**An Evaluation Framework for Linked Data Quality**

Abstract

In the last few years we have noticed the Semantic Web gaining a huge momentum supported by the introduction of many related initiatives like the Linked Open Data. The standardization of the Semantic Web technologies and specifications has resulted in a staggering volume of data being published by both public and private sectors. This data is of great benefit when properly integrated. However, to do so the data provided should be *fit to use.* The fitness of data is commonly conceived as an indicator of its quality. Principles of web data quality have been thoroughly investigated in the literature, however the Semantic Web introduced new data quality issues that need to be addressed and thoroughly investigated.

In this paper, we identify the main quality issues surrounding Linked Data. We list all the principles that are involved at all stages of the data management process from modeling and analysis to storage and presentation. Additionally, a comprehensive list of objective quality indicators and metrics is presented accompanied with the existing tools that measure the various dimensions of the Linked Data quality landscape.

The aim of this paper is to provide researchers and practitioners with a comprehensive understanding of the issues surrounding the Linked Data quality and the existing tools that will help ranking data sources. Moreover, we propose a methodological framework that will incorporate various tools and systems in order to issue a quality certificate that will ensure better data integration thus more informed decisions.

However, the heterogeneous nature of data resources reflects directly on the data quality as these sources often contain inconsistent as well as misinterpreted and incomplete information.

The Linked Open Data cloud is a gold mine for organizations who are trying to leverage external data sources in order to produce more informed business decisions.

However, the nature of document-oriented web requires varying quality measures and assessment techniques. Web information quality can be indirectly defined via algorithms like Page Rank \cite{ Lawrence981}. On the other hand, structured information has more concrete and measurable quality indicators. Assuring data quality in LOD is a challenge especially that it does not only focus on the instances data, but also on the quality of the underlying models, ontologies and vocabularies and links.

Despite the fact that LOD quality is a trending and highly demanded topic, few efforts are currently trying to standardize, track and formalize frameworks to issue scores or certificates that will help data consumers in their integration tasks.

Afterwards, we list the results of our research for existing tools and frameworks that assess or address certain attributes and identify the gaps in these tools. In the end, we propose a comprehensive evaluation framework to measure the quality of Linked Data resources and issue a certificate

We are entering an era where open is the new default. Governments, universities, organization even individuals are publishing huge amounts of open data to the public. This openness should be accompanied with certain level of trust or guarantees about the data quality. Currently, we only know of one certificate that data publishers can have to describe the quality of their data. The ODI certificate provides guarantees about a data set quality in plain English. It acts as a mark of approval that helps understand the quality of data and how it can be used. It gives the ability for publishers to provide assurance and support and enable them to measure the quality of their data and improve that clearly through an ascending scale.

It is a questionnaire on which data publishers have to answer certain questions that will reflect certain characteristics about their data; the questions are classified into the following criteria:

* **General Information**
  + Data title
  + Documentation URL (where is the data described)
  + Data publisher (who is the publisher of this data)
  + Publisher URL/ dataset URL (where is this data published)
  + What kind of release is it
    - One-off release of a single dataset: A single file release with no future plans of publishing similar files in the future
    - One-off release of a set of related datasets: A collection of related files about the same data with no intention to publish similar collections in the future
    - Ongoing release of a series of related datasets: A sequence of datasets with planned periodic updates
    - A service or API for accessing open data): Live web service that exposes data through a query-able interface
* **Legal Information**
  + Indicate if the publisher has the rights to publish the data as open
* **Licensing Information**
  + URL of the copyright and database rights statement
  + License information (under which license can people reuse this data)
* **Privacy Information**
  + State whether individuals can be identified from this data or not. For example road traffic flow data combined with individual commuting patterns can reveal identity. However this can be overcome by applying some statistical analysis techniques and aggregating data.
* **Practical Information**
  + how easy is it to find the data, they define a three clicks threshold from the home page
  + If the data listed within a collection. They highlight the fact that data is generally easier to find when it is in relevant data catalogs
  + If the data is referenced from the publisher's publications such as reports, presentations or blog posts.
* **Quality Information**
  + URL that pints out to where the publishers documents issues with the quality of his data, they believe that this an important to metric for people to judge on the data reliability
  + Description of the quality control process. Provide the users with links to know about how publishers perform automatic or manual checks on the data. This is a measure of the publisher’s seriousness and willingness to improve the data quality.
* **Reliability Information**
  + How long will the data be available for? Is it going to:
    - Disappear at anytime
    - Available experimentally but should be around for another year or so
    - In medium-term plans, should be around for a couple of years
    - A part of the publishers day-to-day operations which means that it will be published for a long term
* **Technical Information**
  + If the data provided is presented in a machine readable format
  + If the data is in a standard open format CSV, XML, JSON ...
  + The type of the published data. Is the data meant for:
    - Human consumption like policy documents, white papers or reports and meeting minutes. This can give an indication about the structure of data as these types are mostly unstructured and textual
    - Statistical data. This data can be census results, traffic flow information or crime statistics for example
    - Geographic information such as points, lines and boundaries
    - Other kinds of structured data like event details, contact information or anything that can be interpreted as data and analyzed and presented in multiple ways
  + Is the data presented using persistent identifiers? This is really important if data coming from different resources use the same persistent and unique identifier to refer to the same things. At that point, people can combine these resources easily. These identifiers can be GUIDs or DOIs or URLs
  + If the publisher provides machine-readable provenance of the data. This helps in building trust in the data published as people can trace back and know how it was handled
  + URL for the verification process to ensure that the data they receive come from the data publisher
* **Social Information**
  + Do the data formats use vocabularies or schemas like CSV, JSON, XML or turtle
  + Are there any codes used in this data? does it refer to things like geographical areas or diseases for example
  + Who should people contact with questions about the data?
  + Where can people find out how to improve the way the data is published
  + Who should users contact with questions regarding the data privacy
  + Do you use social media to connect with users who use the data
  + Is there a URL for a forum or a mailing list where people can talk about the data
  + Are there any active communities built around this data
  + Where do you list tools or recommended usage of the data

Different requirements in these levels map to three main badges that reflects the conformance of a dataset to them. The badges are:

* **Pilot**: Data users receive extra support from and can provide feedback to the publisher
* **Standard:** Regularly published open data with robust support that people can rely on
* **Expert:** An exceptional example of information infrastructure

<https://certificates.theodi.org/>

Although ODI is a great initiative, the issued certificates are “self-certified”. This means that the ODI does not verify or review submissions in any way, while retaining the right to revoke a certificate at any time. We believe that the ODI approach comes short especially when dealing with large data sets.The dynamicity of the Linked Data makes it very difficult to keep track of changes and updating the issued certificates manually, especially when these changes are frequent and affect large and diverse data points.

We believe that we need an automatic certification approach that is not only able to rank the LOD datasets but also keep up with its dynamic nature and frequent updates. This certification process should exist side by side with a manual process where the publishers specify several aspects discussed in the ODI certificate.