



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

Typologie des anomalies, un cadre pour l'action : le cas du machine learning

Posted on 2022-10-18 by [Isabelle Boydens](#)

Isabelle Boydens(*) et Gani Hamiti(**)

(*) Data Quality Expert, Research Team

(**) Data Quality Analyst, Databases Team

Nederlandstalige versie



La qualité d'une donnée désigne son adéquation aux usages et objectifs visés (« fitness for use ») (Boydens, 1999, [Boydens 2014](#)). Dans cet article nous allons voir comment une typologie rigoureuse des anomalies offre un cadre pour l'amélioration de la qualité des données, dans de nombreux domaines, dont le machine learning. A propos du ML, dans un article ultérieur, nous montrerons comment cette technique peut améliorer les fonctionnalités d'un "data quality tool", par exemple dans les opérations de matching, comme noncé dans notre article de décembre 2021.

Toute base de données relationnelle opérationnelle bien conçue repose sur une hypothèse, celle du « monde clos » : le domaine de définition spécifie l'ensemble des valeurs admises au sein du modèle ou du schéma de la base de données (les contraintes d'intégrité) ; les « règles métier » peuvent aussi se décliner dans le code applicatif et contribuer ainsi à la définition des données. En vertu de cette hypothèse, une valeur non incluse dans le domaine de définition est considérée comme erronée et doit être rejetée de la base.

Par anomalie au sein d'une base de données, nous entendons ici une erreur formelle (par exemple, valeur obligatoire non complétée) mais aussi une présomption d'erreur demandant une interprétation humaine (par exemple : présomption de doublons entre enregistrements fortement similaires, émergence d'une nouvelle catégorie d'activité non prise en compte dans les tables de référence, etc.).

Ajoutons qu'une base de données empiriques évolue dans le temps avec l'interprétation des valeurs qu'elle permet d'appréhender (Boydens, 1999, 2011, Bade, 2011). Dès lors, il n'y a jamais de projection biunivoque entre une base de données et le réel observable représenté. La qualité totale n'existe pas. Ceci rend d'autant plus complexe la mise en place d'une stratégie d'évaluation et d'amélioration de la qualité des données, en fonction de leurs usages tels le machine learning (de Valeriola, 2020, 2021), dans le domaine de la justice, de la reconnaissance faciale, du traitement des maladies ou encore du journalisme, ... appliqué à des usages eux aussi bien particuliers (Redman, 2018, Dierickx, 2022).

If Your Data Is Bad, Your Machine Learning Tools Are Useless

by Thomas C. Redman

April 02, 2018



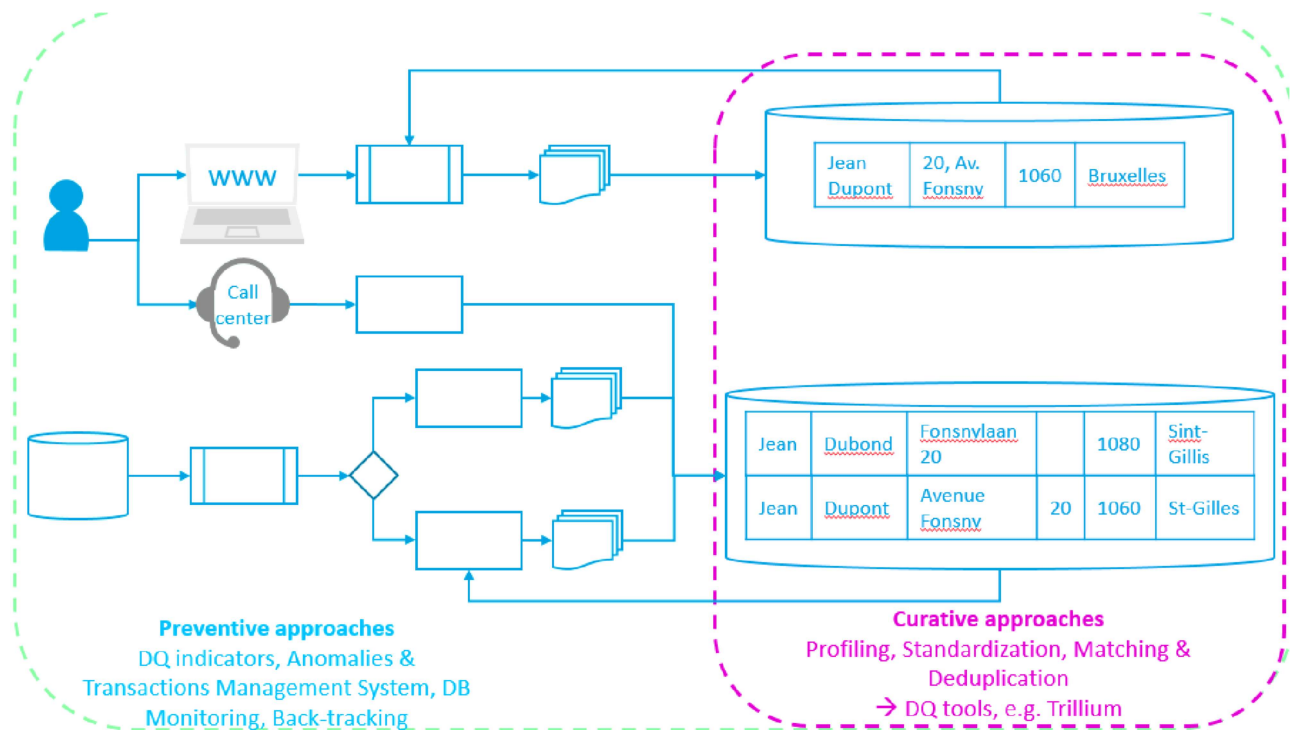
Source: <https://hbr.org/2018/04/if-your-data-is-bad-your-machine-learning-tools-are-useless>

Citons Redman et ses exemples éloquentes :

« Yet today, most data fails to meet basic “data are right” standards. Reasons range from data creators not understanding what is expected, to poorly calibrated measurement gear, to overly complex processes, to human error. To compensate, data scientists cleanse the data before training the predictive model. It is time-consuming, tedious work (taking up to 80% of data scientists’ time), and it’s the problem data scientists complain about most. Even with such efforts, cleaning neither detects nor corrects all the errors, and as yet, there is no way to understand the impact on the predictive model. What’s more, data does not always meet “the right data” standards, as reports of bias in facial recognition and criminal justice attest.”

(...)

“Increasingly-complex problems demand not just more data, but more diverse, comprehensive data. And with this comes more quality problems. For example, handwritten notes and local acronyms have complicated IBM’s efforts to apply machine learning (e.g., Watson) to cancer treatment.”

Figure 1. Approches préventives et curatives

Outre son intérêt évident pour la qualité des données, l'étude des anomalies est également importante en raison de leur pourcentage élevé qui affecte structurellement les systèmes d'information : jusqu'à 10 % du volume des données (Boydens, 2011, Van Der Vlist, 2011). Or, quand les enjeux (humains, sociaux, financiers, juridiques, scientifiques, médicaux, etc.) le demandent, ces anomalies doivent faire l'objet d'un examen semi-automatique, voire manuel, souvent lent et fastidieux, sans programme ad hoc recourant à des mesures préventives et curatives, **Figure 1**, (Boydens, 2014).

La typologie que nous proposons peut être utile dans toutes les disciplines ayant recours aux Data : Database Management, Master Data Management, machine learning (Dierickx, 2022, Redman, 2018), ... en tant que cadre global pour l'action Data Quality et aide à l'identification du traitement le plus approprié.

D'où viennent les anomalies, quelle en est la typologie et de là, comment les gérer au mieux ? Afin de répondre à ces questions, il convient de revenir préalablement sur la notion de donnée telle que nous l'avons posée dès 1999 et reprise notamment en 2021 (Boydens I., Hamiti G. et n Eeckhout R., 2021).

DONNÉES DÉTERMINISTES ET DONNÉES EMPIRIQUES

Dans le monde des bases de données, une donnée est un triplet (i, d, v) composé des éléments suivants :

- un intitulé (i) , renvoyant à un concept (une *catégorie d'activité administrative*, par exemple) ;
- un domaine de définition (d) , composé d'assertions formelles spécifiant l'ensemble des valeurs admises dans la base pour ce concept (une liste contrôlée de valeurs alphabétiques d'une longueur maximale l , par exemple), complétées éventuellement de règles métier se trouvant dans le code applicatif (voir plus haut, hypothèse du monde clos).
- et enfin, une valeur (v) à un instant t (le *secteur de la chimie*, par exemple).

On distingue alors les *données déterministes* des *données empiriques* (Boydens, 1999, 2011).

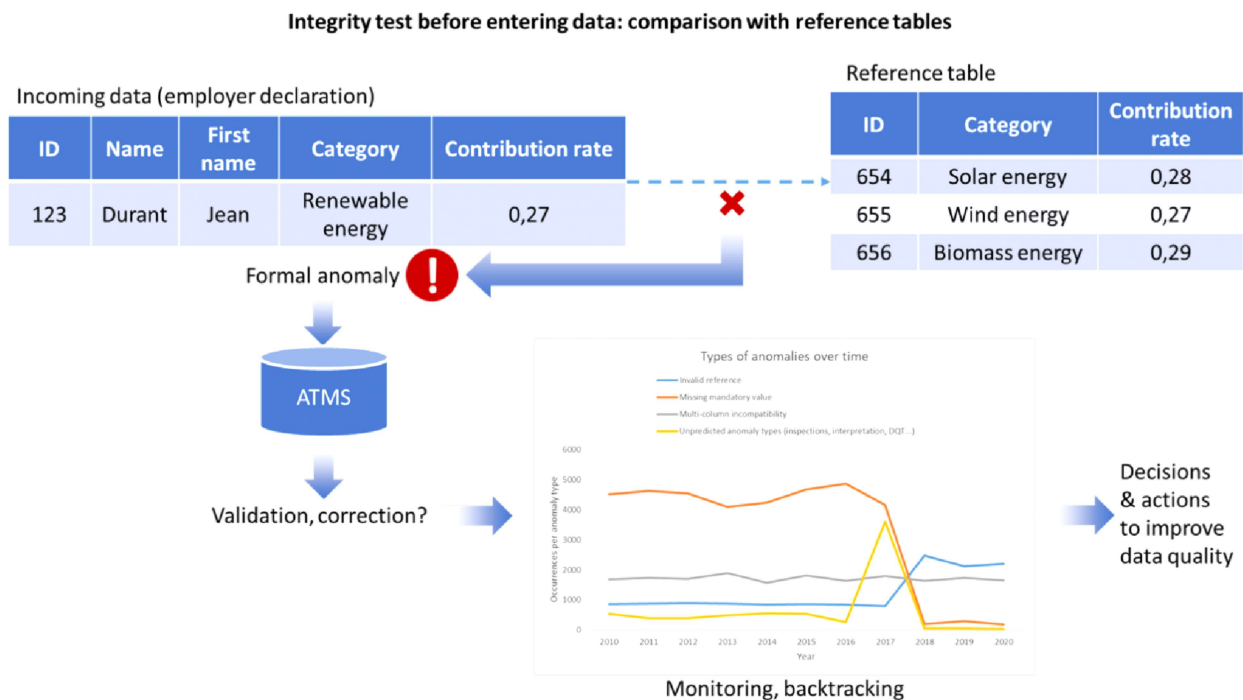
- Les premières se caractérisent par le fait que l'on dispose à tout moment d'une théorie qui permet de décider si une valeur v est correcte ou pas. Ainsi en est-il d'une opération algébrique simple portant sur un objet lui-même déterministe, comme la somme de valeurs relatives à tel champ numérique d'une base de données à un instant t . Les règles de l'algèbre n'évoluant dans le temps, on peut savoir à tout moment si le résultat d'une telle somme est correct ou pas. On dispose en effet d'un référentiel stable à cette fin.
- En revanche, en ce qui concerne les données empiriques, sujettes à l'expérience humaine, la norme évolue dans le temps avec l'interprétation des valeurs qu'elle permet d'appréhender. Ainsi en est-il par exemple du domaine médical (où la théorie évolue au fil des observations sur les patients atteints par une pathologie, comme en témoignent les recherches actuelles sur le coronavirus) mais aussi des domaines juridiques et administratifs où l'interprétation des concepts légaux se transforme avec l'évolution continue de la réalité traitée et avec celle de la jurisprudence. Comment en évaluer la validité en l'absence de référentiel absolu à cette fin ?

TYPLOGIE DES ANOMALIES ET TRAITEMENTS POSSIBLES

Une typologie des anomalies se profile alors, en fonction de leur cause potentielle et de la manière de les envisager :

- erreur formelle certaine : par exemple, un champ obligatoire non complété lors de l'encodage manuel des données par un humain;
- présomptions d'erreurs formelles : par exemple a) présomptions de doubles dues à des processus de capture de données redondantes en amont, ou encore b) une incohérence avec une table de référence dont on ignore si elle est à jour, par exemple dans le domaine de l'énergie renouvelable (Figure 2) ;
- erreur indétectable formellement *a priori* : par exemple, omission d'une mise à jour.

Figure 2. Gestion des anomalies, ATMS, Monitoring & Back tracking



Les deux derniers cas de figure de la typologie qui précède peuvent dénoter de cas d'anomalies dues à l'évolution dans le temps du domaine empirique représenté et à l'émergence de nouveaux concepts non pris en compte (**Figure 2**). Par exemple : un test d'intégrité avant l'entrée des données dans la base de données principale détecte une anomalie formelle. Le traitement de l'anomalie (validation ou correction) est stocké dans l'ATMS (permettant le suivi des anomalies et de leur traitement dans temps, comme indiqué plus loin dans cet article) et alimente un tableau

de bord lequel, moyennant un monitoring des anomalies et traitements aidera à la prise de décision en vue d'améliorer la qualité des données.

Selon les besoins du métier, on décidera de considérer ces anomalies comme :

- bloquantes : elles sont rejetées de la base de données en vertu de l'hypothèse du monde clos précédemment évoquée ;
- non bloquantes : les valeurs sont tout de même intégrées selon des modalités variables au sein du système d'information avec l'enregistrement correspondant, pour deux familles de raisons :
 - les rejeter du système ralentirait le processus métier (par exemple, le prélèvement des cotisations sociales) et elles ne sont pas considérées comme « stratégiques »;
 - les prendre en considération dans le système d'information est indispensable, car elles sont considérées comme stratégiques et sont liées à des données empiriques dont la définition est potentiellement évolutive. À partir d'un certain seuil à évaluer par les spécialistes du domaine, leur traitement demande une interprétation humaine, car elles peuvent dénoter de l'émergence de phénomènes nouveaux qu'il importera de prendre en considération dans le système d'information (**Figure 2**), moyennant une gestion de versions. En outre, elles trouvent potentiellement leur origine dans les flux alimentant la base de données, problématique qui, une fois identifiée, pourra être structurellement résolue avec le back tracking (Boydens, 2018, Boydens et al., 2021).

La décision consistant à identifier les anomalies empiriques « non bloquantes » est sensible en ce qu'elle relève d'une connaissance prévisionnelle des réalités traitées à un instant t , élément lui-même évolutif susceptible de faire l'objet d'une adaptation concertée au sein du système d'information. Ceci nous renvoie à la question épistémologique de la « boucle herméneutique »

La démarche herméneutique consiste à envisager les phénomènes empiriques en termes d'interactions par rapport à un cadre conceptuel plus général construit en vue de leur conférer un sens. Cependant, toute marche interprétative soulève un paradoxe : celui du « cercle

herméneutique » (Aron, 1969). Chaque observation ne prend sens que confrontée à un ensemble, à une « précompréhension ». Or, la sémantique de l'ensemble repose elle-même sur l'interprétation des éléments qui le constituent. Le processus de construction que suppose l'herméneutique est par nature toujours inachevé. Il convient toutefois d'y poser ponctuellement un arrêt en connaissance de cause afin de livrer des résultats provisoires (Boydens, 1999).

Comment prendre en considération les « anomalies non bloquantes » et leurs traitements, sans affecter ni la performance, ni l'intégrité des données en production ? Avec l'ATMS, ou *Anomalies and Transactions Management System* (Boydens et al., 2021) – éventuellement couplé aux Data Quality Tools (Boydens et al, 2021b) – associé au Back Tracking (Boydens, 2018), on passe de « l'hypothèse du monde clos » à celle d'un « monde ouvert » sous contrôles automatisés au sein des bases de données de gestion. Et l'on bâtit, ce faisant un programme consolidé permettant d'évaluer la qualité des données et d'y remédier structurellement. Ce programme n'est jamais un « one shot » car il doit être mené dans la continuité, incluant un processus de maintenance.

APPLICATION AU MACHINE LEARNING : PERSPECTIVES

Avant la mise en place du cycle de ML

Dans le cas du machine learning (de Valeriola, 2020, Dierickx, 2019, 2022), le programme (analyse et solutions associées) – exposé plus haut dans cet article de blog – peut s'appliquer en amont aux données mobilisées pour entraîner un modèle prédictif et en assurer la maintenance dans le temps avant la mise en place du cycle de ML.

Au coeur du cycle de ML (Figure 3)

Au coeur du système de ML, on peut construire les indicateurs de qualité adéquats aux algorithmes et modèles de ML choisis, par exemple, indicateurs de la qualité :

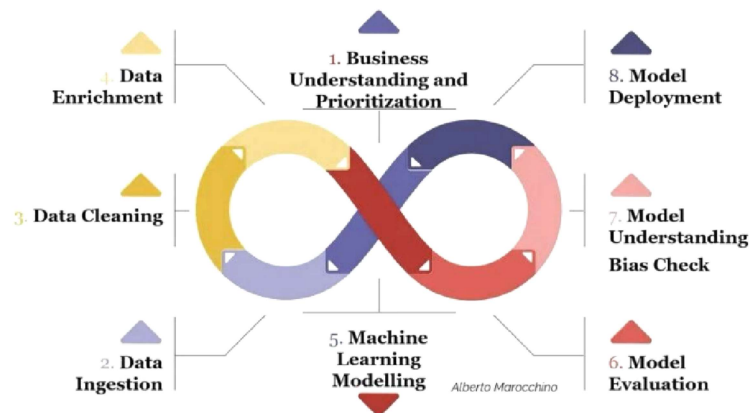
- des annotations potentiellement obtenues par crowdsourcing, dans le cas de modèles supervisés (Northcutt et al, 2021, Gupta et al, 2021)
- des biais (ou marges d'erreur) tolérés ou pas selon les usages déterminés en synergie avec l'IT et le business (concepteurs des

données, utilisateurs du modèle prédictif).

Le système préconisé dans (Gupta et al, 2021) pourrait être enrichi en ayant recours à un ATMS à la source, comme mentionné plus haut et par ailleurs, suggère d'agir sur la qualité de données déjà pré-traitées pour le ML.

Cette référence illustre *la problématique de données de qualité insuffisante au sein du modèle de ML*. Par exemple, la *mauvaise distribution d'un attribut peut constituer un biais* qui, à son tour, soulèvera des problèmes éthiques : dans le cas de l'évaluation des demandes de crédit en fonction du risque, en cours d'examen par l'UE en septembre 2022 : "... les algorithmes peuvent comporter des biais. Par exemple, un modèle est susceptible de refuser fréquemment un crédit aux individus de 30 ans et jamais aux personnes de 57 ans. Pourquoi ? Simplement parce que les dossiers des trentenaires étudiés par la machine lors de son entraînement étaient nombreux, donc la probabilité d'y trouver des défauts également, alors qu'il n'y avait qu'un ou deux exemples de candidats de 57 ans et qu'ils ont à chaque fois payé l'intégralité de leur crédit (une erreur toutefois si grossière qu'elle est généralement anticipée, mais d'autres biais peuvent être plus pernicious)."

Cycle de vie des données en ML



RÉFÉRENCES

- Gudivada, V., Apon, A., & Ding, J. (2017). Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations. *International Journal on Advances in Software*, 10(1), 1-20.
- Gupta, N., Patel, H., Afzal, S., Panwar, N., Mittal, R. S., Guttula, S., ... & Saha, D. (2021). Data Quality Toolkit: Automatic assessment of data quality and remediation for machine learning datasets. *arXiv preprint arXiv:2108.05935*.
- Gupta, N., Mujumdar, S., Patel, H., Masuda, S., Panwar, N., Bandyopadhyay, S., ... & Munigala, V. (2021, August). Data quality for machine learning tasks. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (pp. 4040-4041).

Figure 3. Source : Dierickx, 2022

L'explicabilité des prédictions du ML : législations, questions éthiques et de data quality, nouveaux domaines de recherche

Dans d'autres cas évoqués par l'UE, *l'opacité des modèles d'apprentissage profond* est telle que même les ingénieurs ne peuvent plus motiver précisément les prédictions produites en résultat.

Une loi est en cours de préparation pour tenter de réguler les procédures utilisées, *l'AI Act*, censé renforcer le RGPD. S'agissant de procédures analogues aux demandes de crédit, les experts estiment que *"Cela ne veut pas dire pour autant que les explications données seront sûres et certaines (impossible avec certains modèles), mais qu'elles seront fortement probables – des ingénieurs parlent d'ailleurs plus d'interprétabilité que d'explicabilité."*

Le programme évoqué plus haut (ATMS, data quality tools) avant le cycle du ML, couplé à la prise en compte d'une évaluation de la qualité au coeur du processus de ML, le tout devant être documenté, pourrait s'appliquer aussi afin de donner plus d'informations ciblées aux utilisateurs des données prédictives sur "l'explicabilité" et la qualité relative de ces dernières, la qualité totale n'existant pas.

De nouveaux domaines de recherche se présentent peu à peu dans ce sens : Data Centric AI ou encore, Causal AI. La problématique se pose donc en amont, au coeur et en aval du cycle de vie des données en ML.

A propos du ML, dans un article ultérieur, nous montrerons par ailleurs, comment cette technique peut améliorer les fonctionnalités d'un "data quality tool", par exemple dans les opérations de matching, comme annoncé dans notre article de décembre 2021.

Références

Aron, R., 1969. La philosophie critique de l'histoire. 1969. Édition Librairie philosophique J. Vrin. Collection Points – Sciences humaines. ISBN 2560848158182.

Bade D., It's about Time!: Temporal Aspects of Metadata Management in the Work of Isabelle Boydens. In *Cataloging & Classification Quarterly*

(The International Observer), volume 49, n° 4, 2011, pp. 328-338. ([lien vers l'article](#)).

Boydens I., *Informatique, normes et temps*. Bruxelles : Bruylant, 1999, 570 p. (Cet ouvrage s'est vu décerner le prix de la Fondation L. Davin, conféré par l'Académie Royale des sciences, des lettres et des beaux-arts de Belgique, 1999). (Introduction et Première partie, pp. 30-126) – bibliothèques

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 ([chapitre 7](#)).

Boydens I., *Dix bonnes pratiques pour améliorer et maintenir la qualité des données*. Bruxelles, Smals, Research Section, post de blog, 16/06/2014 (dernière mise à jour : décembre 2021). [/dix-bonnes-pratiques-pour-ameliorer-et-maintenir-la-qualite-des-donnees/](#)

Boydens I., « *Data Quality & Back Tracking : depuis les premières expérimentations à la parution d'un Arrêté Royal* ». Bruxelles, Smals, Research Section, post de blog, 14/05/2018. [/data-quality-back-tracking-depuis-les-premieres-experimentations-a-la-parution-dun-arrete-royal/](#)

Boydens I., Hamiti G. et Van Eeckhout R., *Data Quality : "Anomalies & Transactions Management System" (ATMS), prototype & "work in progress"*. Bruxelles, Smals, Research Section, post de blog, 8/12/2020. [/data-quality-anomalies-transactions-management-system-atms-prototype-work-in-progress/](#)

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.

<https://www.insee.fr/fr/information/5398691?sommaire=5398695>

Boydens I., Corbesier I. et Hamiti G., *Data Quality Tools : retours d'expérience et nouveautés*. Bruxelles, Smals, Research Section, post de blog, 07/12/2021. [/data-quality-tools-retours-dexperience-et-nouveautes/](#)

Brown S., Why it's time for "data Centric Artificial Intelligence" ? MIT Management Sloan School, juin, 2022. <https://mitsloan.mit.edu/ideas-made-to-matter/why-its-time-data-centric-artificial-intelligence>

De Valeriola S., *L'ordinateur au service du dépouillement de sources historiques. ´ Eléments d'analyse semi-automatique d'un corpus diplomatique homogène*. In *Histoire & Mesure*, 35, 2 (2020), 171–196.

De Valeriola, S. *Can historians trust centrality ? Historical network analysis and centrality metrics robustness*. In *Journal of Historical Network Research* 6 (2021), 45–85.

Dierickx L., « *Apprentissage automatique : les challenges de la qualité des données dans la perspective d'une adéquation aux usages*», Conférence, Groupe de contact FNRS « Analyse critique et amélioration de la qualité de l'information », ULB, mai 2022.

<https://mastic.ulb.ac.be/2022/02/reunion-du-groupe-de-contact-fnrs-analyse-critique-et-amelioration-de-la-qualite-de-linformation-numerique-%EF%BF%BC/>

Dierickx, L. (2019, February). Why news automation fails. In *Computation+ Journalism Symposium, Miami, FL*.

Gupta N, Patel H, Afzal S, et al. (2021) Data Quality Toolkit: Automatic assessment of data quality and remediation for machine learning datasets. arXiv [cs.LG]. Available at: <http://arxiv.org/abs/2108.05935>.

Northcutt CG, Athalye A and Mueller J (2021) Pervasive label errors in test sets destabilize machine learning benchmarks. arXiv [stat.ML]. Available at: <http://arxiv.org/abs/2103.14749>.

Redman T. C., If Your Data is Bad, your Machine Learning Tools are useless. Harvard, Business Review, avril 2018. <https://hbr.org/2018/04/if-your-data-is-bad-your-machine-learning-tools-are-useless>.

Chai S. et al., The Case for Causal AI, Stanford Innovation Social Review, summer 2020.

https://ssir.org/articles/entry/the_case_for_causal_ai

Van Der Vlist, E. 2011. Relax NG. Mai 2011. Édition O'Reilly Media.
ISBN: 0596004214

Ce post est une contribution d'Isabelle Boydens, Data Quality Expert, Research Team et de Gani Hamiti, Data Quality Analyst, Databases Team. Cet article est écrit en leur nom propre et n'impacte en rien le point de vue de Smals.

atms

back tracking

data quality

data quality tools

information management

machine learning

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

Newsletter & webinars:

Dutch French

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](#)