Comité stratégique information et communication numérique
Juin 2015

ATTENTES ET IMPACTS DE LA NORMALISATION

Livre blanc

Données massives - **Big Data**
Impact et attentes pour la normalisation

Animateur du groupe de travail Big Data :
Charles HUOT (TEMIS/APROGED)

Rapporteur : Jean-François LEGENDRE

afnor NORMALISATION

SOMMAIRE

Synthèse du livre blanc 4

Public visé par ce livre blanc 8

Introduction 9

Domaine d'application 10

Définitions 11

1 Enjeux 15

1.1. Perspectives et opportunités 15

1.2. Les freins 17

2 État de l'art 19

2.1. Un cadre conceptuel pour une architecture « Données massives/ Big Data » 19

6.1.1. La vue « utilisateur » : rôles et sous-rôles 20

6.1.2. Liste des activités identifiées 22

2.2. Architecture fonctionnelle 24

6.2.1. Lien avec la R&D : les projets des pôles de compétitivité, la R&D communautaire, les initiatives en code source libre (open source) 25

2.3. Architecture technique et interfaces 26

6.3.1. L'état de l'art 26

6.3.2. Exigences en découlant pour la normalisation 30

2.4. Les cas d'usage 31

3.3.3. Les besoins fonctionnels 34

3.3.4. Liens avec les secteurs 34

2.5. Les questions d'intérêt pour les acteurs français 37

6.5.1. Les formats de données 37

6.5.2. La qualité des données 38

6.5.3. La propriété de données 38

6.5.4. Les licences et l'intégration de données hétérogènes 39

6.5.5. L'évolution des ontologies 40

3 L'enjeu stratégique de la gouvernance des données 43

3.1. L'administrateur général des données 43

3.2. L'organisation des données 44

3.3. La gouvernance des référentiels de méta-données 44

3.4. Aspects de protection des données à caractère personnel 45

4 La réglementation 47

5 Cartographie 51
5.1. Travaux et initiatives en cours relatives à la normalisation du big data 51
5.1.1. Les responsabilités d'acteurs français au niveau des instances de normalisation 53
5.1.2. Les acteurs français participant aux instances de normalisation et leur degré de participation 53
5.2. État de l'art sur les normes de sécurité internationales applicables au traitement Big Data 55
5.2.1. Les normes de la série ISO 27001 55
5.2.2. La norme ISO 29100 56

6 Recommandations pour la normalisation 57

ANNEXES 61

Bibliographie 62

Les contributeuts aux travaux 63

Liste des logiciels 64

[Synthèse]

BIG DATA : IMPACT ET ATTENTES DE NORMALISATION VOLONTAIRE

Livre blanc du Comité stratégique AFNOR « information et communication numérique » - Juin 2015

CONTEXTE

En traduisant n'importe quelle donnée en bits informatiques, partageables sur des réseaux sociaux, stockables à distance avec le cloud computing, les nouvelles technologies de l'information et de la communication ont ouvert une nouvelle ère, celle du big data : la production et la gestion de « données massives » (l'équivalent français de l'expression). Par « massives », on entend de grandes quantités d'informations (exprimées en milliards d'octets), produites sur des supports variés (capteurs, téléphones, etc.), avec des outils extrêmement véloces permettant d'envisager des décisions en temps réel.

Ainsi défini, le big data constitue un enjeu fort pour tous les acteurs économiques. Pour les entreprises, le sujet est vu comme un moyen de mieux maîtriser leur marché, de conquérir de nouveaux prospects ou de mieux cerner les attentes de leurs clients actuels. L'information devient un actif stratégique. Par ailleurs, le big data est lui-même un objet de création de richesse, avec des sociétés se positionnant sur des activités de collecte, de vérification, de traitement, d'exploitation, d'archivage de données.

Le big data ne saurait exclure les acteurs publics, qui disposent là d'une opportunité pour proposer de nouveaux services aux citoyens. En particulier dans les domaines en rupture technologique, comme la ville intelligente, l'e-santé ou les smart grids. Mais rapidement, émergent plusieurs autres enjeux : celui du contrôle, de la fiabilité, de la propriété intellectuelle des données, et celui de l'interopérabilité tant au niveau du processus de collecte que de l'extraction de l'information et la restitution des résultats. La normalisation volontaire est un moyen efficace d'y parvenir.

UN LIVRE BLANC POUR QUI, POUR QUOI ?

Le néophyte en big data appréciera ce livre blanc pour comprendre les enjeux du sujet, identifier les acteurs politico-économiques en présence et les normes volontaires à sa disposition.

Les experts du sujet y trouveront des analyses pour orienter leurs choix de développement et identifier les opportunités de normalisation dans lesquels investir pour apporter des réponses aux nombreuses questions posées.

MÉTHODE

Le présent livre blanc a été élaboré au nom du Comité stratégique AFNOR information et communication numérique[1] par un groupe de travail de 32 personnes impliquées dans l'écosystème big data (liste en page 63). Ce groupe était animé par Charles Huot (Temis/Aproged), avec comme rapporteur Jean-François Legendre (AFNOR).

Il s'appuie notamment sur une étude qualitative des besoins de normalisation volontaire au moyen d'un questionnaire, diffusé en 2014 au sein de la communauté française du big data, en particulier le réseau Alliance Big Data. 43 questionnaires ont été remplis et retournés.

QUE FAUT-IL RETENIR ?

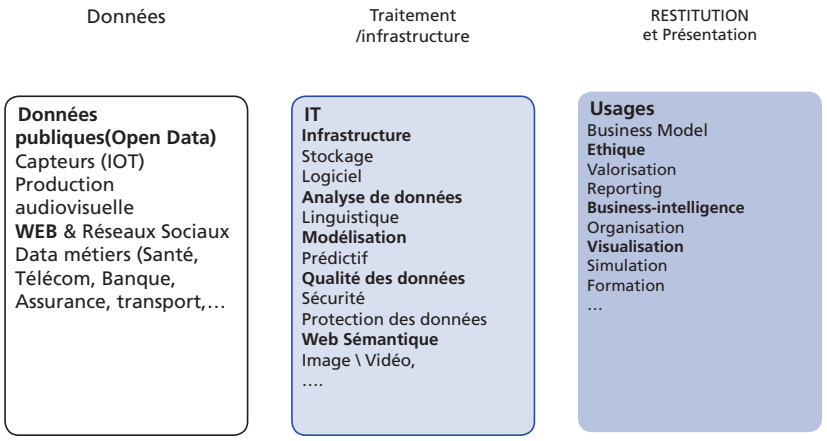
- Le secteur du big data se structure en trois grands métiers : production et collecte, traitement et infrastructure, restitution et présentation.
- Ne pas maîtriser les processus dessine un risque de monopole de la part des grands acteurs du numérique concernant la gestion des données massives et les métiers de la relation client.
- La normalisation volontaire est souhaitée, pour sa capacité à proposer des interfaces, des pratiques et des modes d'organisation partagés. Un point fondamental quand le big data exige qualité, traçabilité, interopérabilité.
- Les métadonnées, c'est-à-dire les données servant à décrire ou documenter d'autres données, nécessitent plus encore de parler un langage commun, car ce langage désigne des données diverses provenant de métiers divers.

1 -Le Comité stratégique assure la gestion collective des programmes de normalisation volontaire dans le secteur concerné.

LES CONCLUSIONS DU LIVRE BLANC

Un écosystème complexe

Le big data est gouverné par la règle des 4 V : volume, vélocité, variété, véracité. Le défi est d'accéder et de pouvoir exploiter ces données en s'affranchissant des contraintes liées à la forme et à l'origine de l'information (données internes des entreprises, internet des objets, web et réseaux sociaux, voix, image, etc.). Le monde du big data est structuré en trois grands métiers (production et collecte, traitement et infrastructure, restitution et présentation) et se travaille en vue utilisateur (avec rôle et sous-rôles).



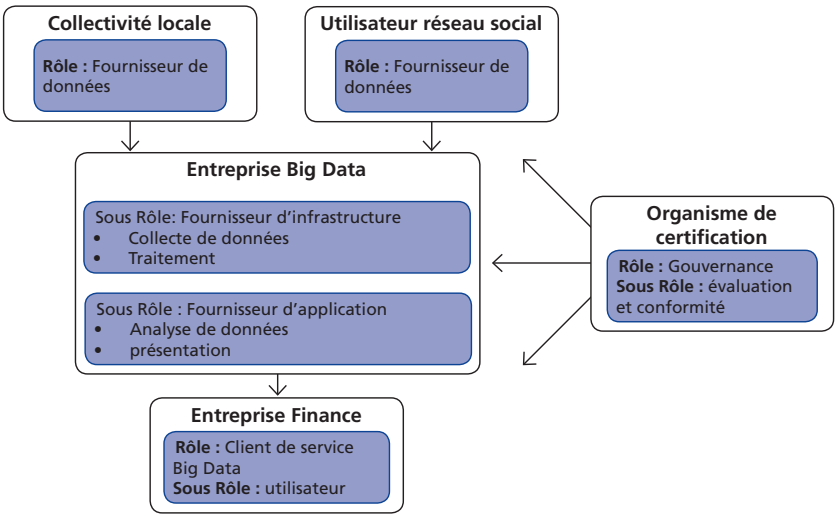
Les trois domaines du big data (p. 19)

Des architectures et interfaces basées sur le cloud

Les premiers projets industriels de big data sont à l'initiative des spécialistes de la requête sur le web. Sont ainsi apparus Google BigTable et surtout Apache Hadoop, la technologie de référence des big data, en java open source. Chaque opérateur utilise cette librairie pour apporter sa propre valeur ajoutée : IBM, EMC, Hortonworks, Oracle, SAP, etc. Au centre des architectures big data figurent aussi le modèle noSQL, qui s'affranchit des bases de données orientées en colonnes. Enfin, le big data met fin au modèle de la gestion de données internes à l'entreprise, où des statisticiens et des data-analystes géraient des « data warehouses » : l'externalisation est désormais la règle, grâce à l'apport du cloud computing. Les données migrent sur des serveurs distants, les applications et traitements également.

Des données hétéroclites à structurer

Comme l'indexation en temps réel d'information peu ou pas structurée est une caractéristique forte du big data, l'apport des technologies sémantiques est déterminant : des informations de toute nature et de tout format sont capturées et structurées à la volée en s'appuyant sur des référentiels métiers (banque, édition, santé, etc.) et des relations sémantiques. C'est ce qu'on appelle les ontologies de domaine. Se pose alors la question de la norme volontaire : dans cet univers hétéroclite, tout le monde a-t-il besoin de parler un langage commun ? En matière d'infrastructures, l'absence de normes volontaires ne semble pas être un obstacle majeur à l'adoption généralisée d'Hadoop. Les systèmes NoSQL souffrent assurément d'une absence de normes volontaires et il serait utile de définir un langage de base unifié pour la requête. Des acteurs ont essayé de pousser en ce sens.



Exemple d'un écosystème big data (p. 21)

Des besoins exprimés pour la normalisation

Pour réaliser ce livre blanc, une étude qualitative des besoins a été menée au moyen d'un questionnaire diffusé auprès des porteurs de projets big data. Ils ont exprimé des besoins de définir collectivement des repères partagés sur des sujets

très variés : besoin de recourir à des prestataires internes ou externes, diversité des données traitées, grand nombre de processus, confidentialité, protection intellectuelle et respect de la vie privée ressortent comme des enjeux largement cités. La pseudonymisation des données personnelles ou industrielles est un point clé : comment garantir des algorithmes non réversibles ? En matière de traitement, les enjeux soulignés concernent la définition des métadonnées, pour faciliter l'exploitation et la catégorisation, ainsi que la traçabilité des opérations effectuées sur ces données.

L'interopérabilité sémantique

Dans un contexte marqué par l'essor de cinq types de données non structurées (texte, image, image animée, son, données de capteurs), l'enjeu de la fusion des données devient majeur. Deux niveaux d'interopérabilité sémantique apparaissent : l'un relatif aux contenus, l'autre aux contenants (schémas XML, etc.). L'initiative internationale Research Data Alliance, à laquelle participe le pôle de compétitivité français Cap Digital, est ici à souligner pour mieux cadrer les ontologies et la sémantique des contenus. La normalisation peut apporter des solutions complémentaires, à l'image des normes volontaires de systèmes d'enregistrement que développe le comité technique ISO TC 46 sur la documentation (ISAN, ISBN, etc.). Dance ce contexte, il y a lieu de mener un travail sur l'élévation des données que permettent les ontologies, c'est-à-dire l'extraction de leur format d'origine pour leur conférer l'interopérabilité. Il existe certes le LOV (Linked Open Vocabulary), une sorte de catalogue d'ontologies de plus de 450 références (ontologies du tourisme, de la météo, de la santé, etc.), mais il ne fait l'objet d'aucune norme volontaire et n'est pas encore reconnu dans le cadre de l'ISO.

L'enjeu stratégique de la gouvernance des données

Le big data fait émerger une fonction nouvelle et transverse aux métiers au sein des organisations, celle d'administrateur des données. À lui de répondre aux questions suivantes : qui est producteur de la donnée ? Qui en apprécie le degré de qualité ? Qui en garantit la qualité, la pérennité, l'accessibilité ? Quel circuit de validation mettre en place ? Cette organisation suppose des prérequis : existence d'un identifiant, utilisation de celui-ci par les différents producteurs, activation de cet identifiant. Dans le domaine des données culturelles, par exemple, de nombreuses normes volontaires existent en matière de métadonnées et d'identifiants. Pour les entreprises, des référentiel existent par secteurs mais ne résultent pas de démarches de normalisation et ne sont pas interopérables : PLIB pour les données industrielles, IFC dans la construction etc. Cet administrateur des données doit maîtriser la réglementation applicable aux données, qu'elles soient publiques ou privées. Pour ces dernières, il n'existe pas de régime juridique unifié encadrant la propriété.

Les recommandations pour la normalisation

L'organisme de normalisation américain NIST[2], relayé par l'ISO/CEI JTC 1[3], travaille sur l'état de l'art et les démarches de normalisation à mener dans le domaine du big data, de même que l'UIT[4]. Du côté de l'ISO, plusieurs normes volontaires de sécurité se prêtent à l'encadrement de l'écosystème big data :

- la série ISO 27001 (systèmes de management de la sécurité de l'information) ;
- la norme ISO 29100 (sécurité technique pour la protection des données).

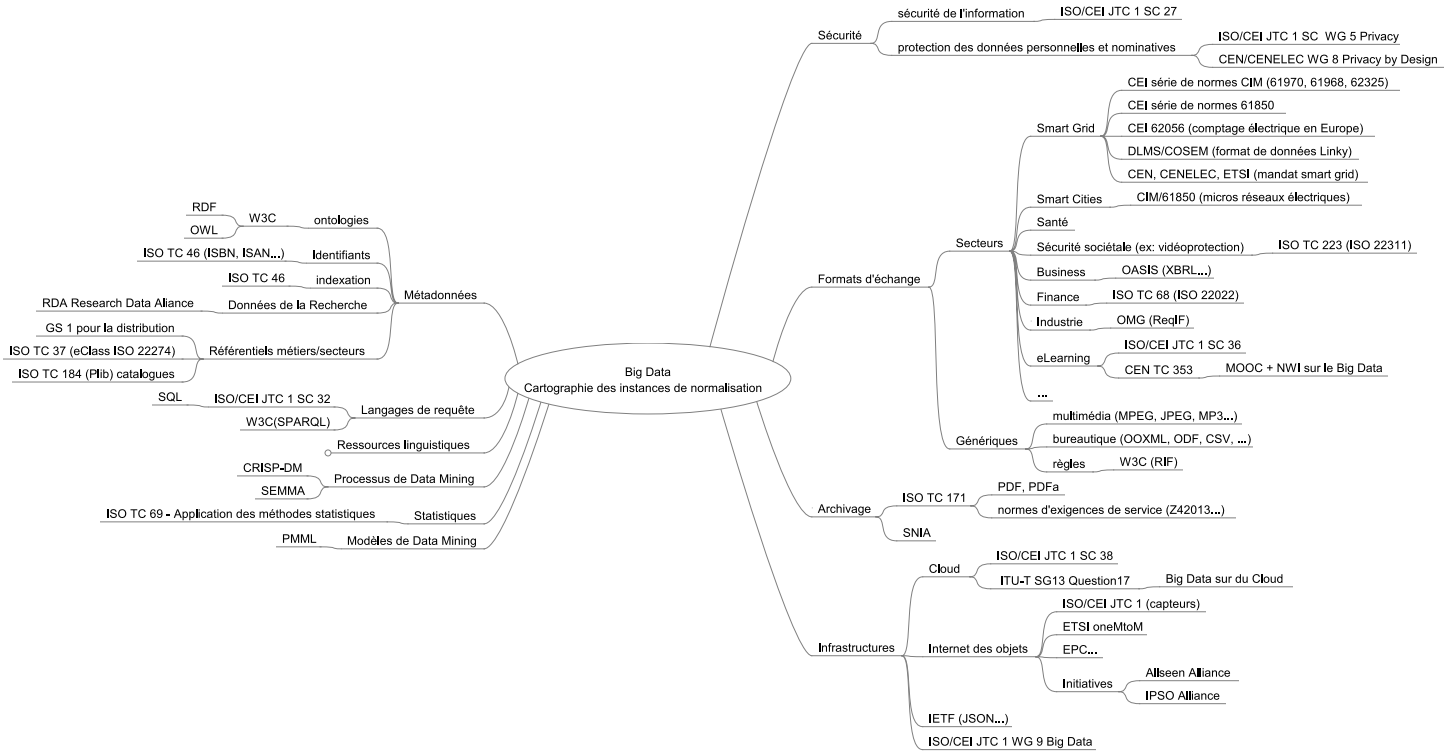
Mais la normalisation doit aller bien au-delà. En particulier, il est nécessaire de pousser une norme internationale cadrant l'architecture de référence et le vocabulaire du big data. Au total, six axes de développement ont été identifiés :

- la gouvernance de la donnée,
- la qualité et l'identification,
- les données ouvertes (open data),
- les opérateurs d'infrastructures,
- les opérateurs de service,
- la normalisation technique.

2 - NIST : National Institute of Standards and Technology

3 - Le Joint Technical Committee 1, créé en 1987 par convention entre l'ISO (Organisation internationale de normalisation) et la CEI (Commission électrotechnique internationale) est l'organe de référence pour la normalisation des Technologies de l'Information au niveau mondial.

4 - UIT-T : l'Union internationale des télécommunications est une agence des Nations unies



Cartographie des instances de normalisation (p. 52)

Participer à la normalisation volontaire

Dans un contexte de mondialisation et de concurrence accrue, les entreprises doivent à la fois renforcer leur organisation et leurs pratiques pour gagner en efficacité ; anticiper les nouvelles normes volontaires (et si possible les initier) pour s'adapter plus vite et innover ; et promouvoir leur agilité pour se différencier.

AFNOR est l'organisme français de référence sur la vie des normes volontaires. Elle recense toutes celles qui existent, anticipe celles à venir, et accompagne leur création aux niveaux français, européen et international.

Une norme volontaire est un cadre de référence, positif et vertueux, qui vise l'amélioration continue des produits, services ou pratiques, au service de l'intérêt de tous : des consommateurs, des entreprises et de la collectivité au sens large.

Elle définit les exigences et fixe les standards en matière de qualité, de sécurité, de performance.

Tout le monde peut participer à sa création. Elle est élaborée par consensus entre l'ensemble des parties intéressées. Elle est volontaire, tout acteur peut ou non s'y référer.

En véritable moteur de cette démarche, AFNOR accompagne celles et ceux qui, par leur expertise, veulent poser les bases de l'économie et de la société de demain. Un soutien unique et indispensable pour favoriser le progrès et faire rayonner la France à l'international.

www.afnor.org

Contact :
Jean-François LEGENDRE
jeanfrancois.legendre@afnor.org
01 41 62 83 57

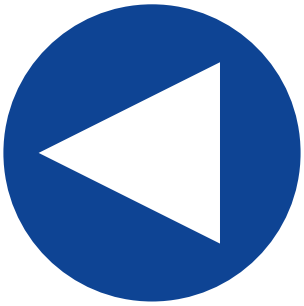
PUBLIC VISÉ PAR CE LIVRE BLANC

Ce livre blanc est destiné avant tout à sensibiliser les entreprises ainsi que les pouvoirs publics aux enjeux des données massives et à l'impact potentiel de la normalisation dans ce domaine.

Il propose quelques recommandations pour la normalisation dérivées des besoins des acteurs français et vise notamment à soutenir l'écosystème industriel qui se constitue autour de l'exploitation des données massives.

Bien qu'élaboré dans le cadre du comité stratégique sur l'information et la communication numérique, ce document concerne tous les secteurs, car les enjeux des données massives se situent bien au-delà des seuls aspects technologiques et impactent également l'organisation de nombreux métiers.

INTRODUCTION



Les perspectives économiques, industrielles, techniques et sociétales, associées à la collecte et l'exploitation de données de plus en plus massives, représentent d'importantes opportunités, mais aussi des risques qu'il convient d'appréhender, d'anticiper et de maîtriser.

L'une des caractéristiques des données massives ou *Big Data* est qu'elle concerne tous les secteurs et qu'elle interfère avec tous les domaines qui connaissent des situations de rupture technologique comme la ville intelligente, l'e-santé, les moyens de production, les réseaux intelligents, les objets connectés, etc.

Le défi est d'accéder et de pouvoir exploiter l'ensemble de l'information dans un monde de plus en plus complexe, connecté et diffus. Cela suppose de s'affranchir des contraintes liées à la forme et à l'origine de l'information.

Les technologies classiques de gestion des données, basées depuis des années sur des mécanismes transactionnels à partir du stockage et de l'interrogation de données structurées, ne suffisent plus car elles présentent des limites avec l'avènement de l'internet social (social web), de la mobilité, des « smart phones » et des tablettes, des capteurs et des objets connectés.

Pour la prise de décision, force est en effet de constater que l'analyse des données transactionnelles est rétrospective et que les analyses effectuées sur ces ensembles de données caractérisent plutôt des compréhensions de faits réalisées ou d'opinions sur le passé.

De nouvelles technologies prennent en compte les caractéristiques des données massives et permettent leur analyse en temps réel ou quasi réel. Cela rend par exemple possible la modélisation des phénomènes avec toute l'agilité dont on a besoin aujourd'hui.

En termes d'infrastructures de services, l'informatique en nuage « cloud computing » apporte la capacité d'ingérer, de stocker et d'analyser les données pour permettre aux organisations de relever les défis associés aux données massives.

Cette conjonction d'innovations offre une opportunité pour de nouveaux acteurs, ceux issus de la révolution du numérique, de pénétrer les secteurs économiques en apportant de la valeur ajoutée. Ils peuvent ainsi se positionner en arbitre ou intermédiation dans la relation entre des clients et des fournisseurs.

Tous les secteurs sont concernés, aussi bien dans la sphère publique et notamment les collectivités territoriales que dans la sphère privée y compris l'industrie.

Développer l'exploitation des données massives demande cependant des leviers.

Les normes sont des outils volontaires venant en appui des entreprises pour apporter des solutions aux enjeux des données massives. Elles sont un facteur d'ouverture des marchés et de confiance entre partenaires. Leur développement de même que leur adoption est un enjeu concurrentiel.

Dans un environnement national, européen et international de plus en plus complexe et multiforme, il importe que les acteurs prennent conscience de l'importance de ce levier et réfléchissent à une stratégie appropriée dans le domaine clé des données massives pour en faire des outils efficaces au service des différentes parties intéressées.

DOMAINE D'APPLICATION

Ce livre blanc traite des données massives ou « *Big Data* ». Par données massives, on entend l'exploitation de grandes masses d'informations (teraoctets) composées de données souvent hétérogènes (multimédia, capteurs, réseaux sociaux, téléphonie, etc.) avec des outils extrêmement véloces (permettant d'envisager des décisions en temps réel), ce qui implique le cas échéant des moyens non conventionnels (exemple : base de données NO SQL).

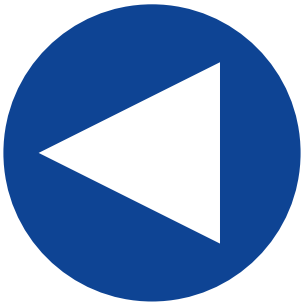
Le présent document analyse les enjeux pour la normalisation associés à la mise en œuvre de processus de collecte, traitement et exploitation de grandes masses de données, souvent hétérogènes et non structurées, par exemple en provenance d'internet, de réseaux sociaux publics ou d'entreprises, de réseaux de communication, de capteurs associés à des réseaux intelligents, des objets mobiles, des dispositifs de sécurité ou des sites de production industrielle, etc.

Pour ce qui est des enjeux normatifs, on ne se limite pas aux questions technologiques, bien qu'elles soient importantes, mais on considère l'outil normatif dans sa capacité à proposer aux entreprises et aux acteurs des interfaces, des pratiques et des modes d'organisation partagés et adaptés aux changements profonds qu'imposent les modèles économiques des données massives.

Ce document se place dans une optique générique et intersectorielle.

Ceci étant, les enjeux étant fortement liés à des besoins et des spécificités des secteurs d'application, il importe donc de prendre en compte cette dimension multisectorielle à travers un certain nombre de cas d'usage sans pour autant prétendre à rechercher une exhaustivité qui serait impossible à atteindre compte tenu de la complexité et de la diversité des approches possibles dans le « *Big Data* ».

DÉFINITIONS



Pour les besoins du présent document, le glossaire suivant a été établi :

ANALYSE DE CONTENUS (« DATA ANALYTICS »)

Selon le livre blanc de janvier 2013 de l'APROGED, l'analyse de contenu / « Content Analytics » est l'analyse et la représentation multidimensionnelle des données issues des données massives après un traitement d'extraction des contenus non structurés afin d'en faire ressortir des entités nommées, des relations inter-entités, des thématiques, des opinions.

ANONYMAT/PSEUDONYMISATION

L'ISO/CEI 29100 : Information technology - Security techniques - Privacy framework, propose les définitions suivantes :

ANONYMAT

Caractéristique d'une information qui ne permet pas l'identification directe ou indirecte du porteur à l'origine de cette information à caractère personnel.

Characteristic of information that does not permit a personally identifiable information principal to be identified directly or indirectly

PSEUDONYMISATION

Processus par lequel une information à caractère personnel est altérée de façon irréversible, de sorte que le porteur de cette information à caractère personnel ne peut être directement ou indirectement identifié, que ce soit par un automate traitant l'information à caractère personnel seul ou en collaboration avec tout autre dispositif.

Process by which personally identifiable information (PII) is irreversibly altered in such a way that a PII principal can no longer be identified directly or indirectly, either by the PII controller alone or in collaboration with any other party

ANONYMIZED DATA

Donnée qui a été produite en sortie d'un processus d'anonymisation de données à caractère personnel.
Data that has been produced as the output of a personally identifiable information anonymization process

DONNÉE

Ensemble des indications enregistrées en machine pour permettre l'analyse et/ou la recherche automatique des informations`` (CROS-GARDIN 1964).

(Ortoland - http://www.cnrtl.fr/definition/donnée)

DONNÉES À CARACTÈRE PERSONNEL

Selon la directive européenne 95/46/EC relative à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données (note : actuellement en vigueur, mais révision en cours à un état avancé) :

« Une donnée à caractère personnel est définie comme toute information concernant une personne physique identifiée ou identifiable (personne concernée); est réputée identifiable une personne qui peut être identifiée, directement ou indirectement, notamment par référence à un numéro d'identification ou à un ou plusieurs éléments spécifiques, propres à son identité physique, physiologique, psychique, économique, culturelle ou sociale. » (art. 2a)

DONNÉES MASSIVES[1] (« *BIG DATA* »)

Selon l'UIT-T, les Données massives « *Big Data* » sont définies comme un ensemble de technologies et services permettant la collecte, le stockage, le partage, l'analyse et la présentation de données fortement massives, véloces et variables.

("a category of technologies and services where the capabilities provided to collect, store, search, share, analyse and visualize data which have the characteristics of high-volume, high-velocity and high-variety".)

Un « livre blanc publié par l'UIT-T en novembre 2013 » précise les quatre caractéristiques essentielles des données massives:

- **1) Les volumes :** des tera à peta octets aujourd'hui, sans doute beaucoup plus demain.
- **2) La vélocité :** elle caractérise la rapidité avec laquelle une information est générée, délivrée, stockée et finalement enlevée puis effacée. Elle se mesure en des évènements de l'ordre de la milliseconde et la capacité à permettre des décisions en temps réel ou proche du temps réel, ce qui apparaît nécessaire pour permettre la flexibilité des organisations.
- **3) La variété :** on traite tout type et toute structure de données : textes, données de capteurs, enregistrements, cartes, son, image, vidéos, données liées issues des réseaux sociaux, fichiers informatiques et plus. Les sources des données sont également diverses.
- **4) La véracité :** la qualité, la précision et la crédibilité des données gouvernent la capacité à élaborer des décisions avec certitude dans un contexte de sélectivité entre différentes sources d'acquisition des données.

1 - Le vocabulaire préconisé par le dispositif ministériel d'enrichissement de la langue française est le terme « mégadonnées »

DONNÉES OUVERTES (« OPEN DATA »)

Une donnée dite « ouverte » est une donnée numérique d'origine publique ou privée dont la diffusion est organisée de manière structurée selon une méthodologie et une licence ouverte garantissant son libre accès et sa réutilisation par tous, sans restriction technique, juridique ou financière.

Cette notion est souvent associée aux données produites par une collectivité ou un service public (service éventuellement délégué).

MÉTADONNÉE

Une donnée sur / à propos de la donnée, une donnée servant à décrire ou à documenter une autre donnée.

ONTOLOGIE

Une ontologie est la spécification d'une conceptualisation d'un domaine de connaissance.

Une ontologie inclut généralement une organisation hiérarchique des concepts pertinents et des relations qui existent entre ces concepts ainsi que des règles et axiomes qui les contraignent. L'ontologie définit ainsi des concepts (principes, idées, catégories d'objets, notions potentiellement abstraites) et des relations.

TRAITEMENT (PROCESSING)

Selon la directive européenne 95/46/EC relative à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données :

Traiter / processing signifie *" toute opération ou ensemble d'opérations effectuées ou non à l'aide de procédés automatisés et appliquées à des données à caractère personnel, tels que la collecte, l'enregistrement, l'organisation, la conservation, l'adaptation ou la modification, l'extraction, la consultation, l'utilisation, la communication par transmission, diffusion ou toute autre forme de mise à disposition, le rapprochement ou l'interconnexion, ainsi que le verrouillage, l'effacement ou la destruction;"*. (art. 2 b)

ENJEUX

1.1 PERSPECTIVES ET OPPORTUNITÉS

L'exploitation des données massives bénéficie de plusieurs ruptures technologiques rendant possibles l'intégration, le traitement, l'interprétation et la représentation de données hétérogènes provenant de différentes sources.

Ceci permet de placer désormais les données au cœur des modèles économiques et d'apporter aux organisations une forte valeur ajoutée par un meilleur ciblage de leurs services à travers à la fois une meilleure connaissance de leur environnement et une optimisation de leurs processus.

Une convergence entre les domaines d'affaires est ainsi possible et ceci marque sans doute le début de nouveaux modèles économiques qui redéfinissent les relations entre les producteurs, les distributeurs et les consommateurs ou les biens et services.

Ceci étant, la complexité liée à ces relations s'est considérablement accrue rendant la prise de décision extrêmement difficile pour une organisation donnée.

La nouveauté dans les données massives réside dans le besoin d'exploiter de gigantesques volumes de données liés à la diversité et multiplicité des sources qui sont désormais accessibles et en particulier :

- Les données internes des entreprises,
- Les données issues de capteurs qui se multiplient avec l'internet des objets soit dans une approche B to B, soit dans une relation B to C,
- Les données issues du Web et des médias sociaux.

Le traitement permet de générer de l'information à forte valeur ajoutée pour l'ensemble de l'entreprise. La finalité est d'aider à la prise de décision, dans un contexte où l'information est devenue l'actif stratégique majeur des entreprises.

Pour les entreprises, mettre en œuvre une démarche de données massives est un enjeu de compétitivité avec la possibilité de générer de nouveaux profits et de se positionner dans de nouvelles activités.

Pour les acteurs publics, c'est disposer d'une capacité d'optimiser leur fonctionnement et de proposer de nouveaux services aux citoyens.

Pour les citoyens, c'est une possibilité d'être acteurs conscients dans l'écosystème des données massives, et de bénéficier de nouveaux services qui amélioreront leur qualité de vie.

Pour bénéficier de ce nouveau modèle économique, il est nécessaire de décloisonner les silos et de jouer sur la transversalité entre les métiers, ce qui impose une démarche de transformation de l'entreprise. Il s'agit en effet d'être en mesure d'appréhender l'information mais aussi de prendre conscience du patrimoine informationnel au sein des entreprises et des organisations.

Ceci suppose une compréhension et une maîtrise de la valeur des données, de ce qui est partageable et de ce qui ne l'est pas, dans quelles conditions, avec des enjeux en termes de propriété intellectuelle, de connaissance de la réglementation notamment en matière de données à caractère personnel.

Pour les acteurs publics, la donnée devient essentielle pour le pilotage des Territoires. La donnée se retrouve au cœur de nombreux concepts émergents (Open Data, *Big Data*).

Elle représente un enjeu pour les collectivités à travers :

- Un meilleur accès aux données qui permet de contrôler les missions confiées aux délégataires ou partenaires,
- Une connaissance de l'utilisation et de l'état des services publics qui facilite l'optimisation des ressources, pour les investissements ou la maintenance,
- Une meilleure information des consommateurs, notamment à des fins de sensibilisation.

Les entreprises comme les collectivités territoriales sont cependant confrontées à des questions techniques :

En premier lieu, la qualité de l'information résultant d'une démarche *Big Data* est directement liée à la qualité des jeux de données en entrée.

La qualité des données est un enjeu important en raison de l'exploitation de données de différentes sources souvent non homogènes, ce qui a des répercussions sur les processus de traitement analytique et sémantique.

Un effet induit peut être de noyer les données pertinentes au sein de données inutilisables ou inutiles.

En second lieu, il est utile de pouvoir mettre en œuvre une interopérabilité qui se traduit dans les processus de capture des données, dans la mise en œuvre de référentiels de métadonnées, dans les processus de filtrage et d'extraction de l'information ainsi qu'au niveau de la restitution des résultats.

En troisième lieu, il est nécessaire d'assurer une sécurité qui concerne les données afin d'en garantir l'intégrité et la confidentialité, et qui concerne le processus « workflow » afin notamment, de disposer de mécanismes de non-répudiation (à la source) et d'approbation (du résultat).

Tout ceci nécessite un apport de compétences spécifiques (analystes, statisticiens, juristes de la donnée, …), mais aussi des infrastructures et des technologies particulières.

1.2

LES FREINS

La constitution de nouveaux monopoles

Les grands acteurs du numérique comme Google, Apple, LinkedIn, Facebook, Amazon, etc. dont aucun n'est français, offrent aux entreprises un accès direct aux données du grand public et parfois même à leurs propres données (accès aux intranets des entreprises…).

Le risque pour les acteurs traditionnels qui ont besoin des outils proposés par ces fournisseurs pour se développer est que ces grands acteurs utilisent cet actif et leur position dominante pour se positionner en inter-médiateur dans la relation clients, puisqu'ils cherchent à imposer aux clients leurs propres services par exemple dans l'assurance, le crédit, etc.

Le contrôle sur les données

Les entreprises sont conscientes de l'importance des données qu'elles détiennent. Le caractère stratégique peut imposer que ces données ne puissent être partageables. Ceci limite intrinsèquement leur exploitation dans un contexte de données massives du fait que celui-ci accroît le risque de fuite d'informations.

D'une façon plus générale, les organisations sont susceptibles d'avoir une sensibilité renforcée aux évènements et aux incidents de sécurité, puisque les dispositifs d'exploitation de données massives imposent des systèmes très intégrés (même si l'infrastructure technique est largement distribuée) et font appel à des prestataires spécialisés.

Lorsqu'elles sont partageables, les données posent la question des licences d'exploitation et de la propriété intellectuelle associée, ainsi que celle de la traçabilité de leur usage. Les notions de droits d'usage des données conduisent aussi à des questions de non-répudiation.

L'enjeu de traçabilité des données et des traitements est d'autant plus important dans un contexte de données ouvertes. Ceci concerne en particulier l'exploitation de données culturelles, celle des données de la Recherche, et induit celle de l'identification des auteurs ou/et des chercheurs. Il existe déjà des pratiques, des règles et des normes, mais dans des domaines limités que le « *Big Data* » bouleverse.

Pour les acteurs publics, la question de la propriété Intellectuelle peut dériver sur celle de l'acceptation des utilisateurs et des exploitants pour délivrer les données dans un entrepôt public. Ainsi, les conditions de gestion (exemple : mutualisation de Systèmes d'Information Géographique) et de transmission des données deviennent un élément déterminant d'un projet d'exploitation de données massives entre entités publiques, ou publiques et privées.

La sécurité juridique et l'éthique

La manipulation de données à caractère personnel conduit à des enjeux sur les processus d'anonymisation et soulève la question de leur cryptage non réversible.

D'autre part, le risque d'image peut être important pour certains acteurs (collectivités publiques, organismes financiers, etc.) en raison des effets et des dérives « Big Brother » liés à la collecte massive de données.

Ce risque vaut aussi pour les grandes entreprises qui souhaitent maîtriser leur réputation sur les réseaux sociaux et à travers celle-ci leur relation client.

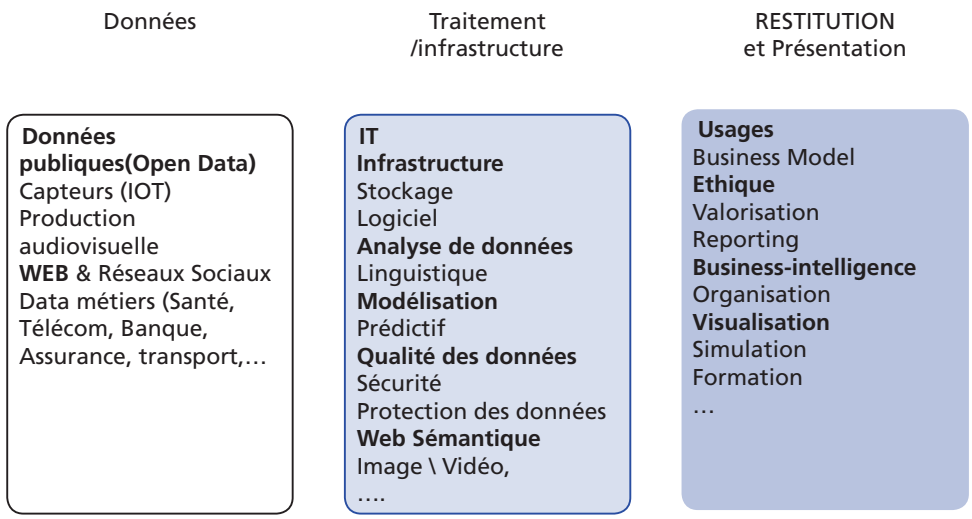
Les enjeux sociétaux peuvent aussi se traduire par des contraintes sur les processus de présentation de l'information et leur usage.

ÉTAT DE L'ART

UN CADRE CONCEPTUEL POUR UNE ARCHITECTURE « DONNÉES MASSIVES - *BIG DATA* »

2.1

Le contexte des données massives se structure en trois grands domaines :



Cette représentation peut être complétée par la définition plus formelle d'une architecture de référence pour traiter les données massives. Ce qui suit s'inspire d'une méthodologie déjà utilisée pour modéliser une informatique en nuage ou « cloud computing » afin de mettre en évidence les besoins fonctionnels et mieux préciser les besoins normatifs en tenant compte de l'existant déjà exploitable.

Cette représentation fait intervenir :

- une notion d' « acteur» « partie intéressée » ou « partie prenante »
- une notion de « rôle »
- une notion de « sous-rôle »
- une notion d' « activité »

Elle permet dans un deuxième temps de définir le modèle fonctionnel en précisant les couches d'abstraction concernées (représentation couramment employée dans les télécommunications et le numérique).

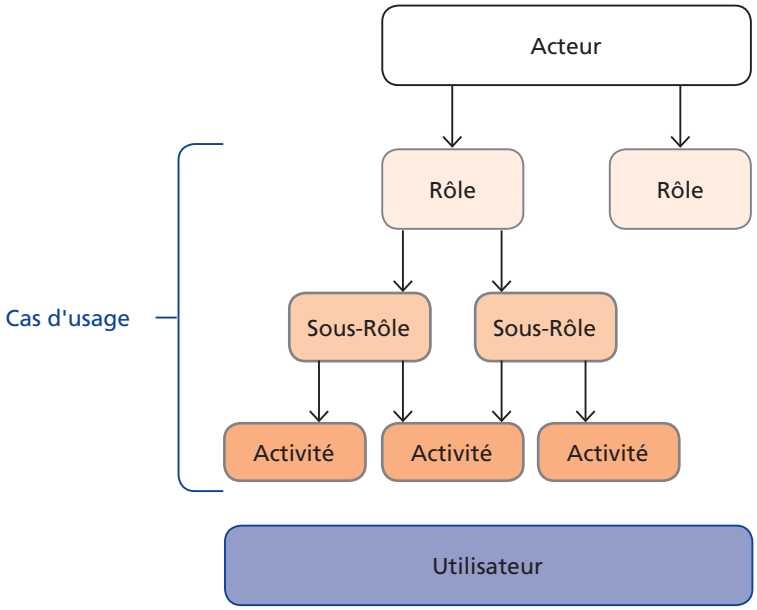
Dans cet effort de modélisation de systèmes complexes (les données massives en sont un exemple), il est utile de référencer des cas d'usage et de vérifier que la modélisation proposée en permet la représentation.

2.1.1

2.1.1 - LA VUE « UTILISATEUR » : ROLES ET SOUS-ROLES

L'écosystème relatif aux données massives est décrit suivant le modèle de la vue « utilisateur » illustrée ci-après.

La vue « utilisateur » consiste à définir l'écosystème des « données massives - *Big Data* » en distinguant les notions d'acteurs et de rôles joués par ces mêmes acteurs suivant un cas d'usage.



Vue utilisateur

C'est ainsi que sont définis quatre rôles principaux dans la fourniture de services « *Big Data* » et ce indépendamment du type de services considéré :

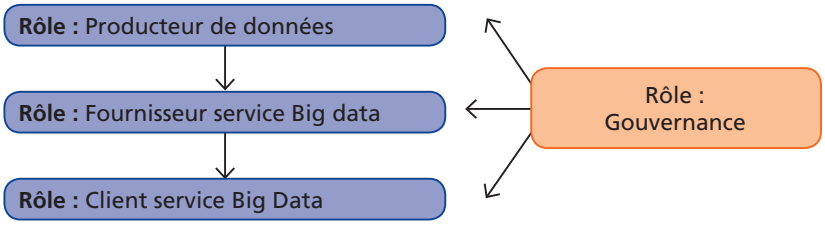
- le fournisseur de données ;

- le fournisseur de service « *Big Data* » ;

- le client de service « *Big Data* » ;

- la gouvernance (pour la protection de données personnelles).

Ecosystème Big Data/Données massives



Ces quatre rôles se décomposent en sous-rôles correspondant chacun à un jeu d'activités (ou ensemble de tâches) intervenant lors de la réalisation de cas d'usage.

Exemples typiques de sous-rôles :

Le rôle « fournisseur de service *Big Data* » comporte deux principaux sous-rôles, à savoir :

- Le « fournisseur d'infrastructure *Big Data* » : ce sous-rôle offre des capacités de stockage, de calcul et de réseau, assure la collecte des données ainsi que leur traitement et la sécurité associée;
- Le « fournisseur d'applications *Big Data* » : ce sous-rôle couvre les activités d'analyse et de présentation des données et assure également la protection et la sécurité des données.

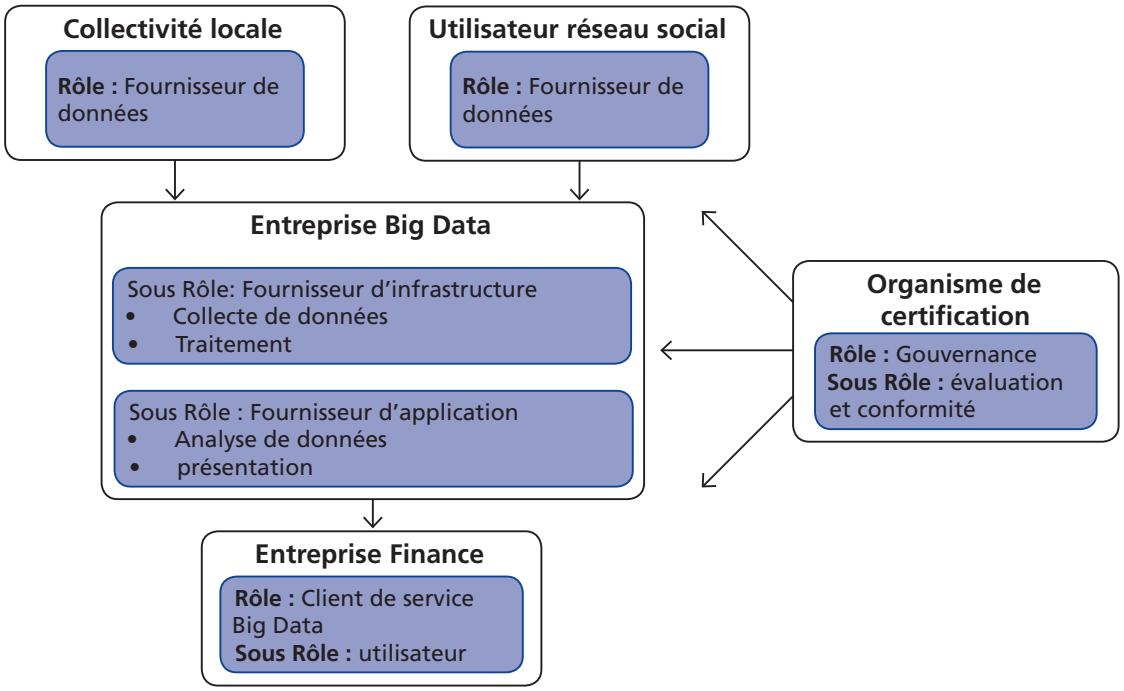
Le rôle « fournisseur de données » consiste à exposer les données de manière sécurisée et protégée et le modèle économique associé.

Le rôle « client de service *Big Data* » comprend les sous-rôles d'utilisateur des applications « *Big Data* » et d'administration des comptes utilisateurs.

Le rôle « gouvernance » comprend les sous-rôles de :

- « conseil en matière d'expertise juridique ou technique »,
- « autorité de régulation pour l'autorisation, le contrôle et les sanctions vis-à-vis de la régulation »,
- « évaluation et conformité (audit, certification) »,
- etc.

La figure suivante illustre un exemple d'écosystème « *Big Data* » :



2.1.2 - LISTE DES ACTIVITES IDENTIFIEES

Un exemple concernant les activités relatives aux rôles :

- Rôle : Fournisseur de données
  - Capture de la donnée (capteurs, applications, SI, …),
  - Mise à disposition d'une place de marché de la donnée (catalogue de données et leurs caractéristiques (ex : métadata), négociation contractuelle avec le fournisseur de service *Big Data*…),
  - Contrôle de la qualité de la donnée à la source,
  - Acquisition du consentement de l'individu pour la mise à disposition des données personnelles,
  - Pseudonymisation des données personnelles ou commerciales avant mise à disposition,
  - Autorisation/consentement sur la mise à disposition des données (licences),
  - Mise à disposition informatique des données.
- Rôle : Fournisseur de service *Big Data*
  - Sous Rôle : Fournisseur d'infrastructure *Big Data*
    - Collecte des données auprès des fournisseurs,
    - Gestion des métadonnées (catalogues, ontologies, …),
    - Contrôle et nettoyage des données,
    - Pseudonymisation des données (si désigné comme tiers de confiance),
    - Indexation, agrégation, intégration multi-sources (ex. synchronisation temporelle, mise à la maille spatiale, …) et multi-types des données (ex. traitement automatique des langues, analyse des images, …),
    - Historisation des données,
    - Stockage des données,
    - Sécurisation des données (accès, pérennité),
    - Gestion du cycle de vie de la donnée (archivage, destruction),
    - Traçabilité de l'historique des opérations sur les données.
  - Sous Rôle : Fournisseur d'application *Big Data*
    - Mise à disposition de données (catalogue, gestion des licences gratuites ou payantes, accès aux données, facturation),
    - Visualisation de données,
    - Services à valeur ajoutée sur les données : data mining exploratoire et prédictif, traitement en flux, …
- Rôle : Client de service *Big Data*
  - Acquisition des données nécessaires au service (auprès d'un fournisseur de service *Big Data*),
  - Acquisition des données du client nécessaires au service,
  - Opération et facturation du service (service métier et non un service lié à la gestion des données),
  - Mise à disposition des données acquises au travers du service (cf. devient producteur de données),
  - Passe un contrat avec le fournisseur de service,
  - Donne/refuse son consentement sur l'utilisation de ses propres données,
  - Approuve/valide le résultat de l'analyse.

- Rôle : Gouvernance
  - Protection de la confidentialité des données,
  - Régulation de l'utilisation des données,
  - Certification/labellisation,
  - Désignation des tiers de confiance,
  - Sous-rôle Conseil : expertise technique et juridique en protection de la confidentialité des données,
  - Sous-rôle Autorité : Régulateur de l'utilisation des données,
  - Sous-rôle Evaluateur : Conformité/Audit/Certification.

2.1.2.1 Processus de collecte, traitement et analyse de contenus

Liste des activités mises en évidence par l'étude (questionnaire en annexe) :

- Extraction (langage d'extraction),
- Collecte,
- Filtrage,
- Indexation,
- Définition des référentiels (métadonnées),
- Élévation (ontologies),
- Pseudonymisation des données personnelles (ou industrielles),
- Rapprochement des sources, sédimentation, agrégation, « data cleaning »,
- Stockage, archivage,
- Normalisation/harmonisation/tokenisation (des données),
- Traitement à la volée (stream processing),
- Conversions (exemples : voix en texte, OCR pour des documents papiers),
- Exploration des données,
- Analyse linguistique, traitement automatique des langues, text mining,
- Analyse prédictive / Data Mining (ex: service de prévision),
- Analyse des réseaux sociaux,
- Effacement des données,
- Autres (ex: compression...).

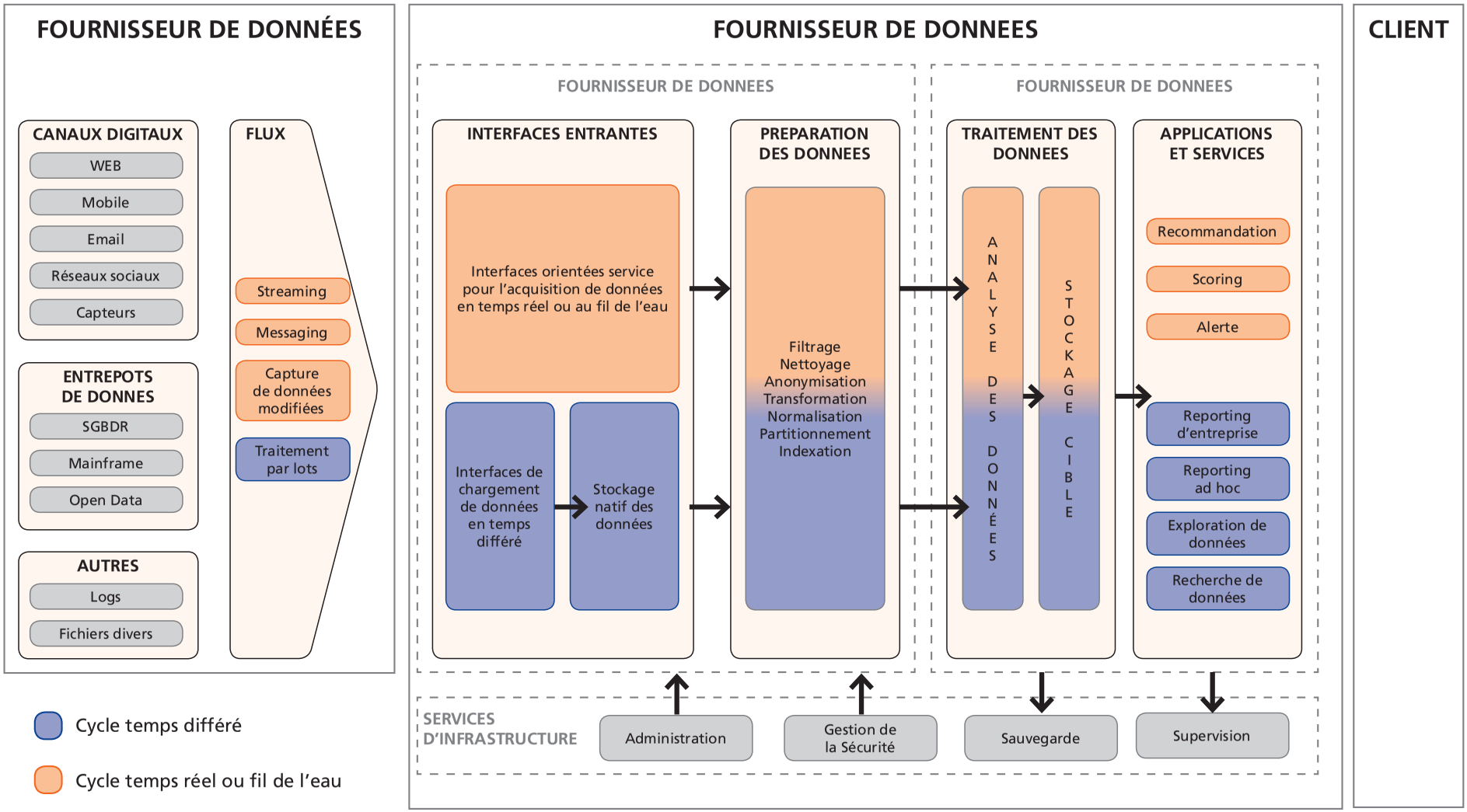
2.1.2.2 Processus de restitution, représentation, visualisation post traitement de l'information

Liste des activités mises en évidence par l'étude (questionnaire en annexe):

- Compte-rendu (reporting),
- Approbation (meilleure compréhension) du résultat de l'analyse du contenu,
- Aide à la décision,
- Représentation,
- Visualisation des graphes (sociaux, sémantiques, autres),
- Représentation géographique (sur une carte), temporelle, chronologique,
- Simulation 3D (représentations multicritères temporalisées),
- Intégration des résultats dans une application (ex : recommandations sur un site web).

La figure ci-après[2] décrit une architecture fonctionnelle typique d'un système « *Big Data* » :

2 -Figure inspirée de la documentation associée à un projet dans le domaine de la santé – source BUSINESS & DECISION



Cette architecture est composée par de modules fonctionnels en lien avec les activités des rôles et sous rôles décrits précédemment. Les modules fonctionnels du fournisseur de services *Big Data* permettent la collecte et l'analyse de données à la fois très variées (structurées ou non structurées) et volumineuses en s'appuyant sur un ensemble élargi de systèmes d'acquisition capables de répondre aux exigences techniques et aux besoins des métiers les plus contraignants.

La gouvernance s'appuie sur une fonction de monitoring ou d'hypervision étendue à l'ensemble de l'architecture et ne figurant pas sur le schéma.

Cette architecture peut être entrevue comme l'aboutissement final de l'assemblage planifié d'un ensemble de modules fonctionnels permettant au fournisseur de services *Big Data* de dégager, graduellement dans le temps, la valeur recherchée dans la transformation de la donnée brute en information exploitable sous forme d'actions et de prises de décisions.

D'autre part, certaines sources de données ont, du fait d'une nature évènementielle de portée éphémère dans le temps et des débits élevés avec lesquels elles sont produites, un intérêt à être collectées en temps réel (ou en mode juste à temps). D'autres sources de données ont, de par leur pertinence dans la durée avec en contrepartie une évolution souvent plus lente dans le temps, un sens à être collectées par lot en mode différé (mode batch).

On distingue alors naturellement deux modes d'acquisition de données complémentaires :

- un mode d'acquisition temps réel (en rouge sur le schéma) dont le principal intérêt est de favoriser la prise de décision immédiate en réaction à une suite d'évènements avec une forte dimension prédictive (ex : détection de fraudes, recommandation webmarketing, rupture produit en linéaires) ;
- un mode d'acquisition temps différé (en bleu sur le schéma) offrant des approches d'exploitation plus classiques, telles que les activités de reporting, la fouille de données non supervisée (exploration de données), la recherche d'information.

Suite à une phase initiale d'identification des sources de données éligibles en première approche et de mise en œuvre du socle technique de collecte des données d'historique à partir des différents canaux, l'exploration de données doit être réalisée pour élaborer les modèles statistiques et analytiques sur lesquels s'appuieront les analyses de données réalisées en temps réel.

Un processus d'amélioration itératif des modèles doit être immédiatement envisagé pour prendre en compte les retours d'exploitation (robustesse des modèles) faits par les utilisateurs des services *Big Data* (par exemple le responsable du pilotage du développement commercial du client des services *Big Data*).

De par sa modularité, l'architecture fonctionnelle *Big Data* offre la capacité d'intégrer, avec une grande agilité, de nouvelles sources de données participant ainsi à l'enrichissement et à l'optimisation des modèles d'analyse reposant sur la découverte de corrélations de données souvent insoupçonnées.

2.2.1. - Lien avec la recherche et developpement (R&D) : les projets des poles de competitivite, la R&D communautaire, les initiatives en code source libre (open source)

De nombreux appels à projets sont identifiables autour du *Big Data* et en particulier une initiative Paris-Saclay Center for Data Science vient d'être lancée (2014) avec pour objectif d'organiser l'exploitation des données scientifiques dans un contexte multidisciplinaire. Il s'agit d'un projet d'infrastructure (cluster) dont l'une des retombées sera de proposer de nouveaux modèles pour disséminer dans un contexte d'ouverture la connaissance sous forme de données et d'outils associés. Un projet en cours financé par l'appel *Big Data* PIA 2012, Teralab, piloté par l'Institut Mines Telecom et le GENES (Groupe des Écoles Nationales d'Économie et Statistique), offre une plateforme technique évolutive (hadoop et serveur in-memory) pour offrir un service dédié à l'innovation en matière de *Big Data*.

Au niveau applicatif, le *Big Data* devient un outil de la Recherche.

Lorsqu'on étudie les appels à projets récents, on constate que certains projets de R&D autour du *Big Data* ont une relation directe avec les standards et les normes. Ainsi, le descriptif de 25 projets récents de R&D portant sur des secteurs très variés citent explicitement le mot norme ou standard. Sur ces 25 projets, 5 expriment l'intention de développer une stratégie normative pour valoriser les résultats de leur recherche et 3 autres envisagent cette possibilité mais sans s'étendre sur la nature du livrable normatif espéré. Les 17 projets restants de cet échantillon se positionnent dans l'exploitation de normes existantes dans leur domaine d'application.

A titre d'exemple, les thèmes de normalisation listés portent sur l'apport des standards du web sémantique pour une application à un domaine particulier (la santé), la mise en évidence de processus permettant l'amélioration de normes existantes (elearning, protocoles de communication multimédia), la définition de formats d'interopérabilité (robotique de service, communication avec des objets internet en milieu avionique).

L'intention des projets cités n'est cependant pas de contribuer au développement de normes (ou standards) spécifiquement dédiées au *Big Data* – à l'exception peut-être du projet traitant de Web sémantique dans le domaine de la santé. Du point de vue de la normalisation, le *Big Data* est donc plutôt compris ici comme un ensemble de moyens permettant de répondre à une problématique de recherche dans un domaine d'application.

2.3 ARCHITECTURE TECHNIQUE ET INTERFACES

2.3.1 - L'ÉTAT DE L'ART

L'état de l'art qui suit s'appuie sur un état de l'art réalisé dans le cadre d'une journée d'étude organisée par le GFII et sur la base de la présentation de Jean Delahousse[3]

Les premiers projets industriels de « *Big Data* » remontent au début de la décennie 2000. Ils sont à l'initiative des acteurs du « Search » sur le web, alors confrontés au problème de «scalabilité » des systèmes, c'est-à-dire de leur capacité à « changer d'échelle » de performance pour accroître ou diminuer leur capacité de calcul afin de s'adapter aux rythmes de la demande et suivre la montée en charge. Cette capacité à ajuster les ressources selon les besoins souvent à performance égale doit par ailleurs être regardée comme un paramètre d'entrée.

Google « BigTable »

En 2004, Google lance à l'interne le projet « BigTable » : une plateforme « haute performance » pour le stockage et le traitement de vastes ensembles de données semi-structurées. L'application, qui repose sur une architecture distribuée (serveurs répartis en « grappes »/« clusters »), est conçue pour pouvoir répondre avec des temps de réponse très courts aux requêtes émanant simultanément de plusieurs milliers d'ordinateurs clients. BigTable est aujourd'hui l'épine dorsale de l'infrastructure Google qui l'utilise pour faire tourner la plupart de ses services en ligne : indexation, crawl, moteur de recherche, GoogleNews, GoogleAlerts, GoogleMaps, Gmail, GoogleBooks (Google a fait son entrée sur le marché de l'informatique décisionnelle alliée aux *Big Data* ces derniers mois. Le service BigQuery, lancé au printemps dernier aux Etats-Unis, propose aux développeurs une plateforme IaaS pour le chargement et le traitement de "données de masse". Si le moteur avait été précurseur dans le

3 -« *Big Data* : exploiter de grands volumes de données : quels enjeux pour les acteurs du marché de l'information et de la connaissance ? » Dossier de synthèse de la journée d'étude du Groupement Français de l'Industrie de l'Information (GFII), Maisons de l'Europe, Paris, Vivien Mann, Juillet 2012

domaine du calcul distribué avec « BigTable », il n'avait pas encore capitalisé sur ces investissements pour se positionner directement sur la BI (Business Intelligence) et peut être considéré comme un nouvel entrant sur ce secteur. On notera que, à l'occasion de son lancement en France et en Europe début octobre 2011, Google a noué un partenariat avec la startup française "We are Cloud", éditrice de la solution BIME, permettant aux entreprises de concevoir leurs propres tableaux de bord, et l'analytique qui les accompagne, à partir de données métiers préalablement chargées dans BigQuery.

MapReduce

« BigTable » repose en partie sur l'utilisation de MapReduce : un formalisme pour le développement de langages de programmation et d'applications optimisées pour le traitement de « données de masse » et leur « mise à l'échelle ». Les librairies MapReduce ont été implémentées dans de très nombreux développements orientés « *Big Data* » par la suite, notamment Apache Hadoop.

Apache Hadoop

Créée en 2004 par Douglass Cutting pour Yahoo, Apache Hadoop est la technologie matricielle de l'écosystème des « *Big Data* ». Il s'agit d'un framework Java en Open Source destiné à faciliter le développement de solutions optimisées pour le traitement de gros volumes de données (les environnements Hadoop permettent d'utiliser, en couche supérieure (top level programming language), des langages de programmation simplifiés par rapport au formalisme Map/Reduce, dont la syntaxe se rapproche de celle des langages de développement connus (Java, SQL, …) : Pig, Hive, Giraph, Sqoop…
(http://developer.yahoo.com/blogs/hadoop/posts/2008/).

Le projet débouche sur le lancement en 2008 de l'application « Yahoo ! Search Webmap 10,000 core Linux Clusters » : à l'époque première et plus importante application opérationnelle de la librairie open source, permettant de faire tourner de plus de 10 000 nœuds (serveurs linux) pour crawler l'ensemble du web.

Modèle No-SQL (Not-Only-SQL)

Au centre des architectures « *Big Data* », il y a la notion de bases de données non-relationnelles, affranchies des contraintes du modèle SQL, notamment les bases de données orientées colonnes, permettant le stockage de très grandes tables (les informations sont stockées par colonnes et non par lignes ; le modèle NoSQL permet de s'affranchir des contraintes de l'architecture relationnelle classique, où les tables et les relations qui les unissent constituent l'unité logique). Parmi les composantes essentielles des environnements Hadoop, on trouve ainsi les applications Hadoop HDFS (Hadoop Distributed Files System) et Hbase qui forment un système de gestion de bases de données orientées colonnes projetées sur des serveurs distribués en clusters.

In-Memory

Les échanges entre niveaux de mémoire occasionnent des pertes de temps soit parce que les mémoires n'ont pas de temps d'accès et de lecture/écriture homogènes, soit parce que les canaux qui les relient ont des débits pénalisants. Plusieurs facteurs d'accélération sont alors possibles par une définition architecturale appropriée. A ce titre, s'il est possible de cantonner le traitement au niveau mémoire le plus rapide, alors le gain de temps est maximal.

Les traitements en mémoire vive (RAM, disques SSD ou mémoire flash) ont des temps d'accès un millier de fois plus rapide que pour les disques durs. La diminution du prix des mémoires vives permet d'en mobiliser une très grande quantité rendant possibles les solutions In-Memory qui peuvent désormais adresser le transactionnel.

Un écosystème complexe, dominé par des « géants »

La librairie Hadoop est le point de départ à la création d'un écosystème « *Big Data* » dans lequel chaque opérateur, à l'image de Yahoo !, utilise la librairie open source pour apporter sa propre valeur ajoutée : IBM, EMC², Hortonworks, Oracle, SAP, Amazon, Microsoft, Cloudera, Datasax, …

Cet écosystème est très créatif. Chaque acteur entretient sa spécificité et se positionne sur la chaîne de la valeur : il y a ceux qui développent et intègrent les bases de données, ceux qui les hébergent et les maintiennent, ceux qui apportent la puissance de calcul et ceux qui les utilisent. Mais l'innovation reste essentiellement californienne, bien que des acteurs européens se détachent, comme SAP en Allemagne (spécialisé dans les technologies de calcul « In-Memory »), ou Quanta en Chine (fournisseur de datacenters), d'où l'enjeu des projets de « cloud souverain » pour développer des majors du secteur en France et en Europe.

Lien vers une cartographie des relations entre acteurs de l'écosystème Hadoop :
http://gigaom2.files.wordpress.com/2012/06/hadoop_ecosystem_d3_photoshop.jpg

Les ruptures technologiques, d'usages et organisationnelles
Temps-réel / non-structuré

Au-delà du buzz word, les « *Big Data* » impliquent plusieurs changements de paradigmes (ruptures technologiques et d'usages) :

1. Un changement quantitatif : l'échelle des volumes à traiter explose, toute la chaîne de création de valeur est bouleversée.

2. Un changement qualitatif : - On ne traite plus des données préalablement échantillonnées et structurées, mais hétérogènes et éparses, structurées et non-structurées (texte, image, multimédia, traces numériques…). - On ne traite plus les données en différé mais en temps réel : on passe d'une logique de silos (batch, tables, ..) à une logique de flux. L'apport des technologies de visualisation est à ce niveau décisif. On peut résumer ces changements par la formule des « 3V » : Volume - Variété – Vélocité. Ceci implique le déploiement d'infrastructures capables de supporter des applications « haute performance ». La comparaison avec la Business Intelligence (BI) traditionnelle permet de saisir les changements organisationnels. Avant, les entreprises développaient à l'interne des entrepôts de données (data warehouse), formaient des statisticiens et des « data analystes » pour lancer des campagnes de fouilles prédictives. Désormais, l'externalisation est nécessaire pour suivre la chaîne des traitements qui se complexifie : il y a les prestataires spécialisés dans l'hébergement de plateformes, ceux spécialisés dans l'intégration et la maintenance des bases de données, ceux qui équipent et apportent la puissance de calcul, les entreprises qui utilisent les applications…

Cloud Computing

Le modèle du Cloud est traditionnellement décrit en plusieurs couches de services sur lesquelles chaque acteur se positionne (IUT-T Y.3500 I ISO/IEC 17799):

- 1. Data Storage As A Service (DaaS) : les données, au centre de l'écosystème, apportées par des producteurs et des fournisseurs de données
- 2. Software as a Service (SaaS) : les logiciels pour traiter les données, fournis par les éditeurs de solutions
- 3. Plateforme as a Service (PaaS) : les plateformes pour héberger et intégrer applications et données
- 4. Infrastructure As A Service (IAAS) : les équipements hardware, fournis par les équipementiers réseaux et fournisseurs de datacenters.

Le développement du « Cloud Computing » est étroitement lié à celui des « *Big Data* ». Des architectures plus agiles et plus puissantes sont requises pour optimiser les ressources et assurer la capacité des infrastructures à tenir la montée en charge sans faire exploser les dépenses d'investissement et de maintenance (scalabilité). Avec le Cloud, les DSI sont en capacité de faire évoluer les infrastructures progressivement sans qu'il y ait besoin d'un « Big Bang ». L'hébergement et les opérations critiques (migration, maintenance) peuvent être externalisés. Les sources d'économie et de ROI sont nombreuses. Avec le « Cloud », tout devient « service ». Pour autant, la notion ne doit pas être réduite à l'externalisation des objets sur des serveurs distants (ex : GoogleDocs). Il n'y pas que les données qui migrent : les applications et les traitements migrent aussi. Le passage à l'échelle et le calcul distribué font exploser la logique d'unité centrale à haute disponibilité, de « supercalculateur » (mainframe).

L'apport des technologies de visualisation

La visualisation temps réel (visual analytics), appuyée par l'analyse sémantique des contenus, apparaît comme une technologie clé du « *Big Data* ». Seule la restitution visuelle permet d'atteindre le niveau d'abstraction nécessaire pour appréhender les « données massives » et leur donner du sens. Des corrélations entre les données permettent d'extraire des connaissances nouvelles qui resteraient tacites et inexploitables sous une autre forme (données tabulaires, linéarité textuelle). L'enjeu est aujourd'hui de développer des technologies permettant de visualiser des flux massifs en temps réel, sans travailler à partir d'échantillons préconstruits, dans une logique de monitoring. Il s'agit aussi de développer l'analytique temps réel autour des tableaux de bords (dashboards), car la restitution visuelle à elle seule ne suffit pas (indicateurs, chiffres clés, …).

Collecte, Stockage et indexation « temps-réel »

L'indexation / catégorisation en temps réel d'information non structurée ou faiblement structurée est un des secteurs les plus porteurs du « *Big Data* ». Ceci dans la mesure où il constitue la brique de base pour la création de services plus élaborés. L'apport des technologies sémantiques est ici déterminant : l'information hétérogène (format, nature) est capturée et structurée à la volée en s'appuyant sur des référentiels métiers et des relations sémantiques (ontologies de domaine).

Le stockage et l'indexation « temps-réel » exigent le développement d'infrastructures réseaux et logicielles à «haute performance » : distribution massive sur des corps de machines distants, utilisation des algorithmes HADOOP / MapReduce pour assurer le passage à l'échelle linéaire, de très puissantes capacités d'analyse et d'agrégation, des SGBD non relationnels mais aux propriétés aussi « ACID » que possible …

A noter : les propriétés ACID désignent les 4 capacités traditionnelles des bases de données relationnelles : Atomicité – Consistance – Isolation – Durabilité. Ces 4 propriétés assurent la stabilité et la cohérence des transactions clients / serveurs dans le modèle relationnel. Mais ces contraintes sont trop fortes pour assurer le passage à l'échelle linéaire des « *Big Data* ». On ne parle plus d'un million d'enregistrements mais de plusieurs milliards. Pour traiter ces « données de masse », il faut lâcher certaines de ces propriétés pour garantir la performance opérationnelle des systèmes. Il faut aussi changer d'unité logique par rapport au modèle SQL (de la table à la colonne), et sortir de stock, pour entrer dans une logique de flux, propre aux SGBD non-relationnels.

Analyse et découverte

Cinq familles de technologies clés ont été identifiées pour ce secteur des « *Big Data* » : text-mining, graph-mining, machine-learning, data-visualisation, représentation de connaissances (ontologies). Les compétences et le niveau de spécialisation propres à ces segments diffèrent considérablement de l'un à l'autre. Mais au-delà du fourmillement des approches, toutes convergent vers un même objectif : simplifier l'analyse de vastes ensembles de données (donner du sens) et permettre la découverte de nouvelles connaissances.

En text-mining, deux approches cohabitent : l'extraction de groupes nominaux (entités nommées), et l'extraction de relations. Dans les deux cas, il s'agit d'extraire de nouvelles connaissances à partir de données faiblement structurées (fichiers logs, conversations sur des réseaux sociaux, forums). On parle de plus en plus de « sens-mining».

Le graph-mining consiste en l'extraction de connaissances nouvelles sous la forme de graphes relationnels. Cette approche repose sur l'isolation de clusters d'information et leur catégorisation, le calcul des « hubs » et « autorités » (les points nodaux), et le calcul de leur positionnement (ranking). Cette méthode est utilisée par les algorithmes de classification des moteurs de recherche sur le web (Page Rank), mais aussi par de nombreuses applications métiers spécialisées. Par exemple, outils d'extraction au sein de bases de données en biologie moléculaire, analyse des réseaux sociaux (détection de communautés, d'influenceurs, …), fouille en environnement brevets (détection de collèges d'innovation).

Le machine-learning est une technologie pivot pour le secteur des « *Big Data* ». Il s'agit de concevoir des systèmes apprenants capables de raisonner rapidement sur des données faiblement structurées. L'apport de l'IA et des approches statistiques sera ici décisif. Les pistes de recherche et développement consistent à développer des algorithmes simulant le fonctionnement du raisonnement humain : inférences bayésiennes, réseaux de neurones, mémorisation, conditional random fields, case-based reasoning… Les formalismes existent mais le passage à l'échelle linéaire reste un défi. Il s'agit de sortir d'une logique d'apprentissage statique pour entrer dans une logique d'apprentissage dynamique. Les ressources (index, arbres, dictionnaires, …) ne préexistent plus à la requête mais sont construites au fil de l'eau, et n'existent que dans le temps de la requête. Dans le domaine de la résolution de problèmes, les gains de performance sont majeurs. Par exemple, la plateforme IBM Watson est aujourd'hui capable de battre des humains au jeu télévisé Jeopardy en analysant en temps réel des énoncés (technologie « speech-to-text »), et en établissant des inférences logiques à partir de données factuelles extraites de DBpedia8.

La « data-visualisation » est un champ désormais bien balisé dont l'apport au secteur des « *Big Data* » est indiscutable. L'enjeu technologique est aujourd'hui de savoir comment faire passer ces technologies à l'échelle de milliards d'unités d'information. Il s'agit aussi de développer une culture de la donnée, de ses logiques de production / exploitation, et de son interprétation visuelle.

La représentation de connaissances (Knowledge Representation) sera aussi un des moteurs du « *Big Data* ». La création de référentiels métiers (ontologies de domaines, de marques, …) doit servir de socle aux développements d'applications métiers à destination de contextes professionnels ciblés. Dans ce domaine, il existe de multiples approches « normatives » ou « émergentes » : taxonomie (RDF), règles logiques (OwL), classification supervisée /non-supervisée, clusterisation, … Toutefois, l'avenir du « *Big Data* » sera dans le croisement des approches et la convergence de ces technologies qui doivent se nourrir l'une l'autre. Sans cela, le « *Big Data* » restera « Small Data » !

2.3.2

2.3.2 - EXIGENCES EN DECOULANT POUR LA NORMALISATION

En matière d'infrastructure, l'absence de normes ne semble pas un obstacle majeur à l'adoption généralisée d'Hadoop. Toutefois, un certain nombre de fournisseurs émergent autour d'Hadoop et offrent leurs versions personnalisées de Hadoop, ce qui conduit à un risque de divergence qui nuit à la portabilité des solutions et à l'interopérabilité au niveau technique.

L'engouement pour Hadoop pourrait cependant évoluer pour aller vers des solutions moins gourmandes en mémoire.

Force est de constater par exemple qu'HDFS n'est pas une interface entièrement portable au sens où elle n'est pas conforme à la norme POSIX, ce qui signifie que les administrateurs système ne peuvent pas interagir avec lui de la même manière qu'ils le feraient avec un système Linux ou Unix.

Dans le même temps, de nombreux projets Hadoop exigent une personnalisation poussée et il manque de guides de pratiques communément admises pour la mise en œuvre de ces environnements. En outre, l'intégration des grappes Hadoop avec les systèmes existants est une tâche complexe et lourde.

Par ailleurs, les systèmes NoSQL (Not Only SQL) souffrent de l'absence d'acteurs du marché en position dominante, mais aussi d'une absence de normes.

Il y a une cause technique à l'absence de normalisation dans ce domaine : un langage de requête destiné à une base de données de graphe n'est pas nécessairement exploitable pour une base de données basée sur une structure de paires de « clé-valeur ». Ceci étant, un point commun est que la plupart de ces bases de données supportent une certaine forme de requête.

Il serait cependant utile de définir un langage de base unifié pour la requête et des acteurs ont essayé de pousser en ce sens, sans résultat probant jusqu'à présent.

Enfin, les entreprises n'ont pas nécessairement besoin d'exploiter une infrastructure aussi complexe qu'Hadoop en permanence, sauf pour des fonctions spécifiques de traitement par lots qui impliquent de très grands ensembles de données. Aussi, le retour sur investissement (ROI) d'un projet de déploiement sur site peut ne pas être très attrayant.

La question de la mutualisation de services *Big Data* via des prestataires de services et de la définition d'interfaces adéquates s'en déduit. Or, le portage de ces environnements vers le Cloud est aussi un enjeu car l'architecture a été pensée initialement pour des grappes de machines ayant des caractéristiques semblables, ce qui est rarement le cas dans un environnement Cloud, même avec l'adoption de machines virtuelles.

En matière d'interfaces, les besoins concernent notamment la représentation de connaissances (Knowledge Representation) sous l'angle de la sémantique comme indiquée ci-dessus, mais également sous l'angle de la capacité de l'utilisateur à manipuler aisément cette représentation.

2.4 LES CAS D'USAGE

Un ensemble de cas d'usage est désormais identifié et référencé par la normalisation. Par exemple, l'organisme US NIST a référencé 51 cas d'usage lors de travaux préliminaires à l'étude de l'ISO/CEI JTC 1 sur le *Big Data* réalisée courant 2014.

Pour ce livre blanc, la méthodologie de recueil de cas d'usage a consisté en une étude qualitative de besoins auprès des acteurs français du *Big Data* via la diffusion d'un questionnaire. Cette étude a été complétée par un travail d'étude de quelques cas d'usage qui a été exploité notamment pour proposer une approche de modélisation conceptuelle.

L'étude de besoins

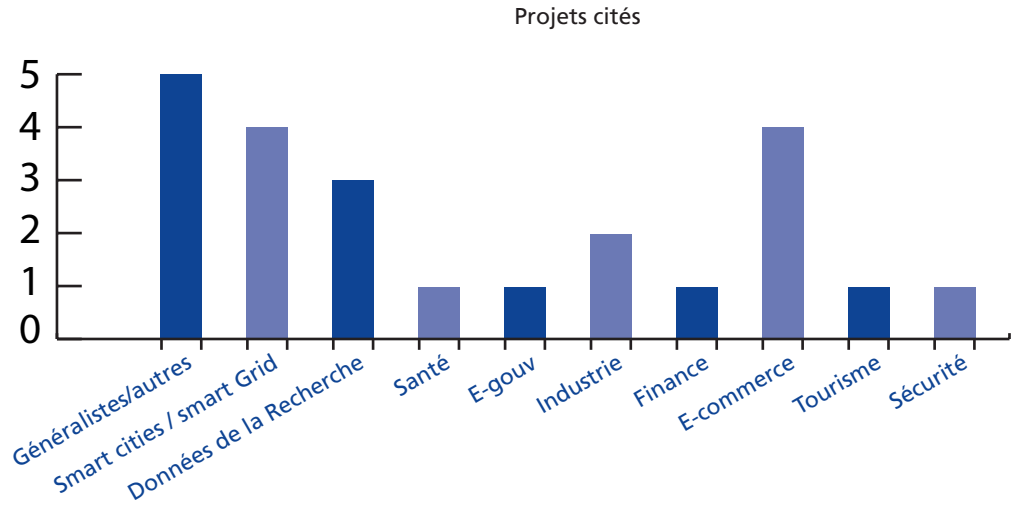
Elle a donc été réalisée en 2014 auprès de la communauté normalisation en France et auprès des communautés autour du *Big Data*, en particulier le réseau Alliance *Big Data*[4]. Elle a été relayée auprès des projets identifiés dans le cadre du plan industriel sur le *Big Data* et auprès des démarches filières industrielles potentiellement concernées.

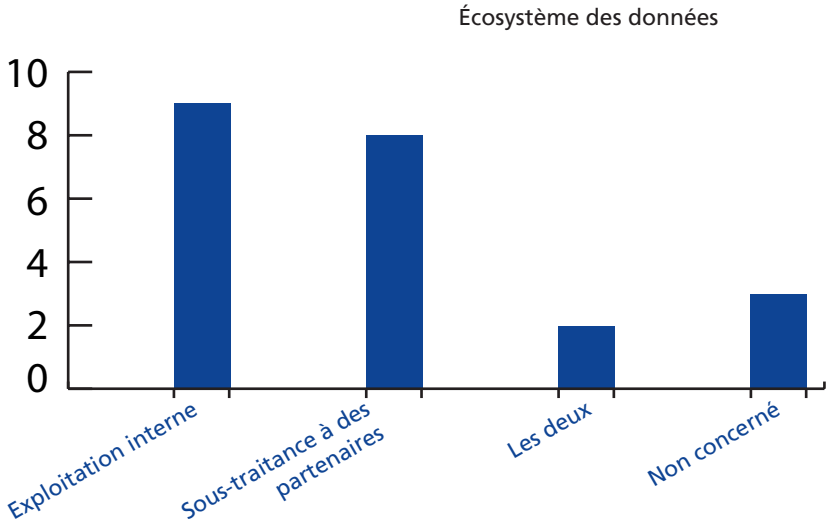
Le questionnaire soumis ne demandait pas une description formelle de cas d'usage en France[5] . Ceci étant les réponses à l'étude de besoins, à travers les 43 questionnaires reçus (209 consultations enregistrées), apportent des indications très intéressantes sur les sphères d'intérêts des communautés et leurs besoins :

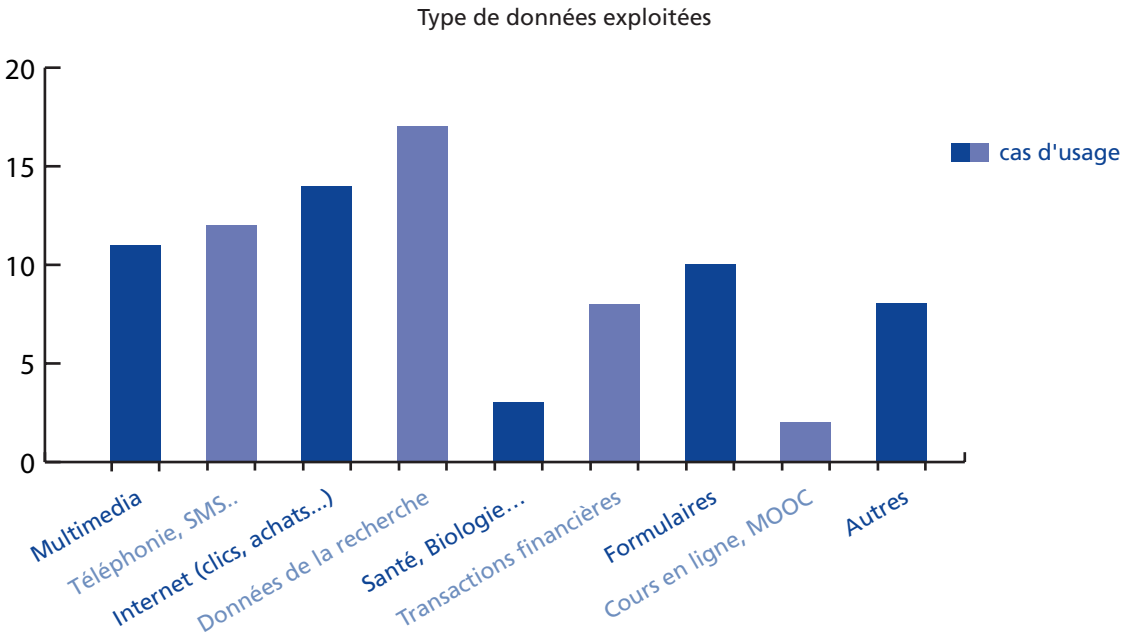
4 - http://alliancebigdata.jamespot.pro/

5 - au sens d'une fiche de description de cas d'usage telle que celle proposée par l'organisme nord-américain NIST

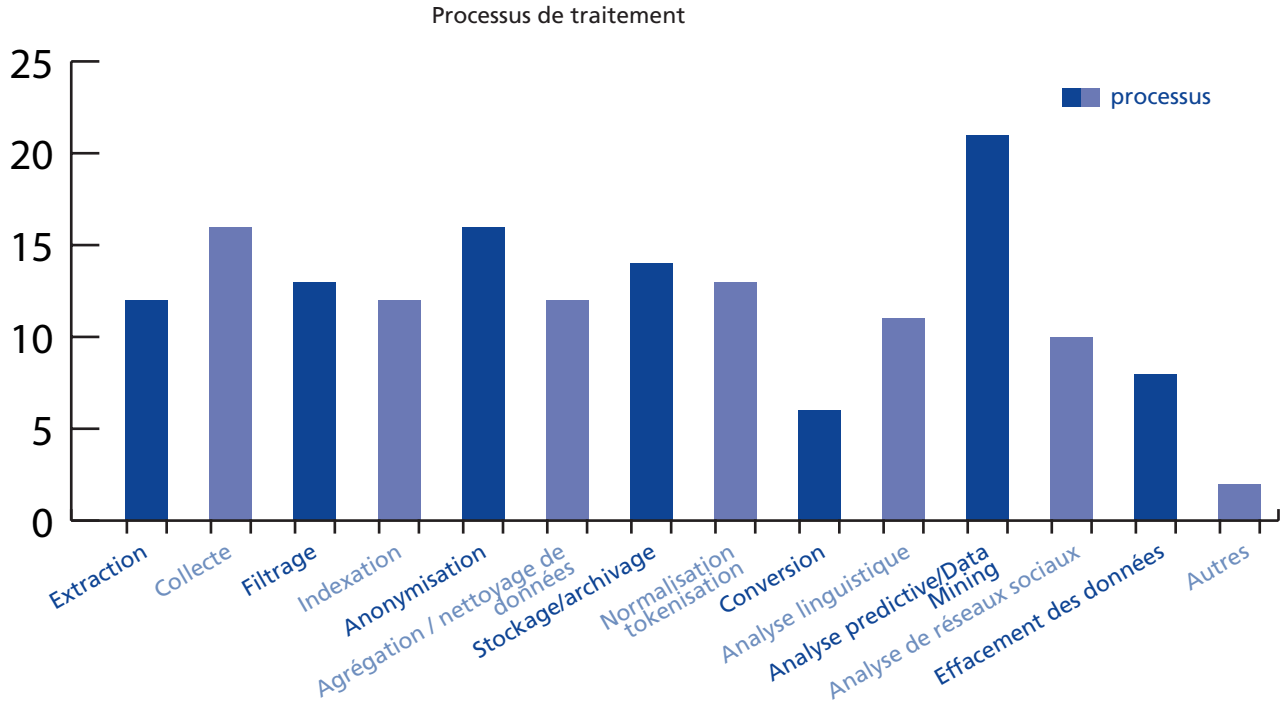
- Domaines des projets – applications envisagées

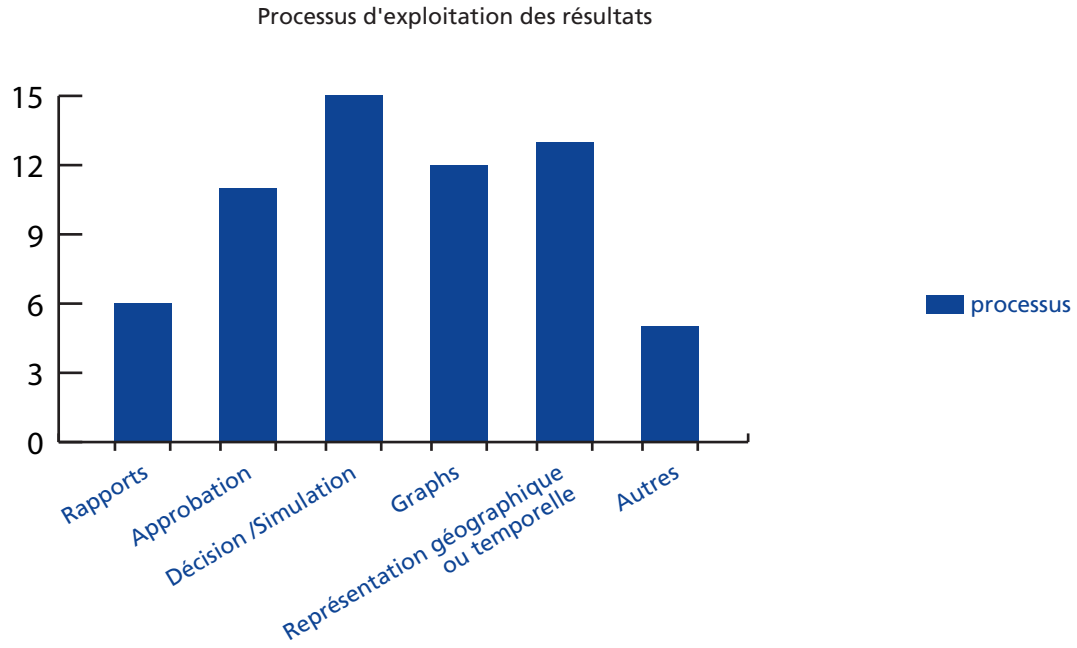






► Processus mis en œuvre



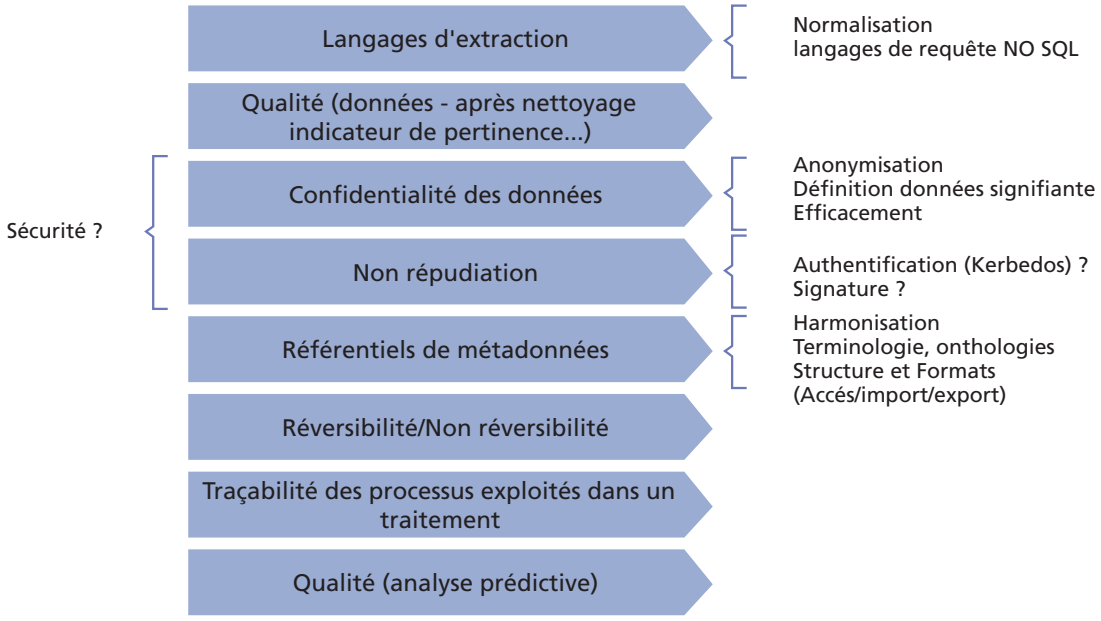


Ces travaux ont également permis de déduire une liste des rôles et des activités qui est décrite dans les chapitres relatifs à l'approche modélisation.

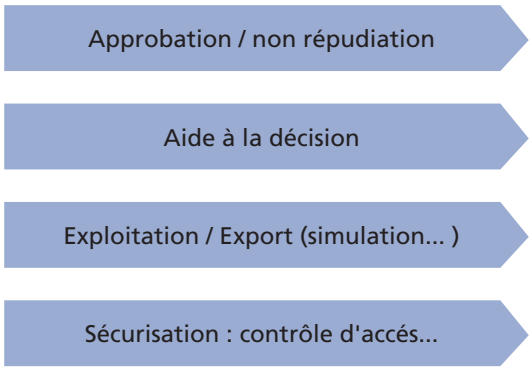
2.4.1 - LES BESOINS FONCTIONNELS

Les besoins fonctionnels qui ressortent de l'enquête qualitative menée par AFNOR ont été regroupés en distinguant ceux associés aux activités liées aux processus de collecte, traitement et analyse de contenus des besoins fonctionnels associés à la restitution, la représentation, et la visualisation post-traitement de l'information ainsi que le présentent les figures ci-après:

Besoins fonctionnels associés aux activités liées aux processus de collecte, traitement et analyse de contenus



Besoins fonctionnels associés aux activités liées aux processus de restitution, représentation, visualisation post traitement de l'information



Le besoin de recourir à des prestataires internes ou externes, la diversité des données traitées ainsi que le grand nombre de processus mis en jeu sont des enjeux largement cités.

Il en est de même de ceux associés à la confidentialité des données, à la protection intellectuelle ainsi qu'aux enjeux juridiques associés au respect de la vie privée.

La pseudonymisation des données personnelles (ou industrielles) est en soit une question technique d'importance : comment s'assurer d'algorithmes non réversibles ? Définir ce qu'est une donnée signifiante qui impacte la "privacy" serait souhaitable, mais également standardiser la manière de présenter et exprimer un opt-in.

En ce qui concerne les formats des données, l'hétérogénéité des formats de données ne constitue pas vraiment une difficulté. Toutefois, des enjeux sont cités en ce qui concerne l'interopérabilité de même que la performance des traitements (plusieurs aspects associés OLAP, versus ROLAP, MOLAP, etc.).

Pour être en capacité à traiter l'ensemble de l'information se posent des enjeux en matière de qualité des données (problèmes récurrents : interprétation des données manquantes, séparateur décimal, données inférieures à la limite de détection...) et de véracité des données. Ceci intervient dès la collecte des données hétérogènes.

En matière de sécurité, le processus devra introduire des concepts de non-répudiation afin de pouvoir fournir à terme des preuves pénales.

En matière de traitement, les enjeux mis en exergue concernent notamment :

- La définition des référentiels (métadonnées) pour faciliter l'exploitation et la catégorisation/filtrage ainsi que le rapprochement des sources, la sédimentation, l'agrégation et le nettoyage des données "data cleaning", les ontologies par métier,
- La traçabilité des opérations effectuées sur les données : on pense souvent à développer toutes les analyses effectuées mais pas à expliquer tout le travail en amont sur les données depuis la réception, le nettoyage de la base, l'interprétation...

La propriété Intellectuelle et l'acceptation des utilisateurs pour délivrer les données dans un entrepôt public sont également des questions liées au volet gouvernance.

Aspects techniques mis en lumière :

L'interopérabilité des systèmes et leur modularité (pour pouvoir facilement et rapidement changer de solution et pour pouvoir mutualiser différentes informations qui sont récupérées par des solutions différentes) est un enjeu cité à plusieurs reprises.

Il en est de même de l'analyse linguistique, traitement automatique des langues, text mining pour ce qui est des conversions (exemples : voix en texte, OCR pour des documents papier) et le sémantique web indispensable avec l'arrivée de l'ère des échanges machine to machine.

La visualisation d'un grand nombre de données qui constitue un enjeu en lui seul !

L'extraction et les langages associés : il faudrait inventer (paradoxalement) le langage SQL du « NoSQL… Collecte » !

Les systèmes de sécurité et d'authentification sont aussi très variés, bien que les mécanismes d'authentification autour de Kerberos semblent devenir un standard de fait parmi les solutions proposées à l'utilisateur/l'intégrateur.

2.4.2 - LIENS AVEC LES SECTEURS

2.4.2

De nombreuses initiatives font référence au « Big Data » dans les secteurs, mais il n'est pas aisé d'identifier ce qui ressort effectivement du « *Big Data* » de ce qui relève d'une problématique de la donnée en général, sans pour autant présenter des spécificités propres au « *Big Data* » (vélocité…).

A titre d'illustration, un focus est donné ci-après sur les besoins du « *Big Data* » dans un contexte de mise en place de « réseaux et territoires intelligents ».

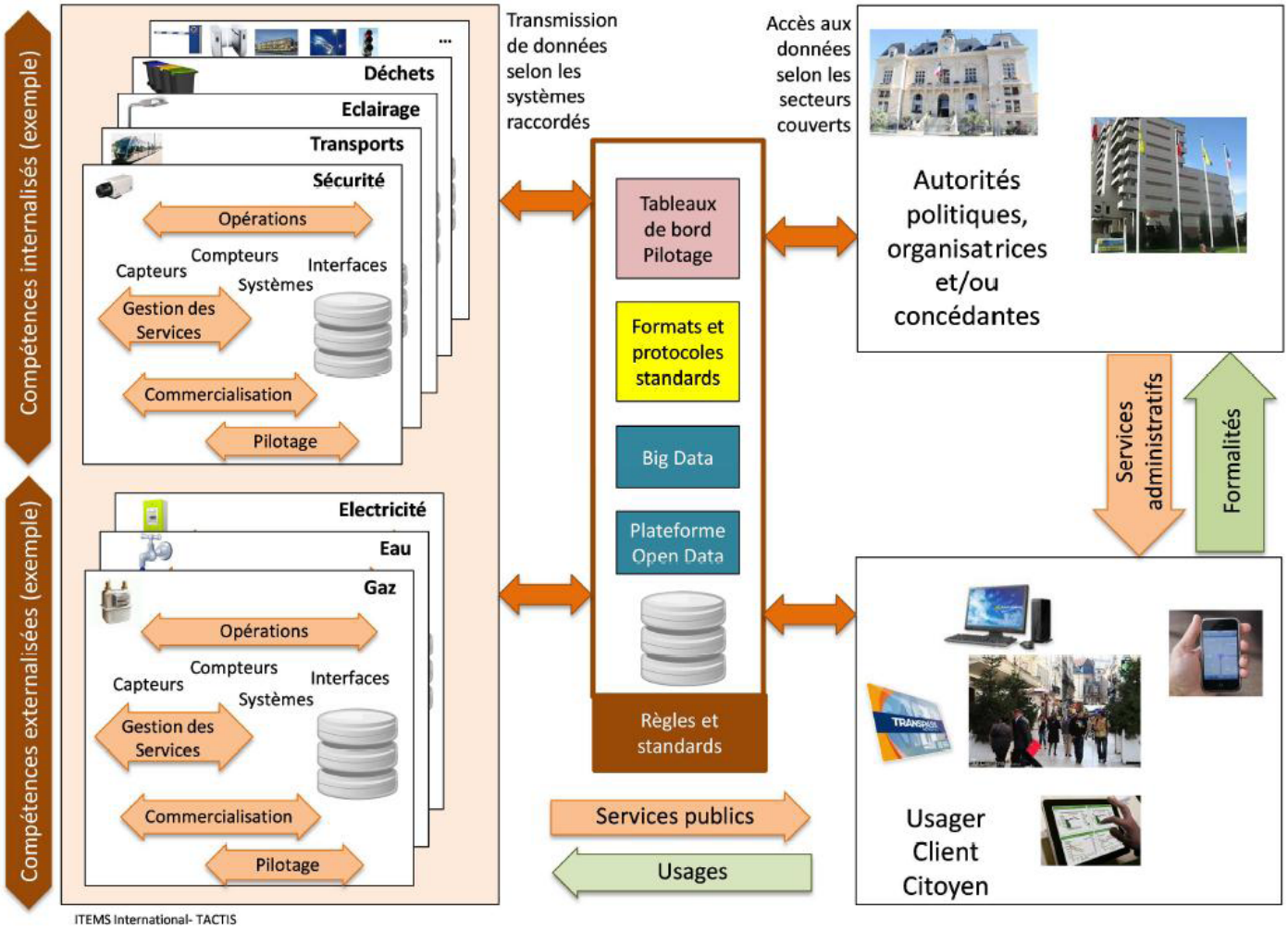
La Fédération Nationale des Concédants de Réseaux (FNCR) a ainsi regardé sous l'optique du « *Big Data* » les besoins exprimés à travers une étude réalisée en 2013 sur les réseaux intelligents (smart grids)

La complémentarité des services publics et du numérique ressort en effet comme un élément déterminant pour la mise en place de démarches de territoires intelligents.

Le *Big Data* présente selon cette étude des opportunités importantes :

- Partage de ressources (capteurs, collecte télécom, stockage, etc.) offrant des perspectives d'optimisation des coûts (CAPEX/OPEX),
- Interopérabilité entre systèmes pour permettre le croisement de données et le développement de services innovants,
- Valorisation renforcée des données.

Ce cas d'usage peut s'exprimer à travers la figure ci-après (source FNCR):



L'étude FNCR fait ressortir 3 scénarios pour la gestion des données des différents réseaux gérés par une autorité publique :

- le premier correspond à la situation actuelle où les données sont gérées en silos sans interopérabilité et sans réelle perspective d'exploitation de type « *Big Data* »,
- le second représente une situation « cible » faisant intervenir une partie « tierce » qui serait une autorité concédant dépositaire des flux de données et qui serait garante de l'interopérabilité,
- le troisième est un scénario intermédiaire qui permettrait de regrouper plusieurs silos de données suivant leur typologie (exemple de rapprochement : les données relatives à l'eau et au gaz).

Les scénarios 2 et 3 mettent clairement en évidence l'importance d'un rôle de gouvernance de la donnée qui a été intégré dans la modélisation présentée dans ce qui précède.

De même, l'approche modélisation proposée présente une généricité permettant la représentation d'une situation avec un acteur tierce partie gestionnaire des données tel qu'évoqué dans l'étude de la FNCR.

En revanche, la question de l'interopérabilité entre des silos de données ne pourra se résoudre dans la seule mise en place d'un projet *Big Data*.

2.5 LES QUESTIONS D'INTÉRÊT POUR LES ACTEURS FRANÇAIS

2.5.1 - LES FORMATS DE DONNEES

Les données se subdivisent en données structurées et en données non structurées. Les premières ont été prises en compte très tôt par l'informatique et concernent à titre d'exemple le transactionnel et les modèles.

Les secondes montent en puissance avec 5 grands types possibles :

- 1.Le texte
- 2.L'image
- 3.L'image animée
- 4.Le son
- 5.Les données de capteurs

La fusion des jeux de données s'avère d'une complexité importante, ce qui est devenu un enjeu pour l'exploitation des données massives.

Ceci a conduit de longue date les spécialistes de la documentation à distinguer deux concepts:

- la donnée : une description élémentaire d'une réalité,
- la métadonnée : une donnée sur / à propos de la donnée, une donnée servant à décrire ou à documenter une autre donnée. On peut à ce sujet distinguer les métadonnées descriptives (du contenu) des métadonnées d'administration et de gestion des données, y compris les aspects juridiques associés.

Il convient donc de distinguer deux niveaux d'interopérabilité sémantique : l'un relatif aux contenus, l'autre aux contenants (schémas XML, etc.).

C'est au niveau des métadonnées que les enjeux se focalisent car identifier des données de toute nature dans différents domaines métier concerne la recherche, mais également les entreprises qui développent des services autour du « *Big Data* ».

Dans le domaine des données de la recherche, l'initiative internationale RDA (Research Data Alliance) à laquelle participent activement les acteurs français autour du pôle de compétitivité Cap Digital, devrait permettre de fixer des objectifs dans le domaine des ontologies et de la sémantique des contenus en lien avec les pratiques du « *Big Data* ».

Ceci ne sera sans doute pas suffisant!

Pour illustrer le besoin d'aller vers/au-delà de la standardisation et de la normalisation, considérons un enjeu spécifique portant sur les adresses universelles URI qui permettent notamment de pointer des ontologies. La question de la qualité et de la maintenance de ces catalogues d'adresses d'ontologies disponibles publiquement sur l'internet se pose si l'on considère que de nombreux projets « *Big Data* » – y compris dans l'industrie - exploitent ces ontologies en pointant ces URI sans trop se poser de questions sur leur origine, qui les maintien et dans quelles conditions !

Actuellement, force est de constater que la communauté privilégie une approche organique et systémique de ce maintien (crowdsourcing, rating, controverses, etc.), mais cela sera-t-il suffisant à l'avenir ?

Si ces approches s'avèrent insuffisantes à l'avenir, la normalisation peut apporter des solutions complémentaires à l'image des normes de systèmes d'enregistrement que développe le comité technique ISO TC 46 sur la documentation (ISAN, ISBN, etc.).

2.5.2

2.5.2 - LA QUALITE DES DONNÉES

La qualité des données est nécessaire pour l'ensemble de la chaîne des traitements *Big Data*. Un enjeu sera de parvenir à qualifier cette qualité pour être en mesure d'avoir confiance dans les traitements, ce qui ne justifie pas nécessairement de viser une perfection.

Pour des analyses *Big Data*, il peut être en effet plus important d'avoir le maximum d'informations (même imparfaites, même partielles), si tant est que le fournisseur qualifie son niveau de confiance / qualité perçue.

Il apparaît donc important de définir des méthodologies[6] pour qualifier la qualité de la source d'une part, et la confiance dans l'information d'autre part.

La mise en place de normes dans ce domaine permettrait de faciliter les relations entre les acteurs de la chaîne et d'avancer vers un ensemble de métriques de référence[7].

2.5.3

2.5.3 - LA PROPRIETE DE DONNÉES

Les questions juridiques et de droits d'usage associés sont également d'importance pour les entreprises qui souhaitent mettre en œuvre des processus de qualité dans une configuration pérenne. On constate sur ce sujet des analogies avec les enjeux associés à l'usage des logiciels libres en entreprise.

Les données, dès lors qu'elles circulent, qu'elles soient copiées, transformées, etc. deviennent des objets d'échange. Si, de surcroît, elles ont ou gagnent de la valeur, alors il devient important d'en connaître le propriétaire. La question peut paraître simple, mais ce n'est qu'un « à priori ».

Qui est le propriétaire ? Celui qui produit les données ? Celui qui possède l'infrastructure qui produit des données ? Celui qui opère l'infrastructure qui produit des données ? Celui qui diffuse des copies ? Celui qui transforme ?

Une fois diffusés, le propriétaire conserve-t-il la propriété ? Une donnée peut-elle avoir plusieurs propriétaires ?

6 - Il serait par exemple possible de s'inspirer des qualificatifs de la Gendarmerie Nationale sur le renseignement.
7 - Il faudra analyser par la suite dans quelle mesure cette action relèverait de comités existants tels que l'ISO/IEC JTC 1 SC 7, SC 40 ou SC 32

Les traitements de données massives confèrent de la valeur aux données, donc aux données utilisables. La notion de propriétaire est-elle dépassée par celle d'utilisateur ? En tout cas, on ne peut plus parler de la seule propriété des données. Dans les réseaux sociaux comme Facebook, ou chez Google, on ne parle plus que d'utilisation, car il semble acquis du point de vue de ces sociétés qu'elles possèdent les données ! La propriété des données est donc un enjeu crucial de notre société. Laissera-t-on les grands groupes mettre la main sur les données ? Il est très probable que nous nous engagions dans une époque de plus forte régulation. Quelle soit nationale, ou transnationale, une autorité doit dire le droit en matière de propriété et cela n'est pas sans conséquence dans l'architecture des systèmes de traitement « *Big Data* ». Il faut aussi définir les lieux où la régulation doit s'exercer.

Si l'on considère les types de contenus traités dans le cadre d'un processus de « *Big Data* », la situation au regard des systèmes d'enregistrement n'est pas la même : si les publications exploitent de longue date des systèmes de registres normalisés (exemple : IBSN), la situation est différente et beaucoup plus incertaine pour la musique et l'audio-visuel (bien qu'il existe un identifiant d'œuvre normalisé ISAN, celui-ci n'est que partiellement mis en œuvre[8]).

Les contenus non structurés de diverse nature : réseaux sociaux… sont quant à eux soumis à des régimes variés à étudier au cas par cas.

Il apparaît donc nécessaire de réfléchir à la stratégie de normalisation la plus efficace pour intégrer les informations de traçabilité sur la propriété et l'origine dans les métadonnées pour les différents types de contenus susceptibles d'être exploités dans un projet *Big Data*.

Ce qui suit n'est indiqué qu'à titre illustratif de la complexité de ces questions et des approches possibles.

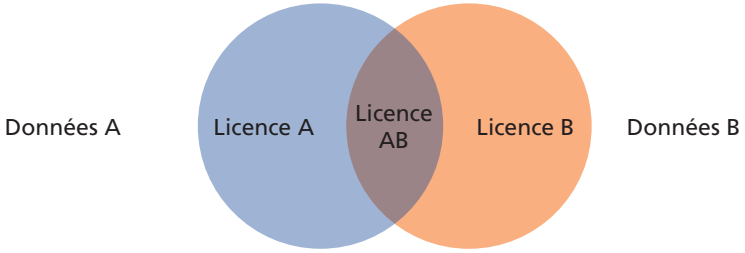
2.5.4 - LES LICENCES ET L'INTEGRATION DE DONNÉES HÉTÉROGENES

2.5.4

Dans le Web des données, qui est celui du « *Big Data* », plus les données manipulées sont volumineuses, plus le coût de leurs traitements augmente et les questions soulevées par les usages de leurs résultats sont importantes. Il n'est plus envisageable de laisser pour compte le problème de la détermination des licences.

Les données, régulièrement constituées en jeux ou flux de données, sont des biens immatériels représentant un capital ou recélant une valeur qui peut être extraite moyennant des traitements. L'ouverture des données a posé de nouveau le problème de leur propriété et celui des conditions de leur copie, diffusion et réutilisation. Des nouvelles licences ont été créées pour répondre à un ensemble croissant de cas.

Mais, il faut désormais ajouter un nouveau volet à cette problématique. En effet, prenons un exemple : si on intègre des données ouvertes, libres d'usage et de réutilisation, avec des données elles aussi ouvertes, mais dont l'utilisation est soumise à une déclaration particulière, que doit-on faire de données obtenues par le croisement de ces deux sources de données ? Il paraît évident qu'il faudra soumettre ce nouvel ensemble de données à la déclaration particulière. Mais, ce n'est pas toujours aussi simple à trancher. La fusion de deux jeux de données peut conduire à trois sous-ensembles soumis à autant de licences différentes, celle de trois jeux à sept sous-ensembles, etc. et la progression suit une loi quadratique.



8 -Étude à paraître réalisée pour le compte du Ministère de la Culture en 2014 par Bearing Point

Il est possible que certaines combinaisons conduisent aussi à des impasses lorsque les contraintes sont contradictoires. Mais on pourra aussi souhaiter trouver la licence qui s'applique à l'ensemble pour ne pas distinguer les parties. Les licences, à l'instar des données qu'elles conditionnent, doivent donc elles aussi être « intégrées » si les données le sont. La question est complexe car, nous l'avons dit, les licences applicables aux données (jeux ou flux) sont nombreuses et la combinatoire qui découle des traitements de croisement possibles est d'un ordre de grandeur supérieur.

Les résultats de recherche effectuée dans le cadre du projet Datalift ont montré qu'une démarche novatrice[9] apportait la solution à la question posée grâce à l'examen de deux caractéristiques pertinentes des données que sont leur provenance et leur licence. La première fournit des instruments pour définir, puis vérifier l'identité du fournisseur des données (l'identification de la provenance est particulièrement pertinente car corrélée à la confiance). La seconde offre des outils pour définir les conditions dans lesquelles les données peuvent être diffusées, et quelles sont les conditions de réutilisation. Le rapport présente plusieurs approches qui ont leurs avantages et leurs inconvénients en fonction des domaines d'application. En décomposant sous forme de prémices logiques les informations de provenance et de licence, on obtient la base d'un jeu de construction qui permet d'assembler, selon des règles propres, de nouvelles licences.

Il est crucial pour le développement du Web de données et celui du Web des objets de fournir des outils permettant aux fournisseurs de données de définir les conditions d'utilisation et de réutilisation de leurs données.

2.5.5

2.5.5 - L'ÉVOLUTION DES ONTOLOGIES

Il existe plusieurs façons d'attribuer du sens aux pures données. L'une d'elle consiste à les faire se référer à une ontologie, c'est-à-dire à un vocabulaire, ou une représentation de connaissances. Celle-ci s'exprime à l'aide d'un langage[10] qui permet de capter ladite connaissance.

Il existe ainsi des ontologies du tourisme, de la météorologie, des organisations, de spécialités médicales, etc. Elles permettent l'élévation de données, c'est-à-dire l'extraction de leur format d'origine pour leur conférer l'interopérabilité. Si des données produites dans des circonstances différentes mais relatives à un même domaine de connaissances sont élevées à l'aide de la même ontologie alors l'hétérogénéité due à leurs origines disparaît, les données résultantes étant exprimées avec le même vocabulaire. L'élévation[11] favorise l'intégration des données.

Même si l'entreprise n'est pas évidente, il se crée de plus en plus d'ontologies et on peut certainement considérer cela comme un bien. Toutefois, ce n'est pas sans inconvénient. En effet, si une ontologie est utilisée pour décorer des données, cela crée une dépendance qu'il faut assumer comme l'illustrent ces quelques questions. Si une ontologie évolue, les données obtenues avec la version précédente, à quelles conditions sont-elles toujours valides ? Si on utilise une ontologie imparfaite, peut-on la modifier pour son propre usage et comment le faire pour conserver le bénéfice de l'interopérabilité ? Que faire si le créateur d'une ontologie ne la maintient plus (on parle ici d'ontologie orpheline) ?

9 - Serena VILLATA, State of art on data provenance and data licencing, 2011, www.datalift.org, https://gforge.inria.fr/docman/view.php/2935/7606/DataLiftD3.5-v1.1-2011-08-08.pdf

10 - Le langage OWL est le plus répandu http://www.w3.org/standards/techs/owl#w3c_all.

11 - Le terme « élévation », ou « élévation de données », est aujourd'hui utilisé dans la communauté du Web des données, il devra certainement être défini dans le glossaire du livre blanc.

Il n'existe pas d'autorité transnationale régulant les ontologies. Et si certaines ontologies peuvent être regardées comme des standards de fait, c'est par pure convention, c'est-à-dire qu'une large communauté les utilise. Malgré leurs défauts, ou leur manque de précision, il est souvent préférable d'utiliser des ontologies bien connues et référencées que d'en développer de nouvelles. Le propre d'une ontologie est de partager des connaissances.

Il faut avant tout se prémunir de l'écueil des données liées orphelines. Tout va bien tant que des données « sémantisées », importantes par leur volume et leur valeur sont liées à leurs ontologies de référence et que celles-ci sont en ligne. C'est la raison pour laquelle le registre « Linked Open Vocabulary (LOV) »[12] a été créé. Il s'agit d'un catalogue d'ontologies. Il est aujourd'hui reconnu au niveau international par les communautés scientifiques et industrielles pour les raisons suivantes : les ontologies qui lui sont soumises pour catalogage sont d'abord évaluées par un comité de curateurs indépendants. Chaque ontologie est décrite, ses différentes versions sont renseignées et un moteur de recherche permet de rechercher et naviguer dans le catalogue. Les données de catalogage sont elles-mêmes sémantisées. Enfin, pour détecter et informer sur les mises à jour, le catalogue est reconstruit tous les jours.

Les données élevées à l'aide d'ontologies cataloguées dans le LOV sont donc garanties interopérables. Le LOV est un dispositif open source et actif depuis 3 ans, il contient actuellement plus de 450 ontologies de référence. Il ne fait en revanche l'objet d'aucune norme et n'est pas reconnu dans le cadre de l'ISO.

12 - Le LOV est un des résultats du projet de recherche Datalift. http://lov.okfn.org/dataset/lov/

L'ENJEU STRATÉGIQUE DE LA GOUVERNANCE DES DONNÉES

Ce qui précède met en évidence l'importance stratégique de la gouvernance des données, bien sûr pour des raisons liées au respect de la réglementation, notamment en matière de protection des données à caractère personnel, mais aussi surtout pour faciliter une meilleure utilisation des gisements de données des entreprises et afin d'optimiser les processus au sein des organisations dans ce but.

L'ADMINISTRATEUR GÉNÉRAL DES DONNÉES

3.1

Il en dérive une fonction nouvelle et transverse aux métiers qui est la fonction d'administrateur général des données, à placer au plus haut niveau, comme peuvent l'être des fonctions stratégiques au sein des organisations.

L'Etat a bien perçu l'aspect stratégique de ce rôle et la nomination récente d'un « chief data officer » de la France[13] confirme l'importance qu'une organisation – l'Etat en l'espèce – doit accorder à la question de la gouvernance des données.

Cet « administrateur général des données » doit avoir des prérogatives particulièrement étendues.

13 - Nomination en 2014 d'Henri Verdier, le directeur d'Etalab, en tant que nouveau « chief data officer » de la France par Marylise Lebranchu, la Ministre de la Décentralisation, de la réforme de l'Etat, et de la fonction publique.

Dans son communiqué, le ministère précise ainsi qu'il sera « autorisé à connaître les données détenues par l'administration de l'État et ses opérateurs [... et] aura pour mission :

- d'organiser une meilleure circulation des données dans l'économie comme au sein de l'administration, dans le respect de la vie privée et des différents secrets légaux ;
- de veiller à la production ou à l'acquisition de données essentielles, de lancer des expérimentations pour éclairer la décision publique ;
- de diffuser des outils, des méthodes et la culture de la donnée au sein des administrations et au service de leurs objectifs respectifs ».

A leur niveau, les entreprises investissant dans les données massives seront également amenées à mettre en place une fonction d'administration présentant des similitudes.

3.2 L'ORGANISATION DES DONNÉES

Les questions qui se poseront à toute organisation concernent en particulier :

- qui est producteur de la donnée ?
- qui en qualifie la qualité et la complétude ?
- qui en garantit la qualité, la pérennité, l'accessibilité ?
- quel circuit de validation mettre en place ?

Cette organisation suppose des prérequis :

- existence d'un identifiant,
- utilisation de celui-ci par les différents producteurs,
- identifiant « activable » et par qui ?

Dans le domaine des données culturelles par exemple, de nombreuses normes existent en matière de métadonnées et d'identifiants. Toutes ne sont pas nécessairement aisément exploitables et la question de passerelles ou de métanormes facilitant l'interopérabilité est une question ouverte[14].

À ces questions se surajoutent des sujets spécifiques au *Big Data* telles que la gouvernance des référentiels de métadonnées.

3.3 LA GOUVERNANCE DES RÉFÉRENTIELS DE MÉTADONNÉES

L'importance des référentiels de métadonnées a déjà été mise en évidence dans ce document car ces référentiels gouvernent dans une large mesure la qualité des processus et des résultats d'une étude *Big Data*.

Il apparaît donc nécessaire d'organiser la production, la diffusion et la maintenance des référentiels de métadonnées. Ceci concerne tant les référentiels internes aux entreprises que ceux générés par un écosystème autour des métadonnées.

Pour les données de la Recherche par exemple, on peut penser que des initiatives récentes telle que RDA, si elles aboutissent, pourront s'avérer particulièrement utiles.

14 - Cf. journée BNF AFNOR du 27 juin – exposé de Katell Briatto.

Pour les entreprises, il faudra sans doute s'appuyer sur les nomenclatures et les initiatives dans différents domaines et ce qui existe dans différents secteurs, sachant qu'il n'existe pas de solutions universelles ni interopérables à ce jour et que les efforts de normalisation restent partiels – exemple : PLIB pour les données industrielles, IFC dans la construction, etc.

Les données linguistiques en sont un aspect particulier, mais qui se révèle essentiel dans le cas de l'analyse de données hétérogènes qui peuvent nécessiter des processus d'extraction sémantique, de conversions de langues, de transcription oral vers écrit, etc.

ASPECTS DE PROTECTION DES DONNÉES À CARACTÈRE PERSONNEL

3.4

Les traitements de données à caractère personnel mis en œuvre dans un contexte de « *Big Data* » sont soumis à la loi 78-17 du 6 janvier 1978 modifiée en 2004.

Les principes de la loi Informatique et libertés s'appliquent à ces traitements de données, même si les données peuvent provenir de sources publiques (réseaux sociaux, etc.). Ces principes incluent par exemple l'obligation de sécurité (article 34 de la loi), ou encore les droits dévolus aux personnes auxquelles les données font référence : recueil du consentement, droit d'information, opposition, etc.

Le cadre réglementaire qui s'applique en matière de localisation de données complique leur exploitation et leur stockage. En effet, la règlementation peut différer d'un pays à l'autre, même en Europe. Il convient de noter toutefois que le projet de règlement en matière de données personnelles vise à uniformiser les règles au sein de l'Union Européenne.

Ces difficultés s'accentuent davantage dans un contexte de *Big Data*, d'où l'intérêt de faire appel à une notion de tiers de confiance qui interviendra pour faciliter la gestion des processus complexes de protection des données privées.

La prise en compte de ces enjeux doit se traduire à différents niveaux :

- technique : mise en œuvre d'une stratégie d'analyse et d'évaluation des risques, d'une logique de conception « privacy by design » prenant en compte à la source la protection des données à caractère personnel et tout au long de la chaîne. En particulier, des processus sont à définir dès le départ comme par exemple des processus de recueil du consentement par ceux, consommateurs ou/et citoyens, qui fournissent leurs données (opt. in versus opt. out),
- organisationnel : mise en œuvre d'une organisation prenant en compte la gestion du risque en matière de données personnelles au niveau des différents intervenants : fournisseurs de données, opérateurs d'infrastructure, opérateurs de services, utilisateurs des résultats.

LA RÉGLEMENTATION

Données publiques

Le terme « donnée publique » n'est pas défini par la loi française. Dans la loi n°78-753 17 juil. 1978 dite CADA (Commission d'accès aux documents administratifs), on utilise le terme « informations publiques ».

Pour les informations publiques, le principe de « disponibilité » est inscrit dans la loi depuis 1978. La Loi CADA susvisée pose le principe de la liberté d'accès aux documents finaux pour le citoyen, et d'une obligation de communication pour les administrations. La mise en place d'une politique de diffusion est laissée à l'appréciation des administrations.

Le principe de « réutilisation » des données publiques a été inscrit beaucoup plus tard, par l'Ordonnance n°2005-650 du 6 juin 2005 et le décret n°2005-1755 du 30 décembre 2005, en transposition de la Directive européenne 2003/98/CE.

Ce nouveau cadre réglementaire consacre les modifications de la loi CADA et apporte de nouvelles obligations pour les administrations :

- Désignation d'un responsable à la réutilisation des données (correspondant PRADA),
- Possibilité de conditionner la réutilisation au versement d'une redevance,
- Condition : non altération, non dénaturation, mention de la source et date de la dernière mise à jour,
- Plusieurs données sont exclues du champ du périmètre : jugements, données des EPIC, SPIC, données culturelles 15 …

Le principe de « gratuité » a été inscrit par le décret n°2011-577 du 26 mai 2011 et la Circulaire du 26 mai 2011.

Ce nouveau cadre réglementaire énonce que :

- La réutilisation des données à d'autres fins que la mission de service public en vue de laquelle les documents ont été élaborés ou sont détenus est permise. La réutilisation à des fins commerciales est donc consacrée ;

- L'autorité compétente décide si la réutilisation doit donner lieu à une redevance ou non, dont le calcul doit être transparent et proportionné aux coûts de collecte et de mise à disposition des données pour les administrations détentrices. Une licence doit être délivrée si l'utilisation donne lieu à une redevance.

Il convient de rappeler que la loi CADA (article 10), exclut la possibilité de réutiliser les données :

- *« a) Dont la communication ne constitue pas un droit en application du chapitre Ier ou d'autres dispositions législatives, sauf si ces informations font l'objet d'une diffusion publique ;*
- *b) Ou produites ou reçues par les administrations mentionnées à l'article 1er dans l'exercice d'une mission de service public à caractère industriel ou commercial ;*
- *c) Ou sur lesquelles des tiers détiennent des droits de propriété intellectuelle. »*

La création de la Mission Etalab en charge de la coordination des politiques d'ouverture des administrations centrales et de l'alimentation du portail data.gouv.fr, ainsi que la mise sur pied de la « Licence ouverte de l'Etat », consécutive à la Circulaire du 26 mai 2012, marquent la reconnaissance institutionnelle du mouvement « Open Data » en France. Ce mouvement milite pour l'ouverture des données publiques la plus large possible (et la « gratuité » de la réutilisation comme principe général), ceci pour encourager la transparence des administrations.

Données privées

Les « données publiques » sont relativement perméables aux principes du « *Big Data* », dont elles sont un des principaux gisements. Pour les « données privées », la situation est plus complexe. Il n'y a pas de régime juridique unifié encadrant la propriété des données. Selon la nature des données et le contexte, le réutilisateur doit aller vérifier le régime applicable.

- **Les données à caractère personnel,** sont régies par la loi 78-17 du 6 janvier 1978 (Informatique et libertés) qui leur est applicable.

Selon cette loi, une donnée à caractère personnel est toute donnée permettant d'identifier directement ou indirectement une personne physique. Il s'agit par exemples des noms et prénoms ou d'un numéro de téléphone. L'adresse IP, donnée importante dans le contexte de *Big Data* et qui identifie - en principe - une machine, est considérée par la Commission Nationale de l'Informatique et des Libertés (CNIL) comme une donnée à caractère personnel et donc soumise aux dispositions de la loi Informatique et Libertés.

Tout traitement de données à caractère personnel suppose le respect de l'ensemble des dispositions de la loi Informatique et Libertés, y compris la réalisation des formalités préalables (déclaration auprès de la CNIL voire une demande d'autorisation pour certains types de traitement).

Il convient de rappeler que la notion de « traitement » de données à caractère personnel au sens de la loi Informatique et libertés est très vaste. En effet, le traitement comprend « toute opération ou tout ensemble d'opérations portant sur de telles données, quel que soit le procédé utilisé, et notamment la collecte, l'enregistrement, l'organisation, la conservation, l'adaptation ou la modification, l'extraction, la consultation, l'utilisation, la communication par transmission, diffusion ou toute autre forme de mise à disposition, le rapprochement ou l'interconnexion, ainsi que le verrouillage, l'effacement ou la destruction ».

Ainsi, la simple consultation, communication ou rapprochement de données à caractère personnel constituent un traitement soumis aux dispositions de la loi Informatique et Libertés. Il convient de rappeler aussi que le non-respect des obligations du responsable de traitement (déclaration/autorisation, sécurité des données, etc.) peut être sanctionné sur le plan pénal.

- **Les données des réseaux sociaux**

Les traces numériques laissées sur les différents médias sociaux par les utilisateurs sont un des gisements des « Big Data » à forte valeur ajoutée, mais là encore, les limitations à la réutilisation sont nombreuses. Le principe général pose qu'un profil d'utilisateur est un espace réservé et donc privé. Ce principe peut être tempéré par un autre, qui pose que la confidentialité des espaces personnels est relative aux paramétrages de l'utilisateur (« paramètres de confidentialité »). Les plateformes sociales ont obligation légale d'informer les utilisateurs de l'état de leur paramétrage et de toute réutilisation de leurs données. Mais la modularité des espaces personnels, comme le degré de complexité des paramétrages et des politiques de confidentialité de certains services complique la donne : qui détient, de l'utilisateur ou du réutilisateur, une donnée « personnelle » publiée sur un espace a priori « fermé » mais dans une configuration « ouverte » ? Ce que j'écris sur un «mur» et qui me concerne peut-il être considéré comme une donnée « publique » ?

En tout état de cause, l'élément contractuel a toute son importance dans le cadre des réseaux sociaux pour déterminer les limites de l'utilisation des données des utilisateurs.

- **Les données non personnelles**

Il s'agit de toute donnée qui ne répond pas à la définition des données à caractère personnel au sens de la loi Informatique et Libertés. Ces données peuvent être protégées par le droit d'auteur si elles correspondent à une œuvre de l'esprit au sens du Code de la propriété intellectuelle (critère jurisprudentiel de l'originalité). A défaut, les données peuvent être défendues sur le fondement du parasitisme, qui est un principe jurisprudentiel qui sanctionne le fait de bénéficier du travail d'un acteur économique sans bourse déliée et sans raison légitime.

- **Le droit des bases de données – le droit « sui generis »**

Dans le contexte de *Big Data*, le moyen juridique qui paraît le plus pertinent pour protéger ses investissements est la loi dite « sui generis ». Cependant, pour pouvoir bénéficier de la protection sui generis, il y a des conditions spécifiques à respecter.

La protection sui generis correspond à la loi n°98-536 du 1er juillet 1998 (en transposition de la Directive du 11 mars 1996). Cette loi protège précisément les investissements du producteur de la base de données.

La base de données est définie comme un «recueil d'œuvres, de données ou d'autres éléments indépendants, disposés de manière systématique ou méthodique, et individuellement accessibles par des moyens électroniques ou par tout autre moyen » (art. L.112-3 Code de la Propriété Intellectuelle).

Le producteur, quant à lui, est défini comme la personne qui prend l'initiative et le risque des investissements correspondants, bénéficie d'une protection du contenu de la base lorsque la constitution, la vérification ou la présentation de celui-ci atteste d'un investissement financier, matériel ou humain substantiel.

Il s'agit donc de l'investisseur.

Le producteur de bases de données a le droit d'interdire :

- *« 1° L'extraction, par transfert permanent ou temporaire de la totalité ou d'une partie qualitativement ou quantitativement substantielle du contenu d'une base de données sur un autre support, par tout moyen et sous toute forme que ce soit ;*
- *2° La réutilisation, par la mise à la disposition du public de la totalité ou d'une partie qualitativement ou quantitativement substantielle du contenu de la base, quelle qu'en soit la forme ».*

Il convient de préciser que le producteur d'une base de données peut ne pas être propriétaire des données de sa base et bénéficier tout de même de la protection accordée par la loi sui generis. Cela pourrait être le cas de Google par exemple ou Twitter…

Outre la protection par la loi sui generis, si la structure de la base de données est considérée comme une œuvre de l'esprit, elle peut bénéficier, en même temps, de la protection par le droit d'auteur.

CARTOGRAPHIE

5.1 DES TRAVAUX ET INITIATIVES EN COURS RELATIVES A LA NORMALISATION DU *BIG DATA*

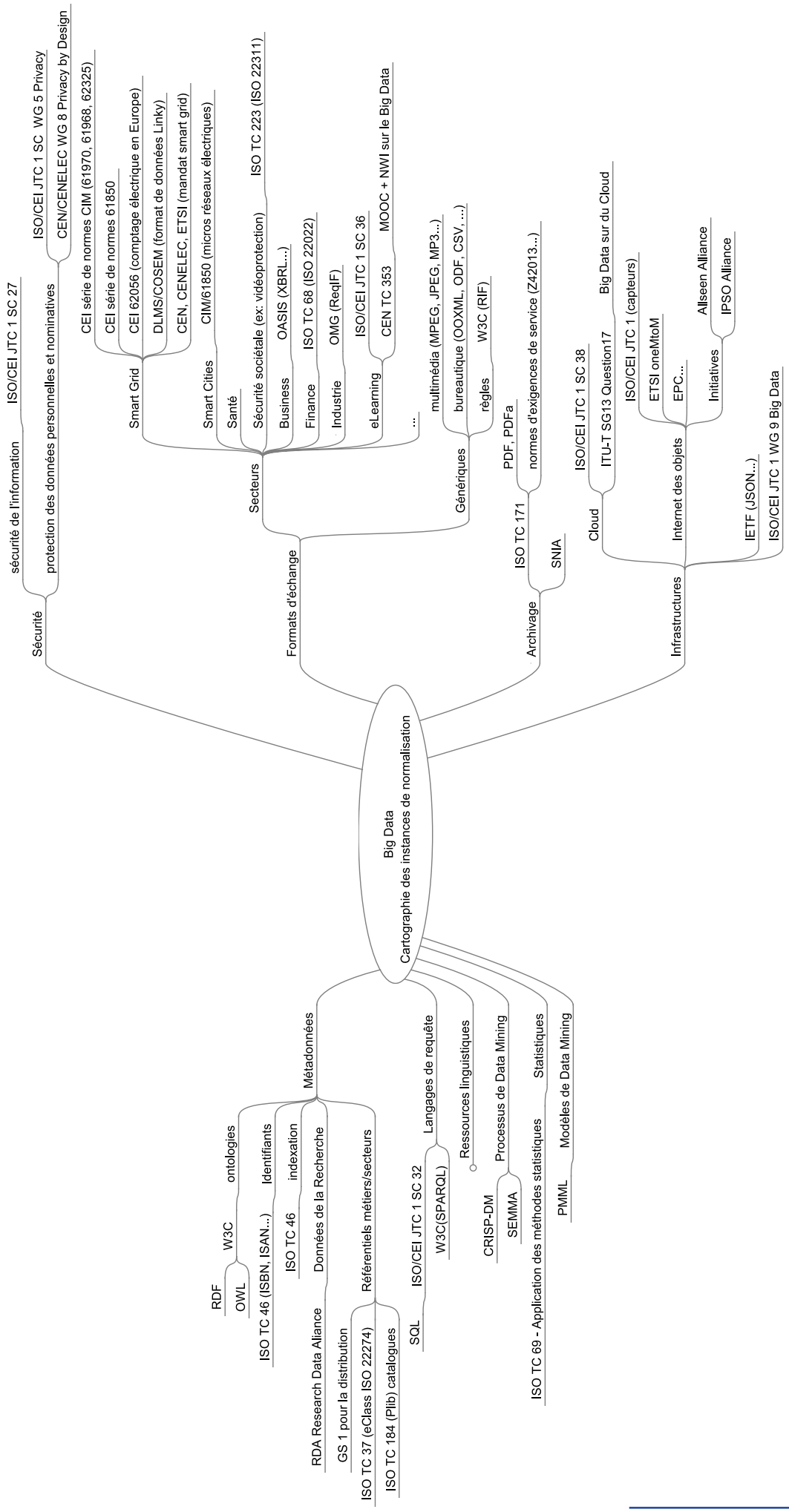
L'engouement pour les données massives ayant pris une importance croissante au niveau mondial, il n'est pas étonnant que la plupart des instances de normalisation s'y intéressent activement. Il convient de souligner que l'organisme US NIST relayé par l'ISO/CEI JTC 1 qui gouverne la normalisation des technologies de l'information, ont engagé un travail stratégique d'état de l'art et de programmation de travaux dans le domaine du « *Big Data* ».

L'organisme international UIT-T de son coté a engagé lui-aussi un travail sur le « *Big Data* », mais orienté dans un contexte de Cloud computing.

En Europe, l'ETSI s'intéresse à ces sujets et en a fait en 2014 le thème d'une journée de travail de son conseil technique (« Board »). Il est cependant peu probable que cet organisme lance un projet normatif dans ce domaine dans l'immédiat.

La Commission Européenne avec le relais de la plateforme Multi-parties prenantes sur la standardisation est désormais sensibilisée et elle pourrait proposer des actions dans le cadre de la révision de son plan roulant sur la normalisation ICT ou dans le cadre de ses actions de soutien à la R&D.

La cartographie suivante des instances potentiellement ou directement concernées par la normalisation (et la standardisation) dans le domaine du « *Big Data* » a été établie :


Big Data
Cartographie des instances de normalisation
Sécurité
sécurité de l'information
ISO/CEI JTC 1 SC 27
protection des données personnelles et nominatives
ISO/CEI JTC 1 SC WG 5 Privacy
CEN/CENELEC WG 8 Privacy by Design
Formats d'échange
Secteurs
CEI série de normes CIM (61970, 61968, 62325)
CEI série de normes 61850
Smart Grid
CEI 62056 (comptage électrique en Europe)
DLMS/COSEM (format de données Linky)
CEN, CENELEC, ETSI (mandat smart grid)
Smart Cities
CIM/61850 (micros réseaux électriques)
Santé
Sécurité sociétale (ex: vidéoprotection)
ISO TC 223 (ISO 22311)
Business
OASIS (XBRL...)
Finance
ISO TC 68 (ISO 22022)
Industrie
OMG (ReqIF)
eLearning
ISO/CEI JTC 1 SC 36
CEN TC 353
MOOC + NWI sur le Big Data
...
Génériques
multimédia (MPEG, JPEG, MP3...)
bureautique (OOXML, ODF, CSV, ...)
règles
W3C (RIF)
Archivage
ISO TC 171
PDF, PDFa
normes d'exigences de service (Z42013...)
SNIA
Infrastructures
Cloud
ISO/CEI JTC 1 SC 38
ITU-T SG13 Question17
Big Data sur du Cloud
Internet des objets
ISO/CEI JTC 1 (capteurs)
ETSI oneMtoM
EPC...
Initiatives
Allseen Alliance
IPSO Alliance
IETF (JSON...)
ISO/CEI JTC 1 WG 9 Big Data
Métadonnées
RDF
W3C
OWL
ontologies
ISO TC 46 (ISBN, ISAN...)
Identifiants
ISO TC 46
indexation
RDA Research Data Aliance
Données de la Recherche
GS 1 pour la distribution
ISO TC 37 (eClass ISO 22274)
Référentiels métiers/secteurs
ISO TC 184 (Plib) catalogues
Langages de requête
ISO/CEI JTC 1 SC 32
SQL
W3C(SPARQL)
Ressources linguistiques
Processus de Data Mining
CRISP-DM
SEMMA
Statistiques
ISO TC 69 - Application des méthodes statistiques
Modèles de Data Mining
PMML

5.1.1 - LES RESPONSABILITES D'ACTEURS FRANÇAIS AU NIVEAU DES INSTANCES DE NORMALISATION

5.1.1

Les comités suivants pilotés par des acteurs français méritent d'être soulignés :

- UIT-T SG13 question 17 ➔ Animateur Jamil Chawki – Orange
- ISO TC 46 « Documentation »➔ Président Gaëlle Béquet – ISSN
- ISO TC 171 « Applications en gestion des documents»➔ Présidence Gérard Cathaly-Prétou (jusqu'à fin 2014)
- ISO TC 184 « Systèmes d'automatisation et intégration »➔ Présidence Alain Digeon – Schneider
- ISO TC 290 « réputation en ligne/E-reputation » ➔ M. Laurent Petit – OXYLANE (groupe DECATLON)
- ISO/CEI JTC 1 SC35 « Interface pour l'utilisateur » ➔ M. Khalid Choukri –ELDA
- CEN/CENELEC JWG 8 « Privacy »➔ Mme. Claire Waast-Richard - EDF

5.1.2 - LES ACTEURS FRANÇAIS PARTICIPANT AUX INSTANCES DE NORMALISATION ET LEUR DEGRE DE PARTICIPATION

5.1.2

Différents niveaux d'intervention sont possibles au niveau des instances internationales :

Animateur de groupe → A

Editeur de norme → E

Expert contributeur → C

Observateur → O

- Sécurité
  - » sécurité de l'information ➔ AEC
    - ISO/CEI JTC 1 SC 27
    - CEN/CENELEC/ETSI CS CG ➔ A
  - » protection des données personnelles et nominatives ➔AEC
    - ISO/CEI JTC 1 SC 27
    - CEN/CLT JWG8 ➔ A
- Métadonnées
  - » ontologies
    - W3C
      - › RDF
      - › OWL
  - » Identifiants →AEC
    - ISO TC 46 (ISBN, ISAN...)
  - » indexation →AEC
    - ISO TC 46
  - » Données de la Recherche
    - RDA Research Data Aliance
  - » Référentiels métiers/secteurs →EC
    - GS 1 pour la distribution
    - ISO TC 37 (eClass ISO 22274)
    - ISO TC 184 (Plib) catalogues

- Formats d'échange
  - Secteurs
    - Smart Grid →AEC
      - CEI série de normes CIM (61970, 61968, 62325)
      - CEI série de normes 61850
      - CEI 62056 (comptage électrique en Europe) →EC
      - DLMS/COSEM (format de données Linky) → AEC
      - CEN, CENELEC, ETSI (mandat smart grid)→ EC
    - Smart Cities → AEC
      - CIM/61850 (micros réseaux électriques)
      - CEN, CENELEC, ETSI (coordination)→A
    - Santé → O (informatique de santé)
    - sécurité sociétale (ex: vidéoprotection) → AEC (EC sur certains sujets)
      - ISO TC 223 (ISO 22311) →A
    - Business →O
      - OASIS (XBRL...)
    - Finance →AOC (France animateur sur le SC 7 services)
      - ISO TC 68 (ISO 22022)
    - industrie
      - OMG (ReqIF) → ? (quelques industriels français moteurs à l'OMG)
    - elearning → EC
      - ISO/CEI JTC 1 SC 36 → EC
      - CEN TC 353 →C
        - MOOC + NWI sur le *Big Data*
    - ...
  - Génériques
    - multimédia (MPEG, JPEG, MP3...) →EC
    - bureautique (OOXML, ODF...) →O
    - règles
      - W3C (RIF)
- Langages de requête →O
  - ISO/CEI JTC 1 SC 32 →O
    - SQL
- Archivage
  - ISO TC 171
    - PDF →C
    - normes d'exigences de service (Z42013...) →E
  - SNIA
- infrastructures
  - Cloud
    - ISO/CEI JTC 1 SC 38 →EC
    - ITU-T SG13 Question17
      - *Big Data* sur du Cloud →AEC
  - Internet des objets
    - ISO/CEI JTC 1 (capteurs) →O
    - ETSI oneMtoM →A
    - EPC... → C
    - Initiatives
      - Allseen Alliance → ?
      - IPSO Alliance → ?
  - IETF (JSON...)

- Ressources linguistiques
  - ISO TC 37 SC4 →O (EC au TC37)
  - ELDA ➔ ?
- Processus de data mining
  - CRISP-DM ➔ ?
  - SEMMA ➔ ?
- Modèles de Data Mining
  - PMML ➔ ?
- Statistiques
  - ISO TC 69 application des méthodes statistiques

5.2 ÉTAT DE L'ART SUR LES NORMES DE SÉCURITÉ INTERNATIONALES APPLICABLES AU TRAITEMENT *BIG DATA*

Au regard de la sécurisation du *Big Data* et en particulier de la question de la protection des données à caractère personnel, il a été jugé important de faire mentionner quelques normes particulièrement importantes

5.2.1 - LES NORMES DE LA SÉRIE ISO 27001

Selon la norme ISO IEC 27000 (Systèmes de management de la sécurité de l'information – vue d'ensemble et vocabulaire), la sécurité de l'information est la préservation de la confidentialité, de l'intégrité, et de la disponibilité de l'information. C'est suivant cette définition que l'on protègera les données d'une manière générale.

D'autre part, les traitements de données susceptibles de générer de la valeur impliqueront dans de très nombreux cas des données personnelles. La loi française [Loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés] en donne la définition suivante : *« Constitue une donnée à caractère personnel toute information relative à une personne physique identifiée ou qui peut être identifiée, directement ou indirectement, par référence à un numéro d'identification ou à un ou plusieurs éléments qui lui sont propres.»*.

Par ailleurs, l'article 34 de la Loi dispose que *« Le responsable du traitement est tenu de prendre toutes précautions utiles, au regard de la nature des données et des risques présentés par le traitement, pour préserver la sécurité des données et, notamment, empêcher qu'elles soient déformées, endommagées, ou que des tiers non autorisés y aient accès ».*

Dans le cas de traitement par un sous-traitant (article 35) :

« Le sous-traitant doit présenter des garanties suffisantes pour assurer la mise en œuvre des mesures de sécurité et de confidentialité mentionnées à l'article 34. Cette exigence ne décharge pas le responsable du traitement de son obligation de veiller au respect de ces mesures »

Ainsi, on voit que la loi impose que des mesures de sécurité soient prises de façon appropriée au regard des risques présentés par le traitement.

L'organisation responsable d'un traitement *Big Data* doit donc décider et mettre en œuvre des mesures de sécurité, en considérant au minimum celles qui sont communément admises dans l'état de l'art et effectivement pratiquées par l'industrie. C'est justement le propos des normes d'établir un référentiel de mesures qui doivent être considérées lors de la mise en œuvre et de l'opération d'un traitement informatique.

La norme ISO IEC 27001 spécifie les exigences pour la mise en place et l'opération d'un ISMS (Information Security Management System) dans le contexte d'activité de l'organisation et des risques encourus. On voit que cette approche répond aux principes de la loi mentionnés plus haut.

27001 constitue une référence en sécurité informatique. Suivant les besoins et les secteurs, d'autres normes de la série 27000 pourront être utilisées dans le cadre d'un ISMS. Par exemple, la norme ISO IEC 27018 s'applique pour la protection des données personnelles dans le cas des services de cloud computing opérant comme sous-traitants d'un responsable de traitement.

5.2.2

5.2.2 - LA NORME ISO 29100

Dans les cas où le traitement implique des données personnelles, l'implémentation de 27001 sur le processus de traitement requiert la mise en œuvre de mesures spécifiques à la protection de ces données. Ces mesures doivent suivre les principes recueillis dans la norme 29100 «Security techniques - Privacy Framework».

Les principes de 29100 sont largement inspirés des lignes directrices régissant la protection de la vie privée et les flux transfrontières de données de caractère personnel adoptés par l'OCDE le 23 septembre 1980.

Il s'agit des principes de la qualité de l'information, de la spécificité des finalités, de la limitation de l'utilisation, des garanties de sécurité, de la transparence, de la participation individuelle, de la simplicité de l'exercice du droit d'accès et du principe de responsabilité.

Ces principes se retrouvent également dans la directive européenne de 1995 et l'article 6 de la loi informatique et libertés et peuvent être résumés ainsi :

- 1° Les données sont collectées et traitées de manière loyale et licite ;
- 2° Elles sont collectées pour des finalités déterminées, explicites et légitimes ;
- 3° Elles sont adéquates, pertinentes et non excessives au regard des finalités pour lesquelles elles sont collectées et de leurs traitements ultérieurs ;
- 4° Elles sont exactes, complètes et, si nécessaire, mises à jour ;
- 5° Elles sont conservées sous une forme permettant l'identification des personnes concernées pendant une durée qui n'excède pas la durée nécessaire aux finalités pour lesquelles elles sont collectées et traitées ;
- 6° Les traitements font l'objet des formalités préalables requises.

RECOMMANDATIONS POUR LA NORMALISATION

Pour que l'écosystème et des modèles économiques adaptés aux données massives se développent, il convient de prendre en compte des problématiques clefs relatives aux données :

- **L'interopérabilité des données et des systèmes d'information** est essentielle pour permettre une utilisation de la masse d'informations disponibles, dans une démarche de pilotage global.
- **La sécurisation des données** doit être prise en compte dans les spécifications techniques des systèmes de collecte, de stockage et de traitement de l'information. Elle concerne aussi le niveau juridique pour ce qui est de la propriété des données et des résultats ainsi que l'exploitation des licences d'exploitation.
- **Le respect des contraintes relatives aux données personnelles,** encadrées par la CNIL, doit être pensé en amont de l'implémentation du système. Une pseudonymisation des données est l'une des techniques clés pour faciliter leur exploitation.
- **Les conditions de gestion de transmission des données**, entre entités privées, publiques ou publiques et privées doivent être précisées par une organisation spécifique.

La normalisation est un outil pouvant apporter des solutions opérationnelles. Il est nécessaire de pousser une norme internationale cadrant l'architecture de référence et le vocabulaire du *Big Data*.

En complément, 6 axes de développement ont été identifiés :

1) La gouvernance de la donnée

Spécifier un rôle de gouvernance de la donnée au sein des entreprises est l'une des conséquences de l'avènement du *Big Data*. Ce rôle de gouvernance est transverse au sein de l'entreprise dans la mesure où les projets autour de la donnée concernent plusieurs métiers. Il s'oriente suivant trois dimensions :

- La nécessité pour l'entreprise d'exploiter au mieux les entrepôts de données qu'elle détient et ceux externes auxquels elle a accès, tout en se préservant des fuites d'information.
- La maîtrise de la qualité des données qui représente un enjeu dont dépend la qualité des résultats des traitements *Big Data*. En ce qui concerne les métadonnées notamment, cet enjeu rejoint un besoin de traçabilité et de persistance associé à l'exploitation des bases d'ontologies existantes.
- La sécurité juridique qui est associée aux politiques relatives à la conformité pour laquelle des référentiels de management seront nécessaires avec en ce qui concerne le *Big Data*, une sensibilité particulière à l'exploitation de données à caractère personnel.

La normalisation doit appuyer une prise en compte des besoins du *Big Data* au travers de normes de management déjà existantes, voire l'élaboration d'une norme de management autour de la donnée.

Il s'agira en complément d'incorporer des exigences du *Big Data* dans les référentiels de compétences en cours de normalisation qui serviront de briques pour préciser des nouveaux métiers tel que l'analyste de données.

Enfin, les processus en lien avec la relation client tels que les méthodes d'effacement de données pourront donner lieu à des développements normatifs dans le cadre du nouveau comité international sur la réputation en ligne et où se situent les projets portés par la France sur l'avis fiable de consommateur.

2) La Qualité et l'identification des données

Qualifier la qualité nécessaire des données est un enjeu important de la confiance dans les traitements *Big Data*. Certains traitements augmentent la confiance, on pense aux techniques de recoupements. D'autres peuvent être sensibles à une erreur qui vient corrompre tout un ensemble. Lorsque le *Big Data* suppose l'exploitation d'une grande quantité d'informations, même imparfaite ou incomplète, la qualité n'est pas intrinsèque, elle dépend des traitements opérés. Il faut donc pousser à la normalisation de critères et de méthodologies pour qualifier les sources et l'information en termes de qualité perçue et de confiance dans un contexte donné et en permettre la communication.

Faciliter l'administration de référentiels de métadonnées par l'exploitation de registres d'identifiants normalisés peut compléter des approches organiques et systémiques déjà mises en œuvre sur les réseaux spécialisés (rating, controverses, etc.).

La capacité à disposer de métadonnées dont la définition est stable devient une question cruciale. Cette question se pose notamment pour les structures de représentation des connaissances (taxonomies, thesaurus, ontologies).

Certains registres normalisés existent déjà dans plusieurs domaines (les données textuelles structurées notamment) mais une réflexion plus globale est à mener, notamment en ce qui concerne l'indexation des données non (ou peu) structurées venant des réseaux sociaux ainsi que les données associées à la mobilité et aux capteurs.

Ceci étant, le repérage de certains types (ou sous-types) de données, par exemple les URIs pointant sur des ressources sémantiques, reste peu satisfaisant. Il faudrait par exemple disposer d'agences chargées d'identifier, répertorier et maintenir ces catalogues d'ontologies, ce qui supposerait le développement (ou la mise à jour) de normes d'identifiants adaptées.

Ces travaux pourraient entrer dans le périmètre des commissions de normalisation sur la documentation sachant que la France dispose dans ce domaine d'un leadership international et qu'une collaboration existe dans ce domaine avec les instances de l'internet que sont l'IETF et le W3C.

D'autres instances pourraient être concernées pour traiter des flux spécifiques à l'exemple des activités sur les normes JPEG et MPEG pour ce qui est des flux multimédia. Ainsi, l'ETSI ou l'UIT-T peuvent sans doute avoir un rôle à jouer dans le domaine des flux de données de téléphonie, etc.

3) Données ouvertes – Open Data

De nombreuses organisations souhaitent mettre à disposition publique des données et organiser l'Open Data est nécessaire. Par exemple, les données culturelles et les données de la Recherche représentent dans le contexte de l'open data un enjeu important à la fois intellectuel et économique.

Il s'agit de sujets complexes et il faudra suivre les travaux issus d'initiatives de R&D tels que le programme IDEX[15] (Cap Digital et Systematic) en tenant compte de l'avancée des consortiums tel que RDA. Une option sera d'étudier s'il faut favoriser à terme un portage des résultats de ces travaux vers la normalisation.

Un outillage normatif est par ailleurs déjà disponible pour certains contenus structurés mais la situation semble peu stabilisée pour l'essentiel des contenus traités par le *Big Data*[16] . En ce qui concerne les contenus non structurés notamment, rien n'existe à ce jour pour construire des métadonnées d'administration.

Il apparaît donc pertinent de revisiter les normes du domaine de la documentation sans l'angle des enjeux des données massives et de l'Open Data.

4) Opérateurs d'infrastructures

Adapter au *Big Data* les infrastructures de services comme le Cloud Computing pour le stockage et les architectures massivement parallèles (clusters, calculateurs haute performance HPC, Plateforme de services - PaaS) pour les traitements est prioritaire pour répondre aux enjeux des entreprises et des collectivités.

Une normalisation est en cours sur les services distribués, notamment à l'UIT-T et à l'ISO. Il est donc essentiel de prendre en compte par la normalisation les nouvelles exigences du *Big Data* pour que le Cloud et les traitements distribués soient en mesure de supporter ces nouveaux services.

Ceci concerne notamment des questions d'interopérabilité, de sécurité des données et des processus (acquiescement, non réputation, identification et authentification des intervenants), de la traçabilité des processus (y compris archivage et effacement des données), mais aussi les sujets en lien avec la réglementation en matière de protection des données à caractère personnel notamment par la prise en compte de modèles de gouvernance de la donnée, etc.

5) Opérateurs de service « *Big Data* »

Permettre qu'un écosystème se développe autour de services de la donnée est un nouvel enjeu de développement économique.

La normalisation visera l'interopérabilité technique et portera sur les schémas de fonctionnement, par exemple les processus d'acquiescement, la mise à disposition des ressources, la traçabilité des traitements, en prenant en considération des questions spécifiques telle que la linguistique dont notamment les processus de conversion, la sécurité avec, entre autres, l'effacement des données, et les questions de conformité règlementaire pour la gestion des données et des droits associés (recueil de consentements…), la qualification des opérateurs, etc.

La normalisation d'API s'avère aussi nécessaire pour pousser le traitement au plus tôt dans la chaîne.

Des travaux pourront ainsi être entrepris au sein de comités spécifiques.

15 -IDEX – voir le lien

16 - Étude 2014 en cours de réalisation pour le compte du service du livre et de la lecture du Ministère de la Culture indiquerait que l'usage d'identifiants normalisés comme l'ISAN pour les contenus multimédia serait peu répandu et entrerait en compétition avec d'autres standards.

La question de la représentation de données complexes suppose que l'utilisateur dispose de moyens de manipulation. La normalisation d'interfaces pourrait permettre une meilleure modularisation des processus de présentation, y compris sous la dimension simulation.

L'intérêt de standardiser les SLAs est par ailleurs une question d'importance qui pourrait être adressée par des comités de l'ISO/CEI JTC 1 sachant que le sujet des SLAs pour le Cloud computing y fait déjà l'objet d'études normatives.

6) Normalisation technique

Des enjeux normatifs importants sont ressortis de cette étude.

Il s'agit en premier lieu des processus et des méthodologies autour de la réversibilité des algorithmes de pseudonymisation. Il faudra examiner de façon plus approfondie si cette question peut être traitée de façon générique, en tant que norme de sécurité de l'information, ou si elle demande à être adressée au cas par cas dans un contexte d'application sectorielle, par exemple pour ce qui est des données de connexion téléphonique.

Autre question d'intérêt : l'évaluation de la performance des systèmes notamment pour des environnements comme HADOOP demanderait des méthodes stabilisées.

Un besoin majeur concerne par ailleurs les langages de requête NoSQL.

Enfin, il faudrait codifier les processus de visualisation et de manipulation des résultats *Big Data*.

Ces travaux devraient être adressés par certains comités techniques au sein de l'ISO/CEI JTC 1 et il convient d'être attentif aux recommandations que va produire fin 2014 son groupe d'étude dédié à la normalisation du *Big Data*.

ANNEXES

ANNEXE 1 : BIBLIOGRAPHIE 58

ANNEXE 2 : LES CONTRIBUTEURS AUX TRAVAUX 59

ANNEXE 3 : LISTE DE LOGICIELS LIBRES AUTOUR DE LA GALAXIE APACHE HADOOP 60

ANNEXE 1 : BIBLIOGRAPHIE

- ISO/IEC/ JTC1-SWG3_N0435_SC_32_Preliminary_report_on_big_data_.pdf
- ISO/IEC/ JTC 1 groupe d'étude Big Data - rapport à paraître
- "Big Data : the next frontier for innovation, competition, and productivity", McKinsey Global Institute, Juin 2011.
- « Big Data: la vision des grandes entreprises », cigref, réseau de grandes entreprises, Opportunités et enjeux, Octobre 2013
- « Analyse des Big Data : Quels usages, quels défis », Commissariat général à la stratégie et à la prospective, La note d'analyse n°08, Marie-Pierre Hamel et David Marguerrit, Novembre 2013
- « Big Data : Comment les 3V bousculent les codes », lettre de veille #15, Cap Digital, 2013
- « Big Data et Réseaux Sociaux: Mythes & Réalités » Livre blanc Jamespot, Alain Garnier, 2013
- « Big Data : Seizing opportunities, preserving Values », The white House, Washington, May 2014
- President's Council of Advisors on Science & Technology, Big Data and Privacy: A Technological Perspective, The White House, May 1, 2014.
- Livre blanc Big Data UIT-T novembre 2013
- Etude FNCCR sur Réseaux et Territoires intelligents – septembre 2013
- 2013-Big-Data-Vision-grandes-entreprises-Opportunites-et-enjeux-CIGREF
- Dossier de synthèse de la journée d'étude du GFII – juillet 2012
- Livre blanc APROGED sur content analytic – juillet 2013
- IEEE conférence October 6-9, 2013, Santa Clara, CA, USA dont je n'ai pas le compte rendu
- Le Telemanagment Forum (TMForum) a publié au premier semestre 2014 un guide de définition et d'implémentation des analyses de données massives "GB979 Solution Suite 2.0" qui comprend :
  - un modèle de référence d'architecture fonctionnelle ;
  - 50 scénarios d'usage ;

ANNEXE 2 : LES CONTRIBUTEURS AUX TRAVAUX

AIT-DAOUD, Sanaa - DIGITAL & ETHICS (France)

BAUDOT, Franck - CNIL

BENHABILES, Nora - CEA SACLAY

BERTHAULT, Denis - LEXISNEXIS SA

BONNET Laurent - BUSINESS & DECISION

BOUJEMAA, Nozha - INRIA

BRUNESSAUX, Stéphan - CASSIDIAN SAS - EADS FRANCE

CAPITAINE, Philippe ALLDC - ASSO LEO LAGRANGE DEF CONSOMMATEURS

CHAWKI, Jamil - ORANGE

CHOUKRI, Khalid - ELDA

COTTE, Dominique - OUROUK

DE SOUSA, Maria - CFONB

DECLAIRIEUX, Valérie DGA - INGENIERIE DE PROJETS

DELEZOIDE, Bertrand - CEA SACLAY

DELOUCHE, Stéphane - CAP DIGITAL

GIBERT, Paul-Olivier - DIGITAL & ETHICS

GIROUX, Patrick - CASSIDIAN SAS - EADS FRANCE (France)

GOUTTAS, Catherine - THALES GLOBAL SERVICES SAS - EPM

GRUSON, Manuel - DASSAULT SYSTEMES

HEBRAIL, Georges - EDF R&D

HOUZE, Paul - MICROSOFT FRANCE

HUOT, Charles - TEMIS SA

KEPEKLIAN, Gabriel - ATOS INTEGRATION

LARHER, Tanguy - DGE / SCIDE / SQUALPI

LEGENDRE, Jean François – AFNOR

MAMOUN, Firas – cabinet ITEANU

MORA, Cédric- DGE

RAVIX, Philippe SOGETI - HIGH TECH

SALLABERRY, Jean-Luc - FNCCR

SCHÜCK, Stéphane - KAPPA SANTE

SOULIE, Françoise - Consultant

STEPHAN, François - INST RECHERCHE TECHNOLOGIQUE SYSTEM X

YVON, François CNRS - LIMSI CNRS

ANNEXE 3 : LISTE DE LOGICIELS LIBRES AUTOUR DE LA GALAXIE APACHE HADOOP

H Base	A key-value pair database management system that runs on HDFS.
Hive	A system of functions that support data summarization and ad hoc query of the Hadoop MapReduce result set used for data warehousing.
Pig	High-level language for managing data flow and application execution in the Hadoop environment.
Mahout	Machine-learning system implemented on Hadoop.
Zookeeper	Centralized service for maintaining configuration information, naming, providing distributed synchronisation and group services.
Sqoop	A tool designed for transferring bulk data between Hadoop and structured data stores such as relational databases.
Mobius	une API Générique en JAVA, à même de traiter des données de haut niveau, et qui se place au-dessus d'un cadre Apache Hadoop. Cette API prend par exemple en charge des fonctions de chaînage et des opérateurs de haut niveau tels que rejoindre (interne ou externe) ou groupement. Elle prend en charge également du filtrage. Elle est utilisée en interne chez eBay pour différentes applications à caractère scientifique.

Contact :

Jean-François LEGENDRE
jeanfrancois.legendre@afnor.org
01 41 62 83 57

S15 02 123 - ADE - Studio - DMK GROUPE AFNOR

AFNOR Normalisation
11, rue Francis de Pressensé,
93571 La Plaine Saint-Denis Cedex

afnor
NORMALISATION