



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

## Linked Open Data Quality Around-The-Clock

Posted on 2011-05-04 by [Isabelle Boydens](#)

Rock Around the Clock. Depuis plus d'une décennie, la qualité de l'information (voir [Data Quality Competence Center](#) de Smals) est considérée comme stratégique dans le monde des bases de données et, en particulier, dans [l'égovernment](#).

En 2011, la question devient maintenant un domaine de recherche à part entière s'agissant du [Web sémantique](#) et du [Linked Open Data](#). Après l'enthousiasme que suscite toujours une nouvelle technologie "hype", apparaissent inévitablement les questions opérationnelles et pratiques liées au "boomerang du réel".

Le jeune et provocateur [Pedantic Web Group](#), source d'information récente sur la question, l'illustre en ces termes :

*"You publish RDF data on the web, and thereby contribute to our shared passion: the emerging global information space that we call the web of data. Thank you for that! Thank you for sharing your data!  
But your data is broken. Syntax errors, unescaped characters, encoding problems, broken links, ambiguous identifiers, undefined vocabulary terms, mismatched semantics, unintended inferences: If you publish anything on the web of data, chances are that there is some problem. "*

Dans le cadre du [Data Quality Competence center](#) de Smals, nous avons, dès 2004, étudié la technologie et annoncé les risques liés à la qualité des données sémantiques (Research Note 5

["Du web sémantique au web pragmatique"](#)) et sur ce blog, en 2009, à [l'oppos de la norme SKOS](#), Simple Knowledge Organization System, norme d'interopérabilité reposant sur le langage [RDF](#).

Fin 2010, Christian Fürber (Universität der Bundeswehr München) introduisait en ces termes le Quality Constraints Library (Open Source Semantic Data Quality Tool reposant sur le le SPARQL Inferencing framework) : *"The Semantic Web allows the publication of data by anyone for everybody. A tremendous amount of data is already available on the web which may be used to retrieve information or integrate data into applications to enhance automation. If we do not care about the quality state of the published data, we will soon have a lot of poor data on the Semantic Web leading to disappointing experiences."*

La question du temps est centrale pour comprendre la problématique de la qualité des données, en particulier sur le Web. En effet, dans le cadre d'Internet, environnement ouvert et non contrôlé, on assiste à une prolifération de liens hétérogènes. C'est ainsi que le décalage temporel entre les données et le réel observable correspondant prend dans les faits une ampleur considérable. On trouve ainsi, pour ne citer qu'un exemple, sur DBpedia différentes températures minimales pour la planète Terre, les valeurs Kelvin et Celsius ayant été converties de manière hétérogène à des moments différents.

L'intitulé du 7eme Programme cadre en cours que la Commission européenne consacre à la question va dans ce sens : "Linked Open Data Around the Clock" . Le premier point du programme a en effet pour objectif de déployer une infrastructure *"to continuously monitor and improve the quality of data links within the Linking Open Data cloud"*.

Si de nombreuses similitudes existent entre l'approche "data quality" relative aux bases de données et au Web sémantique, s'agissant de ce dernier, plusieurs éléments complémentaires méritent d'être signalés car la question de la qualité de l'information est y est plus complexe encore à maîtriser. En effet, la cohérence des données est d'autant plus problématique que les triplets RDF peuvent inclure des informations issues de sources hétérogènes et mises à jour à des rythmes différents, ce qui donne lieu, par exemple, à des "dead links" ou erreurs de format.

Pour remédier ces questions, plusieurs "best practices" sont conseillées. Citons à titre d'exemple :

- le recours à des vocabulaires de mise en cohérence comme, par exemple, I'Open Provenance Model Vocabulary Specification

- l'association aux graphes RDF d'une signature digitale : un des problèmes majeurs résidant dans l'identification des sources à l'origine de l'information et de leur crédibilité
- le recours à des "linked data principles" proposés par Tim Berners Lee, en personne
- le recours à des "data quality tools open source" dédiés permettant d'effectuer du "profiling" et de la "standardisation" sur des données RDF, comme la "Quality Constraints Library" publiée par de Christian Fürber fin 2010, disponible sous la GNU Lesser General Public License.
- Plus que jamais, le principe du "fitness for use", de l'adéquation aux usages, reste d'actualité. Comme le note C. Furber, "*Before you can identify data quality problems in your data set, you need to define "what means high data quality for your domain?"*". La réponse à cette question passe encore et toujours par un travail d'analyse préliminaire incontournable. Celui-ci est d'autant plus coûteux que la qualité des données n'a pas été prise en compte au seuil d'un projet sémantique et qu'on doit traiter la question a posteriori.

A l'heure où les projets liés à l'interopérabilité sémantique et à "l'open data linking" se développent tant dans l'egovernment que dans le domaine de l'ehealth, la question de la qualité du réseau sémantique revêt des enjeux stratégiques pour Smals, société ICT au sein de laquelle la gestion de l'information est centrale.

egov

information management

standards

---

#### 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 &amp; 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](#)