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

Beek, W.; Groth, P.; Schlobach, S.; Hoekstra, R.

DOI
10.1145/2615569.2615654

Publication date
2014

Document Version
Final published version

Published in
WebSci’14: proceedings of the 2014 ACM Web Science Conference: June 23-26, 2014, Bloomington, IN, USA

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
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 on the Web in an open fashion with the explicit goal of facilitating ease of reuse by others.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage, and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s). Copyright is held by the author/owner(s).

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