Le Master Data Management (MDM) et la qualité des données de l'entreprise : synergies digitales et collaboratives

Dominique Mariko

La data en perspective, INTD-CNAM, 5 avril 2016

Cette introduction au MDM et à la DQM est réalisée grâce aux échanges permis par les groupes de travail d'ExQi (Excellence Qualité Information), mes remerciements vont aux membres du GT Big Data (Delphine CLÉMENT, Sebastiao CORREIA, Soumaya BEN HASSINE, Hervé HUSSON et Khalil EL IDRISSI) et du GT Valorisation, tout particulièrement à Xavier HENDERSON et Isabelle SIPMA pour la richesse de leurs débats sur les notions d'information et de valorisation des données en lien avec l'usage. Merci également à Nathalie BARTHÉLÉMY pour ses précisions relatives à GRDF et la bienveillance de ses commentaires.

dmariko.pro@gmail.com
2016.04.13 23:27:35 +02'00'

Table des matières :

I - Introduction

Cette introduction à la gestion des données de référence (Master Data Management, MDM) et à la gestion de la qualité des données (Data Quality Management, DQM) propose de mettre en perspective les trois objets digitaux que sont les données, les informations et les documents.

Quelles définitions retenir :

Donnée : description élémentaire d'une réalité, d'un objet, d'une personne ou d'un événement, voire d'un lieu. Cette réalité est inscrite dans les systèmes d'information en tant qu'enregistrement ou record (sous forme de n-uplet constitué d'une liste d'attributs ou sous forme d'objets).

Information : au sens ontologique, une information est un ensemble d'éléments porteurs de sens. Cette définition permet un rapprochement avec les méthodes développées pour la compréhension des systèmes complexes[1] : le lien entre données et informations est considéré ici comme non-linéaire et métastable, nécessitant des traitements basés sur des logiques conjonctives, récursives et la recherche d'invariants en probabilité (cf. §3.1). Les traitements strictement informatiques de l'information (considérée comme un ensemble de 0 et de 1) ne seront pas évoqués.

Document : Pour faire le lien avec le sujet seul le document numérique, comme construction numérique sur laquelle s'exercent des traitements calculatoires[2] , sera considéré, en tant que jeu de données.

II - Eléments de Master Data Management

C'est une technologie de gestion des données de référence, aussi appelées données maîtres, qui vise à obtenir une information unique et partagée dans l'entreprise.

On appelle données maîtres - ou données de référence, les données essentielles à la performance (et donc à la survie de l'entreprise), c'est-à-dire, essentiellement : les données clients et fournisseurs, les données produits, et les données de reporting (données BI). Selon les process dans l'entreprise le MDM peut gérer également d'autres types de données.

Chez GRDF par exemple les données d'organisation interne du réseau sont également considérées comme données de référence ; les données de nomenclature (liste des communes, unités de mesure, etc.….) sont aussi traitées comme données de référence mais de moindre enjeu que les autres master data.

La finalité du MDM est d'optimiser la synchronisation et l'interaction des bases de données du SI. C'est une problématique familière aux gestionnaires de l'information, qui la considèrent plutôt sous l'angle de la gestion des ressources ou des actifs informationnels de l'entreprise. Le MDM vise donc à garantir l'unicité des données (une donnée référencée par plusieurs applications mais centralisée de façon à être la même pour tous) et la cohérence des données (une donnée est visible partout de la même façon, quelle que soit l'application qui l'appelle).

Le but : augmenter la performance de l'entreprise (en ajustant la valeur des données) et diminuer les coûts liés au traitement et à la gestion des données maitres.

1- Les aspects techniques du MDM : systèmes et outils

Ils sont gérés par les DSI ou par des services dédiés à la gestion des données.

1.1. Une architecture centrée MDM : l'entrepôt de données (Data Warehouse)

Traditionnellement, une architecture de gestion des données de référence urbanisée est une architecture distribuée dévolue à la Business Intelligence (BI), ou les différentes applications sont considérées comme les clients d'un entrepôt central (le Data Warehouse). Les données référentielles, au lieu d'être gérées dans les applications, sont centralisées et gérées dans le Data Warehouse, puis distribuées régulièrement aux différentes applications, ce qui évite les incohérences, les conflits, etc. On dit alors qu'on n'a plus qu'une seule version de la vérité (a single version of the truth).

Si aujourd'hui les Data Warehouses sont moins valorisés qu'au cours des 20 dernières années, en particulier à cause des exigences de traitement en temps réel et de contextualisation des données multisources[3], ils sont encore très présents dans les grandes entreprises, il y a donc de nombreux sujets relatifs à leur évolution et à la manière dont il est possible de les adapter au traitement des grandes masses de données (Clusterisation et partitionnement, curation de données dans les Data Lake, etc.).

1.2. Les outils de MDM

Un outil de MDM doit permettre de construire des modèles de données, d'assurer leur qualité et d'effectuer un contrôle sur celle-ci. De nombreux éditeurs logiciels proposent des solutions d'intégration, de modélisation et de traçage des données qui permettent de réaliser ces opérations[4]. Les problématiques associées à la modélisation des données concernent la définition de hiérarchies sur les données ou la définition de familles de données, la maintenance de ces hiérarchies (le système doit conserver la possibilité de mettre à jour les liens hiérarchiques), la définition de typologies de données et de toutes les métadonnées associées aux traitements, ainsi que les choix techniques concernant les dérivations (héritage) de ces données et métadonnées.

1.3. Les systèmes d'enregistrements (Systems of Records)

Quelle que soit l'architecture source, les outils de MDM doivent permettre d'établir des Systems of Records (SOR)[5], c'est à dire des systèmes chargés de maintenir la représentation la plus complète et la plus digne de confiance d'un jeu de données (notions parallèles en RM et GED). Chez GRDF on parle par exemple de Golden Record pour qualifier la vision du client, constituée des différentes données des applications de GRDF mises en cohérence par le MDM.

Un SOR doit pouvoir gérer les difficultés liées à l'écriture des transactions et aux éventuels conflits d'accès aux données en lecture et en écriture : par exemple une même donnée client peut être écrite au même moment par l'application CRM et l'application dédiée au marketing. Il faut pouvoir identifier le type de données concernées, le référencement des systèmes clients (ERP, DW, CRM, PLM, etc.), identifier les connecteurs possibles pour l'intégration (flux XML, API, Web services, etc.).

L'architecture d'un système d'information, et plus généralement les choix technologiques, seront des influenceurs de la qualité de données (de même que les modèles et les besoins métiers).

2- Les aspects fonctionnels : analyse et cycle de vie.

Pour l'analyste qualité le travail consiste à établir un référentiel standardisé des données, quel que soit l'outil choisi, qui va piloter l'évolution des données maîtres et la mise à jour des bases de données et des applications métiers.

Avant même de penser l'ingénierie du système, il faut idéalement travailler avec les métiers pour reconstituer et surtout cartographier le cycle de vie des données, savoir comment elles sont produites et à quoi elles servent ou serviront. Le cycle de vie se définit généralement autour de fonctions telles que :

- la découverte et le profilage des données
- l'acquisition et l'intégration des données
- la maintenance des données
- l'usage des données
- l'archivage et la destruction des données

Une notion essentielle en qualité de données apparaît au cours de l'examen du cycle de vie : l'usage des données. Une bonne qualité de données se définit par rapport aux traitements métiers de ces données, on dit alors que la qualité est appropriée à l'usage qu'on en fait (fitness for use).

L'usage courant des données peut être déterminé par analyse statistique et en atelier avec les métiers, ce qui permet d'évaluer leur importance à un instant t, leur impact et donc leur valeur.

L'usage futur des données est plus délicat à appréhender. Cette possibilité apparaît avec la capacité à traiter aujourd'hui de grandes masses de données : on est capable de collecter et stocker des big data, mais il est difficile d'établir leur valeur, car on n'envisage pas au moment de la collecte quels pourront être les usages futurs de ces données, c'est un des défis actuel de la gouvernance des données. Une solution possible consiste ajouter des métadonnées de production et d'utilisation sur les données (data lineage) afin de faciliter leur accessibilité et leur exploitabilité pour des usages en devenir (pour un projet de norme internationale voir SDMX).

Pour aider à définir les usages et les contraintes techniques, sont déterminées dans l'entreprise des « fonctions » data :

- les chief data officers qui, entre autre, déterminent et garantissent les principes de gouvernance sur les données. Les CDO interviennent au niveau exécutif et non au niveau de la gestion des données.[6]
- les data owners, c'est à dire des propriétaires de la données, qui ont la priorité sur les données et sont chargés d'autoriser ou restreindre l'accès aux données, ils sont aussi responsables de leur exactitude, de leur intégrité et de leur fraîcheur,

- les data stewarts qui sont des « intendants » de la donnée, souvent des gestionnaires de données,

Les rôles inclus dans ces différentes fonctions dépendent en fait du degré de maturité de l'entreprise en matière de gouvernance des données**.**

- Sur les projets, on trouvera aussi des business analysts, qui analysent les processus métiers pour déterminer les processus de création de la donnée en collaboration avec les architectes de données.

III - Les aspects théoriques et pratiques de la qualité de données

1- Convergences

La gestion des données de référence peut s'inscrire dans des projets plus étendus d'amélioration de la qualité des données (DQM). La définition de niveaux de qualité de données est considérée comme un vecteur de la création de valeur visant à améliorer les performances globales de l'entreprise. Ce type de projet ne concerne plus uniquement les master data mais leur articulation avec les données externes, les dark data, l'open data ainsi que les données réputées « non-structurées ». Les projets visent alors à minimiser l'impact du « garbage in, garbage out » en proposant des politiques d'amélioration continue de la qualité des données (DQM) afin de faciliter la maintenance et l'analyse de ces données. Ils ont pour objectif de minimiser les risques liés à la perte des données, les coûts opérationnels et d'éviter les retraitements préalables à l'analyse de données dont on ne peut faire l'économie sur des données disparates et de qualité inégale, retraitements qui occupent aujourd'hui entre 50 et 80% du temps des data scientists.

Les aspects théoriques de la qualité de données et les couches logiques qui entrent en jeu lors de la définition des modèles de données assemblent des notions similaires à celles utilisées en gestion de l'information. Pour arriver à faire émerger une qualité de données adaptée à l'usage, on peut commencer par travailler un concept qu'on rencontre aussi en modélisation et en gestion des bases de données : la concordance au réel. On cherchera par exemple à analyser à quel point les objets du SI représentent les objets du monde réel en travaillant les dimensions de complétude et d'exactitude. Les approches par les modèles[7] (MDA, Model Driven Architectures) proposent des paradigmes de résolution de la qualité de données intégrant les besoins en robustesse du SI et les besoins en agilité des métiers.

2- Mise en œuvre de la qualité de données :

Traditionnellement le traitement de la qualité de données dans les SOR est effectué au moment de l'intégration[8]. Avec les traitements big data, il pourra être exécuté également en sortie ou à la volée en fonction des demandes[9], ce qui pose des problèmes de synchronisation et d'interaction (Pour un projet open source de gouvernance des données big data, voir Falcon). Dans les data lake, par exemple, cohabitent différents niveaux de qualité de données, les traitements qualité nécessiteraient d'intégrer dans les outils de gestion de ces lacs de données des fonctions de data modeling et data cleansing, qui permettraient de qualifier les données au fur et à mesure de leur usage.

L'implémentation de la qualité des données est effectuée à travers un ensemble de règles de qualité.

Ces règles de qualité sont elles même dérivées :

- des règles métiers
- ainsi que de l'impact attendu par les métiers de ces règles de qualité des données (en termes de valeur)

Ex : il est inutile de forcer la complétude d'un jeu de données qui n'est employé que pour un usage restreint. Ainsi il sera coûteux mais pas nécessairement utile de conserver l'ancienne adresse d'un étudiant qui déménage dans une base de données administrative, par contre conserver l'ancienne adresse d'un client dans un CRM peut aider à dé-doublonner en cas d'homonymie. On peut donc trouver une valeur d'usage à conserver cet historique.

« *La qualité des données est définie par un ensemble de fonctions évaluables sur les données appelées* ***métriques de qualité*** *(ou indicateurs de qualité), qui permettent de mesurer l'adéquation entre la qualité effective des données et la qualité attendue de celles-ci.* » [10]

Pour déterminer ces indicateurs il faut donc considérer au moins trois éléments :

- *La donnée comme représentation d'un objet du monde réel (sa « valeur absolue »)*
- *La valeur d'usage de la donnée*
- *La valeur attendue de la donnée*

(Pour un retour d'expérience sur la création d'un référentiel d'indicateurs DQM, voir[11]).

3- Qualité de données : usages et valeurs de la donnée en tant qu'actif informationnel.

3.1. Une définition systémique de la valeur de l'information :

La valeur d'une information est inversement proportionnelle à la probabilité d'occurrence de l'objet qu'elle décrit.

Autrement dit : une information a potentiellement de la valeur si elle met en évidence un objet (évènement, lieu, personne, etc. : une donnée) inconnu ou méconnu au moment de son énonciation. En fonction du contexte (besoin en information) dans lequel cette information est énoncée sa valeur sera renforcée ou diminuée.

Cette définition est mise en œuvre dans les approches probabilistes de Berthier et Teboul[12], qui questionnent la valeur et la véracité des données : quelle est la probabilité de la véracité d'une donnée, connaissant son historique, son émetteur et sa réputation. (A mettre en relation avec les interrogations récurrentes en DQM sur la « crédibilité des données »[13]).

3.2. Propositions du Gartner[14]

Infonomics de Doug Laney : distinguer les modes d'analyse de la valeur en financier et non financier :

Non financier :

- valeur intrinsèque : qualité et facilité d'utilisation de la donnée versus quelle est la probabilité que d'autres à l'extérieur de l'organisation possèdent cette même donnée ?
- valeur métier de l'information : la donnée est elle utilisable par un métier ou un processus métier ? Avec quelle rapidité l'entreprise peut elle obtenir des données fraiches pour améliorer la performance de ce processus ?
- valeur de performance de l'information : combien une unité d'information contribue à faire en sorte que l'entreprise optimise ses indicateurs de performance ?

Financier :

- valeur du coût de l'information : quel serait le coût de remplacement des données si l'entreprise les perdait et devait les racheter.
- valeur de marché : à quel prix un partenaire d'affaire serait prêt à acheter les données
- etc.

Dans ces propositions apparaissent de manière récurrente des outils probabilistes en lien avec les usages des données. Ce sont deux éléments clés de la recherche en valorisation des données.

Pourquoi insister sur la définition de la valeur : c'est cette valeur (ces traitements de la valeur) affectée à la donnée qui va permettre d'établir les seuils des indicateurs qualité de données. On définit les règles permettant d'établir ces indicateurs en déterminant les dimensions qualité à prendre en compte pour les calculer. Les dimensions sont donc un support logique, une aide conceptuelle pour définir la qualité et interpréter les résultats de l'analyse[15].

4- Le rôle des dimensions qualité :

Les propositions de dimensions qualité de l'information de Harrathi & Calabretto[16] et les propositions de dimensions de qualité de données de Wang & Strong[17] ont comme point commun la répartition des dimensions en familles, hiérarchisées ou non. Cette répartition en familles permet de visualiser les contours des différentes dimensions pour ensuite choisir celles qui correspondent le mieux à l'activité dans l'entreprise considérée.

Quelques exemples de dimensions courantes :

- Complétude : du modèle de données autant que des données stockées, a-t-on bien toutes les représentations dont on a besoin ?
- Fraicheur : dimension temporelle à réviser selon les architectures : dans une architecture lambda (Nathan Martz pour Twitter) toutes les données ont un timestamp, elles sont donc toujours vraies car historisées : la dimension véracité est articulée aux dimensions temporelles.
- Cohérence : des données sont cohérentes entre elles si elles satisfont les règles syntaxiques et sémantiques qui leur sont associées : par exemple on peut mesure la cohérence entre un âge et une date de naissance[18].
- Quelques dimensions prennent de plus en plus d'importance dans les traitements big data, en particulier à cause des exigences liées à l'intégration des données multisources : la crédibilité, qui peut être un scoring de confiance sur la source, ou bien définie au moment de l'usage de la donnée, ainsi que la provenance, qui permet la traçabilité des données.

En conclusion :

- la manière dont ces dimensions sont définies permet d'établir des règles et des indicateurs qualité, et cette définition n'est porteuse de valeur que si elle est établie en amont avec les utilisateurs de la donnée – pour être en accord avec les usages et répondre aux besoins.

- ces dimensions sont interdépendantes, quand on en modifie une il faut évaluer l'impact que cela aura sur les autres.
Ex : si une donnée n'est pas suffisamment fraîche, elle est inexacte dans 70% des cas, inversement améliorer l'exactitude peut prendre du temps et nuire à la fraicheur des données[19].

Ce ne sont là que quelques éléments de définition de la qualité de données, centrées sur l'usage et la valeur de l'information (il y en a beaucoup d'autres), qui permettent d'appréhender les objets du digital. Si la gestion des master data demeure une exigence (il est impossible de facturer un client dont on n'a pas les coordonnées), les enjeux qualité de données s'articulent autour des notions d'enrichissement, de contextualisation et de mise en valeur des données.

Les projets MDM peuvent être des points d'entrée pour les projets DQM, car ces deux types d'engagement nécessitent à la fois des savoir-faire techniques et technologiques et la mise en œuvre de collaborations fortes dans l'entreprise. Aussi, la réussite des projets data, nécessaires à la transformation digitale des entreprises, est très dépendante des réflexions menées autour du développement de l'intelligence collective, de la gestion du changement et de la capacité des organisations à rendre opérantes des collaborations transversales en favorisant ces synergies digitales et collaboratives.

En la matière, des projets GED et DQM qui seraient définis autour de processus métiers communs auraient beaucoup à partager.

Notes et références :

[1] Pour une approche par les systèmes complexes voir notamment les travaux de l'Institut Praxeme, également : BARRAU Delphine et BIZINGRE, Joël : *Contribution PxData : Politique de la donnée, procédé et formalisation* **[en ligne]**. (Modifié le 06/04/2016). Disponible sur : < http://blog.conix.fr/> (Consulté le 06/04/2016)

[2] CROZAT, Stéphane, Chaînes éditoriales et rééditorialisation de contenus numériques **[en ligne]**, in Lisette Calderan, Pascale Laurent, Hélène Lowinger and Jacques Millet. *Le document numérique à l'heure du web*, ADBS, pp.179-220, 2012. (Sciences et techniques de l'information). Disponible sur : <https://hal.inria.fr/hal-00740268> (Consulté le 04/04/2016)

[3] CLÉMENT, Delphine, BEN HASSINE, Soumaya, COEUGNIET, Sébastien et al. : La gestion de données multi-sources : de la théorie à la mise en œuvre dans le cadre d'un référentiel client unique, in Laure Berti-Équille (dir.), *La qualité et la gouvernance des données au service de la performance des entreprises*, Lavoisier, Paris, 2012, pp.177-216.

[4] GOETZ, Michele: *MDM : highly recommended, still misunderstood* **[en ligne]**. (Modifié le 03 juin 2014) Disponible sur :<http://blogs.forrester.com/michele_goetz/ > (Consulté le 04/04/2016)

[5] TALBURT, John R. and ZHOU, Yinle : Entity Information Life Cycle for Big Data : Master Data Management and Information Integration, Morgan Kaufmann, 2015

[6] PEYRET, Henry : *Do not confuse Data Governance with Data Management* **[en ligne]**. (Modifié le 7 février 2016) Disponible sur : <http://blogs.forrester.com/henry_peyret/> (Consulté le 04/04/2016)

[7] SI-SAID CHERFI, Samira, AKOKA, Jacky, COMYN-WATTIAU, Isabelle : La qualité des modèles de données, in Laure Berti-Équille (dir.), *La qualité et la gouvernance des données au service de la performance des entreprises*, Lavoisier, Paris, 2012, pp.75-115.

[8] EVELSON, Boris: *Make data confidence part of your BI architecture* **[en ligne]**. (Modifié le 29 novembre 2012) Disponible sur : <http://blogs.forrester.com/boris_evelson/> (Consulté le 04/04/2016)

[9] HENDERSON, Xavier : *La qualité des données et le Big Data* **[en ligne, série de 4 articles]**. (Modifié en février 2014). Disponible sur : <http://www.gouvinfo.org/IAI/la-qualite-des-donnees-et-le-big-data/ (Consulté le 04/04/2016)

[10] Les dimensions de la qualité de données, fiches ExQi COMPRENDRE **[en ligne].** (Modifié en aout 2013). Disponible sur : <http://exqi.asso.fr/site/medias/> (Consulté le 04/04/2016)

[11] CLÉMENT, Delphine et LABOISSE, Brigitte : Création d'un référentiel d'indicateurs de mesure de la qualité des données CRM **[en ligne]**, in *Actes du 3e Atelier Qualité des données et Connaissances*, 23 janvier 2007, Namur, Belgique, pp. 5-14. Disponible sur : < https://conferences.telecom-bretagne.eu/qdc2007/> (Consulté le 04/04/2016)

[12] BERTHIER, Thierry et TEBOUL, Bruno : Valeur et Véracité de la donnée: Enjeux pour l'entreprise et défis pour le Data Scientist **[en ligne]** in *Actes du colloque «La donnée n'est pas donnée» ,* École Militaire–23 mars 2015. Disponible sur : <https://hal.archives-ouvertes.fr/> (Consulté le 04/04/2016)

[13] REDMAN, Thomas : *Data's credibility problem* **[en ligne].** (Modifié en décembre 2013) Disponible sur :< https://hbr.org/2013/> (Consulté le 04/04/2016)

[14] LANEY, Doug: *Applied Infonomics: Why and How to Value Your Information as an Asset* **[en ligne].** (Modifié le 15/10/2015) Disponible sur : < http://blogs.gartner.com/doug-laney/> (Consulté le 04/04/2016)

[15] Les dimensions de la qualité de données, op. cit

[16] HARRATHI, Rami et CALABRETTO, Sylvie : Un modèle de qualité de l'information, in *EGC'2006*, Lille. pp. 299-304.

[17] WANG, Richard Y. and STRONG, Diane M. : Beyond Accuracy: What Data Quality Means to Data Consumers, in *Journal of Management Information Systems, Vol. 12, No. 4*, Spring, 1996, pp. 5-33.

[18] Les dimensions de la qualité de données, op. cit.

[19] Les interdépendances de la qualité de données, fiche ExQi COMPRENDRE **[en ligne].** (Modifié en aout 2013). Disponible sur : <http://exqi.asso.fr/site/medias/> (Consulté le 04/04/2016)