groupement français de l'industrie de l'information

Dossier de synthèse de la journée d'étude du GFII

« Big Data : exploiter de grands volumes de données : quels enjeux pour les acteurs du marché de l'information et de la connaissance ? »

3 juillet 2012 - Maison de l'Europe, Paris.

Sommaire

GFII - 43 rue Beaubourg - 75003 Paris - Tél. : 01 43 72 96 52 - Fax : 01 43 72 56 04
mél : gfii@gfii.asso.fr - site web : http://www.gfii.fr

1. Cadrage général : définitions, les acteurs des « Big Data »

Jean Delahousse est consultant en ingénierie sémantique chez Knowledge Consult, fondateur de la société Mondeca, et animateur du groupe de travail « web sémantique » du GFII. Il est intervenu pour proposer un cadrage général sur la notion de « Big Data », en ouverture de la journée. Présentation disponible à l'adresse suivante :

http://www.slideshare.net/jdelahousse/bigdata-introduction

1.1 Le contexte : l'explosion des données disponibles

Le développement du web des objets, et avant cela du web contributif (2.0), génère une explosion des données disponibles sur les réseaux. La multiplication des capteurs d'« *intelligence ambiante* » (« informatique pervasive ») favorise l'émergence de nouvelles sources d'information : les hommes et les objets, interconnectés et interagissant, deviennent *de facto* des générateurs de traces numériques. La diffusion des technologies sans contact (puce RFID, QR codes, …) offre ainsi de nouveaux gisements d'information sur les produits industriels, les biens de consommations et les individus. La captation et le croisement de ces traces deviennent hautement stratégiques pour révéler des connaissances nouvelles. Les possibilités pour la création de valeur sont énormes tant le phénomène est transversal et impacte tous les domaines : aménagement urbain, e-commerce, grande distribution, marketing, e-réputation, recherche scientifique, …

Pour les acteurs, l'enjeu n'est plus seulement de capter et détenir l'information stratégique – celle-ci peut désormais être interprétée à partir des corrélations établies entre des données « ouvertes » - mais bien d'être en capacité de pouvoir traiter et interpréter ces nouvelles sources disponibles en (sur)abondance.

1.2. « Big data » : historique de la notion et concepts fondamentaux

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.

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 *« grappe »/« 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 …. [1]

[1] **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 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

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

Crée 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[2]. 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[3].

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[4]. 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*.

1.3. 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, EMC[2], 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* »),

nouvel entrant sur ce secteur. On notera que, à l'occasion de son lancement en France et en Europe début octobre, Google a noué un partenariat avec la start-up 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.

[2] Les environnements Hadoop permettent d'utiliser, en couche supérieure (*top level programming langage*), 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, ….*

[3] http://developer.yahoo.com/blogs/hadoop/posts/2008/02/yahoo-worlds-largest-production-hadoop/

[4] 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.

ou Quanta en Chine (fournisseur de *datacenters*),[5] d'où l'enjeu des projets de « cloud souverain » pour développer des majors du secteur en France et en Europe.[6]

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

1.4. 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*** [7]. 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 service sur lesquels chaque acteur se positionne :

1. **Data 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

[5] http://www.usinenouvelle.com/article/google-equipe-ses-datacenters-directement-en-chine.N171995?utm_source=feedburner&utm_medium=feed&utm_campaign=Feed%3A+a-la-une+%28Usine+nouvelle+-+A+la+une%29&utm_content=Google+Reader#xtor=RSS-215

[6] http://www.lesechos.fr/entreprises-secteurs/tech-medias/actu/0202251831665-orange-et-thales-emboitent-le-pas-a-sfr-et-bull-avec-cloudwatt-359651.php (Cloudwatt) http://www.lesechos.fr/entreprises-secteurs/tech-medias/actu/0202246630614-sfr-et-bull-propulsent-numergy-dans-le-cloud-358813.php (Numergy)

[7] François Forgeron ajoutera « Valeur » dans son intervention

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 de masse » 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, …).

1.6. Conclusion : quelle création de valeur ?

A ce stade, on peut distinguer au moins deux axes de création de valeur pour les industries de l'information et de la connaissance :

1. **Marketing** : écouter la voix du client, mieux comprendre les comportements d'achat, les besoins des usagers et affiner sa stratégie produit.
    - Anticiper sur les résiliations d'abonnement
    - Identifier les lacunes d'une collection d'ebooks ou d'un fonds documentaire …
2. **Création de nouveaux services :**
    - Nouveaux services d'informations cartographiques
    - Nouveaux services analytiques temps-réel (ex : analyses financières, presse, réputation,
    - Applications métiers sur mesure (ex : ontologies spécialisées)

2. Les défis technologiques du Big Data

François Bourdoncle est CTO d'Exalead, qu'il a cofondé en 2000 avec Patrice Bertin, après son expérience sur le projet LiveTopics d'ALTAVISTA et chez Digital Equipment Corporation à Palo-Alto. Exalead a été rachetée par Dassault Systèmes en juin 2010. Exalead a rejoint la communauté Dataconnexions, portée par la mission EtaLab, en tant qu'hébergeur « Gold » en février 2012.

2.1. Les « Big Data » : contexte et enjeux

« Les chiffres et les lettres »

D'un point de vue quantitatif, les réseaux ont franchi un gap. L'interconnexion des objets et des terminaux mobiles, le développement des réseaux sociaux et des contenus « user generated », ainsi que l'enregistrement permanent des transactions ont fait exploser le volume de données disponibles. Le web comprend aujourd'hui des dizaines de milliards de pages HTML soit plusieurs pétaoctets (1 pétaoctet = 1000 Téraoctets = 1 million de Gigaoctets).

La problématique fondamentale que cherchent à résoudre les acteurs du « Big data » est alors : comment traiter des données non structurées ou faiblement structurées ? La Business Intelligence (BI) traditionnelle construisait une représentation de la réalité à travers des agrégats de données consolidées et structurées dans des entrepôts (des chiffres). Les outils « Big Data » proposent d'analyser des données qui sont **le reflet du monde réel** (*« des chiffres et des lettres »*).

D'un point de vue qualitatif, les réseaux ont également franchi un cap. Le calcul distribué et les logiques de virtualisation permettent d'atteindre des performances systèmes inégalées. Mais les données crées croissent plus vite que la vitesse des processeurs, malgré la célèbre loi de Moore sur le doublement des capacité de calcul tous les 18 mois. Dans ces conditions, le recours massif à des infrastructures distribuées s'impose pour répartir la charge sur un très grand nombre de serveurs afin de gérer les pics d'intensité et d'exécuter les traitements au plus près des données.

La force du calcul distribué réside dans la **flexibilité des architectures** : en cas de saturation des systèmes, de nouvelles machines (nœuds) peuvent très facilement être ajoutées pour doper les performances de l'infrastructure en suivant une échelle linéaire, épousant en temps réel les besoins en termes de consommation de ressources. Cette flexibilité permet de réaliser d'importantes économies. Les technologies « Big Data » doivent ainsi s'adapter à la croissance exponentielle de l'information numérique sans que les coûts de traitement n'explosent en retour.

A partir de quand peut-on parler de « Big Data » ?

La notion de « Big Data » désigne une nouvelle frontière entre les secteurs traditionnels de l'industrie de l'information et les secteurs émergents. La BI, le *data mining*, ou le *data marketing* traditionnels ne sont pas à proprement parler du « Big Data » (bien qu'ils convergent vers ce dernier). On peut parler de « Big Data » dès lors que :

- Les volumes à traiter atteignent des tailles « plus grandes » que les problèmes courants : *Peta (web), Terra, Exa, Zettaoctets, ….*
- Le problème ne peut pas être traité par les outils existants : SGBD relationnels, moteurs de recherche, …

⇨ Développement de la mouvance NoSQL / Not-Only SQL depuis 2010.

2.2. Secteurs et technologies clés des « Big Data »

D'un point de vue schématique, la chaîne de traitement des « Big Data » est identique à la chaîne des fonctions documentaires classiques (cf. Paul Otlet). En réalité, chaque étape représente un segment d'activité à part entière, avec ses problématiques et ses technologies propres, tant le niveau de complexité des traitements est élevé. C'est en ce sens que l'écosystème des « Big Data » est une promesse de croissance et d'innovation : chaque acteur peut se positionner sur une étape et apporter sa valeur ajoutée.

2.2.1. 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 porteur 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 la logique de stock, pour entrer dans une logique de flux, propre aux SGBD non-relationnels.

2.2.4. Analyse et découverte

On peut identifier 7 familles de technologies clés pour ce secteur des « Big Data » : ***text-mining, graph-mining, machine-learning, data-visualisation*, TALN (*Natural Langage Processing*), traitement automatique de l'image (*Image Processing*), 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 R&D 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 DBpedia[8].

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.

En **TALN**, deux approches coexistent : les technologies *« speech-to-text »* (transcription automatique de discours livrés sous forme orale) et les technologies de *« machine translation »* (traduction automatique de discours écrits). Traditionnellement, les systèmes TALN nécessitent des ressources linguistiques importantes pour tourner. L'avenir du TALN réside dans l'intégration des systèmes auto-apprenants, capables de constituer eux-mêmes leurs dictionnaires au fil de l'eau. Des « pierres de Rosette » numériques doivent permettre aux systèmes de produire de nouvelles règles, d'établir des correspondances, en limitant au maximum l'intervention humaine.

Dans le domaine de l'**Image Processing** (traitement automatique de l'image), deux secteurs émergent : l'indexation automatique de flux d'images et de fichiers vidéo, de la reconnaissance faciale et de la reconnaissance d'objets. Les flux visuels comptent parmi les plus difficiles à traiter du fait de leur absence de structuration et d'éléments descriptifs. Pour rendre le contenu des images et des vidéos « compréhensible » par les systèmes, il s'agit d'indexer ces flux au moyen de métadonnées sémantiques sur lesquels ceux-ci s'appuieront pour extraire des

[8] Dossier du Mondeinformatique.fr sur les technologies Watson appliquées au Big Data : http://www.lemondeinformatique.fr/actualites/lire-les-technologies-de-l-ibm-watson-appliquees-au-big-data-33901-page-1.html

informations. Le domaine de la reconnaissance faciale est quant à lui en plein essor, comme en témoigne le récent rachat de la société Face.com par Facebook[9], ou l'inclusion de cette technologie dans Google+[10]. L'explosion des banques d'images contributives (Flickr, Picasa, Tumblr, …) et des pratiques de publication visuelle sur les réseaux sociaux (Pinterest) offre à ces acteurs de nouveaux gisements de connaissance pour le marketing et le ciblage client. Si les freins techniques au développement de la reconnaissance faciale sont peu à peu levés[11], l'acceptabilité sociale de la reconnaissance faciale reste un enjeu majeur.[12]

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. Quels enjeux pour l'avenir ?

Les technologies du « Big Data » sont aujourd'hui robustes et atteignent des niveaux de performances opérationnelles satisfaisants. Il faut néanmoins travailler sur ces axes d'amélioration :

- ⇨ Réussir l'imbrication des approches « normatives » (où l'auteur définit le sens et les usages des contenus) et les approches « émergentes » (où le sens et l'utilisation sont définis par les usagers) pour insuffler une cohérence globale au développement du web sémantique.
- ⇨ Approfondir l'analyse des graphes : graphes relationnels sur les réseaux sociaux et graphes métiers au sein des entreprises (relations métiers entre entités).
- ⇨ Réussir le passage à l'échelle linéaire : il faut de nouveaux algorithmes plus puissants (algorithmes en 0(n), ou 0(n.log n) au maximum).
- ⇨ Développer le *machine learning*, et les capacités des systèmes à auto-apprendre pour réduire les coûts en réalisant des économies sur la consommation de ressources.
- ⇨ Le calcul « in-memory » : une nouvelle rupture ?

A noter : le « ***Calcul en mémoire »***, permet de stocker les données dans la mémoire vive des cœurs de processeurs et non pas dans la « mémoire externe » des serveurs distribués. Cette approche s'oppose en cela à la parallélisation. Certains estiment que cette technologie sera le prochain « gap technologique » dans le domaine du Cloud[13].

[9] http://www.lemondeinformatique.fr/actualites/lire-facebook-rachete-facecom-49368.html

[10] http://lexpansion.lexpress.fr/high-tech/google-lance-la-reconnaissance-faciale_274545.html?xtor=EPR-237-[XPN_11h]-20111212--508129@186625655-20111212112619

[11] http://www.futura-sciences.com/fr/news/t/technologie-1/d/reconnaissance-faciale-1-seconde-pour-reperer-un-visage-parmi-36-millions_38662/

[12] http://www.lemondeinformatique.fr/actualites/lire-facebook-va-supprimer-toutes-les-donnees-de-reconnaissance-faciale-en-europe-50567.html

[13] Par exemple, la technologie SAP « HANA » peut ainsi parcourir deux millions d'enregistrements en une milliseconde,, tout en produisant 10 millions d'agrégations complexes par seconde et par cœur. Des traitements qui demandaient auparavant 5 jours peuvent aujourd'hui être exécutés en quelques secondes : http://www.developpez.com/actu/24798/SAP-leve-le-voile-sur-sa-technologie-SAP-in-memory-qui-peut-diviser-par-1200-le-temps-de-traitement-de-certains-scenarios/

2.4. Questions / échanges avec la salle

<u>On parle souvent de quantité des données, mais quid de la qualité ?</u>

C'est un problème essentiel. Il faut l'envisager sous l'angle des boucles rétroactives et miser sur la capacité des systèmes à faire remonter des défauts de qualité à travers l'analyse de très grosses quantités de données. Mais ceci nécessite malgré tout de lourds investissements méthodologiques et techniques en amont pour agréger et homogénéiser les données afin de disposer d'une base d'informations fiable pour l'interprétation. Contrairement aux méthodologies de BI traditionnelles, les outils « Big Data » ne prétendent pas fournir de vérité (des prédictions) mais des interprétations pour aider à la prise de décision. La nouveauté des « Big Data » est que cette interprétation ne s'appuie pas sur une construction de la réalité (des échantillons de données structurées en silos dans des entrepôts) mais sur de très vastes ensembles de données qui, par leur diversité et leur hétérogénéité, sont le reflet du monde réel (des chiffres et des lettres). Une fois le problème de la qualité des données résolu, il reste une question fondamentale : comment créer de la valeur à partir des connaissances extraites ?

<u>Quel arbitrage entre les investissements en hardware et en software ?</u>

Le développement des modèles en Cloud (SaaS, PaaS, IaaS, …) permet d'arbitrer beaucoup plus facilement qu'auparavant. Il est désormais possible de payer à la consommation pour des besoins ponctuels sans avoir à investir dans des infrastructures *hardware* extrêmement coûteuses. Un laboratoire en astrophysique dont le *mainframe* ne parvient plus à traiter les flux de données satellitaires qu'il reçoit à l'entrée peut désormais louer une tranche à un *datacenter* pour lancer les computations les plus lourdes sans avoir à s'équiper. Pour les instituts de recherche, le passage d'une **logique d'équipement (acquisition)** à une **logique d'externalisation (location)** est un bouleversement sans précédent dans la chaîne de la production scientifique.

3. Quel cadre légal pour l'exploitation des « Big Data » ?

François Forgeron est Directeur du pôle Informatique & Droit du cabinet Alain Bensoussan, spécialiste du droit informatique depuis 1986, ancien chargé d'enseignement à l'Université d'Assas, et chargé d'enseignement à Paris Dauphine.

3.1. Introduction

Il n'y a pas de cadre légal homogène pour l'exploitation des « Big Data ». Lorsqu'une notion émerge, le travail du juriste consiste à définir les concepts et leur périmètre, identifier les briques de services et les contextes d'usages et à mettre en mouvement la réflexion sur ces questions : *Quelle(s) appropriation(s) ? Quelle(s) commercialisation(s) ?*

Définition juridique des « Big Data » proposée par François Forgeron : *« ensemble des données que nous produisons ou que nos équipements produisent, en temps réel, et qui sont d'origine diverses et souvent non-prédictibles ».*

- ⇨ Pour le juriste, ***« l'origine diverse »*** des données est un point d'intérêt essentiel dans la mesure où l'origine, la provenance, pose la question de la disponibilité des données.

Les « Big Data » obéissent aux principes des 4V : **Volume, Variété, Vélocité** et **« Valeur »**.

- ⇨ Pour le juriste, la notion de ***« valeur »*** est un autre point d'intérêt car elle pose la question de la monétisation des services associés aux données.

Les « Big Data » sont l'enjeu informatique de la prochaine décennie. Cet enjeu est fondamentalement décisionnel : la maîtrise des flux des « Big Data » doit faciliter la prise de décision et limiter l'exposition aux risques, réduire les « zones d'incertitude ». Cet enjeu est qui plus est transversal. Les « Big Data » traversent tous les secteurs : marketing, santé, cybercriminalité, gouvernance publique et privée …

3.2. Typologie des données

Deux oppositions permettent d'appréhender le massif des *datas* :

- Données VS données personnelles
- Données publiques VS données privées

Données à caractère personnel

Définition d'une « donnée » : *« Fait, notion ou instruction représentée sous forme conventionnelle convenant à la communication, à l'interprétation ou au traitement par des moyens humains ou automatiques »* (Afnor - http://www.afnor.org/; http://www.clusir-rha.fr).

Définition d'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 »* (Données à caractère personnel -loi n°78-17 du 6 janvier 1978).

En termes de R&D, les données à caractère personnel peuvent être qualifiées de *« carburant de l'innovation »*[14]. Or, ceci pose problème car ce sont les données qui ont le plus de valeur qui sont le plus protégées et les moins réutilisables. Par ailleurs, des données non personnelles peuvent présenter un caractère personnel une fois recoupées. C'est un des grands enjeux pour le développement des services analytiques : comment encadrer le **croisement de fichiers** ?

Données publiques

Pour les données publiques, le ***principe de « disponibilité »*** est inscrit dans la loi depuis 1978. La Loi CADA (Commission d'accès aux documents administratifs - n°78-753 17 juil. 1978) pose le principe de la liberté d'accès aux documents finaux pour le citoyen, et d'une obligation à 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 dec.2005, en transposition de la Directive européenne 2011/20/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'autre fin 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 détentrice. Une licence doit être délivrée si l'utilisation donne lieu à une redevance.

La création de la **Mission EtaLab** en charge de la coordination des politiques d'ouvertures 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, 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

[14] Référence aux propos de Neelie Kroes : « Data is the new oil », dans une interview accordée au magasine européen LeTaurillon le 4 janvier 2012 : http://www.taurillon.org/Interview-de-la-Commissaire-europeenne-Neelie-Kroes-data-is-the-new,04730 (04/01/12). L'expression a largement été réutilisée dans la presse. Elle est d'ailleurs reprise par Isabelle Falque Pierrottin, présidente de la CNIL, dans la présentation de sa stratégie pour 2012-2013 : http://www.cnil.fr/la-cnil/actualite/article/article/les-perspectives-pour-2012-2013-la-regulation-des-donnees-personnelles-au-service-dune-verita/ (10/07/12)

[15] Un des grands enjeux de la révision de la Directive PSI au niveau européen, attendue pour la fin 2012, est l'inclusion des données des établissements publics à caractère culturel dans le champ de l'obligation.

et la création de services à partir des données. A noter : auparavant placée sous l'autorité du 1er Ministre, EtaLab sera désormais placée sous l'autorité du Secrétariat Général pour la Modernisation de l'Action Publique[16]

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 les contextes, le réutilisateur doit aller vérifier dans des régimes spécifiques s'il existe des clauses concernant la diffusion et la réutilisation des informations. On sait malgré tout que l'appropriation libre par un tiers n'est pas autorisée. Ceci pour poser des garde-fous et protéger les entreprises de la concurrence déloyale, des actions en contrefaçon, de l'espionnage industriel ou de tout autre forme de parasitisme économique. Il s'agit aussi, pour les entreprises, de protéger la vie privée des salariés et d'être en conformité avec le cadre réglementaire en matière de protection des données personnelles (déclaration des traitements à la CNIL). Par ailleurs, pour certains secteurs, il existe encore un devoir légal de secret : secret médical, secret professionnel, secret d'Etat, secret des affaires.

A ce sujet, il faudra suivre les discussions autour de **la proposition de Loi du député Bernard Carayon**, présentée en première lecture à l'Assemblée en janvier 2012[17]. Inspiré par le Cohen Act US, le projet instaure un cadre légal unique pour la protection du secret des affaires. Si il est adopté, il permettra aux entreprises ayant respecté un référentiel de protection de l'information, de poursuivre quiconque aurait été appréhendé en train de chercher à reprendre, piller ou divulguer frauduleusement ses informations sensibles. Le projet crée ainsi un délit spécifique pour sanctionner l'atteinte au secret des affaires des entreprises d'une peine de trois ans d'emprisonnement et de 375 000 euros d'amende[18].

Données des réseaux sociaux

Les traces numériques laissées sur les différents média 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 » ?

Le réutilisateur pourra s'appuyer sur deux précédents fournis par la Jurisprudence :

1. **Arrêt de la Cour d'Appel de Besançon,** 15 novembre 2011 (RG n° 10/02642) : les propos sont publics et peuvent justifier un licenciement s'ils sont tenus sur le « mur » de

[16] http://www.gouvernement.fr/presse/opendata-le-gouvernement-confirme-la-poursuite-des-demarches-d-ouverture-des-donnees-publique
[17] http://www.fondation-prometheus.org/Ressources/20110112_ppl_protection_info_eco.pdf
[18] http://www.usinenouvelle.com/article/la-loi-sur-le-secret-des-affaires-est-une-occasion-manquee.N167282

l'un des membres du réseau auquel tout un chacun peut accéder si son titulaire n'a pas apporté de restrictions.

2. **Tribunal Correctionnel de Brest**, 1er octobre 2010 : condamnation de la personne qui a publié sur son profil Facebook des propos reconnus comme outrageants envers une personne dépositaire de l'autorité publique

Le droit des bases de données

La Loi n°98-536 du 1er juillet 1998 (en transposition de la Directive du 11 mars 1996) définit une base de données 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).*

Ce référentiel légal protège les détenteurs de base de toute atteinte à la sécurité des infrastructures et de toute **extraction et réutilisation illicite** des données contenues dans les bases. Le contenu des bases relève du droit *sui generis (*protection contre une extraction ou réutilisation d'une partie substantielle). La structure de la base est considérée comme une création de l'esprit et relève du droit d'auteur.

3.3. Typologie des traitements

Qu'est-ce qu'un traitement de données ?

Au regard de la loi « Informatique et Libertés »

La loi « Informatique & Libertés », art. 2, fournit une définition d'un **« traitement de données à caractère personnel »** : *« Constitue un traitement de données à caractère personnel 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 (…) ».*

Il existe un grand nombre de référentiels légaux « Informatique & Liberté » sur lesquels le réutilisateur peut s'appuyer pour arguer de la légalité d'un traitement :

- Traité de Lisbonne sur le fonctionnement de l'UE (art.16) du 1er nov. 2009
- Convention n°108 du Conseil de l'Europe
- Directive 95/46/CE du 24 oct.1995
- Charte des droits fondamentaux de l'UE (article 8) 18 dec.2000
- Loi n°78-17 « Informatique et libertés » du 6 janvier 1978

Ces référentiels imposent tantôt des « conditions de fond », tantôt des « conditions de forme ».

Conditions de fond :

- La collecte doit être loyale, licite, pour une finalité déterminée, explicite et légitime.
- Les données collectées doivent être adéquates, pertinentes, exactes, complètes et non excessives au regard de la finalité du traitement
- Le réutilisateur a pour obligation d'informer la personne dont les données sont collectées et de fournir des garanties sur la sécurité et la confidentialité des informations.

Conditions de forme :

- Le traitement doit être déclaré à la CNIL (régime de déclaration normal ou simplifié)
- Un régime d'autorisation et nécessaire pour certaines données.

Dans un contexte marqué par l'**interconnexion généralisée** entre les données et l'**extraterritorialité** des services et des dispositifs, la mise en œuvre de ces conditions peut s'avérer d'une extrême complexité.

A noter : Aux Etats-Unis, un contentieux a opposé la Société IMS Health à l'Etat du Vermont en 2011 sur la question de la légalité des pratiques d'agrégation et de fouille de données personnelles[19]. IMS Health est leader mondial de l'information de santé, et produit des bases de données marketing et des services de profilage sur les patients et les médecins à destination des laboratoires pharmaceutiques. L'affaire avait été portée devant la Cour Suprême, l'Etat du Vermont ayant voté une législation *« d'opt-in »* obligeant les sociétés comme IMH Health à recueillir le consentement formel et explicite des patients et des médecins avant de procéder aux traitements. Les Etats du Maine et du New Hampshire avaient auparavant adopté des législations similaires pour les mêmes motifs. Ceci revenait dans les faits à interdire les éditeurs comme IMS Health de produire des services de ce type dans la mesure où ces bases de données ne peuvent être alimentées qu'au moyen de procédures automatiques, le recueil du consentement sur d'aussi vastes volumes étant dans la pratique quasi impossible. Or, la Cour Suprême a déclaré cette législation non-constitutionnelle, en vertu du 1er amendement de la Constitution consacrant la liberté d'expression. Cette affaire constitue une jurisprudence importante aux Etats-Unis pour les acteurs du *data-mining* et des « Big Data »[20].

Qu'est-ce qu'un traitement de données ?

(Au regard des réglementations sectorielles).

Différentes réglementations sectorielles contraignent la disponibilité et la réutilisation des données. On retrouve ici les différentes catégories de secret. Ces réglementations très spécifiques restreignent considérablement les possibilités de traitement et d'ouverture des données :

- **Données bancaires** : principe du secret bancaire sauf exception (réquisition judiciaire) fixé par L.511-33, Code monétaire et financier. La sanction de la violation de cette obligation est prévue au Code Pénal (L.226-13).
- **Données de santé** : le Code la Santé Publique réserve le droit d'accès aux informations relatives à la santé du patient aux professionnels et au patient lui-même (Art.L.1111-7) Les numéros de sécurité sociale des patients peuvent être utilisés par les professionnels de santé dans les échanges avec les différents organismes d'assurance maladie (Art.R.115-1 à R115-3).
- **Données relatives à la sécurité nationale** : Art. 413-9 Code pénal, Circulaire de la DACG n° CRIM 08-01/G1 du 3 janvier 2008.
- **Informations commerciales sensibles (ICS)** : protégées par les différentes briques du droit du secret des affaires.

[19] Voir la Dépêche du GFII *« La Cour suprême américaine tranchera sur la légalité du data-mining de données nominatives »*, publiée par Michel Vajou sur AMICO, le réseau social du GFII le 03/05/2011 (réservé aux membres)

[20] http://www.policymed.com/2011/06/us-supreme-court-sorrell-vs-ims-inc-court-upholds-data-mining.html

3.4. Les garanties

Le droit à l'oubli ?

Aujourd'hui, il n'y a pas encore de consécration par les textes du *« droit à l'oubli »* numérique. Il existe un projet de règlement européen pour harmoniser et unifier le régime de protection des données personnelles au sein des Etats-Membres. Ce projet pourrait inscrire le *« droit à l'oubli »* dans le cadre réglementaire. Aujourd'hui, des dispositions se rapprochant du principe du *« droit à l'oubli »* peuvent être trouvées au sein de plusieurs directives. Mais contrairement à un projet de Directive, un règlement européen s'applique à tous les Etats de manière identique. Il n'y a pas de place pour l'interprétation. Si un tel règlement finit par voir le jour, Google et Facebook seront tenus de détruire les données personnelles des usagers sur simple demande.

A noter : une charte du droit à l'oubli numérique intitulée « « *Droit à l'oubli numérique dans les sites collaboratifs et les moteurs de recherche* » a été signée par les représentants du secteur et des acteurs de la société civile sous l'égide du Secrétariat d'Etat à la prospective et au développement de l'économie numérique le 13 octobre 2010. Ni Google ni Facebook n'ont signé cette charte[21].

Quels engagements de la part des prestataires de service ?

Le secteur des « Big Data » fournit aujourd'hui des outils (Bases NoSQL, environnements HADOOP, algorithmes Map/Reduces) offrant toutes les fonctionnalités nécessaires en termes de traitement, ainsi que des infrastructures et des modèles pour les distribuer (SaaS, PaaS, IaaS, …). Dès lors, quels engagements peut-on attendre des prestataires de service et des fournisseurs de solutions ?

L'expérience du **« Yield Management »**, déjà ancienne, suggère quelques pistes de modèles dont il est possible de s'inspirer. Pour rappel, le domaine du « Yield Management » (ou *revenu management*) désigne les méthodes et systèmes permettant une gestion fine en temps réel d'infrastructures complexes visant l'optimisation du chiffre d'affaires (ex : gestion des réservations de vol dans les compagnies aériennes).

Les notions de **« service »** et de **revenus partagés** sont au centre de ce modèle : prestataires et fournisseurs reçoivent une partie des bénéfices générés côté utilisateur par l'utilisation de leurs solutions. On verra peut-être se développer des opérateurs qui vendront des services « Big Data » comme le font aujourd'hui ceux du « Yield Management », avec des modèles de services fondés sur la participation aux résultats (ROI) et les « revenus complémentaires ».

3.5. Questions / Echanges avec la salle

Quelles avancées sur les données personnelles au-delà de l'UE?

L'UE est considérée comme une zone à haute protection des « données personnelles ». Mais les Etats-Unis commencent à rattraper leur retard sur la question : initiatives « Safe Harbor » et « decrees » signées avec Google, Facebook, Microsoft. La FTC (Federal Trade Commission) est désormais très vigilante par rapport aux grands acteurs de l'internet. De même, l'Asie a construit à marche forcée un environnement juridique favorable aux transferts internationaux de données personnelles au sein de l'APEC (marché commun asiatique). Pour l'UE, la **révision de la directive 1995 sur la protection des données personnelles** est en cours. Il s'agit de mieux

[21] http://www.alain-bensoussan.com/avocats/charte-droit-a-loubli-numerique-dans-les-reseaux-sociaux/2010/10/30

adapter le cadre légal à la complexité des environnements interconnectés et de régler les problèmes liés à l'extraterritorialité des services et dispositifs. La Commission devra aussi se pencher sur le problème du croisement de fichiers (données anonymes qui présentent de fait un caractère personnel après recoupement). Un juste milieu devra être trouvé entre la nécessité d'ouvrir les données pour stimuler l'innovation et la compétitivité, et la préservation de l'avantage concurrentiel acquis par l'UE en matière de protection de la vie privée[22]. Il y aura aussi un gros effort à fournir sur la création d'un **environnement de confiance** entre les détenteurs et les réutilisateurs de données personnelles.

Peut-il y avoir à la fois des données personnelles et publiques ?

Lorsque le recoupement de données « publiques » permet l'identification indirecte d'une personne, les données tombent sous le régime de protection des données personnelles.

Quid de l'extraterritorialité et du Cloud Computing ?

Il est aujourd'hui possible d'identifier une personne vivant en France à partir de données stockées sur des serveurs Google implantés en Irlande ou aux Pays-Bas. Pour cette raison, le Cloud a longtemps suscité la répulsion des juristes. Cependant, plusieurs signaux montrent que les lignes sont en train de bouger :

- ⇨ Les **recommandations de la CNIL** sur le Cloud Computing sont beaucoup moins « protectionnistes » qu'on aurait pu le penser : des paramétrages sont possibles pour développer la filière sans attenter à la vie privée des citoyens et consommateurs[23].
- ⇨ Le développement d'une filière française portée par des projets de « **Cloud souverain »** peine à convaincre. La France devra s'adapter à la concurrence internationale sur le marché du Cloud. Or, le Cloud est d'abord un marché de gros acteurs, qui bénéficient de ressources suffisantes pour développer de « très gros équipements ».
- ⇨ Les Chinois sont moteurs sur **la question de la normalisation** des technologies et process du Cloud, car ils souhaitent être leaders internationaux sur l'équipement réseau et les *datacenters*. La normalisation du secteur devrait permettre d'harmoniser ces questions en termes de sécurité de l'information, de confidentialité des données et d'intégrité des systèmes. La France et l'Europe devront renforcer leur présence dans les agences de normalisation internationale, notamment l'ISO[24].

[22] http://www.gfii.fr/uploads/docs/MidisduGFII_CNIL_VF.pdf
[23] http://www.cnil.fr/dossiers/technologies/cloud-computing/
[24] http://www.les-infostrateges.com/actu/12031392/la-premiere-norme-iso-sur-l-informatique-dans-les-nuages-le-cloud-computing-en-preparation http://www.afnor.org/groupe/espace-presse/les-communiques-de-presse/2012/avril-2012/cloud-computing-faire-face-aux-enjeux-geostrategiques-grace-a-une-norme-internationale

4. Y a-t-il un ou des modèles économiques ? Comment créer de la valeur à partir des « Big Data » ?

Gabriel Képéklian est responsable R&D chez ATOS ORIGIN, spécialiste des technologies de l'information émergentes à destination du secteur public, auteur de l'ouvrage *Déployer un projet web 2.0 : anticiper le web sémantique »,* paru chez Eyrolles en 2008.

4.1. Introduction

Les « Big Data » : un lieu d'innovation

Big Data : *« Désigne les masses de données auxquelles sont confrontés les acteurs du secteur privé comme du secteur public et qu'ils veulent/peuvent exploiter pour générer des nouveaux business et/ou être plus efficaces ».* Les « Big Data » sont un lieu d'innovation. Comme tous les lieux d'innovation, la valeur s'y crée en puissance sans que le ROI soit garanti. Comme tous les investissements en R&D, l'innovation dans les « Big Data » doit déboucher sur un produit ou un service à commercialiser sur un marché. Il s'agit alors d'identifier le ou les *Business Models* (BM) susceptibles de **créer**, **capturer** et **délivrer** la valeur à partir des « Big Data ».

Le Business Model : un lieu d'innovation

Intuitivement, l'innovation en matière de modèles économiques semble davantage dans l'ADN des start-up que celle des grands groupes industriels où les freins sont nombreux : poids de l'existant qu'il faut rentabiliser, turn-over des équipes, impact négatif sur le cours des actions pour les groupes cotés en bourse… Une équipe de 3 personnes, sans charges fixes, saura être plus agile qu'une multinationale. Toutefois, il manque aujourd'hui aux start-up une capacité d'innovation dans les modèles économiques. La maîtrise de l'innovation technologique ne suffit plus à l'heure où les marchés sont structurés par les usages et où tout devient « service ».

La notion de BM (« modèle économique », ou « modèle d'affaire ») trouve son origine dans les travaux du théoricien du management Peter Drucker (US)[25]. Elle repose sur deux principes forts :

1. **La recherche de l'optimum :** il existe une meilleure manière de faire ; et tout travail, produit, service peut être analysé, décomposé puis recomposé pour être amélioré.
2. **La recherche de la simplicité :** plus un modèle de revenu est simple, plus il est rentable ; plus une entreprise concentre ses efforts sur ce qu'elle sait faire, plus elle est performante.

Si la notion de BM remonte aux années 50, elle n'est devenue une réalité métier qu'à partir des années 2000. Auparavant, on parlait de « développement stratégique » ou de « business plan » (BP). Cependant, le BM correspond à la logique de création de valeur pour les clients et de revenus pour une entreprise, tandis que le BP désigne la mise en œuvre, sur le plan stratégique et opérationnel, du modèle. Un BM est un « outil » qui décrit la façon dont un acteur économique ou un écosystème économique va : **créer de la valeur** (là où il n'y en a pas assez), **capturer la valeur** (pour ne pas la perdre), **délivrer la valeur** (pour la vendre, la commercialiser, la faire circuler).

[25] *The Practice of Management*, 1952

Les challenges du « Big Data »

Défis technologiques : les « Big Data » promettent aux DSI une baisse drastique des coûts de production et de stockage des informations. Les « Big Data » promettent également aux acteurs industriels de la R&D (start-up, laboratoires, …) des innovations disruptives nombreuses dans la création de technologies de pointe dans le domaine du traitement de l'information.

Défis économiques : les « Big Data » promettent aux Directions Générales de faciliter la prise de décision. Les entreprises sont prises dans un déluge de données à l'intérieur duquel les situations à risque doivent être détectées : risque de sécurité, risque commercial, comportements frauduleux, fiabilité… Pour les acteurs de l'IE et du décisionnel, les « Big Data » représentent un champ d'opportunités pour incuber de nouveaux services facilitant la transformation de la donnée en information stratégique, en connaissance.

Défis dans les usages : les « Big Data » font aux utilisateurs de solutions la promesse d'une utilisation simplifiée d'applications de pointe. Avec le « Cloud », l'utilisation devient « transparente » : sans interférences avec les problèmes de maintenance, d'espace ressources, de sauvegarde des données… Cependant, l'acceptabilité sociale des « Big Data » sera déterminante pour leur développement. Sur les réseaux classiques, l'utilisateur reste partiellement détenteur de ses données, mais la convergence du Cloud et des « Big Data » va pousser les logiques de découplage déjà éprouvées dans l'industrie des télécoms. Demain, données, traitements, infrastructures et applications pourront être entièrement externalisés et l'accès déterminé par un modèle de location / abonnement. Les utilisateurs n'y consentiront que si le bénéfice à l'usage est avéré, et que les prestataires offrent des garanties à travers des « environnements de confiance » : sécurité, intégrité des données confidentielles, données personnelles, ….

4.2. L'émergence de Modèles Economiques dans l'économie des « Big Data »

Raisonner en « business model » est devenu une nécessité

Depuis les années 2000, une série de ruptures ont contraint les acteurs économiques à adopter une approche « BM » et à penser des modèles de de commercialisation radicalement différents à partir des années 2000.

Ruptures technologiques : les nouvelles technologies ont bouleversé la manière de concevoir la rentabilité d'une entreprise, de penser le prix d'un bien. Dans l'économie numérique, tout produit, toute œuvre, peut être reproductible potentiellement à l'infini. La valeur ne peut plus être corrélée à la rareté.

Ruptures économiques : la « nouvelle économie » est marquée par la croissance effrénée, parfois mal maîtrisée, de certains acteurs, et un bouleversement du contenu des activités économiques. A l'intérieur de celles-ci, tout change : ce que l'on produit et la manière dont on le produit ; ce que l'on consomme et la manière dont on le consomme ; ce que l'on finance et la manière dont on le finance. Par conséquent, le nouvel indice de performance des entreprises est bien la valeur, et l'ensemble des opérations qui la font vivre (création, capture, gestion, libération).

Ruptures de marché : le marché n'est plus structuré par les logiciels, les machines et les réseaux (le *hardware*), mais par les usages, les services et les utilisateurs. La valeur de la *data* a dépassé celle du « *hardware* », mais les acteurs continuent d'investir en priorité dans le « *hardware* » au lieu d'investir dans la gouvernance des données alors que cet effort peut être sous-traité. Par ailleurs, on passe d'une logique de ROI (*Return-on-Investment)* à une logique de ROD (*Return-on-Data*) : avec le Cloud et les modèles IaaS, la matière première (les données), l'infrastructure (les plateformes) et le calcul (les traitements) peuvent être externalisés. A l'avenir, les raisonnements en ROI (machine, logiciel, infrastructure, réseaux, …) cèderont peut-être la place à des raisonnements en ROD (Volume, Variété, Vélocité et Valeur).

Ruptures réglementaires : l'entreprise est entrée dans un jeu de régulations et de dérégulations démultiplié sous l'effet combiné de l'apparition de nouvelles possibilités pour générer des revenus et de la complexification des relations entre parties prenantes. Les états ont été obligés d'imposer certaines règles pour limiter les abus, mais ces dispositions sont toujours intervenues après coup. La bulle internet n'a pas été anticipée. Dans ce contexte d'hyper-concurrence, la nécessité de sécuriser chaque étape du développement stratégique impose aux décideurs économiques de raisonner en termes de « modèles économiques ».

Le Business Model Canevas (BMC)

Le BMC est un modèle de référence destiné à fournir une représentation simplifiée des différents modèles économiques existants. Le BMC a été pensé par Osterwalder, et s'appuie sur l'étude approfondie des similarités existantes entre un grand nombre de conceptualisations de BM (+ de 450). Neuf briques de base forment ainsi la matrice de tout BM :

1. **SC : les Segments de Clientèle :**
   - ⇨ A qui je m'adresse ? Qui je cible ? Quels besoins ?
     Dans le cas des « Big data », tous les secteurs sont impactés : Administration, Banque, Industrie, Santé, Universités et Instituts de Recherche, Etablissements culturels, Distribution, …
2. **PV : Les Propositions de Valeur**
   - ⇨ Qu'est-ce que j'apporte ?
     Dans le cas des « Big Data » : de la fourniture de service ? de la donnée ? de la puissance de calcul ? de l'infrastructure ? du conseil ? de l'expertise ?
   - ⇨ Quelle offre ?
     Dans le cas des « Big Data », l'offre est très diversifiée : collecter des données, traiter des fichiers logs, des applications verticales, des solutions d'analyse, de visualisation, d'interprétation, des bases de données structurées, du stockage ….
3. **CD : Les Canaux de Distribution**
   - ⇨ Le marché des « Big Data » suit essentiellement des logiques de distribution B2B et A2C.
4. **RC : La Relation-Client**
   - ⇨ La relation entre ce que je propose (PV) et ce que le client veut (SG)
   - ⇨ Dans le cas des « Big Data » : self-service, formation, support-maintenance, …
5. **R€ : Les Flux de Revenus**
   - ⇨ Soit les flux générés par l'ensemble des transactions autour de ma proposition de valeur : comment les réinvestir ? En R&D ? Dans des dépenses de fonctionnement ? à l'investissement ? Dans l'amélioration de l'existant ?
   - ⇨ Dans le cas des « Big data » : valorisation des données, Ventes / reventes, Lot / transaction, PI, Licences, loyers …

6. **RES : Les Ressources Clés**
   - ⇨ Compte tenu de la complexité des projets « Big Data », il faut s'attendre à ce que l'expertise devienne un important lieu de création de valeur, et le conseil une activité clé.
   - ⇨ Ressources physiques : Débit, Performances, Espace, …
7. **AC : Les Activités Clés**
   - ⇨ Les « Big Data » segmentent les activités clés, qui peuvent être accomplies par différents prestataires spécialisés : conseil, calcul, stockage, filtrage, nettoyage.
   - ⇨ Toutefois, les activités comme le conseil en gouvernance de données ou le nettoyage et l'enrichissement seront centrales car les données sont au centre de l'écosystème.
8. **PC : Les Partenaires Clés**
   - ⇨ Quels partenaires et qu'apportent-ils ? Quelles sont leur PV ? De la fourniture de service ? de la donnée ? de la puissance de calcul ? de l'infrastructure ? du conseil ? de l'expertise ?
   - ⇨ Avec les « Big Data », il est impossible de tout faire seul. La coopération économique et les contextes d'Open Innovation sont encouragés. La dépendance mutuelle entre acteurs de la chaîne est forte. Rôles et pouvoir sur la chaîne de la valeur sont redistribués selon les grandes fonctions : hébergeurs, datacenters, HPC (High Performance Computing), constructeurs (de serveurs), intégrateurs, opérateurs Cloud : ceux qui sont au plus près des données sont gagnants !
9. **C€ : La Structure des Coûts**
   - ⇨ Maintenance de la plateforme
   - ⇨ Abonnement
   - ⇨ Développement
   - ⇨ Exploitation

4.3. Zoom sur les segments de clientèle

Tous les secteurs d'activités sont, à une échelle variable, impactés par les « Big Data ».

Santé : chaque hôpital détiendra au moins 150 TB et jusqu'à 650 TB en 2015 (imagerie médicale, données).

Service client : il y a 4 ans, 59% des clients quittaient leurs fournisseurs s'ils avaient une mauvaise prestation, aujourd'hui on est passé à 86%.

Assurances, administrations : font face à des fraudes en quantité croissante.

Services financiers : un des secteurs d'activités les plus précocement confrontés à la problématique des Big data, avec le développement du « trading algorithmique » et de l'information « temps-réel ». Par exemple, un agrégateur de flux de données comme Dow Jones Factiva : + 19000 « news par jour ».

Grande distribution : un autre secteur « historiquement » confronté aux Big Data. Les ventes ratées pour défaut en stock représentent 170 M$ aux USA, l'analyse des tickets de caisse.

Télécommunication : 5 milliards d'abonnés au portable qui attendent des services personnalisés.

Ministères : détiennent de prodigieuses quantités de données sur la population de chaque pays, …
Bibliothèques : l'importante croissance des fonds patrimoniaux et d'archives engendrée par la numérisation confronte les institutions culturelles à la notion de Big Data : Comment conserver les fonds et les données? Comment organiser l'accès ? Comment relier les données et décloisonner les fonds ?

4.4. Les propositions de valeur (PV)

A l'image de l'écosystème « Big Data », l'offre de solutions et services pour le traitement de gros volumes de données est bien diversifiée, partagée entre une minorité de gros éditeurs de solutions « verticales » et de petites start-ups innovants sur des produits très spécialisés :

Les applications de traitement des logs : *IP-Label (la qualité perçue), Wallix, Splunk, Loggly, SumoLogic, …*
Les applications verticales (ex: big data marketing): *BloomReach, Adobe Social Marketing, Dataguize…*
Les applications de Business intelligence*: SAS, Oracle, SAP, IBM, GoodData, MicrosStrategy, Cognos, Talend …*
Les applications d'analyse et visualisation *: Pikko, GreenPlum, Palantir, Visual.ly, EMC², Karmasphere, Teradata, Datameer, …*
Les fournisseurs de données *: GNIP, INRIX, DataSift, DataPublica, …*
Les fournisseurs d'analyse d'infrastructure *: ATOS, MapR Technologies, Hortonworks, Cloudera, Kognitio …*
Les fournisseurs d'infrastructures opérationnelles : *CouchBase, Teradata, 10gen, Hadapt, Informatica, Couchbase, ATOS, …*
Les fournisseurs IaaS : *Amazon web services, Infochimps, WindowsAzure, ATOS…*
Les fournisseurs de SGBD *: Sybase* SAP, Oracle, MySQL, SQLServer, SkySQL, Armadillo …
Des technologies *:* Hadoop, HiBase, Cassandra, …

4.5. Les Modèles Economiques existants

Les BM classiques du web
Le web a été le terrain d'expérimentation de nombreux modèles économiques en rupture avec l'économie de l'information traditionnelle :

- Gratuité et publicité
- Abonnement à des API
- Abonnement à d'autres fonctions
- Vente de data de qualité (premium)

Le modèle de FREE
Le modèle de l'opérateur FREE est à fort potentiel pour les « Big Data ». Il repose sur 3 piliers :

1. **La gratuité** : principe de l'appât et de l'hameçon. L'accès au service basic est gratuit, mais l'utilisateur accepte en contrepartie que ses données soient collectées. La gratuité permet de faire jouer l'effet de masse pour « capturer la valeur ».
2. **Modèle multi-face :** les clients de FREE sont aussi bien les usagers finaux en aval que les annonceurs en amont. Chaque segment de clientèle est gagnant.

3. **Offre Premium :** le couplage d'une offre gratuite « basique » pour un large panel de client, et d'une offre payante « enrichie » a fait ses preuves dans l'économie du web. Il permet d'équilibrer la structure des coûts et de continuer à investir en R&D.

Le modèle Dégroupé

L'industrie des Télécoms a fait émerger des modèles reposant sur le **principe du découplage** : l'infrastructure et sa maintenance sont externalisés (France Télécom), les opérateurs privés louent l'accès aux réseaux et proposent leurs services sur abonnement. Le secteur des « Big Data » se prête « nativement » assez bien aux modèles découplés. A cette différence près que ce ne sera plus celui qui détiendra les réseaux et l'infrastructure qui sera au centre, mais celui qui détiendra les données autour desquels les autres acteurs vont graviter (ceux qui fournissent l'accès, l'intégration, le calcul, l'hébergement, le conseil…).

5. L'exploitation des données scientifiques

Mark Asch est Chargé de mission Interdisciplinarité et Calcul au CNRS, Professeur de Mathématiques, LAMFA, UMR 6140, Université de Picardie Jules Verne. Son intervention illustre la manière dont les communautés scientifiques commencent à intégrer les « Big Data » à travers des projets opérationnels du CNRS. Elle vise aussi à souligner le rôle moteur que peut jouer le CNRS dans ce nouvel écosystème, et l'exemple que constitue la création d'une Mission pur l'interdisciplinarité.

Lien vers la présentation de l'intervention :

http://www.gfii.fr/uploads/docs/BigData_Donn%C3%A9esScientifiques_ASCH.pdf

5.1. L'interdisciplinarité : un enjeu stratégique

Depuis 3 ans, le CNRS est structuré en 10 instituts disciplinaires. En juillet 2012, une Mission pour l'interdisciplinarité est créée, placée sous l'autorité directe de la Direction Générale Déléguée à la Science, également responsable des projets de *Très Grands Equipements* (Isidore, Adonis). Son objectif : décloisonner les activités, mettre à la disposition des communautés des infrastructures et des outils techniques et impulser des projets dans le domaine du traitement des données « gros volumes ».

Il est naturel que cette mission coiffe les projets « Big Data » car ce dossier dépasse les barrières disciplinaires. D'un point de vue économique comme scientifique, la valeur ajoutée des « Big Data » réside dans le croisement des silos et la découverte de nouvelles connaissances à travers l'exploration des « données de masse ».

Le décloisonnement disciplinaire apparaît aujourd'hui comme **un enjeu stratégique** dans la **concurrence internationale** : les Etats-Unis jouissent d'une avance significative sur les « Big Data ». Non seulement parce qu'ils ont la culture de l'innovation technologique, mais aussi parce que l'organisation de la recherche scientifique répond depuis longtemps à un schéma de mutualisation et d'Open Gouvernance. Les états européens commencent seulement à se familiariser avec ces notions.

Les Etats-Unis investissent massivement dans les technologies « Big Data » pour la recherche fondamentale et appliquée. En mars 2012, le NIS (National Institute of Health) avait débloqué 10 millions de $ pour financer des projets de TGE dans le domaine des géosciences[26]. Début octobre, c'est 15 millions de $ qui ont été débloqués en partenariat avec le NIH (National Institute of Health) pour financer des projets de recherche fondamentale dans le domaine de l'extraction de connaissance à partir de *« Large Scale Datasets »*[27].

5.2. Organisation et actions de la mission

La mission est gérée collégialement via un comité de pilotage dans lequel un représentant de chaque institut et son directeur sont présents. Ses actions se répartissent en 4 axes :

- Les défis technologiques

[26] http://www.nsf.gov/news/news_summ.jsp?cntn_id=123607
[27] http://www.nsf.gov/news/news_summ.jsp?cntn_id=125610&WT.mc_id=USNSF_51&WT.mc_ev=click

- L'interdisciplinarité en réseau
- L'interdisciplinarité sur site
- Autres actions

5.3. Zoom sur les défis scientifiques

Les défis scientifiques recouvrent des programmes de recherche stratégiques, c'est-à-dire menés sur le temps long (> 5 ans), à risque, et dont l'objectif est de favoriser l'émergence de nouveaux champs disciplinaires en faisant jouer la « fertilisation croisée » entre les disciplines et les communautés.

5 grands défis ont été inscrits dans la feuille de route de la mission en 2012 :

1. NEEDS (Nucléaire, Energie, Environnement, Déchets et Santé)
2. Défi SENS : insuffisances perceptives et suppléance personnalisées
3. **MASTODONS** : Très grandes Masses de Données Scientifiques
4. Biologie Synthétique
5. Nano-G3N : Graphènes, Nouveaux Paradigmes, Nanomédecine, Nanmétrologie.

Chaque défi est financé à hauteur de 500 000 à 1 million d'euros par an. Stratégiquement, le développement de la R&D française sur les « Big Data » est placé sur le même plan que les nano-technologies ou la convergence des industries nucléaires et de la protection environnementale.

5.4. Enjeux de la R&D sur les « Big Data »

5.4.1. Caractéristiques de la R&D sur les « données de masse »

La R&D sur les « Big Data » est un domaine transversal, par nature interdisciplinaire : informatique, cybernétique, Intelligence Artificielle, algorithmique, statistique, sciences cognitives… Les champs de recherche sont extrêmement vastes et périodiquement contraints à se repositionner car la R&D sur les « Big Data » est principalement tirée par le développement de nouvelles applications industrielles, par l'émergence de nouvelles technologies et de nouveaux usages, et par l'apparition de nouvelles problématiques à résoudre. Parmi ces champs de recherche, on citera : la **modélisation** et la **simulation**, l'**apprentissage statistique**, l'**inférence logique** ou le **Calcul Haute Performance**.

Autre caractéristique, cette R&D est très majoritairement dominée par les laboratoires industriels, avec une grande perméabilité entre le monde académique et le monde industriel (surtout aux US). Cette domination se traduit par l'avance significative du monde anglo-saxon et du secteur privé dans les développements industriels des « Big Data ». Aux Etats-Unis, la recherche et la R&D des « Big Data » sont d'abord impulsées par les « géants » : IBM, Oracle, Microsoft, Sun, AT&T, Bell Labs, Google, Yahoo!... Ceci pose le problème de la souveraineté de la recherche scientifique publique[28].

5.4.2. Objectifs généraux de la R&D sur les « données de masse »

En termes d'objectifs généraux, la R&D sur les « Big Data » œuvre à trouver des solutions sur les grands verrous existant à la gestion de données « gros volumes » :

[28] Voir par exemple le projet « **1000Genomes Project** », impulsé par le NIH (National Institute of Health) et Amazon visant à faciliter la cartographie de l'ADN humain : http://bits.blogs.nytimes.com/2012/03/29/amazon-web-services-big-free-genetic-database/?scp=1&sq=Genomes&st=cse (29/03/12)

- La virtualisation du stockage et de l'accès et des applications (Cloud).
- L'intégration de données.
- La gestion d'événements, de séries temporelles et de flots de données (*event processing, data streams*).
- L'analyse complexe à grande échelle.
- La qualité et la protection des données.
- La visualisation/navigation dans les masses de données.
- La préservation des données.

5.5. Les challenges de la R&D sur les « données de masse »

5.5.1. Les challenges du Cloud Computing

Les bénéfices du Cloud commencent à être bien compris par les directions des grands groupes : *scalabilité* des infrastructures (modèle « *pay as you go* »), stockage massif de données, réduction des coûts de stockage et d'utilisation, accès *anytime* / *anywhere* via Internet, amélioration de la qualité de service (disponibilité, sécurité), élasticité des infrastructures…

Sur le plan technologique, le Cloud doit encore relever un certain nombre de défis relatifs à :
- L'indexation intelligente (apport du web sémantique)
- La sécurité et la confidentialité des données (*privacy*, trust)
- Le développement du calcul intensif, la cohérence et la qualité des données.

Avec les « Big Data », on sort de l'ère *« One Size Fits All »* : il faut offrir des architectures de données flexibles, avec des services de gestion de données adaptables à chaque type d'application/type de données. Les SGBD ne sont plus visibles en tant que systèmes intégrés et cohérents : les services de gestion de données sont enfouis dans des systèmes à plus forte valeur ajoutée (services métiers).

Par ailleurs, il reste un important travail de sensibilisation à accomplir envers les PME, les administrations et les établissements de recherche publics. Face à la complexité du sujet, les directions des grands groupes sont d'abord séduites par l'argument de la réduction des coûts opérationnels sans forcément voir les bénéfices indirects : création de nouveaux services, nouveaux usages, ….

5.5.2. Les challenges de l'analyse complexe à grande échelle

L'analyse complexe à grande échelle fait la valeur ajoutée des solutions « Big Data » pour la recherche. Il faut encore réaliser des investissements important en R&D sur l'analyse en **temps réel** de flots continus de données émanant de sources multiples et hétérogènes. Le développement du **calcul intensif** et le passage à l'échelle seront ici déterminants. Parmi les applications possibles : la découverte et la compréhension de patterns concernant les comportements clients / utilisateurs (ex : campagne de fouilles prédictives au sein de fichiers « logs », web mining).

Il faut aussi développer la **réactivité** des solutions à des **événements d'alerte**. Parmi les applications possibles : solutions de luttes contre la cybercriminalité, de gestion de crise, de risk management, e-réputation…

Il faut enfin améliorer le requêtage multidimensionnel à des fins d'extraction de connaissances sur des grands ensembles de données. Parmi les applications possibles : découverte et compréhension de patterns analysant le comportement d'une population[29].

5.5.3. Les challenges de la gestion de flots et d'évènements

La R&D doit encore fournir des développements dans les domaines de la gestion des flots de données et d'événements sur les parties :

- Capture d'événements, ex : détection / simulation d'événements rares
- Réaction aux événements, ex : couplage transactionnel
- Bufferisation, ex : taille des fenêtres temporelles
- Historicisation, ex : indexation intelligente

5.5.4. Les challenges de la visualisation

Les chercheurs ont aujourd'hui besoin d'interfaces de navigation intuitives et contextualisées dans les données. Il faut pouvoir modéliser et visualiser des phénomènes complexes et souvent imperceptibles à l'œil nu car situés dans des ordres de grandeur micro (séquençage du génome humain, interactions entre les particules…) et macro (trous noirs, climatologie…). Il faut également pouvoir découvrir des connaissances nouvelles qui seraient restées tacites sans traduction visuelle, en corrélant des phénomènes décorrélés dans l'expérience immédiate.

Les réservoirs d'innovations sont nombreux pour la visualisation :

- Inventer de nouvelles métaphores graphiques en étudiant leur impact cognitif ;
- Développer des algorithmes de graphes à haute performance pour passer d'une logique de visualisation en différé à une logique de visualisation temps réel
- Développer la *clusterisation* et les statistiques de graphes.
- Faciliter l'adaptation des solutions et la restitution des données aux terminaux en tenant compte de l'évolution des usages (ex : mobilité).

Les apports de l'IA, du web sémantique, du calcul intensif et des technologies « *in memory* » seront ici décisifs.

5.5.5. Les challenges de la préservation et du stockage des données

Le Cloud Computing est encore trop souvent analysé sous l'angle de l'exploitation des données, beaucoup plus rarement sous celui de leur préservation. Il faut développer des architectures capables de préserver :

- Les données dont la durée de vie est potentiellement illimitée : données scientifiques, culturelles, archéologiques, environnementales…
- Les donnée dont la durée vie est longue mais pas illimitée : patrimoine informationnel des entreprises, données personnelles, données publiques …

On ne pourra pas non plus faire l'économie d'une réflexion sur les données prioritaires pour l'archivage, et sur les coûts de leur préservation : coût de conversion dans des nouveaux formats,

[29] Voir par exemple le batterie d'indicateur « **MarketPsych »** développées par Thomson Reuters à destination des investisseurs financiers, qui permettent de jauger en temps réel la psychologie des marchés en se fondant sur l'analyse de l'opinion publique sur les réseaux : http://thomsonreuters.com/content/press_room/financial/2012_06_25_thomson_reuters_adds_psychological_analysis_to_machine_readable_news (25/06/12)

coût de migration vers d'autres systèmes, coûts de maintenance des plateformes et coûts de maintien des technologies de niche.

Il faut s'attendre à ce que les masses de données soient de plus en plus hétérogènes et complexes à traiter : on ne pourra pas tout conserver. Ces réflexions seront bien sûr fonction de la capacité des **instances de normalisation** à garantir l'i**nteropérabilité technologique et juridique** entre les acteurs du secteur[30].

5.6. Zoom sur le défi MASTODONS

5.6.1. Organisation / moyens

Le défi MASTODONS est coordonné par Mark Asch et Mokrane Bouzeghoub. Il est financé à hauteur de 700k€ pour l'année 2012. Il n'y a pas eu d'appel à projet *stricto sensu* dans la mesure où il n'existe pas encore de programme de recherche structuré sur les « Big Data » en France, plutôt un appel à intérêt envers les communautés scientifiques ayant des problématiques communes à résoudre.

MASTONDONS a pour objectif de rattraper le retard des instituts de recherche français sur la recherche et la R&D en « Big Data » par rapport au monde anglo-saxon. Par ailleurs, on constate que les laboratoires sont techniquement et scientifiquement dépassés. Les technologies d'acquisition et de captation des données ont explosé, les volumes à traiter sont colossaux. A titre d'exemple, le volume d'information à ce jour cartographié dans le génome humain représente 1 Gigabyte[9]. Sur certains sujets, il y a un continuum entre le micro et le macro. Par exemple, en recherche sur l'ADN humain, l'étude d'un simple échantillon du génome permet de modéliser le fonctionnement de tout un écosystème vivant. Les communautés ont besoin de technologies capables de faire la navette entre ces ordres de grandeur.

Les premiers résultats sont très positifs. 74 unités de recherche ont répondu favorablement à l'appel et soumis 37 projets. A l'arrivée, 18 projets ont été sélectionnés impliquant 43 laboratoires. Après fusion, 16 projets sont ou seront financés par l'enveloppe MASTONDONS. Certains projets peuvent impliquer jusqu'à 7 unités de recherche. La démarche est véritablement décloisonnante, ce qui représente un défi sur le plan opérationnel. Il faut créer des passerelles entre les institutions de tutelle et faire converger des laboratoires aux cultures parfois très différentes (informatique, mathématiques).

5.6.2. Axes de travail

La mission pour l'interdisciplinarité a défini 7 axes de réflexion pour MASTODONS, en lien avec les grands objectifs de la R&D sur les « Big Data » (cf *supra*) :

1. **Stockage et Gestion des données**
   Ex : Cloud, sécurité et confidentialité des données
2. **Calcul Haute Performance**
   Ex : technologies « in memory », virtualisation des applications, clusterisation des serveurs, parallélisme dirigé par les données, …

[30] Sur ce sujet, lire le Livre Blanc du JISC *« Curation In the Cloud »*, publié en septembre 2012 visant à informer les institutions de recherche sur les opportunités, les risques et les facteurs à prendre en compte pour l'archivage des données de recherche dans le Cloud : http://www.jisc.ac.uk/whatwedo/programmes/preservation/CurationCloud.aspx

3. **Visualisation**
   Ex : comment passer d'une logique de visualisation en différé sur des données préconstruites à une logique de visualisation « temps réel » ?
4. **Extraction**
   Ex : Data mining et apprentissage, apports de l'IA, …
5. **Propriété**
   Ex : problèmes de propriété, droit à l'usage, droit à l'oubli, …
6. **Préservation / archivage**
   Ex : Utilisation du Cloud à des fins patrimoniale, encore peu développée.
7. **Exploitation**
   Ex : Bases de données scientifiques, réseaux sociaux, corpus littéraires (digital humanities)…

5.6.3. Focus sur des projets

Projet CrEDIBLE

Dans le domaine des « Sciences de la Vie », on trouve les projets CrEDIBLE, « Séquençage et phénotypage haut débit » et SABIOD. Le projet CrEDIBLE vise à résoudre un certain nombre de défis computationnels complexes liés à la fédération de données et de connaissances distribuées en imagerie biomédicale. Il montre l'apport des technologies du web sémantique au « Big Data » : comment aligner et fédérer des entrepôts de données hétérogènes ? Comment sémantiser les données d'imagerie médicale complexes au moyen d'ontologies de domaines multiples ? Comment lancer des requêtes multidimensionnelles et analyser en temps réel des flots de données d'imagerie biomédicale sur des infrastructures distribuées (cloud, grid) ?

Projet « séquençage et phénotypage haut-débit »

Le projet *« Défis Computationnels des séquençages et phénotypage haut-débit en sciences et vie »* illustre le caractère hautement stratégique des technologies du Big Data pour la recherche biomédicale. A terme, le déploiement de dispositifs de collecte de données de terrain, extrêmement couteux, pourra être économisé. Par exemple, à partir d'un simple échantillon de génome, il sera possible de modéliser le comportement de tout un écosystème. Toutefois, il faudra au préalable avoir relevé un certain nombre de défis computationnels complexes concernant l'algorithmique, l'indexation et le rapprochement sémantique des séquences, l'exploitation des architectures parallélisées (grid, cloud) ou le partage de la fouille de données à grande échelle pour établir des prédictions d'événements biologiques.

Projet SABIOD

Le projet SABIOD montre l'apport des données de capteur dans l'étude de la biodiversité. Des capteurs sonores posés sur des orques renvoient des flux de données très importants via Wifi. L'analyse et la modélisation de ces flux permettent de comprendre leurs comportements, rythmes, interactions. Là encore, l'observation d'un minimum d'individus permet de comprendre le comportement d'une espèce entière et son adaptation à l'environnement. Il s'agit dès lors d'être en capacité de traiter ces flux de données de capteurs : passage à l'échelle, clustering en ligne, classification, fouille de données, modélisation probabiliste nonsupervisée, analyse Bayésienne , adaptation de modèles en ligne, statistique de masse, indexation multi-échelle, interprétation écologique, fusion de connaissances hétérogènes, …

Projet GAIA

En astrophysique, le projet GAIA vise à cartographier en 3D l'intégralité de notre galaxie et sa distribution. Ceci nécessite de mettre en place des outils d'analyse multidimensionnelle de flux de données satellitaires extrêmement complexes et hétérogènes (données astrométriques, photométriques, spectrophotométriques, spectroscopiques). La collaboration entre informaticiens, astrophysiciens, mathématiciens et statisticiens est nécessaire.

Projets DEEPHY et PREDON

Dans le domaine de la physique des particules, on trouve les projets DEEPHY (diapo 30) et PREDON en lien avec le LHC (Large Hadron Collider) du CERN. Le LHC renvoie au CERN des quantités de données colossales (plusieurs dizaines de pétabytes / an) ce qui pose la question de leur exploitabilité et de leur préservation. Le projet DEEPHY fait collaborer physiciens et informaticiens pour faciliter la gestion du cycle de vie des données, leur analyse à grande échelle (algorithmes MCMC), et leur intégration dans l'infrastructure de grille européenne EGI. Le projet PEDRON doit concerner plus largement l'archivage à très long terme des données de la recherche en physique nucléaire. L'exploitabilité et la préservation des données issues du LHC sont fondamentales car elles irriguent tout un pan de la recherche internationale en physique subatomique.

Projet ARESOS

En SHS, on trouve le projet ARESOS dont l'objectif est de modéliser, à travers l'analyse en temps réel de vastes corpus textuels, les réseaux sociaux et les circuits de circulation de l'information (implicites ou explicites) qui peuvent exister entre des producteurs de contenus appartenant à des sphères sociales très différentes (tweets, blogs, pages web, articles de journaux, articles scientifiques, corpus juridiques), ceci pour mesurer et visualiser les dynamiques sociales à l'œuvre dans notre société. Ce projet montre l'apport des technologies « Big Data » dans le champ des SHS « analytics », et doit résoudre les questions liées à la taille des données, l'hétérogénéité et la complexité des données, la dynamicité des informations, la variété des échelles de temps et des sphères sociales impliquées.

Conclusion / perspectives

MASTODONS montre que les « Big Data » sont un levier pour construire la science de demain. C'est toute l'organisation de la recherche qui doit être repensée en termes de process, de culture. 2012 est l'année de la mise en place de l'interdisciplinarité au CNRS et du lancement de projets « exploratoires ». 2013 sera l'année de l'essaimage : ouverture à d'autres organismes de recherche (INSERM, INRA, INRIA…) et aux entreprises et industries de l'information. Il est nécessaire d'avancer vite compte tenu du retard national sur la question.

5.7. Questions / échanges avec la salle

Quelle(s) infrastructure(s) pour le Cloud ?

Un appel à projet est en cours pour déployer une infrastructure en Cloud propre au CNRS mais les ressources sont déjà largement mobilisées par les projets MASTODONS. La Direction de la Science du CNRS est consciente cependant qu'il faudra tôt ou tard investir dans un Très Grand Equipement orienté pour Cloud et « Big Data » pour répondre à la demande des laboratoires.

6. MIA : A Market place for Information and Analysis. An example of a "Big Data" project in Germany.

Stefan Geissler est Directeur de TEMIS Germany et co-fondateur de TEMIS, après une expérience au centre R&D de IBM Heidelberg dans le domaine de l'informatique linguistique. Il a présenté le projet MIA dont TEMIS Germany est partenaire.

6.1. Contexte

Avec le développement du web 2.0, les données non-structurées sur le web germanophone explosent (+ 1 million de tweets / jour, + 500 articles de presse allemande et 500 000 commentaires par jour). Avec plus de 6 milliards de pages en allemand, **le web germanophone est second en termes de nombre de pages web**, juste après le web anglophone.

Ces gisements, la plupart du temps gratuits, sont aujourd'hui peu ou pas exploités – seulement par les leaders du marché – en raison des coûts d'accès à la puissance de calcul et des coûts d'investissement en R&D pour relever les défis technologiques du traitement des « Big Data ».

Cette situation est dommageable :

1. Le potentiel pour la création de services à haute valeur ajoutée par des **PME** innovantes existe dans de multiples domaines : *market research, trend research, sentiment analysis, BI, …*

2. Il y a une **demande croissante des professionnels** : pouvoir mesurer la valeur d'un objet sur le web (produit, opinion, marque) et suivre ses variations en temps réel en collectant et en analysant des « données de masse ». On peut penser qu'un domaine comme l'analyse de tonalité (*sentiment mining*) sera amené à se développer de manière arithmétique avec la multiplication des contenus « *user generated* ».

3. Les plateformes de *sourcing / monitoring* traditionnelles ne répondent pas à ces nouveaux besoins car elles n'ont pas la **puissance de calcul**.

4. Un « gap » technologique est en passe d'être franchi : les industriels sont en capacité de traiter et organiser de larges flux de données ce qui était encore impossible il y a 10 ans. Les coûts d'équipements et de maintenance baissent grâce aux modèles Cloud (SaaS, IaaS, PaaS, …). Les acteurs sont libérés de la tyrannie du Hardware et peuvent se concentrer sur la création de valeur (services, usages, modèles économiques innovants) : ***« The datacenter is the computer »***.

6.2. Présentation de MIA

Le projet MIA (*Market Place for Information and Analysis*), a été initié par le Ministère de l'Economie et des Technologies allemand en novembre 2011. MIA est une déclinaison industrielle du programme de recherche THESEUS (*New technologies For The Internet Of Services*). Lancé en 2007 en partenariat avec la DFG (la plus grosse agence de financement de la recherche publique allemande), THESEUS est le plus important programme de recherche du

Ministère de l'Economie, visant à doper l'économie des services numériques en soutenant le développement des technologies « Big Data »[31].

L'Université de Berlin pilote et coordonne le consortium MIA[32], dont TEMIS Germany est un partenaire parmi d'autres : *Neofonie[33], Empulse[34], VICO Research & Consulting[35], Fraunhofer Institut für Rechnerarchitektur und Softwaretechnik[36].*

Le projet vise à développer un prototype de **place de marché pour des start-up innovantes** spécialisées dans le domaine de la production, du raffinage, de l'analyse de données et de la création de services à partir des données (en priorité des données publiques, ou gratuites issues du web).

Il s'adresse particulièrement aux PME allemandes ayant la capacité d'innovation mais pas les ressources financières nécessaires pour accéder à la puissance de calcul des *datacenters.* L'écosystème fédéré par MIA regroupe tous les acteurs de la chaine de la valeur : *data providers, developpers, applications vendors, professional analytics*.

L'interdépendance des acteurs de cet écosystème étant forte, il apparait essentiel de :

1. **Réduire les coûts** d'accès aux technologies, d'exécution des traitements et de maintenance tout en proposant aux acteurs un modèle économique soutenable sur le long terme.
2. **Créer un « environnement de confiance »** (*trusted cloud programm*) en fournissant aux partenaires une infrastructure stable et sécurisée pour stocker, vendre, échanger et traiter des données.

A terme, le Ministère de l'Economie et des Technologies souhaite faire de MIA un *« data pool »*, aux avant-postes de l'innovation européenne en matière de création de services en « intelligence de données ».

6.3. Les enjeux technologiques

Le projet est animé par une forte dimension R&D et entretient des liens étroits avec le monde académique. Les résultats des travaux menés par le groupe de recherche DIMA (*Database Systems and Information Management Group)* de la Technische Universität Berlin infusent le projet : modélisation de données, optimisation des requêtes, sémantisation de données non-structurées ou semi-structurées, impact des architectures parallélisées sur le management de l'information, …

Le cœur technologique de la plateforme MIA réside dans le développement de **langages de requête**, d'**algorithmes** pour le traitement et de **connecteurs** permettant d'extraire et d'interroger de vastes volumes de données textuelles sur le web. Le dispositif est clairement positionné sur les approches *« discovery »* et les logiques de *« mining »*. Il s'agit de dépasser les logiques de RI *« full-text »* qui ne permettent pas de faire de l'extraction de faits à partir des données. Il faut notamment construire des index enrichis sémantiquement dans le temps de la requête.

TEMIS est leader dans le domaine de l'analyse et l'enrichissement sémantique de contenus textuels. La société apporte cette brique au projet MIA.

[31] http://theseus-programm.de/en/about.php
[32] Database Systems and Information Management Group (DIMA), Technische Universität Berlin
[33] Le France Telecom allemand
[34] Editeur de logiciel et intégrateur de solutions e-commerce et business 2.0
[35] Spécialiste de la collecte et l'analyse de vastes volumes de données
[36] Le CNRS allemand

6.4. Quel stade de développement ? Quelles perspectives ?

Pour l'instant, l'infrastructure est en cours de développement. La plateforme est ouverte aux partenaires du projet en version béta pour observer des cas d'usages dans les domaines de l'analyse de marchés et de l'analyse média. Une attention toute particulière est portée sur les problématiques de gestion des **données à caractère personnel** (*data privacy*) et la **sécurité des données** (*data security*). Progressivement, la plateforme s'ouvrira à d'autres applications et aux acteurs internationaux.

A terme, MIA devrait s'internationaliser en intégrant la **dimension multilingue**. TEMIS est une société internationale dont les technologies prennent déjà en compte le multilinguisme. Le fait de restreindre le périmètre sur les usages d'analyse textuelle (et pas multimédia), permet d'envisager des développements pour les principales langues européennes : hormis l'anglais, le français, l'espagnol et l'italien.

7. Big Data & Open Data

François Bancilhon mène une double carrière en tant que chercheur à l'INRIA et d'entrepreneur. Il a fondé plusieurs start-up (Data Publica, Mandravi, Xylène, …). Data Publica est née en 2010 dans le cadre de l'appel à projet web 2.0 du Ministère des Finances. Depuis 2011, la société est co-financée par IT-Translation, le fonds d'investissement de l'INRIA et par des *business angels.* Data Publica est spécialisée dans le référencement et l'agrégation de données publiques et privées, dans le développement et la revente de jeux de données publiques « sur mesure » et sur étagère.

7.1. Quelques rappels sur l'Open Data

On constate une baisse de popularité du mouvement « Open Data » au profit des problématiques « Big Data ». En réalité, il y a une réelle convergence entre les deux thèmes. L'Open Data est fondamentalement « Big Data » : avec la Directive de 2001, les administrations sont la source d'un déluge de données dans lequel elles sont elles-mêmes prises.

On peut définir l'Open Data comme un mouvement visant à rendre accessibles et réutilisables les données produites, détenues et maintenues par les organismes publics dans le cadre de leur fonctionnement et de l'exécution de leurs missions (administrations centrales, collectivités, institutions culturelles, …).

L'Open Data recouvre deux volets :

1. **L'accessibilité des données** dans un objectif de transparence (pour les citoyens, les journalistes, les acteurs de la société civile, les ONG, …)
2. **La réutilisation des données** dans un objectif de création de valeur (pour les opérateurs économiques et les organismes publics qui sont eux même réutilisateurs).

L'accès et la réutilisation sont deux problématiques distinctes. La simple mise en accès des données repose sur une logique de communication / diffusion des administrations : le format PDF suffit. La réutilisation des données à des fins de créations de services nécessite de livrer aux réutilisateurs des jeux de données exploitables, c'est-à-dire **interopérables** et **structurés** (à minima du .CSV, idéalement du XML).

7.2. « Editeur de données » : une fonction émergente

En natif, les données publiques ne sont pas conçues pour être publiées, encore moins réutilisées. Ceci nécessite de la part des administrations des investissements importants dans la préparation des données, dans un contexte d'austérité économique. La carte des initiatives Open Data en France montre qu'il y a un vrai **besoin de mutualisation** des dispositifs, moyens et bonnes pratiques, notamment au niveau des collectivités territoriales.

Dans cet écosystème Open Data qui se crée, **il y a donc une place pour des « intermédiaires »** capables de recenser, collecter, consolider et raffiner les données fournies par les administrations aux créateurs de services. Une fonction émerge : celles **d'éditeur de données.** Les appellations sont nombreuses : *datastores, data vendors / suppliers, data curator, ….*

Data Publica est pionnier dans ce domaine en France. Avec plus de 13 000 jeux de données recensés, incluant les 2700 jeux de data.gouv.fr[37], Data Publica est le plus important annuaire de

[37] Chiffres au mois de juillet 2012

données publiques et privées en France et en Europe. La plateforme agrège non seulement des données publiques, mais aussi d'autres données « ouvertes », disponibles sur le web (ex : données crowdsourcées). Chaque jeu indexé est qualifié et préparé pour être réutilisable : **les contenus sont décrits, structurés et livrés dans des formats interopérables**. Les jeux sont téléchargeables. Des visualisations sont également disponibles pour certains à travers des interfaces de **navigation cartographique**. En termes de disponibilité, les jeux présentent différents niveaux de complexité selon lesquels se fondent la tarification : certains jeux sont gratuits, d'autres payants.

Data Publica n'est pas la seule entreprise à se positionner en grossiste de données qualifiées. Une société comme DataMarket a également essayé de construire une place de marché pour l'achat et la vente de données retraitées sur un modèle similaire (alignement de données temporelles, géographiques, …). Les jeux préparés sont décomposés et vendus aux réutilisateurs selon différentes modalités de tarifications, licences.

7.3. Quelle valeur de la donnée publique ?

Le **calcul du ROI** est une réelle difficulté pour l'Open Data. Il est encore trop tôt pour pouvoir quantifier précisément cette valeur. En termes d'innovation, il faut réussir à se projeter dans les possibles. Aujourd'hui, il y a un focus important sur les applications de transport développées par les collectivités car ce sont d'excellentes vitrines pour le marketing territorial, avec une utilité directe pour les citoyens. Il faut cependant aller plus loin : les données publiques peuvent servir à la création de services et de produits d'information à haute valeur ajoutée dans de très nombreux domaines (études de marchés, services info trafic en temps réel, …). Mais pour proposer des réutilisations innovantes, **le croisement de données publiques avec d'autres gisements** est nécessaire : données ouvertes, crowdsourcées ou privées. Là encore, c'est une problématique fondamentale des « Big Data ».

Wall-Mart

Un réutilisateur a compté les flux entrants et sortants de voiture sur les parkings de magasins Wall-Mart afin de réaliser des **prédictions** du CA journalier et du cours de la bourse de l'enseigne. La méthode est artisanale mais donne une idée de ce qui peut être fait en termes de services analytiques à partir des données de capteurs.

Données de la Marine Marchande

Chaque navire de la marine marchande est équipé d'un transpondeur AIS : un émetteur signalant sa provenance / destination et le type de cargaison qu'il transporte. Ces données sont réutilisées pour créer des services d'information sur le trafic maritime ou pour créer des indicateurs économiques.

Données de mobilité

Orange « ping » les GSM des automobilistes toutes les 3 secondes. A partir des **données de géolocalisation** collectées, il est possible d'inférer la vitesse moyenne des automobilistes sur un itinéraire donné et l'état d'encombrement du trafic. Les prédictions sont presque aussi fiables que celles réalisées à partir des données collectées par les infrastructures lourdes des équipementiers d'autoroutes : le taux de succès de détection des perturbations est estimé à 95%, avec une dynamique temps-réel équivalente aux capteurs routiers ; l'écart de vitesse moyen par rapport aux capteurs routiers est estimé à moins de 9 km/h[38]. Orange commercialise ces données auprès des collectivités locales et des compagnies autoroutières à travers le service d'info trafic en temps réel « TraficZen ». Cet exemple montre que le **recyclage de données internes** peut ouvrir de nouveaux débouchés économiques aux acteurs.

[38] http://www.transport-intelligent.net/IMG/pdf/traficZen-view-php-1.pdf

Twitter / Netflix

Netflix propose des services de VOD en streaming sur le web, et de location de DVD à distance. Pour la VOD, le poids des fichiers multimédia requiert une infrastructure réseau très performante afin de garantir la qualité et la continuité du service face aux pics de charge. La distribution des architectures complique la détection des pannes (les prestataires type Netflix sont la plupart du temps hébergés chez des prestataires en Cloud)[39]. En analysant les flux Twitter des abonnés, Netflix obtient en temps réel une image précise de ses performances réseaux et gagne en réactivité en cas de panne. Ici encore, on voit les **économies possibles**.

7.4. Quelle valeur des intermédiaires ?

En tant qu'intermédiaire spécialisé dans la revente de données, il faut d'un côté convaincre les administrations de jouer le jeu de l'ouverture, et de l'autre les réutilisateurs de la valeur des données et du travail d'intermédiaire. Les « éditeurs de données » ont aujourd'hui des difficultés à générer un chiffre d'affaire à la hauteur des promesses économiques de l'Open Data.

Pourtant, il existe une réelle demande. En raison de l'explosion des volumes, les réutilisateurs ont besoin de lisibilité sur l'écosystème des sources. Un développeur d'application n'a pas nécessairement l'expertise pour identifier les sources pertinentes (les vecteurs de l'innovation de demain), et pour les rendre exploitables dans une perspective industrielle. C'est le métier des éditeurs de données.

Les traitements sont complexes et nécessitent une ingénierie de pointe : de **nouvelles compétences** sont requises en indexation, alignement, croisement, mise à jour, et parfois même en sémantisation de données. Il faut aussi maîtriser les nombreux référentiels pour aplanir des données hétérogènes. A ce niveau, il y a une **pénurie de compétences** dans les domaines de la **modélisation** et de l'**ingénierie des connaissances**. Les données publiques ne sont qu'une source parmi d'autres. Les éditeurs de données ne doivent pas se limiter à ce périmètre. Il faut croiser et enrichir les gisements.

7.5. De l'Open Data au Big Data

Aujourd'hui, il faut réussir à **monétiser la valeur du travail des intermédiaires**. Dans cette optique, la maîtrise des technologies « Big Data » doit servir de levier. Construire une liste des hommes politiques français peut être fait manuellement. Pour un réutilisateur, il n'y a pas de valeur ajoutée. En revanche, proposer un index structuré des appels d'offres publics passés sur une année est nettement plus complexe avec + de 1000 appels par jours, + de 8000 sources.

C'est sur la **maîtrise des volumes** que les « éditeurs de données » comme Data Publica ou DataMarket apportent une valeur ajoutée.

Création de l'annuaire Data Publica

Pour créer son portail, Data Publica a réalisé un travail d'inventaire des données publiques produites et disponibles en France. La société a recensé 1362 organismes producteurs, plus de 2000 sites web, 6 millions ½ de jeux disponibles, à plus de 95% du PDF : 481 844 jeux en XML, 148 509 en XLS, 26 518 en CSV, et seulement 369 jeux en RDF (standard du web sémantique). Pour exploiter de telles quantités, la maîtrise des technologies « Big Data » est indispensable.

Réutilisation d'Eurostat

Eurostat est une direction générale de la Commission européenne chargée de l'information statistique à l'échelle communautaire. La base est alimentée par le SSE : Système Statistique Européen qui collecte les données des instituts statistiques nationaux des états-membres, et les

[39] http://cestpasmonidee.blogspot.fr/2012_07_01_archive.html

données de diverses institutions européennes et mondiales (OCDE, FMI, banques et banques mondiales, ONU, …). Les formats sont extrêmement variés.

DataPublica a aspiré 5000 jeux pour les stocker après structuration et harmonisation dans une base de données NoSQL représentant plusieurs terra-octets. Les jeux doivent être maintenus quotidiennement (plus de 20 modifications par jour dans la base Eurostat). A travers ce projet, on voit que la simple création d'une base de données alignées à partir de sources « Open data » atteint très rapidement des contraintes de stockage, de puissance et d'intégration qui nécessite de maîtriser les *process* « Big Data ».

8. « Big Data » et contenus multimédia

Daniel Teruggi est Directeur du Département Recherche et Expérimentation à l'INA.

Lien vers la présentation : http://www.gfii.fr/uploads/docs/BigData_crossmedia_TERRUGI.pdf

8.1. L'Institut National Audiovisuel

L'INA a la responsabilité des archives radiophoniques depuis 1933 et TV depuis 1949. En 1985, l'archivage des productions audiovisuelles de fiction sort du périmètre des attributions de l'INA qui se concentre sur la concentration de l'ensemble des productions audiovisuelles publiques et privées (news, sports, divertissement). Le dépôt légal des productions radios et TV est institué en 1995.

En tant qu'EPIC, l'INA emploie plus de 1000 salariés, avec un budget annuel moyen de 149 millions d'euros, et un financement à hauteur de 68% par l'Etat grâce aux redevances TV. Pour le reste, l'INA se finance à travers ses différentes activités industrielles et commerciales : **vente de contenus d'archives**, **production de contenus,** et **formations** (masters en gestion multimédias, formation continue, …).

L'INA réalise un chiffre d'affaire annuel moyen de 13 millions d'euros sur la vente de ses contenus d'archives. Environ 200 000 ayant-droits ont été identifiés depuis 1933. Le service de gestion des contentieux emploie 30 salariés à temps plein. Tous les contrats de cession de droits de production sont scrupuleusement archivés depuis 1933.

8.2. Chiffres Clés, volumes, ordres de grandeur

L'INA administre deux fonds d'archives très différents :

1. **Les archives nationales radios et TV :**

   - + 1 350 000 heures de flux audiovisuels (700 000 heures de radio, 650 000 heures de TV)
   - + 850 000 heures numérisées et accessibles en ligne à ce jour
   - + 300 000 heures par an ajoutées à l'archive
   - Encodage en « haute qualité »
   - L'INA hérite des droits de production sur les contenus (pas les droits d'auteur) :

   ⇨ **Diffusion aux publics autorisée**, **réutilisation commerciale autorisée**.

2. **Les archives « dépôt légal » :**

   - + 3 millions d'heures, + 100 chaînes TV, +20 radios.
   - + 800 000 heures ajoutées par an à la base
   - Depuis 2009 : dépôt légal des sites web (6000 à ce jour)
   - Contenus encodés en « basse qualité »
   - Pas de droit de production

   ⇨ **Pas de diffusion**, réutilisation commerciale interdite
   ⇨ Accès restreint aux chercheurs via la BNF et d'autres institutions partenaires.

L'« heure » est une unité difficile à appréhender. Quelques rapprochements donnent une idée des ordres de grandeur impliqués. L'INA fait depuis longtemps du « Big Data » sans le savoir.

- 30 000 heures représentent en moyenne 260 000 documents, dont 160 000 documents à caractère commercial / publicitaire.
- 1 année représente 8760 heures
- 1 million d'heures représente 114 années de visionnage / écoute en continu
- Numériser 100 millions d'heures coute environ 102 millions d'euros
- 100 000 heures numérisées représentent 18 péta-bytes de données à 50 MB (HQ)
- Documenter 1 heure représente 4h de travail en amont (traitement documentaire, montage et mise en ligne / archivage …).
- Documenter 1 million d'heures représente 4 millions d'heures de travail, soit 456 années à temps plein.
- Hors litiges et contentieux, la gestion de la disponibilité des droits (*right clearance*) pour 1 heure représente 1 journée de travail.

8.3. La fonction documentaire face aux données de masse

Une indexation à réorienter « archive »

La pertinence et l'efficacité de la valorisation d'un fonds d'archive dépendent de la qualité de sa documentation. Le traitement documentaire orienté « archive » est le cœur de métier historique de l'INA. La plupart des documents audios anciens reçus par l'INA sont mal, voire très mal documentés. Pour les documents vidéos, les documents entrants sont documentés « à la main », y compris les parties droits d'auteur et droits de production. Pour les chaînes TV et radio, cette documentation est généralement sérieuse : la question de la ***« trouvabilité »*** de l'information est cruciale pour les broadcasters, surtout pour les « News » ou les contenus d'archives et les séquences illustratives doivent être retrouvées très facilement dans de vastes banques. Il y a généralement un plan de classement, et les métadonnées essentielles sont bien renseignées (noms de lieux, noms de personnes, …). Cependant, cette documentation n'est pas orientée « archive » mais « usages »: il y a des manques importants sur les **données temporelles**, ou sur les **métadonnées de gestion** (droit d'auteur, licences, droit de production, …).

R&D : automatiser et sémantiser l'indexation

Aujourd'hui, l'INA maintient une **« archive de masse »**. Le pôle indexation regroupe 180 documentalistes qui traitent les flux entrants en grande partie manuellement ou de manière semi-automatique. Les moyens humains sont très insuffisants par rapport aux quantités. En R&D, les enjeux portent sur :

- L'**automatisation** de tâches-clés dans la chaîne documentaire : segmentation, description, résumé des flux
- La description des contenus au niveau du sens
- La **structuration** d'une archive de masse en thèmes, collections et autres critères
- La classification automatique ou semi-automatique (**clusterisation**)
- L'optimisation de la **granularité** des outils de « search » : il faut passer de l'échelle du document à celui de l'extrait, identifier et retrouver un passage d'un contenu au sein de plusieurs millions.
- Développer les **logiques exploratoires** (« discovery »).

Les technologies ***« speech-to-text »*** sont efficaces pour décrire un contenu à la volée de façon très générale, mais elles ne sont pas une solution miracle pour les archives. La langue évolue et les dictionnaires des outils sont rapidement dépassés lorsqu'il s'agit de traiter un contenu d'époque (changements de vocabulaires, accents, inflexions, …). Il faut aujourd'hui pouvoir analyser automatiquement et à très grande échelle les flux entrants. Les algorithmes doivent permettre de qualifier de manière univoque les contenus et de les décrire au niveau du concept. Les **technologies sémantiques**, **d'apprentissage automatique (IA)** et de **reconnaissance d'objets** (détection de contours) sont, par rapport au « speech-to-text » la voie à creuser. L'enjeu

est fondamental. Sur 850 000 heures numérisées (portail INA on line), seulement 4500 sont consultées (2%). Il faut améliorer la trouvabilité et la « compréhensibilité » des contenus par les moteurs pour que les fonds vivent.

Ainsi, pour l'INA, l'élévation des traitements au niveau sémantique apparaît comme une nécessité afin de relever les défis des « Big Data ». Un partenariat avec un industriel leader du domaine été noué pour mettre au point une solution d'**extraction automatique de descripteurs** sur 300 000 heures de flux sonores et vidéos (soit plusieurs mois de calcul). Il s'agit de prélever en flux continu (logique de *streaming*) des mots-clés qui serviront de métadonnées pour décrire les contenus en lien avec les ressources sémantiques utilisées pour l'indexation et la classification (vocabulaires, thésaurus). Ces métadonnées seront ensuite encapsulées dans les fichiers pour rendre leur sens / contenu « compréhensible » par les outils d'accès.

8.4. Problématiques d'archivage de masse

Optimiser le stockage

Les fichiers multimédia sont lourds : passer aux **architectures distribuées** (Cloud / Grid) et à la **virtualisation** est une nécessité pour gagner en espace de stockage et en puissance de calcul. L'INA a fait le choix d'externaliser le stockage de secours des 2 pétabytes disponibles en ligne chez un prestataire.

Sécurité et intégrité des données lors des migrations

La notion de ***« migration »*** est centrale à l'INA. Elle désigne l'ensemble des méthodes pour assurer la lisibilité de l'existant (données / SI) dans le futur. Plus précisément, la migration de données désigne le fait de modifier l'intégralité d'un ensemble de données enregistrées dans un système source (présent) dans le but de le rendre lisible par un système cible (futur).
Parmi les points durs, il s'agit de garantir l'intégrité et l'accessibilité des données lors des migrations au long terme. Il faut des **formats standards** et des **systèmes interopérables**. Il faut des identifiants pour assurer la permanence des liens vers les objets. La sécurité des données est une priorité de l'INA : il faut en garantir la **traçabilité** et surveiller quelles utilisations sont faites des bases et des contenus, dans le respect du droit de propriété intellectuelle et des bases de données. Il faut aussi **limiter la dégradation** du contenu. Celle-ci est inévitable : les données se dégradent au fil des téléchargements, des sollicitations et des migrations.

Pour les politiques de migration, il y a trois approches possibles, trois choix qui ont tous leurs avantages et inconvénients :

1. Migrer maintenant / récupérer plus tard ;
2. Migrer continuellement ;
3. Ne pas migrer.

En termes opérationnels, il y a deux méthodes principales :

1. La migration à l'entrée ;
2. La migration par paquets.

La migration à l'entrée est le « graal » de l'archiviste. C'est la méthode la plus sûre, mais elle nécessite des modèles de préservation suffisamment universaux pour que les données soient encore lisibles par le système cible **(UPF: *Universal Preservation Format*, UVC: *Universal Virtual Computer*)**. La migration par paquets impose de changer de SI et de formats de données tous les 4 à 5 ans, mais les données sont davantage sollicités, les risques d'altération sont plus importants. Il existe une voie tierce : **la migration sur l'accès**, sans doute la meilleure mais qui nécessite une excellente description des contenus, extrêmement difficile à obtenir sur de très gros volumes de données.

Quelle que soit la méthode choisie, l'apport des technologies du « Big Data » et du Cloud sont et seront de plus en plus nécessaires. Sur ce sujet, les travaux de normalisation du Cloud doivent être suivis avec attention.

8.5. Zoom sur l'initiative « Mémoires Partagées »

L'INA a lancé le 3 juillet 2012 l'initiative « Mémoires Partagées », visant à collecter les vidéos amateurs crowdsourcées pour sauvegarder le patrimoine collectif des territoires. La démarche est collaborative : les particuliers, associations et entreprises sont invités à transmettre à l'INA leurs fichiers qui seront diffusés gratuitement sur la chaîne Dailymotion « INA - Mémoires Partagées ». La réutilisation à des fins commerciales est interdite.

Pour l'INA, il s'agit d'une évolution majeure. Historiquement, l'archive s'inscrit dans une logique « top down », en collectant et en conservant les contenus TV et radio des « broadcasters ». Mais les contenus « user generated » (***UGC)*** représentent aujourd'hui des pans entiers de mémoire collective qui, faute de politique de conservation, tombent dans l'oubli. Pendant longtemps, les émissions TV et radio étaient considérés comme les parents pauvres du cinéma. Le regard de la société a changé avec l'instauration du **dépôt légal**. Ces contenus sont désormais perçus comme le reflet d'un état social, et leur préservation comme nécessaire au travail des chercheurs. Les contenus crowdsourcés prennent le même chemin.

Avec « Mémoires Partagées », l'INA entre dans la **logique contributive** du web 2.0. Sa présence sur le web sera considérablement renforcée. C'est un moyen de se rapprocher des hébergeurs comme YouTube ou Dailymotion. Le **métier et les problématiques changent aussi**. L'INA a déjà dû apprendre à numériser 850 000 heures de documents multimédia. Aujourd'hui, son cœur de métier est devenu la gestion de flux, leur indexation et leur trouvabilité dans un contexte de stockage de masse. Diffuser les vidéos sur Dailymotion implique de les documenter richement en amont pour garantir leur référencement dans les moteurs de recherche.

Mais les contraintes sont nombreuses. La collecte doit se faire dans le strict respect des droits d'auteur. Ainsi, un partenariat a été établi avec la société **VidéoForever**, chargée de la collecte et de la conversion des films au format numérique. L'INA se charge de sélectionner les fichiers et de contacter les auteurs pour contrôler les aspects légaux.

Il faut aussi maintenir un **taux de sélection fort**. Ne doivent être conservés que les contenus présentant une réelle valeur patrimoniale. L'Institut souhaite ainsi créer un **patrimoine de proximité « labellisé INA »**.Dans cette optique, le projet est pour l'instant testé en phase pilote dans la région Aquitaine[40].

8.6. Question / Echanges avec la salle

Pourquoi tout conserver ? Quelle sélectivité de la mémoire ?

Avec les « Big Data », c'est la première fois que la société se pose la question de la conservation avec autant d'acuité. La numérisation de l'information décuple l'accessibilité, la reproductibilité et l'utilisabilité des données mais a rendu les contenus volatiles et périssables. A terme, et contrairement aux idées reçues, conserver au format numérique coûte plus cher que sur support analogique, avec des pics, notamment lors des migrations. Les gains réalisés sur les espaces de stockage physiques ne compensent pas les coûts de traitement et d'archivage des données de masse qui croissent de manière exponentielle. Sélectionner l'information à conserver dans l'ensemble des flux coûte aussi plus cher. La question de la sélectivité du patrimoine doit être corrélée à celle du **modèle économique** des acteurs, qui doit être soutenable.

[40] http://www.institut-national-audiovisuel.fr/actualites/memoires-partagees.html

Quel impact des fournisseurs de services d'hébergement et de partage de vidéos ?

Le développement de services d'hébergements comme YouTube ou Dailymotion a déstabilisé le modèle des archives audiovisuelles. Si le conflit juridique avec YouTube a récemment trouvé une issue[41], la question du statut de YouTube n'est toujours pas résolue : est-il simple **prestataire de services** (sans responsabilité légale sur les contenus hébergés) ou faut-il le considérer comme un **éditeur de contenus** (avec une responsabilité légale) ? Aujourd'hui, il est difficile de faire accepter la responsabilité légale des hébergeurs sur les contenus par les tribunaux[42].

Mais il faut aussi envisager les **complémentarités possibles** entre ces services et les archives comme l'INA. YouTube bénéficie d'une puissance de stockage, d'indexation, et de gestion des flux en temps réel bien supérieure à celle de l'INA. Les volumes sont sans commune mesure[43]. Si le service respecte le copyright des détenteurs des droits sur les contenus, l'opportunité pour la valorisation des fonds est considérable. Par ailleurs, ces services évoluent et se positionnent de plus en plus comme des prestataires d'hébergement de contenus contrôlés, autorisés. Des offres structurées à destination des producteurs de contenus se développent pour valoriser les vidéos originales sur des chaînes en bouquet.[44]

Le crowdsourcing peut-il être une opportunité pour la conservation ?.

Il y a des freins importants. L'intégration des **contenus UGC** pose le problème de la quantité et de la qualité des contenus à archiver. Il faut une démarche structurée et un filtrage. C'est le choix de l'INA à travers l'initiative « Mémoires Partagées ». Par ailleurs, les documentalistes ne veulent pas renoncer à la fonction documentaire et à leur savoir-faire. Mais on peut penser que la réalité du « Big Data » les obligera à concevoir des plans de classement et des systèmes d'indexation automatique tenant compte des ***folksonomies*** et des ***tags***. L'enjeu sera alors de maintenir la cohérence de l'accès et de limiter le bruit. La fonction documentaire va très probablement évoluer pour se positionner très en amont de la chaîne des traitements, certaines tâches ayant vocation à s'automatiser. Les documentalistes joueront un rôle de contrôle – coordination, et paramétrage des applications.

[41] L'INA a signé en mars 2012 un accord avec Google, sur le modèle du partenariat établi avec Dailymotion, mettant fin au contentieux engagé en 2008 : l'INA diffuse 57000 vidéos protégées sur la plateforme (watermarking), les recettes publicitaires générées par la lecture des vidéos sont partagées. En échange, YouTube s'est engagé à rechercher activement les vidéos « piratées » sur sa plateforme et à les supprimer. Source : http://www.numerama.com/magazine/22139-l-ina-signe-un-traite-de-paix-avec-youtube.html (21/03/12)

[42] Le contentieux avait abouti à la condamnation de Google par le Tribunal de Grande Instance de Créteil en décembre 2012, l'obligeant à verser 150 000 euros de dommages et intérêts à l'INA et à installer un système de filtrage sur ses contenus. Source : http://www.numerama.com/magazine/17611-l-ina-fait-condamner-youtube-pour-contrefacon.html (16/12/2012)

[43] Pour les 7 ans de la plateforme en mai 2012, Google a dévoilé les chiffres clés de l'infrastructure : 72 heures de vidéos sont ajoutées chaque minute ; 800 millions d'internautes visionnent chaque mois 3 milliards d'heures de vidéos : http://youtube-global.blogspot.fr/2012/05/its-youtubes-7th-birthday-and-youve.html (mai 2012)

[44] Salar Kamangar, le patron de You Tube US, a ainsi annoncé en juin dernier sa volonté de développer une offre payante à destination des broadcasters afin de valoriser les contenus originaux qu'il diffuse à travers ses propres chaînes : http://www.reuters.com/article/2012/06/14/net-us-media-tech-summit-youtube-idUSBRE85D1NQ20120614 (14/06/12) Par ailleurs, YouTube a développé un modèle de « contrat de sponsoring » pour ses chaînes originales afin d'optimiser la monétisation des contenus propriétaires, toujours sur un principe de recettes partagées : http://www.inaglobal.fr/communication-publicite/article/le-sponsoring-de-youtube-fait-fructifier-les-chaines-originales (20/04/12)

9. Table ronde et synthèse animée par Michel Vajou

Michel Vajou est consultant (MV études & conseil), et rédacteur de *La Dépêche du GFII.*

9.1. « Big Data » : les ruptures au-delà du « Buzzword »

Ruptures et continuités

Jusqu'au début des années 2000, la notion de « données de masse » concernait des domaines au périmètre bien délimité. On parlait de ***Very Large Scale Data*** dans certains champs de recherche (génomique, astrophysique, physique des particules, …), de ***data warehouse*** dans le secteur des banques et des assurances, de ***trading algorithmique*** pour l'exploitation des flux d'information financière par les opérateurs boursiers[45]. Certains segments de l'industrie de l'information sont depuis longtemps confrontés à la problématique du traitement des données de masse (information marketing, financière, recherche génomique, information météorologique[46] …), mais aujourd'hui, tous les secteurs sont potentiellement impactés.

Par ailleurs, il y a bien des changements de paradigmes à la fois technologiques, économiques et écologiques. On passe de l'analyse de données structurées à l'analyse de **données non structurées** capturées dans des formats hétérogènes. On passe dans une logique d'analyse d'ensemble préconstruits à une logique de monitoring de flux en **temps réel**. L'architecture client-serveur classique éclate avec la distribution des ressources. Des formes de computation radicalement nouvelles émergent, comme le calcul « *in memory* »[47].

Pour les industriels de l'information, ces évolutions changent la donne et les obligeront à se positionner dans un écosystème radicalement différent au sein duquel ils n'ont plus le leadership.

Zoom sur l'information financière

Les opérateurs boursiers s'équipent depuis une trentaine d'années en terminaux et logiciels de passages d'ordres visant à automatiser les opérations financières grâce à des algorithmes pour la capture, l'interprétation et l'émission d'ordres de plus en plus complexes. A l'époque, on parlait déjà de **temps réel** et les problématiques étaient, sur des volumes moins importants mais déjà considérables, les mêmes : augmenter la puissance de calcul, réduire le temps de latence (développement du ***High Frequency Trading,*** ou ***Low Latency Trading***).

Les technologies du « Big Data » ont permis de franchir un cap. Aujourd'hui, un service comme Thomson Financial peut traiter, avec sa plateforme *Velocity Analytics*, plus de 700 millions de cotations à la seconde, et le temps de latence est de l'ordre du 1000ème de seconde contre 1 seconde il y a 30 ans[48]. La **notion de temps réel a évolué** pour atteindre des niveaux de performances inédits. A cette nouvelle échelle, l'**unité de mesure de référence n'est plus la seconde mais la microseconde.** La configuration géographique des réseaux doit être optimisée pour réduire la durée des transferts de flux d'un point à un autre. Le développement de modules

[45] Sur les vices et les vertus du *trading algorithmique* : http://www.capital.fr/enquetes/documents/comment-l-informatique-a-pris-le-controle-de-wall-street-604813/%28offset%29/1 (09/05/11)

[46] Sur les progrès et le challenge que représente la prédiction météorologique assistée par ordinateur : http://www.nytimes.com/2012/09/09/magazine/the-weatherman-is-not-a-moron.html?pagewanted=all (07/09/12)

[47] « Calcul en mémoire », permettant de stocker les données dans la mémoire vive des cœurs de processeurs et non pas dans la « mémoire externe » des serveurs distribués. On estime que cette technologie est le *« next big thing »* dans le domaine du Cloud. La technologie SAP « HANA » peut ainsi parcourir deux millions d'enregistrements en une milliseconde,, tout en produisant 10 millions d'agrégations complexes par seconde et par cœur. Des traitements qui demandaient auparavant 5 jours peuvent aujourd'hui être exécutés en quelques secondes : http://www.developpez.com/actu/24798/SAP-leve-le-voile-sur-sa-technologie-SAP-in-memory-qui-peut-diviser-par-1200-le-temps-de-traitement-de-certains-scenarios/

[48] Voir la Dépêche du GFII *« Thomson Reuters lance sa plate-forme Velocity Analytics »*, publiée par Michel Vajou sur AMICO le 04/04/11.

de visualisation proprement temps réel devient absolument nécessaire : génération des visuels et de l'analytique qui les accompagne dans le temps de l'analyse des flux, sans pré-agrégation préalable, dans une logique de **monitoring** (développement des ***Visual Analytics***) [49].

Zoom sur l'information marketing

Il y a 30 ans, les producteurs d'information marketing comme Nielsen fonctionnaient sur une logique de panel et d'échantillon étendu (ex : analyses des ordonnances en officine pour les industries pharmaceutiques). Les technologies des « Big data » permettent aujourd'hui de passer d'une logique d'échantillonnage à une logique de captation exhaustive. IMS Health a ainsi pu mettre des projets d'envergure inédite dans le domaine du suivi longitudinal des patients. Auparavant, les producteurs de données avaient la capacité de suivre sur le temps long (15 à 30 ans) une centaine de patients. Aujourd'hui, c'est plusieurs centaines de milliers de patients qui peuvent être suivis.

9.2.Les facteurs d'inertie

La convergence en cours entre l'industrie de l'information et celle des « Big Data » est ralentie par un certain nombre de facteurs d'inertie.

La capacité d'innovation

La problématique des « Big Data » n'est pas native des industries de l'information mais des leaders de l'informatique du contenant et des fournisseurs de solutions de BI (IBM, Oracle, SAP). Aujourd'hui, les acteurs qui impulsent la dynamique à l'écosystème « Big Data » sont les pure-players du web (Google, Amazon, Yahoo !…). Ces acteurs ont la force de frappe pour se positionner sur le premier cercle de différentiation dans cette nouvelle économie : l'infrastructure et la capacité d'innovation technologique.

A ce sujet, un projet comme le « **1000genomesProject »** conduit par le NIH américain (National Institute of Health) est emblématique[50]. La recherche génomique est depuis longtemps associée à la collecte et à l'analyse de très vastes quantités d'informations avec pour enjeu la cartographie du séquençage de l'ADN humain. Pendant longtemps, la technologie était un obstacle à la réalisation d'un tel outil. Aujourd'hui, cet objectif est réalisable mais il est significatif qu'un organisme de recherche public comme le NIH, qui coordonne la collecte d'information auprès de 75 organisations scientifiques, s'associe à un hébergeur privé comme Amazon pour y parvenir.

Du document à la donnée, du « search » au « mining »

Le paradigme du « SEARCH » a pendant 50 ans structuré l'industrie de l'information. Il s'agissait de développer la ***« trouvabilité »*** des documents enregistrés dans de grandes banques d'informations structurées, et la pertinence des systèmes permettant d'y accéder (pertinence mesurée selon différentes métriques comme le silence / rappel).

Ce paradigme existe toujours, mais avec les « Big Data », on assiste à la montée en puissance des logiques de « mining ». L'intérêt se déporte de l'objet « document » vers la « donnée ». L'objectif n'est plus seulement de retrouver des documents au sein de corpus (fournir une liste de résultats), mais d'exploiter ces corpus pour extraire des **connaissances nouvelles**.

Désormais celui qui détient les données et la capacité à les analyser devient central. Les pouvoirs sont redistribués sur la chaîne de la valeur. La stratégie d'un producteur de données comme Thomson Reuters montre que la fourniture de services d'agrégations, de liens et de gestion des références vers les datas devient un positionnement possible pour les éditeurs. Thomson développe depuis plusieurs années des bases de citations et des services analytiques associés (indicateurs) : *article citation index*, *book citation index*, *proceedings citation index*, et le dernier

[49] Voir « *La visualisation analytique au cœur du pilotage temps réel des plates-formes d'information financière* », le GFII 360 publié par Michel Vajou sur AMICO le 13/05/11

[50] http://www.nih.gov/news/health/mar2012/nhgri-29.htm

en date : le *Data Citation Index*, un service d'agrégation, de citation et de liens vers les *datasets* de la recherche.

Pour les éditeurs, il s'agit d'un réel défi : quels nouveaux services d'informations imaginer ? Les « Big Data » deviennent le pétrole de l'industrie de la connaissance, mais il manque aux industriels du secteur la capacité d'innovation pour les exploiter. De lourds investissements en R&D devront encore être consentis.

Quel modèle économique ? Quel impact économique ? Comment chiffrer le ROD ?

Les acteurs du secteur doivent passer d'une logique de ROI (*Return-On-Investment*) à une logique de ROD (*Return-On-Data*). Les grands éditeurs et producteurs de données (Thomson, Elsevier, IMS …) développent des applications analytiques pour exploiter leurs corpus et s'assurer des revenus complémentaires. On sait aujourd'hui que ces services ont une très forte valeur ajoutée mais on est encore incapable de chiffrer précisément leur impact économique.

Or, on assiste à un double mouvement parallèle : d'un côté, une course à l'application d'analyse à haute performance dans des secteurs de niche, et de l'autre, une dévaluation de l'information, de plus en plus considérée comme une « matière première » (***commodity***). Pour les éditeurs, il n'est pas encore certain que les coûts d'accès élevés aux services analytiques qu'ils commercialisent compensent d'une part les lourds investissements réalisés en R&D et d'autre part, la diminution des revenus sur leur métier historique liés à la vente d'information.

La production de métadonnées est un **« pari sur le futur »**[51]. Elle se fait dans la perspective d'usages à venir qu'il n'est pas toujours facile d'anticiper et pour lesquels il est difficile de déterminer un modèle économique. Les acteurs sont confrontés à un dilemme : faut-il investir dans ces usages émergents en ignorant quel sera leur calibrage et renoncer à la vente d'information (cœur de métier des éditeurs) ?

Pénurie de compétences : des nouveaux métiers

De plus en plus, les éditeurs spécialisés acquièrent un savoir-faire dans la production, la collecte, le traitement et la vente de données. Ils jouissent d'une avance certaine dans ce domaine par rapport aux éditeurs généralistes, car ils ont pour la plupart une expérience déjà ancienne dans la vente de banques de données. Ce qui manque aujourd'hui, ce sont des compétences en **« intelligence de données » :** comment interpréter les données ? Comment les exploiter ?

La fonction de **« data analyste »** est déjà très recherchée. Elle nécessite de nouvelles compétences à la confluence du marketing, de l'ingénierie documentaire (modélisation), de la programmation informatique, des mathématiques / statistiques et de la stratégie commerciale. Les compétences ne sont pas seulement techniques. Il faut aussi une culture et un regard critique sur la donnée *:* comment est-elle produite ? Par qui ? Pourquoi ? Quelles sont ses limites et quels services complémentaires peuvent être proposés pour y remédier ?

De même, la fonction de **« data designer »** va être amenée à se développer avec le développement de la visualisation. Il faut des compétences poussées en algorithmique, en design informationnel, en ergonomie et en graphisme pour proposer des services faisant sens. Chaque mode de représentation introduit des biais dont il faut connaitre l'impact cognitif pour proposer des visualisations pertinentes.

Il y a une réelle pénurie de compétences dans ces domaines pour les entreprises. Selon une étude du cabinet McKinsey de mai 2011, 1.5 millions de postes de techniciens développeurs et 190 000 postes d'analystes décideurs seront à pourvoir aux Etats-Unis pour couvrir les besoins

[51] *"Metadata is a Lovenote to the Future",* Voir les *« Tendances, silences et signaux faibles du salon OnLine Information 2011 »,* publiés par Michel Vajou sur le site du GFII :
http://www.gfii.fr/fr/document/tendances-silences-et-signaux-faibles-du-salon-online-information-2011

croissants de ce secteur en 2018[52]. En France, 20 000 à 30 000 postes seraient concernés.[53] Le profil **"data expert"** sera désormais aussi recherché que celui de « trader », et les salaires grimpent en conséquence. [54]

Les secteurs du conseil et de la formation connaîtront vraisemblablement une croissance importante pour pallier ce déficit. Pour les PME, qui ont l'agilité pour innover sur des services ciblés, il y a aura de nombreuses opportunités à saisir.

Le problème des données personnelles
La commission européenne, comme la CNIL, le répète : les « données personnelles » sont aujourd'hui le « carburant de l'innovation numérique » car elles permettent de développer des services d'information ciblés. Mais le développement des « Big Data » sera étroitement lié à **l'acceptabilité sociale** de ces technologies et à la résolution de points durs comme le **croisement de fichiers**. Les outils de mining et de cartographie permettent désormais d'identifier des personnes physiques ou morales en corrélant des données anonymes. Les usages devront être encadrés, en accord avec les attentes de la société en termes de *« privacy »*.

Soutenabilité des Big Data
La culture du risque et de la sécurité invite entreprises et institutions à tout sauvegarder. Or, la réflexion de l'INA montre que le développement des « Big Data » doit être corrélé à un projet de société. Les données explosent, mais l'impact cognitif et écologique de cette ultra-disponibilité des données n'est jamais mesuré. Il faut une réflexion systématique sur le **« No Data »**, comme pendant systématique aux « Big Data ». Des fonctionnalités d'oubli devront être développées dans les projets de conservation. De plus en plus, le métier d'archiviste sera la **« destruction intelligente » de données**. Le temps de filtrage, de tri et de destruction devra être budgété à l'avance. Les auteurs, éditeurs et producteurs de données auront peut-être un rôle à jouer en amont.

9.3. Questions / Echanges avec la salle

Comment se posent ces questions de convergence côté client ?
L'innovation dans les « Big Data » émane de nouveaux besoins des clients. Il y a deux angles d'attaque :

1. **Quantitatif :** les besoins utilisateurs restent les mêmes mais l'échelle et les volumes ont explosé. La réponse est essentiellement technologique : comment innover sur la vélocité, la puissance de calcul et les architectures pour répondre à la demande ?
2. **Qualitatif :** les utilisateurs pressentent qu'il y a un bénéficie certain à exploiter de nouveaux gisements de données (sur les réseaux sociaux, les plateformes d'e-commerce, …). La réponse est sur l'innovation de service : comment capturer, traiter et exploiter ces données pour leur donner du sens ?

Quel questionnement pour l'innovation côté éditeur de solutions ?
La chaîne des « Big Data » reste bien structurée en segments d'activité sur lesquels les acteurs peuvent se positionner : hébergement, calcul, collecte, traitement, enrichissement, analyse… Mais les niveaux de spécialisation et de technicité augmentent. Pour les éditeurs, le vrai défi reste la capacité d'innovation à partir des données : que faire ? L'innovation progresse par accident (*serendipity*). Il y a besoin non seulement d'investissements, mais aussi d'une volonté.

[52] http://www.mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation
[53] http://www.channelnews.fr/expertises/tendances/12841-big-data-des-milliers-demplois-en-perspective-aux-confins-de-la-technique-et-du-fonctionnel-.html
[54] http://www.informationweek.com/software/business-intelligence/big-data-talent-war-10-analytics-job-tre/232700311?pgno=1

Y a-t-il une spécificité des « Big Data » propre au milieu éditorial (VS corporate) ?

Les processus diffèrent moins dans les principes que sur les finalités. Un éditeur utilise des bases de données pour fabriquer de nouveaux produits, services d'informations et les commercialiser. Une entreprise utilise des bases de données à des fins de gestion interne. L'objectif n'est pas de produire des informations qualifiées à commercialiser, mais d'extraire des informations stratégiques à des fins décisionnelles. Les enjeux sont différents : pour une entreprise, la question de l'intégration des canaux d'échanges internes et externes est fondamentale (RSE, CRM, …), tout comme celle de la mise en place d'une gouvernance des données structurée par des référentiels métiers. Pour un éditeur, l'enjeu sera davantage dans la maîtrise de nouvelles fonctions liées à la collecte et au traitement des données de masse pour en sortir des produits d'information et des services qualifiés.

Avec l'Open Data, y a –il un métier en émergence : éditeur de données ?

Les étapes du processus restent assez proches du métier d'éditeur de contenus : collecte, traitement, diffusion. Certes, les technologies évoluent, les niveaux de spécialisation et de granularité augmentent avec l'édition de données, mais c'est aussi le cas pour l'édition de contenu. Là où le métier change fondamentalement, c'est sur la partie « intelligence de données ».