View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Bibliothèque numérique de l'enssib

UNIVERSITÉ DE LYON

Diplôme national de master

Domaine - sciences humaines et sociales

Mention – sciences de l'information et des bibliothèques

Parcours – archives numériques

Mémoire de fin d'étude / 2017 - 2018

Archiver les Big Data : un enjeu pour l'archiviste d'aujourd'hui et de demain ?

Alexandre Vieira

Sous la direction de Laurent Duplouy
Maitre de Conférence – ENSSIB

enssib
école nationale supérieure des sciences de l'information et des bibliothèques

Remerciements

Mes remerciements s'adressent en premier lieu au personnel de l'ENSSIB et en particulier à Mr Laurent Duplouy, qui fut un directeur de mémoire attentif et prodigue en conseil. Merci également à lui de m'avoir encouragé à interroger les Big Data sous l'angle de l'archiviste, afin de mieux montrer en quoi notre métier pouvait réclamer ses « lettres de noblesse » parmi les spécialistes des sciences des données.

Je remercie en second lieu les acteurs du domaine de l'EIM : les gestionnaires de données, développeurs, architectes consultants de la société Capgemini, qui m'ont à la fois donné accès à une base de connaissance conséquentes, mais également permis de participer à des conférences et des projets déterminants lors de la rédaction de ce mémoire. Ainsi, un grand merci à Vincent Hacard, Thierry Bourges, Alphonse Vivanloc et Emmanuel Angles pour leur expérience au sein de l'EIM, et d'avoir donné de leur temps à me former en tant que consultant spécialisé dans la gestion de l'information. Je remercie également Eva Paltz, ancienne étudiante de l'école que j'ai rencontré lors de mon stage à Capgemini, pour son soutien et ses conseils de rédaction.

Enfin je remercie mes camarades de promotion et amis de l'ENSSIB. Grâce à eux, j'ai non seulement eu la chance de passer de très belles années à Lyon ; mais aussi de rencontrer des personnalités qui me sont chères.

Droits d'auteur réservés. OU (cc) BY-NC-ND

Résumé : Le numérique s'est imposé dans la plupart des secteurs d'activités. Le métier de l'archiviste ne fait pas exception. Si ce dernier à bien su prendre le fameux « tournant du numérique » (si bien qu'on parle aujourd'hui d'archivage numérique), quant est-il de son rapport au phénomène des Big Data ? Composante incontournable de notre société ou tout est de plus en plus quantifiable, le Big Data challenge l'archiviste dans la mesure ou il représente un vrac informationnel énorme à gérer et organiser, afin d'extraire des signaux d'informations présent dans ces grands ensembles de données à priori dépourvue de signification. Nous apportons ici une esquisse d'archivistique des mégadonnées : d'une part afin de proposer un modèle fonctionnel pour l'archivage des Big Data, et d'autres part pour montrer en quoi l'archiviste à tout à fait sa place parmi les acteurs plus connu des sciences/management des données.

Descripteurs : Archive numérique, Big Data, données, documents, OAIS, VITAM, Conservation sur le long terme, Records management, NoSQL, Data Management.

Abstract : Digital has emerged in most industries. The job of the archivist is not an exception. If he has engaged the famous "digital turning point", what about his relationship to the phenomenon of Big Data? Large data challenge the archivist in that fact it's a matter of control organization, in order to extract signals of information among these large sets given meaningless at the first glance. In this study, we aimed to build a sketch for the digital archiving of big data: on the one hand to provide a functional model for the archiving of Big Data, and other to show how archivist are pertinents actors to manage and bring meaning to Big Data.

Keywords : Digital Archiving, Big Data, Data, Documents, OAIS, VITAM, Long Term Preservation, Records Management, NoSQL, Data Management

Droits d'auteur réservés. OU (cc) BY-NC-ND

Droits d'auteurs

Droits d'auteur réservés.

Toute reproduction sans accord exprès de l'auteur à des fins autres que strictement personnelles est prohibée.

OU

Cette création est mise à disposition selon le Contrat : « **Paternité-Pas d'Utilisation Commerciale-Pas de Modification 4.0 France** » disponible en ligne http://creativecommons.org/licenses/by-nc-nd/4.0/deed.frou par courrier postal à Creative Commons, 171 Second Street, Suite 300, San Francisco, California 94105, USA.

Sommaire

Droits d'auteur réservés. OU (cc) BY-NC-ND

Droits d'auteur réservés. OU (cc) BY-NC-ND

Sigles et abréviations

BDD : Bases de données

BDOA : Bases de Données Orientées Agrégats (ou base de données orientées documents)

EIM/ECM : Enterprise Information Management/Enterprise Content Management

GED : Gestion électronique des documents

HDFS : Hadoop Distributed File System

JBOD : Just a Bunch of Disk

MDM : Master Data Management

NoSQL : Not only SQL

OAIS : Open Archival Information System

SAE : Système d'Archivage Electronique

SI : System d'Information

SGBDR : Système de Gestion de Base de Données Relationnelles

RAID : Redundant Array of Independent Disks

Droits d'auteur réservés. OU (cc) BY-NC-ND

Droits d'auteur réservés. OU (cc) BY-NC-ND

INTRODUCTION

Pendant la Seconde Guerre Mondiale, le gouvernement britannique, en collaboration avec les services secrets polonais, dépêcha une équipe composée de jeunes mathématiciens, statisticiens et linguistes pour résoudre un problème estimé insolvable : décrypter le code de la machine électromécanique des allemands qui chiffre leur communication, Enigma. Parmi cette équipe de brillants esprits, Alan Turing, prodige des mathématiques, et futur grand fondateur de la science informatique telle que nous la connaissons de nos jours, déploie l'étendue de ses talents pour créer une « machine universelle », capable de décrypter la machine allemande. La célèbre équipe dite de « Bletchley Park » doit alors faire face à plusieurs grandes difficultés, qui constituaient déjà en réalité des problématiques que nous qualifierions aujourd'hui de « Big Data ». Voyons plus en détail en quoi. D'une part, le fonctionnement de la machine pour crypter les communications de l'armée allemande procède par combinaison de possible : chaque jour, il y a 159 000 000 000 000 000 000 de possibilités. La mécanique repose sur un changement de la position des lettres, et son paramétrage est modifié chaque jour à minuit. Une masse de données donc, extrêmement difficile à appréhender par son volume, sa densité. Autre enjeu important, celui de la vitesse. Ces 159 trillions de données doivent être traitées en un temps record, c'est-à-dire avant le re-paramétrage de la machine, ce qui laisse à l'équipe environ 18 heures, soit 20 000 000 d'années de possibilités, qu'il faudrait pouvoir traiter en 20 minutes pour que ce décryptage soit viable. Enfin dernière difficulté, et pas des moindres, celle inhérente à la variété de ces données. Les messages encryptés reposaient en effet sur un principe combinatoire. Les Alliés interceptaient aisément les messages, mais ne pouvaient pas les comprendre : les textes faisaient référence à de la propagande nazi, la météo du jour, des slogans comme « Heil Hitler », etc … . Soit des ensembles de messages qui n'avaient à priori aucun sens. Cependant, cette variété dans les messages était construite selon un modèle, une structure. Il est bien connu que les chercheurs ont apporté une solution à Enigma en repérant ce *pattern*, par-delà la diversité, afin de rendre le modèle de fait, prédictible. Mais cette prouesse intellectuelle n'aurait jamais eu lieu sans une prouesse technologique.

Droits d'auteur réservés. OU (cc) BY-NC-ND

Le héros du film *Imitation Game*[1], Alan Turing, est un précurseur des ordinateurs et de l'intelligence artificielle. La scène qui montre les dissensions entre l'équipe de Bletchley et de Turing est celle d'un conflit de méthode : Hugh Alexander parvient à décrypter un certain nombre de messages des allemands en analysant la fréquence de distribution des lettres. Ce à quoi Turing répond ; « *Oh, even a broken clock is right twice a day* »[2]. Cela signifie que, comme une horloge cassée qui donne l'heure juste par hasard en une journée, cette solution est aléatoire et ne saurait fonctionner sur le long terme. De plus, cela suppose que les scientifiques revoient manuellement leurs calculs et algorithmes chaque jour, car les paramètres d'Enigma changent eux aussi de façon journalière. Turing pense alors à construire une machine qui permette de décoder ces messages au jour le jour. C'est une solution sur le long terme, qui repose sur un automatisme, et un axiome : *« Only a machine can defeat another machine* ». A titre de comparaison de nos jours, nous pouvons aussi affirmer que seul une machine, ou plutôt un ensemble de moyen technologique (impliquant parfois l'intelligence artificielle, *le deep learning, cloud computing,* etc…) peuvent affronter le Big Data, qui ne peut être appréhendé aisément avec les moyens traditionnels (SGBDR classique, par exemple). Cependant, une fois Enigma percée à jour, il incombait encore aux scientifiques le rôle de faire parler les données récoltées et comment orienter la stratégie de l'armée britannique par ces mêmes données.

Les données récoltées par l'équipe de Turing et sa machine permettent de prédire les prochaines attaques de U-Boat sur les convois britanniques. Une scène du film du réalisateur norvégien montre les scientifiques face à une grande problématique Big Data : ils réalisent en effet qu'ils peuvent empêcher le sabotage d'un navire dans les minutes qui suivent, et donc sauver des vies, mais en prenant le risque de révéler aux allemands que leur machine a bien été « crackée ». Cela reflète une problématique que peuvent rencontrer des data analystes ou des professionnels du marketing des données aujourd'hui : quelle est la décision à prendre en fonction de l'interprétation des données ; ou plus globalement, quelle stratégie adopter à

[1] TYLDUM, Morten, HODGES, Andrew, MOORE, Graham, DESPLAT, Alexandre, CUMBERBATCH, Benedict, KNIGHTLEY, Keira et GOODE, Matthew, 2015. Imitation game [en ligne]. Universal Studio Canal vidéo [distrib.], 2015. [Consulté le 25 août 2018]. Disponible à l'adresse : https://bibliotheques.paris.fr/Default/doc/SYRACUSE/1012218/imitation-game

[2] *Ibidem*

Droits d'auteur réservés. OU (cc) BY-NC-ND

partir de l'analyse des données ? Turing et son équipe décident de ne pas empêcher l'attaque, car cela serait agir sous le coup de l'émotion au détriment de toute logique et vue sur le long terme. Sauver des vies certes, mais aussi laisser comprendre aux allemands qu'Enigma a été résolue, et donc leur permettre de revoir leur système, et conséquemment, perdre la capacité de prédiction que donne la machine de Turing. Dans une scène du film, le mathématicien énonce alors sa méthode au chef du MI6 :

> *The minimal number of actions it would take for us to win the war … but the maximum number we can take before the Germans get suspicious* »[3].

Cela signifie que l'équipe de *Bletchley* n'exploite pas les données de façon hasardeuse : elle déploie des outils d'analyse statistique pour permettre à l'armée britannique de manœuvrer le plus stratégiquement possible. Ainsi on peut affirmer que l'armée britannique, en quelque sorte, agissait selon une démarche dirigée par les données (*data-driven*).

Ainsi, nous pouvons à partir de ce fait historique bien connu de l'histoire des sciences et de la seconde guerre mondiale, immédiatement donner une définition sommaire du concept de Big Data. Le Big Data, c'est la convergence des mégadonnées (en terme de Volume, Variétés et Vitesse – c'est la célèbre théorie des « 3V » - illustrée dans l'exemple ci-dessus des données produites par Enigma) avec de nouveaux moyens technologiques pour les traiter (ci-dessus, la machine de Turing) ; cela afin d'améliorer les processus de décision, grâce à des modèles prédictif (un management « éclairé » par les données, en l'occurrence ci-dessus le gouvernement britannique qui aligne sa stratégie sur les analyses prédictives fournies par *Bletchley Park)*. La prédiction du monde, par la donnée, la *data*, dont le flux perpétuel et inexorable trouve son intérêt économique et stratégique dans l'instantanéité de son traitement et de son exploitation. Le temps réel et court est donc, à priori, directement privilégié au temps long, celui de l'Histoire et du « passé ». L'instantanéité est ici mère d'efficacité et de rentabilité. On peut de fait se demander que vient faire l'archiviste, celui que beaucoup encore imagine dans le fond d'une cave sombre, à emmagasiner jalousement papiers anciens et manuscrits dans le seul but de thésauriser les documents - que vient faire l'archiviste, dans les faits, sur le territoire des acteurs célébrés de la data sciences : *data scientist* ou autre

[3] *Ibidem*

Droits d'auteur réservés. OU (cc) BY-NC-ND

data analyste ? Davantage, l'archiviste a-t-il encore un rôle à jouer, une plus value à ajouter à l'ère du numérique et des grands et complexes volumes de données ? Si les archivistes ou les professionnelles de la documentation ne sont certainement jamais appelés à devenir informaticien ou statisticien, quel rôle peuvent-ils jouer cependant sur l'immense chantier des Big Data ; et ou peuvent-ils se situer dans ce type de chantier : en amont, en aval, au centre, et de quelle manière ? Premier constat : les données du Big data, par leur volume (mégadonnées) et leur hétérogénéité (Variété), demandent de déployer un projet d'archivage singulier, dans la mesure où il s'agit de prendre en compte un environnement technique et informationnel nouveau et complexe. L'accroissement constant des données de type non structurées – et le *document* est compris dans cette typologie – est un phénomène bien existant, auquel sont confrontés les spécialistes de l'EIM/ECM[4], pour qui aujourd'hui les solutions GED classiques n'apparaissent plus toujours comme une solution viable pour assurer un management de l'information qui non seulement à affaire à la problématique de la volumétrie, mais aussi de la mise en valeur de cette information , ou pour être plus précis – de comment tirer de l'information des données non structurés qui submergent aujourd'hui les entreprises, et dont précisément la valeur reste difficilement extractible avec la puissance analytique des outils de GED classiques. En tant que spécialiste de l'information, l'archiviste fait lui aussi face à ce problème de gestion des volumétries des données archivées, et de comment dégager de la valeur (ou de l'information, mais pas seulement) de ces données/documents. A cette problématique s'ajoute, du côté de l'archiviste, la question clé de la conservation. Question qui nous semble d'autant plus urgente qu'elle constitue un impensé des spécialistes actuels du domaine : ingénieurs, statisticiens, managers ou marketeurs, par leur formation universitaire ou leur parcours professionnel, sont aujourd'hui essentiellement intéressés par les enjeux techniques et économiques immédiats du Big Data. Pourtant, si il est indéniable que la problématique des données constitutives de l'environnement Big Data relève d'enjeux sociaux, stratégiques et économiques réels, il n'en est pas moins vrai que les moyens pour atteindre une politique d'optimisation par les données massives doit prendre en compte l'aspect de la conservation – aspect, qui n'est pas réductible à notre sens aux simples enjeux patrimoniaux et mémoriels,

[4] Trigramme correspondant à Enterprise Information Management/Enterprise Content Mangement

Droits d'auteur réservés. OU (cc) BY-NC-ND

mais aussi rejoint les objectifs économiques et financiers des organisations confrontées à la gestion des mégadonnées. C'est un fait, les acteurs de la *data* semblent peu s'inquiéter de la conservation de leur « or numérique » : les problématiques de pérennisation et de durée légale de conservation n'est pas directement l'affaire des analystes et des développeurs. Dans un même temps néanmoins, il s'avère également que la communauté des archivistes, sauf de rares exceptions au moment où ces lignes sont écrites, n'est pas davantage présente sur le sujet.

Ainsi parmi les archivistes, les plus pessimistes voient dans les Big Data une sorte de crépuscule des archives ; de la notion même d'archives. En effet, les méthodes analytiques appliquées sur les grands ensembles de données, structurées comme non structurées – travaillent en permanence sur ces données, qui ne sont finalement plus archivées :

> *Big data analytics, which applies a set of technologies to examining ongoing trends of multiple and otherwise unrelated data sets, sees no data as archival. Instead, all data is active and has value in day-to-day business decision making or problem resolution. Archives have no real meaning in a framework like this*[5].

Le besoin de traiter la donnée en permanence et dans une forme d'instantanéité (évoquée plus haut) sonne le glas de la notion d'archive. Il n'y a pas de donnée à archiver, puisque la donnée doit toujours être active et soumise à traitement. D'autres néanmoins, se montrent plus nuancés, mais encore très prudent sur le rapport archive et Big Data, ou plus précisément, du rôle que peut jouer l'archiviste au sein de problématique Big Data :

> Il faut selon moi rester modeste quant à notre rôle dans le big data, qui est aujourd'hui loin d'être le nouveau challenge des documentalistes ou des gestionnaires de l'information pour l'emploi[6]

avec cependant l'idée que le professionnel de l'information peut se situer en amont et en aval des projets big data, dans une perspective gestion de la donnée (identification,

[5] TOIGO, Jon, [sans date]. How is big data changing data archiving strategies? SearchStorage [en ligne]. [Consulté le 26 décembre 2017]. Disponible à l'adresse : http://searchstorage.techtarget.com/answer/How-is-big-data-changing-data-archiving-strategies

[6] JOST, Clemence, [sans date]. Ghislaine Chartron : « Je ne transformerai pas mes étudiants en data scientists ». Archimag [en ligne]. [Consulté le 26 décembre 2017]. Disponible à l'adresse : http://www.archimag.com/veille-documentation/2015/11/26/ghislaine-chartron-transformer-etudiants-data-scientists

Droits d'auteur réservés. OU (cc) BY-NC-ND

qualification, classification et gestion des plans) ou encore de *dataviz.* Enfin, d'autres travaux et réflexion des professionnels de l'information essayent d'investir plus pleinement le champ Big Data, qui pour eux relève d'une problématique inévitable et de la plus haute importance pour notre métier :

> Comment se positionner par rapport à l'exploitation des Big Data, qui sont des points problématiques pour l'avenir de la profession ? […] Quelle place pourraient ou devraient trouver tenir les archivistes dans la valorisation de ces nouvelles archives ?[7].

Il nous semble qu'il est important de combiner ces points de vues pour montrer que le Big data est effectivement une problématique qui doit entrer dans le champ de l'archivistique et « challenger » le métier de l'archiviste et de ses approches : déployer une politique d'archivage dans un environnement technologique destiner à gérer de forte volumétrie, et de dégager ensuite de l'information de cette volumétrie par des procédés analytiques nouveaux (rendu possible par les base de données NoSQL), tout en gardant en vue la question de la conservation de la donnée archivée, mais aussi de comment organiser l'archive, en fonction du public cible (question de l'accès aux big data archivés – ou en terme OAIS, de la question de l'intelligibilité des données par une communauté cible). Ainsi par exemple, Claude Minotto affirme que les documents, dès à présent mais encore davantage à l'avenir, sont confrontés à « quelques grands V »[8] , notamment avec l'important enjeu des données non-structurées (un document, par définition, est une donnée non structurée ou semi structurée)[9] :

> Une publication de l'APROGED en 2013 indique que, dans les organisations, 80 % de l'information est disponible sous forme non-structurée, c'est à dire des outils usuels de bureaucratique tels que les texteurs de la suite Microsoft Office ou autre […] c'est précisément cet ensemble que constitue les documents non-structurés qui souffre d'une gestion déficiente sinon chaotique

[7] SERVAIS, Paul et MIRGUET, Françoise, Archivistes de 2030: réflexions prospectives. Louvain-la-Neuve : Academia-l'Harmattan. Publications des Archives de l'Université catholique de Louvain, 32. ISBN 978-2-8061-0214-0. 020, page 143

[8] SERVAIS, Paul et MIRGUET, Françoise, Archivistes de 2030: réflexions prospectives. Louvain-la-Neuve : Academia-l'Harmattan. Publications des Archives de l'Université catholique de Louvain, 32. ISBN 978-2-8061-0214-0. 020, page 349

[9] CHABIN, Marie-Anne, 2018. Des documents d'archives aux traces numériques: identifier et conserver ce qui engage l'entreprise la méthode Arcateg. Bois-Guillaume (Seine-Maritime) : Klog éditions. ISBN 979-10-92272-26-0. 658.403 8, page 11

Droits d'auteur réservés. OU (cc) BY-NC-ND

ce qui entraîne des pertes de données et d'informations pour les organisations, pouvant affecter son business ou sa stratégie[10]. L'archiviste du 21ème siècle doit affronter le défi d'archiver l'énorme masse des données non structurées : échanges par mail, document texte éparpillés, fichiers joints, etc … . Minotto rappelle à ce titre que « 60 % des décisions dans les organisations se prennent par courriels ou fichiers joints »[11]. Un article récent écrit par Edward Hladky[12] (Président Directeur Général d'Iron Mountain France), s'inscrit dans la direction déjà dessinée par Claude Minotto, en pointant le fait que les entreprises peuvent tirer des avantages économiques de leurs archives à l'heure du Big Data, mais à condition que l'archivage réponde à une vraie stratégie, et non à un simple stockage ou sauvegarde, auxquels par facilité et abus de langage, les DSI ou autres acteurs des SI nomment « archivage ». Selon une étude menée par le cabinet IDC, seul 38% des sociétés aujourd'hui exploitent leurs archives selon un objectif décisionnel. L'étude souligne qu'une organisation favorisant une politique d'archivage de ses documents/données dégagerait environ 7,5 millions d'euros de revenu: de par les économies réalisées (par exemple en termes de stockage dans les *data lake*, mais aussi en terme de temps – dans le cas de réalisation d'une documentation métier de référence pour l'entreprise par exemple). Enfin, pour se convaincre définitivement que les mégadonnées ont investi le champ des sciences de l'information et de l'archivage, le dernier salon Documation qui se tenait à Paris ces derniers mois (édition du 20 – 22 mars 2018) avait pour thématique l'IA et le Big Data.

Notre recherche se découpera donc en deux grands axes : Il s'agira tout d'abord de définir précisément les termes du sujet, en décomposant et cernant précisément la notion complexe de Big Data, que l'on définira comme un véritable « ensemble notionnel » pour reprendre l'expression de J. Gillium. Cela sera aussi l'occasion pour nous d'évoquer les nouveaux outils essentiels qui caractérisent aujourd'hui des projets qui ont pour but de gérer les mégadonnées : nous expliciterons ce qu'apporte les bases de données non relationnelles (NoSQL) et les *framework* Hadoop et MapReduce, et comment ces solutions deviennent essentielles pour certains projets de management de l'information (Notamment dans la gestion des documents crées et gérer par les institutions financières). Nous articulerons ensuite cette définition avec ce qu'on entend par « archivage

[10] SERVAIS, Paul et MIRGUET, Françoise, Archivistes de 2030: réflexions prospectives. Louvain-la-Neuve : Academia-l'Harmattan. Publications des Archives de l'Université catholique de Louvain, 32. ISBN 978-2-8061-0214-0. 020, page 320

[11] *ibidem*

[12] HLADKY, Edward, [sans date]. A l'heure du Big Data, vos archives valent des millions. usine-digitale.fr [en ligne]. [Consulté le 26 décembre 2017]. Disponible à l'adresse : https://www.usine-digitale.fr/article/a-l-heure-du-big-data-vos-archives-valent-des-millions.N346120

Droits d'auteur réservés. OU (cc) BY-NC-ND

numérique », et surtout comment la transition numérique assumée par la profession archivistique soulève un certain nombre de défis pour les archivistes, car une transition qui consiste aussi à s'orienter davantage vers des environnements Big Data. Ainsi l'archiviste dans ce cas-là, doit être capable de composer sa partition dans un cadre technique nouveau que nous expliciterons, et de déployer modèles fonctionnels qui conviennent à l'environnement des mégadonnées. Ce sera pour nous l'occasion d'exposer ce qu'est un archivage pérenne des données Big Data, en termes de préservation et de conservation avant tout. Essayons en somme, d'esquisser une *archivistique* des mégadonnées.

Droits d'auteur réservés. OU (cc) BY-NC-ND

LES SCIENCES DE L'INFORMATION DANS LE CONTEXTE DES MEGADONNEES

LE(S) BIG DATA : MYTHE MARKETING OU REALITE INEVITABLE A L'HEURE DU TOUT NUMERIQUE ?

De prime abord, il est difficile de cerner correctement le concept de Big Data, tant il semble employé différemment selon la profession et le champ d'activité des personnes, ou même du contexte dans lequel ces personnes écrivent. Pour certain, ce terme se réduit même à une notion floue, employée parfois à tort et à travers pour décrire un phénomène que l'on a du mal à saisir, et que, par paresse intellectuelle et pour sacrifier à une « mode » issue du marketing, on qualifiera de Big Data. Le terme, et la récurrence de son apparition dans l'espace public (dans la presse et sur le Web), pourrait donc se réduire à une simple tendance marketing[13]. Cependant, au-delà de cet aspect purement « effet de mode » (qui est également à notre sens bien réel), le concept recouvre sans doute une certaine effectivité et se révèle très utile pour décrire ce que d'aucun aujourd'hui nomme les « mégadonnées ».

Le Big Data, un « ensemble notionnel » [14]

Le terme de Big Data trouve son berceau dans le monde de la recherche et de ses publications. Il est probablement employé pour la première fois en 1997 par deux chercheurs de la NASA[15], Michael Cox et David Ellsworth, qui étaient directement confrontés à la problématique de gérer et de visualiser un grand nombre de données produites par super ordinateur lors de simulation de flux d'air produit par avion. Plus récemment, en 2008 ; c'est un numéro, retentissant, de la revue *Nature* qui

[13] BUSINESS, BFM, [sans date]. Serge Abiteboul : « Le big data est avant tout un effet de mode ». BFM BUSINESS [en ligne]. [Consulté le 29 décembre 2017]. Disponible à l'adresse : http://bfmbusiness.bfmtv.com/01-business-forum/serge-abiteboul-le-big-data-est-avant-tout-un-effet-de-mode-572981.html

[14] GILLIUM, Johann, 2015. Big data et bibliothèques: traitement et analyse informatiques des collections numériques, Mémoire sous la dir. de Monique Joly, ENSSIB, page 13

[15] PERRET, Xavier et JACQUEMELLE, Guy, [sans date]. Comprendre le Big Data à travers les films de cinéma. OpenClassrooms [en ligne]. [Consulté le 26 décembre 2017]. Disponible à l'adresse : https://openclassrooms.com/courses/comprendre-le-big-data-a-travers-les-films-de-cinema

Droits d'auteur réservés. OU (cc) BY-NC-ND

ramène le sujet sur le devant de la scène dans la communauté des chercheurs[16]. Aujourd'hui, il inonde l'espace du web. Il suffit de lancer une simple recherche pour s'en convaincre : journaliste, spécialiste du web, juriste, informaticien etc., chacun apporte sa propre définition au concept, rendant sa définition toujours plus riche, mais aussi toujours plus complexe et difficile à cerner.

En fait, l'apparition de ce phénomène peut être analysée selon plusieurs variables qui relèvent avant tout de progrès techniques, déjà mises en exergues par le problème auquel était confronté les chercheurs de la NASA, ou encore pour remonter plus loin dans le temps, l'équipe d'Alan Turing. D'une part, la puissance des micro-processeurs, qui ne fait qu'augmenter (c'est la célèbre et controversée loi de Moore) : les CPU (composants qui dégage la puissance de calcul de la machine) de nos ordinateurs aujourd'hui peuvent posséder jusqu'à 16 cœurs ; les puces voient augmenter drastiquement la densité de leurs transistors (de 1,8 milliards à 2,3 milliards) ou encore la mémoire vive des processeurs, qui fonctionnent majoritairement désormais sur une architecture 64bits. Le stockage lui aussi est un facteur important dans l'émergence des Big Data : ses coûts déjà, qui sont de plus en plus faible (à titre d'exemple: 1 millions de dollars il y'a 50 ans pour 1 gigaoctet contre 1 dollar pour le même espace de stockage aujourd'hui) ; mais aussi la vitesse de transfert des données qui a été multipliée par 25 (une vitesse de transfert qui est passée de 5Mo/s à 130Mo/s) ; ou encore bien entendu la densité des disques dur (de 100Mb/pouce² à 500Gb/pouce², soit une multiplication par 5000 de l'espace de stockage). Evoquons enfin la vitesse des réseaux, qui elle aussi ne fait qu'augmenter (il y a 20 ans, la circulation des données était estimée à 9600 bits/s alors que de nos jours, elle atteint 10Gb/s). Le Big Data, en tant que phénomène caractéristique de notre monde numérique, n'a donc pu se développer que par le développement de la puissance des ordinateurs et des réseaux, combinée à des coûts de stockages de l'information toujours plus bas, couplée à un accroissement de l'espace de stockage disponible[17]. Néanmoins, nous avons identifié ici seulement ce qui constitue le « soubassement » technique qui a permis aux mégadonnées de proliférer. Définir le concept dans ses dimensions stratégiques et applicatives est une tâche beaucoup plus

[16] Big data: The next Google, 2008. Nature. Vol. 455, n° 7209, pp. 8-9. DOI 10.1038/455008a.

[17] PERRET, Xavier et JACQUEMELLE, Guy, [sans date]. Comprendre le Big Data à travers les films de cinéma. OpenClassrooms [en ligne]. [Consulté le 26 décembre 2017]. Disponible à l'adresse : https://openclassrooms.com/courses/comprendre-le-big-data-a-travers-les-films-de-cinema

Droits d'auteur réservés. OU (cc) BY-NC-ND

complexe, tant ici le terme de Big Data recouvre de significations variées. Dans cette perspective, et pour reprendre l'expression de Johann Gillium, il est bon de rappeler que le Big Data se défini d'abord comme un

> Ensemble notionnel, où […] le terme de mégadonnées peut être à la fois appliqué à un objet – qui serait le volume de données dont la grandeur justifierait son appellation - ou aux méthodes d'analyses spécifiques qui lui sont appliquées pour en tirer une information exploitable[18].

Cet « objet » Big Data est souvent passé au crible des fameux « 3V » qui servent à circonscrire de façon assez schématique l'ossature globale des mégadonnées. En premier lieu, V comme Volume, qui peut être illustré concrètement par l'évocation de plusieurs chiffres éloquents. Ainsi en 2010, Éric Schmidt (alors PDG de Google) déclare lors d'une conférence que tous les deux jours, nous créons autant d'information que ce qui a été créé entre le début de la civilisation et 2003, ce qui correspond à environ 5 exaoctets de données[19]. Une étude de la société de conseil Gartner, prévoit sur le marché d'ici 2020 la présence de milliards d'objets connectés (anciens et nouveaux), produisant une masse de données toujours plus abondante et complexe. Selon l'EMC encore, en 2010, le volume mondial de données numériques a atteint les 1.2 zettaoctets, et l'on estime, selon le cabinet IDC, que le monde stockera de 40 à 44 zettaoctets de données d'ici 2020[20]. Le grossissement de ces volumes s'explique par la numérisation de la société. Certes en 2000, on ne comptait que 20 % de l'information sur support numérique, le reste encore sur support analogique (cassette VHS, photo polaroid, radio, CDROM pour la musique etc… .), mais aujourd'hui, c'est 98 % de l'information qui est numérique, et parmi cette masse, 90 % de cette information a été créé ces dernières années. La numérisation du monde a donc enclenché une croissance exponentielle des données et de l'information, produite en grande quantité grâce aux outils connectés : les ordinateurs certes, mais aussi les smartphones, tablettes ou encore plus récemment, les données produites par les maisons intelligentes, des assistants maison (Google

[18]GILLIUM, Johann, 2015. Big data et bibliothèques: traitement et analyse informatiques des collections numériques, Mémoire sous la dir. de Monique Joly, ENSSIB. page 13

[19] SIEGLER, M.G. Eric Schmidt: Every 2 Days We Create As Much Information As We Did Up To 2003, [sans date]. TechCrunch [en ligne]. [Consulté le 26 août 2018]. Disponible à l'adresse : http://social.techcrunch.com/2010/08/04/schmidt-data/

[20]BRASSEUR, Christophe, 2016. Enjeux et usages du Big Data. 2e édition. Paris : Hermès Lavoisier. Management et informatique. ISBN 978-2-7462-4758-1. 658.05, page 10

Droits d'auteur réservés. OU (cc) BY-NC-ND

Home) ou encore des données produites par les compteurs Linky, dans le domaine de l'Energie. Au sein de cette masse, certaines organisations font figures de poids lourds. Citons entre autres : Le New York Stock Exchange (la plus grande bourse au monde) qui génère entre 4 et 5 téraoctets par jour ; le réseau social Facebook, qui accueille plus de 240 milliards de photos, c'est-à-dire 7 pétaoctets de données par mois ; dans le domaine de la recherche, le Grand collisionneur de hadrons (Genève) produisant 30 pétaoctets de données chaque année ou encore, pour citer un organisme que les archivistes connaissent bien, l'archivage du web mené par Internet Archive, qui stocke sur ses serveurs environ 18.5 pétaoctets de données depuis sa création[21]. Notre temps est donc bien manifestement celui des mégadonnées, ou encore pour employer un autre terme qui revient fréquemment, celui de la « datamasse », qui souligne bien d'une part l'immensité d'un écosystème, mais aussi son caractère hétérogène[22].

D'où ensuite un autre V, celui de la Variété. L'utilisateur produit un grand nombre de données certes ; mais aussi des données très variées. D'une part, elles sont produites dans tout type de domaine : santé, culture etc., cette variété constitue une sorte de puzzle, représentatif de chaque utilisateur dans toute sa complexité. Davantage, « la variété est […] la caractéristique d'un âge informatique où tout est donnée, contenu structuré ou non structuré ».[23] Cette notion de « contenu non structuré » ou donnée non structurée, est au centre de ce qui constitue le phénomène des Big Data, et on le verra, centrale également dans notre approche d'archivage des Big Data (l'archiviste ayant souvent à faire à du *contenu*, contenu qui correspond très souvent à un ensemble de données non structurées, ou au moins semi – structurées). De fait, un texte brut, un document en format .word ou .odt, un .pdf ou encore un fichier .csv, une image ou un son, font parties de ces données non structurées (ou semi – structurées) au cœur de l'approche Big Data. On peut se représenter les données non structurées comme des valeurs comprises dans un tableau Excel imaginaire, avec un nombre illimité de lignes et de colonnes, sans titre

[21] WHITE, Tom, 2015. *Hadoop - The Definitive Guide 4e-*. 4. Beijing : O'Reilly. ISBN 978-1-4919-0163-2, pages 3-4

[22] LEMBERGER, Pirmin, BATTY, Marc, MOREL, Médéric, RAFFAËLLI, Jean-Luc et GÉRON, Aurélien, 2016. *Big data et machine learning: les concepts et les outils de la data science*. 2e éd. Paris : Dunod. InfoPro. ISBN 978-2-10-075463-2. Q325.5, page 3

[23] GILLIUM, Johann, 2015. *Big data et bibliothèques: traitement et analyse informatiques des collections numériques*, Mémoire sous la dir. de Monique Joly, ENSSIB, page 15

Droits d'auteur réservés. OU (cc) BY-NC-ND

pour la plupart, et dont ces valeurs n'ont à priori pas de lien entre elles. A contrario, on parle de données structurées

> « Lorsque toutes les valeurs présentes ont un lien logique entre elles, connu à priori, et qu'elles appartiennent à des catégories bien précises, dont le nombre et les définitions sont connues »[24]

dans ce cas, à l'instar d'un tableur Excel bien réel. Ainsi, si le Big Data dépasse l'informatique traditionnelle c'est en cela, que cette dernière était exclusivement focalisée sur la donnée, alors que le Big Data amène l'information au centre de ses préoccupations[25] : une de ses raisons d'être est de pouvoir dégager du sens, de l'information au sein de données qui n'ont a priori aucun rapport entre elles, mais qui une fois traitées avec les outils et méthodes adéquates, dégagent une ou des information(s) pertinentes. Gilles Bobinet file une autre métaphore, qui a le mérite d'insister sur la notion d'information :

> [..] Il faut plutôt se figurer le Big Data comme un torrent de montagne, dont chaque goutte est un chiffre [...]. En apparence tout cela parait extrêmement désordonné et sans vraiment de sens ; pourtant il est possible d'en extraire une quantité d'information impressionnante.[26]

La science des données qui constitue le soubassement méthodologique de notre phénomène est bien une science de l'interprétation. Précisons néanmoins que le V de la Variété est aussi aujourd'hui le « V » le plus difficile à gérer, puisque le caractère hétéroclite des sources et des formats de données relève à chaque fois de situations particulières, difficiles à appréhender avec une méthode universelle[27].

Enfin nommons le dernier V, pour Vitesse, qui est peut-être aussi le critère de caractérisation qui nous concerne le moins pour notre présente étude. Il peut s'agir par exemple, pour un *pure player*, de faire baisser le temps de recherche des utilisateurs sur

---

[24] DENOIX, Antoine, 2018. Big Data, Smart Data, Stupid Data... : Comment (vraiment) valoriser vos données. Dunod. ISBN 978-2-10-077351-0., page 5

[25] LEMBERGER, Pirmin, BATTY, Marc, MOREL, Médéric, RAFFAËLLI, Jean-Luc et GÉRON, Aurélien, 2016. Big data et machine learning: les concepts et les outils de la data science. 2e éd. Paris : Dunod. InfoPro. ISBN 978-2-10-075463-2. Q325.5 , page 3

[26] BABINET, Gilles et ORSENNA, Erik, 2016. Big Data, penser l'homme et le monde autrement. Paris : Le Passeur. ISBN 978-2-36890-492-3., page 32

[27] LEMBERGER, Pirmin, BATTY, Marc, MOREL, Médéric, RAFFAËLLI, Jean-Luc et GÉRON, Aurélien, 2016. Big data et machine learning: les concepts et les outils de la data science. 2e éd. Paris : Dunod. InfoPro. ISBN 978-2-10-075463-2. Q325.5, page 17

Droits d'auteur réservés. OU (cc) BY-NC-ND

le web. A titre d'exemple, la technologie Google Instant a ainsi fait passer le temps de recherche utilisateur de 9sec à 7/6sec[28]. Dans notre cas en tant que spécialiste de l'information, le paramètre de la vitesse trouve son importance lorsqu'il est question de rechercher et obtenir un document rapidement, dans de très grands ensembles de données et ou l'indexation est parfois difficile et impose ses contraintes en termes de vitesse. Nous verrons plus loin que les bases de données NoSQL optimisent les performances de vitesse de « requêtage » par rapport à l'indexation classique des bases de données relationnelles.

On peut résumer les 3 V de cette façon en affirmant que le Big Data est le point de rencontre entre la multiplication des données non structurées, les besoins d'analyse de ces données et les progrès de la technologie[29]. Néanmoins, tous les spécialistes du Big Data ne s'accorde pas sur cette définition des « 3V ». En effet, pour certain, c'est là une définition lacunaire, qui ne décrit pas bien la réalité du phénomène :

> Les données caractéristiques du Big Data manquent d'une définition sérieuse […]. Le slogan des « 3V » rappel les « 3C » du reengeniring d'il y a 20 ans et n'éclaire pas davantage.[30]

, nous dit Pierre Delort. En fait, il s'agit aussi de percevoir le Big Data sous l'angle de la méthode. En effet, le besoin d'analyse montre que le Big Data est donc aussi perçu, et se déploie comme tel sous la forme de méthodes qui conjuguent statistique et informatique : « Le Big Data consiste à créer en exploratoire et par induction sur des masses de données à faible densité en information des modèles à capacités prédictives. »[31]. Le raisonnement par induction souligne assez l'aspect expérimental du Big Data, qui plutôt que de se baser sur des modèles déterminés à priori classiques (le cas de l'informatique décisionnel pratique), préfère avancer à coups d'essais et d'expérimentations, décelant des modèles prédictifs au sein de grands ensembles de données très variées, à faible densité informationnelle. Nous retrouvons ici encore la notion d'information – couplée à la notion de densité. Affinons ici notre propos :

---

[28] PERRET, Xavier et JACQUEMELLE, Guy, [sans date]. Comprendre le Big Data à travers les films de cinéma. OpenClassrooms [en ligne]. [Consulté le 26 décembre 2017]. Disponible à l'adresse : https://openclassrooms.com/courses/comprendre-le-big-data-a-travers-les-films-de-cinema

[29] BRASSEUR, Christophe, 2016. Enjeux et usages du Big Data. 2e édition. Paris : Hermès Lavoisier. Management et informatique. ISBN 978-2-7462-4758-1. 658.05 , page 10

[30] DELORT, Pierre, 2015. Le big data. Paris : Presses Universitaires de France. Que sais-je ?, n°4027. ISBN 978-2-13-065211-3. 005.7, page 5

[31] DELORT, Pierre, 2015. Le big data. Paris : Presses Universitaires de France. Que sais-je ?, n°4027. ISBN 978-2-13-065211-3. 005.7, page 42

Droits d'auteur réservés. OU (cc) BY-NC-ND

L'information dans l'environnement Big Data est extraite de données à signaux faibles : derrière cette expression, il est bien question d'une masse de données hétérogènes, difficile à exploiter du fait de leurs disparités et de leur volume – et donc à priori sans lien. Pour sacrifier à l'expression populaire, nous pourrions affirmer que trouver l'information intéressante et pertinente dans ces grands *data set* est un peu comme chercher une aiguille dans une botte de foin. C'est également ici qu'interviennent les outils et méthodes d'analyses.

Droits d'auteur réservés. OU (cc) BY-NC-ND

La prédiction du monde par les algorithmes

> Le Big Data est souvent associé au traitement de très gros volumes de données. Ce n'est pas tant la quantité qui est déterminante, mais l'association dans une même analyse de données variées afin d'en déduire des informations qu'il aurait été impossible de mettre en évidence avec les analyses classiques de données structurées,

Ou encore :

> Le critère du big data est la combinaison des informations structurées des bases de données avec des informations semi-structurées de logs et des informations non structurées d'enregistrement de conversations téléphoniques pour décider d'une action en temps réel[32]

De ces deux citations ressortent des notions fondamentales pour saisir le Big Data comme méthode. Relevons notamment les mots « association », « déduire », « combinaison ». Concrètement, les méthodes d'analyses déployées dans ces grands champs de données diverses et variées ont pour but de dégager des corrélations à l'intérieur de grands ensembles disparates dont les multiples composants n'ont, à priori, aucun rapport entre eux. En ce sens, le Big Data est créateur d'information, une information qui est produite à partir du travail des données. Le pattern dégagé ensuite à travers ces corrélations doit posséder une valeur prédictive, du fait notamment de sa récurrence. Agir selon une méthode Big Data, c'est être capable d'établir des modèles probabilistes, afin d'apporter des prévisions bâties sur les grands ensembles de données à disposition. En fait, les méthodes et outils déployés pour permettre d'élaborer de telles fonctionnalités relèvent de ce qu'on appelle le *machine learning*. Le *machine learning* ou apprentissage automatique, c'est en somme les algorithmes qui permettent aux machines d'apprendre une tâche à partir de grands ensembles de données. Cette pratique a été mise sur le devant de la scène en 2006, suite à un article de Geoffrey Hinton de l'Université de Toronto : « *A fast learning algorithm for deep belief nets* » (traduit comme « un algorithme rapide pour

[32] COINTOT, Jean-Charles et EYCHENNE, Yves, 2014. La révolution big data: les données au cœur de la transformation de l'entreprise. Paris : Dunod. Stratégies et management. ISBN 978-2-10-071142-0. 658.406, page 36

Droits d'auteur réservés. OU (cc) BY-NC-ND

les réseaux bayésiens profonds »)[33]. Informatique et statistique sont ici étroitement mêlées pour développer pleinement les potentialités des Big Data, alliées à un environnement technologique nouveau élaboré pour des besoins propres : évoquons la parallélisation de traitement, automatisé notamment par le modèle MapReduce de Google (construit spécifiquement pour automatiser l'indexation du web) ou encore la technologie open source Hadoop (fondation Apache) et la mise en place de bases de données non relationnelles pour encadrer ces nouvelles technologies (nous reviendrons sur ces points par la suite). Les cas d'usage du *machine learning* sont nombreux. Nous pouvons citer parmi les plus fréquents : la détection de fraude en cas de comportements inhabituels lors de transaction bancaire en ligne, l'estimation de la capacité d'un site marchand à vendre ses produits en se basant sur le comportement client (nombre de clic sur une page web, comportement lors de la navigation etc …) ou encore conception d'un algorithme de recommandation qui s'adapte au comportement de chaque prospect en particulier. Nombreux sont les cas d'usages, rendus possibles grâce à la disponibilité des données en quantité massive et aux nouvelles technologies mises en place. L'objectif final d'un bon modèle de *machine learning* étant d'obtenir une approximation de la fonction (c'est-à-dire un ensemble de variables prédictives), cela à partir de plusieurs observations[34]. Cependant, l'algorithme de prédiction développé n'est pas suffisant à lui seul. La démarche de l'apprentissage automatique est vaine si elle ne s'appuie pas sur sa matière fondamentale : les données. En effet, l'algorithme choisi et implémenté par le *data scientist* doit se nourrir d'un jeu de donnée (c'est la phase dite d'entrainement), à partir duquel il est capable de dégager des corrélations et conséquemment, de dégager de l'information. La phase de construction du jeu de donnée est très importante. La quantité des données récoltées est capitale pour éviter certaines erreurs courantes que peut rencontrer un *data scientist* dans sa démarche. Il ne s'agit certes pas pour nous ici de rentrer dans les détails du métier du spécialiste des données, mais citons tout de même les risques d'*Overfitting*, d'*Underfitting* ou

[33] LEMBERGER, Pirmin, BATTY, Marc, MOREL, Médéric, RAFFAËLLI, Jean-Luc et GÉRON, Aurélien, 2016. Big data et machine learning: les concepts et les outils de la data science. 2e éd. Paris : Dunod. InfoPro. ISBN 978-2-10-075463-2. Q325.5, préface

[34] LEMBERGER, Pirmin, BATTY, Marc, MOREL, Médéric, RAFFAËLLI, Jean-Luc et GÉRON, Aurélien, 2016. Big data et machine learning: les concepts et les outils de la data science. 2e éd. Paris : Dunod. InfoPro. ISBN 978-2-10-075463-2. Q325.5, page 111.

Droits d'auteur réservés. OU (cc) BY-NC-ND

encore des fameux Biais ou d'évènements rares[35]. Ainsi, une fois le jeu de donnée bien défini, il s'agit de déployer l'analyse selon un type d'apprentissage. Les deux modes les plus courants sont l'apprentissage non supervisé et l'apprentissage supervisé[36]. Le premier, parfois nommé *clustering* , est une méthode d'apprentissage automatique. Le logiciel, pour traiter un ensemble de données hétérogènes, divise ces données en sous-groupes : les données estimées similaires se trouvent dans un même groupe ; alors que les différentes dans des groupes distincts. A partir de ces données, on peut extraire une connaissance organisée. Contrairement à l'apprentissage supervisé**,** l'organisation de ces données n'est pas déterminée à priori, mais de façon dynamique : les catégories de classement s'élaborent en même temps que ce classement. L'apprentissage supervisé suppose donc un étiquetage des données d'apprentissages en préalable. Ensuite, le *data scientist* choisira un algorithme qui correspond le mieux à ses besoins. Là encore, nous n'entrerons pas dans les détails extrêmement techniques qui s'éloignent par trop de notre sujet, mais citons tout de même parmi les plus connus : la régression linéaire, les « k plus proches voisins » ou la classification naïve bayésienne (*Naive Bayes Classifier*). Ce type d'analyse de type « Big Data » peut s'appliquer à différents domaines et servir différents acteurs et organisations. Ainsi par exemple, les pompiers de New York possèdent un logiciel, intégrant un socle analytique puissant, qui permet de cartographier et d'indiquer les zones les plus susceptibles d'être victimes d'incendies[37]. Cependant, il est important de rappeler que l'usage de ces techniques relèvent avant tout d'usage commerciaux, propulsés par les *pure players* du Web et autres grandes structures.

Un des exemples bien connu dans ce domaine est celui du géant du streaming Netflix, dont l'algorithme de recommandation fonctionne sur le principe de l'apprentissage non supervisé (Contrairement à l'apprentissage supervisé, l'organisation de ces données n'est pas déterminée à priori, mais de façon dynamique : les catégories de classement s'élaborent en même temps que ce classement) : la plateforme, grâce aux

[35] L'*Overfitting* (sur – apprentissage) est un modèle trop spécialisé (trop riche), qui ne fait pas ressortir l'information parmi le bruit des données d'apprentissages. A l'inverse, l'*underfitting* (sous – apprentissage) est un modèle pauvre, qui s'appuye sur un biais fort de sa modélisation.

[36] PERRET, Xavier et JACQUEMELLE, Guy, [sans date]. Comprendre le Big Data à travers les films de cinéma. OpenClassrooms [en ligne]. [Consulté le 26 décembre 2017]. Disponible à l'adresse : https://openclassrooms.com/courses/comprendre-le-big-data-a-travers-les-films-de-cinema

[37] GILLIUM, Johann, 2015. Big data et bibliothèques : traitement et analyse informatiques des collections numériques, Mémoire sous la dir. de Monique Joly, ENSSIB, page 8

Droits d'auteur réservés. OU (cc) BY-NC-ND

données récoltées sur chaque utilisateur/clients, est capable de prévoir quelle série est susceptible de plaire, individuellement, à chaque profil utilisateur. Le système repose sur la collecte de données massives et variées : historique du profil utilisateur, actions produites sur le site (pause, replay, temps passé devant une vidéo), commentaire et évaluation des vidéos, « tag » autour duquel gravite le plus d'audience etc. La société américaine, grâce à ces analyses, peut également déterminer ses futurs programmes à succès.[38] D'autres géant du web utilisent ce type d'algorithme : Amazon par exemple développe également un algorithme de recommandations pour orienter ses utilisateurs vers des produits qu'ils sont susceptibles d'acheter. Nous pourrions accumuler les exemples de société ou service qui utilisent des algorithmes pour diriger leur stratégie et fidéliser leur client : un moteur de recherche qui investit massivement dans la R&D (Recherche et Développement) pour que l'utilisateur trouve toujours plus facilement ce qu'il recherche (perfectionnement de l'algorithme) ; une autre société qui se spécialise dans la confection de publicité ciblée, un appareil photo qui permet la reconnaissance faciale et détermine si nous connaissons la personne ou non etc. Google et Facebook sont les deux entreprises qui développent leur stratégie sur les données utilisateurs dans une perspective bien précise : fidéliser un maximum les utilisateurs pour récolter un maximum de données sur ces mêmes utilisateurs, afin de les monétiser par la suite auprès de leurs annonceurs. Ainsi Google : Détient 90 % de la recherche sur le net, donc tout site doit être référencé sur Google, sinon il n'a guère de chance de toucher un public. Google aujourd'hui, c'est 3,5 Milliards de recherches/jours, plus d'un milliard d'utilisateur sur YT et autant sur l'OS Android, mais c'est aussi plus de 15 millions de livres scannés ! De même pour Facebook : en 2013, l'entreprise américaine est à 1,4 Milliards d'utilisateur, cumulé à ses services tiers comme Whatsapp avec 700 millions d'utilisateurs et 300 millions sur Instagram. Dans ce contexte, le terme de Big Data recouvre une dimension singulièrement anxiogène. Le fait que des grandes organisations amassent et analysent un grand nombre de données liées à notre activité sur le web ou offline, et que ces mêmes institutions possèdent la puissance de calcul et les moyens technologues pour analyser ces mégadonnées, n'est pas sans rappeler parfois les plus illustres dystopies de la littérature du XXème siècle.

[38] DUMONT, Olivier, 2014. L'algorithme de Netflix, un cerveau à la place du cœur [en ligne]. [consulté le 26 décembre 2017]. Disponible à l'adresse: http://www.lemonde.fr/televisions-radio/article/2014/09/12/l-algorithme-de-netflix-un-cerveau-a-la-place-du-c-ur_4486880_1655027.html

Droits d'auteur réservés. OU (cc) BY-NC-ND

Big Data : Big Fear ?

En 2054, John Anderton dirige une unité spéciale, la division « Précrime », dont les membres ont la particularité de témoigner du don de clairvoyance. Les « précogs » ont en effet le pouvoir de prédire les crimes qui auront lieu dans le futur, et donc de connaitre *a priori* qui commettra le crime, et où ce crime aura lieu. Le commandant Anderton doit sa réputation à sa rigueur et sa foi sans faille dans le système des « precogs », jusqu'au jour où une vision lui révèle sa propre image, en tant que prochain criminel à poursuivre. Le livre de l'auteur de science-fiction américain Phillip K. Dick exprime la vision d'un futur ou la technologie est omniprésente : les voitures sont sans chauffeurs, les objets connectés utilisables sans connecteurs analogiques etc … . Mais ce qui nous intéresse ici dans l'œuvre de K.Dick, c'est la capacité de la Loi à juger coupable d'un crime un individu, qui précisément n'est pas encore passé à l'acte. Les *data scientist* ont l'habitude aujourd'hui d'utiliser les algorithmes de classification. Ces algorithmes, comme la classification naïve bayésienne, fonctionne selon l'apprentissage supervisé[39]. Un cas d'utilisation simple est la détection de spam par un filtre antispam : le filtre classe les mails en spam et non – spam, donc dans deux catégories pré – déterminées. L'algorithme du filtre, en se basant sur un certain nombre de variables, détermine donc dans quelle catégorie doit être classifiée l'objet entrant. A l'instar des « précogs », la classification de Bayes effectue une prédiction (probabiliste) pour décider à quelle classe appartient tel jeu de données. Dans le cas de la détection de spam, cela n'a pas de très grande conséquence, mais dans le cas ou cette approche implique des traitements de données dont l'objectif est de prédire des comportements humains, cela devient plus sensible. Et à la dystopie de K.Dick de mettre en exergue la foi trop grande placée dans la prédiction, comme si celle-ci se suffisait à elle-même, infaillible, dogmatique. En effet, la Loi dans Minority Report n'a aucune distance critique par rapport aux « précogs ». De façon similaire, une foi absolue dans l'algorithme de classification ferait de cet algorithme une boite noire, dont les résultats seraient suivis comme les paroles d'un oracle, en faisant fi des problématiques de faux positifs/faux négatifs et vrais positifs/ vrais négatifs. Il s'agit bien de ne pas faire des outils et techniques du Big Data des « *weapons of math*

[39] LEMBERGER, Pirmin, BATTY, Marc, MOREL, Médéric, RAFFAËLLI, Jean-Luc et GÉRON, Aurélien, 2016. Big data et machine learning: les concepts et les outils de la data science. 2e éd. Paris : Dunod. InfoPro. ISBN 978-2-10-075463-2. Q325.5, page 125

Droits d'auteur réservés. OU (cc) BY-NC-ND

destruction », pour reprendre le fameux jeu de mot de Cathy O 'Neil[40]. Mathématicienne et ancienne analyste financière dans plusieurs grandes institutions financières américaines, Cathy O'Neil déploie une réflexion critique (mais pas sceptique) sur le Big Data : les algorithmes ne sont pas des objets issus d'esprits purs et désincarnés, mais bien aussi des concepts porteurs de biais ou de présupposés très humains :

> *many of these models encoded human prejudice, misunderstanding, and biais into software systems that increasingly managed our lives*[41].

O'Neil pointe plusieurs affaires qui ont impliqués une approche Big Data avec des conséquences catastrophiques. La méthode « *stop and frisk* » (interpellation et fouille), construite à l'origine pour la police de New York, possède un outil (CompStat) qui permet au service de police d'avoir une approche *data driven.* Concrètement, l'outil, qui fonctionne sur un algorithme de classification, est typiquement utilisé comme une boite noire par la police new yorkaise. Selon la mathématicienne qui a étudié de près le logiciel, environ 85% des arrestations concernaient des individus afro-américains ou latino–américains. Davantage, seulement 0.1% des arrestations étaient liées à un crime violent, le reste pour des crimes mineurs (possession de drogue ou ivresse sur la voie publique). La politique de « *stop and frisk* » conduisit aussi à un taux important de faux positifs[42]. L'auteur évoque également l'effet de la « *nasty feedback loop* »[43], à savoir que les arrestations effectuées enrichissent le jeu de données, qui lui-même justifie (de façon insidieuse) la démarche. Le logiciel en quelque sorte, enrichit le crime qu'il est censé endiguer, en condamnant des personnes pour des crimes mineurs, personnes qui ensuite seront incarcérées, et conséquemment davantage « catégorisées » comme criminel potentiel. Et ici de retrouver un effet « *minority report* » :

> *As stop and frisk grew, the venerable concept of propable cause was rendered virtualy meaningless, because police were hinting not only people who might*

[40] O'NEIL, Cathy, 2016. Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. New York : Crown. ISBN 978-0-553-41881-1.

[41] *Ibidem*, page 3

[42] *Ibidem*, page 92

[43] *Ibidem* , page 93

Droits d'auteur réservés. OU (cc) BY-NC-ND

have already commited a crime, but also those who might commit one in the future[44].

Même si le programme s'est démontré efficace (depuis la mise en place de CompStat, on note en effet une chute de la criminalité à New York), force est de constater que sa logique ne s'inscrit aucunement dans le cadre légal : la technique du « *stop and frisk* » foule aux pieds le quatrième amendement, et notamment la défense des minorités. Le pourcentage des interpellés est révélateur en cela, que pour les mêmes crimes, une personne blanche avait moins de chance de subir une interpellation de la part de la police New Yorkaise, orientée ou non par l'algorithme de CompStat, *data driven* ou non. Ainsi, non seulement la démarche violait le droit des minorités, mais en se justifiant elle-même par l'algorithme, elle confirmait le sort de ces minorités.

La question de la régulation des modèles mathématiques, qui aujourd'hui impactent de plus en plus la société, est donc fondamentale. Une réflexion critique et éthique doit accompagner tout spécialiste de la donnée, dont les compétences techniques et scientifiques ne se suffisent plus à elle mêmes. Le bagage du *data scientist* doit comporter des éléments de droits et de réflexion sur le métier du spécialiste des données. Mais cela n'est pas suffisant. Cathy O'Neil évoque à juste titre l'enjeu de transparence des algorithmes afin que ceux-ci ne soient pas de simples boites noires aux yeux des utilisateurs, mais bien des

models that have a significant impact on our lives, including credit scores and e-scores, should be open and avaible to the public »45, ou encore, dans la perspective du modèle européen, que « any data collected must be approved by the user, as an opt-in […] also prohibits the reuse of data for other purposes[46].

De très récents scandales, en tête celui concernant Facebook et Cambridge Analytica, ont montré la faible emprise que les utilisateurs ont sur leur donnée personnelle, et comment – avec un socle technologique et une stratégie appropriée – des organisations pouvaient s'en servir pour développer à notre insu des stratégies marketing ou économique, avec une réelle incidence sur le plan social et politique. Le nouveau règlement général sur la protection des données (RGPD) s'inscrit dans

[44] *Ibidem*, page 94

[45] *Ibidem, page 215*

[46] *Ibidem*, page 214

Droits d'auteur réservés. OU (cc) BY-NC-ND

cette volonté de l'Europe et de la CNIL de protéger les données personnelles des utilisateurs. Le règlement promulgue par exemple un droit à l'oublie numérique, une responsabilisation des entreprises plus forte quant à la sécurité et à la conservation des données à caractères personnelles, la mise en place d'audits et d'analyse conformité ou encore la création d'un poste de DPO (*data protection officer*). Cette batterie d'obligations doit engager les entreprises à mener une politique des données plus qualitative et plus transparente[47]. Le spécialiste de l'information a son rôle à jouer dans ce vaste chantier. L'archiviste par exemple, aura à charge de veiller à ce que certaine donnée soit conservée, selon des normes de sécurité claires et conformes aux exigences de la loi, selon une durée de conservation clairement définie. Ainsi les données qui sont déplacées de la source primaire, pour aller dans un stockage d'archive, doivent respecter une politique de conformité à la loi, d'intégrité, de fidélité et ne doivent pas être perdues (c'est l'importance d'assurer la pérennité de la donnée). L'archiviste, dans son double rôle de records manager et de « conservateur », est à même d'apporter son expertise de gestionnaire de l'information pour optimiser l'environnement Big Data des institutions.

UN ENVIRONNEMENT TECHNIQUE NOUVEAU : DE LA GED AU CONTENT LAKE ?

Le SI (systèmes d'informations) d'une organisation est un univers complexe, qui manipule des informations porteuses de nombreux enjeux et souvent sensibles. On parle souvent aujourd'hui d'urbanisation du SI et de gouvernance de l'information ; notions que l'archiviste doit connaître afin d'évaluer le rapport qu'entretient son environnement de travail avec ces concepts[48]. Le spécialiste de l'archivage numérique est directement concerné par cette urbanisation, en ce qu'il doit s'assurer que son propre « *pool* d'application » entre dans la démarche d'urbanisation (par exemple, afin d'éviter des problèmes de compatibilité) ; de plus en tant qu'expert transverse, il doit comprendre la logique de fonction entre les différentes applications :

[47] BRASSEUR, Christophe, 2016. Enjeux et usages du Big Data. 2e édition. Paris : Hermès Lavoisier. Management et informatique. ISBN 978-2-7462-4758-1. 658.05, page 42

[48] BÉCHARD, Lorène, FUENTES HASHIMOTO, Lourdes et VASSEUR, Édouard, 2014. Les archives électroniques. Paris : Association des archivistes français. Les Petits guides des archives. ISBN 978-2-900175-06-4. CD974.4, page 18

Droits d'auteur réservés. OU (cc) BY-NC-ND

> Les problématiques portées par le concept de gouvernance de l'information ne sont ni plus ni moins que celles qui structurent le travail des gestionnaires des documents d'activités et des archivistes[49].

Mais dans le cadre d'un écosystème Big Data, le spécialiste de l'information comme l'archiviste doit se confronter à un SI en perpétuelle évolution et avec des enjeux bien particuliers : l'information doit échapper aux logiques de silo et être exploités à travers tous types de données mises en relation. Les SGBDR (Système de Gestion de Bases de Données Relationnelles) classiques ne répondent plus tout à fait aux besoins analytiques des organisations orientées *data driven*. De même les GED et SAE traditionnels sont confrontés à de nouveaux enjeux, auxquels le spécialiste de l'information (et en particulier l'archiviste) doit être capable de répondre.

Les bases de données NoSQL

Les bases de données classiques, ou SGBDR (Système de gestion de bases de données relationnelles) constituent le modèle d'architecture le plus utilisé encore dans l'architecture des SI aujourd'hui. Le SGBDR peut être composé d'une ou plusieurs bases de données, avec ses logiciels associés. Les solutions les plus présentes sur le marché sont les bases de données MySQL, Oracle et PostgreSQL. Une base de données est composée de plusieurs tables ou relations, tables qui seront gérées par l'algèbre relationnel (intersection et indexation, jointure externe/interne, produit cartésien etc …). Les bases de données stockent les données opérationnelles : les données actives, fréquemment et rapidement interrogées, généralement par un grand nombre d'utilisateurs. Des bases de données dédiées, synthétisant les informations et pour répondre à des besoins précis, sont aussi installées : ce sont les *datawarehouses*. Un *datawarehouse* regroupe des données hétérogènes, provenant de multiples applications du SGBDR. Cela peut être des fichiers plats comme Excel, fichier texte ou encore XML. Une des fonctions importantes du *warehouse* est l'historisation, qui se base sur le principe de conservation de la donnée : chaque donnée présente dans celui-ci possède une date ou un numéro de version, afin d'éviter les doublons par rapport à la BDD, et ainsi obtenir une certaine traçabilité (il y a la une notion d'historisation). Cependant, le coup pour le SI est énorme : les BDD stockent difficilement les très grands volumes de données, et le temps de

[49] *Ibidem*

Droits d'auteur réservés. OU (cc) BY-NC-ND

traitement des données entrantes peut être très long (à cause de l'indexation). En outre, les entrepôts de données classiques ne prennent pas en compte les données non-structurées : données XML hiérarchiques, journaux web non structurés ou graphiques point à point ne sont donc en général pas gérés[50].

Ainsi, pour répondre aux besoins de traitement des mégadonnées, tant en termes de volumétrie que de variétés et de vitesse, certains SI se dotent de bases de données dites NoSQL. Les bases données NoSQL répondent à un certain nombre d'exigences : stockage et traitement sont distribués dans des clusters de serveur à bas coût, liés entre eux par plusieurs milliers de nœuds. L'intégrité des données est moins importante que la performance et la disponibilité. Surtout, ce sont des bases de données qui peuvent gérer et traiter à la fois des données structurées et non structurées. La grande majorité des solutions de type NoSQL sont hautement dynamiques, dépourvu de tout schéma ou modèle de base de données construit en préalable (il n'y a pas de jointure entre les tables par exemple), cela précisément afin d'accueillir et de traiter des données peu ou non structurées. Ces systèmes ont en outre l'avantage d'être open source, pour favoriser une forte interopérabilité au sein des SI. Il existe plusieurs types de solutions NoSQL sur le marché aujourd'hui. Ces solutions ont souvent une orientation propre, et une structure différente pour satisfaire à différents besoins. Nommons par exemple les bases de données orientées agrégats (BDOA) : le concept de table (propre aux SGBD classiques) est remplacé par la notion d'agrégat, qui va contenir les données les plus sollicitées[51]. Ensuite, selon la nature des agrégats manipulés, on distingue 3 sous catégories, soient les suivantes : les entrepôts clé – valeur, les bases de données orientées documents, les bases de données orientées colonnes. Les bases de données NoSQL orientées documents sont en fait bien connues des professionnels de l'informations puisqu'elles peuvent se connecter à des solution ECM comme Sharepoint ou Alfresco, tant il est vrai que « Les applications qui manipulent naturellement des documents comme les systèmes de gestion de contenu ou les plateformes de blogs

[50] BRASSEUR, Christophe, 2016. Enjeux et usages du Big Data. 2e édition. Paris : Hermès Lavoisier. Management et informatique. ISBN 978-2-7462-4758-1. 658.05, page 78

[51] LEMBERGER, Pirmin, BATTY, Marc, MOREL, Médéric, RAFFAËLLI, Jean-Luc et GÉRON, Aurélien, 2016. Big data et machine learning: les concepts et les outils de la data science. 2e éd. Paris : Dunod. InfoPro. ISBN 978-2-10-075463-2. Q325.5, page 40

Droits d'auteur réservés. OU (cc) BY-NC-ND

[CMS] pourront […] utiliser avec profit une BDOD »[52]. En quoi ? La base de données NoSQL orientée document fonctionne sur le couple clé – valeur, avec pour valeur des entités semi structurées ou non structurées, c'est-à-dire des documents en général en format XML ou JSON, mais aussi en format Microsoft Office (MS Word, Excel etc …), ou encore PDF. La BDOD possède plusieurs avantages sur le SGBDR pour le traitement des documents : les objets documents peuvent être de format différent (un aspect plus complexe à gérer dans un SGBD) tout en étant situé sur le même emplacement. En d'autres termes, c'est comme si plusieurs documents, possédant chacun une structure différente, pouvaient être stockés dans une même table, sans laisser de champs vides. Conséquemment, la base de données non relationnelle est plus malléable et dynamique, ses champs s'adaptant à différentes structures de documents. De même, cela permet à un utilisateur de modifier la structure d'un document à la volée, sans la contrainte d'un modèle pré-établi. En « greffant » des fonctionnalités métiers sur ce type de base de données pour la gestion électronique des documents, le spécialiste de l'information peut tirer parti de l'élasticité de pareil système, notamment lors de scénario type « Big Data ». A titre d'exemple, une grande banque française est confrontée à la problématique de la gestion de milliards de flux AFP[53] – volumétrie exponentielle, que les GED en place ont à encaisser et à indexer. Une base de données NoSQL, couplée à un type de stockage efficace, peut répondre au défi de l'ingestion de plusieurs milliards de documents. Pour soutenir cette gestion des grands et importants volumes de données, il faut aussi déployer une solution de stockage des bits appropriés : c'est ici que le *framework* Hadoop entre en scène.

Les *frameworks* Hadoop et Mapreduce

MapReduce et Hadoop sont également des programmes inévitables de l'univers technique des big data. Le premier est une solution open source, crée par Doug Cutting (célèbre également pour être le père de la librairie java Lucène, sur laquelle s'est développée le moteur de recherche et d'indexation Solr) en 2005, et dont le développement est poursuivi aujourd'hui par la communauté Apache. Ce

[52] LEMBERGER, Pirmin, BATTY, Marc, MOREL, Médéric, RAFFAËLLI, Jean-Luc et GÉRON, Aurélien, 2016. Big data et machine learning: les concepts et les outils de la data science. 2e éd. Paris : Dunod. InfoPro. ISBN 978-2-10-075463-2. Q325.5, page 44

[53] Le format AFP (*Advanced Function Presentation*) est un format de stockage crée par IBM. Il est constitué comme une archive compressée, qui peut contenir un grand nombre de données non structurées de type textuel ou documentaire.

Droits d'auteur réservés. OU (cc) BY-NC-ND

modèle dispose de plusieurs propriétés fondamentales pour mettre en place une stratégie Big Data, lorsqu'il est implémenté dans un cluster de serveurs[54]. Citons parmi les propriétés du modèle : une grande tolérance aux pannes couplées à un faible coût d'entretien des infrastructures. Hadoop se déploie en effet sur des serveurs bon marché (cela peut être plusieurs milliers de serveurs), supporté par une mécanique complexe de notification (logique maitre – esclave entre les différents nœuds du serveur, dans le vocabulaire Hadoop, on parlera de *job tracker/task tracker*), qui vient aussi assurer la gestion des métadonnées des fichiers systèmes. Par-dessus cette sécurité, Hadoop respecte aussi le principe de réplication, notamment grâce à un système HDFS (*Hadoop Distributed File System*) ou système de fichier distribué. Dans l'environnement des volumes extrêmes, une seule machine ne suffit pas pour supporter la quantité de données en présence. Ainsi un système de fichiers distribués est mis en place sur tous les nœuds du réseau : il y a réplication et distribution des données ; si bien qu'en cas de panne d'un disque dur, les données ne sont pas perdues mais bien répliquées sur un autre nœud du réseau. Ce mécanisme de réplication («*fault tolérance*»), qui peut faire penser RAID, est aussi plus performant. La configuration de Hadoop, basée sur JBOD (« just a bunch of disks»), délivre des opérations (de type lecture/écriture) sur les disques qui s'avèrent 10% à 30% plus rapides qu'avec un RAID 0[55]. En outre, selon la configuration JBOD, le HDFS peut continuer à fonctionner même si un disque tombe en panne, ce qui n'est pas le cas avec le RAID : si un seul disque tombe en panne, le nœud de ce disque n'est plus disponible (et les données sont donc perdues). De plus, le HDFS a la capacité d'héberger des fichiers de très grande taille : un bloc, ou quantité de données sur un disque, peut atteindre jusqu'à 64 Mo (contre 512 octets pour un système de fichier ordinaire)[56]. Il est important de noter que Hadoop est hautement interfaçable, et ainsi peut facilement se connecter à d'autres applications. L'outil intégré Sqoop permet par exemple de transférer facilement des données entre Hadoop et différents systèmes de SGBD classiques ; et donc également vers une application d'archivage, par exemple, qui adopte la stratégie de rendre la donnée

[54] LEMBERGER, Pirmin, BATTY, Marc, MOREL, Médéric, RAFFAËLLI, Jean-Luc et GÉRON, Aurélien, 2016. Big data et machine learning: les concepts et les outils de la data science. 2e éd. Paris : Dunod. InfoPro. ISBN 978-2-10-075463-2. Q325.5, page 65

[55] WHITE, Tom, 2015. Hadoop - The Definitive Guide 4e-. 4. Beijing : O'Reilly. ISBN 978-1-4919-0163-2, page 285

[56] LEMBERGER, Pirmin, BATTY, Marc, MOREL, Médéric, RAFFAËLLI, Jean-Luc et GÉRON, Aurélien, 2016. Big data et machine learning: les concepts et les outils de la data science. 2e éd. Paris : Dunod. InfoPro. ISBN 978-2-10-075463-2. Q325.5, page 197

Droits d'auteur réservés. OU (cc) BY-NC-ND

archivée rapidement disponible, donc non plus stockée sur un support annexe (cassettes, espace de stockage indépendant), mais immédiatement au même niveau que les données les plus actives, les plus « chaudes ». Il s'agit en fait de pouvoir accéder rapidement à la donnée archivée, et cela à bas prix. Hortonworks (société de logiciel informatique spécialisé dans la distribution d'Hadoop), en lien avec Composite Software (Cisco), a par exemple développée une solution qui a permis à une grande banque d'affaire américaine d'obtenir accès aux données critiques auparavant indisponibles car trop « anciennes », mais néanmoins importantes pour ses stratégies d'investissement[57]. Les données en question ont été chargées sur Hadoop depuis des bandes magnétiques, ce qui correspondait à environ 5 ans de données de « trading » et de marchés. Pour ce faire, Cisco à mis en place une plateforme de virtualisation de donnée, grâce à l'outil Cisco Data Virtualisation Suite qui est un logiciel d'intégration de données permettant de connecter tous types de données sur l'ensemble du réseau et de les faire apparaître dans une vue consolidée, comme si elles étaient en un seul endroit. La suite est fondée sur Cisco Information Server, qui permet d'accéder aux données de manière non invasive et fédérer les données. Grâce à cette suite il est tout à fait envisageable de puiser à différentes sources de données, en l'occurrence dans notre cas, les données archivées, utiles à la banque pour trouver des *business insights*, de réduire les risques d'investissements et d'améliorer sa prise de décision. Hadoop peut donc recevoir et traiter des données et/ou des documents archivés. Un expert Big Data comme Dirk Ross (IBM) parle même d'Hadoop comme une destination privilégiée pour les données à archiver[58]. Le HDFS est un bon candidat pour assurer le stockage des données massives à archiver, de façon efficace et sécurisée. En effet, selon les bonnes pratiques de stockage appliqué à l'archivage,

[57] SENSMEIER, Lisa, 2013. Modernizing Data Archiving with Hadoop. Hortonworks [en ligne]. 29 août 2013. [Consulté le 26 décembre 2017]. Disponible à l'adresse : https://fr.hortonworks.com/blog/modernizing-data-archiving-virtualization-for-big-data-analytics/

[58] DEROOS, Dirk, [sans date]. Hadoop as an Archival Data Destination - dummies. [en ligne]. [Consulté le 26 décembre 2017]. Disponible à l'adresse : http://www.dummies.com/programming/big-data/hadoop/hadoop-as-an-archival-data-destination/

Droits d'auteur réservés. OU (cc) BY-NC-ND

> de manière générale, un système de stockage doit préserver l'intégrité des données qui lui sont confiées et en garantir l'accès dans un délai convenu par un contrat de service[59].

Un cluster HDFS complait également aux exigences d'extensibilité (possibilité d'ajouter des modules, sans remettre en cause l'architecture), de stabilité matérielle, d'ouverture (open source), de volumétrie accepté (très haute, dans notre cas), d'outils de contrôle et de mesure de sécurité en cas de perte (la réplication) et aussi d'excellente performance pour les opérations de lecture et d'écriture. Le paradigme MapReduce peut parfaitement répondre à ce dernier point. MapReduce a été décrit pour la première fois en 2004, par deux ingénieurs de Google, Jeffrey Dean et Sanjay Ghemwat dans un article désormais célèbre nommé : « *MapReduce : Simplified Data Processing on large Clusters* ». Les deux abstractions « *Map* » et « *Reduce* » permettent de faciliter les calculs parallèles sur de grands ensembles de données en assurant un gain de temps conséquent dans la phase de traitement. C'est un système qui permet de dégager du sens des données archivées précédemment par exemple, sur disques ou cassettes : *«[MapReduce] changes the way you think about data and unlocks data that was previously archived on tape or disk* »[60]. En somme, le système de batch dégagé par MapReduce réduit le temps d'accès la donnée. MapReduce convient parfaitement aux applications dont les données sont destinées à être stockées selon la technique WORM, et pour performer une analyse à grande échelle dans ce large set de données (en mode batch)[61]. L'archivage opéré dans Hadoop peut ainsi convenir aux besoins de chercheurs ou de *data scientist* qui auraient besoin d'accéder quasiment en temps réel à la donnée archivée, pour des besoins analytiques[62].

Julien Masanès, qui était chargé de piloter les projets expérimentaux de la BNF relatif à l'archivage du web, est aujourd'hui CEO de Internet Memory Research (IMR), start up française qui s'inscrit dans les pas d'Internet Archive : archiver le

---

[59] BANAT-BERGER, Françoise, DUPLOUY, Laurent et HUC, Claude, 2009. L'archivage numérique à long terme: les débuts de la maturité ? Paris : la Documentation française Direction des archives de France. Manuels et guides pratiques. ISBN 978-2-11-006942-9. Z699, page 97

[60] WHITE, Tom, 2015. Hadoop - The Definitive Guide 4e-. 4. Beijing : O'Reilly. ISBN 978-1-4919-0163-2, page 6

[61] *Ibidem*, pages 8-9 (voir tableau – annexes)

[62] DEROSA, Guy, [sans date]. Utilizing Hadoop & HDFS as an Active Archiving & Storage Framework. [en ligne]. [Consulté le 26 décembre 2017]. Disponible à l'adresse : http://www.aptude.com/blog/entry/utilizing-hadoop-hdfs-as-an-active-archiving-storage-framework

Droits d'auteur réservés. OU (cc) BY-NC-ND

web grâce à des technologies pouvant répondre à ce besoin. L'archivage du web, de données donc nativement numériques, confronte directement le professionnel de l'information au Big Data. Ainsi, les fondateurs du projet IMR sont bien confrontés aux « V » du Volume et de la Variété de l'information dans leur démarche d'archivage du web. Les ingénieurs et les archivistes mettent en place des modèles probabiliste pour traiter les archives et les informations, faisant ressortir des modèles récurrent (association de mot, de concept) capable de révéler des tendances.

> Ce mouvement du Big Data concerne particulièrement les archives et les bibliothèques confrontées depuis longtemps aux mêmes problèmes (organiser et permettre l'analyse de l'information) dans le monde analogique et qui doivent se les approprier et les développer pour remplir leur mission dans le monde numérique[63]

On retrouve également le socle technologique décrit dans ce paragraphe : IMR a déployé un cluster de serveur pour gérer une masse de données archivées conséquentes et variées, basées sur une architecture HDFS (*Hadoop Distributed File System*). En plus de réduire le risque de défaillance critique (si un nœud tombe en panne, les autres peuvent continuer à fonctionner pour maintenir le système à flot), il est particulièrement commode pour les archives : d'un point de vue sécurité (création de copie sur plusieurs nœuds) et de puissance analytique : l'analyse peut opérer en cas de panne grâce aux copies précédemment évoquées : « il permet de gérer de manière incrémentale l'investissement dans le système de stockage, et parce qu'il permet de construire de construire un système d'adressage et de nommage lié à l'archive et non à son instanciation matérielle (nom du serveur physique).»[64] . En outre, le traitement de la donnée s'effectue par le patron d'architecture distribué par le *framework* Hadoop, c'est-à-dire de nouveau ici, MapReduce. Celui-ci permet de faciliter le traitement (notamment sa rapidité d'exécution) : procéder à la création de table et d'index, non pas grâce aux techniques d'indexations des SGBD relationnel (Access, Mysql), mais avec des bases de données non relationnelles comme Hbase pour les tables distribuées et ElasticSearch pour les index distribués. Le nommage se fait dynamiquement ici, grâce à des fonctions de hachages adéquats.

[63] LAURENT, Pascale, LOWINGER, Hélène, MILLET, Jacques et CALDERAN, Lisette, 2015. Big data: nouvelles partitions de l'information actes du séminaire IST Inria, octobre 2014. Louvain-la-Neuve [Paris] : De Boeck ADBS. ISBN 978-2-8041-8915-0. 004, page 33.

[64] *Ibidem*, page 34

Droits d'auteur réservés. OU (cc) BY-NC-ND

Les nouveaux volumes d'archives, tout comme leur variété, et les outils pour composer avec ces nouveaux défis propres à un environnement Big Data peuvent donner une nouvelle mission à l'archiviste : « cela nécessite que les archives commencent à créer des index et des tables à partir de leur collection pour permettre ces analyses, ces calculs et non plus seulement pour trouver les documents unitaire » : l'archiviste obtient un rôle organisationnel au niveau du document et de la donnée; qu'il doit organiser, structurer aux mieux pour permettre de déployer une démarche macro analytique sur ces ensembles. C'est à l'archiviste aujourd'hui de comprendre les enjeux du Big Data par rapport aux archives, alors que le terme recouvre ici une grandes diversité d'éléments (données, documents, etc …). Il s'agit de valoriser l'archive sous l'angle du Big Data, afin que celle-ci soit valorisée et puisse apporter une valeur significative à ceux qui peuvent en tirer partie (pour la recherche, mais aussi pour toute organisations, privée ou publique).

Cette combinaison des bases de données NoSQL avec HDFS (ou un autre moyen de stockage très performant, comme le Cloud – mais c'est là un sujet qui mériterait presque un autre mémoire) dépasse les solutions de GED ou d'archivage traditionnel, souvent limités dans leur approche Big Data. Le déplacement progressif du paradigme du document vers celui de la donnée – en tant que support de l'information – demande des ressources que les solutions traditionnelles de gestion documentaire ne sont souvent pas à même (dans la majorité des cas) de fournir. En outre, dans un contexte ou une politique de gouvernance des données est de plus en plus essentielles, les solutions de type NoSQL sont davantage adaptées que les SGBD classiques. En effet, le schéma d'une base de données non relationnelle se définit dynamiquement, et non à priori ; conséquemment, chaque projet de gouvernance peut être appliqué sans pour autant avoir à redéfinir un modèle de donnée (ce qui peut s'avérer fastidieux et coûteux) :

> *In general, traditional relational databases are not good for data governance because you need to define your schema in advance*[65]

On peut estimer que pareille solution trouverait un intérêt pour l'archiviste records manager, dont le métier a un fort aspect gouvernance informationnelle[66] :

[65] CASTENADO, Frederico, 2018, Understanding Data Governance. O'Reilly, ISBN, 978-1-491-99076-6, pages 9

[66] BÉCHARD, Lorène, FUENTES HASHIMOTO, Lourdes et VASSEUR, Édouard, 2014. Les archives électroniques. Paris : Association des archivistes français. Les Petits guides des archives. ISBN 978-2-900175-06-4. CD974.4, page 18

Droits d'auteur réservés. OU (cc) BY-NC-ND

> Les problématiques portées par le concept de gouvernance de l'information ne sont ni plus ni moins que celles qui structurent le travail de gestionnaires de document d'activité et des archivistes.

Ainsi, il s'agit peut-être aujourd'hui pour le professionnel de l'information, de songer à des socles technologiques nouveaux. Conséquemment, essayons d'esquisser maintenant ce que l'on entend par la notion de « Content Lake ».

Vers le Content Lake ?

Dans un article volontairement clivant[67], Phillip Goupil avance :

> Les solutions pour gérer les documents tels que nous les connaissons vont disparaître à moyen terme. Les éditeurs vont devoir évoluer, se faire acheter, ou s'éteindre. Quant aux archivistes et aux informaticiens concernés, c'est certain, leurs fonctions vont évoluer ou disparaître.

Nous assisterions ainsi au chant du cygne des systèmes de gestion documentaires traditionnels. L'explosion des volumes de documents, mais aussi des données et métadonnées n'est plus gérable seulement par les GED ou SAE nativement orientée vers les bases de données relationnelles. De fait,

> Les solutions capables de répondre au big data vont prendre la place des solutions actuellement en place. Les bases de données vont progressivement abandonner SQL pour le NoSQL et l'architecture va scinder les données et les documents, des traitements.[68]

De plus, comme nous l'avons esquissé plus haut, le besoin de traitement des documents et des données doit être désormais pris en compte avec beaucoup d'attention. Le système de gestion documentaire traditionnel – qu'il soit une GED ou une SAE – est organisé par trop en silo. C'est-à-dire que la plupart des solutions ne sont pas adaptées pour faire partie d'un SI au sein duquel les différentes composantes collaborent entre elles (notamment via des API), cela afin de pouvoir appliquer des analyses entre différents types d'informations (structurées et non structurées). Dans cette perspective, certaines études affirment la mort du concept

[67] GOUPIL, Phillip, 2017. La GED, c'est fini !. Archimag [en ligne]. 11 juillet 2017. [Consulté le 26 décembre 2017]. Disponible à l'adresse : http://www.archimag.com/demat-cloud/2017/04/11/ged-fini

[68] *Ibidem*

Droits d'auteur réservés. OU (cc) BY-NC-ND

de l'ECM (Entreprise Content Management) et annonce l'avènement du « Content Service », ou encore du « Big Content » :

> *[ECM] been replaced by the term Content Services, a strategic concept that covers three aspects, namely Content Services Applications, Platforms and Components*[69]

Des exemples récents au sein de grandes structures françaises témoignent de ce besoin de délaisser des solutions de GED historiques – coûteuses et ne répondant plus toujours aux enjeux business - pour passer vers des outils de type Content Services. L'éditeur MarkLogic, leader dans le domaine des bases de données, se positionne dans la gestion de contenu selon une approche orientée à la fois gestion de très importants volumes de données et de documents, mise en relation de type de données très variées et optimisation dans la façon d'accéder à cette information. Marklogic propose ainsi un véritable « Data Hub », dont les fonctionnalités marquantes sont : un moteur de recherche puissant (recherche plein texte et suggestion), une mise en relation des données issues des applications du SI et des documents, de nombreux plans de classements selon le besoin métier, une gestion des rôles utilisateurs (différents niveaux de droits) et la traçabilité (journalisation) du contenu. MarkLogic se présente comme une solution unique, qui fait à la fois office de GED, de serveur applicatif et de base de données (NoSQL, schéma agonistique et orientée documents). Le Crédit Agricole (Consumer Finance) a ainsi décidé de remplacer ses 12 GED Filenet (IBM) internes et externes par un seul produit Marklogic[70]. Les données de type documentaire de la GED ont ainsi été pleinement dans la stratégie transformation digitale et Big Data (reposant sur Hadoop – MapR) du groupe. De même, AXA a également décommissionnée ses GED FileNet pour privilégier la solution de Document Management as a Service proposée par MarkLogic. Dans ces scénarios, le HDFS de Hadoop était utilisé pour recevoir les données métiers extraites des documents.[71]

[69] WOODBRIDGE, Michael, 2017. The Death of ECM and Birth of Content Services [en ligne]. [consulté le 26 décembre 2017]. Disponible à l'adresse: https://blogs.gartner.com/michael-woodbridge/the-death-of-ecm-and-birth-of-content-services/

[70] DECAUDIN, Frederic, VAIDIE, Pierre, 2018. MarkLogic, spécialiste de l'intégration de silos de données. In : Conférences des Architectes, Capgemini. 2018. pp. 38.

[71] *Ibidem*

Droits d'auteur réservés. OU (cc) BY-NC-ND

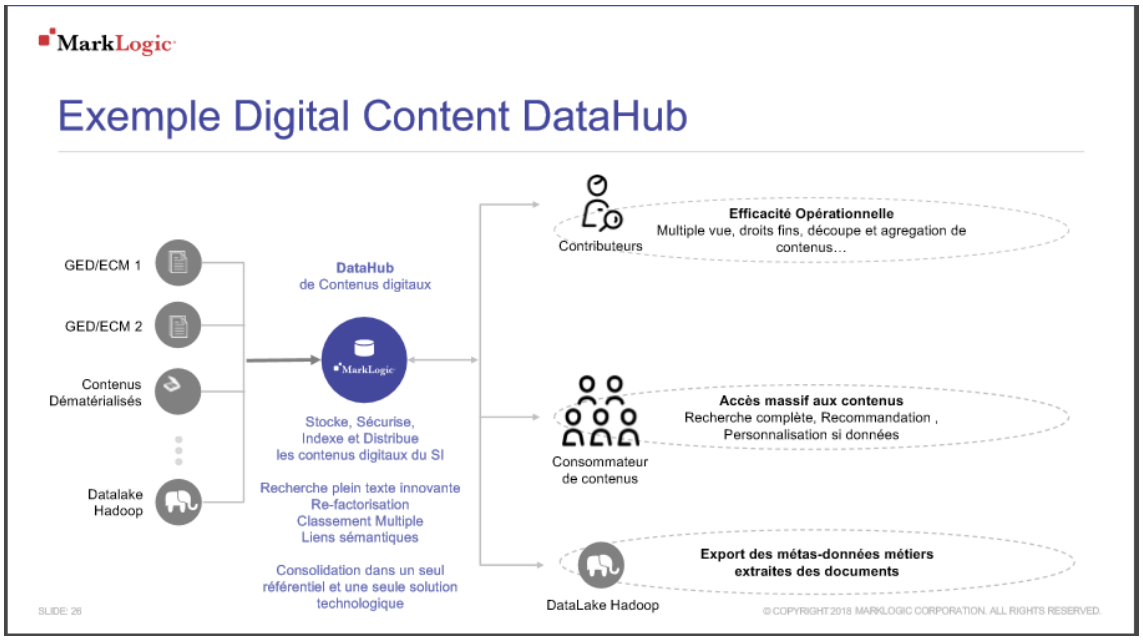

Figure 1:principe du Big Content illustré par MarkLogic

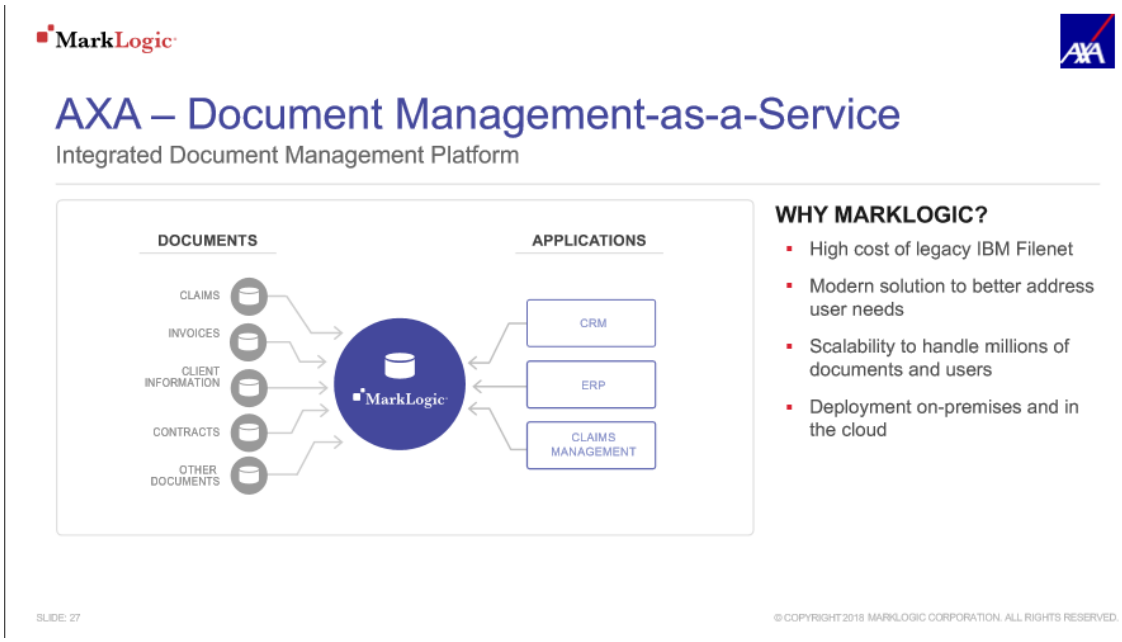

Figure 2: cas de Axa, qui a remplacé ses 12 GED par une solution MarkLogic

L'évolution des besoins pousse les solutions de gestion de contenu documentaire à évoluer. Les utilisateurs choisiront probablement des solutions *custom*, c'est-à-dire des logiciels spécifiques s'appuyant sur des outils NoSql offrant de très bonnes performances et une grande scalabilité. Néanmoins, affirmer que la GED est morte est sans doute excessif. Sans aucun doute, cette assertion mérite d'être quelque peu nuancée. Tout d'abord, si notre monde numérique est prompt à évoluer, cette évolution est parfois dictée par une tendance marketing, qui consiste à apporter de

Droits d'auteur réservés. OU (cc) BY-NC-ND

la nouveauté ou de la « disruption » (pour employer un terme plus à la mode) sans autre finalité que d'apporter cette nouveauté ou de créer cette disruption :

> on pourrait voir dans cette litanie d'enterrements et de baptêmes une course folle à l'innovation où l'innovation serait sa propre finalité ce qui laisse perplexe.[72]

Dans cette configuration, les besoins métiers de l'utilisateur sont moins pris en compte que l'obsession de réaliser des technologies plus performantes. Ainsi, le « Content Lake » ou les tentatives d'archivage dans le HDFS évoqué ci-dessus n'intègrent pas nativement les fonctionnalités métiers que possède une SAE ou une GED. En d'autres termes, ces logiciels se montrent techniquement très performants, mais pauvre fonctionnellement. De plus, ce sont là des outils qui en aucun cas pour le moment ne saurait remplacer l'humain, qui reste encore détenteur du savoir métier et « fonctionnel ». Cela s'avère d'autant plus vrai en ce qui concerne le métier de l'archiviste, pour qui déjà « un simple logiciel ne saurait en aucun cas remplir toutes les fonctions nécessaires à l'archivage électronique »[73]. En fait, l'approche orientée Content Lake ou de l'archivage dans Hadoop semble une approche exclusivement portée par les DSI, c'est-à-dire une approche très orientée « outil ». La notion d'archivage employée ici semble davantage être synonyme de stockage. On rappellera utilement qu'archiver, au sens archivistique, n'est pas stocker :

> le stockage et les opérations de sauvegardes n'impliquent en aucun cas que les documents/données soient organisés et accessibles de manière durable[74].

L'archivage se distingue du stockage en ce qu'il se déploie comme une méthode : on croit encore trop qu'archiver, c'est entreposer des documents ou des données numériques comme de vieux cartons dans une cave humide…. L'archivage n'est donc aucunement réductible à la simple problématique du stockage ou à celle de l'obsolescence des supports de l'information[75].

---

[72] CHABIN, Marie-Anne, 2017. Technologies de l'information : rupture ou continuité ? [en ligne]. [Consulté le 26 décembre 2017]. Disponible à l'adresse : http://www.arcateg.fr/2017/07/07/technologies-de-linformation-rupture-continuite/

[73] BÉCHARD, Lorène, FUENTES HASHIMOTO, Lourdes et VASSEUR, Édouard, 2014. Les archives électroniques. Paris : Association des archivistes français. Les Petits guides des archives. ISBN 978-2-900175-06-4. CD974.4, page 10

[74] BÉCHARD, Lorène, FUENTES HASHIMOTO, Lourdes et VASSEUR, Édouard, 2014. Les archives électroniques. Paris : Association des archivistes français. Les Petits guides des archives. ISBN 978-2-900175-06-4. CD974.4, page 8

[75] *Ibidem*, page 9

Droits d'auteur réservés. OU (cc) BY-NC-ND

> L'archivage est une activité qui consiste à gérer et organiser l'information dans le temps, quel que soit son support, pour la rendre accessible durablement, bien au-delà de la durée de vie des supports.[76]

Ainsi, il faut que l'archivage des Big Data s'incarne lui aussi dans un *framework,* que l'archiviste, dans un univers numérique dominé par la donnée et dont les outils s'automatisent davantage et tendent vers l'analytique, porte afin de faire montre de ses compétences pour construire une archivistique des mégadonnées (au risque peut être de mal négocier son « tournant numérique » et de ne plus être pris au sérieux au sein des organisations). Ainsi par exemple, la société Mosaik.ly, éditeur de la solution éponyme MOZAIK, permet également de réaliser la valorisation du patrimoine informationnel de l'entreprise via des outils de visualisation puissant, qui exposent les liens entre différentes données. Il est intéressant de noter que les fonctionnalités de l'outil et son plan de visualisation (en « étoile ») ont été déterminées depuis la méthode Arcateg (archivage par catégorie)[77]. Arcateg est une méthode d'archivage qui propose une architecture de référentiel, afin de mieux gérer le cycle de vie des actifs informationnels (document et données) dans le temps. La méthode d'archivage propose une structure solide pour la gouvernance de l'information en entreprise : lié à un outil de visualisation comme Mosaik, les utilisateurs peuvent trouver facilement l'information voulue avec sa durée de conservation[78]. Mener un archivage des Big Data n'est pas une simple affaire de technologie : le savoir-faire de l'archiviste, son approche et ses méthodes, sont complémentaires.

[76] *Ibidem*, page 8

[77] CHABIN, Marie-Anne, 2018. Des documents d'archives aux traces numériques: identifier et conserver ce qui engage l'entreprise la méthode Arcateg. Bois-Guillaume (Seine-Maritime) : Klog éditions. ISBN 979-10-92272-26-0. 658.403 8, page 116 - 117

[78] L'outil Mosaik propose une implémentation de la philosophie d'Arcateg pour identifier les documents sensibles et mieux s'assurer de leur bonne gouvernance. Il s'agit d'une part bien de donner des statistiques sur le nombre de documents concernées, mais aussi d'identifier les différentes actions effectuées sur les documents par les différents acteurs. Ainsi, Mosaik propose une expérience *dataviz* complète des traces laissées par les activités de l'entreprise sur des données déterminées.

Droits d'auteur réservés. OU (cc) BY-NC-ND

L'ARCHIVAGE : LA PLUS-VALUE STRATEGIQUE POUR LES ENTREPRISES (L'ARCHIVE ET LE QUATRIEME V : VALEUR).

L'ARCHIVAGE : UNE GRANDE POUSSEE VERS LE NUMERIQUE

> Pour les archives, l'irruption du numérique est indubitablement une chance qui nous est donnée de gagner une visibilité accrue dans nos sociétés et de convaincre nos interlocuteurs que les archives sont les outils d'une meilleure gouvernance dans les organismes[79]

Depuis ces dernières années, l'archiviste est confronté au numérique : les décideurs, publics comme privés, s'attachent à prendre correctement le tournant du tout numérique afin de rester à flot et exploiter les technologies de l'air du digital. Les services producteurs numérisent donc leurs archives papiers, mieux, le document ou la donnée nativement numérique, rencontrent aussi le besoin de l'archivage ; et rencontrent donc aussi ses problématiques d'ordre techniques, comme juridiques. L'archivage du numérique a bouleversé les fondements de l'archivistique[80]. Dans le prolongement du constat de Françoise Banat Berger, nous affirmons aujourd'hui que la donnée, comme paradigme supplantant celui du document, apporte un nouveau challenge dans le domaine de l'archivistique, et plus globalement, dans le domaine des sciences de l'information.

> Le Big Data est omniprésent dans les médias. […] Ces données massives sont-elles du seul domaine des informaticiens, statisticiens, des politiques et des créateurs d'entreprises ? Les professionnels de l'information-documentation n'ont-ils pas un rôle à jouer dans ce nouveau paysage : identification, qualification, archivage, classification ?[81]

[79] BANAT-BERGER, Françoise, 2010. Les archives et la révolution numérique. Le Débat. 2010. N° 158, pp. 70-82. DOI 10.3917/deba.158.0070., page 82

[80] *Ibidem,* page 71

[81] Big data: nouvelles partitions de l'information actes du séminaire IST Inria, octobre 2014. Louvain-la-Neuve [Paris] : De Boeck ADBS. ISBN 978-2-8041-8915-0. 004, introduction

Droits d'auteur réservés. OU (cc) BY-NC-ND

La gestion de l'information concerne les documents comme les données, dans la mesure ou le mot information regroupe les données (de type structurées) comme les documents (semi structurées ou non structurées)[82]. L'information documentaire étant une composante essentielle du Big Data, ses professionnelles issues des domaines des bibliothèques, du management de l'information ou de l'archivage numérique ont un rôle à jouer. A ce titre, il est intéressant de noter que les premières initiatives des sciences de l'information dans le champ de Big Data vient des bibliothécaires.

Approche et outils du Big Data dans les bibliothèques

Plusieurs chercheurs ont tâché de montrer comment les bibliothécaires perçoivent le phénomène des Big Data, et comme ils se situent par rapport à celui-ci dans leur métier. Il s'agit dans un premier temps de savoir en quels termes, selon quelle compréhension, les bibliothécaires qui utilisent le concept de Big Data ? Il semble possible de discriminer deux catégories : d'une part, le terme Big Data fait référence à une compétence particulière, orientée ; d'autre part, le terme fait davantage référence à une définition plus centrée sur les data elles-mêmes[83]. Lorsque le mot est dit « orienté pratique ; compétences », le Big Data fait ici clairement référence aux outils, processus, technologies et idées déployées pour maîtriser de très grandes quantités de données. Lorsque le terme est davantage « orienté donnée » : le Big Data est défini comme donnée et/ou information. L'auteur souligne que la distinction entre donnée et information n'est pas clairement établie ici, considérant de fait l'information comme « des données qui sont traitées pour être utiles », donc que dans ce cas, Big Data désigne à la fois la donnée et l'information[84]. Les auteurs ont aussi cherché à voir ce qui caractérise le mieux les données du Big Data pour les bibliothécaires dans leurs publications, à l'aune des 4V :

[82] CHABIN, Marie-Anne, 2018. Des documents d'archives aux traces numériques: identifier et conserver ce qui engage l'entreprise la méthode Arcateg. Bois-Guillaume (Seine-Maritime) : Klog éditions. ISBN 979-10-92272-26-0. 658.403 8, page 11.

[83] *Ibidem*, page 5

[84] *Ibidem*

Droits d'auteur réservés. OU (cc) BY-NC-ND

Table 4. Summary of big data's characteristics in the LIS definitions.

No. of def.	Volume	Velocity	Variety	Veracity
D1	★			
D2	★	★	★	
D3	★	★	★	★
D4	★		★	
D5	★			
D6	★		★	★
D7	★	★	★	
D8	★		★	★
D10	★			★
D11		★		★
D12	★			★
D13	★	★		★
D14	★			
D15	★			
D16				
D17	★			
D18	★			★
D19	★			
D20	★			★
D21				
D22	★		★	
D23		★	★	★
D24	★	★	★	
D25	★	★	★	
D26	★			
D27	★			
D28	★	★	★	★
D29	★			★
D30	★			★
D31	★	★		
D33	★	★	★	
D34	★			★
D35	★	★		
	29	12	12	14

Figure 3: résultat de l'enquête portant sur la perception du Big Data dans les publications des bibliothécaires

Il en ressort que le V de Volume est la caractérisation première, la plus récurrente. Globalement, les analyses font ressortir 5 aspects de l'attitude des bibliothécaires à l'égard du Big Data : A quoi correspond le Big Data ; la caractérisation de la donnée ; comment la donnée est sollicitée par les acteurs, les challenges du Big Data et les bénéfices du Big Data. Dans la pratique, pour les bibliothécaires et leur recherche, le Big Data est conçu dans un premier temps orienté donnée, c'est à dire correspondant à des caractéristiques de volume, variété, vélocité. Et dans un deuxième temps, orienté pratique, c'est à dire que le Big Data renvoi à des technologiques particulières pour gérer le data déluge : « *it could be concluded that big data is considered data to be processed with developed technologies in*

Droits d'auteur réservés. OU (cc) BY-NC-ND

librarianship »[85]. Certaines études évoquées ici montrent que le Big Data est un très gros challenge pour les bibliothèques. Un des aspects marquant de ce challenge, est révélateur pour nos recherches sur l'archivage des Big Data, est le manque de métadonnées dans les grands sets de données. Cette absence de contextualisation rend complexe, d'un point de vue descriptif, documentaire et archivistique, d'accéder, de gérer et de décrire les données issues du Big Data. De plus, la communauté des bibliothécaires (par extension, de spécialistes de l'information) souffre d'un manque de compétences Big Data :

> *From a value creation point of view, librarians need skills tocreate useful knowledge or information from big data. From the resource management point of view, librarians need skills to collect, store and maintain the data of volume, velocity, variety and veracity. In addition, the skills to handle privacy issues or data security are also needed.*[86]

Une idée qui est confirmée par les spécialistes de la profession, avec en tête certains conservateurs qui affirment bien que « le Big Data en tant que tel est rarement abordé dans les bibliothèques »[87]. Pour autant, il revient de mentionner des interrogations et des esquisses de réflexion sur le couple Big Data/Bibliothèque. Certains mettent ainsi en exergue les *big data skills* des bibliothécaires. Il y est question par exemple, de la transformation de la bibliothèque vers un modèle 4.0, c'est à dire une bibliothèque intelligente, capable de déployer des capacités d'analyse sur les besoins des utilisateurs, voire de leur proposer des suggestions de lectures[88]. En effet, il serait simple pour les algorithmes d'apprendre depuis les larges sources de données présentent dans les bibliothèques. Les recherches de Johann Gilium portent par exemple sur les techniques analytiques tel que le texte et *data mining* comme enjeux majeurs pour les bibliothèques. Aussi pour certain, la collecte de données massives en bibliothèques a le double avantage d'assurer la conservation de ces données (en parallèle des collections matérielles), mais aussi, dans une perspective plus « performative », de renseigner davantage le besoin des utilisateurs (collecte des données dans le sens de l'informatique décisionnelle ici –

[85] *Ibidem*, page 7

[86] *Ibidem,* page 9

[87] BELLIER, Luc, 2017. Organisation des données, organisation du travail en bibliothèques universitaires à l'heure du Big Data, Mémoire sous la dir. de Nathalie Marcerou-Ramel, ENSSIB, page 10

[88] ZHAN, Ming, 2017. Understanding big data in librarianship. Journal of Librarianship and Information Science. 13 décembre 2017. DOI 10.1177/0961000617742451., page 1

Droits d'auteur réservés. OU (cc) BY-NC-ND

c'est-à-dire, élaborer l'expérience utilisateur et anticiper ses besoins par l'analyse des données)[89] :

> Avec des objectifs différents, les bibliothèques partagent avec les entreprises [les GAFA] la nécessité de mieux connaitre l'usage qui est fait de leur service.[90]

De plus, des logiciels comme ezPAARSE ou PREVU proposent des fonctionnalités d'analyses de logs, afin de mieux cerner le comportement utilisateurs par rapport aux services commun de la documentation, et donc proposer une expérience utilisateur améliorée.[91]Luc Bellier insiste aussi sur les évolutions métiers nécessaires : l'automatisation des process ne remplacera pas le personnel humain des bibliothèques, mais appel bien à une hybridation des compétences d'une part, et à un mangement plus agile, et donc moins tayloriste, d'autres part :

> On y cherche moins la maîtrise et l'efficacité d'un processus parfaitement rodé et préparé pendant des mois pour rendre un service attendu, que la capacité d'expérimenter, d'essayer, pour offrir des outils peut être moins maîtrisés, mais plus en phase avec des besoins labiles.[92]

Dans le contexte des mégadonnées, le bibliothécaire a donc tout intérêt à monter en compétences, pour faire évoluer sa vision métier et ainsi produire des services en phase avec les besoins de son époque. Dans cette perspective, il nous semble bon d'évoquer l'émergence du rôle du « *data librarian* »[93]. Le « bibliothécaire des données » doit posséder des compétences en data management, mais aussi en traitement statistique de la donnée et des outils idoines. Son travail consiste substantiellement en un nettoyage des jeux de données disponibles, afin de faciliter la tâche des chercheurs ou des utilisateurs qui exploiteront ces données. Il existe des formations pour former les bibliothécaires cherchant à monter en compétence dans ces domaines. Sans faire du bibliothécaire un *data scientist*, des compétences en la matière lui serait utile pour exprimer les possibilités de mieux cerner la

[89] BELLIER, Luc, 2017. Organisation des données, organisation du travail en bibliothèques universitaires à l'heure du Big Data, Mémoire sous la dir. de Nathalie Marcerou-Ramel, ENSSIB, page 16

[90] *Ibidem*, page 19

[91] *Ibidem*, page 20

[92] *Ibidem*, page 70

[93] LAPÔTRE, Raphaëlle, 2014. Faire parler les données des bibliothèques: du Big Data à la visualisation de données, Mémoire sous la dir. de Julien Velcin, ENSSIB, page 50

Droits d'auteur réservés. OU (cc) BY-NC-ND

problématique des données, d'améliorer leur qualité (notamment en enrichissant leur contexte) et au final mieux se placer au sein du processus décisionnel des organisations, et cela en cernant mieux le besoin du public visé. Aux Etats Unis, certains services des bibliothèques proposent d'accompagner les chercheurs dans la mise en place de leur plan de gestion des données, comme la bibliothèque du National Institues of Health (NIH Library)[94]. En effet, la politique de National Science Foundation (NSF) contraint tout chercheur de constituer une « feuille de route » décrivant la gestion du cycle de vie des données étudiées pendant le projet de recherche. De plus, ce DMP (Data Management Plan) inclut aussi que le chercheur entre en conformité avec la politique de partage des données, elle aussi déterminé par la NSF. Dans ce cadre là, le service de la NIH Library supporte le chercheur dans le montage de son plan de données : mise à disposition de solutions logiciels (par exemple, la solution *Data management planing tools*), accompagnement dans la manipulation de l'outil, enseignement des langages Python ou R[95] à des fins statistiques et de datavisualisation etc … . De même l'archiviste, en tant que spécialiste de l'information et de sa gestion, est concerné par des enjeux similaires.

L'archivage numérique : la prise en compte des mégadonnées

Dans deux ouvrages prospectifs, Diane Baillargeon et Pierre Fluckiger admettent comme un état de fait que la transition de l'archiviste vers le numérique, ou pour mieux dire, la problématique majeure de l'archiviste d'aujourd'hui et à venir, est de faire face au volume des documents et données à gérer, c'est à dire le problème d'accroissement de la donnée née numérique :

> L'archiviste doit […] construire un argumentaire solide et des solutions facilement applicables, pour que les documents bureaucratiques, mais surtout les données emmagasinées dans les grands systèmes de gestion institutionnels, puissent faire l'objet d'une épuration basée sur des règles de conservation

[94] FEDERER, Lisa, 2016. Research data management in the age of big data: Roles and opportunities for librarians. Information Services & Use. 1 janvier 2016. Vol. 36, n° 1-2, pp. 35-43. DOI 10.3233/ISU-160797., page 39

[95] Python er R sont deux langages de programmations proposant des librairies très utiles pour l'exploration des données. Ces langages sont souvent utilisés dans la recherche scientifique et en sciences humaines.

Droits d'auteur réservés. OU (cc) BY-NC-ND

> adaptées à l'économie des systèmes et non pas calquées mécaniquement sur celles élaborées pour l'univers analogique[96].

Ou encore :

> Au-delà des problèmes techniques pour lesquels il n'y a aucune raison de douter que des solutions soient trouvées, c'est de toute évidence la masse de données à évaluer qui représentera une nouvelle difficulté à surmonter, puisque, faut-il le répéter, la quantité produite est exponentielle et peut donner le vertige.[97]

Les auteurs insistent également sur un aspect important : que désormais l'archive doit être évaluée véritablement pour son caractère fonctionnel. C'est à dire, non pas comme une donnée figée et « patrimonialisée», mais comme une donnée active qui doit être soumise à une bonne gouvernance pour être vraiment impactante (ex : les données de l'administration à Genève). Pour ce faire, l'archiviste devra se doter de compétences numériques certaines et collaborer dans la proximité avec les services informatiques[98]. Ainsi, l'archiviste doit pouvoir se positionner en amont de la phase de conception du SI: « Ces derniers [les archivistes] auront été associés à la conception des systèmes dès leur « idéation » afin que les aspects de gouvernance documentaire soient pris en compte ». Nous retrouvons ici encore la notion de gouvernance. Dans le contexte des mégadonnées, l'archiviste aura plusieurs outils à sa disposition pour mener à bien cette entreprise, notamment grâce aux progrès technologiques : système de classification et indexation semi-automatique bien plus performant, côte de classification comprit par l'utilisateur à l'aide du taxonomie ultra précise, des fichiers proposés depuis le contenu même du document, une épuration des fichiers déclenchée automatiquement (règle de conservation attachée à la classification) etc … . De plus, l'archiviste saura composer avec une certaine diversité des canaux de productions de documents et de données : « l'archiviste de 2030 sera en mesure d'assurer la sélection et la conservation des documents utiles pour l'organisation quel que soit l'appareil qui aura servi à les créer »[99]. Louise Fuentes

[96] SERVAIS, Paul et MIRGUET, Françoise, Archivistes de 2030: réflexions prospectives. Louvain-la-Neuve : Academia-l'Harmattan. Publications des Archives de l'Université catholique de Louvain, 32. ISBN 978-2-8061-0214-0. 020, page 19

[97] SERVAIS, Paul et MIRGUET, Françoise, 2015b. L'archive dans quinze ans: vers de nouveaux fondements. Louvain-la-Neuve : Academia-l'Harmattan. Publications des Archives de l'Université catholique de Louvain, 33. ISBN 978-2-8061-0225-6. Z, page 250

[98] *Ibidem* , page 20

[99] *Ibidem,* page 21

Droits d'auteur réservés. OU (cc) BY-NC-ND

Hashimoto semble aller dans le même sens, lorsqu'elle appelle l'archiviste à davantage se situer en amont et davantage participer aux projets informatiques :

> Les services d'archives font face à un cercle vicieux : on ne travaille pas avec les métiers au quotidien, on traite les archives à posteriori, on n'est pas intégré à notre organisation et comme notre activité se concentre souvent sur la communication/valorisation, on ne travaille pas avec les métiers et ils ne font donc pas appel au service d'archives[100].

La pratique du records management peut changer cela : le record manager doit savoir s'imposer dans les processus métier, et montrer en quoi, de part son expertise, il est une valeur ajoutée au métier. Ainsi, la pratique du records management peut permettre à l'archiviste de pleinement intégrer l'environnement métier, et donc aussi le cas échéant, l'environnement Big Data. Dans ce contexte, on peut imaginer que l'archiviste puisse contribuer par exemple à donner de la valeur à la donnée. Hashimoto évoque les compétences de l'archiviste pour « le développement de politiques transversales », « les tentatives de décloisonnement des SI », « l'essor de l'Open Data », mais aussi « la croissance exponentielle de l'information et le besoin de protection des données personnelles »[101]. Donner de la valeur à la donnée, c'est par exemple s'assurer que cette donnée soit toujours accompagnée d'éléments contextuels, de métadonnées descriptives et techniques.

> le classement et l'identification des archives sont une autre face du métier d'archiviste. Il s'agit de rendre les archives exploitables et ré-exploitables. Dans le cas des données, la mise en place, dès l'origine des métadonnées permet de répondre au problème. Mais […] c'est encore rarement le cas [,,,] Il faut donc que l'archiviste […] les « consolide » en le dotant de métadonnées, de contexte, de sens, de conservation. Ainsi il leur donne la plus value attachée à l'archivage, il réduit les coûts de conservation et facilite leur utilisation par

[100] SERVAIS, Paul et MIRGUET, Françoise, 2015b. L'archive dans quinze ans: vers de nouveaux fondements. Louvain-la-Neuve : Academia-l'Harmattan. Publications des Archives de l'Université catholique de Louvain, 33. ISBN 978-2-8061-0225-6. Z., page 99

[101] *Ibidem* , page 96

Droits d'auteur réservés. OU (cc) BY-NC-ND

> le producteur. Ce sont ces services et bien d'autres que les centres de données proposent à leurs clients dans le cadre des *Big Data*.[102]

Cette valeur ajoutée peut aussi venir de l'expertise de l'archiviste / records manager qui prend garde à ce que les documents et les données les plus sensibles soient conformes aux normes en vigueur[103]. Evoluant dans un environnement technologique ou le paradigme du document cède progressivement sa place à celui de la donnée, l'archiviste peut se positionner au niveau la gouvernance de l'information. En fait, de part certains aspects, son profil peut se rapprocher d'un manager de la donnée capable de mener des projets de Master Data Management (MDM) :

> *With the digitization of virtually everything, the importance of governance rises as we need more context around data that drives decision-making. Context includes lineage and provenance: who created the data? Where did it come from? Has it been versioned? Is it accurate?* [104]

Comme le document numérique, une donnée de qualité doit donc être accompagnée de métadonnées descriptives (renseignement du contexte de création de la donnée), cela afin d'assurer la conservation de la donnée mais aussi sa bonne utilisation. Une politique de gouvernance des données recouvre bien des enjeux de confidentialité (périmètre de sécurité définie pour groupe d'utilisateurs définis), d'intégrité (la donnée n'a pas été altérée par des utilisateurs non autorisés) et d'accessibilités (accessibilité de la donnée pour un groupe d'utilisateur déterminé)[105]. En Australie par exemple, les Archives Nationales (entre autres) imposent la sécurité, le cycle de vie, le délai de rétention et les obligations de suppression appliquées aux documents et données des organisations publiques[106], cela conformément à plusieurs décrets et normes (nommons ici entre autres, *le Freedom of Information Act 1982* et le *Financial Management and Accountability Act 1997* (FMA Act), qui affirment la nécessité de la mise en place de projet records management et de gouvernance). En outre, dans le contexte des mégadonnées, des organisations comme ARMA

---

[102] SERVAIS, Paul et MIRGUET, Françoise, 2015b. L'archive dans quinze ans: vers de nouveaux fondements. Louvain-la-Neuve : Academia-l'Harmattan. Publications des Archives de l'Université catholique de Louvain, 33. ISBN 978-2-8061-0225-6. Z. , page 137

[103] On pensera ici aux normes ISO 15489 ou NF Z42-013 de l'AFNOR

[104] CASTENADO, Frederico, 2018, Understanding Data Governance. O'Reilly, ISBN, 978-1-491-99076-6

[105] *Ibidem*, page 4

[106] *Ibidem*, page 7

Droits d'auteur réservés. OU (cc) BY-NC-ND

préconise une gouvernance cognitive pour le Big Data[107]. On entend ici par « cognitive », l'ensemble des moyens informatiques mis en place pour améliorer la gestion des données, notamment dans l'identification et la maitrise du cycle de vie des données sensibles ou des documents d'activités. Il s'agit bien de trouver l'information à gérer dans les grands sets de données, et donc bien de pénétrer la complexité du Big Data[108]. Les fonctionnalités d'analyses sémantiques qu'embarquent ces outils permettent de dégager au sein des données non structurées les notions de contexte et de sentiments. Le programme informatique est donc ici véritablement interprétatif. Cependant, l'expertise humaine (dans notre cas, celle du record manager) n'est pas pour autant remplacée par la technique : l'apport analytique des programmes permet en fait ici à l'utilisateur de prendre des décisions plus rapidement et plus précisément. L'utilisateur a davantage de temps pour le décisionnel, tandis que le programme informatique automatise les opérations de recherche et d'identification[109]. Par exemple, dans le cadre d'un projet de *E-discovery* ou de GDPR, il peut s'agir de réduire rapidement les cas à un ensemble de données beaucoup plus petit et pertinent, afin de réduire les coûts d'examen ; identifier les quasi-doublons et les documents contextuellement similaires ou encore explorer visuellement les données de cas pour obtenir rapidement un aperçu de ce cas, comprendre les faits clés, les schémas de communication et les fils de discussion, afin de dégager facilement les preuves et agir en conséquence[110]. La catégorisation automatique de l'information constitue une évolution essentielle dans la gestion des documents d'activité pour le record manager, et plus globalement pour faciliter la gouvernance des données dans notre contexte des mégadonnées. Ainsi, Le records management s'associe bien à la gouvernance des données. De plus, le records manager est lui aussi impacté par le Big Data : il est un challenge qui s'impose à lui, et le contraint, si ce n'est à changer de « méthode », à se tourner vers de nouveaux outils d'analyse[111]. Cependant, en terme d'approche métier, il est fort

[107] 2-Hogg-Cognitive-Governance-Opportunities-with-Big-Data.pdf, [sans date]. [en ligne]. [Consulté le 26 décembre 2017]. Disponible à l'adresse : http://www.arma-metromd.org/wp-content/uploads/2017/05/2-Hogg-Cognitive-Governance-Opportunities-with-Big-Data.pdf

[108] 2-Hogg-Cognitive-Governance-Opportunities-with-Big-Data.pdf, [sans date]. [en ligne]. [Consulté le 26 décembre 2017]. Disponible à l'adresse : http://www.arma-metromd.org/wp-content/uploads/2017/05/2-Hogg-Cognitive-Governance-Opportunities-with-Big-Data.pdf, page 9

[109] *Ibidem*, page 22

[110] *Ibidem*, page 26

[111] 4-Big-Data-Issues-for-Federal-Records-Managers.pdf, [sans date]. [en ligne]. [Consulté le 26 décembre 2017]. Disponible à l'adresse : http://www.arma-metromd.org/wp-content/uploads/2017/05/4-Big-Data-Issues-for-Federal-Records-Managers.pdf

Droits d'auteur réservés. OU (cc) BY-NC-ND

à parier qu'archiviste et records manager puissent également se positionner dans les rôles liés au Master Data Management.

Archivage et MDM

Antoine Denoix propose de restituer son métier à l'ère de la donnée, et pour ce faire de l'interroger directement :

> Quel est mon métier ? Et apportez-vous une réponse large, centrée sur le client [dans notre cas, l'utilisateur à qui sont destinées les archives], et indépendante de la façon opérationnelle que vous avez aujourd'hui d'apporter de la valeur.[112]

En tant qu'archiviste, il s'agit donc de repenser ma profession et de la mettre en perspective par rapport aux enjeux de la gestion des données. Nous avons essayé de le montrer ci-dessus, l'archiviste, dans son approche Records Management, à un fort rôle à jouer dans la gestion du cycle de vie des données depuis sa création, et donc, dans le cadre des politiques de gouvernance, de gestionnaire des documents et données. L'archiviste/records manager peut tout-à-fait se positionner de façon transverse par rapport au business, la Direction des systèmes d'information (DSI) et les services juridiques. Il y a fort à parier qu'un archiviste peut endosser également le rôle du « data steward ». Le data steward a pour fonction de capturer l'information et de renseigner, documenter rigoureusement la donnée, cela grâce à un ensemble d'éléments de définition. Il peut ainsi s'agir du renommage des éléments, afin de rendre les données /documents à une communauté cible plus intelligible ; d'identifier la donnée maître, c'est-à-dire de s'assurer que le jeu de données ne compte pas de doublons ; de vérifier si la donnée est obsolète ou non (c'est-à-dire, toujours pertinente pour la communauté cible) ; mais aussi de prêter attention à la provenance de la donnée et en fonction de ce critère, décider d'un niveau de confiance ; ou encore enrichir le contexte des données par l'ajout de métadonnées descriptives (taille , date , provenance etc…). En termes de métier, on retrouve bien ici certaines problématiques auxquelles l'archiviste/records manager peut-être confronté. Les aspects métier du data manager sont aussi assez proches des compétences que peut déployer un archiviste : mener une étude de cadrage pour réaliser une cartographie de l'information (savoir ou l'information se trouve, en

[112] DENOIX, Antoine, 2018. Big Data, Smart Data, Stupid Data... : Comment (vraiment) valoriser vos données. Dunod. ISBN 978-2-10-077351-0, page 21

Droits d'auteur réservés. OU (cc) BY-NC-ND

somme) ; veiller à mettre en place des référentiels : il s'agit ici de comprendre finement les process métiers afin de créer des tableaux qui recensent les données et les métadonnées. Data steward, data manager ou records manager ont finalement une commune mission : « préparer le terrain » aux informaticiens et *data scientist* pour que ceux-ci analysent les données ; mais aussi encadrer la pratique de ces derniers notamment en s'assurant de la conformité de rétention et d'utilisation de ces données dans les cadres juridiques et réglementaires (notamment par rapport aux traitements des données à caractères personnels) :

> Il vous faut dans vos équipes quelqu'un capable de répondre à tous moments à ces questions : De quelle data s'agit-il ? D'où vient-elle ? Ai je le droit de l'utiliser ?[113]

Ce quelqu'un, c'est le data manager ou steward, mais nous affirmons ici que l'archiviste formé aux techniques de records management à tout-à-fait son rôle à jouer également, ne serait – ce que dans son intervention en amont des processus métier de création de l'information et dans tout son cycle de vie et sa connaissance des normes réglementaires des organisations et ISO (on pense évidemment aux normes 15489 ou 30300). L'archiviste/records manager peut aussi apprendre beaucoup des méthodes des métiers de la donnée, afin d'enrichir sa propre approche (nous pensons ici surtout, en terme technique). Dans un article de recherche, l'informaticien Gabriel David va même plus loin en pointant les similarités entre les approches métiers de l'ingénieur spécialisé dans la conception de *data warehouse*[114] et l'archiviste qui analyse un SI pour définir une politique d'archivage pour ce système :

> *here is a parallel in the attitude of a data warehouse designer approaching a database-centred operational information system (IS) to specify a data warehouse (DW) and an archivist analysing a document-centred organizational IS to specify an archiving policy and system.*[115]

En effet, l'archiviste comme l'ingénieur des données doivent comprendre la cartographie du SI de l'organisation, agréger des données de différents outils,

---

[113] *Ibidem,* page 87

[114]

[115] DAVID, Gabriel, 2007. Data warehouses in the path from databases to archives. In : International Workshop on Database Preservation. 2007, page 1

Droits d'auteur réservés. OU (cc) BY-NC-ND

comprendre les logiques métiers englobés dans différents process, établir des plans de classification ; mais aussi éliminer les documents ou données jugées superflus et s'assurer de l'intégrité des documents ou données sélectionnées puis faciliter leur exploitation par des utilisateurs déterminés. En outre, le chercheur propose d'appliquer les technologies des Data Warehouse pour la conservation des documents numériques estimés capitaux pour les organisations. Un Data Warehouse consacré à ce type de documents permettrait ainsi d'isoler l'information essentielle à conserver de la masse d'information[116] d'une part, mais aussi de se baser sur une technologie pérenne et scalable. Nous avons donc la une proposition d'architecture novatrice pour réaliser l'archivage des documents et des données, qui en outre est consciente des problématiques d'archivage que soulève le Big Data. Il convient également de noter que ce modèle pour l'archivage essaye de s'inscrire dans des standards du domaine (par exemple, l'Open Archives Initiative). De plus, l'implémentation du modèle d'archivage basée sur le Data Warehouse se produit bien selon une approche archivistique, selon le point de vue fonctionnel de l'archiviste et non simplement de l'informaticien:

> *The involvement of archivists in this task is essential as it goes through the phases of appraisal, selection, elimination and the automated part of description, typical of the archivist process. The system must support the addition of possible bits of manual description*[117]

Cet archivage basé sur la technologie de Data Warehouse repose sur le principe de la migration des supports, dans le sens de l'OAIS, soit la recopie du contenu d'un support vers un autre support de même type. Le modèle OAIS nous semble en réalité incontournable pour réfléchir à une conservation des données massives sur le long terme, et cela indépendamment de tous types de technologies.

VERS UNE CONSERVATION DES MEGADONNEES

Le *Framework* OAIS

Depuis 2012, l'université Polytechnique de Lausanne et son partenaire européen portent et développent les projets des Times Machine : on parle de Venice Time Machine,

[116] *Ibidem,* page 3

[117] *Ibidem,* page 4

Droits d'auteur réservés. OU (cc) BY-NC-ND

Amsterdam Time Machine, Paris Time Machine, Budapest Time Machine ou encore de Jerusalem Time Machine. Le but est de modéliser l'évolution des tissus urbains au fil de l'histoire, afin de proposer une monumentale carte numérique du passé. Les chercheurs et ingénieurs de Lausane ont ainsi travaillé avec les archives d'état de Venise : 80 km de document en l'occurrence, regroupant 1000 ans d'histoire, avec des forêts de cartes (gigantesque ensemble des cartes de la ville) : c'est en cela que réside véritablement des big data du passé, dans d'une part la quantité du matériel, mais aussi sa variété ; variété qui est le fruit d'une temporalité complexe (des archives qui s'étendent sur plusieurs centaines d'années ; donc diversité linguistique/culturelle), mais aussi en terme de types de documents (contrat de notaire, déclaration commerciale etc …). Le but du projet a été de mettre en place un grand chantier de numérisation de ces archives, afin de pouvoir par exemple trouver facilement des infirmations sur un vénitien particulier en tapant son nom dans un moteur de recherche : en sortie, il s'agit d'obtenir un réseau d'information, après parcours d'une grande masse de documents hétérogènes. Pour ce faire, les ingénieurs et chercheurs de Lausanne ont appliqué des stratégies de *data mining* ou *words spotting* pour repérer des mots identiques (cela s'avère tout à fait utile pour les manuscrits anciens, qui regorgent de « mots - images »), et créer un réseau de formes – mots. Le projet permet de montrer un réseau qui se crée, à priori invisible, mais qui existe : la numérisation apporte ici du sens à la donnée, en la replaçant dans son contexte relationnel. La numérisation permet d'apporter du sens à la donnée, un sens supplémentaire, grâce aux relations qu'elle met en exergue. Mais plus loin encore, le projet de la Time Machine propose de situer ce nuage, ce réseau dans une dimension spatiale et temporelle. Il s'agit de superposer différentes couches topographiques (par exemple, selon les années 2016 – 1800 – 1500) , exercice rendu possible par l'identification de points constants ; soit des points d'ancrages qui permettent les translations temporelles. Les outils mis en place permettent d'obtenir des représentations synchroniques de la ville et de ses activités, mais aussi d'obtenir une représentation diachronique. Le but final est donc bien de développer une sorte de Google Maps du passé. Pourtant, dans la présentation de ce projet par Frederic Kaplan[118], le porteur du projet semble ne pas prendre en compte (du moins, à l'heure ou ces lignes sont écrites) un « impensée » de taille : celui de la conservation de ces Big Data sur le long terme. Les Venice Times Machine ne sont pas, paradoxalement, encadrées par une véritable politique d'archivage.

[118] ENS PARIS, [sans date]. Digit_Hum 2017 [en ligne]. [Consulté le 26 décembre 2017]. Disponible à l'adresse : https://transfers.huma-num.fr/digithum/atelier/2017/

Droits d'auteur réservés. OU (cc) BY-NC-ND

Pourtant, il est tout à fait possible de mener une politique d'archivage pour s'assurer de la pérennité des données numériques, surtout lorsqu'il est question des données de la recherche. La méthode probablement la plus adéquate dans ce contexte est transcrite dans la norme ISO 14 721, dite norme OAIS (Open Archival Information system). La présente norme établie une différence claire entre le concept d'information et de données : l'information est relative à tout connaissance que l'on peut échanger, tandis que la donnée est le conteneur, ou le support, qui est porteur d'une information extractible par des moyens humains ou technologiques[119]. Or, comme nous l'avons mentionné plus haut, le rôle du Big Data est précisément de produire de l'information – une information « perdue » pour ainsi dire, dans une masse de données reliées entre elles. C'est même dans ces relations, parfois ténues, que se trouve l'information à trouver. Mais pour s'assurer de la pérennité de cette information, il faut avant tout s'assurer de préserver son support, et donc le format de données – soit du fichier binaire, mais aussi du support physique stockant les bits. C'est ce que propose le modèle OAIS : un ensemble de fonctionnalités répondant aux besoins de conservations des documents et données pour l'archivage numérique, et cela pour tous les domaines, à tous secteurs d'activités, à tous types d'informations numériques »[120]. Il est aussi intéressant de remarquer pour notre sujet, que la norme OAIS a été construite par le CCSDS (*Consultative Committee for Space Data System*), afin de faciliter le développement et la maintenance de systèmes d'informations des organisations spatiales. Organisations comme on le sait, également singulièrement confrontées à la problématique des mégadonnées. En France par exemple, le CNES (Centre National d'Etudes Spatiales), a participé à la rédaction et incrémenté le modèle OAIS dans sa politique d'archivage des données spatiales[121]. Le centre d'études a bien à faire à de « très grand volume de données » [122] , produites à flux tendu au fil des missions par différents instruments de mesures et pendant plusieurs années. Ce grand volume de données à besoin d'être conservé pour les besoins de la recherche : il doit rester possible de traiter ces données 20 ans après leur recueillement. Au CNES, les responsables de la politique des gestions des données ont mis en place le Service de Transfert et d'Archivage des Fichiers (STAF)[123], qui respecte le principe de ne pas avoir

[119] BANAT-BERGER, Françoise, DUPLOUY, Laurent et HUC, Claude, 2009. L'archivage numérique à long terme: les débuts de la maturité ? Paris : la Documentation française Direction des archives de France. Manuels et guides pratiques. ISBN 978-2-11-006942-9. Z699, page 12

[120] *Ibidem*, page 41

[121] *Ibidem,* page 231

[122] *Ibidem*, page 220

[123] *Ibidem*, page 223

Droits d'auteur réservés. OU (cc) BY-NC-ND

de forte dépendance au système de stockage afin de s'assurer de la lisibilité des données sur le long terme. En cela, le STAF répond bien à la recommandation OAIS, qui préconise d'adopter un système d'archivage le plus ouvert possible.

> En matière de stockage, nous considérons que jusqu'au niveau du pétaoctet, les besoins de préservation physique des fichiers sont résolus avec un niveau de fiabilité satisfaisant.[124]

L'organisation spatiale est donc confiante dans sa gestion et sa conservation des très grands volumes de données (jusqu'au pétaoctet du moins). Le modèle OAIS donne un cadre de réflexion pertinent pour répondre à la problématique de l'archivage des mégadonnées. Les chercheurs a la tête du projet se montre à contrario plus sceptiques devant la problématique des ontologies à évolutions rapides[125]. En effet, l'enrichissement manuel des référentiels de métadonnées est particulièrement complexe lorsque l'ontologie des données évolue souvent et rapidement. Tâche d'autant plus complexe lorsque les données sont produites massivement et rapidement. Ainsi, voyons plus précisément comment l'incrémenter concrètement dans le cadre spécifique d'une problématique de pérennisation des big data.

Un modèle fonctionnel pour l'archivage des Big Data

Une récente étude[126] de deux informaticiens (*Laboratory of Architecture and High Performance Computing* ; São Paulo) expose avec rigueur et scientificité un modèle, très inspiré de l ' OAIS, pour l'archivage des données créees et/ou stockées par les organisations. La notion d'archivage est ici bien comprise en terme archivistique : l'article ne confond donc pas le concept d'archivage avec celui de stockage, mais bien comme méthode et ensemble de process pour assurer la pérennité de l'information. Pour preuve, nous voyons que les auteurs emploient des termes qui renvoient directement au vocabulaire de l'archiviste dans le titre même de leur publication : « *long term archiving* », « *préservation* », « *retrieval* ». Des notions clés que l'on retrouve dans les problématiques d'archivage, dans un objectif de pérennisation de l'information. Les informaticiens adoptent donc ici une véritable

[124] *Ibidem*

[125] *Ibidem*

[126] VIANA, P. et SATO, L., 2014. A Proposal for a Reference Architecture for Long-Term Archiving, Preservation, and Retrieval of Big Data. In : 2014 IEEE 13th International Conference on Trust, Security and Privacy in Computing and Communications. septembre 2014. pp. 622-629.

Droits d'auteur réservés. OU (cc) BY-NC-ND

approche archivistique : partant du principe qu'il y a une expansion toujours croissante des données (volumétrie complexe à gérer), concomitante aux demandes de préservation de ces données (pour des raisons de conformités/juridiques – *E-discovery* aux Etats Unis ou GDPR en Europe pour ne pas les citer), et que ces données sont à la fois de types structurées et non-structurées (Big Data), qu'il devient urgent de proposer « *a reference architecture for the long term archiving , preservation and retrieval of Big Data* ».

Pour déployer le nouveau modèle architectural (et encore bien expérimental) les chercheurs vont s'inspirer d'un modèle archivistique bien connu des professionnels de l'information, en l'occurrence l'approche portée par la norme OAIS. De fait, la notion de préservation (« *data preservation* ») est bien entendue comme une donnée numérique, qui doit être lisible, compréhensible et interprétable nonobstant le passage d'un grand nombre d'années, et indépendamment de son support d'origine. Il faut s'assurer donc, de l'intégrité des bits ; mais aussi de la compréhension globale de l'information ; que cette information en somme, demeure *logique* et compréhensible pour l'entendement humain ou au moins pour une communauté visée (tel que cela est proposé dans l'OAIS).

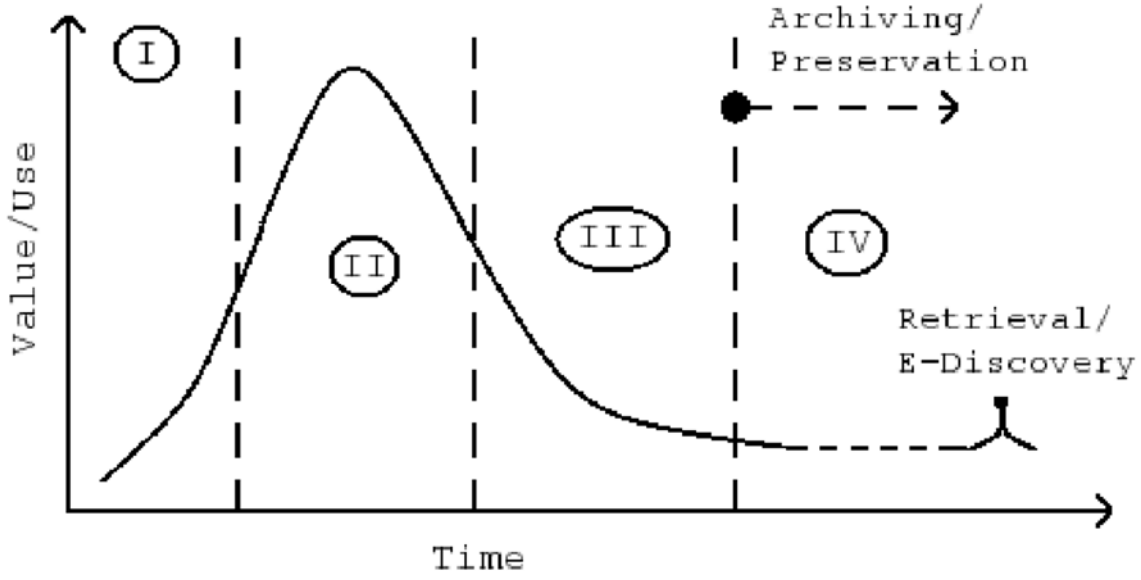

Figure 4: représentation du cycle de vie de l'information.

Droits d'auteur réservés. OU (cc) BY-NC-ND

L'image ci dessus[127] propose sous forme de graphique une représentation du cycle de vie de l'information : en abscisse : la valeur temporelle et en ordonnée : la valeur et l'usage de la donnée. La courbe montre la corrélation entre temps et usage/valeur de la donnée. Une interprétation en étape nous permet de mieux comprendre leur vision du cycle de vie de la donnée. Etape une : la donnée est créée et sa valeur augmente : son support est alors un support chaud, comme un serveur par exemple, ou les données sont massivement manipulées et traitées. Etape deux : C'est le moment ou la donnée est le plus traitée, ou il s'agit d'extraire un maximum d'information/corrélation de cette donnée (traitement analytique par des outils de type Big Data dans notre cas). En étape trois : la data devient « ancienne». Sa valeur et son usage décroissent (c'est-à-dire que l'on accède moins à cette donnée). Les données concernées peuvent alors être transférées vers un autre support de stockage, moins « chaud ». Enfin la dernière étape du cycle de vie, entre l'étape trois et quatre, phase où l'on décide si la donnée est conservée pour l'archivage ou alors supprimée. Ainsi pour les chercheurs, il est bien question d'inscrire leur modèle d'architecture dans ce moment de l'archivage : *If the data are archived, later blips may occur when the data are searched for, temporarily increasing their value* »[128]. Déployer l'archivage des big data à l'aide d'une méthode d'archivage basée sur l' OAIS est considéré comme une solution viable pour les auteurs. Le modèle fonctionnel de référence proposé dans l'étude consiste en un *framework* très proche de l'OAIS, appliqué à l'archivage des big data : « *the design of our référence architecture used as a guideline, which proposes a framework for the design of reference architecture*» : une architecture qui, selon les auteurs, peut supporter les modèles de BDD relationnel comme non – relationnel (NoSQL), prendre en compte des données structurées comme non-structurées. De plus, il est intéressant de noter que les auteurs parlent de paquet AIP pour décrire les paquets d'archives versés et/ou demandés – concepts directement hérités de l'OAIS et ici largement repris dans le *framework* d'archivage des Big Data. De même, lorsque que les auteurs parlent de d'«objet de préservation » ; il s'agit, comparativement dans la norme OAIS, d'une information de préservation (soit, une information de structure ou une information sémantique). La pérennisation est au cœur du modèle d'architecture proposé, tout

[127] VIANA, P. et SATO, L., 2014. A Proposal for a Reference Architecture for Long-Term Archiving, Preservation, and Retrieval of Big Data. In : 2014 IEEE 13th International Conference on Trust, Security and Privacy in Computing and Communications. septembre 2014. pp. 622-629., page 623

[128] *Ibidem*,

Droits d'auteur réservés. OU (cc) BY-NC-ND

comme l'exigence d'interopérabilité, qui est aussi mise en avant par la couche SOA (*service oriented architecture*) et vient introduire un aspect important en urbanisation (la notion de cohérence forte/couplage faible). Ainsi, la couche SOA permet à des applications extérieures au SI de soumettre ou retrouver des paquets sous forme d'archives OAIS. En outre, c'est une architecture « media - Independent » : c'est à dire qu'elle peut intégrer n'importe quel type de support (disque, baie de serveur, cloud etc.). Le modèle des chercheurs a au final bien cela d'innovant, en ce qu'il s'attache au stockage des données non structurées comme à tout flux de données en général.

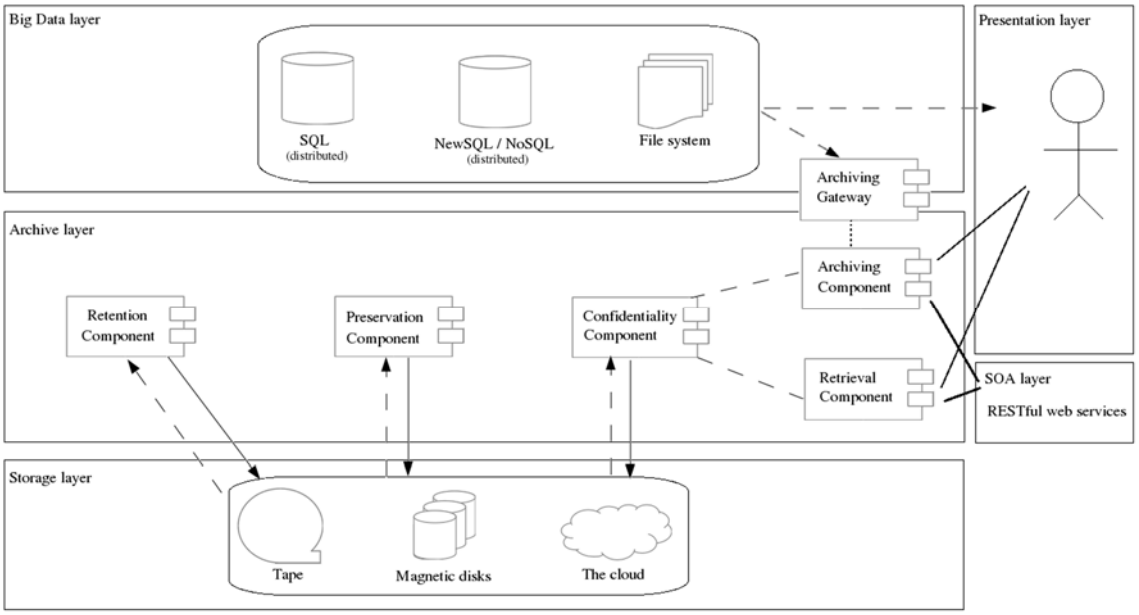

Figure 5: vue globale de l'architecture de réference proposée

L'architecture ici élaborée par les chercheurs à son lexique propre : couche (*layer*), élément (*component*), interface (*interface*), déclinés en plusieurs attributs. Ainsi, la première couche est la couche « Big Data », où sont situés originellement les données structurées (BDD relationnelles – SQL DB) et non-structurées (BDD non relationnelles – DB NoSQL, ou *File System*). C'est l'espace qui contient les données chaudes, stockées dans des espaces fortement sollicités (phase une, deux et trois du cycle de vie). C'est bien entendu une couche abstraite, qui dépend de chaque organisation. La deuxième couche, *archive layer*, est la couche des *components*, qui comporte : *archiving*, *retrieving*, *retention*, *preservation* et *confidentiality.* : les actions d'archivage, de préservation, et de communicabilité des données ont lieu ici.

Droits d'auteur réservés. OU (cc) BY-NC-ND

Les données non structurées et structurées sont empaquetées dans un paquet SIRF[129], qui contient un objet de préservation (soit les données brutes à conserver, plus les métadonnées qui lui sont associées). Ici, l'objet est modélisé sous la forme d'un OAIS AIP (archival information package). Les composants de cette couche ont été établis en s'inspirant des exigences qui caractérisent aujourd'hui une politique d'archivage sur le long terme. Le *storage layer* est une autre couche abstraite, dont l'implémentation matérielle dépend de l'organisation : les *devices* peuvent être des disques magnétiques, des cassettes, ou encore une solution Cloud. Soit dit en passant, les auteurs se montrent davantage critique envers l'archivage sur le Cloud (qui s'apparente plus à une solution de stockage qu'à une réelle stratégie de préservation sur le long terme) ; bien que des auteurs proposent une LDPaaS (« *long-term digital preservation as as service* »), dans la lignée des solutions XaaS. Même si les auteurs ne rejettent pas une solution d'archivage basée sur le Cloud, ils estiment cependant à juste titre que celui-ci ne peut être une solution viable seulement s'il est pensé dans le cadre d'un modèle logique d'archivage cohérent, et non comme simple solution de stockage (qui risque en outre de ne pas satisfaire aux critères de conformité). Toutefois, il est bon de noter que le standard SIRF est compatible avec la solution d'archivage sur le Cloud porté par OpenStack, OpenStack Swift : le principal avantage de cette solution étant que contrairement au stockage sur disque, le document ou la donnée archivée dans la solution StackSwift et empaquetée en SIRF est toujours à disposition. L'Object archivé n'a plus à être considéré comme « froid », il est possible d'y accéder très rapidement, et se trouve donc davantage valorisé[130]. Le cloud OpenStackSwift peut également se coupler avec une solution Cloud compatible OAIS du nom de PDS Cloud[131]. Enfin la dernière couche : la *Presentation Layer* qui correspond à l'utilisateur. C'est lui qui définit les règles d'archivage, qui a à charge de manuellement déterminer quelles données sont à archiver ou encore requêter des données demandées. Cet utilisateur

[129] SIRF (Self-contained Information Rentention Fomat), format d'empaquetage sérialisé pour le stockage sur disque, mais surtout tournée vers l'archivage sur le Cloud. C'est un format maintenu par la SNIA (Storage Networking Industry Association).

[130] DICKINSON, John, [sans date]. Future trends in archival storage. [en ligne]. [Consulté le 26 décembre 2017]. Disponible à l'adresse : https://www.swiftstack.com/blog/2016/09/28/future-trends-archival-storage/

[131] « PDS Cloud is an OAIS-based preservation-aware storage service employing multiple heterogeneous cloud providers » [RABINOVICI-COHEN, Simona, MARBERG, John, NAGIN, Kenneth et PEASE, David, 2013. PDS Cloud: Long term digital preservation in the cloud. In : Cloud Engineering (IC2E), 2013 IEEE International Conference on. IEEE. 2013. pp. 38–45, page 38], Notons de plus que cette solution est une composante du projet européen ENSURE : PDS Cloud y est utilisé pour préserver les données médicales et financières. PDS Cloud est estimé « OAIS - based preservation aware » car il convient à la préservation logique de l'objet (conservation des bits), par sa scalabilité et le fait qu'il n'impose pas de dépendance forte à son environnement technologique.

Droits d'auteur réservés. OU (cc) BY-NC-ND

peut accéder directement à l'archive via les couches SOA et le système des API (ici RESTful), qui exposent les données/documents.

Pour résumer, les auteurs proposent donc ici un *framework* dans lequel chacun peut déployer sa propre solution : ils ne proposent pas de solution outil en particulier (SAE, GED), et évoquent de façon abstraite l'environnement technique. Ce modèle d'architecture, très inspiré de l'OAIS, prend en compte les problématiques de l'archivage numérique, dans ces deux grands aspects : la pérennisation de la donnée et du document, en soulignant l'importance de la contextualisation de la donnée archivée (utilisation de l'AIP), mais aussi sa conservation (des bits) pour satisfaire à des exigences légales[132]. Mises en place par des informaticiens, ce modèle prend cependant bien en compte les problématiques métiers de l'archivistique. Contrairement à l'approche purement informatique (qui comme on l'a vu plus haut, à tendance à trop privilégier une vision par l'outil), les chercheurs ici réfléchissent d'abord au modèle fonctionnel dans lequel déployer un archivage des Big Data. On peut aujourd'hui retrouver un schéma similaire dans quelques projets bien concrets.

Des projets concrets : La Solution VITAM, e- ARK, ENSURE project.

Le projet ENSURE (*Enabling kNowledge Sustainability Usability and Recovery for Economic value*) a été évoqué plus haut comme un projet concret d'archivage des Big Data. Techniquement, le projet s'inscrit dans l'archivage des documents et données grâce à la technologie du Cloud ; fonctionnellement, il s'inscrit dans une réelle démarche d'archivage pérenne des données non structurées et utilise notamment le modèle OAIS. C'est un projet initié par la commission européenne, porté par CORDIS (Service Communautaire d'information sur la Recherche et le Développement), coordonnée par IBM et soutenue par divers organisme privés et publics (Atos, JRC Capital Management, mais aussi les universités de Porto et de Cranfield). L'objet du projet est de s'assurer de l'utilisabilité sur le long terme de données commerciales pertinentes, produites et conservées par les organisations, notamment dans les domaines de la santé, de la

Droits d'auteur réservés. OU (cc) BY-NC-ND

finance ou de l'aérospatial[133]. Pour les porteurs du projet, le cloud à le double avantage d'être hautement scalable et de constituer une sorte « d'archivage as a service ». En outre, la solution basée sur PDS Cloud permet une gestion du cycle de vie des données et documents (une couche métier est intégrée pour satisfaire au monitoring du cycle de vie) ; et se veut également ouverte (car elle n'impose pas de dépendance technologique lorsqu' il est question de changer de support),et également favorable à l'ajout de métadonnées (évolution des ontologies). Néanmoins, il convient de noter que ce projet n'est à l'heure actuelle plus véritablement maintenue : son site web n'est en effet plus accessible et on aura bien difficulté à trouver une actualité sur le projet ces dernières années. Néanmoins, d'autres projets d'archivage des big data ont vu le jour et rencontrés un réel succès. Parmi eux, le projet européen, E – ARK (European Archival Records and Knowledge). Lancé en 2013 sous le signe de l'innovation (projet CIP – *Competitivness and Innovation Framework Programme*), E-ARK est un projet d'envergure européenne. C'est en quelque sorte, une archive pour l'Europe : E-ARK se place comme pilote pour les services d'archives dont la mission est d'assurer l'intégrité, la conservation et l'utilisation des données et documents à travers le temps[134]. Selon les termes des collaborateurs, E-ARK se definit comme

> *a multinational big data research project that aims to improve the methods and technologies of digital archiving, in order to achieve consistency on a Europe-wide scale.*[135]

Big Data en effet, car ici les données et documents à archiver représentent un volume de données aussi gigantesque que varié (le projet prévoit d'archiver tout type de données, depuis n'importe quelles sources). Du point de vue de l'archiviste, E-ARK s'inscrit dans le modèle OAIS (on retrouve de nouveau les instances de SIP – AIP – DIP): la plateforme d'archivage est construite sur un socle technologique très ouvert et favorisant l'interopérabilité entre les systèmes d'information. La norme OAIS se trouve implémentée à travers des applications open source comme RODA

[133] Enabling kNowledge Sustainability Usability and Recovery for Economic value | Projects | FP7-ICT, [sans date]. CORDIS | European Commission [en ligne]. [Consulté le 30 juillet 2018]. Disponible à l'adresse : https://cordis.europa.eu/project/rcn/98002_en.htmlFP7-ICT,ENSURE,Ensuring

[134] BILLENESS, Clive, [sans data]. An E-ARK Overview. [en ligne]. [Consulté le 1 Aout 2018] . Disponible à l'adresse : http://www.eark-project.com, page 4

[135] SENN, Alina, 2016. Enabling 21st century archives. Paths and traces. In : E-ARK Conference 6 december 2016, Budapest.[en ligne]. [consulté le 1 aout 2018]. Disponible à l'adresse : http://www.eark-project.com/resources/conference-presentations/finconfpres/70-asennfinconf-1/file

Droits d'auteur réservés. OU (cc) BY-NC-ND

(Repository of Authentic Digital Record) ou EPP (ESSArch Preservation Platform). La plateforme propose une couche fonctionnelle pour l'archiviste, c'est-à-dire une interface utilisateur qui permet à un administrateur archiviste de gérer le cycle de vie des données et des documents selon la norme OAIS[136], mais supporte aussi d'autres standards comme PREMIS (PREservation Metadata), METS (Metadata Encoding and Transmission Standard) ou encore le Dublin Core. En outre, en fonction du type du contenu binaire à préserver, E-ARK prévoit les briques logicielles reposants sur certains standards internationaux d'archivage, adaptées pour assurer la conservation des bits et leur lisibilité sur le long terme (SIARD2, SMURF, Geo- Data etc…)[137]. De façon plus générale, ces plateformes recouvrent les fonctionnalités attendues dans un SAE, y compris pour s'assurer de la bonne gestion de l'intégrité et de l'authenticité des documents et données (identification et vérification des formats d'entrées, calcul de l'empreinte des documents ou données et possibilité de paramétrer un recalcul selon des intervalles de temps déterminés, horodatage, fonctionnalités de *reporting*, système pleinement auditable etc …). D'un autre côté, l'architecture technique proposée répond également aux enjeux de la scalabilité et s'appuie ainsi sur une brique Hadoop:

> *The backend system of the IPRIP is built on top of an infrastructure that is based on the Apache Hadoop framework*[138]

En effet, comme nous l'avons déjà décrit plus haut, la technologie Hadoop convient singulièrement à l'archivage des données massives car il est possible de « rajouter » de la puissance de calcul (CPU) et de la mémoire de stockage en ajoutant des nouveaux nœuds de calcul (des serveurs à bas coûts) très facilement. Le « redimensionnement » de l'espace de stockage, essentiel lorsqu'une grande quantité de données est ingérée, est réalisé grâce au système HDFS (Hadoop distribued file System, dont les grands principes et les avantages pour l'archivage des données ont aussi été traité plus haut). Dans le projet E-ARK, une base de

[136] Le projet s'inscrivant dans une démarche OpenSource, Il possible de télécharger les fichiers nécéssaires à l'installation de la plateforme ESSArch depuis le dépôt github suivante : https://github.com/ESSolutions/ESSArch_EPP

[137] AAS, Kuldar et AAS, Kuldar, [sans date]. Coping with the data explosion. . pp. 19.[en ligne]. [Consulté le 1 aout 2018]. Disponible à l'adresse : http://www.eark-project.com/resources/conference-presentations/presents2016/62-ica2016-1/file

[138] SCHLARB, Sven , 2016. The Use of Big Data techniques for Digital Archiving. Austrian Institute of Technology [en ligne]. [Consulté le 26 décembre 2017]. Disponible à l'adresse : https://www.bigdata.cam.ac.uk/files/our-digital-future-2016/our-digital-future-2016-slides/sven-schlarb-e-ark-15-march-2016.

Droits d'auteur réservés. OU (cc) BY-NC-ND

données non relationnelles (Hbase[139]) et le modèle MapReduce sont aussi mis en place. L'utilisateur peut rechercher l'archive dont il a besoin via une interface web (E-ARK WEB) : chaque requête, selon une méthode de recherche plein texte ou par mot clé, interroge le répertoire de stockage et renvoie la donnée recherchée via un protocole HTTP. De plus, E-ARK web supporte l'administration complète du processus structuré par la norme OAIS : construit sur une application front end web légère (Python/Django), trois onglets permettent de gérer la création d'un SIP, la conversion d'un SIP vers un AIP et d'un AIP vers un DIP.

Enfin, le projet VITAM, porté par le ministère de la Culture, s'inscrit également dans la lignée des projets cités ci-dessus, mais jouit d'une popularité bien meilleure parmi la communauté des archivistes, et en particulier, de façon logique, des archivistes français. VITAM est un projet libre, dont le socle technique et fonctionnelle est une plateforme faisant office de SAE pour archiver sur le long terme la masse des documents et données de plusieurs Ministères (on compte pour le moment le MAE, le Ministère de la Justice et celui de l'Armée). Le projet est piloté par la DINSIC (Direction Interministérielle du Numérique et du Système d'Information et de Communication de l'État) et le CIAF (Comité Interministériel aux Archives de France), et compte comme partenaires des noms bien connus comme le CINES (Centre Informatique National de l'Enseignement Supérieur)[140]. D'après le site web du projet[141], la solution logicielle (open source) VITAM répond à trois objectifs : permettre la conservation pérenne d'un volume d'archives numériques pouvant correspondre à un volume Big Data ; l'implémentation de la solution logicielle dans trois plateformes, une pour chaque Ministère concerné (Saphir, Adamant et Archipel) et enfin, la réemployabilité des briques logicielles du projet, afin d'étendre l'archivage au-delà des trois Ministères présentement impliqués. VITAM est une solution « sur mesure », hautement modulable et qui s'inscrit dans une certaine philosophie de l'Agilité. Il est bon de noter que la solution possède son « propre » manifeste Agile :

[139] Hbase est la base de données non relationnelle présente par défaut dans l'environnement Hadoop. En tant que base non relationnelle, son utilisation comporte des avantages lorsqu'il est question de très grand volume de données.

[140] Le CINES est depuis 2004 l'organisation chargée de l'archivage des données et documents numériques créés par l'Enseignement supérieur et la Recherche. Le CINES adresse tout type d'archivage : sur le long terme comme à moyen terme, pour des acteurs de l'administration, des bibliothèques ou de la communauté scientifiques.

[141] Focus archivistes · Vitam, [sans date]. [en ligne]. [Consulté le 13 août 2018]. Disponible à l'adresse : http://www.programmevitam.fr/pages/presentation/pres_archivistes/

Droits d'auteur réservés. OU (cc) BY-NC-ND

> Pour les services de l'Administration. Afin de satisfaire les enjeux d'accès dans le temps à leurs informations numériques, Vitam propose une solution logicielle libre d'archivage, évolutive, simple et facilement interfaçable, qui permet la gestion unitaire et sécurisée de milliards d'objets et vise son adoption par le plus grand nombre d'acteurs publics[142].

De fait, les porteurs du projet ont conscience qu'une solution d'archivage particulière, open source, est préférable aujourd'hui à un socle logiciel monté par un éditeur, coûteux, complexe à maintenir sur le temps long et dont les fonctionnalités comme le « moteur » risquent de ne pas parfaitement correspondre aux besoins (d'autant que celui-ci est amener à varier selon la cible – en l'occurrence ici, un contexte d'archivage singulier en fonction des Ministères) :

> Cette vision permet ainsi la réutilisation plus large, tout en assurant la réalisation d'un outil générique intégrable selon les besoins d'acteurs variés dans leur système d'information.[143]

De fait, VITAM est bien pensé comme une brique d'infrastructure, afin de porter les fonctionnalités permettant l'archivage numérique sur le long terme. De plus, pour intégrer pleinement les enjeux de l'archivage numérique, la solution VITAM est conforme aux normes et standards : SEDA est utilisé pour standardiser le transfert des données entre les applications métiers et la plateforme. VITAM est également certifié ISO NF Z42-13. On retrouve aussi, de nouveau, le cadre de la norme OAIS, qui structure fonctionnellement l'architecture de la plateforme des Ministères[144]. VITAM présente encore des fonctionnalités de gestion des DUA, de « désherbage » des documents ou données inutiles à la conservation, d'enrichissement et modification des métadonnées et de migration vers d'autre socle SAE. En cela, la plateforme, en plus de sa grande modularité, répond parfaitement aux critères fonctionnels de ce que l'on attend communément d'une SAE[145]. En outre, l'outil VITAM entre particulièrement en échos avec notre propos en ce qu'il se confronte

---

[142] *Ibidem*

[143] *vitam: Digital Archives Management System developped by French government/Programme interministériel archives numériques ; core system*, 2018. [en ligne]. Java. Programme Vitam. [Consulté le 13 août 2018]. Disponible à l'adresse : https://github.com/ProgrammeVitam/vitam

[144] C'est à dire, selon les 6 fonctionnalités nécessaires à un système d'archivage électronique énoncées par la norme : Entrée, Stockage, Gestion des Données, Administration, Préservation et Accès.

[145] BÉCHARD, Lorène, FUENTES HASHIMOTO, Lourdes et VASSEUR, Édouard, 2014. Les archives électroniques. Paris : Association des archivistes français. Les Petits guides des archives. ISBN 978-2-900175-06-4. CD974.4, page 12

Droits d'auteur réservés. OU (cc) BY-NC-ND

à l'archivage d'une masse de données relevant des Big Data (on parle ici de plusieurs millions ou milliards de données structurées comme non structurées)[146]. Les expérimentations et choix techniques réalisés se sont basés sur l'objectif *in fine* de gérer des milliards d'objets numériques très disparates (données structurées comme non structurées)[147]. Ainsi, le cœur technique du back – office de VITAM a pour matrice 4 modules, répondant à la fois aux exigences techniques d'un SAE classique, mais aussi d'un système de type Big Data[148]. Le moteur de traitement a ainsi été conçu pour traiter de « grandes volumétries d'archives »[149] : il peut s'agir de transférer plusieurs milliards d'objets numériques vers la solution VITAM, et lors de ce transfert, de gérer un certain nombre d'opérations (conversion de format, calcul de l'empreinte numérique etc …). Le moteur de traitement, en « collaboration » avec des programmes greffons (on peut ici imaginer des fonctionnalités d'analyses sémantiques) va permettre de déployer ces actions sur de très gros volumes d'archives. En outre, afin d'optimiser l'indexation et la recherche des métadonnées des archives, les concepteurs de VITAM ont privilégié des bases de données orientées non relationnelles : MongoDB couplé à ElasticSearch. Une preuve du concept a été réalisée, ou

> 200 millions d'items ont pu être pris en charge sur un serveur de type 2 cœurs,/16 Go avec des temps de réponse de recherche de l'ordre de la seconde sur des requêtes simples. Qui plus est la linéarité par ajout de serveurs a été excellente, 8 serveurs amenant à des temps similaires pour 1,5 milliards d'archives.[150]

Nous avons déjà évoqué les avantages d'une base de données orientée document (MongoDB fait partie de ce type de base non relationnelle) dans le cas de gros volumes composés de données non structurées très différentes. Le stockage des binaires n'est cependant pas directement réalisé dans la base de données : le projet a prévu une offre de stockage pour la conservation des bits. VITAM a opté pour un

---

[146] Focus archivistes · Vitam, [sans date]. [en ligne]. [Consulté le 13 août 2018]. Disponible à l'adresse : http://www.programmevitam.fr/pages/presentation/pres_archivistes/

[147] *VITAM_Presentation_solution_logicielle.pdf*, [sans date]. [en ligne]. [Consulté le 13 août 2018]. Disponible à l'adresse : http://www.programmevitam.fr/ressources/DocCourante/autres/fonctionnel/VITAM_Presentation_solution_logicielle.pdf , page 19

[148] *Ibidem*, page 18

[149] *Ibidem,* page 18

[150] *Ibidem*, page 19

Droits d'auteur réservés. OU (cc) BY-NC-ND

stockage sur le cloud, en se basant sur l'offre de type stockage - objet de la fondation OpenStack, Swift. En cela, VITAM s'inscrit parfaitement dans le modèle fonctionnel pour l'archivage des Big Data évoqué plus haut (qui mentionnait également Swift comme solution de stockage possible) : privilégier la scalabilité et le rapport coût/espace de stockage, dans une logique de Storage As A Service. Cet aspect est singulièrement important si l'on considère la croissance exponentielle des données et documents numériques à archiver. De plus, la logique du stockage objet répond aux exigences de l'archivistique en termes de conservation et d'intégrité. En effet, les systèmes de stockages objets sont particulièrement adaptés pour le stockage de données non structurées (mails, vidéos, images, document texte etc …) et dont le contenu ne doit pas être modifié. Contrairement au mode de stockage RAID parfois utilisé en matière d'archivage[151], le stockage objet possède l'avantage de proposer une scalabilité bien supérieure (au-delà de plusieurs centaines de térabytes)[152], et cela à un coût bien inférieur (logique du service, ou de stockage à la demande). En outre, Swift favorise la virtualisation de stockage, séparant l'organisation logique et physique des données, qui elle peut changer au cours du temps, mais grâce à la virtualisation, ne pas perturber l'utilisateur. En dernier lieu, un autre avantage du stockage objet par rapport au stockage fichier classique réside dans la logique d'accès aux données stockées : le protocole REST permet en effet à un nombre presque illimité d'utilisateurs de se connecter en même temps et de n'importe où, et avec une vélocité très élevée (quelques millisecondes, en supposant un débit internet acceptable). En conclusion, VITAM est un projet qui illustre bien notre propos d'archivage des Big data, non seulement du point de vue « informaticien » (qui a tendance à prendre seulement en compte les performances au détriment des enjeux métiers, et d'assimiler l'archivage à du simple stockage), mais aussi de l'archiviste. La couche fonctionnelle du projet VITAM intègre véritablement l'aspect métier pour réaliser un projet d'archivage numérique sur le long terme. Cependant, si une vision Big Data s'est véritablement imposer aux porteurs de la solution d'archivage, cela reste à nuancer : Vitam n'intègre en effet aucune couche analytique, dans la mesure qu'« il n'est pas prévu […] de pouvoir

[151] BANAT-BERGER, Françoise, DUPLOUY, Laurent et HUC, Claude, 2009. L'archivage numérique à long terme: les débuts de la maturité ? Paris : la Documentation française Direction des archives de France. Manuels et guides pratiques. ISBN 978-2-11-006942-9. Z699, page 92.

[152] KAPADIA, Amar, RAJANA, Kris et VARMA, Sreedhar, 2015. OpenStack Object Storage (Swift) Essentials. Birmingham, UK : Packt Publishing Limited. ISBN 978-1-78528-359-8, page 25

Droits d'auteur réservés. OU (cc) BY-NC-ND

effectuer des traitements de masse sur le contenu des archives et d'en déduire des analyses statistiques ou d'utiliser des mécanismes d'intelligence artificielle »[153]. L'absence de besoin analytique a donc exclu Hadoop du projet, réduisant l'aspect « Big Data » de VITAM.

[153]http://www.programmevitam.fr/ressources/Doc0.15.1/html/archi/fonctionnelle-archivistes/black-box/orientation-generale.htm

Droits d'auteur réservés. OU (cc) BY-NC-ND

CONCLUSION

« Doit-on en conclure que le métier d'archiviste et celui de gestionnaire de Big Data sont le mème métier ou deux métiers différents ?»[154]. A cette interrogation de Bruno Delmas, le produit de nos recherches nous conduit à apporter une réponse prudente, sinon nuancée. Ces « gestionnaires de Big Data » évoqués dans l'article de l'archiviste renvoient tout d'abord aux informaticiens et *data scientist*, qui sont les premiers, ou du moins les plus sollicités, par ce phénomène. Les sciences informatiques produisent encore une abondante littérature sur le sujet, qui entre en correspondance avec certaines branches bourgeonnantes des technologies de l'information, comme le *deep* et *machine learning* , ou l'intelligence artificielle. Les Big Data sont alors le plus souvent étudiées et exploitées dans une perspective marketing et financière. Les géants du web américains stockent, exploitent et analysent les mégadonnées dans la perspective de mieux connaitre leurs consommateurs, et ainsi, mieux cerner leur besoin. La notion de Big Data de fait, recouvre également un fort aspect anxiogène : le concept à un air de « Big Brother », ou une société contrôlée par quelques puissantes entreprises perdrait peu à peu de sa liberté, dominée par des boites noires algorithmiques possédées par ces mêmes géants des nouvelles technologies. La mathématicienne et *data scientist* Cathy O'Neil a su monter avec force et rigueur le désastre que peut causer une approche des Big Data excluant toute distance critique ou sens éthique, selon une approche ou les programmes informatiques seraient les oracles infaillibles de notre monde actuel.

De fait ; nous serions tentés d'affirmer que l'archiviste, dont le métier a pour objectif de « gérer et organiser l'information dans le temps, quel que soit son support, pour la rendre accessible durablement, bien au-delà de la durée de vie des supports »[155], n'est en aucun cas un gestionnaire de Big Data. Davantage même, que le Big Data, qui est synonyme de triomphe de l'automatisation et de la donnée manipulée en permanence, rend caduque l'activité de l'archiviste, ce dernier ironiquement, « relégué aux archives ». En effet, si les données se trouvent en permanence sollicitées et traitées, alors il n'y a plus de données ou documents « gelés », donc plus d'archives à proprement dites. Pourtant,

[154] DELMAS, Bruno, 2015. Archivistes de 2030: réflexions prospectives. Louvain-la-Neuve : Academia-l'Harmattan., page 135

[155] BÉCHARD, Lorène, FUENTES HASHIMOTO, Lourdes et VASSEUR, Édouard, 2014. Les archives électroniques. Paris : Association des archivistes français. Les Petits guides des archives. ISBN 978-2-900175-06-4. CD974.4, page 8.

Droits d'auteur réservés. OU (cc) BY-NC-ND

depuis quelques années, le sujet parfois apparait sous la plume des membres de la profession, et suscite réflexion de la part des professionnels des sciences de l'information. L'information d'une part, c'est bien cela même l'objectif du Big Data : une masse disparate de données, sans lien entre elles à priori, mais avec des outils adéquates (couplage de l'information et des statistiques mathématiques), qui révèle des signaux producteurs d'information. En outre, le support de cette information est la donnée, plus précisément pour les Big Data, la donnée non structurée, c'est-à-dire des images, des sons ou bien des formats texte. En somme, la matière souvent manipulée par les spécialistes de l'information, qui s'efforce de la gérer et d'en tirer parti. C'est cette même donnée non structurée qui s'accumule de plus en plus dans les organisations privées comme publiques, et qui vient constituer un important vivier Big Data. Ainsi, les métiers issus de la GED ou encore des bibliothèques s'intéressent aux possibilités offertes par le Big Data, et en quoi les données massives participent à l'évolution de leur profession. On y voit un intérêt pour les technologies NoSQL, mieux à même de gérer les données non structurées ou semi – structurées : nous avons donné l'exemple de plateforme de type Data Hub, comme MarkLogic, qui peuvent se substituer à des GED vieillissantes et difficiles à maintenir pour gérer de très grand volume d'information. Dans le domaine des bibliothèques, des initiatives ont vu le jour dans le domaine de l'archivage du web, mais aussi pour offrir aux bibliothécaires des outils de prédiction améliorant l'expérience utilisateur. L'archiviste n'est pas en reste, du vœu même des spécialistes du métier :

> La compréhension des fonctionnalités des systèmes d'information et la modélisation des flux de travail et des flux documentaires, semblent indispensables à l'archiviste de demain pour lui permettre d'identifier et de documenter les documents à archiver,
>
> et
>
> Loin de sonner le déclin de la profession d'archiviste, la société de l'information se présente comme une opportunité à saisir. Le volume et la complexité de l'information sont tels que celui qui est capable de la gérer et de la préserver présente une grande valeur dans notre société du 21e siècle[156].

[156] DELMAS, Bruno, 2015. *Archivistes de 2030: réflexions prospectives*. Louvain-la-Neuve : Academia-l'Harmattan. Page 169

Droits d'auteur réservés. OU (cc) BY-NC-ND

L'archiviste peut apporter son métier aux Big Data. Et le Big Data, réciproquement, permet à l'archiviste de faire évoluer sa profession et de l'enrichir. Il nous apparait essentiel que l'archiviste ait une conscience aiguë de l'architecture des systèmes d'informations. D'une part, car l'archiviste archives les objets numériques produits par ces systèmes d'informations, et d'autre part, car son métier recouvre des enjeux de gouvernance de la donnée. De plus, dans l'écosystème Big Data, nous trouvons des solutions logicielles qui diffèrent des solutions plus classiques : l'archivage des Big Data par conséquent, implique aussi un changement du socle technologique traditionnel fourni par la plupart des SAE. De même, nous voyons des solutions de stockage évoluer non pas nécessairement vers les systèmes RAID classiques, mais davantage vers des modèles reposant sur un modèle Hadoop ou le Cloud. Si l'archivage des Big Data est bien un archivage qui donne la « part belle » aux informaticiens, il ne doit pas être pour autant une affaire exclusivement réservée à ces derniers. L'archiviste doit réussir à « injecter » sa compétence métier dans ces écosystèmes SI qui s'occupent des Big Data. On a pu voir par exemple, que l'approche du records manager était facilitée par des outils d'automatisation (par exemple, pour détecter les données à caractères personnelles, ou gérer les délais de retentions). Pour autant, le records manager n'est pas évincé, et apporte toujours son expertise fonctionnelle. Il en va de même pour la conservation sur le long terme des données massives. L'archivage des mégadonnées ne demande pas simplement une architecture technique robuste et adaptée aux challenges de gérer des pétaoctets d'informations. Il s'agit bien aussi pour l'archiviste de déployer ses modèles métiers, afin de constituer une couche fonctionnelle « archivistique » robuste, qui vient se greffer sur le socle technique. La solution logicielle Vitam est un bon exemple de réalisation d'archivage du Big Data, qui précisément ne se réduit pas à un simple stockage des données d'un point de vue strictement « informaticien ». Vitam est construit sur des composants logiciels à la pointe, généralement utilisés dans les environnements qui gèrent d'importants et variés volumes de données. On retrouve par exemple les bases de données non relationnelles MongoDB et ElasticSearch , couplé à un stockage sur la technologie Cloud OpenStack Swift pour les fichiers binaires. Ensuite, du point de vue de l'archiviste, la méthode OAIS a été utilisée comme modèle fonctionnel, afin que le système Vitam répondent pleinement aux besoins des archivistes qui assurent la conservation des données sur le long terme. Le standard d'échange des données pour l'archivage (SEDA) a aussi été mis

Droits d'auteur réservés. OU (cc) BY-NC-ND

en place. Ainsi, Vitam constitue, à un certain point (rappelons que le logiciel ne comprend pas de couche analytique), le bon exemple d'une plateforme d'archivage des Big Data.

Le métier de l'archiviste peut apporter une véritable plus-value à la gestion des mégadonnées : il à un rôle à jouer dans leur préservation, et conséquemment, dans l'assurance qu'il peut apporter de leur exploitabilité dans le présent et dans l'avenir. Le phénomène du Big Data allié au « constat de masses documentaires et de données non maitrisées »[157] fait de l'archiviste un acteur pertinent dans la gestion des données massives : son apport, à l'instar d'un métier émergeant tel que *data steward*, réside bien dans le fait qu'il est producteur de qualité pour la donnée. L'approche du records management, qui tend à favoriser l'appréhension du processus d'archivage en amont, s'inscrit parfaitement dans les politiques de gouvernance des SI, si importantes à l'heure des Big Data. De plus, il est très peu probable que la « technologie » à elle seule puisse résoudre le problème du vrac informationnel, et comme le dit non sans humour Marie - Anne Chabin : « investir massivement dans le tri à postériori des données […] équivaut à éponger un appartement inondé par la baignoire qui déborde sans penser à fermer d'abord le robinet »[158]. Nous pensons sincèrement que l'archiviste est largement qualifié pour représenter ce professionnel de l'information qui veille à situer les données dans leur contexte, afin que celles-ci conservent tout leur sens *business*, leur sens métier. Cela est d'autant plus important avec le Big Data, ou la perte de sens des milliards de données est facilitée par leur nombre et leur variété. Si l'approche de l'archiviste fait sens pour tout type d'organisation, et tout type d'activité, nous soulignerons que le milieu de la recherche se prête particulièrement au déploiement d'un chantier d'archivage des données massives, qui combinerait l'approche MDM au modèle OAIS. Cependant à ce titre, il nous semble relativement inquiétant de constater qu'un projet de recherche marqué « Big Data » aussi ambitieux que les Time Machines ne se soit pas encore inquiété de déployer un modèle d'archivage de leur milliard de documents et données. Une opportunité à ne pas manquer pour les futures archivistes ?

[157] CHABIN, Marie-Anne, 2018. Des documents d'archives aux traces numériques: identifier et conserver ce qui engage l'entreprise la méthode Arcateg. Bois-Guillaume (Seine-Maritime) : Klog éditions. ISBN 979-10-92272-26-0. 658.403 8, page 79

[158] *ibidem*

Droits d'auteur réservés. OU (cc) BY-NC-ND

Droits d'auteur réservés. OU (cc) BY-NC-ND

Droits d'auteur réservés. OU (cc) BY-NC-ND

BIBLIOGRAPHIE

Ouvrages

BABINET, Gilles et ORSENNA, Erik, 2016. *Big Data, penser l'homme et le monde autrement*. Paris : Le Passeur. ISBN 978-2-36890-492-3.

BANAT-BERGER, Françoise, DUPLOUY, Laurent et HUC, Claude, 2009. *L'archivage numérique à long terme: les débuts de la maturité ?* Paris : la Documentation française Direction des archives de France. Manuels et guides pratiques. ISBN 978-2-11-006942-9. Z699

BÉCHARD, Lorène, FUENTES HASHIMOTO, Lourdes et VASSEUR, Édouard, 2014. *Les archives électroniques*. Paris : Association des archivistes français. Les Petits guides des archives. ISBN 978-2-900175-06-4. CD974.4

BRASSEUR, Christophe, 2016. *Enjeux et usages du Big Data*. 2e édition. Paris : Hermès Lavoisier. Management et informatique. ISBN 978-2-7462-4758-1. 658.05

BULINGE, Franck et CHOUET, Alain, 2014. *Maîtriser l'information stratégique: méthodes et techniques d'analyse*. Louvain-la-Neuve [Paris] : De Boeck ADBS. Information & stratégie. ISBN 978-2-8041-8914-3. T58.6

CASTENADO, Frederico, 2018, *Understanding Data Governance.* O'Reilly, ISBN, 978-1-491-99076-6

CHABIN, Marie-Anne, 2000. *Le management de l'archive*. Paris : Hermes science publications. ISBN 978-2-7462-0107-1. 0

CHABIN, Marie-Anne, 2018. *Des documents d'archives aux traces numériques: identifier et conserver ce qui engage l'entreprise la méthode Arcateg*. Bois-Guillaume (Seine-Maritime) : Klog éditions. ISBN 979-10-92272-26-0. 658.403 8

COINTOT, Jean-Charles et EYCHENNE, Yves, 2014. *La révolution big data: les données au cœur de la transformation de l'entreprise*. Paris : Dunod. Stratégies et management. ISBN 978-2-10-071142-0. 658.406

DELORT, Pierre, 2015. *Le big data*. Paris : Presses Universitaires de France. Que sais-je ?, n°4027. ISBN 978-2-13-065211-3. 005.7

Droits d'auteur réservés. OU (cc) BY-NC-ND

DELMAS, Bruno, 2015. *Archivistes de 2030: réflexions prospectives*. Louvain-la-Neuve : Academia-l'Harmattan.

DENOIX, Antoine, 2018. *Big Data, Smart Data, Stupid Data... : Comment (vraiment) valoriser vos données*. Dunod. ISBN 978-2-10-077351-0.

KAPADIA, Amar, RAJANA, Kris et VARMA, Sreedhar, 2015. *OpenStack Object Storage (Swift) Essentials*. Birmingham, UK : Packt Publishing Limited. ISBN 978-1-78528-359-8.

LAURENT, Pascale, LOWINGER, Hélène, MILLET, Jacques et CALDERAN, Lisette, 2015. *Big data: nouvelles partitions de l'information actes du séminaire IST Inria, octobre 2014*. Louvain-la-Neuve [Paris] : De Boeck ADBS. ISBN 978-2-8041-8915-0. 004

LEMBERGER, Pirmin, BATTY, Marc, MOREL, Médéric, RAFFAËLLI, Jean-Luc et GÉRON, Aurélien, 2016. *Big data et machine learning: les concepts et les outils de la data science*. 2e éd. Paris : Dunod. InfoPro. ISBN 978-2-10-075463-2. Q325.5

MENGER, Pierre-Michel et PAYE, Simon (éd.), 2017. *Big data et traçabilité numérique : Les sciences sociales face à la quantification massive des individus* [en ligne]. Paris : Collège de France. [Consulté le 29 décembre 2017]. Conférences. ISBN 978-2-7226-0467-4. Disponible à l'adresse : http://books.openedition.org/cdf/4987

O'NEIL, Cathy, 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York : Crown. ISBN 978-0-553-41881-1.

SERVAIS, Paul et MIRGUET, Françoise, 2015a. *Archivistes de 2030: réflexions prospectives*. Louvain-la-Neuve : Academia-l'Harmattan. Publications des Archives de l'Université catholique de Louvain, 32. ISBN 978-2-8061-0214-0. 020.

SERVAIS, Paul et MIRGUET, Françoise, 2015b. *L'archive dans quinze ans: vers de nouveaux fondements*. Louvain-la-Neuve : Academia-l'Harmattan. Publications des Archives de l'Université catholique de Louvain, 33. ISBN 978-2-8061-0225-6. Z.

WHITE, Tom, 2015. *Hadoop - The Definitive Guide 4e-*. 4. Beijing : O'Reilly. ISBN 978-1-4919-0163-2.

Mémoires de recherche

Droits d'auteur réservés. OU (cc) BY-NC-ND

BELLIER, Luc, 2017. *Organisation des données, organisation du travail en bibliothèques universitaires à l'heure du Big Data*, Mémoire sous la dir. de Nathalie Marcerou-Ramel, ENSSIB

GILLIUM, Johann, 2015. *Big data et bibliothèques: traitement et analyse informatiques des collections numériques*, Mémoire sous la dir. de Monique Joly, ENSSIB.

LAPÔTRE, Raphaëlle, 2014. *Faire parler les données des bibliothèques: du Big Data à la visualisation de données*, Mémoire sous la dir. de Julien Velcin, ENSSIB.

Articles de colloques/Conférences

1-Datskovsky-Big-Data-as-an-Asset.pdf, [sans date]. [en ligne]. [Consulté le 26 décembre 2017]. Disponible à l'adresse : http://www.arma-metromd.org/wp-content/uploads/2017/05/1-Datskovsky-Big-Data-as-an-Asset.pdf

2-Hogg-Cognitive-Governance-Opportunities-with-Big-Data.pdf, [sans date]. [en ligne]. [Consulté le 26 décembre 2017]. Disponible à l'adresse : http://www.arma-metromd.org/wp-content/uploads/2017/05/2-Hogg-Cognitive-Governance-Opportunities-with-Big-Data.pdf

3-Olsen-Privacy-in-a-Big-Data-Era.pdf, [sans date]. [en ligne]. [Consulté le 26 décembre 2017]. Disponible à l'adresse : http://www.arma-metromd.org/wp-content/uploads/2017/05/3-Olsen-Privacy-in-a-Big-Data-Era.pdf

5-Isaacs-Predictive-Analytics.pdf, [sans date]. [en ligne]. [Consulté le 26 décembre 2017]. Disponible à l'adresse : http://www.arma-metromd.org/wp-content/uploads/2017/05/5-Isaacs-Predictive-Analytics.pdf

data-warehouse-archiving_white-paper_7082.pdf, [sans date]. [en ligne]. [Consulté le 26 décembre 2017]. Disponible à l'adresse : https://www.informatica.com/content/dam/informatica-com/global/amer/us/collateral/white-paper/data-warehouse-archiving_white-paper_7082.pdf

DAVID, Gabriel, 2007. Data warehouses in the path from databases to archives. In : *International Workshop on Database Preservation*. 2007.

RABINOVICI-COHEN, Simona, BAKER, Mary G., CUMMINGS, Roger, FINEBERG, Sam et MARBERG, John, 2011. Towards SIRF: Self-contained information retention format. In : *Proceedings of the 4th Annual International Conference on Systems and Storage*. ACM. 2011. pp. 15.

RABINOVICI-COHEN, Simona, MARBERG, John, NAGIN, Kenneth et PEASE, David, 2013. PDS Cloud: Long term digital preservation in the cloud. In : *Cloud Engineering (IC2E), 2013 IEEE International Conference on*. IEEE. 2013.

Droits d'auteur réservés. OU (cc) BY-NC-ND

pp. 38–45.

SENN, Alina, 2016. Enabling 21st century archives. Paths and traces. In : E-ARK Conference 6 december 2016, Budapest.[en ligne]. [consulté le 1 aout 2018]. Disponible à l'adresse : http://www.eark-project.com/resources/conference-presentations/finconfpres/70-asennfinconf-1/file

SIRF_Use_Cases_V05a_DRAFT.pdf, [sans date]. [en ligne]. [Consulté le 26 décembre 2017]. Disponible à l'adresse : http://www.snia.org/sites/default/files/SIRF_Use_Cases_V05a_DRAFT.pdf

VIANA, P. et SATO, L., 2014. A Proposal for a Reference Architecture for Long-Term Archiving, Preservation, and Retrieval of Big Data. In : *2014 IEEE 13th International Conference on Trust, Security and Privacy in Computing and Communications*. septembre 2014. pp. 622-629.

Article de revues

BANAT-BERGER, Françoise, 2010. Les archives et la révolution numérique. *Le Débat*. 2010. N° 158, pp. 70-82. DOI 10.3917/deba.158.0070.

Big data: The next Google, 2008. Nature. Vol. 455, n° 7209, pp. 8-9. DOI 10.1038/455008a.

CARLSON, Mark, YODER, Alan, SCHOEB, Leah, DEEL, Don, PRATT, Carlos, LIONETTI, Chris et VOIGT, Doug, 2014. Software defined storage. *Storage Networking Industry Assoc. working draft, Apr*. 2014.

EKBIA, Hamid, MATTIOLI, Michael, KOUPER, Inna, ARAVE, G, GHAZINEJAD, Ali, BOWMAN, Timothy, SURI, Venkata, TSOU, Andrew, WEINGART, Scott et SUGIMOTO, Cassidy, 2015. Big Data, Bigger Dilemmas: A Critical Review. *Journal of the Association for Information Science and Technology*. 1 août 2015. Vol. 66. DOI 10.1002/asi.23294.

FEDERER, Lisa, 2016. Research data management in the age of big data: Roles and opportunities for librarians. *Information Services & Use*. 1 janvier 2016. Vol. 36, n° 1-2, pp. 35-43. DOI 10.3233/ISU-160797.

HOY, Matthew, 2014. Big Data: An Introduction for Librarians. *Medical reference services quarterly*. 14 juillet 2014. Vol. 33, pp. 320-6. DOI 10.1080/02763869.2014.925709.

KHATRI, Vijay et BROWN, Carol V., 2010. Designing data governance. *Communications of the ACM*. 2010. Vol. 53, n° 1, pp. 148–152.

SUKUMAR, Sreenivas Rangan et K. FERRELL, Regina, 2013. « Big Data » collaboration: Exploring, recording and sharing enterprise knowledge. *Information Services and Use*. 1 juillet 2013. Vol. 33, pp. 257-270. DOI 10.3233/ISU-130712.

Droits d'auteur réservés. OU (cc) BY-NC-ND

ZHAN, Ming, 2017. Understanding big data in librarianship. *Journal of Librarianship and Information Science*. 13 décembre 2017. DOI 10.1177/0961000617742451.

Articles sur le Web

AAS, Kuldar et AAS, Kuldar, [sans date]. Coping with the data explosion. . pp. 19.[en ligne]. [Consulté le 1 aout 2018]. Disponible à l'adresse : http://www.eark-project.com/resources/conference-presentations/presents2016/62-ica2016-1/file

BILLENESS, Clive, [sans data]. An E-ARK Overview. [en ligne]. [Consulté le 1 Aout 2018]. Disponible à l'adresse : http://www.eark-project.com

BUSINESS, BFM, [sans date]. Serge Abiteboul : « Le big data est avant tout un effet de mode ». *BFM BUSINESS* [en ligne]. [Consulté le 29 décembre 2017]. Disponible à l'adresse : http://bfmbusiness.bfmtv.com/01-business-forum/serge-abiteboul-le-big-data-est-avant-tout-un-effet-de-mode-572981.html

DEBRAY, Eric, [sans date]. La virtualisation des Données : un atout pour internet de l'objet , Cloud et Big Data. [en ligne]. [Consulté le 26 décembre 2017]. Disponible à l'adresse : https://gblogs.cisco.com/fr/datacenter/la-virtualisation-des-donnees-un-atout-pour-lioe-le-cloud-et-le-big-data/

DEROOS, Dirk, [sans date]. Hadoop as an Archival Data Destination - dummies. [en ligne]. [Consulté le 26 décembre 2017]. Disponible à l'adresse : http://www.dummies.com/programming/big-data/hadoop/hadoop-as-an-archival-data-destination/

DEROSA, Guy, [sans date]. Utilizing Hadoop & HDFS as an Active Archiving & Storage Framework. [en ligne]. [Consulté le 26 décembre 2017]. Disponible à l'adresse : http://www.aptude.com/blog/entry/utilizing-hadoop-hdfs-as-an-active-archiving-storage-framework

DORION, Pierre, [sans date]. Backup vs. archive. *SearchDataBackup* [en ligne]. [Consulté le 26 décembre 2017]. Disponible à l'adresse : http://searchdatabackup.techtarget.com/tip/Backup-vs-archive

DOUETTEAU, Florian, [sans date]. Concilier Big data, IoT, Intelligence artificielle et RGPD. [en ligne]. [Consulté le 26 décembre 2017]. Disponible à l'adresse : http://www.journaldunet.com/solutions/expert/66836/concilier-big-data--iot--intelligence-artificielle-et-rgpd.shtml

FRANCE, Bibliothèque nationale de, [sans date]. BnF - Archives de l'internet. [en ligne]. [Consulté le 26 décembre 2017 a]. Disponible à l'adresse : http://www.bnf.fr/fr/collections_et_services/livre_presse_medias/a.archives_internet.html#SHDC__Attribute_BlocArticle0BnF

FRANCE, Bibliothèque nationale de, [sans date]. BnF - Le Schéma numérique

Droits d'auteur réservés. OU (cc) BY-NC-ND

de la BnF. [en ligne]. [Consulté le 26 décembre 2017 b]. Disponible à l'adresse : http://www.bnf.fr/fr/la_bnf/missions_bnf/s.bnf_schema_numerique.html?first_Art=non

HAQUETTE, Paul, 2017. Le Big Data libère le potentiel des archives d'entreprises. *Blog Big Data & Digital* [en ligne]. 11 juillet 2017. [Consulté le 26 décembre 2017]. Disponible à l'adresse : http://fr.blog.businessdecision.com/bigdata/2017/07/big-data-potentiel-archives-entreprises/

HLADKY, Edward, [sans date]. A l'heure du Big Data, vos archives valent des millions. *usine-digitale.fr* [en ligne]. [Consulté le 26 décembre 2017]. Disponible à l'adresse : https://www.usine-digitale.fr/article/a-l-heure-du-big-data-vos-archives-valent-des-millions.N346120

IBM, Software, [sans date]. benefits of data archiving in data warehouses - Recherche Google. [en ligne]. [Consulté le 26 décembre 2017]. Disponible à l'adresse : https://www.google.fr/search?q=benefits+of+data+archiving+in+data+warehouses&oq=benefits+of+data+archiving+in+data+warehouses&aqs=chrome..69i57.17719j0j7&sourceid=chrome&ie=UTF-8

JOST, Clemence, [sans date]. Ghislaine Chartron : « Je ne transformerai pas mes étudiants en data scientists ». *Archimag* [en ligne]. [Consulté le 26 décembre 2017]. Disponible à l'adresse : http://www.archimag.com/veille-documentation/2015/11/26/ghislaine-chartron-transformer-etudiants-data-scientists

PERRET, Xavier et JACQUEMELLE, Guy, [sans date]. Comprendre le Big Data à travers les films de cinéma. *OpenClassrooms* [en ligne]. [Consulté le 26 décembre 2017]. Disponible à l'adresse : https://openclassrooms.com/courses/comprendre-le-big-data-a-travers-les-films-de-cinema

SENSMEIER, Lisa, 2013. Modernizing Data Archiving with Hadoop. *Hortonworks* [en ligne]. 29 août 2013. [Consulté le 26 décembre 2017]. Disponible à l'adresse : https://fr.hortonworks.com/blog/modernizing-data-archiving-virtualization-for-big-data-analytics/

SCHLARB, Sven , 2016. The Use of Big Data techniques for Digital Archiving. *Austrian Institute of Technology* [en ligne]. [Consulté le 26 décembre 2017]. Disponible à l'adresse : https://www.bigdata.cam.ac.uk/files/our-digital-future-2016/our-digital-future-2016-slides/sven-schlarb-e-ark-15-march-2016

SIEGLER, M.G. Eric Schmidt: Every 2 Days We Create As Much Information As We Did Up To 2003, [sans date]. TechCrunch [en ligne]. [Consulté le 26 août 2018]. Disponible à l'adresse : http://social.techcrunch.com/2010/08/04/schmidt-data/

TECHTARGET, [sans date]. « Selfish » archive strategy key to compliance for the SME. *SearchITChannel* [en ligne]. [Consulté le 26 décembre 2017].

Droits d'auteur réservés. OU (cc) BY-NC-ND

Disponible à l'adresse : http://searchitchannel.techtarget.com/feature/Selfish-archive-strategy-key-to-compliance-for-the-SME.

TOIGO, Jon, [sans date]. How is big data changing data archiving strategies? *SearchStorage* [en ligne]. [Consulté le 26 décembre 2017]. Disponible à l'adresse : http://searchstorage.techtarget.com/answer/How-is-big-data-changing-data-archiving-strategies

Enabling kNowledge Sustainability Usability and Recovery for Economic value | Projects | FP7-ICT, [sans date]. CORDIS | European Commission [en ligne]. [Consulté le 30 juillet 2018]. Disponible à l'adresse : https://cordis.europa.eu/project/rcn/98002_en.htmlFP7-ICT,ENSURE,Ensuring

Focus archivistes · Vitam, [sans date]. [en ligne]. [Consulté le 13 août 2018]. Disponible à l'adresse : http://www.programmevitam.fr/pages/presentation/pres_archivistes/

Présentation · Vitam, [sans date]. [en ligne]. [Consulté le 13 août 2018]. Disponible à l'adresse : http://www.programmevitam.fr/pages/presentation/

vitam: Digital Archives Management System developped by French government/Programme interministériel archives numériques ; core system, 2018. [en ligne]. Java. Programme Vitam. [Consulté le 13 août 2018]. Disponible à l'adresse : https://github.com/ProgrammeVitam/vitam

VITAM_Presentation_solution_logicielle.pdf,[sans date]. [en ligne]. [Consulté le 13 août 2018].Disponible à l'adresse : http://www.programmevitam.fr/ressources/DocCourante/autres/fonctionnel/VITAM_Presentation_solution_logicielle.pdf

Vidéo

ENS PARIS, [sans date]. *Digit_Hum 2017* [en ligne]. [Consulté le 26 décembre 2017]. Disponible à l'adresse : https://transfers.huma-num.fr/digithum/atelier/2017/

TYLDUM, Morten, HODGES, Andrew, MOORE, Graham, DESPLAT, Alexandre, CUMBERBATCH, Benedict, KNIGHTLEY, Keira et GOODE, Matthew, 2015. *Imitation game* [en ligne]. Universal StudioCanal vidéo [distrib.], 2015. [Consulté le 25 août 2018]. Disponible à l'adresse : https://bibliotheques.paris.fr/Default/doc/SYRACUSE/1012218/imitation-game

Droits d'auteur réservés. OU (cc) BY-NC-ND

ANNEXES

Table des annexes

Droits d'auteur réservés. OU (cc) BY-NC-ND

Droits d'auteur réservés. OU (cc) BY-NC-ND

GLOSSAIRE

Apache software Foundation : Organisme sans but lucratif américain qui supporte les projets de logiciel libres du même nom. La fondation rassemble une importante communauté d'ingénieurs et de développeurs, qui s'engagent à réaliser des projets informatiques libres (FOSS – *Free and Open Source Software*), distribués sous la licence Apache.

API : Trigramme pour *Application Programming Interface*. Une API est une interface pour les applications. Elle est constituée pour permettre à des applications de se connecter entre elles. L'API la plus connue et utilisée est l'API « REST » (pour *Representational State Transfer»*), basée sur le protocole web HTTP.

Base de données : Une base de données permet de stocker l'ensemble des informations se rapportant à une ou des activités déterminées. En général, la base de données stocke des données structurées, organisées en tables (une base de données peut comporter une ou plusieurs tables).

Base de données Orientées Agrégats : catégorie de base de données NoSQL. On doit l'expression « agrégats » à l'informaticien britannique Martin Fowler. Les données les plus sollicitées sont stockées non plus dans des tables (comme dans une base de données relationnelle classique), mais au sein d'agrégats. Un agrégat peut être constitué de valeur, soit d'un ensemble d'objet métiers, d'un fichier texte ou encore vidéo ; et d'une clé pour accéder à cette valeur.

Base de données Orientées Documents : la base de données orientées documents est une sous-catégorie des Bases de données orientées agrégats. De type NoSQL, ces bases de données stockent des documents auto – descriptif XML ou JSON. Contrairement au SGBDR, les bases de données orientées documents sont *schemaless*: les documents présent dans la base de données n'ont pas nécessairement tous le même format. De fait, il est possible de modifier la structure des documents sans avoir à redéfinir le format de la base. En termes de disponibilités, les BDOD repose sur un mécanisme de réplication de type maitre – esclave. Les données sont distribuées sur plusieurs instances de base de données, afin d'obtenir un temps de réponse plus rapide lorsque les données sont requêtées (*sharding*). Les BDOD les plus connus sont *MongoDB, CouchDB* et *RavenDB*.

Cloud (computing) : Le cloud computing est un ensemble de ressources partagées, de système configurable et de services de niveau supérieur pouvant être rapidement provisionnés avec un effort de gestion minimale, souvent par le réseau internet. L'informatique en nuage repose sur le partage des ressources pour assurer la cohérence et des économies d'échelle, à l'instar d'un service public. Le modèle de paiement repose en général sur la logique du « as a service ». Il s'agit donc de payer une consommation (une puissance de calcul par exemple), au lieu d'avoir à charge l'entretien de l'infrastructure.

Cluster (ou grappe de serveurs) : technique qui consiste à bâtir une infrastructure regroupant plusieurs ordinateurs indépendants (nommés *nodes* – nœuds), mis en commun afin de répondre à plusieurs défis techniques : augmenter la disponibilité des serveurs, augmenter la puissance calcul, s'assurer d'une meilleure gestion des volumétrie etc .. . Ce patron d'architecture est singulièrement utilisé lorsqu'il est question de réaliser des calculs parallèles (implication de Hadoop/MapReduce).

Droits d'auteur réservés. OU (cc) BY-NC-ND

Data Scientist : Le data scientist est un spécialiste de la science des données. Ainsi, il est un haut responsable de la gestion et de l'analyse des Big Data. Celui-ci doit avoir de bonnes connaissances à la fois dans les domaines de l'informatique (Machine Learning, Deep Learning), mais aussi en mathématique statistique.

Electronic Discovery (ou e-discovery) : Aux Etats – Unis, le terme fait référence à toute mesure d'instruction lors de cas juridique de litige, investgation gouvernementale, ou demande du *Freedom of Information Act* lorsque l'information recherchée est de type numérique. La procédure encourage les entreprises à archiver de façon automatique tout les documents et communications numériques.

EIM/ECM : Trigramme pour *Enterprise Information Management*/*Enterprise Content Management*. Domaine relatif aux sciences de l'information, il recouvre le périmètre couvert par la gestion électronique des documents, l'archivage ou encore l'éditique. Aujourd'hui, l'EIM est également largement concerné par les activités liées à la gouvernance de l'information (Master Data Management).

GDPR/RGPD : Abréviation pour *General Data Protection Regulation*/Règlement Général sur la Protection des Données. La GDPR est un règlement de l'Union Européenne qui fonde le texte de référence en matière de protection des données à caractères personnel. Il est applicable depuis le 25 mai 2018.

GED : Trigramme pour *Gestion électronique des documents*. La GED consiste en un ensemble de pratique et d'implémentation logicielle pour suivre, gérer et organiser l'ensemble des données non – structurées (images, textes brutes, vidéo etc …) d'une organisation. Pour faire simple, c'est l'ensemble de moyens qui doit permettre à la bonne personne de trouver l'information au bon endroit et au moment voulu.

Gouvernance des données (ou de l'information) : Discipline cherchant à amener une meilleure maîtrise de l'information dans les organisations afin de s'assurer de sa qualité globale.

Hortonworks : Entreprise américaine basée à Santa Clara (Californie), qui propose ses propres distributions de logiciels *open source*. Hortonworks est un acteur majeur des Big Data, notamment grâce à ses distributions de Hadoop.

IA/AI : Abréviation pour Intelligence artificielle/*Artificial Intelligence*. Désigne la capacité pour une machine à démontrer de l'intelligence, dans la mesure ou cette intelligence se rapproche de celle dont est capable de faire preuve un être humain (exemple : résolution de problèmes, parcours d'obstacles). La recherche en intelligence artificielle est aujourd'hui bien couverte par le secteur des sciences de l'information et de l'informatique.

JSON : Trigramme pour *Javascript Object Notation*. Langage statique de description à l'instar du XML, JSON est un format de fichier qui permet d'organiser et de décrire des données. Le JSON est compréhensible par l'homme comme par la machine.

Machine Learning : Le Machine Learning (ou « apprentissage automatique », en français) est un domaine des sciences informatique qui, à l'aide de modèles statistiques avancés, permet à un programme informatique d'« apprendre » à partir d'ensemble de données préalablement collectées. Les algorithmes sont utilisés pour réaliser des analyses prédictives, de la détection de fraudes ou encore dans le filtrage des e-mails.

Droits d'auteur réservés. OU (cc) BY-NC-ND

MapR : Entreprise américaine située à San Jose (Californie). Concurrent de Hortonworks, MapR propose également ses propres distributions de logiciels Big Data de la fondation Apache.

MarkLogic : MarkLogic est un éditeur américain qui développe et fournit une solution NoSQL à destination es organisations. La base de données maintenue par la société est une base de données orientée document, apte à gérer des objets XML ou JSON.

MDM : Trigramme pour désigner *Master Data Management*. Ensemble de moyens humains et techniques pour les données critiques des organisations. Il peut par exemple s'agir de monter un référentiel de données, ou encore un dictionnaire de données ou un glossaire de données. Le MDM est en cela un acteur essentiel de la gouvernance des données. Il veille ainsi à la qualité des données présentes dans l'organisation, en assurant la gestion de leur cycle de vie, de leur référencement et de leur valeur métier.

OpenStack Swift : Projet open source lancé par la NASA et la société Rackspace en 2010. OpenStack. La solution logicielle principale proposée est une plateforme de Cloud Computing. Swift en est le modèle de stockage, tourné vers le stockage de fortes volumétries de données non structurées. Le modèle de stockage proposé par Swift s'inscrit dans la logique du stockage objet.

RAID : Abréviation pour *Redundant Array of Independent Disks*. RAID correspond à une technique de virtualisation de stockage pour les données stockées sur disques durs. Elle permet de répartir les données sur plusieurs disques durs, afin d'améliorer la performance ou la tolérance aux pannes du système.

SAE : Trigramme pour *Système d'archivage Electronique*. Un SAE est un socle logiciel qui permet de gérer les documents (*records*) et les données à travers leur cycle de vie, de leur création à leur destruction. Une véritable SAE doit répondre à des exigences de gestion de l'authenticité et de l'intégrité des documents/données , de gestion du cycle de vie, de pérennisation/préservation à long terme ainsi que de sécurité et de stockage.

SI: abréviation pour *Système d'Information*. Un system d'information représente un ou l'ensemble du socle technologique d'une organisation. Le SI est généralement administré par une DSI (Direction des Systèmes d'Information).

SGBDR : abréviation pour *Système de Gestion de bases de données relationnelles*. C'est un système de gestion de bases de données reposant sur le model relationnel mis en place par l'informaticien américain de l'entreprise IBM, Edgar F. Codd. Les données sont donc organisées dans une base de données relationnelles, dans une ou plusieurs tables constituées de ligne et de colonnes. Le langage SQL est utilisé pour interagir avec les données.

SQL : trigramme pour *Structured Query Language*. Langage de programmation destiné à gérer et exploiter les données présentes dans les bases de données relationnelles. Il est très utilisé pour contrôler les données structurées contenu dans les tables de la base de données montrant des relations entre elles. En somme, le langage SQL permet de dialogue avec la base de données, cela grâces à certains opérations implémentées dans le langage (sélection des données, fonctions de jointures, d'indexations, de tries etc …).

XML : Trigramme pour *Extensible Markup Language.* Langage statique de description, le format XML définit un ensemble de règle pour encoder les documents afin que ceux-ci soient compréhensibles à la fois par un humain et une machine. Le

Droits d'auteur réservés. OU (cc) BY-NC-ND

XML est également utilisé pour décrire la structure de données transitant via des services web.

Droits d'auteur réservés. OU (cc) BY-NC-ND

INDEX

Droits d'auteur réservés. OU (cc) BY-NC-ND

TABLE DES ILLUSTRATIONS

Insertion de la table des illustrations

Droits d'auteur réservés. OU (cc) BY-NC-ND

TABLE DES MATIERES

Droits d'auteur réservés. OU (cc) BY-NC-ND