réseaux sociaux (Source : Wikimedia Commons – CC-by – Apache Hadoop)

masse et en vitesse d'une collection de données hétérogènes. Il s'agit de collecter, stocker et d'analyser un volume exponentiel de traces numériques au fil de leur propagation sur le Web. Le succès des dispositifs NoSQL est donc lié à l'évolution du matériel de stockage d'informations, de la baisse des coûts de mémoire et de l'augmentation de la puissance de calcul des serveurs. Logiquement, les entreprises en rapport avec Internet développent leur propre solution selon le modèle NoSQL. En 2004, Google met en oeuvre un environnement de traitement distribué de données collectées : MapReduce. Ce système propriétaire prend le nom d'Hadoop dans sa version Open Source. Il s'agit d'un écosystème logiciel qui permet le traitement massivement parallèle de l'information. L'architecture sur un ensemble de serveurs distribués permet de répartir les données et les opérations de calcul en simultané sur des milliers d'ordinateurs.

Le Big Data d'un point de vue économique : enjeux pour l'Entreprise

L'enjeu du Big Data pour l'entreprise est d'optimiser la valeur des données qualifiées par les outils du big data. La collecte, le traitement et le stockage des données sur un long terme ont un coût que l'Entreprise doit maîtriser. Le big data marketing crée de la valeur économique. L'entreprise est plus compétitive. Elle augmente son chiffre d'affaires par l'exploitation de données big data[4]. Les nouvelles technologies sont sources d'innovations et de création d'emplois. Le marché du big data est estimé à 28 milliards de dollars. En 2021, cette valeur s'élèverait à 66,71 milliards de dollars.

Cette croissance montre que de nombreuses entreprises investissent dans les technologies big data pour optimiser la prise de décision. Cette nouvelle industrie se divise en deux segments : les outils d'analyse et les services. Parmi, les principaux acteurs de ce marché on note, IBM, le géant Microsoft ou encore Amazon Web Services. Néanmoins, le coût d'un projet big data reste élevé. Le ticket d'entrée varie entre 10 000 euros par an pour une solution Opensource, telle que Hadoop, à 60 000 euros et plus pour une solution propriétaire, telle qu'Oracle. Cette réalité n'est pas la même pour toutes les entreprises.

Chaque initiative organisationnelle demande des prérequis et de nombreuses firmes manquent encore de compétences internes pour envisager ce changement. L'enjeu majeur pour supporter la démarche big data reste donc le retour sur investissement (ROI). La création de richesses à partir du traitement d'un large volume de données à faible densité en information fait encore débat. D'autant que la sécurité de l'information et les aspects juridiques sous-jacents à la collecte, au stockage et au traitement de données personnelles sont des coûts non négligeables dans la masse des charges d'un projet informatique. Paradoxalement, la loi Informatique et Libertés[5] peut difficilement répondre en l'état aux critères de volume, variété et vélocité du big data. En outre, l'Open Data[6] associé au big data facilite le recoupement des données personnelles et donc l'identification potentielle d'un individu. Il reste à noter un enjeu économique et stratégique de taille pour l'Entreprise, à savoir le rôle des fournisseurs de traces numériques en masse. Le modèle de traitement du social par les traces numériques est largement diffusé les plateformes GAFA[7]. D'ores et déjà les géants du Web s'imposent comme les « gatekeepers[8]» d'un nouveau marché.

---

[1] La stratégie cross canal s'appuie sur l'usage en simultanée de plusieurs canaux de communication par un consommateur pour faire ses achats.

[2] Not only SQL, en anglais- Le terme né en 2009.

[3] Gros, en anglais

[4] La données big data se caractérise par une faible densité informationnelle.

[5] La loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés – dite loi « Informatique et libertés » – encadre la collecte et l'utilisation des données à caractère personnel. Elle confère des droits aux personnes dont les données sont collectées et traitées et impose le respect de plusieurs obligations aux responsables du traitement de ces données.

[6] Open data : données ouvertes par les administrations dans un premier temps.

[7] Cet acronyme désigne les « géants du Web » : Google, Apple Facebook, Amazon. Il s'agit des sociétés de l'univers du numérique qui se caractérisent par l'énorme volumétrie de leur base utilisateurs.

[8] Les gardes-barrières, en anglais.

BIBLIOGRAPHIE

Big Data, quelle véritable utilité pour les entreprises ?, Adresse : http://www.conseilsmarketing.com/autres-conseils-marketing/big-data-quelle-veritable-utilite-pour-les-entreprises.

Boullier, D. (2015). « Vie et mort des sciences sociales avec le big data »,. Socio. La nouvelle revue des sciences sociales, vol. , n° 4, p. 1937. Adresse : http://socio.revues.org/1259

Bruchez, R. (2015). Les bases de données NoSQL et le big data: comprendre et mettre en oeuvre

Comment exploiter les big data pour le marketing de votre e-commerce ?, 2016, ECN | E-commerce Nation. Adresse : https://www.ecommerce-nation.fr/comment-exploiter-les-big-data-pour-le-marketing-de-votre-e-commerce.

Comment le big data va transformer le monde de la publicité ? | Orange Business Services, Adresse : http://www.orange-business.com/fr/blogs/usages-dentreprise/open-data/comment-le-big-data-va-transformer-le-monde-de-la-publicite

De direct, le marketing est devenu digital pour vivre avec son temps., Adresse : https://www.dunegestion.com/p309-marketing-continue-mue-mettant-digital.html

Delort Pierre, 2015, Le big data, Paris, PUF. Adresse : http://catalogue.bnf.fr/ark:/12148/cb443062267

DOSSIER - Big Data : Les spécificités du marché français | Conseil et analyse en solutions logicielles pour l'entreprise et ses métiers, Adresse : http://www.cxp.fr/content/news/dossier-big-data-les-specificites-du-marche-francais

Enjeux économiques du développement du commerce électronique | Observatoire du numérique, Adresse : https://archives.entreprises.gouv.fr/2015/observatoire-numerique/www.observatoire-du-numerique.fr/enjeux-economiques-du-developpement-du-commerce-electronique.html

Filipone Dominique, « Big Data : comment se lancer sans se ruiner »,. Adresse : http://www.journaldunet.com/solutions/dsi/se-lancer-dans-le-big-data.shtml

Ibekwe-Sanjuan Fidelia, 2014,« Big Data, Big machines, Big Science: vers une société sans sujet et sans causalité? », XIXème Congrès de la Sfsic. Penser les techniques et les technologies: Apports des Sciences de l'Information et de la Communication et perspectives de recherches., , p. 1–10. Adresse : https://hal-univ-lyon3.archives-ouvertes.fr/docs/01/06/62/02/PDF/SFSIC-30605.pdf

Infographie - MARKETING TRADITIONNEL VS MARKETING DIGITAL !, Adresse : https://www.dunegestion.com/p426-duel-marketing-traditionnel-contre-marketing-digital.html

Jouanny Jean-Simon, 05:39:06 UTC, « Big data et marketing : Nouvelles opportunités pour les entreprises »,. Adresse : https://fr.slideshare.net/JouannyJeanSimon/big-data-et-marketing-nouvelles-opportunits-pour-les-entreprises

Le commerce électronique sur Internet : la naissance d'un nouvel univers marchand ? | La Jaune et la Rouge, Adresse : http://www.lajauneetlarouge.com/article/le-commerce-electronique-sur-internet-la-naissance-dun-nouvel-univers-marchand

Sparklane, « Exploiter les potentialités du Big Data et du marketing automation en… »,. Adresse : https://fr.slideshare.net/Zebaz-Data/exploiter-les-potentialits-du-big-data-et-du-marketing-automation-en-b2b

Thibault PAILLIER, 19:32:38 UTC, « Thèse professionnelle - COMMENT LES BIG DATA VONT AMELIORER LE MARKET »,. Adresse :https://fr.slideshare.net/ThibaultPAILLIER/thse-professionnelle-comment-les-big-data-vont-ameliorer-le-marketing-dans-les-entreprises-en-b2b-thibault-paillier

Touria Engohan, 14:21:07 UTC, « Les enjeux du Big Data pour l'Entreprise - These professionnelle »,. Adresse : https://fr.slideshare.net/touriaengohan/les-enjeux-du-big-data-pour-lentreprise

Big Data, Open Data : sources d'intelligence collective ?

■ Valérie Griman

« Dans L'Eau des collines (Jean de Florette, Manon des Sources), […] Jean Cadoret […] se tue à la tâche pour trouver de l'eau, au sens figuré, puis au sens propre. Nul ne lui aura dit que la précieuse eau coulait sous ses pieds et avait été cachée par la cupidité des hommes. C'est ainsi que je me représente la connaissance : une source de savoir, que les hommes choisissent de garder pour eux ou de partager, et dont l'usage permet tout. À l'ère numérique, nous avons sous nos pieds un réservoir immense d'informations produites par l'humanité, et demandant à devenir des savoirs communs. Selon que nous saurons ou non exploiter et partager cette source, le visage de notre siècle sera changé. »

Benoît SILLARD

Avec le Big Data et l'Open Data, nouvel or noir de l'économie, l'accès et le partage de la data représentent un changement dans les pratiques de tout un chacun et les méthodes managériales vers plus de transparence dans l'information.
Peut-on exploiter ces rivières de données pour tendre vers une intelligence collective ?

Big Data, Open Data : naissance de nouvelles pratiques numériques

« *83 % d'internautes en France sur la population totale*
4 h en moyenne par jour à naviguer sur un ordinateur ou un smartphone
107 milliards d'emails envoyés par an à travers le monde
2,46 millions de contenus partagés sur FB par minute » (Monimo)

Le Big Data englobe le traitement de ce volume de données, la collecte, le stockage, la visualisation et l'analyse.

L'Open Data, quant à lui, offre à l'internaute de nouvelles sources d'informations en accès libre.

Les données, aujourd'hui, sont devenues le « carburant de l'économie numérique ». (Monimo). Une information récoltée sur le web peut donner naissance à une connaissance nouvelle, selon comment l'internaute la traite en la croisant avec son propre capital de connaissances.

Les TIC sont en partie à l'origine de cette nouvelle ère car elles suscitent de nouvelles pratiques de travail collaboratives en réseau, de partage... Les outils de communication liés multiplient les échanges thématiques autour d'experts de tout horizon tels que les forums, les blogs, les communautés de pratiques etc...

Nous sommes dans une économie de la connaissance.

« *Ouvrir ses données sur une plateforme c'est bien. En faire des réutilisations et ainsi montrer en quoi ces données peuvent être utiles, c'est mieux. […] La première raison d'être de l'ouverture des données est sans nul doute le bouillonnement créateur qu'elle est capable d'engager.* » (site Opendatasoft)

L'ouverture des données permet un élargissement des savoirs, un accès aux richesses du patrimoine. La connaissance du passé est partagée, préservée. Connaître l'histoire peut permettre de comprendre et d'agir dans l'avenir et surtout de ne pas réinventer la roue.

Pour les organisations, il faut faire face à de nouveaux défis externes : comprendre et cerner leur

environnement et se transformer pour répondre aux exigences de leur nouveau contexte. Comment faire de ce phénomène Big Data et Open Data une source de progrès ?

Des pratiques numériques génératrices d'intelligence collective (IC)

Tout d'abord, tâchons de définir ce qu'est l'intelligence collective. Le site Big Data.fr cite une « *intelligence partagée émergeant de la collaboration, des efforts collectifs ou de la compétition entre plusieurs individus* ».

Les interactions des internautes sur les forums, les communautés de pratiques et la mise en commun par ce biais de compétences, connaissances et expériences variées génèrent des idées nouvelles, de approches novatrices et plus approfondies des sujets traités.

L'IC est donc le résultat d'une « interconnexion » entre les actions humaines et des nouvelles sources et pratiques de traitement de l'information. Les communautés de pratiques et professionnelles sont facilitatrices de la transmission de savoirs basés sur l'expérience et croisés avec les connaissances acquises en formations différentes qui multiplient les angles de vision : une vision systémique des sujets qui tend vers la création d'un savoir collectif.

Cette source d'informations est une opportunité d'amélioration continue dans les organisations voire de résolutions de problèmes. La compétitivité des entreprises repose sur leur capacité à se doter d'une organisation apprenante.

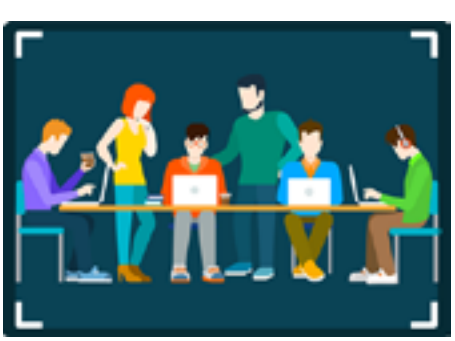
Socrate disait : « Le savoir ne vaut que s'il est partagé. » La création des richesses passe par la transmission, le partage des connaissances.

Quelques exemples

Un des exemples les plus connus est le célèbre « Wikipédia ». Ce wiki est né de l'ouverture publique des savoirs issus d'une démarche collaborative, fruit d'une intelligence collective.
Pour les scientifiques également, Internet devient un groupe de recherche commun avec des sources de données exponentielles. L'ouverture des données fédère un plus grand nombre d'experts en temps réel via les forums de discussion spécifiques.

Pour exemple : le Seoul Innovation Challenge. Ce projet d'appel à l'intelligence collective (citoyens, étrangers, entreprises et universités) est basé sur une plateforme en ligne pour proposer des idées nouvelles pour améliorer les problèmes urbains à Séoul en matière de sécurité, d'environnement et de trafic. Ce projet est enrichi par le partage de données publiques des mesures de capteurs qui permet une corrélation entre l'occupation de l'espace urbain, trafic, pollution air et bruit. Le fruit de cette intelligence collective partagée avec les entrepreneurs et les architectes permettra de mieux penser la ville.

Dans les entreprises, les communautés métiers ou autres se développent. Thales, par exemple, a mis en place des communautés de pratiques pour favoriser le transfert de connaissances métier, via un portail de gestion des connaissances multi-sites et des dispositifs de travail collaboratif comme les wikis.



(Source : http://www.lebigdata.fr)

Quelques bémols toutefois !

Attention à l'éthique !
Cette surveillance des individus via leurs traces numériques : titres de transport, carte de paiement, interactions sur les réseaux sociaux sans leurs autorisations pose un problème quant à la vie privée, l'intime de l'internaute.

Dans le domaine scientifique, le phénomène d'ouverture des données permet donc une collaboration internationale autour d'une échelle d'échantillons nettement modifiés par le Big Data. Mais peut-on résumer une personne à des datas ? Et que fait-on des personnes qui ont encore un usage modéré des TIC ? L'échantillon numérique fourni par l'or noir n'est pas exhaustif et il semble difficile d'identifier les exclusions.

Sur le modèle de « je pense donc je suis », si je ne suis pas une donnée, je n'existe pas ?

Alain Kiyindou va plus loin et évoque une fracture numérique, voire une fracture cognitive pour toutes les populations qui n'ont pas accès à ce réservoir de savoirs, voire celles qui se refusent à appartenir à un réseau social en ligne.

Pour Folon, Big Data et Open Data peuvent également être synonymes d'« océan infini de médiocrité ». Il dénonce le « remplacement des experts, des critiques par des amateurs ». En effet, toutes les publications sont-elles fiables ?

Attention également aux pratiques de désinformation, aux hoax, aux rumeurs. Par exemple, dans l'hôtellerie, comment s'assurer que les critiques sont postées par les consommateurs et non pas par le restaurateur lui-même ou son concurrent ?

« *Jusque dans quel recoin du monde ne volent-ils pas, ces essaims de nouveaux livres ? Leur multitude même est préjudiciable au savoir*

parce qu'elle crée une surabondance et que même dans les bonnes choses, la satiété est néfaste. » Erasme

Erasme, face à l'expansion de l'imprimerie, demandait une censure pour garantir qualité et pertinence des contenus. La question n'est donc pas nouvelle sur la nécessité de validation des publications. Quantité ne signifie pas qualité des datas !

Que faut-il donc retenir ?

Le Big Data et l'Open Data peuvent être des sources d'intelligence collective, à condition de rester critique sur les informations récoltées et de les croiser, de vérifier la légitimité des auteurs. La recherche d'informations doit passer par une phase de validation des sources.

Il faut donc des compétences spécifiques afin de maîtriser ces flux de données pour les rendre intelligentes, les partager collectivement, chacun avec son domaine de compétences et ainsi les faire monter en puissance.
Il faut former et sensibiliser au plus tôt que ce soit dans les entreprises mais également à l'école sur le comment faire de ce réservoir de connaissances un outil d'apprentissage, de progrès, voire d'innovation.

Les technologies évoluant très rapidement, l'interconnexion entre Big Data, Open Data et l'intelligence collective sera synonyme de « formation continue tout au long de notre vie » à l'avenir.

BIBLIOGRAPHIE / WEBOGRAPHIE

Archambault, J.-P. (2004). Les TIC, la formalisation et le partage des savoirs. Vers une économie de la connaissance ?. *Revue de l'EPI (Enseignement Public et Informatique)*. [En ligne] http://www.epi.asso.fr/revue/articles/a0403a.htm

Bastien, L. (2017). Intelligence collective et big data : deux sources de savoirs interconnectées. [En ligne] http://www.lebigdata.fr/intelligence-collective-big-data-2402

Choquet, I. (2015). Web 2.0 et Intelligence collective : émergence d'une nouvelle gouvernance pour les PME ?. Les écosystèmes numériques et la démocratisation informationnelle : Intelligence collective, Développement durable, Interculturalité, Transfert de connaissances.

Folon, J., Le printemps numérique: Comment la révolution digitale a modifié la société, l'enseignement et les entreprises., Cork, Primento Digital Publishing, 2015, [En ligne : http://public.eblib.com/choice/publicfullrecord.aspx?p=2088642].

Ibekwe-Sanjuan, F. & Paquiensegy, F. (2015). Open, Big, Collaboration: Trois utopies de l'innovation au XXIème siècle. Bredoux, E & Chartron, G. Colloque "Document numérique et Société". Open Data, Big Data: Quelles valeurs ? Quels enjeux ?, May 2015, Rabat, Maroc. De Boeck, pp.17-29, 2015, Information & Stratégie.

Kiyindou, A. (2007). De la fracture numérique à la fracture cognitive : pour une nouvelle approche de la société de l'information. In Thémat'IC 2007 « La maîtrise de l'information par les adultes : enjeux et méthodes ». [En ligne] http://www.enssib.fr/bibliotheque-numerique/documents/1948-de-la-fracture-numerique-a-la-fracture-cognitive-pour-une-nouvelle-approche-de-la-societe-de-l-information.pdf

Monino, J.-L. (2015). Big Data, Open Data et valorisation des données – Working papers - RRI Réseau de la Recherche et de l'innovation, n°48

Odumuyiwa, V. (2009). De la recherche sociale d'information à la recherche collaborative d'information. Mohamed Hassoun and Mabrouka El Hachani. 7ème colloque international du chapitre français de l'ISKO - Intelligence collective et organisation des connaissances, Jun 2009, Lyon, France. pp.387-388, 2009.

Pauline (2017). L'open data : ça sert à quoi ?. Récupéré du site Open Datasoft : https://www.opendatasoft.fr/2017/05/24/lopen-data-ca-sert-a-quoi/

Sillard, B. (2011). Maîtres ou esclaves du numérique ? 2049 : Internet, notre second cerveau – Clermont-Ferrand : Editions Eyrolles.

En matière d'élection, tout est permis... Ou presque

■ Béatrice Tofili

Éclairages. Terme bien connu aux Etats Unis, le Big Data électoral fait son apparition en France lors de la campagne aux élections présidentielles de 2017.
Une nouveauté attrayante qui n'est pas sans risque.

L'utilisation du Big Data dans les élections remonte en 2008 avec Nate Silver, statisticien américain sportif, qui avait dû compléter son analyse des sondages pour obtenir des prévisions quasi exactes annonçant Barak Obama, vainqueur de l'élection présidentielle des États-Unis. En 2016, c'est le logiciel, MogIA, créé par Sanjiv Rai en 2004 qui annonça Donald Trump président des Etats unis contre toute attente. Quelle est cette nouvelle méthodologie ? Qu'a-t-elle de particulier ? Et pourquoi prend-elle de l'importance aujourd'hui en France ?

Un contexte favorable à son essor

Trois facteurs sont responsables de son essor en France : la multiplicité des erreurs de prédictions des sondages, le remplacement des outils, méthodes et méthodologies traditionnels et surtout l'usage des réseaux sociaux.
En effet, la remise en question des méthodologies utilisées pour le sondage d'opinion et notamment lors d'élections ne date pas d'hier. Il est rappelé qu'en 2002, les instituts de sondage français avaient surestimé le vote en faveur de monsieur Le Pen laissant entendre qu'il pouvait passer au second tour. Malgré une marge d'erreur d'estimation, « *les instituts n'auraient pas modifié leur méthodologie puisque le problème d'estimation d'intention de vote d'extrême droite resta entier en 2008* » confirmait Claire Durand dans son étude[1].

Outre, les méthodes contestées et contestables des instituts de sondage ou « les dix raisons qui expliquent que les sondages peuvent se planter » rappelait Tronchet, journaliste de France Info, dans son article[2], il faut retenir que les instituts de sondage ont cumulé les erreurs de prédiction sur une courte période : le oui des britanniques lors du référendum sur le Brexit, l'élection de Donald Trump aux États-Unis et dernièrement les désignations de François Fillon et Benoît Hamon lors des primaires en France. Cette situation a permis d'instaurer auprès des citoyens un doute avéré sur les sondages.

En même temps, sont présents sur le sol français des outils et des méthodes innovantes. Pour répondre aux nouveaux enjeux de traitement de très haute volumétrie de données, les équipes de communication politique d'un candidat font de plus en plus appel à des nouvelles machines technologiques équipées de système de traitement innovant. Les outils classiques (ordinateurs, logiciels, outils et méthodes de communication, carnet d'adresse sur papier) sont remplacés par des plateformes web permettant une gestion moderne et mobile de la campagne du candidat. Les solutions spécialisées en Big Data ont « *la capacité à traiter en temps quasi réel d'immenses gisements d'informations, de grandes masses de données structurées, semi-structurées, voire non structurées et disparates pour extraire de leur recoupement de nouvelles valeurs qui demeuraient jusqu'alors inconnues* » (Zolunski, 2015).

Enfin, les données permises par les réseaux sociaux permettent ainsi aux équipes de campagne du candidat d'analyser le comportement de l'électorat permettant ainsi au candidat d'adapter sa campagne selon les besoins de l'électorat à convaincre.

Guenaëlle Gault, responsable numérique de l'entreprise Kantar Sofres, avait testé le Big Data lors des primaires de droite, en effectuant l'agrégation de données disponibles à partir de Facebook et de données publiques (commentaires, partage, like...). Elle soulignait qu'il était possible d'obtenir des informations fiables. Il est rappelé que lors des élections, il y a l'électorat militant et sympathisant, c'est-à-dire ceux qui vont en principe voter pour le candidat et l'électorat potentiel, c'est-à-dire ceux qui font que le candidat peut gagner une élection.
Mieux comprendre qui fait partie de ce dernier électorat devient un réel avantage pour le candidat en lice.

Les solutions Big Data permettent ainsi de :

- Identifier les électeurs perplexes ou susceptibles de changer d'avis

- Sonder les tendances de ces derniers

Pour ainsi adapter leur discours.

Mashable de France24 rapportait fin 2016 que l'application Knockin avait été créée spécialement pour Nicolas Sarkozy, dans le but tout d'abord de localiser d'éventuels électeurs pour ensuite les convaincre de voter en sa faveur lors des primaires de la droite.

Un contexte favorable à son dynamisme

Une année d'expérimentation

En 2017, les électeurs sont appelés aux urnes pour deux scrutins importants : celui des présidentielles et celui des législatives. Outre, l'adaptation du discours, le Big Data électoral doit permettre au candidat le ciblage des catégories d'électeurs à convaincre, le dressage des bureaux de vote prioritaires, l'organisation de l'action publique ou encore la présence à un meeting...

Une disponibilité de données permise par les réseaux sociaux et l'open data

Les électeurs produisent des données comme celles présentes à travers les réseaux sociaux tels que Twitter. Le logiciel recueille les données et affecte une analyse (commandée par le candidat et traduit par les experts en Big Data) pour en extraire un résultat qui permettra de mettre en place la stratégie du candidat.

Le logiciel de NationBuilder (entreprise californienne) permet de récolter toutes les informations utiles sur les électeurs, militants et fans à travers d'évènements ou de formulaires. Les données issues de Twitter sont également prises en compte. Toni Cowan-Brown[3], vice-présidente du développement stratégique de NationBuilder en Europe précise que sans stratégie, les données ne sont qu'une masse importante inutilisable. En exemple, il est possible d'analyser les sentiments à partir d'un tweet et en combinant à un algorithme d'analyse lexicale pour pouvoir dégager le sentiment positif ou négatif. Cette méthode est également utilisée dans l'industrie pour surveiller leur e-réputation.

Du coté de Digital Box, Vincent Monceris, co-fondateur du logiciel explique que celui-ci est alimenté par les données appartenant aux clients. Ainsi, le logiciel centralise les données issues du site web du candidat, des réseaux sociaux, de la newsletter, des militants sur le terrain.

Une mise en relation de bases de données

En exemple, Nicolas Poirier[4], consultant en droit des TIC, avait soulevé la question sur la possibilité donnée à Google de prédire les résultats des présidentielles de 2017. Sa méthodologie portait sur l'analyse par GoogleTrend du volume de données présentes sur le Web et relatives à des candidats à une élection dans un pays donné et pendant une période donnée. Il en résulte que GoogleTrend « prédisait » le candidat vainqueur. En résumé, les sociétés spécialisées en Big Data ne possèdent pas de données. Elles assistent leur client à constituer des bases avec les données déjà existantes (détenues par les équipes communicantes du candidat) ou à trouver les moyens d'en posséder (par le biais de réseaux sociaux comme Twitter, sondages classiques, les bases de données en open data (données mises à disposition du public), les formulaires, les newsletters et le site Internet du candidat, les études des « requêtes » sur Web...).

Vers un ciblage individuel

Outre la capacité de traiter rapidement une variété de données en y appliquant de puissants algorithmes, le second avantage du Big Data est de recueillir des données sans dispositif voyant et sans influence de l'observateur. Ainsi, au contraire des sondages qui recueillent des données « provoquées » et notamment issues d'un échantillon de personnes, le Big Data permet l'exploration de données collectées sans but défini et donne ainsi une vue à 360° de l'électorat au candidat en lice. Il est ainsi mis en exergue que le Big Data électoral ne se limite pas à l'analyse collective de l'évolution du corps électoral, il permet le ciblage individuel.

Un contexte particulier en France

Les séries policières américaines et françaises illustrent bien la différence de réglementation en matière de protection des données individuelles. En effet, lorsque outre Atlantique, les policiers américains utilisent une vidéo équipée de reconnaissance faciale, ici, seule la vidéo surveillance est permise. Ceci est fait aux lois[5] françaises qui encadrent rigoureusement le recueil et l'utilisation des données personnelles et l'accès aux documents administratifs[6].

Un régime dérogatoire existe en matière d'élections

Photo by Arnaud Jaegers on Unsplash

Toutefois, un régime dérogatoire existe et notamment en matière de communication des listes électorales, documents contenant les données personnelles (nom, prénom, date et lieu de naissance, adresse postale et physique) des électeurs sur multi support (pdf, .doc, CD ou USB).

En effet, conformément à l'article L. 28 du code électoral, les listes électorales sont communicables à tout individu, c'est-à-dire qu'un électeur, un candidat ou encore un parti ou groupe politique peuvent être en possession de ce document.
Outre la faible restriction dans l'utilisation définie par la loi qui est l'usage non commercial, engagement express donné par le demandeur qui peut être sanctionné d'un an d'emprisonnement surélevé de 15 000 euros d'amendes pour non-respect, les listes peuvent donc faire l'objet de traitement informatique.

Un profilage au dépit de l'électeur

Avez-vous déjà été confronté à l'appel téléphonique ou l'invitation personnalisée d'un parti politique à un meeting sans que vous ayez fait une demande express auprès d'eux ? Avez-vous aimé une page Facebook, liker un tweet de leur compte ou encore laisser un commentaire sur leur site Internet ?

C'est à partir de ce clic que le « profilage » commence.
Le Big Data a besoin de données et ces données sont récoltées à partir des bases constituées par la plateforme mise en place pour la campagne comprenant les comptes des réseaux sociaux, les formulaires de contact, les mails, ... les sondages et les listes électorales.

Profils, messages, commentaires, émotions, photos, coups de gueule sont mis à la disposition du Big Data. La variété et la multiplicité des données n'est pas un souci pour le Big Data qui analyse en temps réel dans l'objectif d'apporter une information faible. Ainsi, lorsqu'un électeur aime une page d'un parti politique qui souvent lui a été partagée, il est fiché dans les listes des électeurs dits potentiels. L'analyse permise par le Big Data croisant ainsi des bases de données sensées ne jamais se rencontrer permet « même sous pseudonyme » de retrouver l'identité d'un électeur. Goûts, coup de gueule, photos, adresse, profession sont dévoilés au dépit de ce dernier.

Une frontière fragile à surveiller

Information personnalisée ou intrusion dans la vie privée ? Les nouveaux usages des réseaux sociaux et notamment l'arrivée du Big Data déstabilisent le principe du respect à la vie privée. La Commission Nationale de l'Informatique et des Libertés (CNIL) a précisé le 8 novembre 2016[7] les conditions dans lesquelles les données issues des réseaux sociaux peuvent être utilisées et notamment en matière de stratégie électorale. Elle met l'accent sur le caractère illégal de collecte massive de données en l'absence d'information des personnes concernées. Elle rappelle également que la collecte doit être loyale et licite en respectant le droit à l'information et également le droit d'opposition des internautes électeurs.

Ainsi, deux profils sont dressés : le contact régulier et le contact occasionnel. Ce dernier peut devenir un contact régulier si et seulement si, il accepte de manière positive de l'être. La CNIL précise enfin que seules les opinions politiques des contacts réguliers peuvent faire l'objet de traitement. C'est pourquoi plusieurs sociétés proposant des logiciels de Big Data ont été contraintes de s'adapter[8]. La société NationBuilder par exemple, a ainsi retiré de leur plateforme le module fonctionnel permettant d'enrichir les bases de données avec les informations issues des réseaux sociaux grâce au simple renseignement de l'email de l'internaute.

Le croisement de données n'est donc plus permis pour le contact occasionnel et par conséquent, l'électorat potentiel. Outre cet aspect du droit à l'information, la CNIL veille également à la sécurité des données sensibles. En effet, il est rappelé la sanction administrative érigée à l'encontre du parti socialiste suite à la faille de sécurité constatée[9].

Une effervescence mitigée

Porte à porte, appel téléphonique, sondage, open data restent les moyens

complémentaires à l'enrichissement des bases de données.

Outre la réglementation appliquée en France portant sur la protection des données, l'analyste et l'expert en Big Data restent les gardes fous à l'utilisation effrénée des données en matière électorale. Ce premier scrutin a permis de mettre en exergue qu'il est bien sûr absurde de croire que le Big Data après recueil des données, offre une analyse pertinente. Rappelons, la création de l'application Knockin qui avait été créée spécialement pour N. Sarkozy, l'application participative Gov prédisant Fillon au second tour ou encore l'expérimentation de cinq étudiants de l'école Télécom Paris Tech qui ont utilisé le big data[10], les réseaux sociaux et sondages pour prédire un duel Le Pen-Fillon au second tour… ce qui s'est soldé par une erreur de calcul. Même si le Big Data électoral n'a pas trouvé écho en France comme il opère outre Atlantique, il n'en reste pas moins qu'il a permis d'optimiser l'action sur le terrain des équipes de campagne des partis politiques.

---

[1] Durand Claire, « La méthodologie des sondages électoraux de l'élection présidentielle française de 2007 », Bulletin de méthodologie sociologique, pp. 5-17, n°97, 2008

[2] http://www.francetvinfo.fr/elections/sondages/election-presidentielle-dix-raisons-qui-expliquent-pourquoi-les-sondages-peuvent-seplanter_2102031.html

[3] https://cercledescommunicants.com/2017/02/14/dossier-le-recours-au-big-data-en-communication-politique/

[4] https://www.contrepoints.org/2016/08/17/252984-google-peut-il-predire-le-resultat-des-presidentielles-de-2017

[5] Loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés

[6] Loi n° 78-753 du 7 juillet 1978 portant diverses mesures d'amélioration des relations entre l'administration et le public

[7] https://www.cnil.fr/fr/communication-politique-quelles-sont-les-regles-pour-lutilisation-des-donnees-issues-des-reseaux

[8] https://usbeketrica.com/article/presidentielle-comment-la-france-est-entree-dans-l-ere-du-big-data-electoral

[9] https://www.cnil.fr/fr/faille-de-securite-de-donnees-sensibles-en-ligne-avertissement-public-pour-le-parti-socialiste

[10] http://www.lepoint.fr/presidentielle/presidentielle-les-deux-qualifies-pour-le-second-tour-sont-18-04-2017-2120431_3121.php

BIBLIOGRAPHIE / WEBOGRAPHIE

Sermindadaz Sarah, « Le data mining peut-il infléchir l'issue de l'élection présidentielle américaine ? », 7 novembre 2016, Sciences et Avenir, [En ligne] : https://www.sciencesetavenir.fr/politique/comment-big-data-et-data-mining-chamboulent-l-election-americaine_107974

« L'intelligence artificielle proclame Trump prochain président des USA », 28 octobre 2016, Suptnik France, [En ligne] : https://fr.sputniknews.com/international/201610281028435864-usa-trump-ia-president/

Gavard et Wybo, avec Soulas-Gesson, Emmanuel, Gilles, Delphine, « Peut-on encore mesurer l'opinion ? », L'évènement – Enquête, Stratégies, pp. 6-8, n°1880, 17 novembre 2016

Durand Claire, « La méthodologie des sondages électoraux de l'élection présidentielle française de 2007 », Bulletin de méthodologie sociologique, pp. 5-17, n°97, 2008

Tronchet Sylvain, « Election présidentielle : dix raisons qui expliquent pourquoi les sondages peuvent se planter », France INFO, 19 mars 2017, [En ligne] : http://www.francetvinfo.fr/elections/sondages/election-presidentielle-dix-raisons-qui-expliquent-pourquoi-les-sondages-peuvent-seplanter_2102031.html

Zolynski Célia, « Big Data : pour une éthique des données », La revue I2D – Information, données & documents, pp. 24-26, n°2, 2015

Arnaud, Damien, « Dossier : le recours au Big Data en communication politique », Le cercle des communicants francophones, 14 février 2017, [En ligne] : https://cercledescommunicants.com/2017/02/14/dossier-le-recours-au-big-data-en-communication-politique/

Bonnaffé, Hugo, « [Idées] Le Big Data remplacera-t-il un jour les sondages ?, OVH News, 21 décembre 2015, [En ligne] : https://www.ovh.com/fr/a2055.idees-big-data-remplacer-sondages

Mahammedi-Bouzina, Mehdi, « Big data : démocratie augmentée, caricaturée ou en danger ? », Point d'ancrage, [En ligne] : http://pointdaencrage.org/big-data-democratie-augmentee-caricaturee-ou-en-danger/

Tronchet, Sylvain, « Sondages : faut-il changer de méthode ? », France INFO, 17 mars 2017, [En ligne] : https://www.franceculture.fr/politique/sondages-faut-il-changer-de-methode

Lumenogic Gazette, « Prévisions : le marché prédictif Hypermind remporte la médaille d'or », janvier 2017, [En ligne] : https://blog.lumenogic.com/2017/01/

Poirier, Nicolas, « Google peut-il prédire le résultat des présidentielles de 2017 ? » [Replay], 17 août 2016, Contrepoints, [En ligne] : https://www.contrepoints.org/2016/08/17/252984-google-peut-il-predire-le-resultat-des-presidentielles-de-2017

La CNIL : https://www.cnil.fr/fr/faille-de-securite-de-donnees-sensibles-en-ligne-avertissement-public-pour-le-parti-socialiste et https://www.cnil.fr/fr/communication-politique-quelles-sont-les-regles-pour-lutilisation-des-donnees-issues-des-reseaux

Ledit, Guillaume, « Présidentielle : comment la France est entrée dans l'ère du big data électoral », Usbef & Rica, 2017, [En ligne] : https://usbeketrica.com/article/presidentielle-comment-la-france-est-entree-dans-l-ere-du-big-data-electoral

BIG DATA ET ENVIRONNEMENT :

Impact et coût environnemental du Big Data versus apports du Big Data au sein des politiques environnementales

■ Céline Hoarau-Legoux

La gestion de la production, de la collecte et de l'analyse de mégadonnées, ce que nous appelons communément le Big Data, semble à première vue un concept impalpable, l'affaire de spécialistes au sein d'espaces virtuels inaccessibles. Il n'en est rien, tout du moins pour ce qui est de la production de cette masse de données, devenue si précieuse. Nous contribuons tous à construire le Big Data à partir de nos terminaux informatiques ou de nos objets connectés. Mais avons-nous conscience du coût écologique que celui-ci engendre ou de l'incroyable richesse de ressources que cette avalanche de données offre aux analystes, spécialistes de la donnée environnementale ?

La face cachée du Big data

Serions-nous capables de nous passer de nos mails, de nos réseaux sociaux, des consultations de nos données (hébergées sur un cloud ou consultables via une base de données en ligne), depuis notre ordinateur portable, mais plus probablement davantage depuis notre tablette ou smartphone ? Trouvons nous acceptable aujourd'hui que l'accès à ces données soit suspendu, ne serait-ce que quelques heures ? Est-il concevable que l'ensemble de ces données ne soient pas disponibles 24h sur 24, 7 jours sur 7?

La réponse est non, nous n'avons pas l'intention de nous passer de nos data, les géants du web (Google, Amazon, Facebook, Apple,...) l'ont bien compris et mettent d'ailleurs tout en oeuvre pour répondre à cette demande croissante et d'une exigence grandissante.

Les entreprises, les administrations, les établissements publics sont également d'importants producteurs de données et ont subséquemment un besoin élevé de stockage des données, avec une exigence accrue en termes de sécurisation et d'accès à ces données en continu.

La diffusion et le stockage des données : espaces virtuels ou bien réels ?

La plupart d'entre nous avons conscience que la circulation des données ne s'opère pas dans les méandres plus ou moins virtuels de la sphère internet. Pour atteindre leur destinataire ou leur espace de stockage, ces données transitent via des infrastructures physiques (câblages, gaines,..) et sont entreposées sur des serveurs de manière temporaire ou sur du plus long terme.

Mais mesure-t-on réellement le véritable impact environnemental, bien réel, de cette explosion et cette circulation des données numériques, dont le volume double désormais tous les 2 ans. Le volume colossal de données que nous produisons quotidiennement (environ 2.5 trillions de données) nécessite en effet la construction de véritables fabriques du numérique, en d'autres termes de vastes entrepôts hébergeant des centaines, voire des milliers de serveurs par lesquels transitent ou sont stockées nos données numériques.

Data center de Google

Google datacenter
(Source : https://www.google.com)

Ces centres de données, conçus pour être opérationnels en continu, sont extrêmement énergivores, tant pour ce qui concerne leur fonctionnement que pour la régulation de la température générée par toutes ces machines. Les mails envoyés chaque jour sur la planète se comptent en centaine de milliards, ce sont en effet 204 milliards de mails qui ont été envoyés dans le monde chaque jour en 2015[1]. Rapporté à une entreprise de 100 personnes, cela génère une émission de 13,6 tonnes équivalent CO2 par an, soit l'équivalent, en termes de pollution, de près de 14 allers retours en avion entre Paris et New York, selon l'Ademe.

Ces chiffres parlent d'eux-mêmes et permettent de comprendre les enjeux qui pèsent sur la circulation des données numériques. « *Ce que redoutent par dessus tout les fournisseurs d'accès, c'est le crash numérique* »

Pour éviter cela, tous les moyens sont mis en oeuvre pour ne laisser aucune place à quelque dysfonctionnement que ce soit, comprenant par exemple la redondance des équipements, l'installation des centres de traitement des données proches des sources de production électrique.

« *A l'échelle européenne, la Commission, qui estimait cette consommation à 56 milliards de kilowatts en 2008, table sur 104 milliards en 2020.* »[2]

Un rapport de Greenpeace, « Votre cloud est-il net ? »[3] insiste sur le fait que les entreprises américaines, proposant des services de Cloud computing, doivent veiller non seulement à la réduction de leur consommation énergétique, mais aussi au mode de production de cette énergie. En effet, encore trop de géants américains du web sont installés en majorité en Caroline du Nord, proche des centrales électriques au charbon, lesquelles puisent ce minerai en défigurant les paysages environnants. La production d'énergie fossile participe au réchauffement climatique, à la production de gaz à effet de serre : notre utilisation du web, nos actions quotidiennes liées à la production et la diffusion de données ont un impact direct sur la planète.

« *Entrepôt, fabrique : le vocabulaire du champ industriel croise celui du numérique…* »

« *Le réseau de production énergétique devient intelligent pour une approche collaborative et solidaire* »

Comment appliquer l'efficacité énergétique à la révolution numérique ?

Pour des raisons économiques, mais aussi d'impact environnemental, les entreprises du web ont pris conscience cependant qu'il était urgent de se tourner vers les énergies renouvelables pour faire face à ces besoins énergétiques exponentiels.

La course au renouvelable parmi les entreprises du web est lancée. Google et Yahoo développent progressivement les énergies propres pour l'alimentation de leurs data centers. Facebook montre également l'exemple. Le leader des réseaux sociaux a en effet installé l'un de ses data center en Suède afin de bénéficier du climat favorable au refroidissement naturel de ses équipements. Apple a mis en place des fermes solaires destinées à alimenter ses centres de traitement des données. Certaines entreprises, en Europe également, recyclent la chaleur produite par leurs serveurs et la proposent pour le chauffage de bureaux. Ces quelques exemples, au-delà de l'effet Greenwashing recherché par les entreprises, contribuent à la prise en considération du coût environnemental des pratiques liées au net.

Ces entreprises informatiques portent une responsabilité environnementale évidente, mais les utilisateurs de ces technologies n'ont-ils pas aussi un rôle à jouer pour ralentir cette course effrénée vers toujours plus d'accès, toujours plus rapide et continuellement disponible ?

L'Ademe propose des guides à destination des usagers du web pour sensibiliser les professionnels et les particuliers aux coûts

environnementaux des actions de chacun sur internet. Il est sans doute temps de se poser la question de l'intérêt de conserver certaines données ad vitam aeternam sur des serveurs énergivores alors que ces données ne seront plus jamais consultées. La question du tri et de la conservation des données est cruciale dans ce contexte également. Un accès sans interruption à toutes les données sans exception est-il absolument indispensable ?
Toutes ces questions émergent progressivement. Nous vivons désormais à l'ère du numérique, pour le pire mais sans doute aussi pour le meilleur.

Le Big data au service de l'environnement

Production de données sur tous sujets, volume colossal d'informations produites et diffusées, ouverture de ces données, intelligences artificielles de plus en plus performantes capables d'analyser en temps réel les données issues de capteurs, divers fixes ou mobiles. Comment le Big data peut-il contribuer à mieux comprendre notre environnement, à mieux s'y adapter, à mieux le préserver ?

La technologie au service de la production énergétique

Les Smart grids ou réseaux électriques intelligents se développent partout dans le monde. La France est le premier pays européen engagé dans ce type de projets, 100 projets étaient en cours en 2015. Ces Smart grids permettent une meilleure maîtrise de la production et de la consommation énergétique, comprenant des compteurs communicants pour un suivi en temps réel des consommations au plus près des besoins.

Dans le domaine des énergies renouvelables, nous savons produire ce type d'énergie mais ne disposons pas des technologies permettant de le stocker.
Des tests sont actuellement effectués afin de permettre une consommation de l'énergie au moment où cette dernière est produite, l'objectif, via une connexion internet, est de relier en direct les moyens de production énergétique aux consommateurs. Le réseau de production énergétique devient ainsi intelligent pour une approche collaborative et solidaire.

Le « Big data climate challenge » organisé par le programme Global pulse de l'ONU est un autre exemple d'initiative qui parie sur le potentiel d'exploitation des données massives pour développer des stratégies de lutte contre le changement climatique. Les entreprises partenaires de ce challenge mettent à disposition des participants des sets de données anonymes, la société Orange y contribue.

Grâce au Big Data, les entreprises sont désormais en capacité de mesurer précisément leur consommation énergétique et identifier les économies à réaliser en terme d'éclairage des bâtiments par exemple.

Pour en savoir plus sur ce challenge :

« Big data climate challenge » organisé par le programme Global pulse de l'ONU est un autre exemple d'initiative qui parie sur le potentiel d'exploitation des données massives pour développer des stratégies de lutte contre le changement climatique. Les entreprises partenaires de ce challenge mettent à disposition des participants des sets de données anonymes, la société Orange y contribue.

http://www.dataforclimateaction.org/

https://www.orange.com/fr/Engagements/Responsabilite/Environnement/Folder/Changement climatique/Folder/

Rechauffement climatique Big Data pour un immense defi

La collecte des données environnementales au service des citoyens

Les villes intelligentes fleurissent un peu partout en France et dans le monde. La collecte et l'analyse des données produites par des capteurs disposés dans la ville, ou par les smartphones des citoyens, sont autant de sources d'informations permettant d'offrir aux usagers des données utiles pour connaître les niveaux de pollution, de bruit, de qualité de l'air ou de fréquentation des espaces publics, en temps réel dans l'agglomération où ils se trouvent.

Ces données sont utiles également aux pouvoirs publics pour adapter l'action publique afin de réduire ces pollutions ou orienter, désengorger les flux. Les villes connectées sont en mesure de fournir également des informations sur le trafic en continu, les usagers des transports deviennent ainsi capable d'optimiser leurs déplacements. L'Optimod'Lyon est un exemple d'optimisation de la mobilité durable en ville.

L'Open data pour lutter contre la pollution et le changement climatique

Les pouvoirs publics misent de plus en plus sur l'ouverture de leurs données qu'ils mettent à disposition du public et des entreprises pour une réutilisation de ces dernières dans le cadre de projets innovants.

En décembre 2016, la France accueillait le 4e sommet mondial du Partenariat pour un gouvernement ouvert[4]. Notre pays est en effet très impliqué sur les questions d'ouverture des données et sa plate-forme Data.gouv.fr ne cesse de s'enrichir d'informations liées aux différentes thématiques des politiques publiques menées en France.

En février 2016, toujours en France, la ministre de l'écologie, Ségolène

Royale, a lancé GreenTech. Ce programme a pour but de mettre en place un réseau d'incubateurs de startups de la transition énergétique pour une croissance verte[5]. Le numérique est au centre du dispositif puisqu'il doit favoriser et accompagner cette transition énergétique.

L'ouverture des données publiques à ces startups est essentielle dans le but de « créer de la richesse par la construction de nouveaux services au public ». L'incubateur d'Orléans, baptisé Data Centre d'Orléans, a vocation à accompagner tout particulièrement les acteurs du numérique et est axé sur la « science de la donnée environnementale ».

Objets connectés et applications au service du suivi de la consommation énergétique, voire de l'autoconsommation énergétique, suivi numérique de l'entretien des bâtiments, outils de partage de moyens de transports électriques, atlas de la biodiversité alimenté par la science participative, voici quelques exemples de projets amenés à se développer dans le cadre de ces incubateurs.

L'innovation technologique dans le domaine de l'énergie et du transport, très étroitement liée au Big data et

au développement d'algorithmes de plus en plus puissants, est un enjeu majeur de la lutte contre le réchauffement climatique et la réduction de notre empreinte carbone.

Le Big data peut rendre de véritables services aux populations en terme de surveillance de l'environnement, qu'il soit immédiat, local, mais aussi plus global, à l'échelle de notre planète corrélé à la lutte contre le changement climatique. Tout est question d'équilibre entre le coût environnemental et énergétique des infrastructures directement associées au fonctionnement du Big Data, et les atouts, que nous commençons tout juste à explorer, des applications en faveur de l'environnement et de la sauvegarde de la planète que permettent le Big Data.

Des serveurs virtuels, consommant 3 fois moins d'énergie que les serveurs physiques, sont développés aujourd'hui. Ils ont vocation à optimiser les ressources des serveurs physiques pour une consommation énergétique réduite. Une prise de conscience est née relative à l'utilisation des énergies renouvelables dans ce secteur, mais aussi à l'importance de sensibiliser les utilisateurs à réfléchir et modifier leurs pratiques au quotidien.

Ces pratiques personnelles et professionnelles ont un impact qui va bien au-delà de ce que l'on imagine à première vue. Il en va probablement de la responsabilité de chacun, qui plus est dans le domaine de l'information et de la communication. Une prise de conscience est née relative à l'utilisation des énergies renouvelables dans ce secteur, mais aussi à l'importance de sensibiliser les utilisateurs à réfléchir et modifier leurs pratiques au quotidien.

La 3e révolution industrielle est en marche, nous dit Jérémy Rifkin[6] (économiste et consultant auprès de l'Union européenne), gageons que nous saurons en tirer parti.

---

[1] http://www.planetoscope.com/Inte rnet-/1024-emails-envoyes-dans-lemonde.html

[2] http://www.lemonde.fr/planete/article/2013/07/01/les-centres-dedonnees-informatiques-grosconsommateurs-denergie_3439768_3244.html#C7I6sJ EMSywWeUVX.99

[3] https://cdn.greenpeace.fr/site/uploads/2017/02/votre-cloud-est-ilnet.pdf?_ga=1.191404 08.1021068756.1491025039

[4] http://www.archimag.com/vienumerique/2017/02/03/opendata-paris-capitale-mondialegouvernement-ouvert

[5] http://www.developpementdurable.gouv.fr/reseaudincubateurs-greentech-verte

[6] http://www.huffingtonpost.fr/jeremyrifkin/troisieme-revolutionindustrielleenvironnement _b_8396394.html

POUR ALLER PLUS LOIN

Guide pratique de l'Ademe, Internet, courriels : réduire les impacts. Ademe, février 2014

https://www.ademe.fr/sites/default/files/assets/documents/guide-pratiqueinternet-courriels-reduire-impacts.pdf

Le Groupement de services (GDS) EcoInfo du CNRS

Pour une informatique éco-responsable : http://ecoinfo.cnrs.fr

Les ambiguïtés du Big Data

■ Alexandre HUBERT

Le Big Data : un concept assez méconnu très associé à l'informatique et au monde des affaires.
Et si les organisations prenaient conscience de l'importance des « travailleurs du savoir ? »

Big Data : de quoi parle-t-on ?

Le Big Data est un buzzword[1] qui continue de faire couler beaucoup d'encre. Il n'a pas la même signification selon son domaine d'appartenance ou selon son angle d'approche. On finit généralement par prendre conscience qu'il est polymorphe et qu'une grande confusion règne parfois, confusion générée volontairement ou non, par les acteurs et observateurs.

Ici, nous nous appuierons sur les travaux d'E. Schmitt[2] pour entendre « *les Big Data comme l'ensemble des conséquences de la massification de l'informatisation du fait humain, c'est-à-dire de ses activités individuelles, sociales et collectives, parmi lesquelles se trouvent notamment la multiplication des traces laissées par les individus par leur usages de dispositifs tels que le web et l'appropriation de ces traces pour en faire des données exploitables, l'informatisation d'activités productrices de données […] ainsi que l'ensemble des outils, techniques et méthodes développés pour stocker, manipuler, analyser, visualiser ces données, à quoi il faut ajouter la compréhension politique et éthique de ces phénomènes lorsqu'ils sont envisagés dans leur dimension sociotechnique d'outils de surveillance et de pouvoir.* »

Dans cette acception, nous sommes donc dans le phénomène et ses conséquences, ces dernières intégrant les orientations technologiques et leurs nouvelles architectures, les changements de paradigmes économiques, sociaux et sociétaux inhérents. D'un point de vue scientifique et en général, un monde semble s'être ouvert aujourd'hui et semble promettre beaucoup pour demain.

Changements principaux et champs d'application

Changements

Le phénomène Big Data offre en réalité trois changements interdépendants selon V. Mayer-Schönberger et K. Cukier[3] :

- La capacité d'analyser des quantités bien plus grandes de données. Jusque-là, et depuis le XIXe siècle, il était question, avec de gros nombres, de recourir à l'échantillonnage. En comparant les périodes, il y a d'un côté la conjonction de facteurs tels que les méthodes statistiques
avec des règles sur des « petites données » et un univers analogique ; de l'autre, le numérique et les progrès technologiques qui l'encadrent à savoir l'architecture technique, le stockage, la capture des informations et la puissance de calcul (pour faire simple).

- Vis-à-vis de l'époque précédente, une sorte de détachement par rapport à la recherche systématique d'exactitude. La « massification » peut fait ressortir des grandes tendances, des phénomènes. L'exactitude n'est plus forcément le but systématique et ultime recherché par les organisations dans ce nouveau paradigme. Elles se satisfont d'une « notion générale plutôt que de connaître un phénomène jusqu'au plus petit détail ». La compréhension globale au détriment de la précision en quelque sorte.

- « *Une prise de recul vis-à-vis de l'éternelle quête de causalité [de l'être humain]*» qui est conditionné pour raisonner de la sorte avec les biais

qu'elle peut parfois induire; « *dans le monde de big data, en revanche, fini l'obsession de la causalité ; dans les données, nous pouvons découvrir des corrélations qui révèlent des tendances et nous apportent des informations vraiment précieuses, voire innovantes* ». En somme, il s'agit davantage de quoi que de pourquoi : alerter qu'un fait se produit. Faire émerger des informations non visibles jusqu'alors.

Le Big Data marque donc une rupture importante avec la période qui l'a précédée. Quantitativement, l'échelle des volumes à traiter explose et toute la chaîne de création de valeur est bouleversée[4]. Sur le plan qualitatif, on ne traite plus des données préalablement échantillonnées et structurées, mais hétérogènes et éparses, structurées et non-structurées (texte, image, multimédia, traces numériques…). Les données sont traitées non plus en différé mais en temps réel : on passe d'une logique de silos (batch, tables, ..) à une logique de flux[5]. De plus, il consiste[6] à « *créer en exploratoire et par induction sur des masses de données à faible densité en information des modèles à capacité prédictive. Les résultats de ces modèles à capacité prédictive peuvent être exploités soit de manière automatique, soit pour décision* ».

Un écosystème économique

Le Big Data est devenu un écosystème économique[7] . Il touche tous les secteurs d'activité et a de nombreux champs d'application. On peut observer qu'il s'est bien imposé dans le milieu[8] du business. Mieux comprendre le présent ou envisager la prédiction des faits rend les décideurs très attentifs. Parmi ces domaines, on entend beaucoup parler d'informatique, de web, de marketing, de sciences et de surveillance mais la liste est bien entendu très loin d'être exhaustive car l'objectif n'est pas là.

Photo by Mervyn Chan on Unsplash

L'informatique est au premier plan car le phénomène provient clairement de prime abord d'innovation à plusieurs niveaux sur ce périmètre ; on peut voir cela comme le « moyen », c'est-à-dire que les données sont capturées, stockées sur des matériels informatiques et exploitées au sens technique du terme par les informaticiens.

La valorisation des données, pour en extraire de l'information est faite par des profils émanant des sciences de données : les data scientists ont généralement une triple compétence informatique, mathématique et statistique. Ces sciences de données utilisent des méthodes qui s'appuient mais succèdent au data mining (en simplifiant, fouille de données : analyse de données et statistiques exploratoires) qui visent avant tout à exploiter les données collectées dans les entrepôts des organisations[9]. Les data scientists recherchent des motifs qui pourraient se répéter dans les masses de données en vue d'en extraire de l'information ou des éléments pouvant inviter les parties prenantes à une analyse plus fine.

D'autre part, le marketing et la gestion de la relation clients font beaucoup parler car le Big Data est très prometteur dans la connaissance des clients ou des prospects ; ce que nous évoquions précédemment concernant les limites des échantillons est particulièrement applicable dans ces activités ; de plus, sur les portails d'e-commerce, il est très intéressant d'étudier les parcours effectués et les taux de conversion qui s'ensuivent. Un intérêt croissant également pour ce qu'englobent les « longues traînes[11] ». Les marketeurs[12] ont longtemps travaillé avec des moyennes et sur des segments (panier moyen etc.), mais les moyennes cachent des réalités bien plus riches et les situations réelles plus complexes (un sujet appartient à plusieurs segments notamment) : les Big Data peuvent par exemple mettre en exergue que l'augmentation d'une consommation de shampooing en pharmacie correspond à l'arrivée des pollens[13].

Dans des domaines comme la santé pour ne citer que celui-ci, le Big Data ouvre des perspectives très prometteuses pour la recherche mais inquiète également pour des aspects déontologiques (données de santé). Nous ne développerons pas davantage.
Bien sûr que les statistiques et la volonté d'analyse ne sont pas nées avec le déluge Big Data. D'autre part, les promesses et la concrétisation de ces promesses sont deux choses différentes. Toutefois, nous sommes entrés dans une nouvelle ère avec l'avènement de ces Big data.

Management des connaissances

Comment est classé Big data dans les bibliothèques ?

Le langage RAMEAU[14] est un langage d'indexation utilisé par la Bibliothèque Nationale de France (BNF). L'expression « Big data » y est remplacée par « Données massives »

et cette entrée est en outre catégorisée dans les domaines 020 et 621 correspondant respectivement à *Bibliothéconomie* et *sciences de l'information (archives, archivistique, centres de documentation)* d'une part, et *Physique appliquée (électrotechnique, génie civil, génie mécanique, ingénierie appliquée, principes physiques en ingénierie)* d'autre part. Cela montre que ce phénomène est à la fois ancré dans les sciences dites dures et les sciences sociales.

Données, informations, connaissances, savoir ? Help !

C. Deschamp constate[15] que ces termes ne font pas consensus et rappelle au lecteur que connaissance et savoir ont pour seul équivalent knowledge en anglais. Il s'appuie sur les travaux de T. D. Wilson pour proposer des définitions :

- « *Données (data) : faits simples, extérieurs à l'esprit (mind) qui peuvent être manipulés (utilisés, «mixés», groupés, etc.) ;*

-Informations : données devenues signifiantes par leur intégration dans un contexte pertinent pour l'utilisateur/ destinataire. Elles peuvent aussi être manipulées à un niveau de granularité moindre ;

- Connaissances : tout simplement « ce que nous connaissons », car pour chacun de nous cette définition se suffit à elle-même ;

- Savoir : ce terme n'est pas défini par Wilson puisqu'en anglais il est synonyme de connaissance. »

En complément, nous proposons cette définition[16] :

« *Le savoir est constitué d'informations mises en relation, organisées par l'activité intellectuelle du sujet. Contrairement à l'information qui n'est emmagasinable que spatialement, le savoir s'organise dans le temps en s'intégrant au savoir antérieur du sujet. Ainsi le savoir est personnel ou n'est pas, il est avant tout le savoir que le sujet a lui-même construit* ».

Traitement versus médiation

Traitement

En résumant et en simplifiant, les data scientists vont notamment chercher à mettre en œuvre l'automatisation de l'apprentissage des machines (machine learning) dans leur quête de reconnaissance de motifs en fonction d'objectifs qui ont été fixés. L'informatique regroupe de nombreux métiers et nombreuses compétences pointues. Toutefois, dans le périmètre qui nous intéresse ici, du point de vue des données, l'informatique raisonne et se fixe comme finalité le traitement automatisé des données.

Médiation

Pour que les données puissent être transformées en informations et en connaissance et faire sens auprès des usagers ou commanditaires, une action de médiation est requise. Le dictionnaire de l'information[17] la définit comme l' « *action de s'entremettre pour produire un accord ou réguler une situation* », et explique que cette question est au cœur de la société de l'information et de la communication. Les professionnels de l'information ont un rôle à jouer et doivent proposer leur expertise, en amont, dans l'élaboration des systèmes d'information et / ou être en appui des usagers en aval.

Par ailleurs, dans le Big Data et dans la société de l'information d'une façon générale, tout le monde produit de l'information. Dans les organisations, les professionnels précités peuvent créer les conditions pour que les données soient structurées dès le départ en auditant les activités et en étudiant les flux métiers. Manager l'information concerne, par exemple pour le chercheur, la veille, la recherche, le traitement et la diffusion de l'information[18]. Ils peuvent aussi associer les collaborateurs à une démarche de responsabilisation vis-à-vis des données qu'ils produisent et les accompagner dans leur montée en compétence sur le *personal knowledge management*[19] - à un niveau individuel donc - et sur les pratiques attendues dans l'organisation (animation de session de formation, construction de dispositifs intuitifs, architecture de l'information telle que l'expérience utilisateur passe au premier plan etc.). Les collaborateurs ne sont égaux face à la gestion de l'information et n'ont surtout la plupart du temps ni

Photo by William Iven on Unsplash

les mêmes logiques ni les mêmes réflexes (ni les mêmes personnalités en fin de compte !).

Quiconque travaille dans une grande structure ou bien recherche des informations précises sur des portails institutionnels sait au combien il est parfois difficile d'obtenir la bonne information (valable) au bon moment.

Dans les entreprises de taille significative, d'un point de vue

documentaire, c'est le Records Management qui vise à répondre à cette problématique en proposant l'organisation de documents applicables même si l'expliquer de la sorte est très réducteur.

Concernant la diffusion d'information à destination de l'ensemble des collaborateurs ou de division organisationnelle en particulier, on cherche à répondre à ce besoin par des intranets dont les fonctionnalités évoluent aussi au fil des ans pour converger aujourd'hui pour certains vers des réseaux sociaux d'entreprise (RSE).

Le travail collaboratif s'effectue quant à lui sur des serveurs de fichiers traditionnels, logique assez répandue, des espaces de travail collaboratif ou dans des outils collaboratifs spécifiquement mis en œuvre.

La principale limite rencontrée dans « cette coexistence technologique » est la « mise en silo » de l'information et de la documentation et l'absence de lien véritable entre, non pas les technologies, mais les contenus et le sens de ces contenus. Le savoir est en quelque sorte « fragmenté ».

À ceux qui diraient « c'est accessoire, non ? », nous répondons que cela peut constituer un vrai problème ; ou plutôt un véritable axe d'amélioration. Manager avec efficacité l'information s'inscrit complètement dans l'intelligence économique.

Les systèmes en place sont souvent très influencés par les outils présents et prédominants sur le marché et ces derniers sont proposés par des éditeurs se livrant une concurrence impitoyable. Ils mènent donc le jeu de l'interopérabilité et de la richesse fonctionnelle en fonction de leurs propres plans. Les produits sont parfois très peu personnalisés aux besoins véritables des usagers.

Il est difficile voire impossible de quantifier le gain relatif à la valorisation et l'exploitation de ce patrimoine immatériel que constituent les corpus info-documentaires des organisations.

Ces aspects sont souvent mésestimés dans les sciences de gestion et dans les méthodologies, même modernes, visant à configurer puis déployer les systèmes d'information[20]. Concernant les softwares, la culture anglo-saxonne prédomine avec la logique de briques applicatives telles que nous la connaissons aujourd'hui et l' « obsolescence programmée ».

Elle apporte des outils, certes d'une grande qualité et d'une conception remarquable mais cela reste … des outils. D'autre part, les utilisateurs ne connaissent éventuellement que leurs propres besoins lorsqu'ils sont consultés. Or, avec la profusion de données et d'information, il faut plus que jamais raisonner en termes de dispositifs d'information et de communication et ce, avec une vision systémique.

Valoriser, diffuser tout en protégeant l'information stratégique qui est à la base de tous processus de décision au sein des organisations.

L'utilisation des Big Data pour valoriser le patrimoine immatériel

Les organisations ont besoin de professionnels pour gérer, organiser et structurer le patrimoine informationnel. Utiliser les Big data doit permettre d'utiliser des connaissances que l'on peut exploiter. Le seul traitement des données, aussi intelligent et sophistiqué soit-il, n'est pas une fin en soi pour les organisations ou alors il faudra une grille de lecture sur ce que font les algorithmes et on peut supposer que le besoin de médiation prend place à ce niveau ?

Par ailleurs, comme sur le Web, au sein des organisations, le numérique désoriente et provoque un état de surcharge cognitive (information overload) pour l'usager lorsque les informations ne sont pas structurées et organisées au niveau attendu. Diffuser et protéger les informations. Relier et proposer des ressources à celui qui manifeste un besoin informationnel. Apprendre à le comprendre et s'adapter à lui. Capitaliser et pouvoir faire un « état de l'art » en peu de temps sur un sujet déjà initié. Trouver aisément des informations reliées entre elles par le sens.

Avec l'avènement et la généralisation d'internet, les bibliothèques ont eu une phase de remise en cause et ont dû mener une réflexion pour se réinventer. Elles innovent aujourd'hui bien plus qu'on ne le croit avec des offres de services parfois méconnues du grand public (service de questions/réponses, orientation des établissements en « tiers lieu », services personnalisés etc.). À l'image du e-commerce, l'apport des Big Data est également étudié. À un niveau macroscopique, l'analyse du besoin informationnel des usagers alimente la réflexion relative à la politique d'acquisition de nouvelles ressources et du renouvellement des collections.

Dans l'absolu, sous réserve de disposition relative à la protection des données personnelles, la mise en corrélation des profils avec les ressources parcourues et empruntées, ou l'étude entre les termes recherchés et les notices bibliographiques sélectionnées etc. sont des axes intéressants pour s'adapter à sa communauté d'usagers et déceler l'évolution des besoins.

La représentation visuelle des informations

Un phénomène accompagne les évolutions des Big Data, et en fait partie d'après la définition initialement proposée : la visualisation de l'information. Les professionnels de l'information, appelés aussi travailleurs du savoir, s'efforcent de décrire les corpus info-documentaires par l'intermédiaire d'un certain nombre de traitements intellectuels, en s'adaptant aux contextes des groupes d'usagers. Ils sont en outre force de proposition pour décrire, catégoriser, relier, donner un maximum de sens aux ressources ici numériques. La visualisation de l'information est très intéressante car elle permet, lorsqu'elle est bien utilisée, de représenter des contenus et vise ainsi l'apport d'une vue sémantique globale. Elle amplifie en outre la cognition de celui qui l'utilise car il peut alors se concentrer sur l'exploitation des informations perçues. La visualisation de l'information s'appuie sur l'emploi de cartographies qui permettent de :

- « Saisir un problème dans son ensemble.

- Comprendre l'information dans son espace contextuel.

- Co-construire des représentations mentales.

- Changer d'angle de vue.Présenter les connaissances de manière impactante. »

Des avancées récentes dans le domaine de la visualisation permettent de s'intéresser de près à des fonctionnalités telles que la navigation dans les corpus documentaires : parcourir le corpus en fonction des descripteurs et concepts qui auront été au préalable définis lors de l'intégration des ressources dans le dispositif selon des règles.

Conclusion

Tout en cherchant à clarifier ce concept dans une logique de vulgarisation, nous avons cherché à expliquer en quoi le Big Data était une révolution. Il bouleverse tous les pans de l'économie et l'on mesure encore mal ses impacts sociétaux. Pour y faire face, plusieurs types de compétences sont requis : des personnes très techniques – les data scientists – en charge de faire parler des données pour en extraire des informations, et en parallèle, des professionnels capables d'orchestrer les systèmes d'information ou les dispositifs d'une manière plus générale en ne perdant pas l'essentiel pour les usagers : les contenus. Les usagers comptent de plus en plus sur ces professionnels pour bénéficier de leur appui.

[1] Mot à la mode

[2] Schmitt, É. (s. d.). L'ambiguïté épistémologique des big data : le cas de la donnée web en sciences sociales. Consulté à l'adresse http://www.academia.edu/29272069/Lambigu%C3%AFt%C3%A9_%C3%A9pist%C3%A9mologique_des_big_data_le_cas_de_la_donn%C3%A9e_web_en_sciences_sociales

[3] Mayer-Schönberger, V., & Cukier, K. (2014). Big data : la révolution des données est en marche. Paris: R. Laffont.

[4] Groupement français de l'industrie de l'information. (s. d.). Big data, exploiter de grands volumes de données : quels sont les enjeux pour les acteurs du marché de l'information et de la connaissance ? - Journée d'étude - GFII. Consulté 18 juin 2017, à l'adresse https://www.gfii.fr/fr/evenement/big-data-exploiter-de-grands-volumes-de-donnees-quels-sont-les-enjeux-pour-les-acteurs-du-marche-de-l-information-et-de-la-connaissance

[5] *Ibidem.*

[6] Delort, P. (2015). Big Data : concepts et définition. Que sais-je ?, 29-47.

[7] Cartier, M. (s. d.). Le 21e siècle : le Big Data. Consulté 19 mai 2017, à l'adresse http://www.21siecle.quebec/table-des-matieres-2/le-big-data/

[8] Ibekwe-Sanjuan, F., & Paquienséguy, F. F. (2015). Open, Big, Collaboration : trois utopies de l'innovation au 21e siècle. In G. Chartron & E. Broudoux (Éd.), Big data - Open dataQuelles valeurs ? Quels enjeux ? (p. 15-30). De Boeck. Consulté à l'adresse https://hal.archives-ouvertes.fr/hal-01271656

[9] Portail Data Mining - Université Lyon 2. (s. d.). Data mining - Data Science - Big Data Analytics. Consulté 11 juin 2017, à l'adresse http://eric.univ-lyon2.fr/~ricco/data-mining/

[10] Comprendre la conversion d'une visite en vente

[11] L'effet de la « longue traîne » correspond au phénomène selon lequel, dans les ventes en ligne, le chiffre d'affaire issu de la somme de nombreux produits peu populaires dépasse celui généré par quelques produits très à la mode et donc très achetés.

[12] Les Vidéos du Succès. (2012). Comprendre la BIG DATA avec Captain Dash. Consulté à l'adresse https://www.youtube.com/watch?v=u3uH6MYrAmI

[13] *Ibidem.*

[14] Répertoire d'Autorité-Matière Encyclopédique et Alphabétique Unifié

[15] Deschamps, C. (2010). Le nouveau management de l'information : la gestion des connaissances au coeur de l'entreprise 2.0. Limoges: Ed. FYP.

[16] Le site du CEDIP : Compétences et formation. (2006, juillet 26). Savoir. Consulté 2 juin 2017, à l'adresse http://www.cedip.developpement-durable.gouv.fr/savoir-a314.html

[17] Bibent, M., Bouwens, M., Buckland, M., & Cacaly, S. (2008). Dictionnaire de l'information. Paris: Colin.

[18] Galaup, X. (2012). Développer la médiation documentaire numérique. Villeurbanne: Presses de l'ENSSIB.

[19] Gestion des connaissances personnelles

[20] Système d'information devrait être compris dans son sens premier : « *Ensemble cohérent constitué par l'identification et la description des processus liés à une tâche, un ensemble de tâches ou un métier et des produits et services d'information nécessaires au fonctionnement de ces processus* ». C'est par extension que les technologies et moyens informatiques entrent en ligne de compte. Cacaly, S., Le Coadic, Y.-F., Paul-Dominique, P., & Sutter, E. (2008). Dictionnaire de l'information (3ème édition). Paris: A. Colin.

[21] Surcharge d'information (vision subjective)

[22] Card, S. K., Mackinlay, J. D., & Shneiderman, B. (1999). Readings in information visualization : using vision to think. San Francisco, Calif.: Morgan Kaufmann Publishers.

[23] Tricot, C., Roche, C., Foveau, C. E., & Reguigui, S. (2006). Cartographie sémantique de fonds numériques scientifiques et techniques. Document numérique, 9(2), 12-35.

[24] Bulinge, F. (2014). Maîtriser l'information stratégique : méthodes et techniques d'analyse. Bruxelles: De Boeck.

BIG
DATA
kindle

Big Data et gouvernance des données : Vincent Moreno - HUPI

■ Propos recueillis par Valérie Griman, le 6 juin 2017

HUPI est une société implantée à Bidart dans le Pays Basque, spécialisée dans l'édition de logiciels en mode SaaS (Software As A Service) dans le domaine du Big Data. HUPI propose des plateformes opérationnelles de traitement et d'analyse de données en continu afin de créer de « l'intelligence automatisée ».

M. Vincent MORENO est co-fondateur de l'entreprise, créée en 2014.

Pour commencer l'interview, pourriez-vous me décrire votre parcours ? Quel est votre profil de formation et comment est née l'idée d'une entreprise d'édition de logiciels « Big Data » ?

J'ai passé un Master en informatique à l'Université de Bordeaux, puis un MBA à Paris (Master of Business Administration) auprès du CNAM.

L'idée de l'entreprise est venue du croisement des compétences des trois fondateurs : un spécialiste des aspects territoriaux, un profil axé sur le partage des connaissances stratégiques et moi plus technique.

Le concept était d'offrir aux entreprises de nouveaux services dits intelligents pour pouvoir les aider demain à progresser. Nous avions travaillé sur le Big Data dans nos expériences passées et y avons vu une opportunité.

Vous fournissez des abonnements à une plateforme locale Big Data ? Quel en est le principe ?

Il faut d'abord commencer par la définition donnée au Big Data chez HUPI pour expliquer le principe des plateformes. C'est un outil qui va nous permettre de faire des recommandations automatisées en temps réel. C'est-à-dire comment à partir des données et des outils de type machine learning, on est capable de faire de la prédiction, de la recommandation en temps réel et de manière automatisée. C'est la machine qui va préconiser à l'homme quelque chose à faire.

Chez HUPI, nous avons développé des solutions qui sont basées sur notre propre plateforme d'où l'avantage concurrentiel. Nous proposons tout un socle technique qui appartient à HUPI et qui permet de ne pas réinventer la roue chaque fois que nous devons travailler sur une solution de ce type-là pour nos clients.

La plateforme est découpée en trois axes verticaux métier : e-commerce, industrie et transport. L'e-commerce étant notre clientèle principale suivie de l'industrie.

Comment identifiez-vous avec le client potentiel le flux de données brutes ? Sur quels critères ?

Dans les faits, nous faisons le contraire. Avant de dire « qu'est-ce que vous avez comme données ? » et après on va réfléchir, ce qui marche une fois sur 100. La démarche est plutôt de réfléchir avec le client à leurs

usages et leurs problématiques de tous les jours. Une fois ce contexte défini, nous allons réfléchir à la typologie de données qui pourraient potentiellement répondre à ces besoins, en chercher de nouvelles et compléter par d'autres systèmes auxquels nous n'aurions pas pensé directement.

Quand nous allons voir les clients, nous ne pouvons pas parler outils. Si nous leur parlions « machine learning », il n'y a pas de connexion qui se ferait entre leurs métiers et nos métiers. Pour établir cette connexion, nous essayons d'accompagner nos clients sur la réflexion. Nous les faisons parler par un jeu de questionnements sur leur stratégie qui, nous, nous aident derrière à dessiner leurs besoins et associer à nos solutions potentielles ou non.

Exemple de e-commerce : jeu de questions sur le Qu'est-ce vous vendez ? Pourquoi vous le vendez ? Pourquoi avoir choisi tel ou tel canal de ventes ? Quelle catégorie de clients ? Pourquoi leur fournissez-vous telles informations plutôt que celles-là ?

« *Du coup je fais un parallèle avec mon métier principal qui est la Qualité. Mais en bref, vous les aidez à avoir une vision systémique et à cartographier leurs activités pour bien cerner leur fonctionnement ? à identifier leurs risques et leurs opportunités ? C'est exactement ça !* »

Sur votre site, vous écrivez que votre plateforme devient pour le client son « usine de traitement et valorisation analytique des flux de données ». Je trouve le choix du mot « Usine » très intéressant. Mais pourriez-vous expliquer pourquoi ce choix ?

C'est mon collègue qui a eu l'idée pour ce mot. Tout simplement car au début de l'aventure, le mot « plateforme » ne parlait à personne. Alors que le mot « usine », par rapport au Big Data, était beaucoup plus simple à expliquer aux clients. En effet, comme dans une usine, il y a des matières premières qui vont entrer. Comme dans l'usine, vous avez des machines-outils qui vont travailler ces matières premières pour faire un nouveau produit. Et c'est ce nouveau produit que vous allez vendre. Voilà pour l'image !

Ce qui était important pour nous également dans le mot « usine » était d'expliquer cette notion de machine-outil qu'il y avait à l'intérieur, bien calibrée et spécifiée pour faire chacune leur travail.

Pourriez-vous expliquer de façon simple comment vous parvenez à traiter les flux de données hétérogènes (sources multiples, structurées et non structurées) ?

Comment parvient-on à gérer différents flux de données et à les faire interagir entre eux et à leur donner du sens ? La plateforme a été pensée pour ça : comment recevoir des flux de différents formats en temps réel ? Un module basé sur une technologie « code source » qui va permettre de faire des correspondances a été développé et ne fait que ça. Il se nourrit de données de tout horizon et réinterprète ces données pour qu'elles se comprennent entre elles. »

La machine va définir la stratégie des entreprises de demain ?

« Je dirai plutôt que les stratégies de demain ne pourront pas écarter ces outils-là. Pour moi, le Big Data remplacera les cadres moyens dans certaines activités.

Nous donnons des cours dans les écoles d'Ingénieurs informatique et les universités car l'idée n'est pas d'en avoir peur. Il faut essayer de comprendre ce qu'il se passe, d'apprendre à utiliser ces nouveaux outils plutôt que de les rejeter.

Aujourd'hui, c'est une évidence que ces produits sont partout, les IoT sont partout. Dans nos smartphones, il y a plus de 70 % de nos applications mobiles qui utilisent le Big Data, qui utilisent les machine learning pour nous faire des propositions BtoB ou BtoC. Nous expliquons aux étudiants qu'ils devront intégrer ces outils pour ne pas être en retard par rapport à leurs collègues. Il faut apprendre à travailler avec ce genre d'outils. »

Selon vous, la valorisation analytique d'une entreprise peut être une opportunité de différenciation : pouvez-vous développer ?

Quand on parle du Big Data, on parle d'une nouvelle source de richesse : c'est le nouvel or noir. Nous, chez HUPI, notre métier est de faire des recommandations automatisées et en temps réel. C'est-à-dire que l'analytique souvent est entendu par

des personnes qui vont se rapprocher de la Business Intelligence.
L'analytique au sens HUPI regroupe les modèles mathématiques qui peuvent guider sur des choix stratégiques : hausse ou baisse de prix sur un produit.

À partir de l'historique des données d'une entreprise et tout le flux de données qu'elle reçoit, elle peut créer de nouveaux services intelligents qui pourront faire une différentiation par rapport à sa concurrence. C'est là qu'est tout l'intérêt pour une entreprise.

Dernière question : Vous avez participé avec la CCI Pays basque à des présentations sur le « Big Data par et pour les nuls » fin 2016 et cette année ? Qu'est-ce qui a suscité cette démarche ? Quel est votre ressenti sur le niveau d'appropriation du Big Data dans les entreprises ?

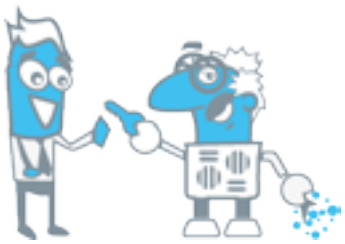
Le sujet est travaillé depuis deux ans et demi avec la CCI. Au départ, c'était davantage une « évangélisation ». Le mot « big data » au début, tout le monde l'entendait mais n'en comprenait pas vraiment les tenants et les aboutissants. Pour beaucoup, il y avait quelque chose qui se passait.

Donc le travail d'évangélisation a été d'expliquer la définition du Big Data pour que tous comprennent ce que HUPI pouvait leur offrir dans leur cadre métier. L'explication sort de l'aspect purement technique, perçu comme très complexe, pour partir des usages et des problématiques actuels métier des entreprises.

Aujourd'hui ce que l'on relève de plus en plus, c'est que les sociétés comprennent, veulent y aller mais ne savent pas encore par quel bout le prendre car elles n'ont pas encore de retour sur investissement. Les dirigeants comprennent aujourd'hui dans quels secteurs d'activités ils pourraient l'appliquer mais sont encore dans la réflexion de comment l'intégrer dans leur processus métier, que ce soient des PME ou des grands groupes. Comment intégrer ces parties là car elles ne sont pas encore très maîtrisées. Les entreprises ne rajoutent pas de nouveaux services sans réfléchir.
Quand on travaille avec un client, il y a deux cas de figure : arriver à l'accompagner dans une problématique connue qu'il n'arrive pas à résoudre. Dans ce cas, nous nous opposons à des experts dans ces métiers-là. Il devient vite très compliqué d'expliquer qu'avec une technologie il est possible de trouver une solution alors qu'eux experts n'y parviennent pas.
Notre astuce c'est de présenter nos services comme étant complémentaires pour accompagner à la prise de décision. Quand on est dans l'accompagnement, les discussions sont beaucoup plus simples. Ce qui permet aux experts de recroiser les données avec leurs expertises, leurs connaissances ou d'autres sources d'informations et donc de les intégrer à la démarche.

Le Big Data et le management de l'information en général devraient aider à repenser et restructurer l'entreprise ?

C'est exactement ça. Aujourd'hui, dans le monde industriel, les dirigeants devront intégrer ces nouveaux services dans leur fonctionnement. Une entreprise qui ne s'intéresse pas à ce sujet et qui ne l'intègre pas au moins pour avoir une



vision globale sur ce qu'il se fait ou ce qu'il pourrait faire, passera à côté d'un sujet qui pourrait être vital pour l'entreprise. Mais aujourd'hui le Big Data est encore pour eux en phase d'expérimentation.

---

[1] Définition selon le site http://www.jeveuxetredatascientist.fr/quest-ce-que-le-machine-learning/ : « *consiste en la mise en place d'algorithmes ayant pour objectif d'obtenir une analyse prédictive à partir de données, dans un but précis* ». C'est en quelque sorte l'apprentissage par l'exemple. Avec le Machine Learning, on cherche davantage à établir des corrélations entre 2 évènements plutôt qu'un lien de causalité.
Exemple: on peut détecter une corrélation entre la consommation de sucre et les maladies cardiaques, sans pour autant dire que l'une est la cause de l'autre.

Site web : http://www.hupi.fr/

Thibaut Britz, fondateur de l'outil de veille nouvelle génération Talkwalker

■ Propos recueillis par Hélène Kaiser

Thibaut Britz est le fondateur et le CTO de Talkwalker, outil de veille et d'analyse du web et des réseaux sociaux.

Quel est votre parcours ? Quand avez-vous commencé à vous intéresser à la veille ? Quel a été le déclic qui vous a fait prendre conscience que la veille des réseaux sociaux pouvait être une valeur ajoutée dans l'information stratégique ?

L'aventure de Talkwalker a commencé en 2009, lorsque j'ai créé l'entreprise avec Christophe Folschette.
J'ai toujours aimé construire de nouveaux outils, et Christophe et moi-même étions tous deux très conscients du potentiel que représentaient les médias sociaux pour les entreprises. Nous voulions leur fournir une technologie avancée qui leur permettrait de collecter et d'analyser automatiquement des quantités importantes de données brutes issues du web avec une précision maximale. Nous proposons aujourd'hui un outil de veille et d'analyse du web et des réseaux sociaux qui aide nos clients à protéger leur marque mais aussi à identifier ce qui est dit à leur propos à travers les blogs, les forums, les sites d'actualité, les réseaux sociaux, et bien plus. L'outil permet aussi à nos clients d'optimiser leurs initiatives sociales sur le web, pour gagner en performance. Il est encore particulièrement utile pour analyser la manière avec laquelle une marque performe par rapport à ses concurrents.

Avec les années, nos clients sont devenus plus matures. Nous avons de plus en plus de gros contrats qui courent sur un ou deux ans. Je pense que dans tous les secteurs, B2C, mais aussi B2B, le potentiel des médias sociaux est bien compris. Ce qui change c'est le nombre grandissant d'entreprises de toute taille qui ont recours à des outils professionnels.

Avec Talkwalker, vous proposez de communiquer de l'information stratégique aux décideurs d'entreprise, pouvez-vous expliquer comment cela fonctionne ? Quelles sont les opportunités du Big Data pour la veille stratégique ?

Le web et les réseaux sociaux représentent une quantité de données phénoménales que les entreprises peuvent exploiter pour répondre à de nombreuses questions qu'elles se posent au quotidien : dois-je investir davantage d'argent dans ma campagne publicitaire ? Comment notre dernier lancement produit a-t-il été reçu ? Comment ma marque est-elle perçue sur ce marché spécifique ? Pour obtenir des analyses encore plus détaillées, ils combinent également nos données avec leurs propres informations internes (chiffres de vente, taux de satisfaction client, cours de l'action, etc.) et obtiennent ainsi des informations stratégiques qu'ils peuvent utiliser pour prendre des décisions informées et atteindre leurs objectifs commerciaux.

À quelles difficultés avez-vous été confronté lors de l'élaboration de Talkwalker ? Comment y avez-vous fait face ? Comment avez-vous perçu l'opportunité de travailler dans un incubateur du centre de

recherche public Henri Tudor ? Qu'est-ce que cette expérience vous a apporté ?

Je pense qu'un entrepreneur devrait pouvoir se concentrer sur les trois choses essentielles dont dépend la réussite de son projet : le produit, la vente et le marketing ainsi que les ressources humaines pour constituer les bonnes équipes. Tout le reste, aussi important soit-il, la logistique, l'administratif, le networking avec des financiers, peut être efficacement apporté par des structures d'accompagnement comme les incubateurs. Ces derniers nous ont permis de rester concentrés sur nos fondamentaux et donc de nous développer plus vite.

Quand et pourquoi avez-vous eu l'idée d'utiliser la datavisualisation pour communiquer les informations ? Est-ce lié au volume des données ou bien à la lisibilité des informations ?

Quand nous avons commencé à développer notre outil, nous avions deux buts importants en tête : il devait être à la fois puissant et facile à utiliser. Nous ne voulions pas créer une plateforme qui serait simplement utilisée par des experts de la donnée - elle devait pouvoir être accessible à toutes sortes d'équipes dans l'entreprise, du département marketing aux relations publiques en passant par le légal et la gestion de risques.

La datavisualisation permet aux membres de ces différentes équipes d'analyser rapidement des quantités importantes de données et de les intégrer en un clic à un rapport qui peut être facilement compris par leurs managers. Ces analyses visuelles sont très populaires, comme par exemple notre carte de viralité, qui analyse la façon dont une publication se propage sur le web en fonction de la plateforme, de la langue et du sentiment.

En chiffres, combien de posts collectez-vous en 24h ? Combien de temps vous faut-il pour les analyser ?

Nous collectons des données au quotidien sur plus de 150 millions de sites web et une dizaine de réseaux sociaux comme Twitter, Instagram, Facebook, Pinterest mais aussi Weibo en Chine et Vkontakte en Russie. Chaque jour, 300 millions de nouvelles données sont ainsi stockées et analysées sur les 1500 serveurs qu'opère l'entreprise pour que nos clients puissent bénéficier d'informations en temps réel.

Pourquoi s'implanter aux États-Unis alors que le web met toutes les données à disposition n'importe où ? Quelle est votre perception du besoin de présence et de rencontre en réel alors que les nouvelles technologies sont présentées comme facilitateur de liens ?

En effet, notre logiciel peut être directement utilisé en ligne, ce qui nous permet de nous déployer très efficacement. Il reste cependant que chaque marché a ses spécificités, avec des besoins et des niveaux de maturité différents. Nos équipes américaines développent leurs propres réseaux de contact, ce qui leur permet de mettre en avant la qualité de notre produit par rapport à nos concurrents.

Avant l'ouverture de nos bureaux new yorkais, nous avions déjà des clients américains, mais au vu de l'augmentation de leur nombre, nous avons décidé d'ouvrir des bureaux locaux, d'abord sur la côte est, puis tout récemment à San Francisco pour leur offrir le meilleur service client et développer notre présence sur l'entièreté du continent. Deux ans après nos premiers pas sur le sol américain, le bilan est plus que positif, autant du point de vue du chiffre d'affaires que de la réputation d'excellence que Talkwalker se construit en Amérique du Nord.

Quels sont vos prochains challenges ? Vos prochains défis ?

Pour Talkwalker, l'enjeu est de toujours innover. Du fait de notre clientèle internationale, notre produit doit rester à la pointe d'un marché en constante évolution et apporter à nos utilisateurs des analyses avec toujours plus de valeur et de détails. Il est pour cela nécessaire d'attirer de nouveaux talents avec les bonnes compétences – notre but cette année est de recruter des dizaines de nouveaux collaborateurs pour atteindre les 180 employés ! D'autre part, des efforts considérables sont consentis en R&D mais aussi pour se positionner au plus proche du client et de ses besoins, afin de lui proposer de nouvelles solutions. Il y a un an, nous avons ainsi présenté notre solution de reconnaissance de logos à travers le web. Grâce aux avancées de notre département spécialisé en machine learning, nous sommes maintenant à même de détecter les logos de plus de 30.000 marques dans les images partagées sur le web, ainsi que d'identifier les objets et les paysages présents dans ces images. Bientôt, nous serons à même d'analyser le contenu des vidéos. Ce développement répond à des attentes de clients mais aussi à des tendances, avec une utilisation grandissante du contenu visuel sur les réseaux sociaux.

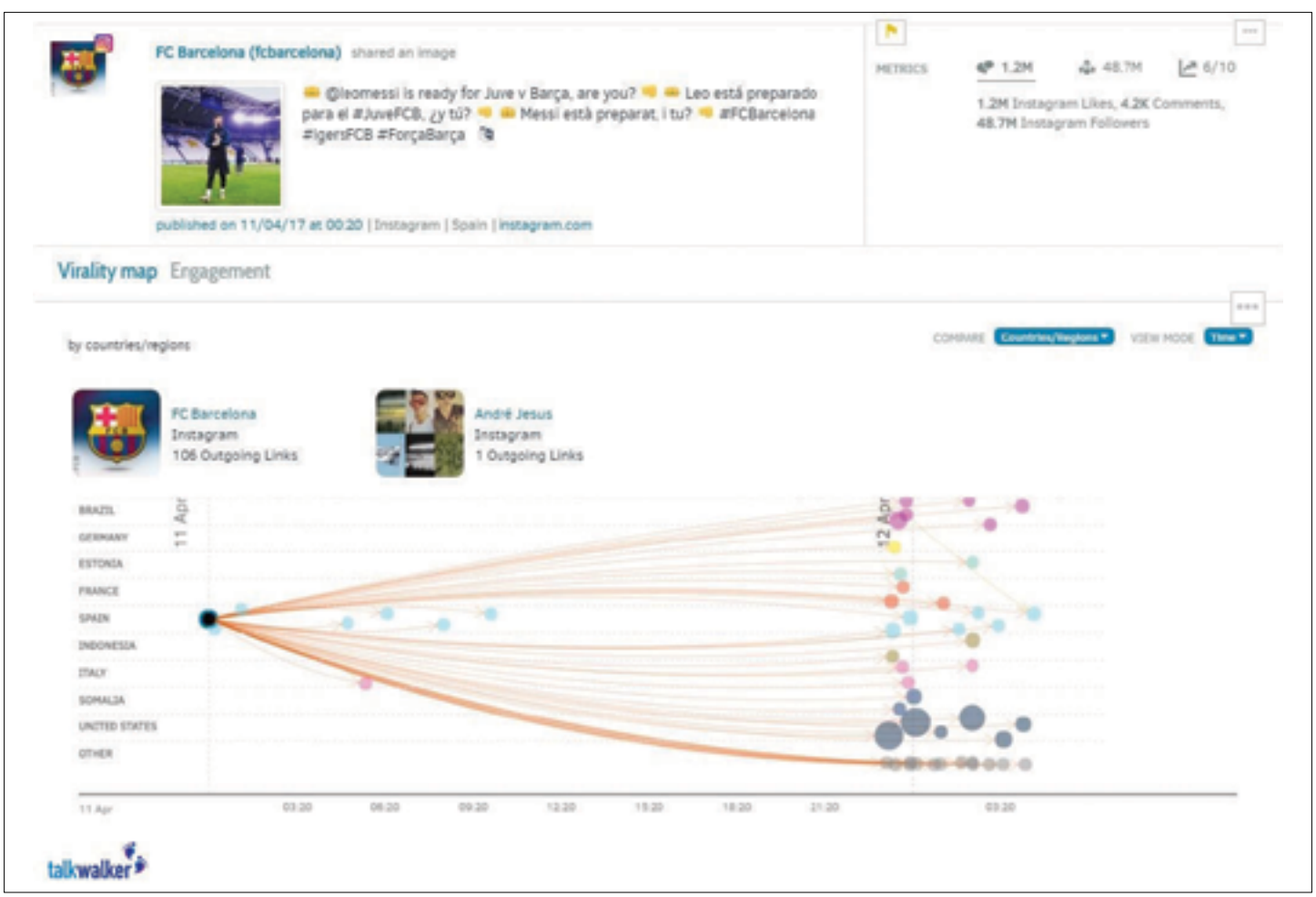
talkwalker

A PROPOS DE TALKWALKER

Talkwalker (www.talkwalker.com) est un outil de recherche, de veille et d'analyse des médias sociaux simple à utiliser et ultra performant. En plus des fonctionnalités incontournables de veille de mots-clés et de hashtags sur les réseaux sociaux, les blogs et les forums, Talkwalker propose des tableaux de bord puissants et flexibles, des rapports sur mesure et une API spécifique selon les besoins clients, permettant aux sociétés et agences de prendre de meilleures décisions stratégiques, chiffres à l'appui.

Talkwalker est un produit certifié par Twitter et a été classé parmi les dix meilleurs outils de veille et d'e-réputation au niveau mondial par Goldbach Interactive (2013, 2014, 2015). Sa technologie de pointe scanne plus de 150 millions de sources web en temps réel en 187 langues.

La plateforme Talkwalker est utilisée par plus de 700 clients dans le monde, comme par exemple des entreprises telles que Volkswagen, Benetton, KPMG et des spécialistes de la communication tels que Publicis, Ogilvy et Weber Shandwick.



Carte de viralité de l'outil de veille Talkwalker

Données de la recherche, Entretien avec Lucie Albaret

■ Propos recueillis par Aurélie Quinodoz le 14 mars 2017

Lucie Albaret est conservatrice de la Bibliothèque Universitaire de Sciences - UGA (Grenoble)

Lucie Albaret, vous êtes conservatrice à la Bibliothèque Universitaire Sciences de Grenoble. Nous nous sommes rencontrées l'an passé alors que vous présentiez les avancées de l'Open Access pour les chercheurs à la librairie Arthaud (Grenoble). Aujourd'hui, la question de la diffusion de la recherche semble s'étendre aux données de la recherche. Pouvez-vous faire un bref retour sur l'Open Access et sur le rôle des bibliothèques dans ce mouvement ?

Je dois dire tout d'abord qu'il s'agit de mon point de vue et qu'il y a beaucoup de littérature sur le sujet. Il y a notamment le blog de Marlène Delhaye qui s'appelle Marlene's corner. Elle a écrit de nombreux billets sur le sujet et sur l'évolution du rôle des bibliothèques et la nécessaire réflexion que les professionnels doivent avoir sur leur métier. Beaucoup de littérature professionnelle a également été écrite sur ces questions.
Les bibliothèques se sont emparées assez tôt du sujet. Dès les années 2004, des écrits ont été publiés sur ces questions, notamment en raison du coût d'accès aux publications numériques.

Pourquoi est-ce important pour les bibliothèques ?

Notre mission est de donner accès à la documentation pertinente et dont les publics ont besoin. On a très vite été confronté aux difficultés de gestion des collections numériques et notamment des abonnements numériques. On s'est transformé en donneurs d'accès avec parfois l'obligation d'interdire des accès à tel ou tel labo qui n'avait pas d'abonnement.

Cela est un changement dans la philosophie de la bibliothèque. L'Open Access nous a semblé philosophiquement intéressant et un des intérêts est de revenir au cœur de l'édition scientifique c'est-à-dire faire circuler vite la connaissance et les découvertes entre des communautés parfois éloignées. Il faut que les gens échangent, critiquent et amendent, sans obstacle financier ou de temps. Plus c'est rapide et mieux c'est. L'Open Access permet aussi une réflexion sur le patrimoine scientifique et de garder la mémoire de la construction scientifique. C'est un peu à l'opposé de l'édition scientifique qui, avec sa concentration de gros monopoles a oublié ses missions de départ de circulation du savoir. Les chercheurs ont par ailleurs perdu, avec les questions d'évaluation, leur capacité à élaborer leur stratégie de publication. Qu'est-ce que je veux faire de ce texte ? Comment je veux le diffuser ? L'Open Access remet le chercheur au cœur de la façon dont il souhaite communiquer avec ses pairs et avec la société puisque c'est l'Etat qui finance la recherche.

Du côté des bibliothèques, elles peuvent assurer la visibilité de la production de son université et ses spécificités dans tels ou tels domaines. Cela converge avec les problématiques de l'Open Access. Avec la mise en place de plateformes institutionnelles ou des portails HAL, les bibliothèques se sont emparées du sujet. Mais elles ne peuvent faire cela toutes seules. Avant tout, cela doit

PLATEFORMES RECOMMANDÉES

Biosharing (médecine) :
https://bioshraing.org/

INRA Science et Impact :
https://www6.inra.fr/datapartage/

SEANOE : http://www.seanoe.org/

Re3data (registre mondial des dépôts de données de recherche) :
http://www.re3data.org/

OpenAire : https ://openaire.eu

rencontrer l'intérêt de la communauté universitaire. Il faut prendre garde à ce que le portail ne devienne pas la chose de la bibliothèque. Cela doit être la chose du chercheur et si des décisions doivent être prises, c'est au sein de la commission Recherche de l'Université. C'est la chose de la communauté.

Maintenant que le mouvement de l'Open Access est largement partagé, l'ouverture des données de la recherche étend encore la diffusion. La loi Pour une république numérique ou encore certains programmes d'Horizon 2020 recommandent, voire imposent, le dépôt du plan de gestion des données. Pouvez-vous nous dire où en est l'Université avec l'ouverture des données ?

Les universités à l'étranger sont plus en avance que les universités françaises. Je pense notamment à la Grande-Bretagne, au Pays-Bas ou encore la Suisse avec l'Ecole polytechnique de Lausanne. Le programme Horizon 2020 a été un coup d'accélérateur au niveau européen. Après le libre accès, il a été très vite question du partage des données, à la fois pour vérifier la qualité de ce qui est produit et parce que certaines données sont très coûteuses à produire comme les données astronomiques. Aussi pour ne pas s'épuiser à trouver des financements pour reproduire des données existantes. Tous les champs sont concernés (sciences sociales et sciences dures).

Ce qui est important ce sont les plateformes thématiques (comme le CDS en astronomie ou encore Huma-num pour les humanités numériques). Il existe des plateformes en sciences sociales, en sciences de l'environnement. En fait, il y a une coexistence de plateformes thématiques, par exemple pour les données médicales comme Biosharing et de plateformes institutionnelles. Les plateformes sont bien référencées par des sites comme Re3data notamment.

Comment cela fonctionne-t-il pour le dépôt ?

Les communautés de recherche sont porteuses des dépôts. On sait normalement où aller. Par exemple, si on travaille sur toutes les données liées aux enquêtes qualitatives, il y a des infrastructures qui gèrent ce type de données. Pour les dépôts des données en sciences de la terre et de l'environnement, on pourra se tourner vers Pangaea. Les infrastructures permettent une relation contractuelle et de bénéficier des outils mis à disposition.

Le dépôt est-il fait seul par les chercheurs ?

Cela dépend des communautés. Dans certains cas, la question ne se pose pas. Cela dépend aussi des financements de type européen avec l'obligation de réalisation d'un plan de gestion des données. La plateforme *Doranum* permet de référencer des outils, des tutoriels qui permettent de se familiariser très vite avec les plans de données.

Et vous, en tant que bibliothécaire, est-ce que vous présentez des plans de données, est-ce que vous aidez à les monter ?

Pour l'instant, cela ne s'est jamais trouvé. À Grenoble, dans le cadre de l'Idex (Initiative d'excellence), la mise en place d'un data centre est en cours de concrétisation avec différentes briques qui correspondent à certaines thématiques (données médicales, humanités numériques, données liées à l'environnement). Il sera équivalent à celui l'Université de Paris-Saclay. Actuellement, nous ne participons pas à la création d'un plan de gestion des données mais cela fait partie des évolutions incontournables de nos missions. Nous sommes néanmoins associés au signalement et au dépôt des jeux de données du Labex Persyval (maths appliquées, informatique appliquée). La bibliothèque a été associée au titre de la participation à HAL. Dans le Grenoble Data Center, la bibliothèque devrait être associée au titre des humanités numériques. Mais nous ne savons pas encore précisément quel sera notre rôle dans cette infrastructure. La formation est au cœur de nos problématiques actuelles pour accompagner les chercheurs dans cette nouvelle étape.

Merci à Lucie Albaret pour son accueil et ce généreux échange !

LES RECOMMANDATIONS DE LUCIE ALBARET

Huma-Num :
http://www.humna-num.fr/

Doranum :
http://www.doranum.fr/

Marlène's corner :
https://www.marlenescorner.net

Sandra Bringay

■ Propos recueillis par Alexandre Hubert

Sandra Bringay Enseignante à l'Université Paul-Valéry Montpellier 3 en Mathématiques et Informatique Appliquées. Vous co-dirigez le Master Mathématiques et Informatique Appliquées aux Sciences Humaines et Sociales (MIASHS). Vous êtes en outre chercheur au Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier (LIRMM) dans l'équipe ADVANSE.

L'objectif du master que vous proposez est de former des spécialistes capables de valoriser l'ensemble des données des organisations, est-ce bien cela ? Pourriez-vous nous en dire plus sur les finalités professionnelles du Master ?

Oui. Typiquement, ce que l'on cherche à former, ce sont des spécialistes de l'analyse des données, qui vont être centraux dans l'entreprise, qui doivent connaître toutes les données produites et savent les valoriser. Ces data scientists sont opérationnels sur les aspects « collecte de données » qui peuvent être du texte, de l'image, du son et de la vidéo. Ils sont opérationnels sur les aspects « structuration » de ces données : cela implique toutes les nouvelles techniques de bases de données. Pour finir – et cela est très important – ils sont opérationnels sur les méthodes d'analyse statistique et informatique qu'il est très judicieux de combiner.

En stat comme en info, les étudiants vont apprendre des méthodes qui permettent d'avoir une vue d'ensemble sur les données et de tester des hypothèses. Nous allons explorer également des méthodes sans a priori, c'est-à-dire pour lesquelles on ne sait pas ce que l'on cherche dans les données mais qui vont permettre d'identifier des régularités. L'originalité de ce master, c'est que l'on ajoute à cette chaine de traitements une brique de « Visualisation ». En effet, on va aller mettre une suite à tous ces algorithmes info et stat pour que la personne pour laquelle on travaille, l'expert, puisse s'approprier de manière interactive les connaissances issues de ces méthodes. [...]

Un autre point fort du Master, c'est qu'il se déroule sous la forme de l'alternance. L'alternance c'est un rythme. Pour nous, deux semaines à l'Université puis deux semaines en entreprise. Pour leur alternance, les étudiants peuvent signer des contrats d'apprentissage, des contrats de professionnalisation et des conventions de stage. Cette organisation sous la forme de l'alternance permet aux étudiants d'avoir une expérience pro. dès le master 1.

Les activités de recherche de l'équipe ADVANCE s'inscrivent dans le domaine des grandes bases de données, i.e. Big Data, et plus particulièrement dans le domaine de l'Extraction de Connaissances et de la Fouille de Données. Dans quelle mesure votre équipe travaille-t-elle avec des linguistes, des professionnels de l'information-documentation ou en équipes pluridisciplinaires d'une manière générale ?

Nous travaillons généralement avec des équipes pluridisciplinaires.
Par exemple, nous travaillons beaucoup sur des méthodes de fouille de textes. Avec ces méthodes, nous

produisons une vue statistique de ce que l'on retrouve dans les textes. Par contre, si l'on veut aller plus profondément dans la compréhension des mécanismes du texte, on est obligé de passer par un travail de linguistique. À titre d'exemple, on a travaillé cette année avec Sacha Diwersy, Maître de conférence, Praxiling, sur une méthode à base de motifs récurrents pour étudier les expressions de sentiments. [...]

Sur les aspects « ontologies », nous travaillons beaucoup avec Clément Jonquet, Maître de conférences à Polytech Montpellier. Il est l'instigateur d'un projet qui s'appelle SIFR[1]. Ce projet vise à développer un serveur permettant de stocker tout un tas d'ontologies en Français, notamment « médicales », mais aussi en agronomie. Par ailleurs, dans le cadre d'une collaboration avec le statisticien Christian Lavergne de l'IMAG et Caroline Mollevi de l'ICM, nous avons proposé une nouvelle méthode pour produire automatiquement une ontologie des termes utilisés par les patients atteints d'un cancer du sein, qui a été intégrée dans Bioportal [...].

On travaille également avec des spécialistes de l'information et de la communication. Par exemple, avec Céline Paganelli du LERASS et Viviane Clavier du GRESEC, nous travaillons sur les réseaux sociaux pour détecter les pratiques sexuelles à risque dans le cadre d'une collaboration avec SIDA info service. Elles interviennent pour qualifier le type d'information à rechercher dans les messages. À partir d'un petit ensemble de données étiquetées, nous pouvons appliquer des méthodes d'apprentissage pour apprendre des modèles puis automatiser la production de ces étiquettes pour une grande quantité de messages.

Nous travaillons également beaucoup avec des utilisateurs finaux de nos chaines de traitements comme par exemple des professionnels de santé. Avec Jérome Azé, Professeur d'informatique de l'Université de Montpellier, nous étudions en collaboration avec des psychiatres du CHU de Montpellier les comportements à risque dans les réseaux sociaux. Avec le professeur Landais du CHU de Nimes, nous travaillons également sur les trajectoires de patients à partir des données médico-économiques du PMSI[2].

Dans ces aspects pluridisciplinaires, nous sommes dans le coeur de nos métiers. Nous faisons sans cesse des allers-retours entre la théorie qui va nous permettre de développer de nouveaux algorithmes et la pratique qui va nous permettre de remonter des problèmes. [...] Nos projets sont donc rarement sans expert.

Vous vous êtes spécialisée dans le domaine des méthodes de sciences des données. Pourriez-vous nous expliquer le champ d'application et le périmètre de ces méthodes ?

Je dirais que le périmètre est très large. De la collecte, au stockage, au traitement, jusqu'à la visualisation. En termes d'application, tous les domaines où l'on produit de la donnée sont concernés. Il y a l'industrie, les banques, les assurances, le web... Dès que de la donnée est produite, il y a de la donnée à valoriser. C'est ce que l'on voit dans les offres d'emploi pour étudiants du Master. Il y a deux types d'offres :

- des offres qui portent sur la production des méthodes d'analyse et de logiciels permettant d'extraire de la connaissance à partir de différentes données,

- des offres, la majeure partie en fait, viennent des producteurs des données issues de tous les domaines. Ce n'est pas la même finalité d'emploi. Dans ce cas, les demandes sont plus « métier ». Il est question ici d'aider les organisations à exploiter les données qu'elles produisent.

On voit d'ailleurs que les grands groupes savent déjà bien formuler ce besoin et emploient couramment le terme *data scientist* dans leurs offres d'emploi.

On peut donc dire que c'est en fonction du besoin perçu, du degré de maturité sur ce besoin et d'un intérêt éventuel pour des projets de Business Intelligence par exemple ?

C'est ça, tout à fait. Je pense que le besoin, il est partout, public, privé. Les données à traiter sont partout. Concrètement, ce n'est pas toujours facile car lorsque nous rencontrons les entreprises, parfois leurs données sont déjà structurées mais la plupart du temps, ce sont des fichiers Excel à agréger sur plusieurs années ! Il y a déjà donc déjà un grand travail à mener en amont pour rendre possible l'exploitation de ces données et surtout automatiser les processus.

Il faut donc que la collecte soit « propre » en amont ?

Oui, c'est ça. En général, on dit que 80% du travail du data scientist concerne le nettoyage des données, avant de pouvoir attaquer la phase plus rigolote de traitement des données. [...]

Le data mining est souvent décrit comme une activité visant à donner du sens aux données. D'après vous, les activités relatives au Big data sont-elles une « extension de ce domaine d'expertise » ? Un « changement d'échelle » ?

Pour moi, ce sont deux choses différentes. D'un côté, il y a le *Big Data* avec ses caractéristiques (Volume, Variété, Vélocité etc.) et puis de l'autre, les méthodes de *Data Mining* qui vont permettre d'analyser les données qu'elles soient petites ou qu'elles soient volumineuses d'ailleurs. Parce que parfois analyser des petits jeux de données, c'est tout aussi compliqué. Le data mining c'est un ensemble de

méthodes d'analyse et le Big Data ce sont les données complexes que l'on analyse.

Ce que recouvrent les Big Data semble mal perçu ou mal compris par des décideurs non spécialistes. Ils sont pourtant d'une manière générale très intéressés par le potentiel de connaissances pouvant être extraites des « masses de données ». Certains acteurs économiques pensent que les projets Big Data ne concernent que les grands groupes. Pourriez-vous nous apporter votre point de vue ?

Effectivement, cela concerne évidemment les grands groupes qui ont à leur disposition de grands volumes de données. Mais cela concerne tout autant les petites organisations qui sont producteurs de données.

Si on prend, par exemple, le cas d'une mairie qui va produire pour sa ville tout un tas de données sur les écoles, les transports, les aménagements... Individuellement ces données sont limitées et parfois non directement exploitables. Mais en les croisant, avec d'autres données de l'organisation ou des données issues d'autres organisations, on pourra envisager des applications originales.

Pour vous, quelle est la principale différence entre la Business Intelligence (informatique décisionnel) et le Big Data ?

Le *Big Data,* c'est les gros volumes de données complexes alors que la Business Intelligence, c'est tout ce qu'on va pouvoir mettre derrière en termes d'intelligence pour valoriser ces données à destination des décideurs et des dirigeants d'entreprise.

Le mot de la fin : comment voyez-vous le métier de data scientist dans une quinzaine d'années ?

Je pense que tous les aspects « stockage » sont déjà bien maitrisés, les aspects « traitement » sont déjà bien développés. Par contre, les aspects « intelligence » sont souvent à inventer. Donc c'est un métier qui existera, je l'espère, encore dans 15 ans ! [rires]

Probablement ! Ce qu'il recouvrira, est-ce que ce n'est pas ça la vraie question ?

C'est ça. Comment il évoluera ? Je pense qu'il évoluera au gré des technologies. Là où il faudra que nos étudiants soient forts, c'est au niveau de leurs choix : savoir s'investir dans les technologies qui vont perdurer, ce qui n'est pas forcément le plus simple. Quels sont les standards qui vont ressortir de l'ensemble des technologies qui ont récemment émergé ? Je n'ai pas la réponse. C'est confrontée aux usages qu'une technologie va faire ces preuves.

Il ne me reste plus qu'à vous remercier pour le partage de votre expertise.

---

[1] Semantic Indexing of French Biomedical Data Resources

[2] Programme de médicalisation des systèmes d'information

Confidences d'Olivier Rajzman, directeur France Docuware

■ Propos recueillis par Béatrice Tofili le 31 mars 2017

Directeur France chez DocuWare Corporation depuis novembre 2011, Olivier Rajzman nous livre sa vision du Big Data à travers la GED.

Pouvez-vous me parler de votre mission au sein de la société DocuWare ?

La mise en place voire l'intégration de la solution de gestion électronique des documents (GED) DocuWare® au sein du système d'information des entreprises clientes est ma mission. La GED a beaucoup évolué ces trois dernières années. En effet, d'un outil de stockage d'archives dans un objectif de sécurité de l'information et du partage, elle devient un véritable épicentre nerveux de l'entreprise.

Vous parlez d'évolutions de la GED ? Quelles sont-elles ?

La première évolution constatée est que la GED permet le workflow (travail collaboratif par le biais de la dématérialisation des processus de demande d'achat, de note de frais, de commande...) en plus de la gestion des documents et de l'information. Chaque acteur du processus possède ainsi un parapheur électronique qui lui liste les tâches à effectuer ou l'état d'un document. La seconde évolution est que la GED s'intègre dans le système d'information de l'entreprise. En effet, afin de coordonner et faciliter tous les processus en entreprise, elle doit interagir avec tous les Entreprise Resource Planning (ERP, en français : progiciel de gestion intégré) de la structure. Pour rappel, les ERP couvrent la comptabilité, la facturation, la gestion de stock, la RH, la gestion de la Relation Client... En somme, les ERP gèrent les data et la GED, les documents.

Concrètement, la GED n'est plus une armoire de stockage de documents mais un système vivant de flux de documents et d'informations au sein de l'entreprise devenant ainsi un véritable outil stratégique. L'information devient ainsi exploitable par des outils de reporting permettant ainsi à l'entreprise d'obtenir plusieurs indicateurs nécessaires à sa décision. La GED permet de gérer la confidentialité des informations au travers de la gestion des droits utilisateurs. Enfin, la GED possède dorénavant la faculté de gérer les archives mails ainsi que les documents sortants des SI (facture client par exemple).

Peut-on dire que la GED précède le Big Data ?

Notre approche est axée utilisateur. En effet, notre mission est de lui permettre d'accéder à l'information le plus facilement possible tant dans l'espace spatio-temporel qu'à partir de l'interface d'accès à la base de données. La GED et notamment DocuWare® permet un accès facile de l'information et notamment d'une information structurée. Le terme de Big Data est une notion abstraite pour

le client et le technicien informatique. Au vu de toute l'information et les documents détenus par les entreprises, on peut dire que le Big Data se retrouve au sein de la GED.

Comment interagissent la GED et le Big Data ?

Le logiciel de GED ainsi que sa base de données sont souvent hébergés on Premise (serveur physique) chez nos clients. De plus en plus de clients optent pour un hébergement en Cloud afin d'accéder à l'information sans frontière temporelle ni spatiale.

Toutefois, une particularité subsiste pour les collectivités de France. En effet, conformément à la réglementation, les data (les données) produites par ces dernières ne peuvent pas sortir du territoire français. Dans ce cas précis et afin de se conformer à la réglementation, la société DocuWare propose des « cloud souverain » c'est-à-dire des serveurs installés sur le territoire national hébergeant lesdites données et l'ensemble du traitement effectué sur ces données. Une clause est prévue dans ses contrats afin de prévoir la destruction desdites données après un délai de soixante jours.

Comment est assurée la protection de l'information au sein de la GED ?

Concernant la protection et la diffusion de l'information, il est rappelé que plusieurs strates ou briques de sécurité sont mises en place tant au niveau du workflow, du logiciel, de la base des données mais également dans le temps avec la gestion du cycle de vie d'un document. En effet, selon les informations et documents gérés par le logiciel et donc conservés dans la base de données, cette dernière doit respecter la réglementation en vigueur (par exemple, déclaration à la CNIL pour la conservation de données personnels, les différents règlements techniques...). Toutefois, il est rare que le logiciel de GED diffuse une information sensible. Cette faille est souvent décelée au niveau du personnel de l'entreprise.

En conclusion que direz-vous de la relation GED et Big Data ?

« Le Big Data dans le papier n'est pas simple ! »

La tendance actuelle est que les entreprises vont petit à petit migrer vers du zéro papier.
D'ailleurs, un arrêté adopté le 22 mars 2017 fixant les modalités de numérisation des factures papier en application de l'article L. 102 B du livre des procédures fiscales et publié au JORF d'hier (le 30 mars 2017) permet aux contribuables qui le souhaitent de désormais numériser leurs factures papier dès l'envoi ou la réception de ces dernières et peuvent ainsi les conserver sous forme dématérialisée jusqu'à la fin de la période de conservation fiscale (soit six ans). Cette mesure a pour objet de favoriser la dynamique de dématérialisation des processus. Elle permet en outre aux assujettis à la taxe sur la valeur ajoutée (c'est-à-dire les entreprises) de réaliser des gains de productivité en ayant recours à un archivage dématérialisé moins onéreux qu'un archivage de document papier. De plus, il est également constaté la normalisation d'un format pour les factures (Cas de l'EDI avec le format XML et la base de donnée *Zukferd* qui existe déjà en Allemagne. Ainsi à travers le portail CHORUS PRO qui permet aux entreprises françaises d'envoyer leur facture normalisée à l'administration à compter du 1er janvier 2018).

Ainsi, l'échange et le partage d'informations qui existaient à l'intérieur de l'entreprise seront également constatés à l'extérieur de celle-ci. La lettre recommandée électronique, le courrier électronique, le contrat de travail ou encore le bon de commande seront ainsi échangés de manière sécurisée à l'aide d'une plateforme (par le système XMS).

Toutefois, il faut bien identifier les points de rupture numérique afin de rendre complète la chaine du numérique. Enfin, dans cet ensemble, les datas vont être de plus en plus nombreuses et il serait essentiel que le Législateur ne tarde pas à prendre les mesures nécessaires à la confidentialité et la sécurité des données de nouveaux genres.

Fondée en 1988, la société DocuWare est spécialisée dans les logiciels de GED.

Elle possède une présence internationale considérable (en Europe et aux Etats Unis).

« *La solution DocuWare apporte une valeur ajoutée considérable au fond[s] documentaire des entreprises quel[le] que soit leur taille ou leur secteur d'activité. Peu importe le format ou la source, les processus métiers et les [workflows sont automatisés] au sein de la GED DocuWare. Les documents liés à un dossier sont accessibles à tout moment en un clic de souris* ».
(Source : docuware.com)

Protection des données à caractère personnel dans le domaine juridique : « le risque de ré-identification est réel »

■ Propos recueillis par Adèle Duret

Perspectives historiques de la protection des données à caractère personnel dans le domaine juridique. Un sujet d'actualité brûlante dont il est pourtant nécessaire d'appréhender l'histoire afin de bien en cerner les enjeux.

Bonjour. Merci de nous accorder cet entretien. Pouvez-vous nous dire en quelques mots quel est votre parcours universitaire et professionnel ?

Juriste de formation « Sciences Po », j'ai un DESS d'informatique documentaire réalisé avant Internet en 1993.

Étant le seul juriste de la promotion, j'ai alors pris part à une aventure très surprenante !

Suite à cela, j'ai été recruté en tant que gestionnaire de système d'information au Conseil constitutionnel. J'ai dû tout créer (toutes les bases de données par exemple) dans un contexte particulier. En effet, il s'agissait d'une époque pendant laquelle un monopole très strict de diffusion de l'information existait ce qui rendait le travail bien plus complexe.

On entend parler de l'open data depuis quelques années. Mais quelle est la genèse, la tradition française en matière d'accès et de diffusion à l'information juridique ?

Quelques repères historiques :

Discours d'Hourtin du 25 août 1997 : Préconisations de Lionel Jospin à propos de la diffusion gratuite des données publiques essentielles. Il s'agit en réalité de l'acte de création de Légifrance. La plupart des parties prenantes étant contre (éditeurs juridiques, avocats...), de nombreuses difficultés sont donc apparues. Un recours a même été déposé par l'Ordre des avocats à la cour d'Appel de Paris.

2002 : Décret n° 2002-1064 du 7 août 2002 relatif au service public de la diffusion du droit par l'internet

Directive PSI 2003 (directive 2003/98/CE du Parlement européen et du Conseil du 17 novembre 2003 sur la réutilisation des informations du secteur public): Possibilité d'imposer la diffusion gratuite du droit, les notions de « Favoriser l'accès au droit et sa réutilisation » commençant ainsi à prendre forme.
Tout ce processus appartenait à un ensemble cohérent largement aidé par certains juges. C'est le cas notamment de ceux du Conseil Constitutionnel auxquels on doit la création de l'objectif de valeur constitutionnelle d'accessibilité et d'intelligibilité de la loi (Décision n° 99-421 DC du 16 décembre 1999).

Ainsi, le Conseil Constitutionnel a porté le droit d'accès au même niveau qu'un droit de l'Homme, la France devenant le premier pays au monde à le concevoir de cette manière.

En pratique, dès le 12 avril 2000 la loi relative aux droits des citoyens dans leurs relations avec les administrations et plus précisément dans son article 2 impose la gratuité de l'accès du droit.

Il s'agit d'une époque où personne ne parle encore de big data et d'open data. Pourtant cette chronologie oubliée est importante car elle constitue la base du processus permettant « la mise à disposition et surtout l'encadrement de la réutilisation du droit ».

Quelle est la tradition française de la diffusion du droit ?

En France, une structure forte existe en matière de diffusion juridique. Les bases de données juridiques françaises sont les plus anciennes du monde, notre pays bénéficiant par conséquent de fondations solides en matière de données juridiques. Ainsi, la base de référence du Journal Officiel a été créée en 1936, et informatisée en 1978 (bases LEX et LEGI).
Déjà à l'époque, l'hypertexte était inventé pour les textes appliquant un autre texte ou en citant un autre puisqu'il était possible de passer d'un texte de loi à l'autre grâce à une ontologie. Tout cela prenait place dans un cadre de relations définis (RDF) créé à ce moment-là, d'ailleurs toujours utilisé dans le XML actuel.

L'accompagnement juridique et institutionnel était très fort, corrélé d'une forte volonté des juges :

Dès 1997, le Conseil d'Etat a établi que « le service public des bases de données est un service public par nature » (Conseil d'Etat, 17 décembre 1997, N° 181611). Cette décision est très importante, elle signifie que ce n'est pas privatisable comme le sont les routes ou les écoles par exemple.

Aujourd'hui, la loi pour la République numérique (LOI n° 2016-1321 du 7 octobre 2016 pour une République numérique) pose la mise à disposition de l'ensemble des décisions sous réserve du respect des droits des personnes.

Cette actualité fait écho à d'autres actualités : Le décret du 7 août 2002 (Décret n° 2002-1064 du 7 août 2002 relatif au service public de la diffusion du droit par l'internet), l'ensemble de la jurisprudence des cours suprêmes essentiellement ainsi qu'une sélection des jurisprudences des cours d'appel et des cours administratives d'appel ont été mises en « accès libre et gratuit ». Observons que la notion de sélection posait problème. Toutefois, tentons de percevoir qu'à l'époque il s'agissait d'une petite révolution car auparavant tout était payant et uniquement destiné à être diffusé commercialement.

Qu'en est-il de l'anonymisation des données personnelles ?

L'anonymisation a fait l'objet d'une recommandation de la CNIL :

1/ non anonymisation quand la diffusion est payante,

2/ anonymisation quand la diffusion est effectuée de manière gratuite.

Cependant, il s'agit seulement d'une recommandation. Notons que les lois en matière de protection des données personnelles restent applicables.

Comment est géré le risque de ré-identification des personnes passées en jugement ?

Jusqu'en 2010/2012, il y avait environ 150 demandes d'anonymisation par an dans Légifrance.

Toutefois, la modification du Journal Officiel n'est pas possible étant donné qu'il est signé. La concordance entre la publication papier et informatique reste en effet primordiale.

Cependant, les données ne sont plus « recherchables ». En réalité, ce qui est modifié c'est une des bases de données Légifrance qui est en fait un portail interrogeant des bases qui ne sont que des numérisations des données officielles. L'information ne disparaissant pas, nous ne pouvons donc pas parler d'anonymisation mais de « droit au déréférencement ».

Quelle est l'actualité de l'anonymisation ?

En ce moment, nous sommes préoccupés par le caractère ambitieux et très large du RGPD (règlement général de la protection des données) qui sera complètement d'application, y compris avec des amendes, dès mai 2018. Il suppose l'établissement d'études d'impacts sur les risques des données à caractère personnel, et impose une analyse de risque à toutes les bases des administrations. Analyse qui devra aussi être effectuée par Légifrance. À noter que ce qui est visible du grand public ce sont les bases de diffusion, a contrario des bases de production.

Concrètement, ces dernières sont des outils pouvant être utilisés par 3 500 personnes qui ont notamment la possibilité de « signer ». Et ce ne sont pas moins de 400 000 personnes susceptibles d'accéder à la procédure normative.

Même si ce n'est pas public, l'impact pour les données nominatives et les données personnelles reste fort. Soulignons ici la possibilité de traiter grâce à l'anonymisation ou la pseudonymisation.

Quel est le problème et risque de ré-identification ?

Ce qui pose problème avec le big data, c'est la ré-identification. Les risques sont importants. Même avec une anonymisation, les données pourront être facilement retrouvées en les croisant grâce aux techniques du big data.

Prenons l'exemple d'une affaire de pédophilie concernant un psychiatre dans une petite ville bien avant le big data. Bien qu'anonymisée, cette décision était facilement retrouvable, cette ville ne comptant en effet qu'un seul psychiatre.

Pour le moment, aucun logiciel n'est en capacité de régler ces problèmes. Les risques *pour l'administration ainsi* que les contentieux sont pourtant considérables. Malgré l'anonymisation des données personnelles, le risque de ré-identification est donc bien réel.

Pierre Delort

■ Propos recueillis par Firouze Betou

Cet article est l'objet d'un travail universitaire. Il s'agit d'une activité de collecte et de restitution d'information sur un thème de recherche : le Big Data. Au-delà de l'aspect pédagogique de l'exercice, la rencontre avec un professionnel de son domaine est également un événement. Il s'agit de confronter l'adéquation de son projet professionnel aux métiers de l'Entreprise. Aussi, la narration est-elle exploitée pour retracer cet entretien. Le récit est un genre. Il s'agit, pour l'auteur du récit, le narrateur, de reporter au lecteur ce qu'il a ressenti et compris. Ce document est une source de témoignage, d'orientation pour d'autres étudiants.

« Dans une bonne histoire il y a trois temps.
Il y a d'abord, celui de l'ouverture décrivant le contexte,
le décor, les circonstances et l'arrivée des personnages... »

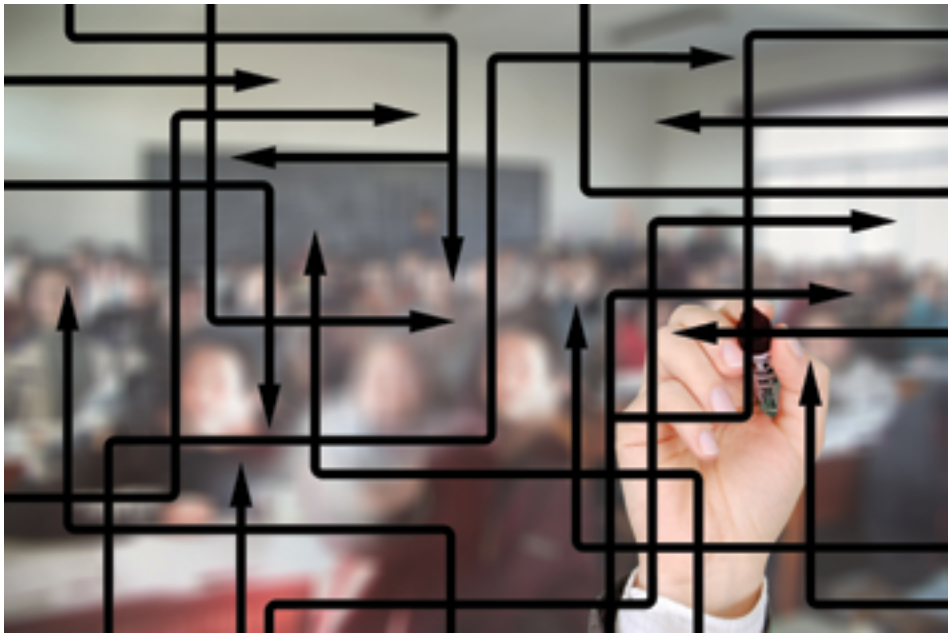
Étudiante de Master 2, parcours Management et valorisation de l'information numérique (Mavinum-Université Paul-Valéry, Montpellier 3), mon projet professionnel s'inscrit dans une démarche prospective.

La question que je me pose est « Quelles sont les incertitudes pour le système d'information (SI)1 dans sa mission d'outil d'aide à la décision face à la masse exponentielle d'information utile ? »

Pierre Delort est président de l'Association Nationale des Directeurs des Systèmes d'Information (ANDSI). Il est membre depuis 12 ans. Auteur d'une thèse doctorale « Information contre institution : workflow, Lean Management et gouvernance de la firme[2] » et d'un ouvrage « Le Big Data[3] », il crée le cours sur ce sujet à l'école Mines (Mines ParisTech) en 2011. Directeur des Systèmes d'Information il dispose d'une expérience multisectorielle notamment de diverses stratégies organisationnelles. Dans ses travaux académiques Pierre Delort a étudié la logique de la transformation numérique des Entreprises selon le concept du Lean management[4].

« Quand le téléphone a sonné, j'ai attrapé le combiné. L'entretien avec Pierre Delort a commencé à cet instant, le lundi 27 février 2017 aux alentours de 10 h 30 ».

Comme toute notion à la mode, le sujet « Big Data » devient difficile à cerner. Le risque est que le lien ne se fasse pas avec la stratégie informatique, au sens large, de l'organisation. L'angle d'approche de cet entretien est la sécurité du système d'information (SI) de la firme. La spécificité du terrain est celui de la prise de décision stratégique. Il convient à ce stade de distinguer les technologies du Web et les technologies de l'information et de la communication (TIC).



« La traduction de "Big Data" est "les mégas données" […]. On ne pouvait pas prendre ce titre, car le préfixe méga est associé à "million", comme dans megaoctet. Nous avons choisi "Big Data" pour éviter cette confusion et l'article défini "le" car pour la langue (français) du livre […] »

Pierre Delort, auteur de
« Le Big Data : Que-sais-je ? »

La première question que je pose sur le thème du Big Data n'est pas claire : *« La traduction est la langue du Web. Qu'en pensez-vous ? »* D'après mes lectures et rencontres, la langue d'Internet est l'anglais. Selon Walter Benjamin dans « *L'oeuvre d'art à l'époque de sa reproductibilité technique*[5] », la technique fait perdre du sens au contenu.

Ainsi, la traduction robotisée est une médiation. Le terme « Big Data » appartient donc à différents répertoires culturels et sa traduction est une médiation du social. Selon, le grand dictionnaire terminologique, le terme « *méga données* » est associé au domaine de l'informatique. Ainsi, **P. Delort** distingue le « *computer science* » et l'Informatique au sens large à savoir, les pratiques, les traitements et les utilisations de l'ordinateur. Ici l'Informatique est à interpréter comme les Technologies de l'information « *Ensemble des matériels, logiciels et services utilisés pour la collecte, le traitement et la transmission de l'information* ».

Le Big Data est un concept. Sa définition n'est pas la même que l'on se situe dans le champ de l'intelligence artificielle, de l'informatique ou de la gestion. **Pierre Delort** propose dans sa définition une convergence entre les domaines.

Le Big Data est le résultat d'un volume considérable de données produites par des activités diverses, comme écouter de la musique via le Net, envoyer un mail à un collègue de travail depuis son bureau, etc... Ces traces numériques sont stockées pour créer, par induction, des modèles et des algorithmes. Il s'agit pour l'Entreprise de sa capacité à collecter, stocker et sécuriser une masse exponentielle de données. On note l'aspect stratégique sous-jacent à la gouvernance informationnelle de cette approche. L'Entreprise a désormais pris conscience de la création de valeur du Big Data, en terme de nouvelles connaissances pour la prise de décision stratégique. Aussi, le SI et ses acteurs sont-ils au coeur de cette transformation digitale.

« Les entreprises se situent loin des réflexions sur le Web sémantique et des travaux du W3C »

P. Delort, Président de l'ANDSI

Les nouvelles technologies du Web ne sont pas un facteur clé de la stratégie informatique d'une firme. Les standards du W3C, les normes AFNOR sont exploitées par les DSI et l'usage de l'ISO 270006 pour la sécurité des SI est connue, mais dans ce cas précis, les techniques de l'ingénieur sont privilégiées. Ainsi, le test de pénétration, « *pentest* » en anglais, permet d'évaluer la sécurité d'un SI selon les attaques d'un « testeur ». Il s'agit d'en découvrir les failles pour mettre en oeuvre des procédures de sécurité sur mesure.

« Les algorithmes et les données sont issus d'une logique du passé […]. La prise de décision se base sur les données du passé et la vision du dirigeant… ».

P. Delort, Président de l'ANDSI

La business intelligence emprunte une nouvelle logique d'informatisation dans une dynamique 2.0. Le système décisionnel s'imbrique d'outils, de technologies et d'algorithmes prédictifs favorisés par un environnement numérique. Ainsi, la prise de décision se base sur des probabilités et non des certitudes. Les risques de pilotage augmentent puisque l'information utile est le produit d'un calcul mathématique. Le Big Data n'est pas un nouveau mode de pensée pour l'Entreprise. Néanmoins, « *les données dans le big* » modifient le paradigme décisionnel. Ainsi, l'informatisation du SI est un facteur partiel de risque pour l'Entreprise.

« Les technologies de l'information et de la communication sont invisibles, une fois le programme codé il est difficile de comprendre ce qui se passe […] »

P. Delort, Président de l'ANDSI

Pierre Delort cite trois facteurs d'influence sur la sécurité et l'intégrité du SI à l'ère du Big Data : l'humain, les technologies et le management. Les risques liés aux innovations technologiques sont internes. Cette affirmation met en garde contre le fétichisme technologique.

L'évolution du SI doit intégrer une gestion du risque dynamique et dans le temps. Dans les sciences humaines et sociales, connaître le passé permet de comprendre le présent. À l'instar, le SI de l'Entreprise ne se construit pas dans un temps unique. Les vagues successives d'évolutions technologiques bousculent logiquement l'organisation.

En 1956, l'équivalent du terme ordinateur est l'expression anglaise data processing machine. La notion de SI apparaît en 1960, avec l'implantation de l'ordinateur dans le monde de l'entreprise : la transformation de la firme est en marche. Les années 70 sont marquées par le refus de l'automatisation du SI et la naissance des Sciences de l'information et de la communication : les SIC. Néanmoins, la vague d'implantation de logiciels informatiques pour répondre aux besoins informationnels dans les Entreprises court jusque dans les années 1990. À cette époque, « La mode » du Workflow, littéralement « flux de travaux » et de la gestion par processus s'installe. Néanmoins, cette distribution du travail limite l'action des individus.

L'approche systémique s'oppose à cette vision cybernétique de l'information. Ainsi, le Big Data est à considérer dans un contexte d'évolution du SI et d'innovations des TIC. L'Entreprise s'adapte à des transformations techniques et sociales. L'ère de la data seconde l'ère de la puce et l'automatisation du SI pose la question de la dictature de la donnée.

« Les algorithmes peuvent-il régler la vie de l'Entreprise ? […] oui mais partiellement »

P. Delort, 27 février 2017-11h40.

Le progrès, les révolutions techniques permettent une multiplicité de relations entre les individus et des systèmes. Aujourd'hui, l'Entreprise observe son environnement au travers des données. Néanmoins, le Big Data repose sur un jeu de corrélations, qui ne représentent pas exactement la réalité. Ainsi, la valeur des données réside uniquement dans l'équilibre de l'intelligence de leur manipulation.

---

[1] Selon Robert Reix, le système 'information est « Ensemble organisé de ressources : matériel, logiciel, personnel, données, procédures...permettant d'acquérir, de traiter, de stocker des informations (sous forme de donnée, textes, images, sons, etc.) dans et entre des organisations ». Systèmes d'Information et Management des Organisations, Paris, Ed.Vuibert, 1995.

[2] Delort, P. Information contre institution : workflow, lean management et gouvernance de la firme. Thèse de doctorat - Paris : L'École nationale supérieure des mines de Paris.

[3] Delort, P. (2015). Le big data. Paris : PUF.

[4] Le Lean management est un modèle d'organisation du travail issu de l'industrie automobile (Toyota).

[5] Benjamin, W. (2013). L'oeuvre d'art à l'époque de sa reproductibilité technique. Paris, Éd. Payot & Rivages.

[6] La norme ISO 27000 formule des recommandations des meilleures pratiques de sécurité du SI pour les entreprises.

La gestion des risques pour l'Entreprise (Source : startupstockphotos.com)

Pierre Savary de Beauregard

■ Propos recueillis par Céline Jourden

Pierre Savary de Beauregard est responsable du pôle Valorisation des données à la Banque Postale.

Vous êtes responsable du pôle Valorisation des données au sein de la Banque Postale. Les activités de votre pôle sont en lien direct avec le Big Data et la gouvernance des données. Pouvez-vous définir plus précisément votre organisation, son positionnement, son activité ?

Mon pôle est rattaché au Chief Data Officier (CDO) de la Banque Postale, c'est-à-dire le responsable de la gouvernance des données qui lui-même est rattaché à la directrice financière.

Le CDO est responsable de la direction Data Management qui contient plusieurs pôles : celui dont j'ai la responsabilité, un pôle de gouvernance des données, qui répond pleinement à cette thématique, un pôle data visualisation et un pôle pilotage commercial.

Initialement, mon pôle a été construit sur les équipes de la maîtrise d'ouvrage du système d'information décisionnel de la banque postale. Les données de ce dernier sont utilisées par l'ensemble des métiers de la banque, comme par exemple la direction marketing, la direction des risques.

Le CDO a été nommé il y a un peu moins d'un an, il a récupéré mon équipe et a décidé d'en élargir les compétences. Il a voulu conserver la maîtrise d'ouvrage du DataWarehouse qui est parfaitement maîtrisée et très utilisée. Par ailleurs, Il a voulu intégrer un système de veille qui a pour ambition de faire des POC (Preuve de Concept) afin d'évaluer ce que la banque postale peut exploiter, développer en matière de Big Data.

La banque postale a donc un énorme pôle concentré tout autour des données, de la gouvernance de ces dernières, à la gestion de leurs évolutions. Le CDO est le chef d'orchestre de cette ambition. Il a le rôle de gestionnaire de la gouvernance des données, par exemple par le biais d'établissement de chartes. Il a aussi un rôle lié à la valorisation de la donnée à savoir définir la valeur de cette dernière afin d'estimer s'il faut l'exploiter.

Pouvez-vous décrire les données massives traitées, utilisées et également décrire leur évolution ?

Le Big Data est considéré comme une évolution de ce que la banque faisait déjà auparavant. La banque a toujours traité des données. Aujourd'hui elles sont plus importantes, plus volumineuses. La Banque Postale est jeune, elle a une dizaine d'années, son DataWarehouse existe depuis le même nombre d'années. C'est un entrepôt de données qui regroupe des données structurées. Cet entrepôt possède des milliers de données, de tables.

En termes de volumétrie, il représente plusieurs dizaines de téraoctets. Au sein de la Banque Postale, nous parlons du Big Data en termes de volume. La banque dispose des données sur l'ensemble de ses clients et de ses contrats. Ces derniers représentent plusieurs dizaines de millions de clients et plusieurs centaines de millions de contrats.

D'autres données conservées sont, par exemple, la fréquentation de nos clients dans les bureaux de postes. Nous avons 17000 points de contact, cela représente un gisement de données énorme que nous exploitons déjà mais que nous voulons encore développer. Notre ambition est également d'étendre l'activité sur le web car ces données sont stockées mais pas suffisamment exploitées.

Aujourd'hui les données structurées sont parfaitement maîtrisées. L'objectif

serait de s'intéresser davantage aux données non structurées en interne comme par exemple les appels et les emails que font les clients aux conseillers, mais aussi celles qui sont externes (les réseaux sociaux). La donnée non-structurée n'est pas encore exploitée. Le CDO a donc mis en place une cellule de veille pour parvenir à mieux appréhender ces données. Nous sommes en mode exploratoire. Par exemple, mon service travaille sur l'ensemble des relevés bancaires des clients, mais nous ne traitons pas le libellé des opérations de ces relevés. Nous travaillons sur le montant, la date et les codes mais nous ne cherchons pas à savoir ce que le client a acheté.

Cela s'explique par le fait que quand le DataWarehouse a été construit, nous ne savions pas exploiter ces données et le volume était un problème à l'époque. Aujourd'hui, nous ne sommes plus limités par le stockage donc nous aimerions exploiter ces libellés car nous pensons qu'il y a un véritable potentiel d'exploitation d'un point de vue commercial, marketing ... Nous souhaitons le faire en respectant la réglementation, en respectant les codes éthiques. La banque postale a de vraies valeurs éthiques.

L'évolution des données massives est réelle. Pour gérer, exploiter et industrialiser les données relatives au Big Data et notamment les données non structurées, le groupe La Poste est doté d'une plateforme Big Data que la banque utilise pour déposer des données et faire des POC, pour élaborer des algorithmes afin de mesurer le taux de churn, d'attrition c'est-à-dire établir des analyses pour anticiper par exemple le départ de nos clients. La plateforme est basée sur le framework hadoop.

Sur cette plateforme, on peut faire la fouille et la préparation des données, puis la préparation et le paramétrage des algorithmes et enfin si les algorithmes donnent des résultats satisfaisants tenter la mise en production. Nous traitons environ dix projets à ce jour.

Le Big Data est donc en évolution constante au sein de la Banque Postale, comment sont gouvernées ses données ?

La Banque Postale dépend du groupe La Poste qui a édité une charte data qui indique comment exploiter la donnée. Cette charte explique que les données doivent être protégées de façon éthique. Les données doivent être respectées et traitées avec la plus grande vigilance. La Poste et la Banque Postale ont donc une volonté forte de gouverner les données.

Aujourd'hui, nous écrivons une charte spécifique à la Banque Postale qui bien évidemment ne contredit pas la charte de La Poste mais qui va la compléter et la développer notamment sur les réglementations qui sont spécifiques au monde bancaire. Cette charte sera validée par le Comex de la banque, elle est en cours de rédaction. Elle décrit ce que nous devons faire et ce que nous voulons faire en termes de maîtrise des données et de qualité de la donnée.

Mettez-vous en place des procédures particulières ? Quels sont les points essentiels à la mise en place de la gouvernance des données ?

La gouvernance des données se fait dans un premier temps en référençant les données. Il faut faire une cartographie, donner une définition métier et mettre en place un contrôle qualité de la donnée. Il s'agit de mettre en place un process pour s'assurer, mesurer et augmenter la qualité des données.

La gouvernance des données ne peut se faire que si les responsables de la donnée sont clairement définis. Les propriétaires des données, que l'on a nommé DataManager, ont été nommés dans chacune des directions de la banque. Leur premier rôle est de déterminer le périmètre de leurs données. Cela peut nécessiter des discussions avec le CDO.

Le but est d'établir le paradigme qu'une donnée a forcément un propriétaire. Ils ont pour mission de donner une définition métier et s'assurer de la maîtrise de ces données. Cela passe par des plans d'action de mise en qualité des données pour s'assurer de la justesse de celles-ci. Par exemple, la date de naissance du client est-elle correcte ? Cette date est saisie par un conseiller, elle est ensuite diffusée dans tout le Système d'information (SI), il faut s'assurer que la saisie et la diffusion soient correctes.

Pour cela dans le SI, des contrôles qualité sont mis en place par le biais de tableaux de bord qui incluent des indicateurs, des KPI (indicateurs clés de performance) qui sont calculés de manière mensuelle et qui ont pour but de vérifier qu'il n'y a pas de déperdition de qualité sur les données. Par contre, nous n'utilisons pas aujourd'hui un outil de Data Quality. Mais la mise en place d'un tel outil ou non est une des missions de CDO.

Nous produisons un reporting nommé COREP qui est propre aux banques et qui a pour objectif de déterminer des ratios de solvabilité calculés à partir de plusieurs dizaines voire centaines de données qui proviennent

de l'ensemble du SI. Dans ce cadre, nous avons dû établir un schéma global des flux de données du SI et donc de positionner les propriétaires des données, savoir s'il y en a bien à chaque flux, noeud, jointure, à chaque transformation.

Ces éléments ont été formalisés dans un dictionnaire de données. La Banque Postale est très avancée sur l'élaboration de ce dictionnaire car il est historique et référence la plus grande partie des données qui sont dans le SI décisionnel. Le dictionnaire est technique et décrit les tables et les colonnes et les définit. Le dictionnaire est plus orienté système décisionnel qu'opérationnel.

La description du système opérationnel est moins structurée et plus éclatée, non centralisée dans un même outil. L'évolution du dictionnaire veut aussi prendre en compte cette lacune. Il s'agit d'une volonté interne de mieux comprendre les données qui nous alimentent.

En plus de cette volonté interne, les banques sont dans un environnement réglementaire très contraint et fréquemment audité, pouvez-vous nous en parler ?

Nous sommes soumis à Bâle III qui a un volet sur la gouvernance et la qualité de la donnée. Nous devons aussi respecter la directive *Solvency*. II car nous sommes une banque assureur. Nous devons par ailleurs prendre en compte les principes BCBS 239 qui ont pour objectif de mettre en place une meilleure qualité des données de reportings réglementaires. Cette norme, édictée par le comité de Bâle est applicable à l'ensemble des banques européennes. La première exigence de la norme est de nommer un CDO pour mettre en place une gouvernance des données qui correspond aux exigences de la norme.

Le règlement général de l'Union européenne sur la protection des données (RGPD) doit aussi être respecté. Il se complète avec les autres règlementations, il est spécifiquement lié à la protection des personnes. Les autres normes vont plus loin car définissent l'ensemble des aspects de la donnée, la qualité.

Le respect de la réglementation sur *Bâle III* et *Solvency II* a été validé suite à un audit. Il y a eu quelques recommandations mais la Banque Postale respecte bien les obligations. Concernant RGPD et BCBS239 qui vont beaucoup plus loin, nous ne sommes pas encore au niveau requis mais c'est également le cas de l'ensemble des banques et ces normes ne sont pas encore applicables (2018). Pour faire face, nous mettons en place un programme de transformation de la banque qui est sous gouvernance du CDO. Le projet est lancé et officialisé.

Plus spécifiquement quelles sont vos obligations ?

On doit anonymiser les données. Dans le DataWarehouse, nous cachons les données nominatives (nom, prénom, mail, téléphone, etc.) mais nous devons pouvoir les rendre disponibles au corps de contrôle qui est composé de l'inspection générale et de la direction de la conformité. Cependant il faut savoir que dans le cadre de la CNIL, nous pouvons tout de même utiliser ces données pour des besoins marketing, si ce besoin est justifié et validé par le CIL (Correspondant informatique et liberté).

Par exemple, quand nous faisons des campagnes marketing, nous avons besoin du nom et du prénom de la personne donc l'autorisation est donnée mais uniquement dans le cadre d'un process avec le CIL. Dans le cadre de RGBD le process ira plus loin car nous ne pourrons faire que ce que le client nous a autorisé spécifiquement de faire. Si nous voulons faire un score marketing d'appétence au crédit consommation, nous devrons faire cocher une case mentionnant l'acceptation d'utilisation des données personnelles. Cela nécessitera un travail important de recueil d'autorisation.

La nouvelle norme impose aussi que le client puisse visualiser les données que les organismes possèdent sur lui. Il faut comprendre la réglementation car comment demander l'autorisation si nous ne pouvons pas utiliser les noms. Il faut donc véritablement s'approprier la norme pour la mise en place opérationnelle.

Pour voir si nous sommes conformes nous serons certainement audités tout comme pour les autres réglementations, l'audit nous permettra de mieux nous situer, d'avoir des recommandations d'amélioration qu'il faudra mettre en place. Nous sommes déjà en ordre de marche, un programme est lancé pour atteindre les objectifs.

Quels seraient vos mots pour conclure ?

La Banque Postale est une banque éthique qui se donne les moyens de mettre en place une véritable gouvernance des données par le biais d'initiatives internes et par le respect des réglementations. L'objectif désormais est d'établir une meilleure exploitation de la donnée non structurée. La gestion des données est un sujet sensible mais une bonne gouvernance peut limiter les dérives. Finalement le véritable risque serait de ne pas saisir l'opportunité, le risque de ne pas faire et du prendre du retard par rapport aux autres banques et aux FinTech.

L'outil BigQuery

BigQuery et Analytics : le combo gagnant !

■ Propos recueillis par Céline Jourden

Si vous faites partie d'une organisation et que la CRM (Customer Relationship Client, ou Gestion de la Relation Client en français) n'a plus de secret pour vous, vous connaissez très probablement Google Analytics. Aujourd'hui utilisé sur plus de 80% du marché mondial, cet outil permet d'analyser l'audience d'un site web ou d'une application : nombre d'utilisateurs, analyse des canaux de conversion, suivi des réseaux sociaux ou encore mots-clés, tout est passé au crible.

La solution Big Query est née en 2012 suite à la trop grande augmentation des données à traiter par les équipes Google (100 milliards de recherches par mois).

Cette solution repose sur la plateforme de Cloud de Google et de ce fait, bénéficie de sa puissance de calcul.

Couplé à Google Analytics dans sa version Premium, BigQuery vous permet d'interroger un grand nombre de lignes de données à l'aide de requêtes SQL en seulement quelques secondes. Plus besoin de télécharger les données pour réaliser vos requêtes, BigQuery se charge de tout !

Non seulement l'outil permet un véritable gain de temps, mais il possède également un espace de stockage illimité (la limite de stockage est la vôtre !), grâce à Analytics Premium vos données vous appartiennent (contrairement à Google Analytics dans sa version standard) et surtout, BigQuery est gratuit, à l'inverse de beaucoup d'autres solutions logicielles de Big Data.

Big Query, un outil multifonctions

La seule limite est votre imagination !

L'outil trouve sa place dans de multiples contextes. Vous pouvez par exemple effectuer des requêtes très précises, savoir combien de visiteurs sont venus sur votre site grâce aux réseaux sociaux sur une plage temporelle très précise, ou encore ajouter du contexte à vos données (leur attribuer une date, un auteur, un type…) afin de mieux les organiser.

Plus impressionnant encore, vous pouvez coupler vos informations hors-lignes avec vos informations en ligne. Prenons un exemple : vous possédez un site web et une boutique physique. Vous souhaitez savoir combien de visiteurs de votre site se sont ensuite rendus dans votre boutique, s'ils ont acheté un produit, lequel, combien de temps ont duré leurs visites, etc.

Cela vous semble impossible ?

Pas avec BigQuery ! Mesurez vos informations hors-lignes et centralisez-les dans Analytics, vous pouvez ensuite faire le lien entre le virtuel et le

physique et avoir une vue détaillée du parcours utilisateurs (sous réserve de se conformer à la réglementation en vigueur sur la protection des données). Impressionnant, n'est-ce pas ?

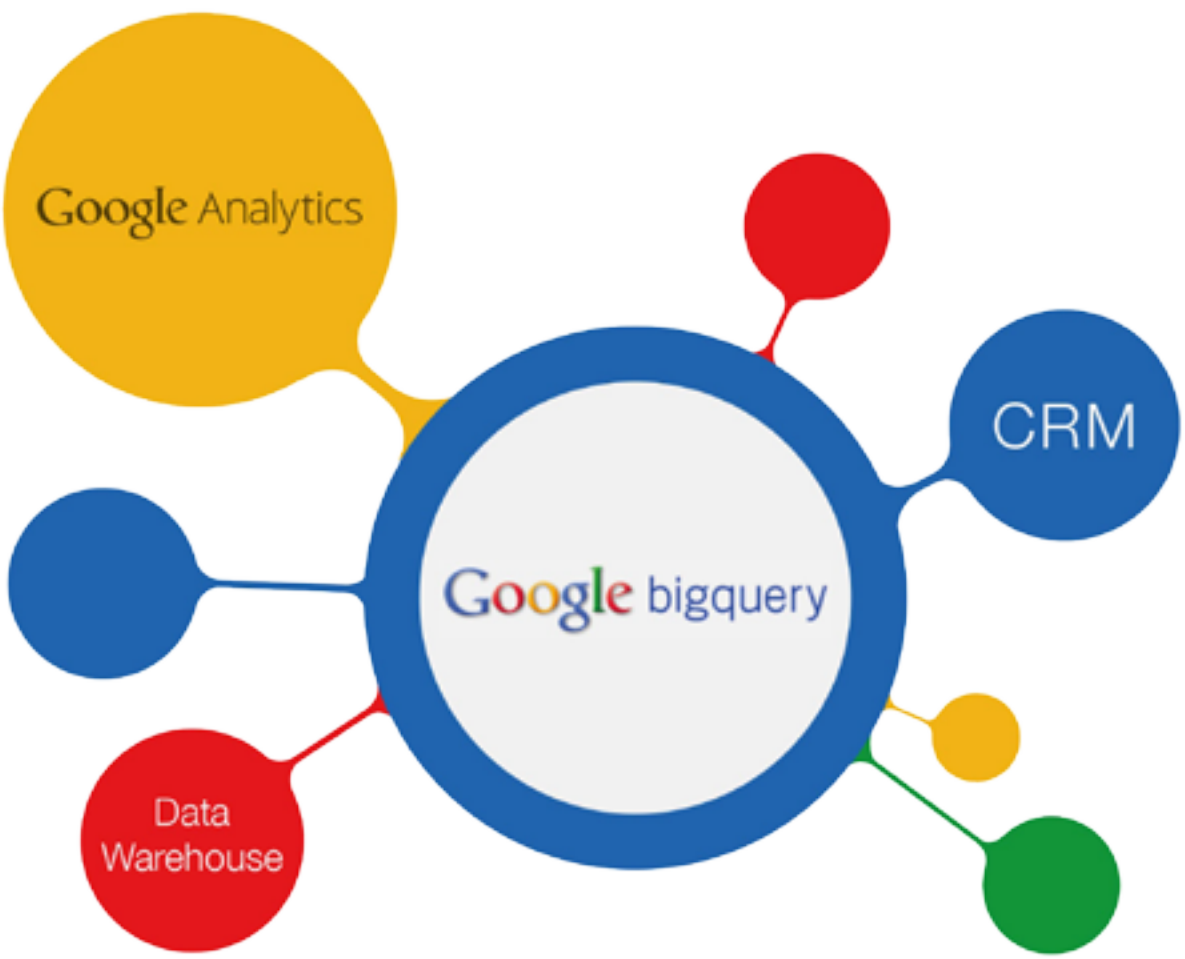
À noter que si vous souhaitez sauter le pas et adopter BigQuery, il vous faudra bien sûr posséder un compte Google Analytics Premium mais aussi vous rapprocher d'un responsable de compte afin de vous aider à configurer l'outil.

OÙ TROUVER L'OUTIL ?
https://cloud.google.com/bigquery/?hl=fr

LES POINTS POSITIFS
Pas de coût de serveur
Rapide
Propriété des données

LES POINTS NEGATIFS
Performants sur d'énormes tables (Go et To)
Nécessite un compte Google Analytics Premium
Nécessite un paramétrage entre l'outil, Analytics, le CRM et la DataWarehouse

POUR ALLER PLUS LOIN
http://www.duchess-france.org/dans-les-coulisses-de-google-bigquery/



Les Big Data à découvert

Sous la direction de Mokrane Bouzeghoub et Rémy Mosseri

L'humanité a toujours collecté et exploité des données. Après l'ère de l'informatique et Internet, nous assistons à une nouvelle révolution sans précédent. Les avis d'experts sur l'état de l'art sont condensés dans cet ouvrage permettant ainsi au citoyen d'en saisir les enjeux. Le but est d'éclairer le citoyen sur la société en pleine mutation.

Paris : CNRS Éditions, 2017

Le big data

Pierre Delort

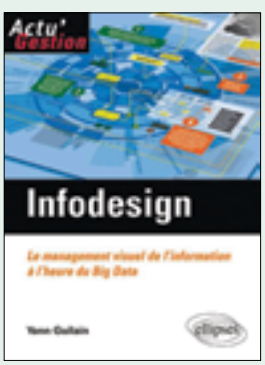
Ce livre s'adresse à des non initiés. Il apporte un premier niveau de connaissances sur le concept du Big Data. L'innovation technologique impacte l'organisation sociale et le développement de l'Entreprise.

Paris : PUF, 2015

Infodesign

Le management visuel de l'information à l'heure du Big Data

Yann Guilain



La production élevée d'informations, ces dernières années, a fait naître l'infodesign qui consiste à réaliser des représentations visuelles synthétiques pour donner à comprendre des volumes importants d'informations. Histoire, typologie, bonnes pratiques et outils y sont présentés.

Paris : Ellipses Marketing, 2013

À quoi rêvent les algorithmes

Nos vies à l'heure des big data

Dominique Cardon

L'auteur souhaite défendre à travers son ouvrage que la personnalisation des calculs est à la fois l'agent et la conséquence de l'individualisation de la société actuelle. En somme, il souhaite mettre en exergue les techniques de calculs et notamment comment les calculateurs s'immiscent dans la vie des individus.

Paris : Seuil, 2015

Big ou Bug Data ?

Manuel à l'usage des datadéontologues

Assaël Adary

L'auteur souhaite mettre en exergue l'urgence de sortir des idées convenues. Il décide ainsi de traiter globalement les Big Data à partir de la dimension éthique. Le présent ouvrage amorce une nouvelle discipline : la datadéontologie.

Paris : Éditions du Pallio, 2017

Big Data, Open Data et valorisation des données

Jean-Louis Monino, Soraya Sedkaoui

Cet ouvrage très complet analyse la révolution du big data puis se penche sur l'open data et détaille toutes les potentialités de ce nouveau mouvement. Enfin, il examine la valorisation de ces données en proposant des outils très concrets et pragmatiques pour les entreprises.

London : ISTE Éditions, 2016

GLOSSAIRE

D

Data journalisme
Le data journalisme, ou journalisme de données, consiste à collecter, analyser, traiter et exploiter des données ouvertes (open data) afin de mettre en valeur des faits intéressants, pertinents ou surprenants. Ceci dans le but de les rendre compréhensibles de tous et de les mettre à disposition du public. Il se pratique aujourd'hui grâce aux nouvelles technologies (traitement de bases de données, utilisation de logiciels spécifiques et de langages de programmation...).

Data storelling
Concept proche de la data visualization, le data storytelling consiste à illustrer des données puis, à l'aide de ces illustrations, à raconter une histoire et à construire des scénarios. Il a pour but d'orienter la prise de décision.

Data visualisation
(ou dataviz) Action de transformer des données en visualisation graphique (infographies, cartographies, graphiques, pictogrammes...) afin de les rendre plus claires et simples à comprendre.

Datafication
Fait d'utliser la technologie pour transformer les différents traces numériques de nos vies en data. Peut aussi désigner la tendance croissante à avoir recours aux data, notamment dans les entreprises.

E

Encapsulage des données (ou encapsulation ?)
Action de réunir les données brutes dans une structure avec les process qui permettent de les lire ou de les exploiter, tout en empêchant la modification de ces données brutes afin d'en garantir l'intégrité.

I

Informatique décisionnelle
(Business Intelligence) Utilisée dans les organisations, l'informatique décisionnelle désigne les techniques, outils et procédures de traitement de l'information mobilisées afin de collecter et analyser des données dans le but d'aider à la prise de décision, notamment pour les dirigeants.

IoT
Internet of Things ou l'Internet des Objets - Représente l'extension d'Internet dans des lieux ou objets de notre vie quotidienne.
Plusieurs exemples :
- les GPS dans les voitures sont connectés à une application web qui analyse notre position pour nous guider vers notre destination finale.
- les montres « connectées » utilisées en running qui enregistrent notre vitesse de course, la distance parcourue, notre rythme cardiaque et les calories dépensées par exemple dans une base de données « en ligne », ce qui permet de faire des statistiques sur notre progression, en temps réel.

L

Lac de données
(data lake) Réservoir informatique qui concentre toutes les données structurées ou non structurées d'une entreprise. Le lac de données sert à la collecte de données brutes en flux continu afin d'en permettre l'exploitation par des applications diverses.

M

Marché prédictif
Le marché prédictif est une nouvelle méthode de prévision venue des États-Unis. Dans l'objectif d'obtenir des informations fiables, il se baserait sur la consolidation d'intelligence collective issue d'un panel de prévisionnistes parieurs. Au contraire des sondages où un besoin d'échantillonnage représentatif de la société est exigé, cette méthode fait appel au volontariat des personnes. Enfin, le marché prédictif pourrait être assimilé à la transition entre les sondages d'opinion et le Big Data.

O

Open data
Littéralement « donnée ouverte », le mouvement de l'open data est devenu une thématique très en vogue. C'est une démarche selon laquelle une donnée publique est mise à disposition de tous et peut être réutilisée librement par les citoyens et les entreprises. L'ouverture et le partage des données publiques contribuent à la fois à la transparence de la vie publique et à instaurer un lien entre les institutions publiques et les citoyens. L'open data présente aussi un potentiel économique fort en stimulant l'innovation et la création de nouveaux services.

www.didaktic.fr

UNIVERSITÉ PAUL VALÉRY MONTPELLIER 3

ITIC INSTITUT DES TECHNOSCIENCES DE L'INFORMATION ET DE LA COMMUNICATION