The WebIndex Data Portal

Jose Emilio Labra Gayo[1], Hania Farham[2], Juan Castro Fernández[1], and Jose María Álvarez Rodríguez[3]

[1] WESO Research Group
{jelabra,juan.castro}@weso.es
[2] The Web Foundation
hania@webfoundation.org
[3] Dept. Computer Science
Carlos III University
josemaria.alvarez@uc3m.es

Abstract. We describe the development of the Web Index linked data portal that represents statistical index data and computations.
In order to empower the Web Index transparency, one requirement was that it should be possible to verify every published data. The resulting portal contains data that can be tracked to its sources so an external agent can validate the whole index computation process. We consider this approach a step towards more quality in linked data publication.

1 Introduction

The WebIndex is a multi-dimensional measure of the World Wide Web's contribution to development and human rights globally. It covers 81 countries and incorporates indicators that assess several areas like universal access; freedom and openness; relevant content; and empowerment [4].

First released in 2012, the 2013 Index has been expanded and refined to include 20 new countries and features an enhanced data set in the areas of gender, Open Data, privacy rights and security.

The 2012 version offered a data portal where the data was obtained by transforming raw observations and precomputed values from Excel sheets to RDF. In this paper, we describe the development of the 2013 version of the WebIndex data portal[5], where we employ a new validation and computation approach that enables the publication of a verifiable linked data version of the Web Index data.

Given that the most important part of a data portal about statistical indexes are the numeric values of each observation we established the requirement that any value published had to be justified either declaring from where it had been obtained or linking it to the values of other observations from which it had been computed.

The resulting data portal [6] contains not only a linked data view about the statistical data but also a machine verifiable justification of the index ranks and computations.

[4] http://thewebindex.org
[5] http://data.webfoundation.org
[6] http://data.webfoundation.org/webindex/2013

2 WebIndex workflow and validation

Two types of data were used in the construction of the Index: existing data from other data providers (*secondary data*), and new data gathered via a multi-country questionnaire (*primary data*) that was specifically designed by the Web Foundation and its advisers. The computation procedure employs common statistical formulae [7]. The raw data was obtained by a team of statisticians in a big Excel file comprised of 184 Excel sheets which contained a combination of raw, imputed and normalized data. That external data was then filtered and converted to RDF by means of an specialized web service called *wiFetcher* [8]. Although some of the imported values had been pre-computed in Excel by human experts, we collected only the raw values, so we could validate the whole computation process.

We implemented an online validation tool called *Computex*[9] which takes as input an RDF graph and checks if it follows the integrity constraints defined by Computex. The validation tool can also check if the RDF graph follows the RDF Data Cube integrity constraints and it can also do the index computation for small RDF Graphs using SPARQL CONSTRUCT queries.

Although this declarative approach was very elegant, computing the webindex using only SPARQL queries was not practical (it took around 15 minutes for a small subset), so the computation process was finally done by the specialized Scala program *wiCompute* [10] which took the raw values and computed the index following the computation steps generating RDF datasets for the intermediary results and linking the generated values to the values from which they had been computed. The new tool took a few seconds to do all the process. The data portal documentation employs a combination of examples and templates using Shape Expressions [11].

Following the linked data principles, we considered that it was necessary to offer not only RDF views but also HTML representations of the different data. We developed a visualization tool called Wesby [12] which takes as input an SPARQL endpoint and offers a linked data browsing experience. Wesby was inspired by Pubby **?** and was developed in Scala using the Play! Framework. Wesby combines the visualization with a set of templates to offer specialized views for different types of resources. [13]. Wesby also handles content negotiation automatically to offer views in HTML, RDF/XML, Turtle, etc.

---

[7] The computation steps are described in http://thewebindex.org/about/the-web-index/
[8] https://github.com/weso/wiFetcher
[9] http://computex.herokuapp.com/
[10] https://github.com/weso/wiCompute
[11] http://weso.github.io/wiDoc/
[12] http://wesby.weso.es
[13] http://data.webfoundation.org/webindex/v2013/country/ESP contains the WebIndex visualization of the country Spain

3 Acknowledgements

We would like to thank Jules Clements, Karin Alexander, César Luis Alvargonzález, Ignacio Fuertes Bernardo and Alejandro Montes for their collaboration in the development of the WebIndex project.