A Web Observatory for the Machine Processability of Structured Data on the Web

Wouter Beek
VU University Amsterdam
De Boelelaan 1081a
Amsterdam, The Netherlands
w.g.j.beek@vu.nl

Paul Groth
VU University Amsterdam
De Boelelaan 1081a
Amsterdam, The Netherlands
p.t.groth@vu.nl

Stefan Schlobach
VU University Amsterdam
De Boelelaan 1081a
Amsterdam, The Netherlands
k.s.schlobach@vu.nl

Rinke Hoekstra
VU University Amsterdam
De Boelelaan 1081a
Amsterdam, The Netherlands
rinke.hoekstra@vu.nl

ABSTRACT

General human intelligence is needed in order to process content on the Web of Documents. On the Web of Data (WoD), content is intended to be machine-processable as well. But the extent to which a machine is able to navigate, access, and process the WoD has not been extensively researched. We present LOD Observer, a web observatory that studies the Web from a machine processor's point of view. We do this by reformulating the five star model of Linked Open Data (LOD) publishing in quantifiable terms. Secondly, we built an infrastructure that allows the model's criteria to be quantified over existing datasets. Thirdly, we analyze a significant snapshot of the WoD using this infrastructure and discuss the main problems a machine processor encounters.

Categories and Subject Descriptors

E.m [**Data**]: Miscellaneous; H.3.5 [**Information Systems**]: Information Storage and Retrieval—*On-line Information Services*

Keywords

Web Observatory; Machine processing; Web of Data; Linked Open Data

1. PROBLEM STATEMENT

There is an increasing amount of structured and semi-structured data available on the Web [4]. This data is made available in a variety of formats[1]. Initiatives such as the Open Knowledge Foundation (OKFN) and the Linking Open Data (LOD) community have promoted the release of data in an open fashion with the explicit goal of facilitating ease of reuse by others.

WebSci'14, June 23–26, 2014, Bloomington, IN, USA.
ACM http://dx.doi.org/10.1145/2615569.2615654.

Table 1: Five Star Linked Open Data

⋆	Make your data available on the web with an open license.
⋆⋆	Make it available as machine-readable structured data (e.g. excel instead of image scan of a table).
⋆⋆⋆	As above, plus use a non-proprietary format (e.g. CSV instead of excel).
⋆⋆⋆⋆	As above, plus use open standards from W3C to identify things, so that people can point at them.
⋆⋆⋆⋆⋆	As above, plus link your data to other people's data to provide context.

While facilitating reuse by humans is important, one of the initial aims of the Web of Data was that data be accessible and usable by machines in an *automatic fashion* [3]. Indeed, the W3C set of standards for exposing data on the Web is specifically designed to enable machine reasoning [5]. However, the data currently available is often far from machine-friendly. Even data made available using Semantic Web standards is rife with quality issues making them difficult to process by machines [6].

2. 5-STAR DATA

In 2006 Tim Berners-Lee published an opinion piece on LOD publishing [2], which became known as the "5 star model". This model (Table 1) quickly obtained the status of a manifesto for *the right way* to publish data: webby, machine readable, non-proprietary, standards conformant and linked.

Now, 7 years later, the community tends to believe that LOD publishing according to the 5-star model has become the de facto standard. There are three problems with this claim, though. First, as the manifesto is just a manifesto it is conceptual rather than operational. This means that there is no simple checklist to which datasets have to comply in order to receive the 5 star predicate. The lack of such explicit criteria implies that investigating adherence to the manifesto is impossible to automate, which then means that claims about success or failure of LOD publishing remain vague and difficult to quantify.

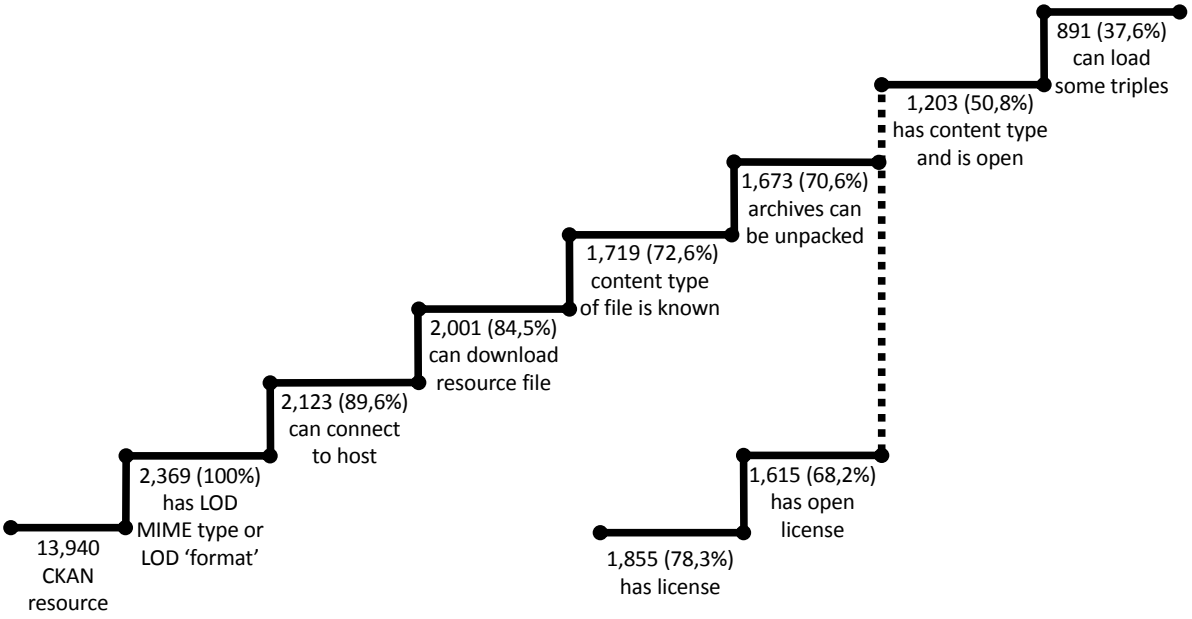

Figure 1: A 'staircase' overview of the succes rates of the various tasks a machine agent has to perform in order to retrieve LOD. The second stair from the left is the sample we have chosen to run our script on (set to 100%); the percentages that appear in the other stairs are relative to this number. The stair in the top right corner shows that 37,6% of the Datahub resources are fully machine-processable.

3. CONTRIBUTIONS

We tackle the problem of providing an up-to-date view on the machine friendliness of the Web of Data. We address this problem in three ways: first we argue for an operationalization of the 5 criteria for LOD publishing from the manifesto. With this operationalization, checking for compliance with LOD principles can now be automated. To this end we implement a Web Observatory, called LOD Observer, which collects and analyses thousands of datasets, and measures and reports on their machine friendliness. This allows us to focus on specific aspects that prevent a dataset from being consumed and processed by a software agent. This analysis (Figure 1) gives a far more detailed and shaded picture of the state of LOD publishing than previous analyses have provided. These results show that in its current state the Web of Data is not yet machine friendly.

We have run our operationlized Web Observatory on a specific CKAN repository: Datahub (`http://datahub.io/`). It contains 13,940 descriptions of data documents, 2,369 of which can be identified as containing LOD based on MIME content type mappings. From this sample, 2,123 data documents (89,6%) have a host to which LOD Observer can connect, and 2,001 (84,5%) have a data file that can be retrieved by using the designated communications protocol (e.g. HTTP(S)). For the sample of 2,369 resources LOD Observer retrieved, 540 (22.8%) did not have a license associated with it. 240 (10.1%) resources have a closed license, and 1,615 (68.2%) resources have an open license. Not all files with an associated open license have syntactically well-formed contents. LOD Observer can load some triples (i.e. one or more) for 891 resources, or 37,6% of the original sample.

4. CONCLUSION

We conclude that in its current state, much of the LOD available on the Web is far from reaching the 5-star level. This is not just a technical issue but a social issue where the dynamics of the Web's social technical system have not reached a point where machine friendly data is widely available. By providing an observatory on the state of the machine processability of Web data, we hope to guide interventions at both the technical and social level. Additionally, this observatory will help in tracking the outcome of those interventions.

5. REFERENCES

[1] M. D. Adelfio and H. Samet. Schema extraction for tabular data on the web. *Proc. VLDB Endow.*, 6(6):421–432, Apr. 2013.

[2] T. Berners-Lee. Rdf schema 1.1. Technical report, 2006.

[3] T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web. *Scientific American*, 2001.

[4] N. Dalvi, A. Machanavajjhala, and B. Pang. An analysis of structured data on the web. *Proc. VLDB Endow.*, 5(7):680–691, Mar. 2012.

[5] P. Hayes and B. McBridge. RDF semantics. Technical report, W3C, 2004.

[6] A. Hogan, A. Harth, A. Passant, S. Decker, and A. Polleres. Weaving the pedantic web. In C. Bizer, T. Heath, T. Berners-Lee, and M. Hausenblas, editors, *LDOW*, volume 628 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2010.