



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

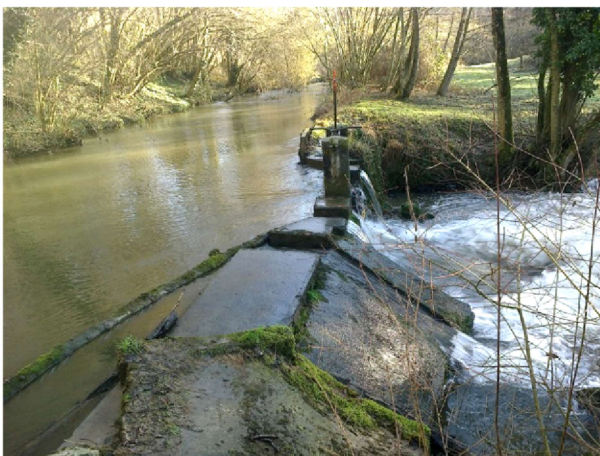
Data Quality & « back tracking » : depuis les premières expérimentations à la parution d'un Arrêté Royal

Posted on 2018-05-14 by [Isabelle Boydens](#)

Thomas Redman compare une base de donnée à un lac, alimenté par des flux aquatiques continus. La métaphore illustre l'approche qui sera évoquée dans ce blog en vue d'améliorer la qualité des données.

1. Les enjeux de la qualité des données : rappel et exemple

En effet, nettoyer "à l'infini" le fond du lac (via des algorithmes de "data cleansing") n'est pas efficace , même s'il faut parfois recourir ponctuellement à cette méthode à moyen terme, car des flux de qualité douteuse continueront sans cesse d'arriver. Il s'agira plutôt d'adapter structurellement les flux d'information et processus qui alimentent la base de données, de manière à remédier aux problèmes identifiés, à la source et durablement.



*« A Database is like a lake...
The stream or streams feeding the lake represent information chains that*

create the data ... »

Thomas Redman, 2001, 2018 (1)

La qualité d'une base de données désigne son adéquation relative aux usages ("fitness for use"), lesquels évoluent dans le temps. La qualité totale n'existe pas. On se trouve nécessairement face à un compromis d'ordre pratique, sous contrainte de budget.

Les enjeux sont d'autant plus importants que la base de données est un instrument d'action sur le réel ; à l'instar des domaines administratifs, environnementaux ou énergétiques, par exemple. Le blog que nous avons publié en 2014, "Dix bonnes pratiques pour améliorer et maintenir la qualité des données", en témoigne.

Ainsi en 2013, le portail gouvernemental destiné à soutenir la politique de l'Obamacare a-t-il connu de graves problèmes d'accès et de qualité durant quelques semaines. Ceci fut dû, entre autres, à un manque de prise en compte des spécifications conceptuelles liées au domaine assurantiel. Ces lacunes ont engendré des pertes d'information en conséquence et de là, un impact évident sur la qualité des données, sur le plan financier, sur l'action politique et sur la crédibilité de celle-ci.

Parmi les méthodes en vue d'améliorer et de maintenir la qualité des données, nous avons évoqué à plusieurs reprises la technique du *data tracking* de T. Redman (1), que nous avons adaptée de manière originale et généralisable au secteur de la sécurité sociale belge, sous l'appellation de *back tracking*. En 2012, un post (Améliorer la qualité de l'information : du stemma codicum au data tracking) et une note de recherche (Data Tracking : le « Return On Investment » de l'analyse des flux d'information) y font référence.

Nous y revenons en 2018 car le succès des expérimentations menées à grande échelle sur le terrain a donné lieu à la parution d'un Arrêté Royal publié le 02/02/2017 qui en généralise la portée à l'échelle de la Belgique.

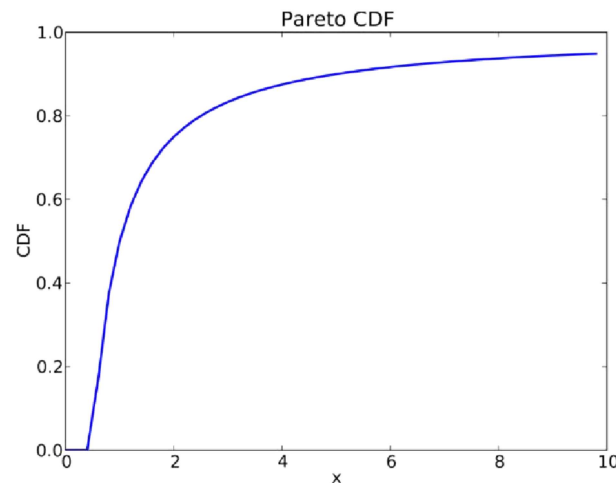
2. Le *back tracking* : rappel des principes et conclusions des premières expériences à grande échelle

Nous rappelons ci-dessous les grands principes de la méthode du *data tracking* de Thomas Redman (2.1.) et ceux du *back tracking*, reposant sur plusieurs innovations apportées par la Data Quality Cel de Smals (2.2.).

2.1. Le data tracking de Thomas Redman (AT&T Labs)

Le *data tracking* proposé par Thomas Redman d'AT&T Labs aux USA à la fin des années 1990 vise à évaluer quantitativement la validité formelle des valeurs introduites dans une base de données et à en améliorer structurellement le traitement. Une base de données s'apparente à un lac, selon Redman. Au lieu de nettoyer ponctuellement le fond du lac continuellement alimenté par des flux et courants externes (comme le préconise le "*data cleansing*", méthode de correction automatique), Redman propose, sur la base d'un échantillon aléatoire de données prélevé en entrée, d'analyser méthodiquement les processus et les flux permettant l'assemblage des données dans la base. Le but final consiste à déterminer les causes des erreurs formelles afin d'y remédier structurellement à la source.

L'opération repose sur l'hypothèse selon laquelle un petit nombre de flux, processus ou pratiques sont à l'origine d'un pourcentage important d'erreurs formelles (ou anomalies). L'approche fait référence au principe empirique de Pareto également appelé « principe 80/20 » : une part importante des cas problématiques (environ 80 %) est engendrée par environ 20 % des causes possibles.



2.2. Les apports originaux du back tracking

Nous avons enrichi cette méthode sur cinq aspects importants (en nous basant toujours sur le principe de Pareto) :

- Le **modèle de la base de données est étendu** et relié à un historique des violations de *business rules* (anomalies) et de leur traitement ;

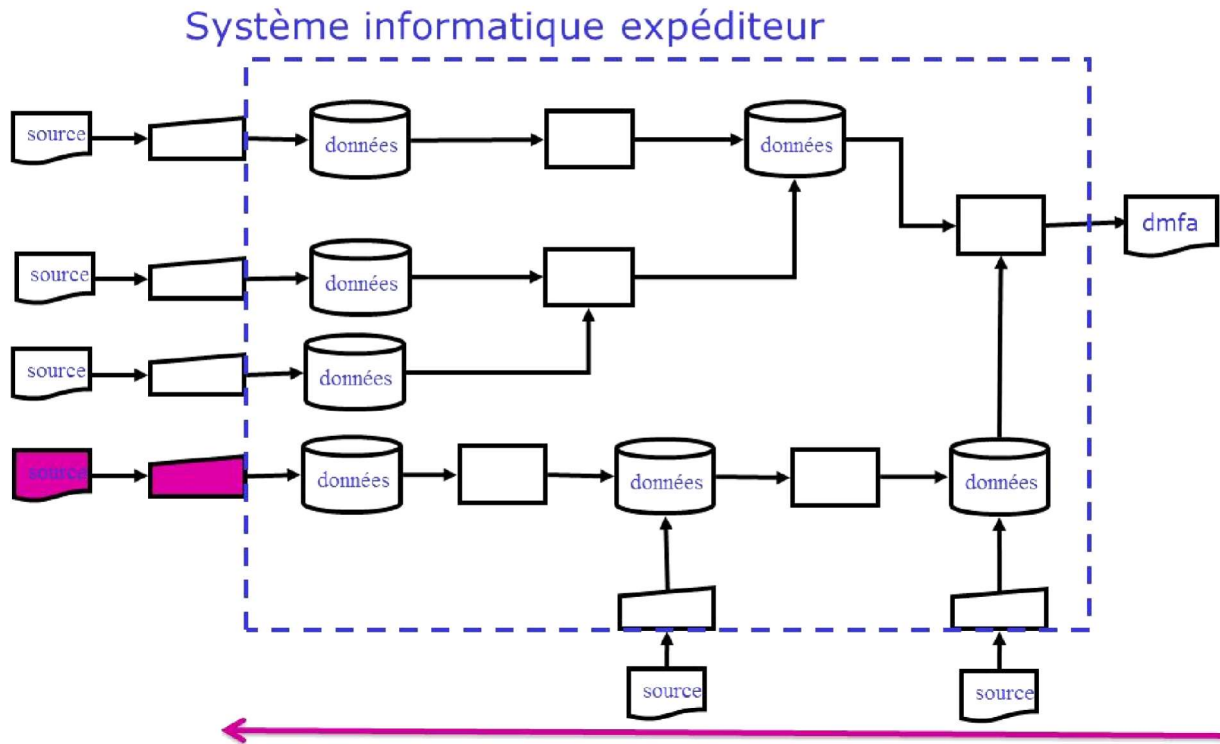
- un **monitoring des cas jugés les plus stratégiques, de façon à faciliter la gestion** de la qualité de la base de données est mis en place. Le suivi des anomalies formelles permet par exemple de détecter, dans les **domaines d'application empiriques** fortement **évolutifs**, l'émergence de nouveaux 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 adaptation 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.
- **au-delà de l'erreur formelle, les questions d'interprétation des données** au fil de l'évolution de la législation et des réalités appréhendées sont également abordées. L'opération se réalise en interaction avec tous les intervenants : fournisseurs de l'information, gestionnaires de la base de données et spécialistes du domaine.
- l'échantillon d'individus et de cas retenus **n'est pas aléatoire** puisque l'on dispose d'une connaissance *a priori* concernant les dossiers problématiques (via l'historique des violations de

business rules et de leur traitement), ce qui permet une sélection précise et plus exhaustive des cas problématiques dès le début de l'opération.

- il s'agit **d'un *tracking arrière* (ou *back tracking*)** : on part de la situation finale pour revenir, étape par étape, en synergie avec les fournisseurs et les producteurs de l'Information, à chaque source et processus qui en a permis l'élaboration, jusqu'à l'identification des causes à la source de cas problématiques. L'objectif est d'éviter le traitement de données ou de flux inutiles pour l'analyse et de travailler de manière plus économe, renforçant le ROI de l'opération. L'objectif est d'éviter le traitement de données ou de flux inutiles pour l'analyse et de travailler de manière plus économe, renforçant le ROI de l'opération. En effet, la recherche des origines structurelles des anomalies prend fin dès que toutes leurs causes par type ont été détectées, sans que tous les flux ne soient inutilement parcourus.

La base de données DmfA, fut le « case study » privilégié vu son ampleur : elle permet en effet en 2017 le prélèvement et la redistribution annuels de 65 milliards d'euros à l'échelle du pays. Mais suite à la complexité et la mouvance du réel observable et de la législation, la gestion des flux administratifs inclut de nombreuses présomptions d'anomalies qu'il est coûteux et difficile de traiter.

« Back tracking » (DmfA)



Chaque trimestre, 10% environ des valeurs transmises doivent être analysés et traités intellectuellement, comme dans la plupart des vastes systèmes d'information empiriques (ainsi est-ce le cas des systèmes bancaires, par exemple). L'extension du schéma de la base de données et le monitoring des violations de « business rules » (ou présomptions d'erreurs) et de leur traitement a permis d'analyser ces cas, avec l'aide d'informaticiens et de juristes, et de remarquer que la majorité de ces informations potentiellement erronées sont récurrentes.

Comme la législation (en matière de sécurité sociale notamment) et la réalité observable (entreprises, secteurs d'activités, ...) changent fréquemment et sont sujets à interprétation, ainsi qu'en témoignent les difficultés de catégorisation des métiers issus de « l'ubérisation », par exemple, on se retrouve parfois face à des situations problématiques (cas inédits, définitions difficiles à interpréter, ...).

La majorité des présomptions d'erreurs proviennent de ce décalage inévitable entre l'évolution de la réalité du terrain, celle de l'administration, des systèmes informatiques et, enfin, de la législation.

L'opération de *back tracking* consiste à détecter, chez l'expéditeur et en partenariat avec celui-ci et l'administration, les éléments à l'origine de la production d'un grand nombre de présomptions d'anomalies systématiques (traitement inadéquat de certaines sources de données, interprétation inadéquate de la législation, erreurs de programmation, etc.). Sur cette base, un diagnostic ainsi que des actions correctrices peuvent être posés (correction de code formel dans les programmes, adaptation de l'interprétation d'une loi, clarification de la documentation administrative ...). L'opération permet notamment :

- d'établir un **partenariat** avec les fournisseurs de l'information en vue d'en améliorer la qualité dans l'intérêt de tous ;
- de mettre en place des **solutions structurelles d'amélioration peu coûteuses**, ne nécessitant aucun développement logiciel d'envergure; les coûts sont en plus dégressifs si l'opération est récurrente ;
- d'accélérer et d'améliorer le calcul précis des cotisations et prestations sociales à destination des assurés sociaux, **incluant un ROI avec une diminution de plus de 50 % du nombre d'anomalies à traiter et de l'effort humain spécialisé associé.**
- d'obtenir des **résultats potentiellement durables**, puisque la cause structurelle des anomalies systématiques, structurelles ou émergentes, est pratiquement identifiée (qu'il s'agisse d'erreurs de programmation ou de problèmes d'interprétation de la législation en matière de temps de travail, par exemple) et peut être théoriquement définitivement réglée.

Toutefois, ce dernier point n'est valable que tant que les conditions "externes" demeurent constantes. Or, toute base de données empirique s'inscrit nécessairement dans un environnement ouvert et changeant au sein duquel l'interprétation de la base évolue avec le traitement des valeurs qu'elle permet d'appréhender.

Pour cette raison, il est conseillé de relancer de manière régulière l'opération de *back tracking* en vue de :

- s'assurer de la permanence des résultats obtenus;
- détecter d'éventuelles nouvelles sources d'anomalies et d'y remédier.

3. L'arrêté Royal du 02/02/2017 : le « back tracking » au cœur d'un baromètre de qualité, régi par la Loi

Depuis 2006, les tests « grandeur nature » menés également en 2012 et réalisés en synergie avec les développeurs et les responsables de l'administration spécialistes du domaine d'application ainsi que les expéditeurs de l'information furent concluants.



Vu les résultats probants en terme de partenariat et de ROI, en 2016, le Conseil des Ministres a donné son feu vert au projet d'Arrêté Royal du 2 février 2017 – voir ci-dessous, "Références" (2) – instaurant une base légale pour le baromètre de qualité destiné aux Secrétariats Sociaux Agréés (SSA) et ultérieurement à tous les prestataires de services.

Le baromètre inclut la méthode du *back tracking* proposée qui y joue un rôle central s'agissant du traitement des anomalies. Testée en 2016 avec plusieurs SSA et avec l'aval de l'Union des Secrétariats Sociaux Agréés, la méthode du *back tracking* est ainsi légalement généralisée depuis l'Arrêté Royal du 2 février 2017 et mise en production, faisant l'objet d'un suivi régulier.

Dans la foulée, l'Université libre de Bruxelles a publié le 4 mai 2017 un communiqué de presse sur la question dans la mesure où la méthode est généralisable à d'autres domaines d'application empiriques (médecine, environnement, énergie, ...).

En décembre 2018, cette méthode a été présentée dans la revue "Le Courrier des Statistiques" de l'Institut National des Etudes et des statistiques français (3).

Références

(1) Thomas Redman, Bell Labs, on DataQualitySolutions, 2018 ("*Data Quality. The field Guide*", Digital Press, 2001, p. 53).

(2) Arrêté Royal – Koninklijk besluit (source : Moniteur belge – Bron : Belgisch Staatsblad) [Link](#)

2 FEBRUARI 2017. – Koninklijk besluit van 2 februari 2017 houdende wijziging van hoofdstuk IV van het koninklijk besluit van 28 november 1969 tot uitvoering van de wet van 27 juni 1969 tot herziening van de besluitwet van 28 december 1944 betreffende de maatschappelijke zekerheid der arbeiders. Publicatie : 2017-02-20. Numac : 2017200919
« ... *De barometer laat toe de problemen en pijnpunten zichtbaar te maken in het permanent proces van verbetering van de algemene kwaliteit en volledigheid van de aangiften. De barometer is tevens een monitoring instrument. Door de problemen in kaart te brengen is het ESS in staat de nodige stappen te ondernemen tot aanpassing van de werkmethodes. De wederkerende kwaliteitscontroles geven hen de mogelijkheid de vooruitgang te bewaken/controleren...* »

2 FEVRIER 2017. – Arrêté royal du 2 février 2017 modifiant le chapitre IV de l'arrêté royal du 28 novembre 1969 pris en exécution de la loi du 27 juin 1969 révisant l'arrêté-loi du 28 décembre 1944 concernant la sécurité sociale des travailleurs. Publication : 2017-02-20. Numac : 2017200919 –
« ...*Le baromètre permet de faire apparaître les problèmes et points névralgiques dans le processus permanent d'amélioration de la qualité générale et de complétude des déclarations. Le baromètre est également un instrument de monitoring. En procédant à une cartographie des problèmes, le SSA est en mesure d'entreprendre les démarches nécessaires en vue de l'adaptation des méthodes de travail. Les contrôles de qualité récurrents lui donnent la possibilité de surveiller/contrôler les progrès réalisés...* »

(3) Pascal Rivière., Utiliser les déclarations administratives à des fins statistiques. In *Le courrier des statistiques*, Paris, INSEE, décembre 2018, n°1, p. 14-23.

back tracking

data quality

data tracking

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

Your email address

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](#)