Comment améliorer la qualité de ses données ?

Dans le prolongement de sa conférence sur la qualité des données organisée en février dernier au Ministère, **Isabelle Boydens,** Data Quality Expert chez Smals et professeure à l'ULB, nous explique ici les principales étapes à suivre pour assurer, de manière curative ou préventive, une meilleure qualité des données.

© Shutterstock

L'impact d'une mauvaise qualité des données est beaucoup plus sérieux qu'on ne l'imagine. Thomas Redman, expert mondialement reconnu dans ce domaine, a évalué l'impact financier de la mauvaise qualité des données en 2016 à 3.100 milliards de $ pour les seuls Etats-Unis, ce qui correspond à 20% du PIB américain ! Sans compter l'impact que cela peut aussi représenter en vies humaines dans certaines domaines (médical, environnemental, du transport, etc.).

Hélas, aucune organisation n'est vraiment à l'abri. On estime entre 5% et 30% le taux d'erreur dans les bases de données. Mais les organisations qui ont décidé de s'attaquer au problème et qui gèrent la qualité des données de manière systématique diminuent sensiblement le risque d'être un jour confrontées aux conséquences, parfois dramatiques, de données erronées.

Comment une organisation peut-elle améliorer la qualité de ses données ? Quelles sont les principales étapes à suivre ?

Isabelle Boydens : « La première étape consiste à identifier dans vos bases de données celles qui sont stratégiques et en quoi elles le sont, et d'y distinguer les champs ou procédures qui s'avèrent être les plus importants en fonction des objectifs.

On évalue ensuite l'impact d'une non-qualité de ces données stratégiques. Une fois le diagnostic posé, on met en place l'équipe chargée d'améliorer la qualité de ces données, sachant qu'elle devra compter des spécialistes IT mais aussi du domaine d'application. La démarche doit aussi obtenir le support du management. Cela ne peut pas se résumer à l'action isolée de quelques bonnes volontés.

L'organisation doit avoir à l'esprit les enjeux du problème et le budget qu'elle peut y consacrer. On va agir sur cette base-là pour être sûr que l'on peut agir dans la continuité.

On adoptera une approche curative du problème quand on est face à des bases de données qui contiennent des données de qualité inadéquate par rapport au but à atteindre, des données manquantes, incohérentes, des présomptions de doublons, etc. On mettra alors en place des mesures d'évaluation et d'aide à la correction des problèmes quand ils surviennent. »

Quels sont les outils disponibles pour mener ce type d'opération curative ?

« Pour le traitement curatif du problème, on utilise des **« data quality tools »**. Il s'agit d'outils spécialisés dans l'évaluation et le traitement de la qualité des données. Ces outils incluent principalement trois fonctionnalités : le data profiling, la standardisation des données et le data matching.

Le **data profiling** consiste à analyser la base de données (et, si elles existent, ses métadonnées) afin d'identifier, par exemple, les valeurs nulles ou non complétées, le nombre de valeurs incohérentes par rapport à d'autres valeurs, le nombre de champs surpeuplés, …

Utilisée pour des données très structurées (par exemple, un n° de compte bancaire ou une adresse postale), la **standardisation des données** permet de comparer la donnée par rapport à un modèle-type et d'apporter la correction requise le cas échéant. Cette fonctionnalité repose parfois sur des bases de connaissance spécialisées par type de données, notamment en ce qui concerne les adresses postales, par région, pays et continent.

Enfin, le **data matching** consiste à comparer entre eux sur un ou plusieurs paramètres donnés (par exemple, la phonétique d'un terme, les inversions de caractères, …) les enregistrements d'une ou de plusieurs bases de données pour identifier des présomptions de duplicats ou d'incohérences. L'opération demande des techniques spécifiques de gestion de la performance (« blocking » ou « windowing »). Au terme de cette opération, on est en mesure, sur la base de règles spécifiées avec le business et documentées,

« Quelle que soit l'approche suivie, curative ou préventive, la gestion de la qualité des données implique de travailler en synergie avec les personnes du terrain et en contact étroit avec le domaine d'application pour lequel la base de données a été conçue »

d'identifier un « golden record », un enregistrement de référence.

Mais si on se limite à une approche strictement curative, on va passer son temps à corriger *ad infinitum* des problèmes qui vont sans cesse ressurgir. Il faut donc aussi mettre en place des mesures préventives qui vont identifier la source de ces cas problématiques. »

En quoi consistent ces mesures préventives ?

« Le traitement préventif dans la qualité des données consiste à repérer, via une extension spécifique du modèle initial de la base, les anomalies stratégiques qui peuvent survenir dans l'utilisation d'une base de données, à établir l'historique de ces anomalies et de leur traitement de façon à en assurer le suivi sur le long terme et à y remédier à la source. Cette opération s'appelle le **« back tracking »**. Elle implique d'identifier tous les processus qui alimentent la base de données (il peut y en avoir des centaines) depuis le moment où la donnée est envoyée jusqu'au moment où elle arrive dans la base de données. On remonte ainsi, en collaboration avec les expéditeurs de l'information et ses gestionnaires, étape par étape jusqu'à la source (d'où le terme **« back tracking »**) et on s'arrête dès qu'on a trouvé la ou les causes et solutions structurelles à l'origine des anomalies (erreur d'interprétation, directive mal documentée, etc.). Une fois les solutions mises en place, on a pu constater que le nombre d'anomalies diminuait drastiquement. Un suivi, dont l'effort est dégressif dans le temps, est toutefois nécessaire en raison de l'évolution inévitable de l'environnement de la base. »

Avez-vous eu l'occasion d'appliquer cette approche sur le terrain ?

« Avec mon équipe, j'ai appliqué cette approche à l'analyse de la base de données « DMFA » *(ndlr. Déclaration multifonctionnelle qui remplace la déclaration ONSS trimestrielle depuis 2003).* Conformément au principe de Pareto qui dit qu'un grand nombre de problèmes sont produits par un petit nombre de causes, nous avons pu observer, lors de plusieurs de ces opérations, que sur les quelques 240.000 employeurs enregistrés dans la base de données DMFA, seulement 50 étaient à l'origine de 80% des anomalies prioritaires. Après plusieurs tests concluants, nous avons alors mis en place une opération structurelle qui est passée dans la loi sous la forme d'un arrêté royal en février 2017.

Depuis février 2017, cette méthode appliquée à la DMFA est inclue dans le baromètre de qualité pour l'ensemble des secrétariats sociaux qui gèrent environ 90% des DMFA en Belgique. C'est une méthode généralisable à tout domaine d'application empirique (médecine, environnement, …).»

Que retenir comme bonne pratique avant de s'engager dans ce type d'opération ?

« Quelle que soit l'approche suivie, curative ou préventive, la gestion de la qualité des données implique de travailler en synergie avec les personnes du terrain et en contact étroit avec le domaine d'application pour lequel la base de données a été conçue (ce sont les services en première ligne qui seront contactés si un problème survient suite à des données défectueuses). Dès le début de la démarche, il faut concevoir l'organisation ad-hoc, identifier les rôles techniques et métiers, les bases de données stratégiques, les anomalies les plus graves, les scénarios de détection et de correction. Une fois que l'on a défini et documenté tout cela, on peut se lancer. »

Propos recueillis par Philippe du Busquiel,
Direction Communication

À quoi sont dus les problèmes de « non qualité » des données ?

L'apparition de problèmes dus à la « non qualité » des données tient principalement aux facteurs suivants :

1. **Une vision à « court terme »** lors de la conception d'un projet, l'accent étant trop souvent exclusivement porté sur les aspects purement techniques, au détriment de l'analyse du domaine d'application qui est négligée (en témoignent les problèmes qu'a connue la mise en œuvre de la réforme « Obamacare » aux USA en 2013 : les blocages du portail fédéral étaient dus à une analyse insuffisante de la complexité du domaine assurantiel). Ce n'est d'ailleurs que depuis peu que la communauté IT s'intéresse de près à la question de la qualité de l'information.

2. Une attention insuffisante accordée :

   A. aux usages et au partage des données (l'adage *« use it or lose it »* illustre le fait que la qualité de données peu utilisées et peu partagées se détériore au fil du temps);
   B. à la documentation des données et des processus;
   C. à la gouvernance des données sur le long terme, pourtant indispensable en raison de la complexité de nombreux domaines d'application empiriques évolutifs (pensons aux domaines législatifs, médicaux, scientifiques, …);
   D. à la génération d'une redondance non contrôlée d'information, faute de source authentique, au sein d'une même entité : le concept de *« ghost factory »* (usine fantôme) désigne le temps et l'argent consacrés par une entreprise à produire des défauts et à les corriger…

(source : « Dix bonnes pratiques pour améliorer et maintenir la qualité des données », Isabelle Boydens, 2014).