



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

## "Mapping the World of Data Problems" : la qualité des données vue par la communauté IT

Posted on 2013-04-03 by [Isabelle Boydens](#)



En novembre 2012, O'Reilly Media a édité un "livre-événement" en matière de "data quality" : [Q. E. McCallum, Bad Data Handbook, Mapping the World of Data Problems](#), O'Reilly Media, 2012, 246 p.

Cet ouvrage collectif sur la qualité des données est inédit car il émane exclusivement de la communauté des *web software developers* (Python, Perl script, Parallel R, NLP, cloud computing, ...), *web predictive analytics et architectes IT* ... Il compte même un *hacker* parmi ses co-auteurs. Ces auteurs n'avaient a priori aucune prédilection pour l'étude des données : « *In fact, I dare say that I don't quite care for data* » (p. 1). Mais, quotidiennement affectés par les problèmes de data quality dans leur job, ils ont programmé une pause entre deux lignes de code pour partager leur longue et douloureuse expérience dans les domaines d'application les plus variés : "*Bad Data .... include data that eats up your time, causes you to stay late at the office, drives you to tear out your hair in frustration. It's data that you can't access, data that you had and then lost, data that's not the same today as it was yesterday...*" (p. 1).

En soi, les principaux apports pratiques de cet ouvrage, en ce qui concerne le thème "Database Quality", sont déjà connus par certains (*"The ideas presented here are born from (often painful) experience and are likely not new to anyone who has spent any extended time looking at data"*, p. 226). Ils sont par exemple plus largement intégrés dans l'approche opérationnelle du Data Quality Competence Center de Smals (voir le data tracking, la gestion intégrée des anomalies, le recours aux "Data Quality Tools", la documentation du système ou encore, la mise en place d'une organisation). S'agissant de l'egovernment, nos travaux sont synthétisés dans un ouvrage coédité à New York chez Springer en 2011 et dans un article paru aux Annales des Mines à Paris en 2012 : ils placent la question de l'évolution de l'information dans le temps au coeur de la réflexion conceptuelle, appliquant la critique historique aux sources informatiques à des fins opérationnelles en termes de coûts-bénéfices et de gestion.

Nous présentons toutefois ici un aperçu de ce "Bad Data Handbook" et des catégories de questions qu'il aborde car il comporte au moins **quatre aspects très intéressants et, en soi, particulièrement innovants** :

- les très nombreux **cases studies** présentés sont extraordinairement riches, inédits et variés dans des domaines d'applications stratégiques (police criminelle, marchés financiers internationaux, chimie urologique, egov, ...);
- c'est la **première fois que la communauté "geek"** des développeurs & architectes IT aborde la question "data quality", sujet sur lequel elle ne publie en général jamais, se concentrant essentiellement sur la complexité technique, algorithmique et mathématique;
- on y trouve une **reconnaissance des impacts financiers énormes** que suscite l'inadéquation des données aux usages ("non qualité") : *"For large enterprises, this could be a multi-million dollar problem"* (p. 163);
- sans aucune référence bibliographique explicite, plusieurs auteurs font preuve **d'une finesse d'analyse et d'une acuité assez impressionnantes sur le plan épistémologique** (certains d'entre eux ont fait leur thèse de doctorat en physique théorique, ce qui explique sans doute que K. Popper ne leur soit pas étranger).

Les apports de l'ouvrage retenus sont ici structurés en deux catégories logiquement liées et utiles non seulement pour les développeurs IT et les architectes mais aussi, la communauté des bases de données, les décideurs et utilisateurs finaux

## A. "Data format, storage & infrastructure" : 5 pistes pour faciliter l'accès aux données



Avant d'aborder la qualité de l'information, ... il s'agit d'abord d'accéder physiquement et logiquement aux données. Or, notre longue expérience en "data profiling" le confirme, c'est souvent l'étape la plus fastidieuse.

Ceci est encore plus vrai dans le cadre du Web, espace ouvert, dynamique et non contrôlé : "*in some (regrettably rare) cases, all the information about the data is provided*" (K. Fink, p. 9); "*the first, and sometimes, hardest part of doing any data analysis is acquiring the data from which you hope to extract information*" (A. Laiacano, p. 69). Ceci amène les auteurs à s'interroger sur l'opacité des Media sociaux dont l'étude soulève de nombreux défis (P. Warden, How to Feed and Care for Your Machine-Learning Experts, ch. 16), qu'il s'agisse d'effectuer une "root cause analysis" des Web sites (R. Draper, Data Traceability, ch. 17) ou encore, de vérifier l'impact des données effacées, de liens en liens, sur les réseaux sociaux (J. Valeski, Social Media: Erasable Ink?, ch. 18). Cela étant dit, voici 5 pistes concrètes en vue de faciliter l'accès aux données.

1. **Eviter, à la source, la production non organisée de volumineux ensembles de données stratégiques dans un format peu lisible par la machine, comme les spreadsheets.** Il est très fréquent que les utilisateurs "business" utilisent de tels formats qui conviennent

bien à la lecture humaine mais génèrent des "silos de données" redondants dont le traitement automatisé ultérieur est ardu.

S'appuyant sur son expérience en matière de statistiques dans le domaine scolaire en Nouvelle Zélande, P. Murrell propose des conseils de développement en R pour coder des données issues de tableurs dans un format réutilisable (P. Murrell, *Data Intended for Human Consumption, Not Machine Consumption*, ch. 3). Dans un autre chapitre appliqué au domaine de la chimie, R. Cotton plaide en faveur de processus de codage organisés, incluant contrôles et gestion des versions (R. Cotton, *Blood, Sweat, and Urine*, ch. 8), proposant une cure de "*Rehab for Chemists (and Other Spreadsheet Abusers)*" (p. 115) et s'exclamant au passage : "*Live Fast, Die Young and Leave a Good-Looking Corpse Code Repository*" (p. 114).

2. **Prendre en considération la variété des systèmes d'encodage hétérogènes sur le web** (ASCII, différentes normes ISO, UTF, ...). J. Levy propose des conseils de programmation ("text processing") en Python à cette fin offrant même au lecteur intéressé une série d'exercices (J. Levy, *Bad Data Lurking in Plain Text*, ch. 4).
3. **Identifier le pattern d'organisation des sites web analysés et en conserver l'historique des versions off line en vue d'un parsing ultérieur.** En raison du caractère imprévisible et dynamique de la mise à jour des sites web, cette démarche est indispensable. A. Laiacano propose plusieurs exemples de parsing et de reengineering du pattern de sites web en Python, Ajax et MATLAB scripts (A. Laiacano, *(Re)Organizing the Web's Data*, ch. 5).
4. **Evaluer les avantages et inconvénients des différents modèles logiques de bases de données, en fonction des usages et des modèles de coûts.** Deux chapitres discutent cette question essentielle pour le stockage et l'analyse des données issues du Web. S'inspirant d'une étude des "social media", l'un plaide en faveur d'un format simple de type "plain text" avec des flat files, lorsque les données sont volumineuses et statiques. Ceci en facilite la préservation à long terme, la rapidité de traitement et la sauvegarde, contrairement à certaines bases de données NoSql reposant sur le MapReduce paradigm (T. McNamara, *When Databases Attack: A Guide for When to Stick to Files*, ch. 12). L'autre évalue les coûts de gestion en terme de performance des différents modèles, reconnaissant la précision du modèle relationnel qui peut cependant être coûteux en

terme de performance, évoquant *"the Delicate Sound of a Combinatorial Explosion..."* (p. 167). Il conseille un modèle en graphe qui constitue une abstraction simplifiée mais utile quand il s'agit de gérer à la fois la complexité des interactions entre données et la performance de leur gestion (B. Norton, *Crouching Table, Hidden Network*, ch. 13).

5. **Utiliser le "cloud computing" avec prudence, en fonction du domaine d'application.** Sur la base d'un exemple réaliste, les risques de perte de performance, de coûts élevés et de pertes de données, lorsque le « cloud computing » est appliqué sans précaution sont évoqués (S. Francia, *Myths of Cloud Computing*, ch. 14).

## B. From "big data" to "long data" : 5 pistes pour faciliter l'interprétation des données



Une fois les données accédées, il s'agit de les interpréter pour les exploiter. Il est impensable d'étudier le phénomène **"big data"** sur le web sans prendre en considération la **question historique du temps**. Dans un blog publié en février 2013 par le journal Le Monde, la notion de **"long data"** est préconisée pour envisager la prise en compte de l'évolution des phénomènes dans le temps. Certains changements "brutaux" et récents (étude de la surpêche, de la déforestation, du climat, ...) prennent par exemple leur source dans des évolutions datant de plusieurs siècles. Mais cette étude est complexe car elle demande l'examen de l'évolution du sens des données et des mots dans le temps et dans l'espace. Dans cet esprit, citons par exemple l'application Google Ngrams, *"qui vise à tracer l'historique de l'usage d'un mot depuis l'an 1500, grâce à une analyse des livres numérisés par Google Books. Évidemment, cela ne commence qu'à l'invention de l'imprimerie et le fonds n'est pas exhaustif. Mais c'est un début qui a lancé*

*un nouveau champ d'études, la culturomique, reposant sur une analyse quantitative des termes étudiés."*

Associant le concept de « big data » à celui de « long data », voici 5 conseils relevés dans l'ouvrage en vue de faciliter l'interprétation des données.

- 1. Prendre en considération le caractère interdisciplinaire d'une approche « data quality », à travers des échanges permanents entre « connaissance métier » et « culture technique ».** Dans son chapitre déjà cité, *"Blood, Sweat, and Urine"* (Ch 8), R. Cotton présente une expérience dans ce sens dans le domaine de la chimie urologique. Pendant une semaine, en tant que développeur IT, il a échangé son poste avec celui d'un chimiste en vue d'un apprentissage réciproque. Dans un paragraphe éloquent, *"How Chemists Make Up Numbers"* (p. 108), **il relate sa stupeur devant l'exigence de précision de l'approche scientifique face à la complexité du réel observable et l'importance des enjeux humains et médicaux associés.** Il en tire avec humour les conclusions hypothétiques pour son propre métier d'informaticien : *"They have an endless list of documents and rules on good laboratory practice, how to conduct experiments, how to maintain the instruments ... The formal adherence to all these rules was a huge culture shock to me. All the chemists are required to carry a lab book around, in which they have to record the details of how they conducted each experiment. And if they forget to write it down ? Oops, the experiment is invalid. Run it again. I sometimes wonder what would happen if the same principles were applied to data scientists. You didn't document this function. Delete. I can't determine the origin of this dataset. Delete. There is no reference for this algorithm. Delete, delete, delete. The outcry would be enormous, but I'm sure standards would improve."* (p. 108). **A l'inverse, cet échange permet à son collègue chimiste, spécialiste du domaine d'application, de tirer des "best practices" quant au traitement des données** (éviter l'encodage intensif et non contrôlé sur des tableurs (cfr supra), à la source de redondance et de "data silos", remplacer le double encodage humain et les phases de réencodage (à la source d'erreurs et coûteuses en terme de manpower) par un workflow structuré organisant tâches humaines de validation et contrôles

automatisés ou encore, associer d'emblée aux données un modèle de base de données auquel correspondent des business rules, des règles de validation et une gestion des versions. L'auteur conclut : « *Sometimes, technology just works...* » (p. 116).

2. **Adopter une approche statistique itérative face à la complexité du domaine d'application incluant des facteurs exogènes imprévus sur le Web.** Dans un chapitre à propos des **taux de consultation des données et du trafic sur le Web**, qu'il s'agisse du **"Pay per click"** ou de la consultation de **Wikipedia**, F. Fink (It Just Me, or Does This Data Smell Funny ?, ch. 2) montre comment aux effets saisonniers qui diminuent structurellement le taux de consultation («*Superbowl Sunday*" aux USA, congés scolaires, week-ends) se mêlent malicieusement des bugs dans les logs de Wikipedia qui complexifient l'interprétation des séries temporelles . On trouve un phénomène analogue dans un chapitre (J. Perkins, "Detecting Liars and the Confused in Contradictory Online Reviews", ch.6) consacré à **l'analyse des sentiments sur le web** (à propos des restaurants, par exemple) où l'auteur découvre des contradictions (apparemment intentionnelles) entre les scores (ratings) attribués et les commentaires associés qui incluent parfois des doubles négations, sources de confusion en langage naturel. Dans l'approche, l'auteur montre comment construire un *"sentiment classifier"* en Python Natural Language sur la base d'un training set et d'une étude itérative en vue de détecter ces "mensonges volontaires".
3. **Face à certaines anomalies non élucidées par le modèle d'observation, ne pas hésiter à retourner sur le terrain pour réinspecter le domaine d'application (quand c'est matériellement possible).** Le chapitre correspondant (P. K. Janert, Will the Bad Data Please Stand Up, ch. 7) est introduit en ces termes : *"there is no such thing as bad weather – only inappropriate clothing ; there is no such thing as bad data – only inappropriate approaches"* (p. 95). L'auteur relate plusieurs expériences d'analyse des données en industrie visant à évaluer, sous contrainte de coût, le nombre d'appels en entreprise ou encore, les critères de production des produits défectueux. **Les modèles statistiques employés (courbe de Gauss, modèle de Poisson), ont chaque fois permis de détecter des exceptions qui ont requis une nouvelle inspection du domaine d'application** (par exemple, au sein de la

chaîne de production, des sources de destruction accidentelles n'avaient pas été intégrées dans la structure de l'échantillon). L'auteur plaide pour une approche empirique scientifique invitant à un réexamen régulier du modèle d'observation et des hypothèses associées : *"It was not the data that was the problem. The problem was de discrepancy between the data and our ideas (assumptions) about what the data should be like ... this discrepancy can lead to a form of "creative tension, which brings with it the opportunity for additional insights"* (p. 104).

4. **Prendre en considération le fait que des données non valides peuvent avoir, à l'insu de l'observateur, un impact (financier, par exemple) sur le réel empirique étudié.** Dans certains cas, l'inadéquation des données au modèle d'observation a un impact direct sur les réalités observées (*S. Burns, When Data and Reality Don't Match, ch. 9*). Ainsi, **les données sur l'état des marchés financiers diffusées sur Internet (Google Finance – Yahoo! Finance)** peuvent faire, en quelques minutes, partie intégrante du marché étudié où l'on observe *"a tight feedback loop where data about the state of the market affects the market (e.g. rising prices may cause people to push prices up further)"* (p. 119). Même si un algorithme de « data cleansing » permet a posteriori de détecter facilement les anomalies, celles-ci ont eu, entre temps, un impact concret sur le marché. Ainsi, le cas s'est-il présenté le 6 septembre 2008, lorsque le spider de Google News a diffusé par défaut à la date du jour des données plus anciennes non datées (et en fait obsolètes) concernant la banqueroute d'une valeur cotée sur le marché. En quelques minutes, cette information a donné lieu à des mouvements de vente massifs de la part des traders, avant que l'on ne se rende compte de l'erreur (p. 125). De tels phénomènes se sont souvent produits dans le secteur financier. Comment considérer le statut de ces données formellement erronées ex post, lorsqu'elles ont agi sur le marché réel ? D'importantes questions d'interprétation doivent être en effet abordées, lorsqu'on étudie un domaine d'application empirique critique, au sein duquel le système d'information est un instrument d'action sur les réalités qu'il représente.
5. **Accepter les compromis, dans le cadre d'un double arbitrage "fitness for use" & "coût-bénéfice".** On déduira facilement des recommandations qui précèdent que la "qualité parfaite" n'existe pas

(Vaisman M., The Dark Side of Data Science, ch. 15) . **Dans le domaine de la police criminelle**, par exemple, au sein du **Chicago Police Department's Predictive Analytics Group** (B. J. Goldstein, Don't Let the Perfect Be the Enemy of the Good: Is Bad Data Really Bad?, ch. 11), les séries statistiques temporelles relatives aux appels d'urgence ("*Reported Crime Information*", "*Sale of Narcotics*", ...) sont exploitées en vue de prévoir l'émergence de crimes par secteur géographique. Naturellement, dans la pratique, certains appels ne donnent pas lieu à la détection d'un délit (parce que les auteurs ont été prévenus entre-temps, par exemple). Ces informations sont toutefois utiles, pragmatiquement. Ainsi, le responsable du département conclut en ces termes : "*In order to make informed strategic and tactical decisions in an environment with imperfect data, one must make compromises. ... Still, I have repeatedly noted that it is better to have an informed decision built on imperfect data than to have decision built on no data at all. When one accepts that imperfection, it opens up the ability to integrate data into all supports of projects and policies*" (p. 148). On trouve le même type d'analyse dans le domaine du recensement aux USA et des enquêtes réalisées par le **Congressional Budget Office** ou la **U. S. Social Security Administration** (J. A. Schwabish, Subtle Sources of Bias and Error, ch. 10). C'est sur cette sage relativité que l'ouvrage se termine, privilégiant le pragmatisme et l'expérience à toute velléité stérile d'une représentation idéale du réel (Q. E. McCallum & K. Gleason, Data Quality Analysis Demystified: Knowing When Your Data Is Good Enough, ch. 19) :

*"Things change (and break)*

...

*Indeed".*



big data

data quality

data quality tools

database modelling

information management

long data

predictive analytics

---

**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](#)