Les Archives à l'Ère des Big Data

Les Enjeux de l'Archivage des Données Numériques Massives

Fatma Ben Amor
Institut supérieur de documentation
Université de la Manuba
Tunis, Tunisia
fatma_bilel@yahoo.fr

Abderrazak Mkadmi
Institut supérieur de documentation
Université de la Manuba
Tunis, Tunisia
amkadmi@gmail.com

ABSTRACT

Big Data is now a cross-cutting research topic in all disciplines related to digital as content and as technology too. They lie in the intersection between all the massive data captured, obtained, created by different means and of various origins. They represent an advanced step in the re-development of information, particularly concerning data management and data retention issues. This upheaval due to these massive data has touched all sectors, particularly the archives. Indeed, given their volume, their speed of creation and their importance to the social, economic, scientific and cultural actors, the sorting, the treatment and the conservation of these data Massive e orts require memory capacity, new methods and techniques for processing, analyzing and managing particular flows. We will try in this article to bring some elements of primary answers on the modalities of generative multiplication. exponential of numerical data and archive in a mass of numeric data, data that are in flux and that occur in a sup speed.

CCS CONCEPTS

• **Information systems → Digital libraries and archives**;

KEYWORDS

Big Data, Digital archiving, long-term conservation, big data access

ACM Reference Format:
Fatma Ben Amor and Abderrazak Mkadmi. 2018. Les Archives à l'Ère des Big Data: Les Enjeux de l'Archivage des Données Numériques Massives . In *Digital Tools & Uses Congress (DTUC '18), October 3–5, 2018, Paris, France.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3240117.3240139

Résumé

Les Big Data constituent aujourd'hui un sujet de recherche transversal touchant toutes les disciplines ayant lien avec le numérique en tant que contenu et en tant que technologies aussi. Elles se situent dans l'intersection entre toutes les donnéées massives captées, obtenues, créées par des différents moyens et de diverses origines. Elles représentent une étape avancée de la révolution de l'information touchant notamment les questions de gestion et de conservation des données. Ce bouleversement dû à ces données massives a touché tous les secteurs et notamment celui des archives. En effet, vu leur volume, leur vitesse de création et leur importance auprès des acteurs sociaux, économiques, scientifiques et culturels, le tri, le traitement et la conservation de ces données massives nécessitent des capacités de mémoire vive, des nouvelles méthodes et techniques de traitement, d'analyse et de gestion de flux particulières.

Nous allons essayer dans cet article d'apporter des éléments de réponses primaires sur les modalités de gérer la multiplication exponentielle des données numériques et archiver dans une masse de données numériques, des données qui sont en flux et qui se produisent en une supère vitesse.

DTUC '18, October 3–5, 2018, Paris, France

ACM ISBN 978-1-4503-6451-5/18/10…$15.00
https://doi.org/10.1145/3240117.3240139

1 INTRODUCTION

Chaque changement ou évolution technologique donne lieu à des mutations effectives surtout dans le domaine de traitement et conservation de l'information. Avec le développement de l'internet, l'augmentation de la puissance des ordinateurs, l'accroissement des capteurs divers et variés des données, l'apparition des nouvelles méthodes de traitement de l'information dues aux « humanités numériques », la production d'une masse exponentielle des informations et des corpus d'archives numériques résultant des opérations d'analyse, d'annotation, d'indexation collaborative des textes, que commence à paraitre, une explosion quantitative des données numériques, ainsi, est né le « Big Data ».

L'appellation de « Big Data » est apparue en 1997 selon les archives de la bibliothèque numérique de l'Association for Computing Machinery (ACM), affirmant la notion d'un « grand ensembles de données» [1] [15]. Le Big Data est qualifié comme une nouvelle révolution industrielle semblable à la découverte de la vapeur (début du 19e siècle), de l'électricité (fin du 19e siècle) et de l'informatique (fin du 20e siècle) ou comme étant la dernière étape de la troisième révolution industrielle, celle de « l'information », le Big Data est un bouleversement profond de la société en général et des archives spécifiquement bien qu'il touche à la question de gestion et de stockage des données.

2 PROBLÉMATIQUE

Les Big Data se situent dans l'intersection entre toutes les données massives captées, obtenues, créées par des différents moyens et de diverses origines. Sous le label Big Data, nous trouvons plusieurs types de données :

[1] "data sets are generally quite large"

- les données numériques captées et enregistrées via des dispositifs particuliers (interfaces mobiles, bracelets numériques, lunettes augmentées),
- les données commerciales collectées par les instituts de sondages,
- les données numériques résultant des archives numérisées ou les données « non nativement numériques »,
- les données numériques issues de questionnaires (statistiques, tableaux, fichiers, page de traitement de texte, d'images, de sons),
- les données publiques qui concernent l'e-gouvernement, l'e-citoyenneté et l'e-administration générées par les administrations de l'Etat et les collectivités territoriales,
- les données produites par le web des réseaux sociaux ainsi que les masses de données textuelles et audiovisuelles produites par les programmes de numérisation et de diffusion des données massives textuelles [25].

Vu leur volume massif, leur vitesse incroyable de création et leur importance bien qu'elles soient à la disposition de plusieurs et différents acteurs sociaux et font le sujet d'usage à la fois politique, économique, commercial, scientifique et culturel, le tri, le traitement et la conservation de ces données massives nécessitent des capacités de mémoire vive, des nouvelles méthodes et techniques de traitement, d'analyse et de gestion de flux particulières.

Cette nécessité se confronte avec un ensemble de problématiques qui touchent à la fois leur nature : données "données" ou données "construites " par l'être humain, leur valeur ou leur sens en tant que "Big Data" ou " smart data " et leur vitesse de création immense qui pose la question de leur "stockage" ou "archivage".

Sous le prisme de « Big Data » se posent, alors, les questions suivantes :

- Comment peut-on gérer la multiplication exponentielle des données numériques et archiver dans une masse de données numériques, des données qui sont en flux et qui se produisent en une supère vitesse ?
- Comment maitriser les contraintes de stockage des données et d'archivage des contenus ?
- Comment les Big Data ont-t-elles modifié le monde d'archivage des données ?

3 INTRODUCTION AUX BIG DATA

3.1 Définition des Big Data

Commençant par l'affirmation de Etienne Ollion [14] le Big Data « est un terme à l'origine contestée, il n'en existe pas de définition stabilisée ». « Le Big Data, littéralement « grosses données », ou méga-donnée[2] , parfois appelées données massives, désigne des ensembles de données devenus si volumineux qu'ils dépassent l'intuition et les capacités humaines d'analyse et même celles des outils informatiques classiques de gestion de base de données ou de l'information ».

Littéralement le Big Data désigne « grosses données » ou mégadonnées [3], ou encore données massives, selon Marie Anne Chabin le concept de « magadonnées » qui traduit l'idée de nombre (gigadonnées, teradonnées,..) reflète une confusion avec l'idée de masse. Ainsi, il présente un risque avec les métadonnées. Quant au « données massive » la notion de masse donne l'impression d'une masse compacte or il s'agit des données complètement dispersées traduites efficacement par le concept idéal « Big Data » « qui relève autant de l'économie, de la sociologie ou de l'archivage que de l'informatique » [7]

Une autre définition renvoie à la notion d'accès à l'information celle donnée par l'avocat Merav Griguer "Le Big Data désigne une démarche particulière, qui consiste à extraire l'information pertinente d'un ensemble de données" [16].

3.2 Les origines des Big Data

Pour faire un bref historique du Big Data, nous pouvons se reporter à l'article de Gil Press daté du 5 mai 2013: "une très courte histoire du Big Data" [15]. Il retrace les origines des Big Data en une trentaine de dates de 1944 à 2012. Ça commence par l'explosion des données en 1944 perçue comme une menace sur la vie privée puis, les données s'étendent systématiquement jusqu'à combler l'espace de stockage. Dans les années 70, la qualité des données est enfin mise en cause: tout est stocké, il n'est plus utile de faire le tri. L'expression « Big Data » fut apparu en octobre 1997 dans la bibliothèque numérique de l'ACM (Association for Computing Machinery).

3.3 Les spécificités des BIG DATA: les 5 V

Les Big Data d'après le Principe des « trois V » présente les 3 propriétés suivantes: Volume, Variété, Vitesse. On parle alors de la règle des 3V qui est devenue par extension, règle des 4V, puis règle des 5V [3].

Le volume de données de plus en plus important 100 To (téraoctets) de données non structurées); la Variété de ces données qui peuvent être brutes, non structurées ou semi-structurées et la Vitesse est liée à la vitesse de traitement crucial et propre à la notion de Big Data. Quant à la Véracité rajouté à la règle V4 désigne la vérification de la crédibilité de la source et la qualité du contenu des données exploitées. La Valeur s'ajoute avec la règle des 5V indiquant la valeur ajoutée pour l'entreprise par l'exploitation des données massives.

4 ARCHIVES ET BIG DATA

Quel couplage entre les archives et les Big Data ? Réfléchir à la manière dont les documents, les principes, les modalités de sélection et de tri des données à traiter et à archiver et les données produites et détenues par les services d'archives pouvaient faire l'objet d'une redéfinition, resélection, retraitement et reconservation par un nouvel outillage issu des Big Data, c'est la question.

4.1 Du document aux data

Si dans le contexte de l'humanité numérique « tout peut devenir document » [6] vu l'élargissement vers la notion de « source » qu'il s'agit des éléments interconnectés, des pages manuscrites numérisées, des textes transcrits et annotés, des images de manuscrits isolées, des liens tissus entre divers éléments de corpus numériques, est-ce que c'est le cas avec l'expropriation énorme des données numériques à l'ère des Big Data ? Est-ce que «tout peut devenir data » ?

[2]Selon la commission générale de terminologie et de néologie, depuis le décret du 22 août 2014.

[3]Terme recommandé en France par la DGLFLF, Journal officiel du 22 août 2014, et au Canada par l'OQLF

Un document renvoie à un ensemble formé par un support et une information (wikipédia) et que l'information est par définition une donnée interprétée. Donc un document en fait, englobe des données organisées et interprétées.

Avec le numérique, le document est décomposé alors qu'il était unique [1]. Il est devenu à la fois une structure (langage informatique), une forme (ensemble des données organisées), un signe (texte, image ou son) et un medium (trace des relations sociales reconstruites par les dispositifs informatiques) [21]. On est alors passé du document indexé permettant la recherche documentaire suivant des méthodes traditionnelles, à la ressource annotée (Web sémantique), puis à la donnée manipulée âĂŤ Big Data [1].

4.2 Des Archives aux « big archives »

Le concept de "big archives" est utilisé par plusieurs auteurs, à l'égard de Seven Spieker dans son ouvrage "The big archive art from bureaucracy" [4]. Dans cet ouvrage, Sven Spieker étudie les archives en tant qu'institution bureaucratique et indice d'évolution des attitudes envers le temps éventuel dans la science et l'art. La notion de "big archive" à cet effet, prend sa place dans un environnement des humanités numériques et prend du concept "Big Data" le mot "big" pour désigner l'importance et la massivité des archives en relation avec l'art. Spieker et contrairement à la croyance selon laquelle les archives capturent l'histoire d'une manière ordonnée, défend l'idée qu'elles sont plutôt des liens de chaos et de contingence [5]. [18]

Si dans la notion classique des archives, l'archive est fondée sur la notion du temps, l'hypothèse étant que le temps progresse de manière linéaire, ouvrant ainsi des possibilités d'archivage, à l'ère des Big Data le temps parait comme une série d'intervalles (par opposition à la linéarité continue). La désorientation, le choc et la non-linéarité deviennent la logique des archives et remplacent selon spieker le « Principe de provenance et de la sûreté de l'ordre original ». [18]. Si l'archive est, comme l'affirme Derrida, le site de commencement (l'originaire) et de commandement (l'autorité) [6] , Spieker prétend que l'archive rationnelle et ordonnée est en réalité entachée par sa prédisposition au désordre. La notion de "big archive" nous permet de penser au-delà des limites des murs du dépôt et de l'ordre bureaucratique, Spieker nous invite à réfléchir sur les possibilités des archives, les espaces au-delà des archives. Les "big archive" ne sont pas simplement un recueil de textes enregistrés, mais incluent aussi des concepts théoriques plus larges comme la métapsychologie, la théorie critique et l'historiographie. [18]

Maria Estiva dans son article "Data Mining for "Big Archives Analysis : a Case Study"[11], et Martyn Jolly dans son article "Big Archives and Small Collections" [20] utilisent le terme "big archive" pour traduire la notion de l'importance du volume, de la variété et de la complexité au niveau des enregistrements qui varient en termes de contenu, de type, de quantité et de modèle d'organisation. « Comme né-numérique et numérisé les archives deviennent plus grandes et plus complexes, l'application des méthodes de calcul pour traiter la quantité d'une manière efficace devient une nécessité inévitable » [11].

4.3 Stockage de données *vs* archivage de données

Le débat entre stockage et archivage se poursuit depuis des années. Une confusion existe souvent entre l'archivage et le stockage de données.

Marie-Odile Charaudeau parle dans son article "Et demain ? Archivage et Big Data" de la différence entre le "big archivage" et le "big stockage" affirmant que l'archivage perçu comme une approche dédiée aux documents en tant que données non structurées alors que le "Big Data" perçu comme une approche dédiée aux données structurées. Une convergence entre les deux approches parait nécessaire? [8].

L'archivage de données protège les anciennes informations qui ne sont pas nécessaires pour les opérations quotidiennes, mais sont encore importantes et nécessaires pour référence dans le futur, et qui doivent être conservées pour la conformité réglementaire et indexées pour faciliter leur localisation et leur récupération.

En revanche, Les stockages des données sont destinés à la reprise après sinistre et l'archivage de données est à découvrir. Un stockage de données est pour restaurer les fichiers perdus ou corrompus.

Les experts en stockage affirment constamment que les sauvegardes ne sont pas des archives. Et vu que les logiciels de stockages traditionnels n'aideront pas à archiver, les fournisseurs de logiciels de stockage et de récupération de données ont commencé à intégrer différentes fonctionnalités dans leurs logiciels incluant la déduplication des données et la gestion du cycle de vie des données avec la hiérarchisation du stockage.

Si les données sont instables et hétérogènes et circulent rapidement, leur conservation ne consiste pas à les faire stocker et les mettre à part mais au contraire, à les laisser circuler « conserver des données numériques, c'est les laisser circuler » [7] [2]et les faire migrer continuellement d'un serveur à l'autre par leur découpage en morceaux. L'archive donc est un « processus contenu » dont le rôle n'est pas seulement de fixer les contenus, mais préserver l'usage et les interprétations qui donnent le plus à ces contenus, les enrichissant et les produisant ainsi d'autres.

Ce processus dynamique rend l'archivage une action qui précède les données, les connaissances, en efft, l'archive « ne suit plus la connaissance mais la précède » [8] [2]on se met alors en archives avant même la production des données qui seront interprétés ultérieurement.

4.4 Du système d'Archivage Electronique (SAE) au data lake

Le Système d'Archivage Electronique (SAE) « est un outil informatique permettant la conservation pérenne et sécurisée des documents électroniques. Une fois intégré dans un SAE, un document n'est plus modifiable et conserve donc sa valeur probante » [9], c'est le résultat d'une application stricte des normes et recommandations

[4]The Big Archive: Art from Bureaucracy by Sven Spieker. Cambridge: MIT Press, 2008. 219 pp. ISBN 978-0-262-19670-6.

[5]Traduit par Google:" contrary to the belief that archives capture history in a well-ordered manner, they are rather sites of chaos and contingency, with the presupposition of the rationality of linear history haunted by the specter of entropy and disorder."

[6]Idixa - Art, pensée, philosophie - La demeure de l'Orloeuvre https://www.idixa.net/Pixa/pagixa-0704300851.html

[7]Propos de Bertrand Muller

[8]ibid

essentielles en la matière, qui sont : ISO 15 489 Records Management ; OAIS ISO 14721 ; NF Z42-013 et NF Z44- 022 MEDONA. [19]. Il s'appuie sur une architecture simple, structurée et traditionnelle respectant les principes d'indépendance, d'autonomie et donc de pérennité des informations et des archives.

Inspirant de la norme OAIS (ISO 14721 2003), le SAE conserve l'information sous une forme pérenne indépendamment des systèmes producteurs. En effet, « une infrastructure d'archivage pour être pérenne doit être le plus possible autonome et commune. Avec pour corollaire que les archives doivent être exploitables indépendamment de leur contexte de production d'origine » [19]. Cela permet de gérer les règles de gouvernance telle que la sécurité, la conservation et l'élimination.

Il convient alors, de créer des entrepôts de données « data warehouses » [9] qui vont permettre de regrouper toutes les données structurées en silos dans les bases opérationnelles. A l'ère des Big Data, un terme vient se greffer au système d'archivage électronique qui est les « data lake » ou les lacs de données là où les données sont stockées dans des lacs plutôt que dans des entrepôts. "Un lac de données est plat, sans structure. Les données sont conservées sur le même plan. La structure est alors créée au moment de l'analyse. On parle de « data lake» mais aussi de «data reservoir», réservoir de données"[23].

Dès lors, avec les data lake, la notion de bases de données relationnelles et structurées s'est changé vers une autre notion à l'égard des systèmes de stockage arborescents comme Hadoop [10] . Dans les Hadoop les données sont stockées sous forme d'une multitude de fichiers distribués et seront regroupées et structurées lors de la phase d'analyse. Cette méthode de conservation est recommandée pour de gros volumes de données qui ne nécessitent pas une structure analytique a priori, c'est le rôle du « data warehouse » "qui reste la structure la mieux adaptée à l'analyse répétitive et comparative des données structurées de l'entreprise"[23].

5 LES BIG DATA DANS LES HUMANITÉS NUMÉRIQUES OU BIG DIGITAL HUMANITIES

5.1 De Big Data au "smart data"

Si l'expression "Big Data" s'est répandue dans les sciences expérimentales et les médias depuis 2011, comme si une quantité accrue de données disponibles et parut comme la prochaine percée scientifique [5], dans les sciences humaines et sociales en contre partie, nous ne pouvons pas parler strictement de Big Data.

Bien que le volume de données ne soit pas comparable à celui actuellement généré par les médias sociaux, les blogs et les grandes entreprises, dans les sciences humaines (et spécifiquement dans les études littéraires) nous ne pouvons parler de Big Data qu'en relation avec les technologies associées à ce phénomène, telles que l'exploration de données, la stylométrie [11] ou le traitement du langage naturel. Christof Schöch [24] pose la question de "smart big" dans les humanités (les données intelligentes). Selon lui, le terme « données intelligentes » n'est pas un terme établi ou bien défini, n'est pas très répandu et n'a pas de signification stable mais il s'agit des « données structurées ou semi-structurées ; il est explicite et enrichi car, en plus des données brutes, il contient du balisage, des annotations et des métadonnées » [12].

Si les Big Data impliquent généralement de gros volumes de texte brut, clair et quelque peu désordonné, les données intelligentes, quant à elles, impliquent généralement de plus petits volumes de texte soigneusement codés et très propres. Les données volumineuses doivent être analysées à l'aide des méthodes statistiques, telles que l'analyse de clusters [13] ou l'analyse en composantes principales et bien d'autres, tandis que les données intelligentes peuvent être analysées avec des outils spécifiques permettant de tirer parti des balisages structurels, linguistiques et contextuels. Les données intelligentes sont «propres», étant donné que les imperfections du processus de capture ou de création ont été réduites autant que possible, ainsi elles ont tendance à être «petites» en volume, parce que leur création implique une intervention humaine et exige du temps [24]. Peut-on parler ici –comme intervention humaine– du rôle du spécialiste de l'information au moment de traitement et préservation des archives numériques, les éditions numériques savantes produites à l'aide des lignes directrices de l'Initiative d'encodage comme le TEI [14] parmi les exemples typiques de « données intelligentes » construites par l'être humain.

5.2 De data au "capta": les données comme une construction humaine

La chercheuse Johanna Drucker (2011) rejette le terme «données» - latin pour «ce qui nous est donné» - et utilise à la place le terme «capta» qui signifie «ce qui a été pris ou collecté» ; il est évident que cette intervention critique met en évidence la nature impartiale et incomplète des données. Les humanistes numériques ont également souligné la temporalité des données - toutes les données ont une date de création et d'expiration - et l'erreur de séparer les données des métadonnées (c'est-à-dire des données telles que titre, auteur, thème, description, date, format, identifiant, source, langue, etc.). Les métadonnées sont tout aussi importantes, sélectives et impartiales que les données parce qu'elles sont produites par des humains (ou plutôt par des algorithmes conçus par des êtres humains).

Les données ne doivent pas être considérées comme des vérités absolues, mais être mises en question de manière critique la croyance que les données sont intrinsèquement quantitatives - évidentes, neutres et indépendantes de l'observateur. Cette croyance exclut les possibilités de concevoir des données comme qualitatives, constituées de manière co-dépendante - en d'autres termes, de reconnaâtre que toutes les données sont capta [10].

[9]Le terme entrepôt de données (ou base de données décisionnelle, ou encore data warehouse) désigne une base de données utilisée pour collecter, ordonner, journaliser et stocker des informations provenant de base de données opérationnelles et fournir ainsi un socle à lâĂŹaide à la décision en entreprise

[10]Hadoop est la principale plateforme du Big Data. Utilisé pour le stockage et le traitement dâĂŹimmenses volumes de données

[11]En analyse littéraire, étude du style par des méthodes statistiques

[12]Traduit par google : Âń Smart data is data that is structured or semi-structured; it is explicit and enriched, because in addition to the raw data, it contains markup, annotations and metadata Âż

[13]Le terme cluster désigne en anglais un groupe consonantique ; il correspond donc à une succession d'au moins deux consonnes dans un mot. Par exemple, dans "fraise" ou "tigre".

[14]Les fichiers TEI considérés comme des donnés intelligentes Âń semi-structurées Âż et qui contiennent non seulement le texte intégral, mais aussi les métadonnées associées au texte (dans la section tei Header) ; le balisage qui rend explicite la structure du texte, en identifiant des parties, des chapitres, des en-têtes, paragraphes, ainsi que les sauts de page et de ligne

Dans ce même sens, et dans cette vision humaniste, la notion des archives elle même dépasse l'idée de trésor "donné", mais c'est un processus [17]. Si les données ne sont plus données mais, construites par l'être humain, donc, elles peuvent être soumises à l'archives qui sont aussi une construction humaine.

Pour résoudre les problèmes de création et de traitement des données massives et pour arriver à les transformer de "Big Data" aux "smart data" le "crowdsourcing" se présente comme un mode d'archivage de masse.

5.3 Le crowdsourcing ou l'archivage par l'être humain

Les "Big Data" nécessitent une puissance de traitement massive, en revanche, de nombreuses organisations ne disposent pas du matériel requis pour traiter ce flux exponentiel des données. Le crowdsourcing parait alors, comme une solution plus simple et moins onéreuse pour résoudre ce problème. Il s'agit de faire impliquer un groupe de personnes dites "foule" dans les activités de traitement de ces données massives (annotation, indexation, transcription,..) La définition donnée par Burger-Helmchen va dans ce sens : «Le crowdsourcing consiste littéralement à externaliser (to outsource) une activité vers la foule (crowd) c'est-à-dire vers un grand nombre d'acteurs anonymes (à priori)». [4]

Le crowdsourcing s'applique dans le cadre des archives participatives qui donnent à l'être humain la main de participer dans l'action archivistique, c'est dans la même vision qui affirme que les données sont le produit de chaine de constructions humaines successives autour des domaines liés mais distincts (sociologie, histoire, archivistique, documentation, informatique), que leur traitement nécessite le recours aux personnes. Ces personnes peuvent être le grand public d'un côté ou des experts d'un autre. Dans ce cas, il s'agit d'appliquer «une stratégie de niche», une démarche qui permet d'impliquer un public "niche". Une niche renvoie à un expert spécialiste dans un sujet donné, il s'agit donc des membres qui possèdent une motivation intrinsèque qui leur permettent de contribuer à haute qualité dans les activités de la production du savoir [22].

6 DISCUSSION

Certes, grâce aux Big Data on arrive à la mise au point d'outils complexes permettant de traiter et de mieux visualiser, analyser et cataloguer les flux énormes de données. Mais, son couplage avec l'open data soulève des nouveaux enjeux sociaux et juridiques touchant aux problèmes d'accès, sécurité publique et de vie privée, Certaines critiques se pointent vers les limites des outils de conservation comme les data lake, le détournement de l'usage des données collectées que peuvent poser les Big Data, ce qui peut atteindre la vie privée et la liberté civile.

6.1 Les limites de la conservation par "data lake"

La conservation des données dans les data lake pose certains problèmes en effet, la data lake stocke la donnée dans sa forme brute dans les systèmes producteurs, elles ne prennent pas en compte les règles de gouvernance (sécurité, conservation, élimination). Elles n'ont aucune idée à priori de la nature des données stockées (leurs spécificités, leurs degré d'importance,..) cela risque de rendre certaines information sensibles accessible à tous [8].

En outre, les data lake permettent de stocker n'importe quel format sans limite de quantité ni l'utilisation d'une hiérarchie ou une catégorisation entre les données, ce qui conduit à de nombreux problèmes d'accès aux données vu l'ignorance des valeurs des données.

Ainsi, la conservation des données dans les lacs de données ne respecte pas leur provenance, les données sont stockées de n'importe quelle façon et quelque soit leur source ce qui risque de perdre leur chemin d'accès et pose des problèmes à l'égard des lois.

6.2 L'accès aux données et l'influence sur la vie privée

La multiplication des données numériques soulève un enjeu concernant l'accès aux sources des données. En effet, les chercheurs travaillant sur les données numériques sont confrontés à des problématiques touchant le droit de la propriété intellectuelle et la protection de la vie privée [14].

Concernant le droit de la propriété intellectuelle, la consultation des bases de données par exemple, doit prendre en compte le respect de droit en vigueur (le droit des bases de données) ; les chercheurs qui sont intéressés par la collecte automatique des informations visibles sur le site peuvent, à partir d'un programme quiconque, collecter des dizaines de milliers des données personnelles, ce qui est illégal. L'accès à ces bases de données doit respecter certaines règles liées notamment à l'établissement d'un contrat, ou l'engagement d'une restitution et enfin suivre une longue procédure pour être mise en place dans le cadre de recherche [14].

Ainsi, « Toute collecte et tout traitement de Big Data dans le contexte socio-économique actuel impliquent une récupération massive d'informations sur les individus » [12] Qu'il s'agit d'informations sur leurs profils, leurs déplacements, leurs dépenses, leurs opinions et comportements, cette facilité de récupérer ces données d'individus est « porteuse de risques de dérives » [12] et menace la vie privée. Une étude juridique montre que 70% des données du Big Data sont des données à caractère personnel, elles sont produites par les personnes soit consciemment ou inconsciemment Certes ces données représentent un enjeu important pour les besoins d'entreprises mais elles constituent en contrepartie une source d'inquiétude pour les internautes et les consommateurs [16].

En contrepartie, et pour faire face à ces problèmes, il y a des règlements qui encadrent l'usage du Big Data. En France la collecte et l'analyse des données sont soumises à la surveillance de la Commission nationale de l'informatique et des libertés (CNIL) et la loi «Informatique et libertés n 78-17 du 6 janvier 1978 » relative à l'informatique, aux fichiers et aux libertés réglementant l'usage des données à caractère personnel. "Elle confère des droits aux personnes dont les données sont collectées et traitées et impose le respect de plusieurs obligations aux responsables du traitement de ces données" [16]

Aussi, parmi les exigences formulées par la CNIL , nous pouvons exploiter les technologies du Big Data en respectant certains principes dans l'utilisation des données personnelles à savoir la transparence sur l'usage des données stockées et analysées , le respect de la confidentialité et de la vie privée, le développement des

systèmes de sécurité pour que les données ne soient pas piratées. D'autres règlements entreront en application concernant la protection des données personnelles, à l'égard du règlement européen sur la protection des données personnelles [15]. [13].

7 CONCLUSION

Certes, avant les Big Data il y'avait les archives numériques. Cependant, si les archives numériques répondaient à des obligations de conservation, avec les big, elles s'engagent à de nouveaux rôles et de nouvelles fonctionnalités à l'égard de la mise à disposition du capital informationnel. « La révolution des données » apporte des nouveaux changements au niveau des concepts archivistiques, conservation et traitement des données, des nouvelles fonctionnalités ont été apportées aux archives à l'égard des recherches rapides sur les métadonnées associées aux fichiers, de l'accès plus rapide aux données et à l'information, de génération automatique des classifications les plus pertinentes.

A l'ère des Big Data, les notions et les concepts sont changés, la notion de documents a dépassé l'idée de ressource annotée, notion répandue à l'ère des humanités numériques, vers la donnée manipulée.

La notion des Archives elle-même s'élargit au " Big Archive" là où le temps s'élargit au "choc", à la désorientation, au non linéarité, remplaçant ainsi le principe de provenance et la sûreté de l'Ordre original. [18].

Sous l'impulsion des Big Data, la notion d'archivage elle-même s'est changée, en effet, avec les données instables et hétérogènes, l'archivage devient une action de conservation en place, traduisant, ainsi, un "processus dynamique" dont le rôle dépasse la fixation de contenu à la préservation de l'usage. Passant d'un système d'archivage électronique au lac des données ou réservoir des données, l'archivage passe alors d'un système structuré de conservation à un lac sans structure avec un ordre arborescent.

Si les Big Data sont perçues comme une approche dédiée aux données non structurées, l'archivage est perçu comme une approche dédiée aux données structurées. La fusion entre ces deux approches peut-t- elle réaliser l'objectif majeur : un archivage pérenne patrimonial à valeur probatoire ? Si les Big Data affirme la notion des "données" qui doivent être conservées telle qu'elles sont, les "smart data" ajoutent la notion des données "construites par l'être humain" justifiant l'action archivistiques de structurer, de gérer, d'annoter, de "méta-donner" les données et donc de valoriser la valeur des archives en tant que construction humaine.

REFERENCES

[1] Bruno Bachimont. 2017. l'archive et la massification des données : une nouvelle raison numérique. *La Gazette des archives* 245 (2017), 27–43. https://chartes.hypotheses.org/790

[2] Michèle Battisti. 2015. Big data et Smart culture. *I2D, Information, données et documents* 52, 2 (2015), 15–17. https://www.cairn.info/revue-i2d-information-donnees-et-documents-2015-2-page-15.htm

[15] Entrera en application le 25 mai 2018. Ce règlement prévoit une transparence accrue sur l'usage des données personnelles collectées. En particulier, de nouvelles obligations s'imposeront aux opérateurs collectant des données personnelles: ceux-ci auront l'obligation de s'assurer du consentement des individus (et d'être en capacité de le prouver) pour la collecte et le traitement de leurs données. Ils devront également mettre en place tous les dispositifs nécessaires pour sécuriser ces données contre des risques comme la perte, le vol ou encore la divulgation

[3] Cédric Bédini, Guillaume Benamma, and Lionel Bender. 2016. Big Data entre risque et opportunité/Groupe de veille et d'analyse- 19 session nationale spécialisée 2015-2016 'protection des entreprises et intelligence économique'. https://inhesj.fr/sites/default/files/fichiers_site/les_publications/les_travaux_des_auditeurs/big_data.pdf

[4] Thierry Burger-Helmchendu. 2011. Crowdsourcing : définition, enjeux, typologie. *Management and Avenir* 41, 1 (2011), 254–269. https://www.cairn.info/revue-management-et-avenir-2011-1-page-254.htm

[5] Antonio Rojas Castro. 2017. *Big Data in the Digital Humanities. New Conversations in the Global Academic Context.* Technical Report 4. AC/E Digital Culture. 62–71 pages.

[6] Marie-Anne Chabin. 2004. Document trace et document source. La technologie numérique change-t-elle la notion de document ? *Revue I3, Information Interaction Intelligence* 4, 1 (2004), 62–71. https://archivesic.ccsd.cnrs.fr/sic_00001020/document

[7] Marie-Anne Chabin. 2014. Les mégadonnées auront-elles raison du big data ? http://marieannechabin.blog.lemonde.fr/2014/08/26/les-megadonnees-auront-elles-raison-du-big-data/

[8] Marie-Oldi Charaudeau, Alexi Fritel, Charles Huot, Philippe Martin, and Laurent Prèvel. 2015. Et demain? Archivage et Big Data. *La gazette des archives* 240 (2015). https://www.persee.fr/doc/gazar_0016-5522_2015_num_240_4_5319

[9] Service d'Archives de CIG. 2014. Système d'archivage électronique: fiche pratique. https://www.cigversailles.fr/sites/default/files/nuxeo/FP9_SAE_mars_2013.pdf

[10] Johanna Drucker. 2011. Humanities Approaches to Graphical Display. *DHQ: Digital Humanities* 5, 1 (2011). http://www.digitalhumanities.org/dhq/vol/5/1/000091/000091.html

[11] Mariam Esteva, Jeffrey Felix Tang, Weijia Xu, and Karthik Anantha Padmanabhan. 2013. Data Mining for 'Big Archives' Analysis: a Case Study. https://onlinelibrary.wiley.com/doi/pdf/10.1002/meet.14505001076

[12] Paola Tubaro et Antonio Casilli. 2017. Enjeux sociaux des Big Data. *Les Big Data à découvert* 36, 7 (2017), 292–293. https://hal.archives-ouvertes.fr/hal-01456369/document

[13] LCL Banque et Assurance. 2017. Big data : Déefinition , Enjeux et application. https://www.lcl.com/guides-pratiques/zooms-economiques/big-data-banque.jsp

[14] Étienne Ollion and Julien Boelaert. 2016. Au-delà des Big Data. *Sociologie* 6, 3 (2016). https://journals.openedition.org/sociologie/2613

[15] Press Gil. 2013. A Very Short History Of Big Data. *Forbes* (2013). https://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/#4c891ea665a1

[16] Merav Griguer. 2017. Les enjeux juridique du Big Data. http://www.tendancedroit.fr/focus-sur-les-enjeux-juridiques-du-big-data/

[17] Eric Ketelaar. 2006. (Dé) Construire l'archive. *Matériaux pour l'histoire de notre temps* (2006). https://www.cairn.info/revue-materiaux-pour-l-histoire-de-notre-temps-2006-2-page-65.htm

[18] Andrew Lau. 2009. Review:The Big Archive: Art From Bureaucracy by Sven Spieker. *InterActions: UCLA Journal of Education and Information Studies* 5, 1 (2009). https://escholarship.org/uc/item/76c6k66f

[19] Jacques Leret. 2017. Une Architecture simple pour la Pérennité des Archives éélectroniques Bonnes Pratiques et Recommandations. http://www.opusconseils.com/var/fichiers/opusconseils-architecturesimplepourlarchivageelectronique.pdf

[20] Jolly Martyn. 2009. Big Archives and Small Collections: Remarks on the Archival Mode in Contemporary Australian Art and Visual Culture. *Public History Review* 21 (2009), 60–80.

[21] Bertrand Müller. 2011. Archives et temps présent : considérations inactuelles. In *Temps présent et contemporanéité.* https://halshs.archives-ouvertes.fr/halshs-00769732/document

[22] Ariane Néroulidis. 2015. *Le Crowdsourcing appliqué aux archives numériques: concepts, pratiques et enjeux, mémoire de recherche à l'ENSSIB.* Master's thesis. ENSSIB.

[23] Philippe Nieuwbourg. 2017. ULe concept de 'data lake' - lac de données : explication de texte. https://www.decideo.fr/Le-concept-de-data-lake-lac-de-donnees-explication-de-texte_a6976.html

[24] Christof Schoch. 2013. Big? Smart? Clean? Messy? Data in the Humanities. *Journal of Digital Humanities* 2, 3 (2013), 2–13. https://hal.archives-ouvertes.fr/hal-00920254/document

[25] Peter Stockinger. 2015. Les apports de la sémiotique dans la valorisation des données numériques : L'exemple des archives audiovisuelles. (2015). https://hal.archives-ouvertes.fr/hal-01227616/document