cigref
Réseau
de Grandes Entreprises

Enjeux *business* des données

Comment gérer les données de l'entreprise pour créer de la valeur ?

cigref
Réseau
de Grandes Entreprises

PROMOUVOIR LA CULTURE NUMERIQUE
COMME SOURCE D'INNOVATION ET DE PERFORMANCE

Enjeux *business* des données

Comment gérer les données de l'entreprise
pour créer de la valeur ?

Octobre 2014

LE MOT DU PILOTE

La donnée est devenue, avec le capital, les ressources humaines, les clients, les processus et le système d'information, l'un des principaux actifs de l'entreprise, quel que soit son secteur d'activité.

Mais au contraire des autres actifs, la donnée n'est pas gérée en tant que tel, de manière professionnelle et organisée tout au long de son cycle de vie dans l'entreprise. Ce constat est partagé par toutes les sociétés qui ont contribué au groupe de travail sur les enjeux *business* des données.

Ce constat est exacerbé par la transformation numérique qui est aussi la révolution de l'information, devenue plus « Volumineuse », plus « Variée » et plus « Véloce », pour reprendre les fameux « 3V » qui définissent le *Big Data*. Mais le *Big Data* est aussi associé à une série de termes qui évoquent des catastrophes : « tsunami », « explosion », « avalanche », « déluge », « infobésité », etc. qui illustrent bien le défi auquel toutes nos organisations sont confrontées : faire des données la mine d'or de l'ère numérique.

L'information ne constituera une nouvelle richesse pour l'entreprise que si on sait l'exploiter et la valoriser. Une donnée peut être vue comme un actif numérique dès lors qu'elle comporte un potentiel de création de valeur pour l'entreprise. La question qui se pose est donc bien de savoir comment on optimise l'exploitation de ses données pour créer de la valeur.

Ce qui est nouveau dans cette réflexion sur les données, c'est la démultiplication des potentiels de combinaison des données dans une perspective de création de valeur pour le *business*. Mais, on ne peut pas prédéfinir la valeur de l'information, on ne peut pas prévoir à l'avance l'usage que feront les utilisateurs des informations mises à leur disposition.

Ainsi, sur cette question de la valeur des données, on cherchera à identifier en quoi les données contribuent à la réalisation d'un des axes stratégiques de l'entreprise. La responsabilité de l'utilisateur portera alors sur la production et la mise à disposition de données de qualité.

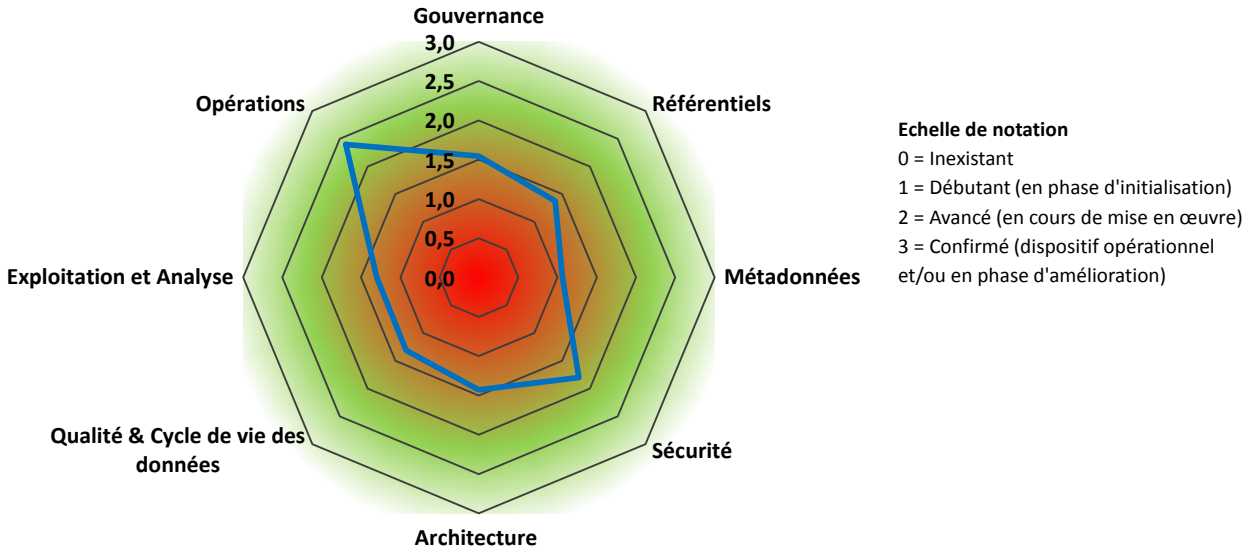
Il existe plusieurs manières pour une entreprise de tirer de la valeur de ses données, qu'elle soit financière ou pas. Dans tous les cas, il faudra s'appuyer sur une bonne maîtrise du cycle de vie de la donnée, et sur quelques prérequis : tout au long de la chaîne, l'information doit être collectée, contrôlée, protégée, partagée et analysée, pour contribuer à mieux décider ou à innover.

Tout cela concourt à la mise en place d'une véritable gestion des données (*Data Management*, *Enterprise Data/Information Management*), d'autant que le cadrage réglementaire et normatif

évolue et continuera d'évoluer sur ce sujet. Cette démarche d'entreprise implique la fonction SI autant que les Métiers, et devrait être pilotée en interne par une instance dédiée.

Le présent livrable propose une méthodologie de gestion des données, avec des exemples concrets de mise en œuvre et de bonnes pratiques, ainsi qu'une ouverture sur une démarche de valorisation des données. Est également associé à ce document un outil d'auto-évaluation de la maturité des entreprises en matière de gestion des données (document Excel. Les critères d'auto-évaluation sont disponibles en annexe pour information. A partir des résultats, présentés sous forme de radar, le lecteur peut facilement identifier ses zones de maturité et de progrès, et le cas échéant définir et mettre en place un plan d'amélioration de la gestion des données au sein de son organisation. A titre d'illustration, le groupe de travail a utilisé cet outil pour évaluer la maturité de leur entreprise sur le management des données. La figure ci-dessous agrège les résultats individuels : on voit bien que si les fondamentaux sont bien maîtrisés (opérations et sécurité), un long chemin reste à parcourir sur la qualité, l'architecture, les métadonnées, la gouvernance et la valorisation…

Maturité de la gestion des données
dans les entreprises du groupe de travail CIGREF



Avec ces deux livrables, le lecteur trouvera sans nul doute matière à valider son niveau d'avancement et à déterminer ses priorités d'action.

Bonne lecture !

Régis DELAYAT
DSI de SCOR et pilote du groupe de travail

Le CIGREF, réseau de Grandes Entreprises, a été créé en 1970. Il regroupe plus de cent très grandes entreprises et organismes français et européens de tous les secteurs d'activité (banque, assurance, énergie, distribution, industrie, services...). Le CIGREF a pour mission de promouvoir la culture numérique comme source d'innovation et de performance.

Titre du rapport : ENJEUX *BUSINESS* DES DONNEES

EQUIPE DU CIGREF

Jean-François PÉPIN – Délégué général
Sophie BOUTEILLER – Directrice de mission
Anne-Sophie BOISARD – Directrice de mission
Josette WATRINEL – Secrétaire de direction
Flora FISCHER – Assistante de mission
Frédéric LAU – Directeur de mission
Matthieu BOUTIN – Chargé de mission
Marie-Pierre LACROIX – Chef de projet
Josette LEMAN – Assistante de direction

REMERCIEMENTS :

Le groupe de travail « *Enjeux business des données* » a été piloté par **Régis DELAYAT, DSI de SCOR et Administrateur du CIGREF.** Nous remercions les personnes qui ont participé aux échanges et contribué à la rédaction du présent document :

Pierre-Etienne BARDIN – AXA GROUP SOLUTIONS
Michel BENARD – GPMT DES MOUSQUETAIRES
Dominique CUPPENS – RFF
Isaure DE CHATEAUNEUF – SAINT GOBAIN
Bruno DUCROS – GEODIS
Jean-Louis GHIGLIONE – RENAULT
Anne-Laure JACQUEMIN – MALAKOFF MEDERIC
Marc MEYER – MINISTERE DE L'ECONOMIE
Patrick MOLE – SCOR
François NGUYEN – SFR
Emmanuel OLIVIER – SAINT GOBAIN
Jean-Luc RAFFAELLI – LA POSTE
Marc RICHARD – HARMONIE MUTUELLES
Mickael ROYER – GIP MDS
Isabelle SIPMA – SCOR
Yao SOGADZI – FONDATION DE FRANCE
Carine TUAL – GALERIES LAFAYETTE
Florence WATHELET – RENAULT
Guy WOTCHOU – MALAKOFF MEDERIC

Ce rapport a été rédigé par Sophie BOUTEILLER (CIGREF), avec la contribution des membres.

POUR TOUT RENSEIGNEMENT CONCERNANT CE RAPPORT, VOUS POUVEZ CONTACTER LE CIGREF AUX COORDONNEES CI-DESSOUS :

CIGREF, Réseau de Grandes entreprises
21, avenue de Messine 75008 Paris
Tél. : + 33.1.56.59.70.00
Courriel : contact@cigref.fr

Sites internet :
http://www.cigref.fr/
http://www.fondation-cigref.org/
http://www.histoire-cigref.org/
http://www.questionner-le-numerique.org
http://www.entreprises-et-cultures-numeriques.org

Droit de propriété intellectuelle

Toutes les publications du CIGREF sont mises gratuitement à la disposition du plus grand nombre, mais restent protégées par les lois en vigueur sur la propriété intellectuelle.
Est autorisée la copie du titre et d'extraits de 500 caractères, suivis chacun de la mention « Source : » assortie de l'url de la publication CIGREF. Toute autre reprise doit faire l'objet d'une autorisation préalable auprès du CIGREF cigref@cigref.fr.

SOMMAIRE

Le mot du pilote 2

1. Introduction 3
   1.1. Donnée, information, connaissance, savoir : pas seulement une question de terminologie 4
   1.2. L'information est un actif stratégique qui doit être identifié et géré comme tel dans l'entreprise 5
   1.3. Contexte 7
   1.4. Approche retenue 9

2. Prérequis : mettre en place une gouvernance et une gestion des données 11
   2.1. Gouvernance des données 11
      2.1.1. À quoi sert la gouvernance des données ? 11
      2.1.2. Rôles et compétences à développer 14
      2.1.3. Standards auxquels se référer 15
   2.2. Référentiels et données clés 15
      2.2.1. Enjeux liés aux référentiels de données 15
      2.2.2. Standards auxquels se référer 18
   2.3. Métadonnées 19
      2.3.1. Données structurées ou non structurées : définitions 19
      2.3.2. Enjeux de la gestion des métadonnées 19
   2.4. Sécurité et conformité 22
      2.4.1. Protection des données: la classification 22
      2.4.2. Le cas des données personnelles 28
   2.5. Architecture 32
      2.5.1. Elaborer un modèle de données d'Entreprise 33
      2.5.2. Définir des règles de propagation des données dans le SI 36
      2.5.3. Standards auxquels se référer 36

3. Cycle de vie des données : collecte, traitement et contrôle de la qualité 38
   3.1. L'impératif pilotage du cycle de vie de la donnée 38
   3.2. La récupération et le traitement des données, en particulier non structurées 38
   3.3. L'obligation de qualité exacerbée 40

4. Créer de la valeur à partir des données 44
   4.1. Exploitation et analyse 44
   4.2. Valorisation 45
   4.3. L'*Open Data* 49
      4.3.1. Le cas des données de l'Administration 50
      4.3.2. L'ouverture des données au sein de l'entreprise 55

5. Conclusion 56

6. Annexe : Critères d'auto-évaluation de la maturité des entreprises en matière de gestion des données 57

1. INTRODUCTION

Dans quelque activité que ce soit, la donnée est devenue un actif clé de l'entreprise. Cela va sans dire dans la finance, où l'information est à la fois la matière première et le produit fini sur lesquels s'appuient les services proposés. Mais c'est aussi le cas dans l'industrie et les services, où l'information sous-tend également toute prise de décision, tout choix d'orientation stratégique. Il n'est pas besoin de convaincre : chacun reconnait aujourd'hui la valeur de la donnée pour les entreprises et les administrations qui s'organisent pour optimiser sa gestion et son usage. Le groupe de travail mis en place cette année par le CIGREF s'est attaché à matérialiser ces « Enjeux Business des données », à décrire les éléments qui contribuent à créer de la valeur, depuis la collecte jusqu'à la prise de décision, en passant par le contrôle de la qualité de la donnée, sa protection, son stockage, son partage entre les experts concernés, … tous ces maillons du cycle de vie de la donnée au sein de nos entreprises qui, mis bout à bout, constituent l'indispensable *« Data Management »* qui reste encore largement à définir et à mettre en place.

La révolution numérique ouvre une ère nouvelle, celle de l'information, dorénavant volumineuse et partagée au-delà des frontières mêmes de l'entreprise ou de l'administration. Les exigences de qualité, de transversalité et de protection apparaissent – ou pour le moins, sont exacerbées, pour des raisons d'efficacité opérationnelle, commerciale et/ou légale.

Comment mesurer cette valeur de la donnée pour le Business ? Les nombreuses entreprises membres rassemblées à l'occasion de ce groupe de travail ont essayé d'apporter une réponse à la question. L'ambition de ce livrable n'est pas de donner à chaque société ou chaque lecteur intéressé une liste d'instructions à suivre à la lettre, mais en fonction du niveau de maturité sur ces sujets, de permettre à chacun d'investiguer plutôt tel ou tel composant de ce puzzle de la gestion de la donnée pour optimiser son usage et sa valeur pour l'entreprise.

Le présent document essaie ainsi de balayer l'ensemble des éléments susceptibles de créer de la valeur, dont l'illustration ultime va pour certains jusqu'à la monétisation. Chacun y trouvera les éléments de réponse à ses préoccupations principales.

1.1. Donnée, information, connaissance, savoir : pas seulement une question de terminologie

Lorsque l'on parle du phénomène qui se développe autour de la donnée, largement stigmatisé par le « *Big Data*[1] », on le fait souvent en reprenant des notions de catastrophes : « tsunami », « avalanche », « explosion », « déluge », « infobésité », etc., plutôt qu'en évoquant la richesse, la mine d'or que cet afflux de données recèle.

Ces termes illustrent bien l'idée que cette information pléthorique générée par la révolution numérique et à notre disposition aujourd'hui est inutile si on ne sait pas l'utiliser, l'exploiter. Au contraire, pour qui réussit à optimiser son usage, la **donnée** devient **information**, puis, bien partagée au sein de l'entreprise, elle se transforme en **connaissance** et constitue son **savoir**. Elle peut être une source de services et d'innovations, notamment lorsqu'on la croise avec d'autres données et qu'elle provient de sources diverses.

Face à cette réalité, l'enjeu pour l'entreprise ou l'administration est de gérer cette chaîne de valeur des données autour desquelles se greffent des contraintes fortes de protection, de sécurité, de qualité et d'éthique.

La sémantique fait débat et peut être génératrice de flou… Nous pouvons considérer que la notion de « donnée » (ou *« data »*) est le fruit de la collecte, et est proche des couches techniques. Les données collectées peuvent être structurées ou non, internes ou externes, en provenance de sources institutionnelles (gouvernementales, *Open Data*[2], etc…), commerciales (marché, clients, etc.), techniques, financières ou diverses (web, réseaux sociaux, etc.). Elles peuvent prendre des formes aussi variées que des bases de données, du texte, des images et du son, des vidéos, etc.

---

[1] *Big Data* : La CSTIC recommande dans le Journal Officiel du 22 août 2014 de préférer l'appellation « mégadonnées » ou « données massives ». Le *Big Data* est une démarche qui consiste à extraire l'information pertinente d'un ensemble de données, cet ensemble de données se caractérisant par le Volume, la Variété, la Vélocité et la Véracité. Le volume à lui seul n'est pas une réelle problématique, étant déjà au cœur des préoccupations de nombreuses entreprises depuis longtemps. Seule la combinaison de deux ou plus de ces axes correspond à une démarche de *Big Data*. Ces « 4V » sont apparus avec l'évolution du web de ces dernières années : les réseaux sociaux, l'internet des objets, l'explosion des vidéos en ligne… Des quantités gigantesques de données (Volume) sont publiées en continu (Vélocité), dans des formats structurés ou non, provenant de sources diverses (Variété) et dont la légitimité n'est pas toujours vérifiée (Véracité) (source : *« Big Data : la vision des grandes entreprises »*, CIGREF – 2013)

[2] Voir plus loin le chapitre consacré à l'*Open Data.*

N'oublions pas non plus que les organisations sont elles-mêmes « productrices » de volumes de données toujours plus conséquents.

La qualification et la contextualisation de cette donnée créent « l'information », sur laquelle vont s'appuyer les décisions, choix, orientations et arbitrages de l'entreprise. Ces différentes actions passent par l'assurance que l'information est correcte, et l'élimination du « bruit », la validité temporelle (non-obsolescence), la corrélation, la mise en perspective constituent autant de défis, exacerbés par le volume. Autres problématiques fortes : la protection de ces données, alors que la cybersécurité et les réglementations récentes ou à venir sur la protection des données à caractère personnel sont devenues l'une des préoccupations majeures de l'entreprise, et l'utilisation proprement dite de données qui sont dispersées dans le système d'information et ses silos applicatifs. Les démarches de classification de l'information peuvent être un bon vecteur pour valoriser les informations et pour gérer les données à leur juste niveau, avec pragmatisme, comme nous le verrons plus loin.

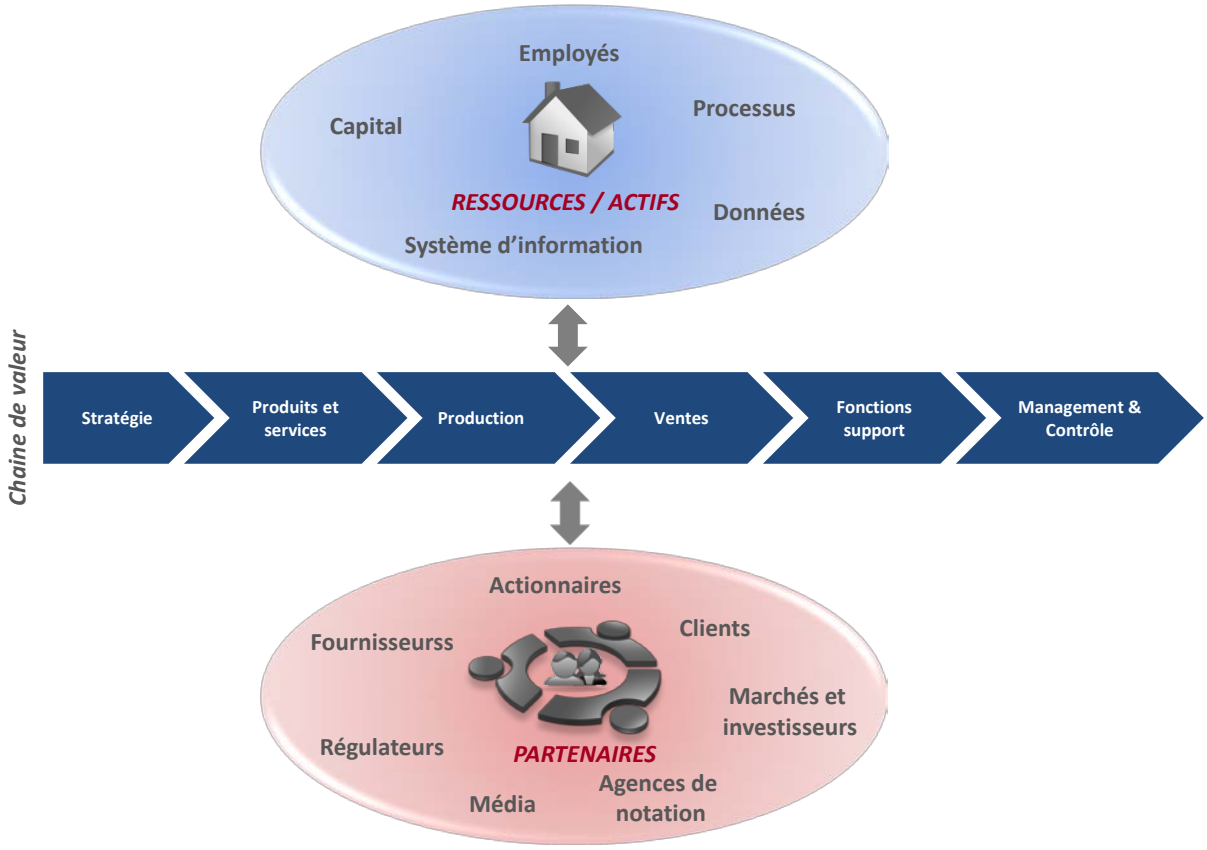
1.2. L'information est un actif stratégique qui doit être identifié et géré comme tel dans l'entreprise

Le groupe de travail CIGREF considère la donnée comme un des actifs majeurs de l'entreprise, au même titre que son capital, ses hommes et femmes, ses clients, ses processus et son système d'information comme le montre la figure 1 ci-après. Cette représentation illustre en effet la chaîne de valeur de toute entreprise, quelle qu'elle soit et quel que soit son domaine d'activité, ses ressources ou actifs, et enfin ses partenaires, de toute nature.

La donnée est donc l'un des principaux actifs immatériels de nos organisations, et pour autant, n'est pas encore gérée avec la même rigueur ni les mêmes moyens que les autres ressources, capital et ressources humaines notamment.

Dans un contexte où elle est devenue critique pour l'activité de l'entreprise, la mise en place d'une gestion structurée et industrielle de la donnée est impérative.

Figure 1 : L'information, un actif stratégique de l'entreprise



Source : Groupe de travail CIGREF, 2014

Il s'agit donc de :

- Traiter l'information comme un actif qui a une valeur quantifiable (le principal actif de Facebook est son information !)
- Réfléchir à la manière dont l'information trouve sa place dans la stratégie (l'information n'est pas un dérivé de l'activité de l'entreprise mais elle contribue à sa performance)
- Penser l'information globalement et penser l'infrastructure pour la gérer : définir un *Information Capabilities Framework*
- Démarrer avec des cas d'usages et en fonction des besoins Métiers (approches par « petits pas »), d'autant que les initiatives de *Big Data* peuvent être freinées par la complexité et le morcellement du paysage technologique, ainsi que par la fragmentation de l'information (favorisée par le *Cloud*).

1.3. Contexte

Le sujet de la valeur *business* des données pose de nombreuses questions contextuelles, complexes et qui dépendent beaucoup de la lecture que l'on en a et de l'importance que l'on y accorde à un moment donné.

Avec la transformation numérique, l'entreprise est résolument ouverte, poreuse et l'information est au cœur de cette transformation portée par le numérique. Selon Gartner, les volumes massifs d'informations complexes deviendront la norme dans l'avenir. La variété d'informations à traiter (structurées et non structurées, *Open Data*, données démographiques, statistiques, provenant des réseaux sociaux, tablettes et autres supports…), ainsi que la vélocité des données (*smart meters*[3], données GPS collectées en temps réel, …) et la consommation de ces données contribuent au développement de la masse des informations disponibles, et sont autant d'éléments à prendre en compte pour innover dans les métiers. Ainsi, ce développement des volumes d'informations disponibles est surtout lié à la révolution internet (avec l'Internet des Objets notamment), à la prédominance de l'information, et au *Big Data*[4].

Cependant, la prise de conscience des décideurs sur la nature stratégique des données est récente et progressive, et les investissements dans ce domaine restent encore limités. Beaucoup d'entreprises sont actuellement en phase de prototypage, de *Proof Of Concept*, mais très peu ont encore mis en place des projets de grande envergure autour de l'exploitation des données.

Pourtant, elles observent toutes que les pratiques et les habitudes de consommation des informations changent, tant chez leurs clients qu'en interne. Face à cette réalité, quelques entreprises ont compris que leur différenciation viendra de leur capacité à innover, et que les données ont un rôle majeur dans cette dynamique d'innovation. Car le *Big Data* permet certes de prendre de meilleures décisions, mais il est aussi (surtout ?!) une opportunité pour créer de nouvelles offres et de nouveaux produits.

---

[3] *Smart Meter* (compteur communicant) : compteur disposant de technologies avancées, dites *« AMR » (Automated Meter Reading)* qui identifient de manière détaillée et précise, et éventuellement en temps réel la consommation énergétique d'un foyer, d'un bâtiment ou d'une entreprise, et la transmettent, par téléphone (ou par courants porteurs en ligne (CPL)

[4] Voir le rapport CIGREF, paru en 2013 : « *Big Data*, la vision des grandes entreprises »

Pour ces entreprises précurseurs, adopter une approche d'investigation sur les données, à travers un processus d'expérimentation (R&D), est désormais une évidence. La démarche basée sur l'expérimentation permet de trouver de nouvelles relations entre les données :

- Comment lier les différentes sources de données ?
- Quelle valeur leur associer ?
- Retenir *vs* libérer les données ?

Ainsi, la capacité à capter des données qu'on ne savait pas saisir jusqu'à présent (par exemple avec la géolocalisation), puis la capacité à les croiser avec d'autres données (l'appétence d'un client pour un produit par exemple) favorise l'innovation commerciale et la personnalisation.

Parallèlement, la possibilité de traiter à faible prix et rapidement de grands volumes de données, disponibles dans des formats variables, élargit le champ des possibles technologiques. D'autant que l'acquisition de ces données peut se faire désormais n'importe où, et presque à l'insu des acteurs, ce qui fait émerger de nouveaux enjeux en matière de protection des libertés individuelles.

Le mouvement est donc en marche mais selon Gartner, 85% des plus grosses organisations (Fortune 500) ne sauront pas tirer un avantage compétitif avec le *Big Data* d'ici 2015. La majorité des grandes entreprises seront encore dans l'expérimentation, mais celles qui auront su anticiper avec le *Big Data* tireront un avantage compétitif de 25% ou plus sur leurs concurrents, issu de l'innovation : mobilité (accès constant à l'information), réseaux sociaux, *Cloud* (infrastructures) et information contextuelle (géolocalisation). L'apparition de nouvelles lignes de produits et de nouveaux services, favorisées par l'exploitation des données, crée des opportunités *business* nouvelles et font évoluer – de manière plus ou moins profonde – les modèles d'affaires des entreprises[5] :

- Dans le domaine de la relation client par exemple, l'internet des contextes et des personnes permet de personnaliser les produits et services, de localiser les clients *via* des capteurs RFID insérés dans leurs cartes de fidélité. Les entreprises personnalisent leur relation client par la contextualisation des interactions qu'elles entretiennent avec eux. Elles créent ainsi de nouvelles lignes de produits et de nouveaux services, et cherchent à fidéliser leur clientèle par le développement d'une expérience unique. Ceci leur permet à la fois d'accroitre la performance de l'acte de vente, mais aussi d'inscrire la relation dans la durée et de s'assurer un potentiel de ventes futures.

[5] Fondation CIGREF, 2012 : « Les Essentiels – Modèles d'affaires dans l'espace numérique »

- La personnalisation de la relation client est d'ailleurs la première motivation évoquée par les entreprises qui investissent dans le *Big Data*[6].
- Dans le secteur de la santé, le séquençage de l'ADN pour moins de 1500 dollars est désormais possible et permet d'identifier les pathologies dont un individu est susceptible de souffrir dans le futur. Ainsi, il offre la possibilité d'anticiper l'apparition d'une maladie par la mise en place éventuelle d'un traitement préventif par exemple.

Si ces quelques exemples soulèvent des questions éthiques d'importance, ils mettent aussi en évidence l'impact majeur que les informations peuvent avoir sur nos prises de décision personnelles et professionnelles.

Dans ce contexte, cet actif clé qu'est la donnée doit impérativement être géré de façon professionnelle et industrielle. Le groupe de travail du CIGREF a entrepris de **poser les bases d'une véritable gestion des données en menant une réflexion sur leur valeur pour l'entreprise, leur exploitation et leur protection.** Il a identifié une liste de thèmes relatifs à la donnée, contribuant à la valoriser et qui sont repris dans les différents chapitres du livrable, chacun illustré par des témoignages de sociétés qui ont participé aux travaux.

1.4. Approche retenue

Le groupe de travail a retenu une approche mixte, articulant interventions d'experts et contributions des participants.

Les enjeux *business* des données couvrent plusieurs dimensions structurantes, que le groupe de travail a regroupées en trois grandes parties dans le présent document.

Ainsi, dans la première partie, nous rappelons les **prérequis** nécessaires à la mise en place d'une démarche de valorisation des données. La dimension patrimoniale des données, qui constituent un des principaux actifs de l'entreprise, est clé. La mise en place d'une gouvernance des données est un moyen d'adresser cette dimension, en identifiant les responsables des données, et les implications de ces responsabilités. La gouvernance permet également d'organiser les actions autour de la donnée dans l'entreprise, afin de garantir sa qualité et sa sécurité tout au long de son cycle de vie. Les référentiels, les métadonnées, la sécurité (protection, classification et localisation des données, conformité réglementaire) et l'architecture sont ainsi les quatre piliers d'une bonne gestion des données.

La deuxième partie du rapport est consacrée au **cycle de vie de la donnée**, qui couvre toutes les activités de traitement (acquisition, transformation, archivage, …) de la donnée, depuis sa

[6] Interview de Matthias Fille (Conseiller TIC à la CCI Paris Île-de-France) sur le *Big Data* et le 4ème Forum Europe qui s'est tenu à Paris le 13 mai 2014 : http://www.entreprises.cci-paris-idf.fr/web/international/4eme-forum-europe-profitez-des-opportunites-offertes-par-le-big-data-le-13-mai-2014

collecte jusqu'à son usage. Cette partie pose la question de l'**accessibilité** des données, notamment des données non structurées, et de la **qualité** des données internes et externes utilisées dans les processus métiers. L'*Open Data* est au cœur de la problématique de valorisation des données : comment capter, traiter et valoriser les données ouvertes, tant en interne qu'en externe ? Nous pouvons associer à cette question celle de la représentation des données, au travers notamment des démarches de modélisation et de gestion des *métadata*.

La troisième partie enfin ouvre sur la **valorisation des données**, objectif ultime de notre réflexion : comment faire du patrimoine informationnel de l'entreprise un outil éthique de « conquête commerciale » ? Comment développer une valeur d'usage au-delà de la valeur patrimoniale ? Dans notre approche, inspirée des travaux d'expérimentation du Pr Bounfour, Rapporteur général du programme ISD de la Fondation CIGREF, l'éthique associée aux démarches de valorisation est apparue comme un atout fondamental dans l'obtention d'un avantage concurrentiel pour une entreprise. Nous avons d'ailleurs identifié quelques bonnes pratiques dans l'utilisation des données externes à l'entreprise, notamment les données personnelles.

2. PREREQUIS : METTRE EN PLACE UNE GOUVERNANCE ET UNE GESTION DES DONNEES

2.1. Gouvernance des données

2.1.1. À quoi sert la gouvernance des données ?

La gestion des données (ou *data management*) consiste à mettre en œuvre l'ensemble des dispositifs relatifs aux informations utilisées au sein de nos organisations afin d'en optimiser l'usage. La valeur de l'actif « information » passe ainsi par l'assurance d'une compréhension commune et partagée, par l'identification et la représentation des circuits internes, par le déploiement de moyens adaptés d'accès, de contrôle de qualité et de traitement, par l'application de règles de protection en ligne avec la sensibilité, etc…

La gouvernance des données en est le composant qui décrit les responsabilités, fixe les règles et contrôle leur application. Cette démarche de gouvernance est portée par une instance dédiée qui la pilote, avec des règles, des guides, des référentiels, des indicateurs, des chartes, une politique (de gestion des données personnelles, de classification, de stockage, de conservation, de brevets, de propriété intellectuelle, …).

L'enjeu principal d'un dispositif de gouvernance des données consiste à déterminer qui sont les propriétaires (*Data Owner* ou *Data Steward*), au sein des métiers, de chaque famille de données. On parle de *« Data Owner »* ou de propriétaire de donnée pour désigner celui/celle qui de par son expertise métier peut définir une donnée, le niveau de qualité requis, les processus d'acquisition de la donnée et les processus de contrôle. Le *« Data Steward »*[7] ou gestionnaire de donnée fait vivre les éléments de documentation sur les définitions, processus, contrôles, etc. Il apporte un savoir-faire méthodologique et peut s'assurer de la cohérence globale de la documentation sur un ensemble de données, en

[7]Définition du *Business Data Steward* (DAMA – DM BOK) : travailleur du savoir et manager, le *Business Data Steward* est reconnu comme un expert de la donnée et se voit attribuée la responsabilité de spécifier les données (commerciales, bases de données, …) et d'en assurer la qualité. A ce titre, il : :

1. Participe aux réunions de comité de pilotage sur la gestion des données. .
2. Identifie et définit les besoins Métiers en matière d'information.
3. Propose, modélise, precise et affine les informations (noms, définitions, …), modèles de données et autres attributs relatifs aux différents Métiers de l'entreprise.
4. Garantit la validité et la pertinence des modèles de données associés aux différents Métiers.
5. Définit les règles Métier à respecter pour les attributs de données et vérifie que les exigencies sur la qualité des données est respectée.
6. Valorise les données de référence et les significations associées.
7. Contrôle la qualité des données de test et participe à la conception des jeux de données de test.
8. Identifie et aide à la resolution de tout problème sur les données.
9. Participe à l'analyse de la qualité des données et à leur amélioration.
10. Contribue à l'élaboration des politiques de données, des normes et procédures.

coordination avec les architectes et maitres d'ouvrage du système d'information. Il joue vis-à-vis du *Data Owner* le même rôle que le Maître d'Ouvrage ou le *Business Analyst* vis-à-vis du *Process Owner*. Ces acteurs doivent savoir en quoi consiste leur rôle : il ne s'agit pas seulement d'utiliser les données comme on en a l'habitude dans ses propres processus, mais bien de les entretenir comme un patrimoine, les documenter, en contrôler la qualité, les mettre à disposition, … pour au final en accroitre la valeur pour l'entreprise. La dimension « conduite du changement » sur ce sujet est clé compte-tenu du caractère très transversal de l'approche : cela nécessite du temps, un travail important de coopération et des acteurs capables de montrer des résultats concrets sur un périmètre restreint, afin de répliquer et diffuser les pratiques.

De plus, qui dit gouverner, dit orienter, fixer le cap et suivre la mise en œuvre adaptée. Le dispositif de gouvernance des données doit donc organiser l'ensemble des décisions relatives aux données, dans une recherche de :

- Valorisation du patrimoine,
- Maximisation des usages de ces données,
- Optimisation de la qualité (la juste qualité pour le juste besoin).

Dans ce dispositif de gouvernance, la protection des données est une question majeure : un juste équilibre est à trouver entre un accès totalement ouvert à l'information et la nécessité de protéger les données (notamment pour les données personnelles) en en restreignant l'accès. Il faut par ailleurs trouver le juste équilibre entre ce qui doit être défini de manière *« top/down »* et être encadré, et ce qui peut relever d'initiatives locales.

Un bon moyen de répondre à cet enjeu est de définir les données clés afin de concentrer le travail de la gouvernance sur ces données essentielles. Pour cela, il peut être pertinent de nommer un *« Chief Data Officer »* (soit un « Responsable des données d'Entreprise » ou encore un « Responsable du patrimoine informationnel » - ou équivalent), positionné au bon niveau dans la hiérarchie de l'entreprise, et d'organiser des instances de gouvernance associant l'ensemble des propriétaires de données et les responsables des différentes fonctions de gestion des données au sein de la DSI.

Dans ce cadre de gouvernance, le rôle de la Fonction SI est de proposer aux acteurs en interne un ensemble le plus large possible de services de gestion des données, et d'assurer directement les fonctions relevant de son périmètre d'activité (le stockage, la sécurité, la modélisation, le développement, …).

Ainsi, la question des contours de l'offre de services, et des priorités de mise en œuvre est clé. La fonction SI est partenaire du Métier dans le dispositif de gouvernance : la nomination d'un responsable du *Data Management* au sein de la DSI peut faciliter la relation avec le

Métier et permettre une coordination des différents acteurs de la DSI impliqués dans la gestion des données.

Témoignage SCOR

Le Groupe a défini des politiques en matière de gestion des données, à des fins d'amélioration de l'efficacité opérationnelle, mais aussi de réponse aux demandes réglementaires, notamment de Solvency 2[8]*, cadre qui régit les sociétés d'Assurance :*

- *Une politique de «* Data Quality *», qui définit notamment les instances de gouvernance, et les rôles et responsabilités des utilisateurs autour des données ;*
- *Des compléments au travers de différentes «* Guidelines *» :* Change Management, Audit Trail, Monitoring *et* Reporting *;*
- *Des politiques définies par la DSI sur l'archivage et la rétention des données d'une part, et la sécurité des données d'autre part ;*
- *La gouvernance comprend également un comité de pilotage de la « protection des données » associant les différentes composantes métiers du Groupe ;*
- *L'identification et le contrôle des données clés véhiculées dans le « modèle interne », application qui évalue les risques et les besoins en capital associés, à partir de paramètres d'activité en provenance des applications* front *et* back office *;*
- *La mise en place d'un* Data Management *au sein de la DSI ;*
- *Un vaste programme de protection des données.*

Ces différentes politiques s'inscrivent dans un cadre global comprenant un code de conduite, une politique générale de contrôle interne, et une politique informatique intégrant une dimension forte sur la sécurité.

Ces politiques seront enrichies dans un futur proche pour y intégrer les différentes dimensions du Data Management *(architecture des données, gestion des métadonnées, …). Au-delà du rôle clé du propriétaire des données (déjà défini dans les politiques relatives à la qualité des données), l'ambition du Groupe est de pouvoir s'appuyer sur des correspondants relais dans les différents métiers.*

[8] La nouvelle règlementation Solvency 2 définit la façon dont doit être déterminé le niveau des fonds propres afin de couvrir les risques pris en charge par les compagnies d'assurance. Cette règlementation s'appliquera à compter du 1er Janvier 2016, et suppose des évolutions importantes des systèmes d'information.

2.1.2. Rôles et compétences à développer

L'enjeu sur les nouvelles compétences dépasse la seule création de nouveaux postes tels que *Data Scientist* ou *Data Steward*, et nécessite d'adapter les organisations et les processus pour que la transformation puisse s'opérer. L'expertise requise pour un usage performant des données est très diverse, à la croisée du métier, des statistiques, des mathématiques, de l'économétrie et de l'informatique. Le défi consistera également à développer plus de transversalité entre les différents silos de données en place dans les entreprises. Il est donc nécessaire de créer des synergies entre les entités afin de permettre une meilleure exploitation du contexte métier.

L'organisation pertinente doit donc permettre un accès riche et varié aux données de l'entreprise ainsi qu'aux données extérieures lorsque cela est possible. Une part importante de l'analyse de la donnée repose sur l'expertise et l'intelligence des décideurs, au service de leur vision stratégique.

Le changement d'organisation peut simplement se matérialiser par la mise en place d'une gestion d'ensemble des données au sein de l'entreprise, mobilisant de fait de nouvelles compétences, au-delà des compétences plus traditionnelles dont dispose notamment la DSI (par exemple, les *DBA - Database Analytics*, les responsables sécurité, les architectes …).

L'un des enjeux est de faire émerger côté métier des **responsables de données (ou *« Data Steward »)*** susceptibles sur un périmètre métier sur lequel ils détiennent une expertise reconnue, de spécifier les exigences sur les données et d'en contrôler la qualité. Ces responsables de données peuvent être positionnés à différents niveaux dans l'organisation, et peuvent être pilotés par des coordinateurs au niveau d'un métier, d'une fonction support ou d'une géographie. Ils peuvent être assistés d'analystes métiers ou de maitres d'ouvrage à l'instar de ce qui se fait la plupart du temps pour la description et la transformation des processus métiers. Faire émerger ce rôle est indispensable pour pouvoir mettre en place une gestion de données dans l'entreprise : les acteurs désignés doivent être à la fois reconnus pour leur expertise métier, mais aussi capable de sortir du cadre organisationnel dans lequel ils évoluent pour développer une vision globale et transversale des données qu'ils manipulent.

Une autre fonction clé nouvelle autour de l'exploitation des données est apparue : le ***« Data Scientist »* (ou expert en analyse de données)**. Il s'agit de disposer de compétences de haut niveau en matière d'analyse de données, en combinant à la fois les méthodes statistiques, mais aussi d'autres connaissances telles que la linguistique, la sémantique, utiles notamment pour travailler sur des données non structurées, sans oublier la bonne compréhension du métier sur lequel on travaille, et de mettre en œuvre une démarche d'analyse itérative, en acceptant de tester des hypothèses sans *a priori* sur le résultat recherché. Il faut donc à la

fois des compétences techniques (informatiques) de manipulation des données, des compétences en analyse de données plutôt dans l'univers mathématique, et des compétences en marketing et gestion (analyse des comportements clients, compréhension des processus métiers, etc.). Les universités et grandes écoles s'attachent aujourd'hui à développer cette alchimie de compétences qui peut aussi s'acquérir sur le terrain, au sein des entreprises, en permettant aux potentiels identifiés d'évoluer.

Côté DSI, en complément de profils plus traditionnels tels que le spécialiste du BI, des Datawarehouse, ou des Référentiels qui restent clés pour la gestion des données, de nouveaux profils émergent pour traiter les sujets de **gestion des métadonnées (*« Metadata Specialist »*)**, ou travailler sur les **données non structurées (*« Chief Knowledge Officer »*),** tandis que **des métiers d'architectes spécialistes des données ou de l'intégration** se développent.

2.1.3. Standards auxquels se référer

Le référentiel *DM-BOK*[9] décrit les rôles clés associés à la gestion des données. Il consacre également un chapitre spécifique à la gouvernance des données, qui est au cœur du modèle d'activité proposé, et comprend 2 familles d'activités principales :

- Le planning : stratégie, politiques, définition des rôles, nominations, définition des projets dédiés à la gestion des données, analyse économique des avantages et des coûts associés à cette gestion ;
- Le contrôle qui, au-delà du contrôle interne, comprend aussi tous les aspects relatifs à la conformité.

Par ailleurs, le référentiel *COBIT5*[10] s'est enrichi d'un chapitre dédié à la gestion des données *(COBIT 5 for information)*.

2.2. Référentiels et données clés

2.2.1. Enjeux liés aux référentiels de données

La gestion de référentiels d'entreprise est souvent une première étape dans la mise en place d'un dispositif de gestion des données. Les référentiels contiennent des données relativement stables, permanentes, auxquelles de nombreux processus peuvent faire référence. Il s'agit à la fois de nomenclatures ou de tables de référence (des codes devises,

[9] Plus d'information sur le DM-BOK : http://www.dama.org/i4a/pages/Index.cfm?pageid=3364

[10] Le référentiel COBIT5 est consacré à la gouvernance SI : il décrit cinq principes et définit les sept inducteurs de base qui le composent. Plus d'informations : http://www.isaca.org/COBIT/Pages/default.aspx?cid=1003566&Appeal=PR

des codes pays, des numéros de comptes …), ou de données de gestion telles que clients, produits, etc. Les référentiels sont les données les plus éligibles à être partagées largement, et il y a donc beaucoup de valeur à investir d'une part dans la documentation de ces données, et d'autre part dans les services qui permettent de les rendre accessibles, de garantir leur qualité et leur adéquation.

La question qui se pose est le plus souvent celle de la frontière entre ce qui doit être géré sous la forme de référentiels d'entreprise, et donc de façon globale et transversale, et ce qui peut rester géré plus localement (au niveau d'un métier, d'une géographie, d'une entité organisationnelle, etc.). La répartition entre les données de référence qui servent de base aux processus opérationnels et celles qui servent de base aux processus décisionnels est un sujet qui a tendance à s'estomper car tous les processus convergent vers des rythmes très contraints, proches du temps réel (et, cette convergence peut être d'ailleurs une bonne cible pour ce que l'on construit aujourd'hui).

Dans ce contexte, l'approche *« Master Data Model »* (*MDM)* est une bonne grille d'analyse des problématiques liées aux référentiels : la séparation entre la notion de « point d'acquisition » (liste des données « brutes » telles qu'elles sont collectées « à la source » au travers des processus métiers) et celle de « point de vérité » (données transformées et contrôlées, de telle sorte qu'elles puissent être diffusées largement dans l'entreprise, avec l'assurance d'un bon niveau de qualité) permet de construire des architectures de référentiels qui tiennent compte de la spécificité des modes de collecte des données d'une part (souvent complètement « embarqués » dans les processus métiers), et des exigences de qualité des données et des contrôles associés d'autre part. Le *MDM* propose d'ailleurs plusieurs architectures possibles selon la nature des référentiels, la nature des contraintes liées à la collecte, la nature des usages, et les contraintes liées à l'existant applicatif (la migration d'un référentiel existant est une tâche longue et complexe). L'architecture peut être soit complètement centralisée (le point d'acquisition et le point de vérité sont assurés par un seul et même système), soit pour partie décentralisée (notamment pour le point d'acquisition), ce qui suppose la mise en place de mécanismes de synchronisation plus ou moins sophistiqués (uni ou bi-directionnel, temps réel ou *batch*) selon les exigences de niveau de service.

Témoignages SCOR et GIP MDS

SCOR

Les déploiements d'un outil de back-office *mondial pour le traitement des contrats de réassurance, puis d'un système comptable mondial se sont accompagnés de la mise en place de référentiels unifiés et partagés. À titre d'exemple, l'identifiant du contrat est géré en maitre par le système* back-office *qui fournit cet identifiant aux systèmes* front-office

qui l'enregistrent. Les référentiels sont donc globaux, et ils sont aussi partagés entre les différents métiers.

De plus, la mise en conformité avec Solvency 2 d'une part, et la poursuite de l'unification des systèmes front-office *d'autre part, conduisent à pousser un cran plus loin le périmètre et l'usage de référentiels partagés, en intégrant les systèmes de gestion des risques et le décisionnel (les* datawarehouses *et les outils de* reporting*).*

L'approche « Data Quality *» nécessaire à Solvency 2 conduit aussi à préciser les responsabilités relatives à la qualité des données de référence, en faisant appel aux concepts MDM de point d'acquisition / point de vérité. Enfin, les services d'accès aux données de référence doivent aussi s'enrichir.*

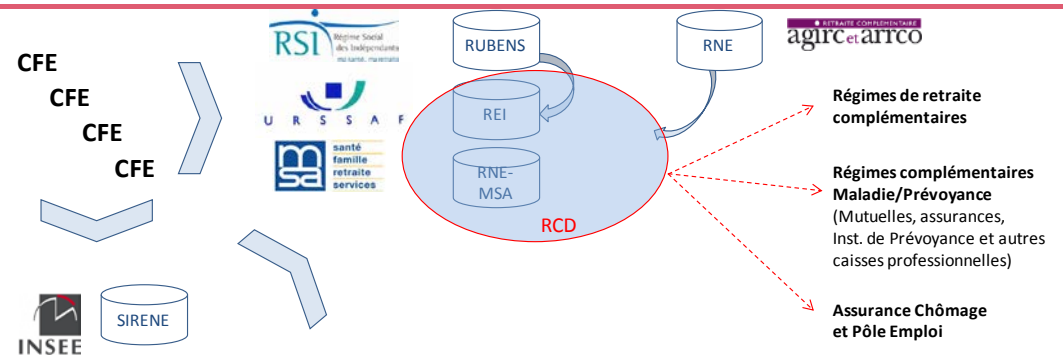
GIP MDS

Dans le secteur public, le schéma stratégique du système d'information de la sécurité sociale et de la protection sociale (SSSI), définit parmi ses objectifs stratégiques la nécessité « d'acter le principe d'un référentiel-socle des employeurs pour la sphère sociale ».

À l'image des données individus, l'identification des employeurs constitue un besoin commun aux acteurs de la sphère sociale. La fiabilité, l'unicité et la mise à disposition de cette donnée sont les ingrédients nécessaires à la mise en œuvre de nombreux processus (déclarations sociales pour différents cas – retraite, maladie, chômage …, recouvrement des cotisations) ainsi qu'au maintien de leur cohérence transverse.

*Cet objectif se traduit par la mise en œuvre d'une architecture de consolidation des données employeurs autour des référentiels socles (REI/ACOSS – **R**éférentiel **E**ntreprises et **I**ndividus et RNE/MSA – **R**éférentiel **N**ational des **E**ntreprises) qui sont « unis » dans le cadre du répertoire RCD – **R**épertoire **C**ommun de **D**éclarants.*

- *L'identification des employeurs pour la sécurité sociale est portée par deux référentiels socles : les employeurs agricoles dans le RNE et les employeurs du régime général dans le REI. Ces référentiels sont articulés avec les informations en provenance des différents Centres des Formalités Entreprises (CFE) et de la base SIRENE de l'INSEE qui consolide les informations.*
- *Pour les besoins liés à la mise en œuvre de la DSN (Déclaration Sociale Nominative), qui est une procédure unique, un répertoire des déclarants (RCD) adossé à ces référentiels a été mis en œuvre, qui enregistre leurs caractéristiques et fait le pont entre la sécurité sociale et les organismes complémentaires.*



*Cette démarche peut être rapprochée d'une démarche de gestion de données de référence (*Master Data Management *ou* MDM*). Les processus mis en œuvre visent à définir, stocker, et diffuser des données référentielles au sein du système d'information de la sphère sociale.*

2.2.2. Standards auxquels se référer

Le *DM-BOK*[11] contient un chapitre dédié aux référentiels et au *MDM*. Il distingue :

- Les *« Reference Data »*, qui correspondent à ce que l'on appelle parfois les nomenclatures (comme par exemple les codes et noms des pays, des devises …). Il s'agit de données plutôt statiques, permettant de classer ou de qualifier les données de gestion.
- Les *« Master Data »* (que nous appelons généralement « données de référence » ou « référentiels »), qui sont des données de gestion pour lesquelles il est nécessaire de disposer d'un enregistrement de référence (ou *« golden copy »* ou « point de vérité »), comme par exemple la description d'un client, d'un produit, ou d'un contrat.

Par ailleurs, le *DM-BOK* donne des exemples assez concrets de solutions et d'architectures à mettre en place.

Citons également l'ouvrage de Franck Regnier-Pecastaing, Michel Gabassi et Jacques Finet, intitulé « MDM : enjeux et méthodes de la gestion des données ».

[11] Plus d'information sur le DM-BOK : http://www.dama.org/i4a/pages/Index.cfm?pageid=3364

2.3. Métadonnées

Ce sujet couvre tout le périmètre des données de l'entreprise, structurées et non structurées.

2.3.1. Données structurées ou non structurées : définitions

Déjà en 2007, le CIGREF affirmait aux côtés de l'APIL et de l'APROGED, dans un Livre Blanc conjoint intitulé « Valorisation de l'information non-structurée », que le traitement de l'information était un défi pour les entreprises, et plus spécifiquement pour les DSI[12], notamment du fait de l'explosion du volume d'informations disponibles. Tout le monde en est aujourd'hui convaincu. Dans ce même Livre Blanc, les 3 associations considèrent que *« L'information non-structurée, et partant, les documents qui la contiennent, possède un cycle de vie intrinsèque, intimement lié aux processus métiers et à la vie de l'entreprise. L'optimisation de ce cycle de vie informationnel est alors un levier majeur d'optimisation des performances de l'entreprise dès lors qu'une approche méthodologique (…) est mise en place »*.

On peut considérer qu'une information est structurée si elle est utilisable directement par un ordinateur pour effectuer un calcul : opérations arithmétiques, comparaison, … S'agissant de l'information non-structurée (textes, images, sons, …), l'enjeu est de la rendre exploitable et d'en tirer de la valeur ajoutée pour tous les Métiers de l'entreprise.

2.3.2. Enjeux de la gestion des métadonnées

Les métadonnées sont des informations descriptives sur les données. La gestion des métadonnées est donc une activité centrale dans la gestion des données car elle contribue à la valorisation des données en en facilitant les usages. La valeur ajoutée de cette fonction est liée au temps qu'elle fait gagner dans la recherche de la bonne donnée pour le bon usage, et dans la capacité à faire gagner du temps aux projets dans la recherche de la compréhension et de la définition des données.

Pour les données structurées, gérer des métadonnées suppose de pouvoir s'appuyer au préalable sur une démarche de modélisation des données : en effet, les métadonnées sont attachées aux données modélisées, quel que soit le niveau de modélisation (global de niveau Entreprise, ou bien au niveau d'un métier voire d'une application). Des métadonnées peuvent aussi être associées à une représentation des flux de données.

[12] APIL-APROGED-CIGREF, 2007 : *« Valorisation de l'information non-structurée »*

Quand il s'agit de données non structurées, les métadonnées peuvent être attachées soit à chaque élément de contenu (un document, une page web, une vidéo, etc.) ou bien à des ensembles plus vastes (un dossier comprenant plusieurs documents, un site web, etc.). Quand elles sont associées à des données non structurées, les métadonnées facilitent la recherche de l'information en proposant un plan de classement.

Les questions clés à traiter lorsque l'on parle de métadonnées sont les suivantes :

- Quelles sont les métadonnées principales à gérer ?
  - Le premier niveau de métadonnées est le dictionnaire, c'est-à-dire la définition de la donnée :
    - Pour les données structurées, cette définition doit être attachée à un niveau de description conceptuel, et gérée au niveau Entreprise en fonction de l'importance de la donnée et de son niveau de partage ;
    - Pour les données non structurées, plusieurs niveaux de définition peuvent être gérés (le titre d'un document, d'un film ou de tout bloc de contenu est un premier niveau de définition, qui peut être enrichi de descriptions plus longues comme par exemple un résumé).
  - Le deuxième niveau regroupe les informations nécessaires à la gouvernance de la donnée, et notamment le nom du responsable ou du propriétaire (*« owner »* ou *« steward »*) de la donnée ;
  - Le troisième niveau permet de qualifier le niveau de confidentialité : s'agit-il d'une donnée hautement confidentielle qui ne doit pas sortir de l'entreprise et ne doit être connue que d'un nombre restreint de personnes (informations sur les résultats avant une publication par exemple) ? S'agit-il d'une donnée dont l'usage doit être strictement interne ? D'une donnée publique ? D'une donnée personnelle, soumise à des contraintes spécifiques (CNIL, …) ?
  - Il est souvent utile de classer les données par type ou par catégorie : classement des données structurées en fonction des types de processus métiers auxquelles elles sont liées (données de référence, données opérationnelles, données décisionnelles) ou classement de documents par nature (présentation/compte-rendu), en fonction du niveau de maturité (document de travail/version intermédiaire/version finale), par thème ;
  - Il est important d'opérer à partir d'une description de niveau Entreprise, une géolocalisation des données dans le système d'information, en s'appuyant notamment sur un métamodèle d'architecture : processus, applications, unités organisationnelles qui mettent à jour la donnée et la consomment, localisation physique des données, etc.

- À quel niveau de description des données gérer les différentes métadonnées (niveau du modèle / niveau des flux, niveau modèle d'entreprise / niveau modèle applicatif, niveau document / niveau dossier, etc.)
  - Des métadonnées sont à associer aux différents niveaux de description des données, mais pour les données structurées clés de l'entreprise, la plupart des métadonnées sont à attacher au niveau conceptuel Entreprise (notamment les définitions et la gouvernance) ; de même que pour les données non structurées, une gestion par ensemble d'éléments de contenus est plus simple à administrer qu'une gestion élément par élément (par exemple au niveau d'un document) ;
  - Il est essentiel de gérer les liens entre les différents niveaux, et de mettre en place des logiques d'héritage entre le niveau le plus haut et les niveaux bas (niveau applicatif notamment), afin de simplifier une gestion qui peut s'avérer très lourde eu égard aux volumes concernés.
- Quelle organisation pour gérer les métadonnées ?
  - Une équipe de gestion, proche des équipes de modélisation et de gestion des connaissances doit apporter du support, fournir des outils et des processus ;
  - La description et les évolutions doivent être autant que possible décentralisées auprès des propriétaires des données, outillées, et faire partie des processus de gestion des changements du SI.
- Comment piloter l'équilibre effort / rétribution ?
  - Faire porter l'effort prioritairement sur des ensembles de données dont le besoin de partage est fort (données de référence par exemple) ;
  - Gérer les métadonnées par ensembles de données, et non *a priori* au niveau de chaque donnée élémentaire ;
  - Tirer parti des projets qui sont des opportunités de mettre en place des métadonnées à moindre coût car celles-ci sont partie intégrante du projet ;
  - Outiller pour limiter l'effort ;
  - Communiquer beaucoup pour promouvoir les usages.
- Quels outils pour gérer les métadonnées ?
  - La plupart des outils de gestion de documents comprennent leur propre gestion de métadonnées, ainsi que des outils de recherche s'appuyant sur les métadonnées gérées ;
  - Pour les données structurées, les outils qui permettent de gérer un référentiel d'architecture permettent en général de gérer un métamodèle et des métadonnées.

Témoignage SCOR

- *Une utilisation extensive d'un logiciel de collaboration pour la gestion de*

documents, et un développement de la GED (Gestion Electronique de Documents), ont conduit à mettre en place au cas par cas des métadonnées facilitant la recherche de documents :

- *Le réseau social d'entreprise déployé conduit à mettre en place des plans de classement de documents sur une échelle plus large ;*
- *Des sites « chapeau » dans le logiciel de collaboration permettent d'élargir les périmètres de gestion et de partage de documents, et proposent une structure plus développée en matière de métadonnées.*

- *Une approche combinant un dictionnaire et d'autres métadonnées a été engagée sur les applications du périmètre Solvency 2 ;*
- *La définition des modalités d'articulation du niveau Entreprise et des descriptions de niveau applicatif est en cours d'analyse dans le cadre de l'initiative « Data Management ».*

2.4. Sécurité et conformité

Rappelons au sujet de la sécurité que le cadrage de l'utilisation de la donnée peut devenir rapidement complexe au regard du respect des droits en vigueur. En effet, il est utile de préciser que les possibilités d'exposition des données sont reliées bien souvent à un cadre « contractuel » qui définit les conditions applicables aux référentiels concernés. Dans la volonté chaque jour grandissante de vouloir partager les données, un travail d'analyse est nécessaire pour connaitre précisément l'origine et les obligations intrinsèques aux données. Les chapitres suivants traiteront de la classification des données, ou comment garantir la protection de certaines données, mais le groupe de travail rappelle que certaines données non confidentielles sont pourtant soumises à des contraintes règlementaires et/ou juridiques. C'est le cas par exemple lorsque des données de localisation ont été acquises auprès des d'organismes certifiants. Ces derniers dressent alors un cadre d'utilisation stricte à l'entreprise, ce qui limite le partage aux seules entités d'une organisation.

2.4.1. Protection des données: la classification

Dès 2008, le CIGREF reconnaissait l'importance de la protection des données et des informations[13], et définissait ainsi la notion de protection : *« La protection de l'information est une démarche consciente visant à protéger, au sein de l'entreprise étendue, ce qui vaut la peine d'être protégé, tant au niveau des données que des supports d'information. Cette démarche implique un système de gestion, une identification des informations sensibles, une analyse de risques, des acteurs, avec des rôles et responsabilités et un programme de réduction des risques. »*

[13] CIGREF, 2008 : *« Protection de l'information : enjeux, gouvernance, et bonnes pratiques »*

Cette protection s'avère d'autant plus nécessaire du fait de l'évolution du contexte qu'ont connu les entreprises depuis 2008 : développement rapide du *Cloud* avec les questions que posent l'externalisation accrue des données et leur localisation par exemple[14], ou encore évolution du contexte règlementaire comme le projet de règlement européen sur les données personnelles qui prévoit des obligations de protection accrues pour les entreprises et un régime de sanctions très lourdes en cas de défaut de protection des données.

La classification des données selon leur niveau de confidentialité est à la base de la protection de l'information. Cependant, quand on aborde des démarches de mise en commun et de ré-utilisation des données, cela ouvre une dimension de risque supplémentaire : des données bien protégées dans une application sécurisée et dont les droits d'accès sont maîtrisés, peuvent se trouver exposées de façon non contrôlée si elles sont mises à disposition d'utilisateurs non avertis, soit par transfert à des applications aval, soit par déversement dans un entrepôt de données collectif. Une réponse peut être d'instituer l'autorisation obligatoire du *Data Owner* pour toute ré-utilisation d'une donnée classifiée.

Une analyse fine des risques sera nécessaire. Par exemple, dans l'étude de mise en place d'un outil de *Business Intelligence*, il faudra distinguer :

- Des cas où l'information élémentaire elle-même est confidentielle ; c'est le cas en particulier pour les données à caractère personnel.
- Des cas où l'information élémentaire n'a pas de valeur en elle-même, mais où la capacité d'agrégation de l'ensemble des données doit être protégée.
- Des cas où l'agrégation des données n'a de sens que si elle est faite selon les règles spécifiques détenues par l'expert (par exemple pour des analyses faites sur des données économiques), et dans ce cas c'est le rapport d'analyse final qui porte la confidentialité.

La classification des données constitue la pierre angulaire de la démarche de protection de la donnée puisqu'elle permet d'identifier les données qui doivent faire l'objet d'une protection et le degré de protection requis. La classification est habituellement utilisée pour gérer la confidentialité des données, à l'instar de ce qui se fait dans le monde militaire (on parle d'une classification « secret défense »). Étymologiquement, le verbe « classifier » peut être utilisé simplement pour classer des informations dans des catégories. En transposant cette approche à l'entreprise, il est possible d'établir plusieurs types de classifications : certaines pour gérer les aspects de confidentialité, d'autres pour gérer les aspects d'ordonnancement comme, par exemple, pour classer les données dans des familles, auxquelles seront

[14] AFAI-CIGREF-IFACI, 2013 : « *Cloud* Computing et Protection des données »

appliqués des modes de gestion communs. La question clé est de déterminer quel est l'usage de ces classifications. Ainsi, pour une classification sur la confidentialité, chaque acteur doit savoir exactement comment se comporter en fonction du niveau de confidentialité attaché à la donnée.

La proposition de loi de 2012 sur le secret des affaires (Sénat français) met en exergue ce besoin de classification des données[15] : *« Constituent des informations protégées relevant du secret des affaires d'une entreprise, […] les […] données […] qui ont, en conséquence, fait* ***l'objet de mesures de protection spécifiques destinées à informer de leur caractère confidentiel*** *et à garantir celui-ci. »*

De même, l'évolution récente de la jurisprudence[16], où dans plusieurs affaires la notion de vol de biens immatériels tend à prendre forme, est une raison supplémentaire pour les entreprises de mettre en place un système de classification des données.

Bien qu'il n'existe pas de système de classification universel des données, l'utilisation d'objectifs de sécurité proposée par les démarches d'analyse de risques[17] (confidentialité, intégrité, disponibilité et traçabilité) apparaît comme un incontournable. La classification des informations, en vue de la protection de celles-ci, nécessite un dialogue entre la DSI et les Métiers :

- Les Métiers, pour l'évaluation des impacts d'atteinte à la sécurité des données, et pour la sensibilisation des opérationnels ;
- La DSI pour l'évaluation des risques informatiques, et pour la proposition de solutions de sécurisation des données.

Témoignages Groupement des Mousquetaires et Renault

Groupement des Mousquetaires

L'approche de classification des données a été initiée à l'occasion d'un projet, suite au besoin exprimé par les métiers de protection de la confidentialité de données jugées particulièrement sensibles.

La démarche adoptée a consisté à évaluer, de manière classique, les impacts pour le Groupement des Mousquetaires, d'une atteinte aux objectifs de sécurité des données : confidentialité et intégrité (les critères de disponibilité et de preuve ne sont pas intégrés à

[15] Lire la proposition de loi : http://www.assemblee-nationale.fr/13/propositions/pion3985.asp

[16] Affaire « Rose » : http://www.institut-ie.fr/bsa/BSA_00_03_2012.pdf

[17] E-bios par exemple

ce jour). L'échelle d'impact, indicative, est celle qui avait été définie pour la gestion des risques métiers du Groupement des Mousquetaires liés à l'informatique. Le périmètre de classification concerne à ce jour les informations gérées par les applications.

Tableau 1 : Échelle d'impact des risques selon 4 perspectives

Niveau d'impact	Impact financier		Impact adhérent du Gpt des Mousquetaires	Impact consommateur	Impact collaborateur
	Résultat net (variation négative N *vs* N-1)	Perte de trésorerie (variation négative N *vs* N-1)			
Critique	Supérieure à 100%	Supérieure à 100%	Propre à remettre en cause le contrat entre l'adhérent et le Groupement	L'image de l'enseigne est en danger	Détérioration durable du climat social, démotivation durable des collaborateurs
Fort	Comprise entre 50 et 100%	Comprise entre 50 et 100%	Service fortement affecté	Évasion durable de la clientèle	Forte démotivation des collaborateurs, détérioration nette du climat social
Moyen	Comprise entre 10% et 50%	Comprise entre 10% et 50%	Service modérément affecté	Un groupe important est affecté	Détérioration des conditions de travail, le climat social s'en trouve tendu
Faible	Inférieure à 10%	Inférieure à 10%	Service peu affecté	Quelques clients sont affectés	Les conditions de travail et le climat social sont peu affectés

Dans un second temps, si au moins un impact dépasse un seuil critique, une analyse de risques, adaptée de la méthode E-BIOS[18], est menée par le RSSI qui préconise un plan de réduction des risques informatiques dans un dossier de sécurité. La décision finale de l'adoption partielle ou totale du plan de réduction des risques revient aux métiers.

La démarche de classification des données gérées par les applications informatiques est en voie de systématisation, elle sera portée par la méthode interne de gestion des projets informatiques. Le résultat de la classification des données sera intégré dans le dictionnaire des données d'entreprise.

Renault

[18] Créée en 1995 par l'ANSSI et régulièrement mise à jour, la méthode EBIOS (Expression des Besoins et Identification des Objectifs de Sécurité) permet d'apprécier et de traiter les risques relatifs à la sécurité des systèmes d'information (SSI).

Depuis 2010, une politique de maîtrise de l'information a été mise en place et déployée progressivement dans toutes les entités du Groupe. Cette démarche s'appuie sur une classification des informations sensibles :

- *La classe de confidentialité détermine les règles d'accès aux informations sensibles, ainsi que les moyens de protection à mettre en œuvre.*
- *La classification des informations est déterminée à partir des enjeux et des risques qu'entraînerait une divulgation non contrôlée des informations sensibles.*
- *La classification doit être sélective et établie au juste nécessaire, à partir de l'analyse des risques, pour éviter de créer des complexités de gestion et des freins inutiles à l'accès aux informations.*

La grille de classification est établie en cohérence avec celle de Nissan, selon une correspondance :

Tableau 2 : Grille de classification

Classification	Interne	Sensible	Critique	Stratégique
Enjeux en cas de fuite	Risque faible, mais une communication non autorisée à l'extérieur représenterait une divulgation gratuite du savoir-faire, ou pourrait perturber la communication externe de l'entreprise	Risque de perte réelle, mais limitée	Risque de perte significative pour l'entreprise, par désavantage commercial, crise d'image, mise en cause judiciaire, ou autre.	Risque de perte ou de crise majeure, de nature à impacter le résultat financier ou la valorisation boursière.

Chaque métier a opéré la classification des données / informations selon cette grille et produit un catalogue des données confidentielles. Les niveaux de classification sont marqués sur les documents, et déterminent les moyens de protection à utiliser.

La classification a ensuite été étendue aux SI, à partir de l'analyse du niveau de classification des données traitées, et en prenant en compte les règlementations sur les données à caractère personnel. La PSSI définit les règles de sécurisations en fonction de la catégorie des données traitées dans chaque SI.

Pour les nouveaux projets, la méthode de projet intègre un passage en comité SSI qui émet un avis de PSSI.

Pour les SI existants, un inventaire des traitements informatiques selon la classification des données traitées est tenu à jour. A partir de cette classification sont ensuite analysées les chaines de confidentialité et les études de sécurisation des SI

2.4.2. Le cas des données personnelles

Les démarches de valorisation des données risquent de rapidement heurter de front deux des principes majeurs défendus par la CNIL :

- L'interdiction d'exploiter les données pour des finalités autres que celles qui ont été déclarées à l'origine,
- Les restrictions particulièrement sévères vis-à-vis des croisements de fichiers.

A partir de là, quand les traitements de données ne peuvent pas être anonymes, ce qui est évidemment le cas quand il s'agit de proposer de nouveaux services au client ou d'améliorer la performance du processus de vente, il faudra veiller à ce que ces traitements s'inscrivent dans le cadre de finalités déclarées de façon assez large, et des consentements demandés aux personnes concernées. Dans les autres cas, on conclura rapidement au besoin d'anonymiser les données.

Par ailleurs, en matière de sécurité, l'anonymisation de données personnelles est un processus clé qui transforme, de manière irréversible, des données concernant des personnes afin d'obtenir un nouvel ensemble de données à partir desquelles il n'est pas possible, par des « moyens raisonnables », de réidentifier ces personnes[19].

Le G29[20] (ensemble des « CNIL » de l'Union européenne) décrit la pseudonymisation des données personnelles comme un processus de transformation de données personnelles, visant à cacher les éléments personnels des données, mais susceptible de permettre néanmoins la réidentification des personnes.

Ces techniques peuvent répondre à différents besoins pour les entreprises :

- La conformité règlementaire,
- La protection du patrimoine de données, lorsqu'il s'agit de faire traiter ses données hors de l'entreprise (développement ou recette externalisés par exemple),
- L'exploitation de données dans un cadre initialement non prévu lors de la collecte des données (échange de données avec des partenaires, etc.).

À titre d'exemple, l'exigence d'anonymisation des données est particulièrement forte dans le domaine de la santé, où le *corpus* règlementaire relève tant de la loi « Informatique et Libertés », que du code de la santé publique ou du code de la sécurité sociale, et où le dialogue doit être instauré avec différentes autorités administratives :

- CNIL,

[19] Adapté de la directive européenne Directive 95/46/EC, Récital 26.

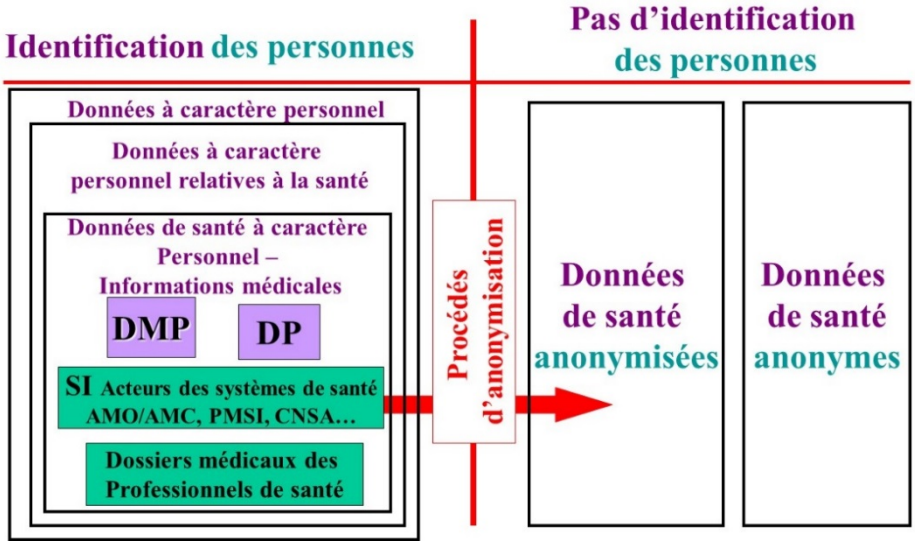
[20] Opinion 05/2014 on Anonymisation Techniques

- Comité consultatif sur le traitement de l'information en matière de recherche dans le domaine de la santé (CCTIRS),
- Institut des données de santé (IDS),
- Agence nationale des systèmes d'informations partagées de santé (ASIP),
- Etc.

Comme le souligne le G29, une anonymisation correcte des données, i.e. une anonymisation qui empêche toute réidentification, est difficile à obtenir et doit être évaluée selon trois axes :

- Quelle est la possibilité de réidentifier un individu à partir d'un jeu de données ?
- Quelle est la possibilité d'inférer des informations sur un individu à partir d'un jeu de données ?
- Quelle est la possibilité de relier des jeux de données différents appartenant à un même individu ?

Figure 2 : Anonymisation des données, cas des données de santé



Le G29 estime également que les risques de réidentification doivent être régulièrement évalués, et les dispositifs d'anonymisation revus si cela est nécessaire.

Avec le développement du *Big Data*, l'entreprise met en place des traitements qui collectent et mettent en corrélation des données de plus en plus volumineuses et de plus en plus granulaires. La valeur créée pour l'entreprise est potentiellement très élevée. Ces croisements doivent pouvoir être réalisés tout en apportant aux personnes concernées la garantie de respect de leur vie privée grâce à l'utilisation de techniques d'anonymisation efficaces.

Le sujet des techniques d'anonymisation est en développement actif, comme l'atteste par exemple le projet d'investissement d'avenir en cours xdata[21], qui explore ce domaine avec l'INRIA et l'équipe Privatics. C'est également un sujet « pointu » pour lequel il existe peu de guides pratiques pour les entreprises[22]. La maturité dans le domaine semble encore faible, et les pratiques défaillantes d'anonymisation font encourir un risque pour les entreprises. On citera ici le cas de Netflix[23] qui avait diffusé des données clients anonymisées dans le cadre d'un concours afin de trouver des algorithmes de recommandation : un expert ayant démontré une anonymisation inadéquate des données des clients de Netflix, la société interrompit son concours et dut conclure un accord pour éviter un procès[24].

Aussi, il apparaît important, tant pour la protection des données que pour leur exploitation, que les entreprises puissent trouver, en relation avec les autorités administratives (CNIL,….) et les laboratoires de recherche, un dispositif permettant de professionnaliser les procédés d'anonymisation des données. À ce titre, l'exemple d'un service d'anonymisation comme le service britannique *UK Anonymization Network*[25] *(UK AN)* pourra être étudié avec attention : financé par l'*ICO* (CNIL britannique) et coordonné par des universités, il permet d'assister les entreprises dans leur initiative d'anonymisation des données.

Sur le plan des systèmes d'information, les exigences de protection de la donnée par anonymisation doivent mobiliser l'ensemble des acteurs des projets informatiques :

- Les Métiers, pour la connaissance des exigences de sécurité des données ;
- La DSI, pour la prise en compte :
  - De ces exigences très en amont dans les démarches projet, idéalement traitées dans le cadre général d'une méthode d'analyse des risques,
  - Des méthodes d'anonymisation et de leurs impacts en termes de modèles de bases de données ou de performances,
  - Des contraintes techniques et légales sur les moyens de stockage de ces données, et sur la sécurisation des canaux de communication.

---

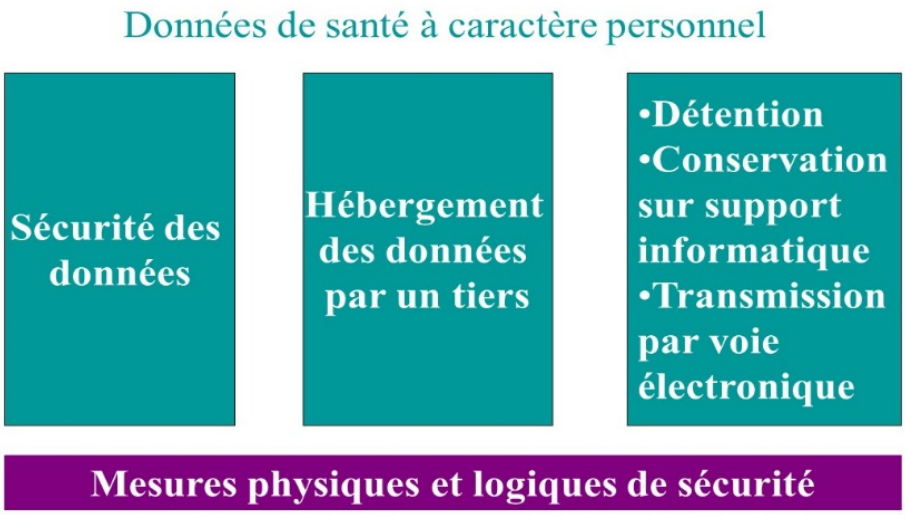
[21] http://xdata.fr/ : participation de l'équipe Privatics

[22] À notre connaissance, seule le CNIL anglaise - l'ICO - a publié un guide sur ce sujet : http://ico.org.uk/for_organisations/data_protection/topic_guides/anonymisation

[23] http://blog.netflix.com/2010/03/this-is-neil-hunt-chief-product-officer.html

[24] http://www.forbes.com/sites/firewall/2010/03/12/netflix-settles-privacy-suit-cancels-netflix-prize-two-sequel/

[25] http://ukanon.net/



Témoignages Groupement des Mousquetaires et Harmonie

Groupement des Mousquetaires

La question de l'anonymisation des données s'est posée avec acuité dans le cadre d'un projet collaboratif « Big Data » pour lequel il a été nécessaire d'échanger un volume important de données personnelles avec les partenaires du projet : industriels et laboratoires de recherche. L'anonymisation s'est avérée indispensable avant tout échange, notamment pour des raisons de conformité règlementaire puisque les partenaires n'étaient pas à l'origine identifiés comme destinataires des données personnelles.

Ce cas d'usage a été l'occasion d'étoffer le portefeuille d'offres du service chargé des recettes d'intégration des systèmes informatiques : ce service a entrepris d'acquérir la connaissance des techniques d'anonymisation, de développer une démarche d'anonymisation afin de conseiller les projets, et de sélectionner un outil propre à automatiser l'anonymisation des données.

Parallèlement, la solution de gestion du patrimoine applicatif a été étendue afin de recenser, en se basant sur le référentiel de la CNIL, les catégories de données personnelles gérées par chacune des applications afin de préparer un plan d'anonymisation des données personnelles pour les environnements informatiques hors production : développement, recette, … Enfin, un jalon sur l'anonymisation des données sera intégré dans la méthode interne de gestion des projets informatiques.

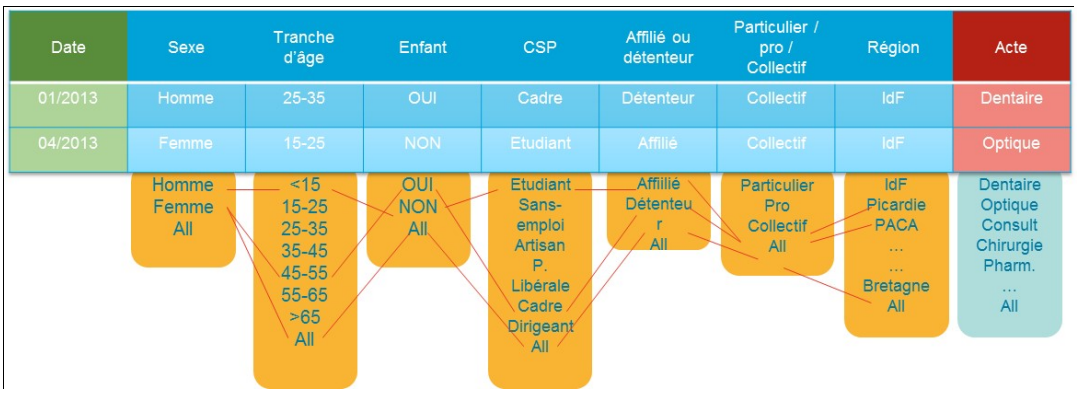
Harmonie

Depuis novembre 2013, Harmonie s'est lancé dans l'expérimentation « Big Data santé » avec un partenaire et le parrainage du laboratoire TNP – Supelec « Big Data dans

l'assurance santé ».

La démarche adoptée pour ces expérimentations a été déclinée comme suit :

Les données étant traitées à l'extérieur de l'entreprise, la question de l'anonymisation s'est bien évidement posée. Une étude prédictive de ce type a nécessité la fourniture d'un volume considérable de données (3 millions de personnes et 190 millions de lignes de consommation médicale). Pour respecter la règlementation tout en conservant un détail suffisant sur les informations, nous avons sélectionné, après extraction, les données adéquates, puis effectué une anonymisation afin de s'assurer qu'une personne ne puisse être réidentifiée (suppression des informations signalétiques et renumérotation aléatoire).



2.5. Architecture

Les méthodes de conception de SI ont longtemps donné la priorité aux processus et aux traitements plutôt qu'aux données, souvent vues comme des sous-produits des traitements

(notions d'entrées/sorties). L'introduction de méthodologies comme RACINE ou MERISE a mis en avant la nécessité de modéliser les données indépendamment des traitements, et selon 3 niveaux d'abstraction : le niveau conceptuel, le niveau logique, le niveau physique.

Aujourd'hui, l'importance des données, leur valeur, et par ailleurs la difficulté à prévoir les traitements qui varient en fonction des comportements des acteurs (un client qui va choisir un canal d'interaction, un employé qui va choisir son *device* …) font de la modélisation des données l'élément central d'une démarche de conception de SI. Comme évoqué précédemment, la modélisation est un prérequis à d'autres activités de gestion des données, telles que la gestion des métadonnées ou la gestion des référentiels. C'est donc une brique essentielle de tout dispositif de gestion des données. L'architecture des données est désormais considérée comme une discipline à part entière par toutes les approches d'Architecture d'Entreprise (TOGAF considère l'Architecture des données comme un élément clé de tout le travail d'Architecture du SI). Parler d'architecture des données, c'est utiliser les techniques de modélisation pour mettre en œuvre plusieurs niveaux de représentations des données dont principalement un niveau « entreprise », ceci afin de promouvoir le partage des données clés de l'entreprise, et de développer de nouveaux usages permettant de créer de la valeur.

Deux sujets apparaissent donc comme prioritaires pour maximiser la valorisation des données au travers d'une démarche d'architecture : l'élaboration d'un modèle de données d'entreprise et la définition de règles de propagation des données dans le SI.

2.5.1. Elaborer un modèle de données d'Entreprise[26]

Concernant le modèle de données, l'enjeu principal d'un travail sur l'architecture des données est de se doter de la capacité à ré-utiliser et partager plus facilement les données, associée à une plus grande efficacité dans l'assurance de leur qualité. En effet, si les données sont contrôlées à la source, et si la qualité n'est pas altérée dans les processus de transmission de celles-ci, alors la qualité peut être plus facilement garantie dans les processus qui se situent au bout de la chaine de traitement, et notamment les processus décisionnels. Définir un modèle de données d'entreprise permet un partage plus aisé des définitions de données, mais aussi de leurs inter-relations : les mêmes données peuvent

[26] Qu'est-ce qu'un modèle de données d'Entreprise ? Il s'agit d'une définition et d'une représentation des principaux objets métiers de l'entreprise (le client, le produit, le contrat, l'employé etc.), de la façon d'identifier ces données, de leurs inter-relations (liaison 1 à 1, 1 à n …). Le modèle de données d'Entreprise peut être formalisé en utilisant des techniques de modélisation habituellement appliquées à la conception de SI (MCD Merise, UML …). Il présente l'avantage de donner une vision d'ensemble des données les plus largement partagées et qui constituent le socle du système d'information, notamment les données de référence.

ainsi être partagées entre les processus opérationnels qui les créent, et les processus décisionnels qui les utilisent pour le pilotage.

Les questions qui se posent lorsqu'on engage un travail de définition d'un modèle de données d'entreprise portent sur les priorités à allouer aux travaux de documentation de l'existant et de définition de cibles en matière de données, en fonction d'une part des projets, mais aussi d'autre part de l'ampleur des enjeux de réutilisation et de partage (cela dépend donc de la nature des données concernées). La sélection de données peu nombreuses, mais particulièrement structurantes et/ou largement utilisées est un facteur de réussite.

Les périmètres à adresser de manière prioritaire peuvent être :

- Les référentiels ;
- Les données décisionnelles ;
- Les données opérationnelles qui constituent le socle du SI et peuvent faire l'objet d'échanges externes (contrats et opérations).

… car il s'agit des périmètres les plus transversaux, et donc porteurs des principaux enjeux de partage et de qualité.

Ces mêmes périmètres devront être ajustés également en fonction des enjeux de transformation du SI (portefeuille de projets).

Le modèle de données d'Entreprise doit en tout état de cause s'enrichir au fil des projets, et pouvoir être relié à la façon dont les données sont gérées concrètement par les applications informatiques au travers d'une démarche de type « géolocalisation des données » (les données sont où ? traitées par quels processus ? quelles applications ? etc.).

Témoignage SCOR

- *Cartographie et localisation des données pour les fonctions cœur métier : souscription et actuariat Vie/Non vie, Investissements*
    - *Permettant de définir les usages de la donnée (services), d'identifier sa valeur et les enjeux de qualité avec le métier ;*
    - *Avec une identification des cycles de vie et des propriétaires métiers ;*
    - *En tenant compte d'un contexte évolutif du fait de transformation numérique (développement itératif d'une capacité d'analyse de données provenant de différentes sources internes et externes).*
- *Élaboration d'un modèle de données d'Entreprise, le GMDM[27] : formalisation des concepts métiers clés et du glossaire, identification des données clés pour les*

[27] Global Master Data Model

métiers)

- *Initiative conjointe métier et IT comprenant la définition et la représentation en UML des données les plus utilisées*
  - *En s'appuyant sur les modèles du marché;*
  - *En gérant des versions du modèle ;*
  - *En utilisant des attributs communs à toutes les fonctions : statut lié au positionnement de la donnée dans son cycle de vie par exemple, définition du niveau de sensibilité (privé ou non), donnée externe ou interne.*
- *Recensement des flux fonctionnels et des applications utilisant les données référencées dans le modèle ;*
- *Définition d'indicateurs sur les niveaux de représentation (connu, normalisé, informatisé, contrôlé, ..).*

L'outil utilisé est « Power Designer *» qui est utilisé également pour la modélisation de données de niveau applicatif (modèle CDM[28], LDM[29], PDM[30]) et la DSI développe plus largement un référentiel d'architecture (intégrant non seulement les données, mais aussi les fonctions et les applications) pour répondre aux différents besoins métier. Les modèles sont publiés sur l'intranet, « versionnés » et archivés via le référentiel.*

[28] *Conceptual Data Model*
[29] *Logical Data Model*
[30] *Physical Data Model*

2.5.2. Définir des règles de propagation des données dans le SI

Les processus mis en œuvre au sein d'une entreprise sont par définition basés sur des échanges d'informations entre des acteurs. Pour que ces échanges soient efficaces, il faut que l'information échangée soit bien comprise, que son niveau de qualité puisse être appréhendé, que toutes les informations nécessaires aux différentes activités réalisées soit disponibles rapidement. Parmi les éléments essentiels à l'efficacité des échanges d'information figurent la mise à disposition des éléments nécessaires à une compréhension réciproque (le langage, et si celui-ci n'est pas acquis *a priori*, il faut des traductions, ou des glossaires permettant de disposer de définitions), et aussi l'identification du propriétaire de l'information, celui qui va pouvoir répondre à des questions si besoin, et qui garantit la qualité de l'information puisqu'il en est la source. Au niveau informatique, les mêmes principes doivent s'appliquer pour garantir la fluidité des échanges entre applications :

- L'application propriétaire d'une donnée dans le SI doit être clairement identifiée (tout particulièrement quand il s'agit d'une donnée de niveau Entreprise, et donc largement partagée/utilisée),
- Elle doit mettre à disposition les données qu'elle détient dans des formats et en utilisant un langage exploitables par les différents consommateurs des données.

Le besoin d'une modularité de l'architecture, et d'une rapidité de plus en plus grande dans la mise en œuvre de nouveaux processus impose de plus en plus que chaque composant logiciel propriétaire de données mette en œuvre les services nécessaires à l'accès aux données qu'il détient d'une façon modulaire, avec des formats variés, et en respectant des règles strictes de découplage inter-applicatifs. Les architectures de type SOA illustrent bien cette façon de procéder.

Au final quels que soient les choix d'architecture opérés, il est essentiel de mettre en place des règles de propagation des données au sein du SI pour d'une part garantir l'intégrité des échanges de données, leur rapidité de mise en œuvre, et ainsi faciliter l'accès aux données de l'entreprise en interne, mais aussi vis-à-vis de clients de plus en plus connectés.

2.5.3. Standards auxquels se référer

Le *DM-BOK*[31] définit l'ensemble des activités relatives au domaine *« Data Architecture ».*

TOGAF[32] considère que l'architecture du Système d'information comprend 2 sous-ensembles : l'architecture des applications et l'architecture des données. En matière d'architecture des données, *TOGAF* définit un certain nombre de principes :

[31] *Data Management Book Of knowledge*

- Principe 10 : *Data is an Asset* (*Data is an asset that has value to the enterprise and is managed accordingly*);
- Principe 11 : *Data is Shared* (*Users have access to the data necessary to perform their duties; therefore, data is shared across enterprise functions and organizations*);
- Principe 12 : *Data is Accessible* (*Data is accessible for users to perform their functions*);
- Principe 13 : *Data Trustee* (*Each data element has a trustee accountable for data quality*);
- Principe 14 : *Common Vocabulary and Data Definitions* (*Data is defined consistently throughout the enterprise, and the definitions are understandable and available to all users*);
- Principe 15 : ***Data Security*** *(Data is protected from unauthorized use and disclosure. In addition to the traditional aspects of national security classification, this includes, but is not limited to, protection of pre-decisional, sensitive, source selection-sensitive, and proprietary information).*

[32] *The Open Group Architecture Framework*

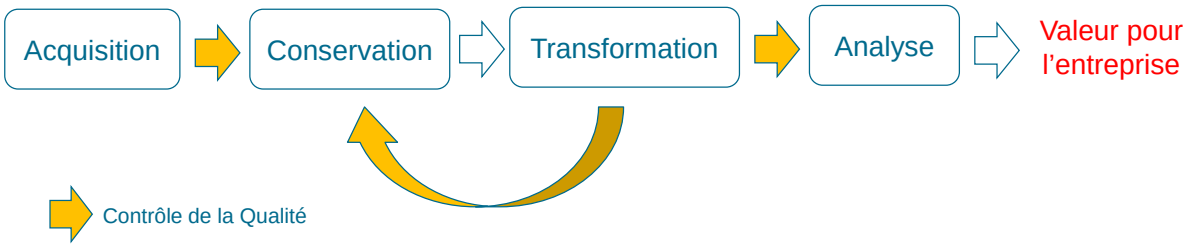
3. CYCLE DE VIE DES DONNEES : COLLECTE, TRAITEMENT ET CONTROLE DE LA QUALITE

3.1. L'impératif pilotage du cycle de vie de la donnée

L'augmentation des volumes, l'explosion du non-structuré, la variété des typologies et des supports, et l'enjeu que représente l'usage des données pour les entreprises imposent une maîtrise de tout le cycle de vie de l'information, depuis son acquisition jusqu'aux décisions ou innovations qu'elle supporte. L'évolution est telle que les entreprises sont de plus en plus en situation de dépendance par rapport à leurs données, créant de fait de nouvelles exigences d'accès, de fiabilité, de sécurité et de capacité d'analyse.

A l'état brut, la donnée n'a pas de valeur. Elle doit être validée, transformée (et de nouveau contrôlée après chaque transformation), rapprochée, analysée, notamment pour constituer une connaissance, un savoir, et apporter de la valeur pour l'entreprise.

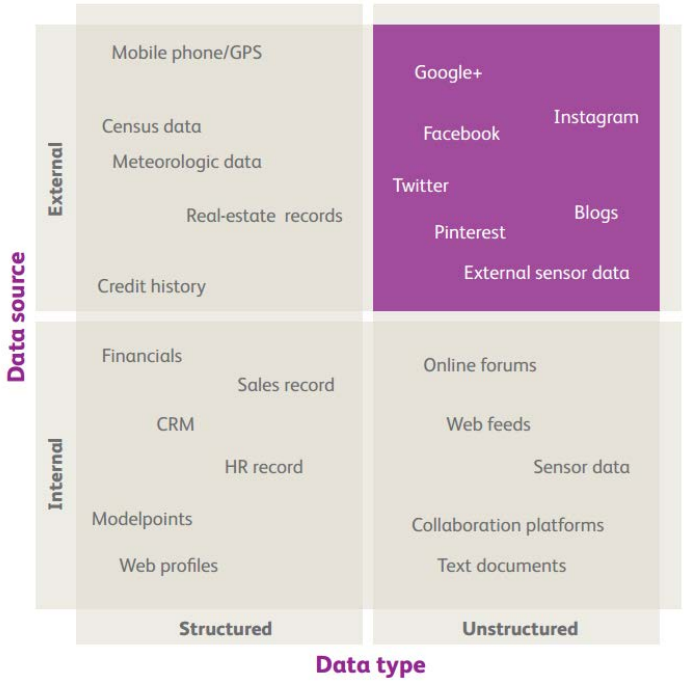
Les entreprises doivent donc envisager une approche « globale » pour leurs données, en garantissant leur qualité, leur protection et leur accessibilité à chaque étape de leur cycle de vie :



Source : Groupe de travail CIGREF, 2014

3.2. La récupération et le traitement des données, en particulier non structurées

La question de la collecte des données pose celle de la récupération et du traitement des données non structurées (données publiques donc mais aussi échanges de données entre partenaires), et en particulier celles contenues dans les réseaux sociaux, blogs, … comme on peut le voir sur le schéma ci-dessous.



Source : BearingPoint Institute, Report Issue 004

Ces données non structurées, échangées de plus en plus largement en dehors de l'entreprise, sont l'illustration du développement de l'intelligence collective et de la connaissance dans le monde numérique. Pour l'entreprise, ces données non-structurées peuvent être une nouvelle source de création de valeur, si tant est qu'elle soit en capacité de la capter et de l'exploiter.

Ainsi, la collecte de données externes à l'entreprise est de plus en plus nécessaire au *business* : les entreprises doivent se mettre en position de traiter ces données en s'adaptant à des contextes très variés. En effet, parfois les données sont non structurées, parfois elles sont structurées mais selon des formats qui ne sont pas toujours exploitables simplement, ou encore, elles sont hétérogènes pour une même catégorie de données. En cela, il est nécessaire de promouvoir les initiatives de normalisation et d'automatisation des échanges au sein d'une profession, ou sur des types de flux largement partagés (ex : flux de paiement), sachant que le champ de ces normalisations est de plus en plus mondial.

Par ailleurs, les données externes doivent être contrôlées pour pouvoir être utilisées de façon pertinente dans l'entreprise : quand il s'agit de données non structurées, cela signifie des vérifications et/ou recoupements de sources, mais aussi un contrôle humain des informations sélectionnées et de leur qualité afin de garantir la pertinence des données au regard des usages.

Dans l'autre sens, il serait utile de disposer d'éléments explicites qualifiant à la fois le degré de qualité d'un document ou d'une information (l'information a-t-elle été vérifiée ? provient-elle d'une source professionnelle ? quel crédit lui accorder ?), ainsi que les conditions de sa réutilisation (notion de *copyright*,risques liés à l'extraction d'une information, et donc à son utilisation « hors contexte » ? …).

Témoignage SCOR

Un dispositif « d'intelligence économique » a été mis en place depuis une dizaine d'années, qui permet, à des fins de développement du business, la collecte de données externes non structurées, et leur mise à disposition ciblée auprès des experts à même de les analyser, après validation de pertinence et classement. Un outil permet de faire un premier filtrage des informations sélectionnées sur les différents thèmes d'intérêt définis, et une équipe est ensuite en charge de vérifier les contenus avant publication auprès d'un réseau d'experts.

*Des POC (*Proof of Concept*) sont en cours afin de mettre en place des capacités de traitement combinées de données internes et externes.*
Une vaste initiative de dématérialisation des échanges de données administratives de réassurance entre assureurs, courtiers et réassureurs est engagée (Initiative Ruschlikon – e-administration in (Re)insurance *– qui décline les normes et standards XML définis par le groupement ACORD pour le secteur de l'Assurance).*

3.3. L'obligation de qualité exacerbée

La qualité des données est directement liée à la nécessité d'avoir un vocabulaire commun, des référentiels, des modèles et une cartographie des données (géolocalisation des données, définir les rôles et responsabilités et quelques règles : puis-je diffuser l'information ? si oui, à qui ? qui décide ? quelles sont les contraintes règlementaires qui s'appliquent aux données que je gère ? etc.). Avoir toutes ces informations permet de répondre en partie aux enjeux de valeur tout au long du cycle de vie.

La question de la qualité des données est au cœur des enjeux relatifs à la valeur des données pour les métiers : comment valoriser des données dont on ne saurait pas garantir la qualité ? En tout état de cause, la définition des marges d'erreur acceptables eu égard à l'usage des données est une étape incontournable dans la démarche de valorisation des données. Il est par exemple essentiel de disposer d'un excellent niveau de qualité des données d'identification des clients (nom, prénom, adresse postale, adresse e-mail, téléphone …) si l'on veut garantir l'efficacité des campagnes marketing. Un fichier client est un actif qui a une valeur importante, et peut la plupart du temps être commercialisé (existence d'une valeur de marché), mais cela suppose de pouvoir en garantir la qualité, notamment au travers d'interactions fréquentes avec les clients, qui sont autant d'opportunités de mise à

jour des données. *A contrario*, certaines données telles que des données calculées (données financières, cotation d'un risque …) peuvent supporter des marges d'erreur (l'approximation n'altère pas la pertinence de la décision). Le pilotage de la qualité des données passe donc par une qualification du niveau de précision requis, et la mise en place des contrôles permettant de vérifier que le niveau de qualité souhaité est atteint.

Elle est également intimement liée à la question du cycle de vie des données : en effet, garantir la qualité des données en évitant des surcoûts liés au contrôle et à la correction permanente des données suppose d'optimiser globalement le circuit de traitement des données, de l'acquisition jusqu'à la restitution. En effet, si les contrôles qualité sont automatisés, et directement intégrés dès les phases d'acquisition, et si par ailleurs on est capable de garantir que la donnée n'est pas altérée lors de son cheminement dans le système d'information, alors les contrôles *a posteriori* peuvent être allégés, et la donnée peut être utilisée simplement par les fonctions de pilotage de l'entreprise.

L'organisation des entreprises doit évoluer pour prendre en compte ces exigences de qualité de l'information. Les Métiers sont nécessairement responsables de la qualité de leurs données. Ils définissent les contrôles à opérer lors de l'acquisition ou de la (voire des) transformation(s), et qui seront embarqués dans les applications. De même, ils sont aussi les seuls à savoir quels contrôles appliquer sur les résultats en sortie pour valider la possibilité de les utiliser pour prendre des décisions adaptées.

Les technologies permettent désormais d'accéder en temps réel à des volumes importants de données (*Big Data*), et notamment de combiner l'utilisation des bases de données du *« legacy »*, et de données contextuelles et volatiles, parfois peu structurées (navigation sur Internet par exemple), et de réaliser des contrôles en temps réel. On peut ainsi mettre en œuvre immédiatement des processus décisionnels qui, auparavant, étaient réalisés en différé, et donc de façon à la fois moins pertinente et moins efficace. La frontière traditionnelle entre les fonctions *« front-office »* (intégrant beaucoup de fonctions d'acquisition de données) et les fonctions *« back-office »* et de pilotage a donc tendance à s'estomper, et la capacité à appréhender dès l'origine les contraintes relatives à l'ensemble du cycle de vie devient un atout important. De même, la difficulté à prédire les usages, combinée à des exigences de plus en plus fortes de temps réel, doivent conduire à penser l'organisation physique des données pour en faciliter l'accès : plus les données seront accessibles simplement, plus elles seront utilisées. Et, plus elles seront utilisées, meilleure sera leur qualité et plus l'entreprise sera en situation d'en tirer de la valeur. Le cycle de vie doit donc être regardé en intégrant dès le départ les questions d'accessibilité, au travers notamment de la mise en place de dispositifs de visualisation des données.

La qualité n'est pas un élément facile à mesurer, et doit nécessairement être mise en perspective des usages. La juste qualité est celle qui convient relativement à l'utilisation que

l'on fait des données : gérer la qualité des données consiste donc à définir des exigences de qualité variées, et intimement liées aux processus Métiers qui vont consommer les données. Une fois les exigences de qualité définies, il faut se mettre en position de mesurer la satisfaction de ces exigences. Cela signifie mettre en place des contrôles, définir des marges d'erreur et vérifier qu'elles sont respectées, et enfin corriger si nécessaire, en veillant à conserver la trace des corrections réalisées (piste d'audit).

Le processus de gestion de la qualité des données est en fait un processus d'amélioration continue de type *« Plan/Deploy/Monitor/Act »* (cf. DM-BOK, page 293). La fréquence des contrôles et le rythme de ce processus doivent être adaptés aux exigences des consommateurs. A l'inverse, des processus, par exemple dans le domaine du décisionnel, qui doivent pouvoir se mettre en œuvre très rapidement, se satisfont la plupart du temps d'une moindre précision. Un équilibre est donc à trouver en permanence entre la rapidité demandée et l'exigence de qualité : les nouvelles technologies permettent cependant de concilier ces deux exigences en permettant l'automatisation des contrôles les plus essentiels pour assurer un contrôle minimal.

Il faut noter par ailleurs que plus les données sont publiques, plus elles sont utilisées largement, plus les exigences de qualité sont fortes. Les données financières d'une entreprise, et plus largement les données publiées sont soumises à des exigences de contrôles importantes, intégrant des normes d'audit externe. Le périmètre de ces données publiées a par ailleurs tendance à s'élargir pour répondre aux exigences de transparence et à l'évolution de la règlementation. On peut par exemple noter que dans le secteur financier (banque et assurance), le nombre d'informations publiées et leur nature ont été largement affectés ces dernières années par les règlementations relatives aux exigences de fonds propres (Bâle 2 puis 3 dans la banque, Solvency 2 dans l'assurance) et par les nouvelles normes comptables (diffusion des normes comptables IFRS en Europe). Notons également que les nouvelles données publiées sont de plus en plus des données extra-comptables, en tout cas élaborées en dehors des processus comptables qui par construction sont porteurs d'un certain niveau de qualité (contrôle d'équilibre lié à la partie double, piste d'audit statique et dynamique …) : ceci conduit à mettre en place des processus de réconciliation, qui peuvent s'avérer très lourds, et peuvent être évités par des processus et une architecture du SI permettant d'obtenir un bon niveau de qualité « à la source ». Les modèles de risques utilisés dans le secteur financier peuvent par ailleurs nécessiter un « jugement d'expert » qui compte tenu des impacts financiers potentiels, doit faire l'objet d'une documentation claire. De plus en plus, les niveaux de contrôles exercés sur les données doivent être explicites, et constituer une forme de labellisation. Cette démarche qui s'applique lorsqu'on publie des données à l'extérieur de l'entreprise, est certainement à développer aussi en interne : en effet, attacher aux données les plus utilisées des attributs

explicitant les différents niveaux de contrôles appliqués à la donnée (un label qualité) facilite le partage et la valorisation.

En résumé, on ne peut imaginer développer une politique de valorisation des données sans traiter cette question de la qualité des données, en mettant en place des processus ajustés en fonction des usages, et optimisés grâce à une organisation, des méthodes et des outils dédiés à cette question. Cette recherche de qualité de l'information partagée et manipulée au sein de l'entreprise relève d'une démarche qui embarque l'ensemble des Métiers.

4. CREER DE LA VALEUR A PARTIR DES DONNEES

4.1. Exploitation et analyse

Qu'elles soient d'ordre réglementaire, économique ou de gouvernance, les activités de suivi, d'analyse ou de pilotage nécessitent une organisation et des outils. Le pilotage de processus, les rapports d'activité et autres arrêtés des comptes n'ont pas attendu les solutions informatiques, et plus précisément le décisionnel, pour être réalisés, mais à quel prix ? Combien de collaborateurs est-il nécessaire de mobiliser pour les produire ? Pendant combien de temps ? Les valeurs obtenues sont-elles cohérentes entre elles ?

Les infocentres ont permis de répondre à un premier niveau de besoin : disposer des informations stockées dans les systèmes de gestion à travers des extractions. Mais comment, en plus de ce puits de données, disposer d'une information en mesure de fournir de la traçabilité, de l'historique et permettant de mettre « en relation » les données ?

Témoignage Fondation de France

Pour répondre à ce besoin de mise en relation des données, la Fondation de France a développé et mis en place un système décisionnel. Basé autour d'un entrepôt de données, le SID (Système d'Information Décisionnel) a pour ambition de répondre à plusieurs objectifs :

- *Fournir une palette de fonctionnalités facilitant le pilotage, le suivi et l'analyse des activités,*
- *Accroître la réactivité des décideurs au sens large,*
- *Permettre à ces derniers d'être focalisés davantage sur l'analyse que sur la création d'information.*

En réunissant les informations clés de l'entreprise, l'entrepôt de données offre des perspectives jusqu'à ce jour non perceptibles, qui dépassent les aspects purement techniques. En effet, il s'agit de mettre en musique les partitions organisationnelles et processus à travers la mise à disposition de données pertinentes, fiables cohérentes et « historisées ».

Œuvrant comme un véritable catalyseur d'information pour la prise de décision, le portail décisionnel accentue l'approche collaborative et collective des données de l'entrepôt, et permet de répondre aux objectifs :

- *D'amélioration de la synergie entre processus et entité organisationnelle,*
- *D'accroissement de la réactivité et de la capacité d'anticipation des décideurs.*

4.2. Valorisation

La question de la valorisation renvoie à celle de la valeur des données, et ce questionnement est aussi générique que celui relatif aux brevets, marques, ou encore aux SI. Mais quels critères associer à la valorisation des données ? Pour une organisation, quelle qu'elle soit, quelle valeur créer à partir de l'exploitation de cet actif ?

Pour une entreprise, un actif est une ressource dont elle contrôle l'usage (à travers des droits de propriété spécifiques). Une donnée peut donc être vue comme un actif numérique dès lors qu'elle comporte un potentiel de création de valeur pour l'entreprise. Ce qui est nouveau dans la réflexion sur les données, c'est la démultiplication des potentiels de combinaison des données dans une perspective de création de valeur pour le *business*. L'*Accéluction*[33] marque le mode de production dominant dans l'entreprise de demain (accélération de liens et plateforme de services). Ce nouveau modèle de création de valeur va au-delà des 3 espaces classiques du *lean production*, que sont les fournisseurs, les clients et les entreprises. Dans ce contexte émergent, ces 3 catégories d'acteurs doivent désormais gérer une dizaine d'espaces différents, dont les données.

Or, la valeur n'est pas intrinsèque à l'information : elle réside dans **l'usage** que les utilisateurs en font. On ne peut pas prédéfinir la valeur de l'information, on ne peut pas prévoir à l'avance l'usage que feront les utilisateurs des informations mises à leur disposition. Ainsi, sur cette question de la valeur des données, un des axes clé est d'identifier en quoi les données à valoriser contribuent à la réalisation d'un des axes stratégiques de l'entreprise. La responsabilité de l'utilisateur portera alors sur la production et la mise à disposition de données de qualité.

Selon le Pr Bounfour, Rapporteur général du programme ISD de la Fondation CIGREF, la donnée est un actif numérique qui peut être de 3 natures différentes :

- Actif numérique propriétaire (carte de fidélité par exemple),
- Actif numérique conjoint (partenariats et alliances),
- Actif numérique non propriétaire sur lequel l'entreprise n'a aucun contrôle (ou actif numérique externe non propriétaire).

[33] L'*Accéluction* est un concept développé par le Pr Ahmed Bounfour, Rapporteur du programme international de recherche ISD de la Fondation CIGREF, et qui caractérise l'émergence d'un « nouveau système de production accélérée de liens », marquée à la fois par « l'extension des espaces de création de valeur et par l'accélération considérable des liens (transactionnels ou organiques) comme source de création de valeur ». In « L'Accéluction en action, programme international de recherche ISD : premier rapport d'étape, une mise en perspective des projets Vague A », Fondation CIGREF – Octobre 2011, p. 8

Mais comment combiner tous ces actifs pour créer de la valeur ? Faut-il ouvrir ces actifs ? La question de la gouvernance des données pose celle de la combinatoire de ces actifs entre eux.

Selon le modèle DVM développé par le Professeur Bounfour, la valorisation des données peut répondre à différents « attendus », comme par exemple :

- Identifier le potentiel stratégique des données et l'impact sur le modèle d'affaires de l'entreprise (créer de nouveaux produits et services)
- Améliorer la connaissance (du client, de l'écosystème, …)
- Définir la potentialité de partage de valeur avec les partenaires (question de la rente)
- Vérifier la « transversalité » de la donnée (une donnée sera d'autant plus valorisée qu'elle peut être utilisée de manière pertinente dans différents contextes : quelle efficience interne ? comment intégrer l'efficience externe ?)
- Améliorer la prise de décision et la gestion en temps réel : réactivité, anticipation, assimilation, proposition de valeur
- Se différencier par l'éthique des usages et la RSE (différenciation dans les usages, l'éthique pouvant être une source de création de valeur)

Si le potentiel de création de valeur des actifs numériques (les données) est confirmé sur une ou plusieurs dimensions, alors il est possible pour l'entreprise de tirer une rente supplémentaire à partir de l'exploitation des données. Ainsi par exemple, pour les entreprises des secteurs de la banque et de l'assurance, une adaptation profonde des métiers est nécessaire, liée à l'impact du numérique sur la donnée. Ces entreprises disposent de volumes très importants de données à valoriser et qu'elles doivent savoir utiliser. L'approche par l'usage est clé, autant que la qualité, et raisonner par rapport aux dimensions précédemment énoncées peut être un moyen d'identifier les gisements de création de valeur.

Par ailleurs, comme le souligne l'OCDE dans un rapport de 2013[34], la convergence des tendances technologiques et socioéconomiques liées au développement de l'utilisation d'internet et à la décroissance des coûts de gestion de la donnée dématérialisée, conduit à la génération d'un volume important de données qui peut être exploité pour la création de nouveaux produits, pour la création de nouvelles richesses et pour la création de valeur nouvelle pour les entreprises. Plus précisément, l'OCDE identifie cinq secteurs porteurs de création de valeur :

- L'amélioration de la recherche et développement ;

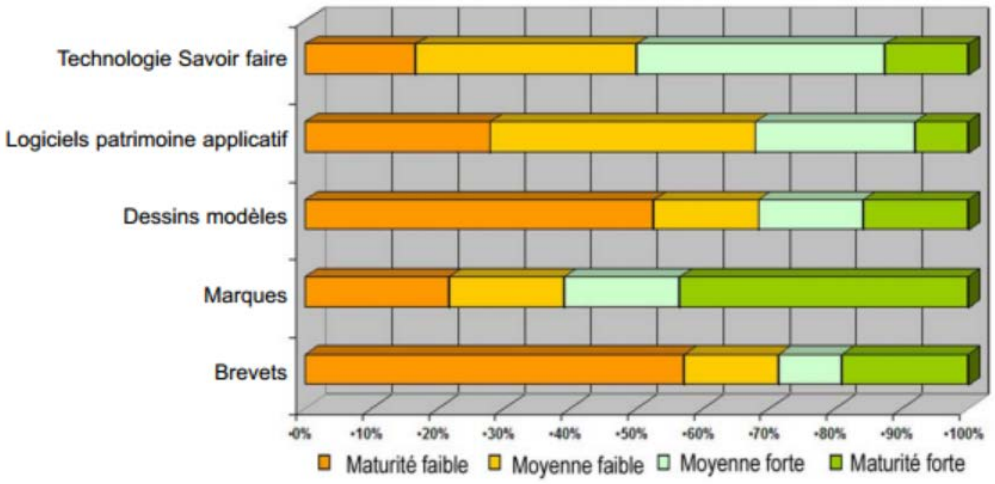
[34] OECD (2013), "Exploring Data-Driven Innovation as a New Source of Growth: Mapping the Policy Issues Raised by "BigData"", OECD Digital Economy Papers, No. 222, OECD Publishing. http://dx.doi.org/10.1787/5k47zw3fcp43-en

- La création de nouveaux produits basés sur la donnée ;
- L'optimisation des processus de fabrication ;
- L'optimisation du marketing ciblé ;
- L'amélioration des approches managériales.

S'il apparaît « évident » aujourd'hui que la donnée doit être considérée comme un actif immatériel essentiel de l'entreprise, force est de constater cependant que peu de modèles existent pour estimer la valeur de ces données. L'OCDE a ainsi récemment évalué différentes méthodes pour calculer la valeur monétaire des données personnelles détenues par les entreprises[35]. Elle conclut que les différentes méthodes d'évaluation sont très dépendantes du contexte, que les méthodes doivent être combinées pour éviter tout biais, et que leurs résultats doivent être utilisés avec prudence. L'OCDE souligne également que la valeur monétaire ne résume pas à elle seule la valeur de ces données.

Si le CIGREF a traité de manière pionnière et extensive le sujet du capital immatériel[36], les études ont essentiellement porté sur les notions de savoir-faire, de patrimoine applicatif, de valeur pour les métiers, sans focus sur les données elles-mêmes.

Figure 3 : Niveau de maturité des entreprises, relativement au traitement de leurs actifs immatériels – SI Liste des actifs immatériels, Cahier de recherche du CIGREF n°4 (2008)



Le think-tank « Futur Numérique »[37] (Institut Mines Télécom / Fondation Télécom) dans son programme « *Business models* numériques du futur », reconnaissant l'importance des

[35] OECD (2013), "Exploring the Economics of Personal Data: A Survey of Methodologies for Measuring Monetary Value", OECD Digital Economy Papers, No. 220, OECD Publishing. http://dx.doi.org/10.1787/5k486qtxldmq-en

[36] CIGREF, 2008 : « Capital immatériel et Systèmes d'information - Cahier de recherche n°4 »

données dans les modèles d'affaires numériques, ainsi que le manque d'outils dans l'évaluation des données a décidé, en 2014, de se concentrer sur la valeur des données.

Par ailleurs, le programme international de recherche ISD[38], piloté par la Fondation CIGREF, traite de la question de la donnée en tant qu'élément central de l'écosystème de l'entreprise de 2020.

Aussi deux constats s'imposent :

- Celui de l'importance de la donnée dans les modèles d'affaires numériques, dans les vecteurs de croissance pour l'entreprise ;
- Celui du manque d'outils conceptuels pour apprécier la valeur de la donnée, qu'il s'agisse de sa valeur d'usage, ou de sa valeur patrimoniale.

Ces deux constats appellent à mener une action afin de développer des modèles d'évaluations des données d'entreprise.

Citons enfin quelques leviers à même de permettre la création de valeur des données :

- Une démarche constante de qualité des données ;
- Une méthode de valorisation des données, et des processus ;
- La « liquidité » de la donnée[39,] i.e. la capacité à diffuser la donnée au sein de l'entreprise (désilotage), mais également à l'échanger avec les partenaires, par des efforts du développement de la standardisation des modèles de données (sémantique des données …) autour de modèles sectoriels.

Témoignage Groupement des Mousquetaires

Le Groupement des Mousquetaires a depuis longtemps déjà abordé la question de la valorisation du capital immatériel[40]. Cependant jusqu'ici, les méthodes ou indicateurs employés étaient axés sur un modèle global de la DSI (méthode IC-dVAL), sur la valorisation des projets informatiques par le criblage, ou sur l'estimation de la valeur créée

[37] http://willstrategy.com/download/MinesTelecom/2014-Will_IMT-BusinessModelsNumeriques.pdf

[38] Présentation du programme, dont le Rapporteur scientifique est le Pr Bounfour : http://www.fondation-cigref.org/programme-isd/

[39] Cf. intervention d'Henri Verdier, Etalab : http://www.youtube.com/watch?v=ehut9DmuTRc

[40] A Bounfour, G Epinette : « Valeur et performance des SI : Une nouvelle approche du capital immatériel de l'entreprise », Dunod, Paris, 2006 ; Publication Cigref, Cahier n°4, Capital immatériel et Systèmes d'Information http://www.cigref.fr/cigref_publications/RapportsContainer/Parus2008/Cahier_recherche_4_SI_Immateriel_2008.pdf

par les applications informatiques (tableaux de création de valeur).

Le Groupement des Mousquetaires a entrepris, en collaboration avec le professeur Bounfour[41]*, de développer un modèle de valeur de la donnée afin de compléter les axes d'analyse de la valeur du capital immatériel. Encore en cours d'élaboration, ce modèle –* Data Value Modelling *– DVM, s'attache à définir de manière rigoureuse la notion d' « actif numérique » à travers des critères d'évaluation thématiques. La définition de ces critères doit permettre d'estimer la valeur des données tant pour les données internes à l'entreprise, que pour la valeur d'usage des données « publiques », ou celles liées à l'ouverture des données de l'entreprise à des partenaires. Le modèle sera éprouvé sur des cas d'usage du Groupement des Mousquetaires (carte de fidélité, …).*

4.3. L'*Open Data*[42]

Le terme *Open Data*[43], littéralement « données ouvertes » ou « données libres de droits », est utilisé pour désigner des données en libre accès et ouvertes à l'utilisation par des tiers.

L'*Open Data* trouve son origine dès 1995 aux États-Unis dans le monde scientifique où le partage de données multiplie les chances de détecter des erreurs et permet une meilleure collaboration entre chercheurs. Puis l'ouverture des données publiques s'est effectuée aux États-Unis « du bas vers le haut » : les villes de Washington, New-York et San Francisco ont ouvert des *« Data Store »* à disposition du public et ont encouragé leur réutilisation en organisant des concours d'applications qui connaissent un succès considérable. Le portail data.gov ouvre en mai 2009 aux États-Unis.

[41] Le Pr Bounfour est Rapporteur scientifique du programme ISD de la Fondation CIGREF et titulaire de la Chaire européenne de Management de l'Immatériel à l'Université Paris Sud 11 : http://www.chairedelimmateriel.u-psud.fr/tiki-index.php

[42] Les informations contenues dans cette partie sont tirées du rapport « Pour une politique ambitieuse des données publiques », de l'Ecole des Ponts Paris Tech (2001)

[43] Une donnée ouverte est une donnée numérique d'origine publique ou privée. Elle est produite par une collectivité ou un service public (éventuellement délégué). Elle est diffusée de manière structurée selon une méthodologie et une licence ouverte garantissant son libre accès et sa réutilisation par tous, sans restriction technique, juridique ou financière. L'ouverture des données (en anglais *open data*) représente à la fois un mouvement, une philosophie d'accès à l'information et une pratique de publication de données librement accessibles et exploitables. Elle s'inscrit dans une tendance qui considère l'information publique comme un bien commun (tel que défini par Elinor Ostrom) dont la diffusion est d'intérêt public et général. En Europe et dans certains pays, des directives et lois imposent aux collectivités de publier certaines données publiques sous forme numérique. (source : Wikipedia)

La Grande-Bretagne adopte une démarche similaire, mais complète le dispositif en ouvrant une « agence pour les données publiques », dont l'objectif est d'accompagner les réutilisateurs potentiels.

4.3.1. Le cas des données de l'Administration

En France, une mission, Etalab[44], coordonne les administrations en matière de mise à disposition de leurs données. La création d'un portail permet d'en faciliter l'accès et donc leur réutilisation par la communauté des entrepreneurs et des *startups*. Dans les collectivités territoriales, Rennes fait figure de pionnier en ouvrant ses données sur le transport, puis en organisant un concours d'applications. La Ville de Lyon ouvre également ses données dans le cadre d'un projet d'optimisation de la mobilité en ville, lancé en 2011 avec 12 partenaires. La Ville de Paris, quant à elle, ouvre ses données publiques aux entreprises dans le cadre de licences *« share alike »* (utilisation commerciale sous réserve de partager leur création dans les mêmes conditions d'utilisation et de mise à disposition).

L'ouverture des données publiques est à la fois une opportunité d'amélioration pour l'État et l'administration, et un outil stratégique pour le développement de notre économie. En effet, en réutilisant les données collectées dans le cadre du fonctionnement des administrations, des entrepreneurs, des chercheurs et des développeurs inventent de nouveaux usages qui contribuent ainsi à constituer une nouvelle relation numérique de l'administration avec les citoyens. De surcroît, l'accès aux données publiques favorise l'essor de l'économie numérique, vecteur d'innovation et de création d'emplois.

La Commission européenne estime la valeur potentielle de la donnée publique à plusieurs dizaines de milliards d'euros, d'autant que les réutilisateurs potentiels de données publiques sont de nature très diverse :

- Chaque citoyen peut récupérer des données publiques et les enrichir avec des données personnelles, ou en les combinant avec d'autres données ouvertes pour délivrer un service nouveau.
- Des développeurs de logiciels, libres ou non, sont également des parties prenantes essentielles au mouvement d'ouverture des données.
- Dans le monde de la recherche, la réutilisation de données produites ou mises en commun par les chercheurs est un élément important du succès de grands projets scientifiques.
- Plusieurs associations ou collectifs se sont donné pour objectif d'encourager la mise en ligne de données publiques dans un contexte de transparence démocratique, de respect de l'environnement, de droits des consommateurs, etc. Le « journalisme de

[44] http://www.etalab.gouv.fr/

données » est aussi un exemple marquant de l'importance potentielle de la réutilisation pour le débat démocratique.

- La réutilisation commerciale des données peut permettre l'éclosion de nouvelles activités économiques, tant au niveau de jeunes pousses innovantes que d'entreprises qui enrichissent leurs produits par l'utilisation des données publiques.

Une opportunité dans l'écosystème du Web

L'écosystème du Web fait aujourd'hui l'expérience d'un véritable « déluge de données ». Ce sont des milliers d'exaoctets qui sont produits chaque année dans le monde. Le *Cloud computing* et le développement d'outils de traitement algorithmique des données à très grande échelle (outils du *Big Data*, par exemple) permettent le stockage, le traitement et l'analyse de masses énormes de ces données, structurées ou non. Les données publiques, dans leur état brut, participent à ce mouvement de fond. Ainsi, l'accès aux données publiques encourage le développement de jeunes entreprises technologiques, dans un processus d'exploration (au sens d'Eric Ries[45] qui définit la jeune pousse comme une organisation temporaire chargée de développer la solution commerciale et technologique qui répond le mieux à un problème supposé, mais qu'il faut découvrir, comprendre et confirmer), sous réserve que l'accès aux données soit aisé, et si possible gratuit.

Il est très difficile de mesurer précisément l'activité économique générée par la réutilisation des données publiques, car elle se matérialise au niveau des usages finaux et non à partir des volumes de données téléchargées. Néanmoins, des gains économiques ont été évalués globalement pour la société. Ainsi, le Royaume-Uni a estimé à 7.9 Md€ pour 2010 et 2011, les bénéfices de l'ouverture des données publiques pour la société britannique. Et ces gains sont maximisés lorsque l'accès aux données devient gratuit. Ainsi, le passage à la gratuité du Référentiel Grande Echelle de l'IGN pour les organismes chargés d'une mission de service public administratif a entraîné une multiplication par 20 des volumes téléchargés, soit un bénéfice social estimé à 114 M€ par an, pour un manque à gagner de 6 M€ de redevance environ.

La place de la France dans l'*Open Data*

La France est considérée comme bien placée par l'*Open Knowledge Foundation*, au sein du G8. C'est aussi ce qui ressort d'une analyse de Capgemini Consulting et de l'European Public Sector Information Platform, comme l'illustrent les figures 2 et 3 ci-après, reprises dans le rapport de Mohammed Adnène Trojette sur l'ouverture des données publiques, pour la Cour

[45] Eric RIES est un entrepreneur de la Silicon Valley, blogueur et auteur reconnu sur le « lean startup movement ».

des Comptes en 2013. Ces enquêtes ne peuvent offrir qu'un indice sur le dynamisme réel ou perçu de la France sur ces sujets.

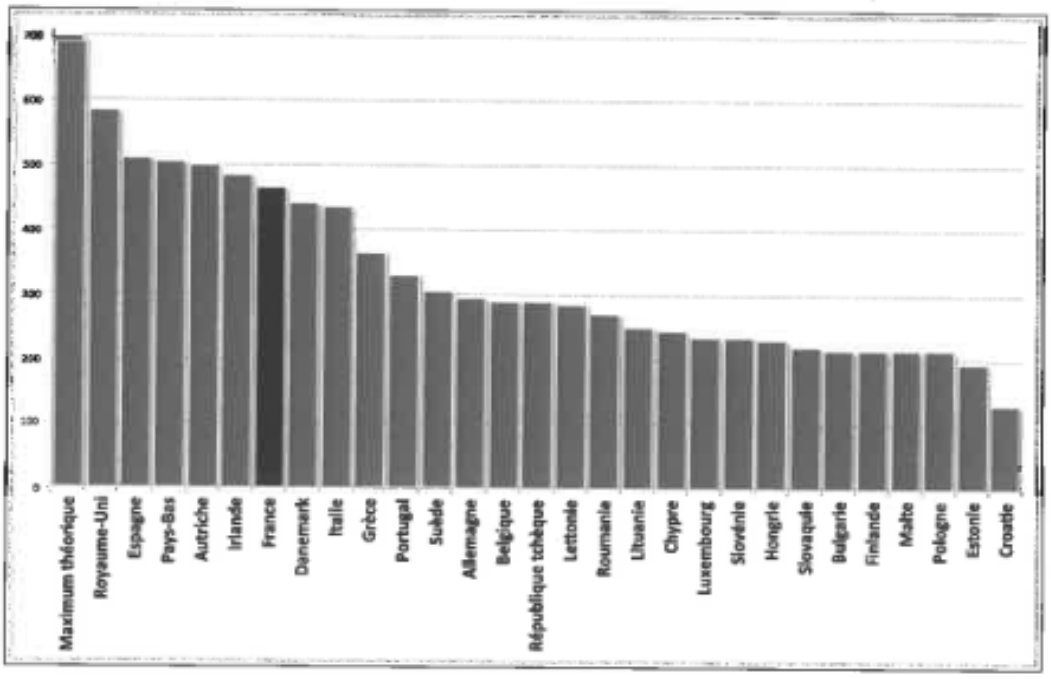
Néanmoins, Capgemini Consulting considère que la France a lancé une politique d'ouverture des données plutôt récemment, mais a engrangé des succès dans un laps de temps réduit, ce qui la place dans le groupe des précurseurs (*Trend Setters*).

Figure 4 : Comparaison des démarches d'ouverture des données publiques (2012)

NB : la taille des « bulles » représente le niveau de soutien gouvernemental.

Source: Analyse Capgemini Consulting.

Figure 5 : Classement collaboratif de l'ouverture des données publiques dans l'UE



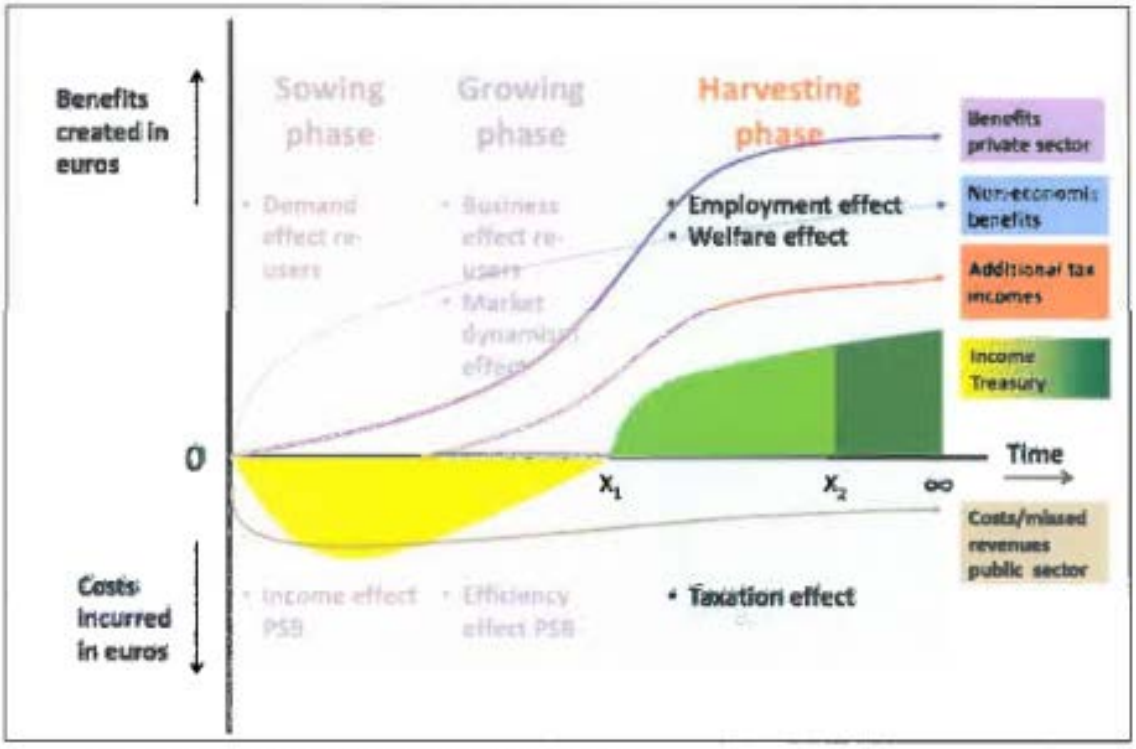
Source : Mission redevances. Données : European Public Sector Information Platform, PSI Scoreboard, juillet 2013.

Des publications récentes visant à estimer l'impact de la mise à disposition des données publiques s'appuient sur une modélisation empirique (voir figure 4 ci-après) en trois phases :

- L'ouverture avec un coût pour l'administration ;
- La croissance des réutilisateurs générant des profits pour la société ;
- L'impact positif sur l'emploi et les finances publiques.

Les précurseurs seraient plutôt dans la seconde phase, la troisième phase reste devant nous.

Figure 6 : Temporalité des effets de l'ouverture des données publiques



Source : Rapport à l'Agence spatiale européenne.

En synthèse

Il est très difficile de mesurer l'impact *Business* réel de l'ouverture et de la réutilisation des données publiques, car ce sont les usages finaux qui en font la valeur, sans lien direct avec les volumes téléchargés. Néanmoins, cette valeur est considérable sur le plan économique direct et est un vecteur d'innovation et de dynamisme pour de nombreux acteurs innovants.

Les clefs de succès sont :

- Un accès aisé aux données brutes (coûts faibles et formats réexploitables),
- La rapidité d'exécution,
- La sécurité juridique.

Faut-il mettre les données à disposition gratuitement ou adopter un modèle payant ? Ce dernier permettrait, selon l'État, d'obtenir un complément budgétaire pour couvrir les coûts de diffusion, voire une partie des coûts de la mission de service public. Il correspondrait aussi à la volonté de proposer une offre, et pour certains de réguler le volume des demandes. Ce point de vue est celui d'un nombre important de producteurs de données publiques. Cependant, la tendance actuelle n'est pas celle-ci car les avantages du modèle de gratuité semblent l'emporter si l'on considère les gains « sociaux et économiques ».

Par ailleurs, l'accent mis sur la rapidité d'exécution au détriment de la complétude, voire de la qualité est un véritable changement culturel puisqu'il s'agit de passer d'une culture administrative de l'offre à une posture axée sur la demande.

4.3.2. L'ouverture des données au sein de l'entreprise

L'enjeu de l'ouverture des données concerne également l'entreprise en interne. Nombre de données sont restées longtemps enfermées dans des référentiels peu ou pas communicants, que ce soit par les solutions qui les abritent, ou par la volonté de contrôler fortement leur utilisation. Le maintien en situation de silo est effectivement justifié sur certains référentiels confidentiels, mais se justifie de moins en moins pour des référentiels permettant la collaboration entre les différentes entités d'une organisation. Qui plus est, la mise en place d'autorisations d'accès à un grand nombre de bases a engendré de grands retards dans les projets transverses et se trouve à l'origine de difficultés diverses dans la maîtrise du savoir contenu dans ces objets métiers.

Les entreprises sont conscientes de ces difficultés, tout en connaissant l'apport évident qu'un meilleur partage des informations structurantes apporterait dans le recoupement des différentes sources. L'*Open Data* se trouve donc décliné au sein de l'entreprise afin d'ouvrir des données utiles pour les nouveaux projets, pour l'essentiel tirant leurs valeurs de la diversité des données et de leur mise en relation.

Témoignage du Groupe La Poste

Le Groupe La Poste est parti du constat que de nombreux projets innovants se heurtaient à des difficultés d'accès aux données des référentiels, certains de ces projets ayant subi des retards de plusieurs mois. Dans un premier temps, le choix de l'Open Data a été moteur dans la création de nouvelles infrastructures industrialisées permettant l'exposition des données et des services à l'extérieur. Au-delà de ce premier pas, ces structures exposent désormais auprès des partenaires, et contribuent fortement en interne à la synergie entre les Directions et les Branches (Métiers). Une organisation s'est créée pour veiller au caractère diffusable des données exposées et garantir la confidentialité des données sensibles et/ou personnelles. Le partage est donc piloté et l'ouverture encadrée.

Afin d'accompagner le mouvement, la branche Numérique a mis en place différents événements d'information et des comités de données afin d'encourager les différents acteurs à valoriser les données de l'entreprise. De manière régulière, ces acteurs se réunissent en comité afin de partager les expériences en cours et déterminer les besoins futurs d'exposition des données pour satisfaire l'émergence de projets innovants.

Cette approche s'adresse à toutes les entités du Groupe et soutient la volonté de la Direction de travailler de concert sur de grands projets transverses.

5. CONCLUSION

Avec la révolution numérique qui est aussi celle de l'information, avec le phénomène *Big Data*, avec le renforcement des exigences de qualité des régulateurs, la donnée est devenue l'un des principaux actifs de toute entreprise ou organisation. Les autres actifs, comme le capital, les ressources humaines, les processus, les clients ou même le système d'information sont l'objet d'une gestion organisée, professionnelle et industrielle.

Puisque la donnée est unanimement reconnue comme la nouvelle richesse de l'entreprise, un dispositif de gestion (*« data management »*) de même nature doit être mis en place, pour accompagner toute démarche de valorisation.

Veiller à la mise en place d'une gouvernance et d'une gestion des données dans l'entreprise est donc une démarche devenue indispensable pour tout dirigeant désireux de tirer de la valeur des données disponibles, internes et externes. Les gisements de valeur contenus dans les données aujourd'hui encore insuffisamment exploitées sont importants, et les entreprises qui sauront anticiper et prendre en compte cet actif désormais stratégique, en tireront un avantage compétitif durable.

Mettre en place un système de *« data management »*, tel que décrit dans le présent livrable, est indispensable pour tirer tout le profit possible de ses données, innover et *in fine*, prendre de meilleures décisions, et proposer de nouveaux produits et services aux clients finaux.

Nous espérons que ces quelques pages auront contribué à vous convaincre de la nécessité de passer à l'action !

6. ANNEXE : CRITERES D'AUTO-EVALUATION DE LA MATURITE DES ENTREPRISES EN MATIERE DE GESTION DES DONNEES

Pour réaliser votre auto-évaluation, merci d'utiliser l'outil Excel disponible en complément de ce rapport sur le site du CIGREF : www.cigref.fr

Axes	Critères d'évaluation
Gouvernance	La donnée est reconnue comme un actif clé au sein de l'entreprise
	Il existe une stratégie et des politiques de la gestion des données
	Les rôles et responsabilités sont clairement définis en matière de gestion des données
	Une organisation spécifique, en charge de la gestion des données, est mise en place
	La propriété des données est établie
	Une gestion prospective des compétences requises autour des données est en place
	Une gouvernance spécifique est en place pour la protection des données
	Des règles d'éthique sont définies dans l'usage des données
	L'entreprise est sensibilisée à l'importance d'une bonne qualité des données
	Le besoin de transversalité dans l'entreprise pour favoriser la valorisation des données est pris en compte (culture orientée "données")
	Une démarche d'identification de la valeur des données est en place
Référentiels	Les données de référence sont gérées spécifiquement
	Chaque donnée de référence est placée sous une autorité responsable de son cycle de vie
	Il existe un mapping des références internes avec les références externes (standards ou partenaires)
	La gestion des données de référence est outillée
	Les données de référence sont utilisées en transverse par l'ensemble des systèmes
	S'il y a plusieurs versions des données de référence, une version maître est identifiée ("golden copy")
Métadonnées	Le cadre standard pour les métadonnées existe
	Les données clés sont définies et leurs attributs identifiés
	Une autorité unique est identifiée pour la définition et la gestion des métadonnées
	Un outillage est en place
Sécurité	Une catégorisation des données a été établie
	Le niveau de protection associé est défini
	Une politique de protection des données est en vigueur
	Les contraintes réglementaires et juridiques sont connues et prises en compte (données à caractère personnel, droits de propriété intellectuelle, …)
	Les contraintes de localisation des données sont connues et prises en compte
	Une gestion des identités et droits d'accès est en place
	Les accès aux données de l'entreprise sont contrôlés
	Une organisation est en place qui garantit la sécurité des données
	Une organisation est en place qui garantit la conformité des données
	La sécurité des données fait l'objet d'un audit régulier

Axes	Critères d'évaluation
Architecture	Les données clés de l'entreprise sont identifiées
	Les données clés sont modélisées
	La gestion des données clés est outillée
	Le modèle des données clés s'impose aux équipes projets et évolue en fonction de leurs besoins
	Un dictionnaire des données clés est en place
	Les flux de données sont cartographiés
	Il existe une cartographie de la localisation des données
Qualité & Cycle de vie des données	Une même donnée est collectée une seule fois
	Une démarche d'identification des sources de données non-structurées utiles existe
	Les données non-structurées sont collectées de façon organisée
	Une politique de qualité des données est en place
	Toute donnée collectée quelle qu'elle soit est contrôlée
	Toute donnée transformée est également contrôlée
	Les défauts de qualité sont pilotés jusqu'à correction
	Le partage des données est facilité
	Une organisation est en place au sein des Directions Métiers pour améliorer/garantir la qualité des données
	Un reporting est en place et fourni sur demande aux autorités de contrôle
	La qualité du reporting est sous contrôle
	Certaines données de l'entreprise sont ouvertes à des tiers (open data, partenariats stratégiques)
	Les données sont archivées et purgées en fonction des politiques et contraintes de rétention
Exploitation et Analyse	Un processus d'anonymisation des données est en place pour permettre leur exploitation
	Des méthodes de *préparation* des données existent pour créer des ensembles de données cohérents
	Un système de reporting de qualité est en place pour l'analyse des données
	Un système d'exploration des données est en place
	Les nouvelles données accessibles sont exploitées (approche *Big Data*)
	Des applications de *Big Data* sont en place
	Des spécialistes de la donnée recherchent des dépendances et explorent des voies d'innovation (data science)
	Une démarche d'open data existe en interne (data democratisation)
	La connaissance des outils d'analyse des données est diffusée de façon organisée au sein de l'entreprise
	La valeur des données est estimée/mesurée
Opérations	Des modèles conceptuels et physiques des données existent
	Les bases de données sont administrées
	Les modules d'accès aux bases de données sont développés (connecteurs, APIs, web services, …)
	Les données sont stockées, potentiellement répliquées, et font l'objet de sauvegardes
	La qualité des sauvegardes est contrôlée
	Les données sont restaurées sur demande
	Des tests de continuité d'activité sont opérés régulièrement

CIGREF

21 avenue de Messine

75008 PARIS

Tel. : +33 1 56 59 70 00

cigref@cigref.fr

www.cigref.fr