



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

# "Data Observability", un nouveau topic dans le paysage "Data Quality" ?

Posted on 2023-09-26 by [Isabelle Boydens](#)

[Nederlandstalige versie](#)

Isabelle Boydens(\*), Isabelle Corbesier(\*\*) et Gani Hamiti(\*\*)

(\*) Data Quality Expert, Research Team

(\*\*) Data Quality Analyst, Databases Team



## Les origines du concept d'"observability"

À l'origine, le terme *observability* en IT est issu des pratiques de software engineering.

L'observabilité est un concept de haut niveau qui consiste à analyser dans un ensemble l'état d'un système de composants particulièrement hétéroclites et nombreux, afin de diagnostiquer les comportements

critiques et de faciliter l'identification de leurs causes (1). En pratique, cela implique de collecter et d'analyser continuellement des données élémentaires d'infrastructure comme l'utilisation dans le temps du processeur ou de l'espace de stockage, mais aussi des logs applicatifs ou de traçage potentiellement plus complexes. Face à la multiplication des technologies et composants des systèmes d'information modernes, obtenir un niveau d'observabilité suffisant peut nécessiter un travail de développement parfois important. Cette difficulté a constitué un terrain fertile pour l'émergence d'outils dédiés à l'observabilité, servant tantôt à exploiter les données déjà produites par un système, tantôt à greffer, sur ses composants, des fonctionnalités de production de données destinées à une meilleure visibilité de son comportement. (2)

La distinction parfois forcée entre monitoring et *observability* peut être questionnée, dans la mesure où, tout comme l'*observability*, les pratiques labellisées comme du "monitoring" ne sont jamais élaborées comme une fin en soi mais visent également à diagnostiquer l'état du système et à en prévenir ou corriger les incidents. Certaines références relient d'ailleurs les deux concepts, puisque le *monitoring* (couramment désigné comme APM ou *Application Performance Monitoring*) et l'observabilité sont repris dans une seule et même définition, qui vise les mêmes produits. Quitte à distinguer les deux concepts, un lien de parenté les lie tout du moins, dans la mesure où un bon niveau d'observabilité requiert un monitoring suffisant, en plus d'une bonne documentation et d'une solide connaissance du système par les équipes chargées de l'observer.

## **Du monde software à celui des données**



Figure 1 – Système d’information : composants hétérogènes et boucles de rétroaction

À l’instar du data profiling (audit de la qualité des données (3) précédant généralement les phases de standardisation et de matching (4) dans les *data quality tools*) dans les années nonante, la *data observability* est une transposition récente (le terme s’est popularisé massivement vers 2022-2023) et assumée reprise au monde de infrastructure. Tout comme l’observabilité dans ce contexte consiste à pouvoir diagnostiquer et améliorer l’état du système sur base de ce qu’on en voit, la *data observability* est la capacité à diagnostiquer l’état général des données d’un système sur base d’une vue détaillée construite à travers ses métadonnées. La *data observability* visera donc à rassembler le monitoring, le suivi et le tri des incidents liés aux données, avec pour objectif final de prévenir ou de minimiser le downtime qui leur est imputable.

Dans un livre coédité en 2022 (5), Barr Moses, la CEO de « Monte Carlo Data », identifie plusieurs piliers de *data observability* qui existaient déjà dès les années 90 nonante dans le « *data profiling* » et qui ont ensuite été largement réutilisés. Nous retenons **les 4 points essentiels** suivants :

- La **fraîcheur (*freshness*)** : la confirmation que les données sont à jour et sont rafraîchies de manière adéquate.

- La **distribution (*distribution*)** : la confirmation que les données se situent dans un intervalle acceptable, notamment via la mesure des valeurs inattendues ou nulles.
- La **complétude (*completeness*)** : la vérification qu'un dataset est complet (nombre de records ou nombre de colonnes) afin d'identifier les problèmes possibles dans les systèmes à la source. Notons que la complétude peut être fondamentalement impossible à mesurer avec certitude ; c'est le cas, par exemple, avec la population totale de personnes atteintes de l'Alzheimer ou d'un cancer, parfois à leur insu dans les premières phases.
- Le **lineage** : la documentation et la compréhension de l'entièreté des systèmes de données d'une organisation, y compris les sources de données en amont et les systèmes cibles en aval. En pratique, on peut également remarquer des boucles de rétroaction où l'exploitation des données en aval (par exemple, lors de projets de machine learning ou BI) amène à des modifications du système en amont (cf. Figure 1). Ainsi, le lineage dépasse les limites des approches techniques et requiert une intervention humaine considérable ainsi qu'un budget à la taille de celle-ci. D'autres obstacles évoqués plus bas peuvent se présenter pour lier des systèmes d'information entre eux.

### **Différences entre « *data observability tools* », « *data quality tools* » (approche curative) et « *ATMS-back tracking* » (approche préventive)**

La documentation relative aux outils de « *data observability* » évoque le "*lineage*" dans un sens technique. Il s'agit d'observer comment les données évoluent à travers les différents composants d'un système ; par exemple, depuis un front-end où les données sont introduites, en passant par une API REST en back-end, une base de données transactionnelle, puis un datawarehouse, jusqu'à des systèmes de business intelligence ou de reporting. Contrairement au « *back tracking* », qui étudie le flux de données dans le système d'information complexe mais surtout en amont et en aval du SI (cf. Figure 2), on parle ici de "*lineage*" entre les composants du SI auxquels les gestionnaires de l'outil d'*observability* auraient intégralement accès afin de suivre la mutation des données en temps réel.

Une panoplie d'outils encore récents portent le label « *data observability* », au sein desquels le monitoring des data coexiste parfois avec le monitoring système (Bigeye, Collibra, Databand (IBM), DataBuck (FirstEigen), Kensu, Metaplane, Monte Carlo, Soda, ...). Il est à noter que certains fournisseurs de *data quality tools* comme Informatica ou Precisely incluent déjà la « *data observability* » à côté du *profiling*. Le terme semble opportun dans l'air du temps. Il sera intéressant de suivre la maturation des outils en question et éventuellement de les tester.

Les *data quality tools* (4), quant à eux, ne se limitent pas à l'observation des données, mais visent à intervenir directement sur celles-ci. C'est en effet le but des fonctionnalités de standardisation (entre autres des adresses) en batch ou en temps réel (via API REST, par exemple) ainsi que du « *data matching* » potentiellement « *fuzzy* » tirant parti de familles algorithmiques dédiées aux *business cases* à traiter.

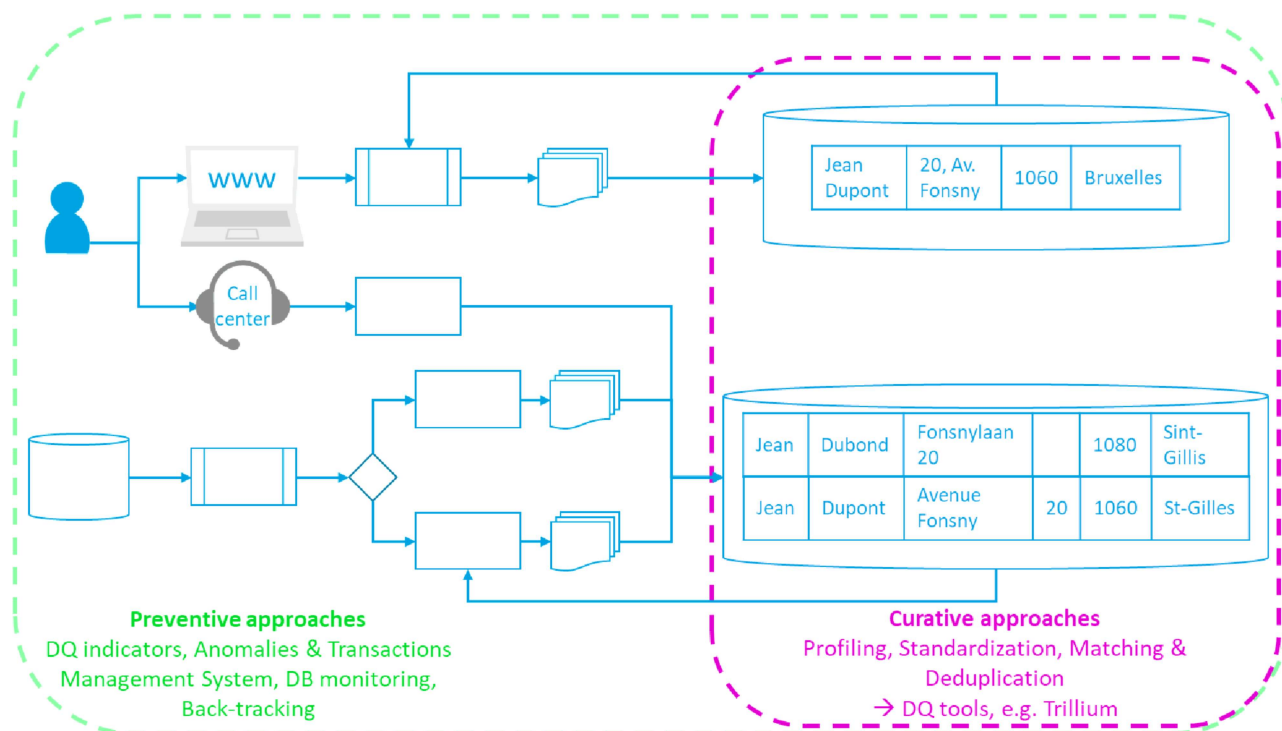


Figure 2 – Approches préventives et curatives

En résumé, les outils de *data observability*, qui s'apparentent à du « *data profiling* » en temps réel, peuvent soutenir l'effort d'amélioration de la qualité des données, en contribuant notamment à identifier et alerter les responsables sur les problèmes de qualité. Il reste cependant nécessaire :

- de faire appel à des outils de *data quality* (4) dédiés en synergie avec le business pour mettre en place des stratégies de résolution des

- problèmes détectés (approches curatives);
- et/ou de disposer d'un ATMS (6), permettant d'effectuer du *back tracking* et d'éviter l'apparition de ces problèmes en amont (approches préventives).

## Conclusion : des outils récents à suivre

De manière générale, on peut déduire que pour que l'observabilité proposée par ces outils ait un impact positif mesurable dans un contexte d'entreprise grande nature :

- Il faut que l'organisation qui emploie l'outil permette structurellement l'accès centralisé à des sources de données de natures et de degrés de criticité divers. Ce point est susceptible de poser un grand nombre de challenges techniques, légaux et organisationnels :
  - La séparation des environnements : toutes les sources de données ne se trouvent pas toujours en production
  - Les questions de réseaux et de flux : faire communiquer tous les composants nécessaires entre eux n'est pas trivial dans la plupart des organisations d'une certaine taille et avec un certain niveau d'exigence de sécurité.
  - quid des droits d'accès et de RGPD ? Pas tant du point de vue de la faisabilité technique de l'implémentation des exigences de RGPD, mais surtout du point de vue organisationnel.
  - La séparation des projets et des équipes : les équipes responsables du système applicatif ne sont pas nécessairement attachées à la même hiérarchie ou aux mêmes pratiques que celles gérant l'infrastructure ou celles qui exploitent les données en aval pour du *datamining* ou de la BI.
- Pour autant que le premier point soit acquis et résolu, il faut également que les gestionnaires du système se penchent sérieusement avec le business sur les indicateurs à monitorer, les seuils critiques, les alertes à configurer et la définition des rôles : qui est responsable de quelles données et quelles sont les réactions attendues en cas de problème ?

Dès lors, nous sommes loin du "plug-and-play" quasi magique vanté par certains contenus en ligne. À cela s'ajoute la réalité de l'investigation des problèmes de données non triviaux, qui dépasse souvent les frontières

d'un système d'information technique, aussi complexe soit-il. Pour ces raisons, lorsque les enjeux le justifient, la qualité des données reste plus que jamais en besoin d'intervention et d'interprétation humaines.

## Références

- (1) <https://about.gitlab.com/blog/2022/06/14/observability-vs-monitoring-in-devops/>
- (2) <https://www.ibm.com/topics/observability>
- (3) Olson Data Quality: The Accuracy Dimension (The Morgan Kaufmann Series in Data Management Systems), 2003.
- (4) BOYDENS I., CORBESIER I. et HAMITI G., [Data Quality Tools : retours d'expérience et nouveautés](#), 07/12/2021.
- (5) <https://www.oreilly.com/library/view/data-quality-fundamentals/9781098112035/>
- (6) 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 2021, n°6, p. 100-122. [Fichier PDF](#) / [Lien vers la Revue et vers l'article](#).

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

data quality

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

