



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

Data Quality Tools : retours d'expérience et nouveautés

Posted on 2021-12-07 by [Isabelle Boydens](#)

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

(*) Data Quality Expert, Research Team

(**) Data Quality Analyst, Databases Team

La problématique de la qualité des données (ou "fitness for use", adéquation aux usages) est maintenant reconnue au plan international comme étant un facteur de succès à prendre en compte dans tout projet impliquant des bases de données. En 2016, T. Redman chiffrait ainsi les coûts de la "non qualité" aux USA, dans son livre "Getting in front on Data" (p. 25) : "\$3,1 Trillions/year in the US, which is about 20 percent of the Gross Domestic Product".

Les "data quality tools" professionnels et commerciaux sont nés dans les années 1980 avec la nécessité pour les entreprises du monde entier de disposer de fichiers d'adresses et de coordonnées précises concernant leurs clients et transactions. Avec le temps, ces outils ont pris un essor considérable, tant du point de vue de l'ampleur des domaines couverts, traitant tout type de chaîne alphanumérique, que des dizaines de milliers d'algorithmes "ad hoc" développés à cette fin. Vu le caractère stratégique de la qualité des données dans les entreprises, organismes internationaux et administrations, la recherche est très active dans le domaine et de nouveaux algorithmes sont régulièrement proposés et intégrés dans ces outils. Pour ces raisons, dans un cadre professionnel, on peut difficilement leur substituer un développement "home made" et leur acquisition est commandée en vue d'une approche "data quality" sérieuse et complète.

Aussi, depuis plus de 10 ans, Smals a acquis un "Data Quality Tool" professionnel, toujours parmi les leaders du marché à l'heure actuelle, dans le cadre de son Data Quality Competency Center. Depuis lors, plus de trente projets d'envergure ont mobilisé cet outil dans le cadre de la sécurité sociale belge et en dehors de celle-ci.

Nous proposons ici, sur la base de l'expérience acquise :

- de rappeler et d'illustrer les fonctionnalités les plus usitées de l'outil, ainsi que quelques bonnes pratiques;
- d'annoncer plusieurs nouveautés qui sont autant d'extensions de l'outil dans le courant de cette année 2021.

1. Retour d'expérience : une approche curative très efficace en vue de traiter les problèmes de qualité au sein des bases de données

En parallèle et en complément des approches préventives destinées à éviter en amont l'émergence de problèmes de qualité de données ("back tracking" reposant sur un ATMS, Anomalies and Transactions Management System) à propos desquelles un article de blog a été publié respectivement en 2018 et en 2020 ainsi qu'un article scientifique édité à Paris par le Courrier des Statistiques (Paris, INSEE, juillet 2021), les "data quality tools" représentent une approche curative très efficace pour traiter les anomalies déjà présentes dans les bases de données (*Figure 1*).

First: identify business priorities, «fitness for use», budget and «cost-benefits»

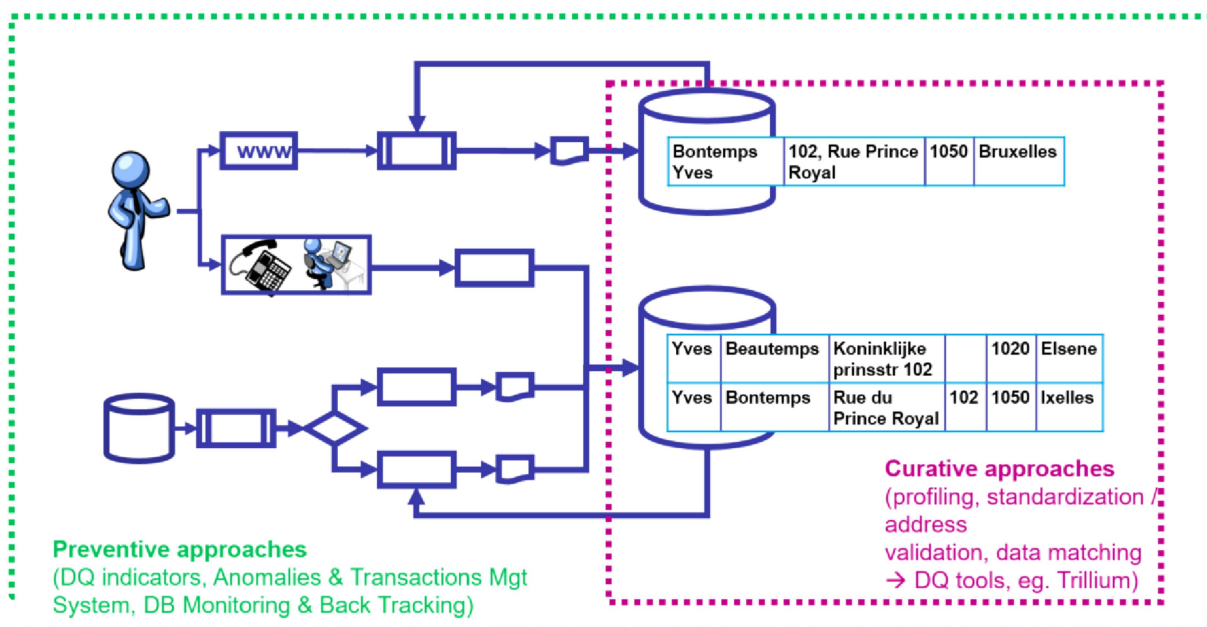


Figure 1. Approches curatives et préventives.

Celle-ci est destinée à l'amélioration semi-automatique de la qualité des données :

- on the fly, "online", par exemple dès la saisie d'un enregistrement dans le système (approche API, voir point relatif aux nouveautés)
- en batch, par exemple en traitant périodiquement l'entièreté d'une ou plusieurs tables dans une base de données préexistante ou un ATMS .

Généralement, les data quality tools couvrent une à trois des grandes familles de fonctionnalités présentées ici, l'outil acquis par Smals incluant les trois dans le cadre d'une suite comportant un grand nombre d'algorithmes de traitement de données génériques ou "ad hoc".

Ceci est fondamental car l'expérience montre que ces trois fonctionnalités (Figure 2) sont interdépendantes de manière cyclique et itérative en synergie avec le business (qui doit être impliqué aux côtés de l'IT, s'agissant d'un outil commercial demandant une importante courbe d'apprentissage). Nous les présentons brièvement :

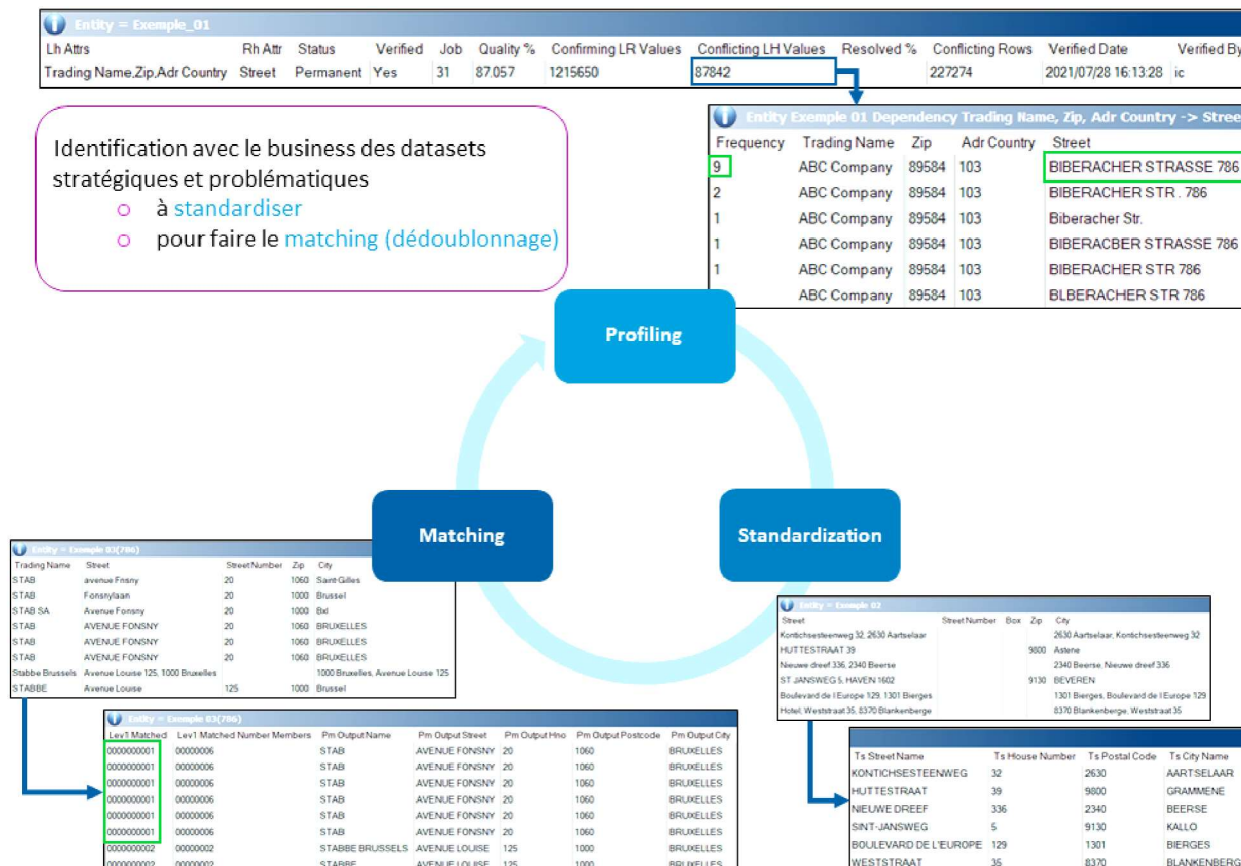


Figure 2 : Profiling, standardisation et matching : les trois familles de fonctionnalités majeures et itératives.

- **profiling** : analyser qualitativement et quantitativement des données pour en évaluer la qualité, isoler ou quantifier des problèmes déjà connus mais dont l'ampleur n'a jamais été évaluée et, souvent, débusquer automatiquement et semi-automatiquement des problèmes inattendus. Exemple : distribution de la longueur des valeurs d'une colonne, inférence de type, vérification ou découverte de dépendances fonctionnelles ;
- **standardisation** : transformer les données en vue des les conformer à un standard défini avec le business ou à un référentiel existant ("data cleansing"), pouvant être fourni avec l'outil. Exemple : nettoyage et uniformisation de la représentation des numéros de téléphone, correction, enrichissement et validation d'adresses postales. L'outil acquis par Smals est particulièrement puissant dans cette dernière fonctionnalité qui couvre la Belgique et est en voie d'extension (voir nouveautés ci-dessous) ;
- **comparaison, détection d'incohérences et dédoublonnage**, via des algorithmes de **matching** (qui se déclinent en familles bien spécifiques sur le plan théorique) : détecter les duplicats et incohérences dans les enregistrements au sein d'un jeu de données ou entre plusieurs (issus potentiellement de bases de données distinctes, en vue d'une intégration ou dans le cadre d'un reengineering, par exemple). La comparaison se base sur des colonnes discriminantes et des algorithmes tolérants à l'erreur (mesure de la distance d'édition, comparaison de l'empreinte phonétique, etc.), déterminés avec le business. Les outils les plus avancés permettent ici de conserver et lier les enregistrements originaux pertinents (après validation par le Business) sans les écraser. Les meilleures valeurs identifiées pour chaque colonne serviront à construire le «survivor » ou « golden record », représentant chaque grappe ainsi repérée et utilisé pour dédoublonner le(s) jeu(x) de données si nécessaire. Notons que la problématique est telle que les règles d'établissement d'un "golden record" sont formalisées dans la loi ou dans des règlements administratifs pour certaines sources authentiques, telles que le Registre National ou la Banque Carrefour des Entreprises belges, par exemple. Enfin, vu le nombre de records à comparer entre eux et

d'opérations associées, des **mécanismes de gestion de la performance** ("blocking" ou "windowing") doivent être utilisés de manière itérative dans les opérations de matching d'envergure.

Deux points importants :

- les fonctionnalités de "**drill down**" de l'outil permettent un échange aisé entre l'IT et le business quand des enquêtes intellectuelles doivent être réalisées sur des résultats donnés, comme le montrent les fonctionnalités de la figure 2, le drill down s'appliquant aussi au matching.
- tout projet "data quality" doit être **documenté**, à différents niveaux de granularité, en fonction du public (IT ou business) visé ; cela vaut donc bien sûr pour les projets reposant sur des data quality tools.

Typiquement, ces outils **interviennent en « batch »**, c'est-à-dire en ciblant, en différé, un ou plusieurs jeux de données déjà existants. Certains permettent cependant également d'intervenir plus en amont, en exposant ces fonctionnalités sous la forme d'une **API** que l'application peut appeler au cas par cas au moment où les données entrent dans le système. Nous revenons sur cette fonctionnalité au point suivant car il s'agit d'une nouveauté qui est en cours de standardisation dans la suite des fonctionnalités dont dispose Smals en complément du batch (*Figure 3*).

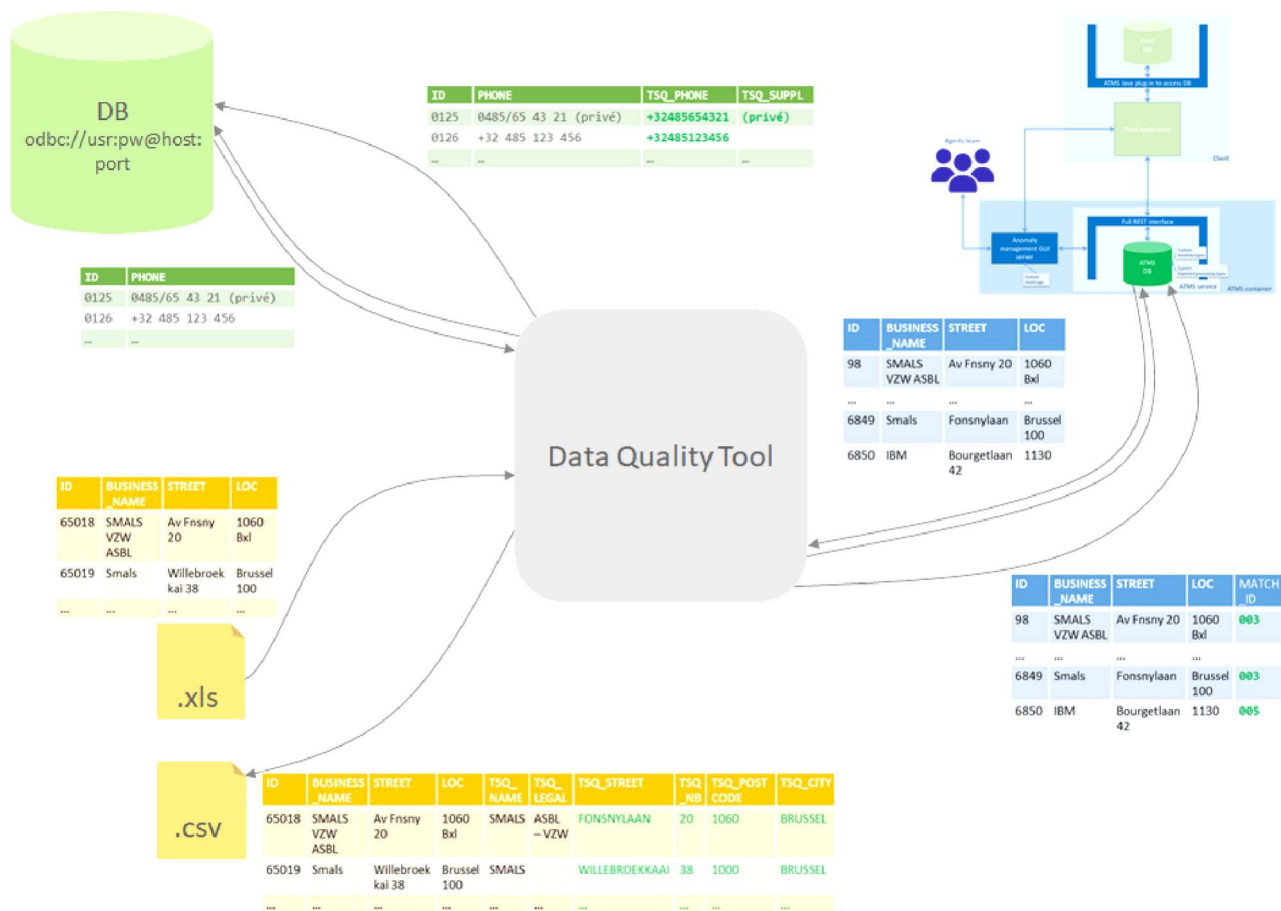


Figure 3. Data Quality Tool : approche batch

2. Nouveautés

2.1. Une REST API standardisée

La suite de fonctionnalités Batch, fondamentale pour les gros volumes de données, est maintenant complétée par une REST API développée en 2020 et en cours de standardisation en 2021 (Figure 4). Celle-ci permet de transformer des valeurs en vue de les conformer à un standard défini ("data cleansing"), de détecter des incohérences ou de dédoubler les données avant leur écriture dans la base et même, si besoin, de conditionner cette écriture par la réussite des opérations qui la précèdent. L'outil implémente ainsi effectivement un pare-feu de données complémentaire au système de détection d'anomalies déjà mis en place par l'application. L'API est par exemple utile pour la standardisation de numéros de téléphone, la validation d'adresse ou la vérification de l'existence de duplicats dans une base de données avant l'insertion d'un nouveau record. L'API `dataQualityImprovement` est documentée sur l'ict-use, au fil de son enrichissement.

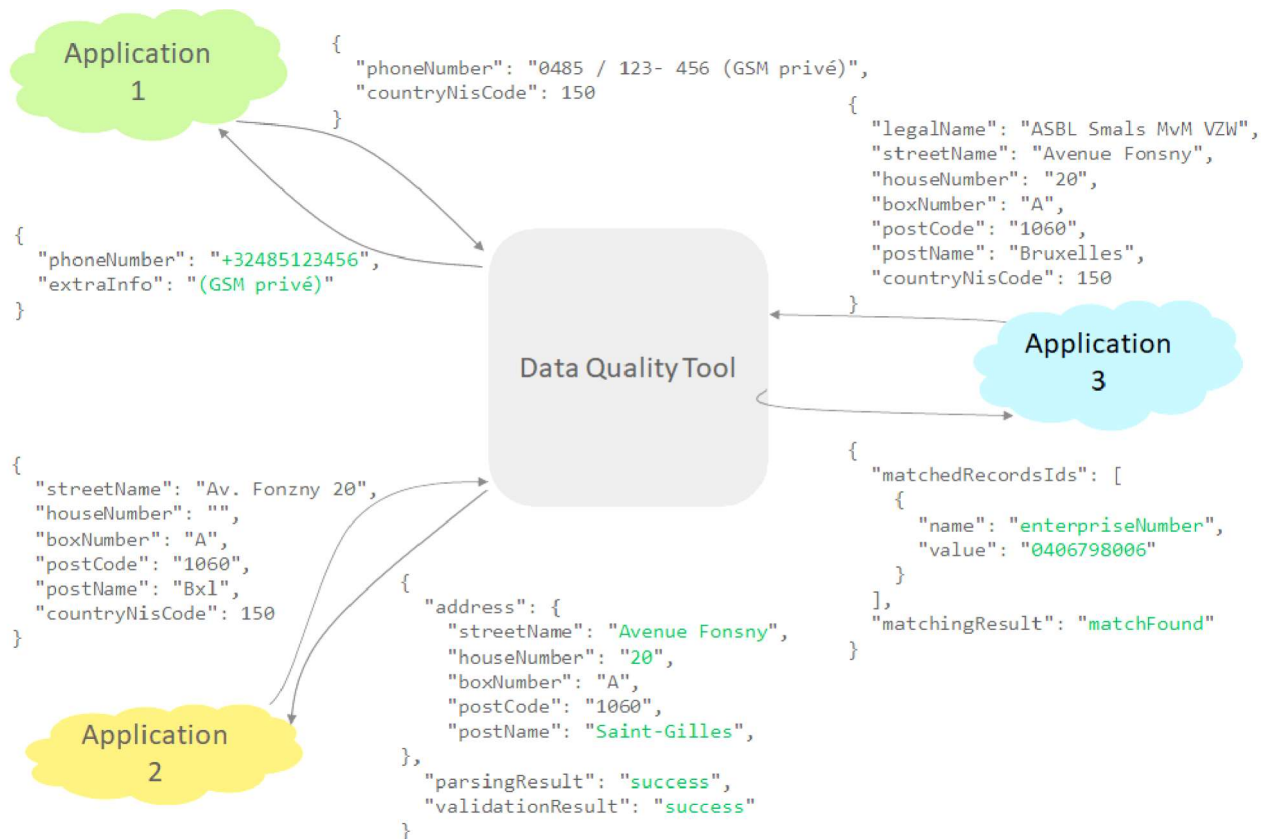


Figure 4. Data Quality Tool : REST API standardisée

2.2. Une extension de la fonctionnalité de validation d'adresses à l'international via OpenStreetMap

Une autre nouveauté en 2021 réside dans l'extension de la fonctionnalité de correction et de validation d'adresses. Particulièrement puissante pour la Belgique dans le Data Quality Tool dont dispose Smals, cette fonction est en cours d'extension pour inclure désormais les adresses internationales via OpenStreetMap, outil de cartographie Open Source et collaboratif pour lequel une API a été développée au sein de la section Recherche.

La qualité des adresses traitées avec ce nouveau module pourra, selon les cas, être moindre que celle concernant les adresses belges, et la performance sera variable en fonction des projets. Cela dit, ce développement n'en demeure pas moins très intéressant pour toutes les bases de données incluant des adresses étrangères et demandant un traitement Batch ou via une REST API en ayant recours aux fonctionnalités du data quality tool évoquées plus haut.

La possibilité d'intégrer les données issues d'OpenStreetMap à l'outil dont dispose Smals permet également d'envisager, à moyen terme, la mise en place d'un service de géocodage en complément des fonctionnalités déjà existantes.

2.3. Data quality et machine learning

Dans un futur proche, les liens entre les "Data Quality Tools" et le Machine Learning seront envisagés pratiquement, notamment pour répondre aux deux questions suivantes :

- Comment améliorer la qualité des "Big Data" en amont afin de rendre les résultats du ML plus adéquats aux attentes des utilisateurs ? Les données alimentant les modèles de ML se distinguent souvent tant par leur volume que par une certaine tolérance aux données aberrantes ; la qualité des données reste cependant un composant critique dans la construction de modèles fiables et durables.
- Comment le ML peut-il enrichir le résultat de certains algorithmes au coeur des "Data Quality Tools" au regard du "fitness for use" ? L'approche déterministe a l'avantage de fournir des résultats relativement prévisibles mais requiert une connaissance préalable des conditions précises auxquelles une opération (par exemple un "match" positif) peut avoir lieu ; il pourrait être intéressant d'étudier dans quelle mesure le recours à des algorithmes "apprenant" eux-mêmes ces conditions pourrait contribuer à des résultats encore plus précis.

En conclusion, le service Data Quality chez Smals ne cesse de s'étendre et les applications concrètes abondent. Un point de contact unique existe désormais pour toute question ou demande relative à ce sujet : dataquality@smals.be

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

data quality tools

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](#)