

Améliorer la qualité de l'information : du "stemma codicum" au "data tracking"

Posted on 2012-06-05 by [Isabelle Boydens](#)



Au Moyen-âge, le processus de transformation de l'information, avant l'invention de l'imprimerie, se déployait de siècles en siècles avec les générations de moines copistes. Afin d'établir l'appareil critique d'un manuscrit (dont on dispose souvent de multiples copies divergentes et dont on a parfois perdu l'original), l'historien construit un stemma codicum (généalogie des données), technique d'analyse comparative des variantes empruntée à la philologie. L'établissement du *stemma codicum*, accompagné d'un travail d'interprétation critique, offre une reconstruction conjecturale argumentée du manuscrit original.

A la fin du XXème siècle, le géant des télécommunications américain, AT&T Laboratories, a déployé à plus grande échelle une technique d'analyse du processus de transformation de l'information ("data tracking") reposant en partie sur un principe analogue. La technique du "data tracking", proposée par [Thomas Redman](#) ("Data Quality : the Field Guide", 2001), vise à évaluer et à améliorer les aspects formels de la qualité des vastes bases de données contemporaines.

La qualité d'une base de données désigne son adéquation relative aux usages ("fitness for use"), lesquels évoluent dans le temps (voir : Boydens I., Informatique, normes et temps, Bruxelles, Bruylant, 1999). On se trouve nécessairement face à un compromis d'ordre pratique, sous contrainte de budget. En tant qu'instrument d'action sur le réel, les bases de données revêtent toutefois des enjeux considérables. L'OTAN en fit l'expérience durant la guerre du Kosovo en mai 1999, après avoir bombardé l'ambassade de Chine à Belgrade. Pressée d'expliquer son attaque, l'organisation incrimina les bases cartographiques utilisées pour guider ses missiles. Celles-ci répertoriaient en effet un plan de Belgrade obsolète ... (voir : Boydens I., Les bases de données sont-elles solubles dans le temps? In La Recherche hors série ("Ordre et désordre"). Hors série n° 9, novembre-décembre 2002, p. 32-34).

Parmi les différentes méthodes d'évaluation et d'amélioration de la qualité des bases de données, nous synthétisons ici brièvement la technique du "data tracking". L'approche se caractérise par le fait qu'elle se concentre sur l'étude du processus de production de l'erreur formelle en vue d'y remédier à la source. Elle a été mise en oeuvre sur le terrain dans le cadre de l'egovernment en Belgique par le Data Quality Competence Center de Smals, en collaboration avec le secteur de la sécurité sociale. Les bases de données traitées permettent le prélèvement et la redistribution annuels d'environ 40 milliards d'euros de cotisations sociales; les enjeux sociaux et financiers qu'elles soulèvent sont importants.

Le « data tracking » vise à évaluer quantitativement la validité formelle des valeurs introduites dans une base de données et à en améliorer le traitement. Une base de données est un fleuve : au lieu de nettoyer ponctuellement le fond du fleuve (comme le préconise le "data cleansing", méthode de correction automatique), Redman propose d'en analyser structurellement les sources et les flux. Traditionnellement, chaque enregistrement d'une base de données est assemblé au terme de plusieurs étapes (ou processus), de la même façon qu'un produit est assemblé dans une usine. La qualité des données dépend donc, entre autres, de la qualité du processus d'assemblage.

Une des caractéristiques du "data tracking" (suivi des données) réside dans le fait que l'instrument de mesure est incorporé aux processus et permet en quelque sorte une analyse continue de leur qualité. Le "data

tracking" repose sur une exploitation de la redondance des données que l'on retrouve dans la plupart des systèmes informatiques en vue d'en évaluer trois aspects :

- la validité formelle de données intégrées dans une seule base de données ;
- la cohérence entre données intégrées dans plusieurs bases de données ;
- la durée des cycles de production et de traitement de l'information.

A titre d'exemple, dès 2006, le secteur de la sécurité sociale belge a appliqué la méthode du *data tracking* afin d'assurer le suivi des processus au niveau du « top 50 » des employeurs commettant le plus d'anomalies (violations de contraintes d'intégrité) dans les déclarations sociales envoyées à l'administration (lesquelles impliquent le traitement trimestriel de plus de quatre millions d'enregistrements auxquels correspondent plusieurs centaines de champs). Le but de l'opération consiste à détecter, chez l'expéditeur et en partenariat avec l'administration, les éléments à l'origine de la production d'un grand nombre 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, ...). Notons qu'en raison du contexte, l'application présentée ici inclut deux adaptations par rapport à la méthode de Redman :

- 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 un historique des anomalies et de leur traitement ;
- il s'agit d'un « tracking arrière » (ou « back tracking ») : on part de la situation finale pour revenir, étape par étape, à chaque source et processus qui en a permis l'élaboration. L'objectif est d'éviter le traitement de données ou de flux inutiles pour l'analyse.

L'opération permet :

- d'établir un partenariat avec les citoyens 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'obtenir des résultats potentiellement durables, puisque la cause structurelle des erreurs systématiques 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 (voir par exemple : Boydens I. et Van Hooland S., Hermeneutics applied to the quality of empirical databases. In *Journal of documentation*, volume 67, issue 2, 2011, pp. 279-289). Pour cette raison, il est conseillé de relancer de manière régulière l'opération de "data tracking" en vue de :

- s'assurer de la permanence des résultats obtenus;
- détecter d'éventuelles nouvelles sources d'erreur;
- mettre en place un processus d'évaluation et d'amélioration de la qualité des données continu, avec un ROI important.

data quality

egov

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](#)