



[Blog](#) [Talks](#) [Publications](#) ▾ [Tools](#) ▾ [Radar / Plan](#) ▾ [Team](#) [About](#)

Dix bonnes pratiques pour améliorer et maintenir la qualité des données

Posted on 2014-06-16 by [Isabelle Boydens](#)

(dernière mise à jour : décembre 2021)

Les bases de données se prêtent aux métaphores financières. Ne les désigne-t-on pas souvent par le terme « banques de données » ? Elles évoqueraient ainsi un capital d'information sur lequel on peut faire des retraits à la demande. A condition que le compte soit correctement approvisionné...(*)

Vu l'actualité des enjeux soulevés, dès lors que l'information est un instrument d'action sur le réel, nous envisageons successivement ici les coûts de la "non qualité" des données, leurs causes les plus fréquentes et ensuite dix bonnes pratiques en vue d'améliorer et de maintenir dans le temps la qualité de l'information (**).

Les coûts de la « non qualité des données »

En 2016, T. Redman chiffrait ainsi les coûts de la "non qualité" aux USA, dans son livre "*Getting in front on Data*" (p. 25) : "**\$3,1 Trillions/year in the US, which is about 20 percent of the Gross Domestic Product.**"

Les chiffres précis et récents sont rares : pour des raisons d'image, les entreprises ne les communiquent pas volontiers publiquement ... On estime que de nombreuses bases de données dans les secteurs financiers, bancaires, médicaux ou administratifs incluent en moyenne 10 % d'anomalies formelles sur la totalité des valeurs répertoriées. La qualité des données désigne leur adéquation relative aux usages et objectifs attendus (on parle de « *fitness for use* »). Elle relève toujours d'un

arbitrage de type « coût-bénéfice ». Ces coûts peuvent être ventilés en terme :

- de frais de *manpower* de vérification de l'information et de traitement des incohérences;
- de traitement des plaintes, procès et réparation éventuelle des préjudices, en cas où des données dotées d'un statut juridique de force probante occasionnent un préjudice à une autre partie ;
- d'investissements humains et techniques lors d'un reengineering ou de l'introduction d'une nouvelle technologie.

Ils incluent également des paramètres critiques non mesurables :

- pertinence du service rendu (application de la législation, soin apporté aux patients, par exemple);
- crédibilité;
- stratégie à long terme.

On pourra chiffrer plus précisément les coûts et le ROI liés à une approche "qualité" lorsque l'on précise le topic abordé, comme dans l'étude "Email address reliability".

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

L'émergence toujours actuelle de problèmes dus à la « non qualité » 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 purement IT s'intéresse de près à la question de la qualité de l'information.
2. Une attention insuffisante accordée :
 - 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)

- à la documentation des données et des processus
- 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, ...)
- 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..

Dix bonnes pratiques

En raison de l'importance de la problématique, nous rappelons dix bonnes pratiques afin d'améliorer et de maintenir la qualité des données dans le temps.

1. **Définir les objectifs et usages des données en fonction des enjeux selon le principe du « *fitness for use* »** évoqué plus haut : dans certains cas, une tolérance à l'erreur sera acceptable (exploitation marketing, statistique, ...) alors que dans d'autres, l'ensemble du système d'information devra être traité avec la plus grande précision (en cas d'impact juridique, médical, financier, ...). Dès que l'on se penche sur un domaine d'application « grandeur nature », ce travail de définition s'avère complexe et demande des choix et des arbitrages explicites. En effet, dans tout domaine d'application empirique (sujet à interprétation humaine), le système d'information est susceptible d'évoluer dans le temps avec l'interprétation des valeurs qu'il permet d'appréhender (ce sera le cas des nomenclatures des catégories d'activités d'une entreprise, par exemple, sachant que l'univers socio-économique évolue de manière continue).
2. **Etablir une organisation pluridisciplinaire impliquant le management, des spécialistes du domaine d'application et des informaticiens en charge du suivi transversal de la qualité de l'information** (à travers toutes les bases de données et processus inclus dans l'entreprise et impactés par les mêmes concepts). Cette organisation doit être **souple et flexible**. Son ampleur varie **en fonction des ressources et des enjeux du domaine**

- d'application.** La mise en place d'une organisation trop lourde ne sera pas suivie d'effet et sera contre-productive.
3. Une fois les objectifs définis et l'organisation mise en place, **identifier les flux d'alimentation, les processus, les champs les plus critiques au sein du système ainsi que les événements principaux susceptibles de les affecter** (suppression, modification de définition, ...) ; l'identifiant unique d'une entreprise ayant plus d'importance que son numéro de fax, par exemple. Une fois ces éléments identifiés, c'est sur ceux-ci que les efforts seront concentrés dans un premier temps.
 4. **Distinguer les bases de données de gestion des sources authentiques** et dans le cadre d'une **gestion collaborative des anomalies**, se référer prioritairement à la source authentique et la traiter en premier lieu pour tous les concepts stratégiques identifiés (voir : J. Bizingre, J. Paumier et P. Rivière, "Les référentiels du système d'information. Données de référence et architecture d'entreprise", Paris, Dunod, 2013).
 5. **Etablir pour ceux-ci un ensemble d'indicateurs de qualité quantifiables** (par exemple, nombre de valeurs absentes pour un champ, nombre de valeurs incohérentes en comparaison avec une source authentique, ...), **dont l'historique sera associé à la base de données** et inclura la prise en compte du suivi du traitement des anomalies formelles (valeurs violant les contraintes d'intégrité de la base de données). A ces indicateurs, permettant d'assurer le suivi de la qualité, il faudra **associer des objectifs à atteindre en fonction des enjeux ainsi que des mesures d'amélioration potentielles** pour remédier aux problèmes « types » connus (par exemple, mieux documenter les instructions de saisie de la base de données quand des problèmes d'interprétation sont à la source d'erreurs massives). Ces mesures pourront s'enrichir avec le temps, en fonction de l'évolution du domaine et des objectifs stratégiques.
 6. **Prendre conscience que la définition des indicateurs, des objectifs à attendre et des solutions associées repose sur des compromis** (tel que l'arbitrage entre la fiabilité de l'information et sa rapidité de sa diffusion, ou encore, entre la précision et l'exhaustivité des données, ...).
 7. **Mettre en place des stratégies de gestion continues transversales reposant sur ces indicateurs et objectifs :**

- 1. le suivi des anomalies formelles permet par exemple de détecter, dans les domaines d'application empiriques fortement évolutifs, l'émergence de nouveaux phénomènes observables demandant une adaptation régulière des contraintes d'intégrité et du schéma** de la base de données en vue de diminuer le nombre d'anomalies fictives à traiter. Ainsi, en Belgique, lors de la mise en place d'une directive administrative en faveur du secteur non marchand, la question s'est posée (au regard de la réalité qui avait été progressivement appréhendée au sein de la base) de savoir s'il fallait inclure dans ce secteur les maisons de repos privées, qui en étaient a priori exclues du fait de leur finalité lucrative. Initialement considérées comme des cas erronés au regard du domaine de définition spécifiant le secteur non marchand, ces entreprises y ont finalement été intégrées après interprétation juridique (ce qui a impliqué une restructuration du schéma de la base de données). **Dans ce cas, la restructuration d'une base de données résulte d'une décision humaine tendant à rendre le modèle conforme (au moins transitoirement) aux nouvelles observations. En l'absence d'une telle intervention, l'écart entre la base de données et le réel se creuserait.** En effet, si l'on omet d'adapter le schéma, les anomalies correspondant à ces cas vont continuer d'apparaître et devenir de plus en plus nombreuses, nécessitant un examen manuel potentiellement lourd et susceptible de ralentir considérablement le traitement des dossiers administratifs. Pour la sécurité sociale belge, la mise en œuvre de cette méthode a permis d'améliorer la précision et la rapidité du traitement des cotisations sociales en **réduisant potentiellement de 50 % le volume des anomalies formelles.**
- 2. une méthode originale de « back tracking »** (inspirée du « data tracking » de Thomas Redman) permettant, sur la base d'un échantillon d'anomalies représentatives, d'en détecter l'origine à la source et d'y **remédier structurellement** (erreur de programmation, d'interprétation de la loi, ...). Les expériences menées à ce propos ont donné le jour à un **ROI** très important ainsi qu'à la parution d'un **Arrêté Royal**, le 2/2/2017, assurant

la mise en place de la méthode dans le secteur de la sécurité sociale belge. **Voir en particulier l'article suivant publié en juillet 2021 dans une revue scientifique à Paris et synthétisant toute l'approche Data quality depuis ses débuts incluant les dernières nouveautés de décembre 2021 sur les "data quality tools"** : Boydens I., Hamiti G. et Van Eeckhout R., *Un service au cœur de la qualité des données. Présentation d'un prototype d'ATMS*. In Le Courrier des statistiques, Paris, INSEE, juin-juillet 2021, n°6, p. 100-122.

8. **Pour les bases de données « legacy » incluant des erreurs (présomptions de duplicats, adresses nationales ou internationales incohérentes, ...) ou les projets de migration, recourir à une approche curative via les data quality tools (et selon les besoins, à leurs fonctionnalités de profiling, standardisation ou matching), on line (via une REST API) ou en batch, en vue de la détection semi automatique des présomptions de doubles et d'incohérences ainsi que de leur traitement. En particulier, il est conseillé de recourir tôt et fréquemment au *data profiling*. Il s'agit d'une approche complémentaire à l'approche "préventive" du back tracking, vue au point précédent.**
9. **Documenter la base de données et les indicateurs dans le temps, dont les définitions sont validées via un workflow d'approbation (dans le cadre des solutions de "case management")**. Un système de documentation performant repose sur :
 1. une organisation transversale (recourant au Master Data Management et éventuellement, à des recommandations européennes, telles que ISA, Interoperability Solutions for European Public Administrations);
 2. un workflow de validation (avec prise en compte du multilinguisme);
 3. une gestion des versions et un historique des modifications;
 4. une conception aussi économe que possible recourant au principe de l'héritage de façon à minimiser l'ampleur du travail de mise à jour manuel et le risque d'erreur;
 5. le principe du « WOPM » « *Write Once Publish Many* » de façon à exploiter chaque mise à jour à des fins opérationnelles (en vue de l'adaptation des contraintes d'intégrités et des tables de

- référence) et documentaires (en vue de la diffusion de ces mêmes mises à jour sous des formats lisibles par l'être humain);
6. une documentation du processus de traitement des anomalies en vue de s'assurer que la base de données est mise à jour de manière homogène et cohérente.

10. Assurer une formation continue des gestionnaires et utilisateurs de la base de données, tirer les enseignements des difficultés rencontrées et communiquer les *success stories*.

Une approche en vue d'évaluer, d'améliorer et de maintenir la qualité d'une base de données est nécessairement continue, itérative et sujette à adaptation dans le temps. Recourir aux dix bonnes pratiques que nous venons d'évoquer, en fonction des enjeux de l'information et des budgets disponibles, permet assurément de tendre vers cet objectif et d'en mesurer les effets.

(*) Extrait de : Boydens I., "Les bases de données sont-elles solubles dans le temps ?". In "La Recherche", Hors série n°9, Paris, 2002, p. 32-34.

(**) On trouvera la définition précise de concepts importants évoqués dans ce post ("données", "base de données", ...) dans plusieurs de nos publications antérieures, telles que : Boydens I. "Strategic Issues Relating to Data Quality for E-government: Learning from an Approach Adopted in Belgium". In Assar S., Boughzala I. et Boydens I., eds., "Practical Studies in E-Government : Best Practices from Around the World", New York, Springer, 2011, p. 113-130 (chapter 7). Voir aussi : Boydens I., Informatique, normes et temps. Bruxelles, Bruylant, 1999. Boydens I., "L'océan des données et le canal des normes", In Carrieu-Costa M.-J., Bryden A. et Couveinhes P., eds, Les Annales des Mines, Series "Responsabilité et Environnement" (thematic issue : "La normalisation : principes, histoire, évolutions et perspectives"), Paris, n° 67, Juillet 2012, pp. 22-29

data quality

data quality tools

information management

MORE POSTS

Je data beschermen tegen beheerders: 'on-premise' Confidential Computing

2026-03

Protéger ses données des administrateurs : l'informatique confidentielle « on-premise »

2026-03

De performance van LLM's: Een vergelijkende analyse tussen Frans en Nederlands

2026-03

Made by Smals Research – Privacyvriendelijk Kruisen van Persoonsgegevens

2026-02

Search

Search

Newsletter & webinars:

Dutch French

Subscribe

Keywords:

[analytics](#) [artificial intelligence](#) [big data](#) [blockchain](#) [bpm](#) [chatbot](#)
[cloud computing](#) [cost cutting](#) [cryptography](#) [data center](#)
[data quality](#) [development](#) [eda](#) [egov](#) [event](#) [gis](#)

information management machine learning managing it costs
methodology mobile natural language processing open source privacy
productivity security social software design
software engineering standards

Smals Research

© Smals Research – License/Disclaimer: [FR](#) / [NL](#)