Baseline Statistics of Linked Statistical Data

Andrea Scharnhorst, Albert Meroño-Peñuela and Christophe Guéret

We are surrounded by an ever increasing ocean of information, everybody will agree to that. We build sophisticated strategies to govern this information: design data models, develop infrastructures for data sharing, building tool for data analysis. Statistical datasets curated by National Statistical Offices (NSO) and international bodies like the OECD and Eurostat are being increasingly published on the Web, using Web standards, as Linked Open Data. We call *Linked Statistical Data* (LSD) all statistical datasets published on the Web as Linked Open Data. These datasets do not only consist of current statistics, but also an inheritage of such statistics conducted in the past by these NSO. Of these, the most prominent statistics are the Historical Censuses [1,2,3,5].

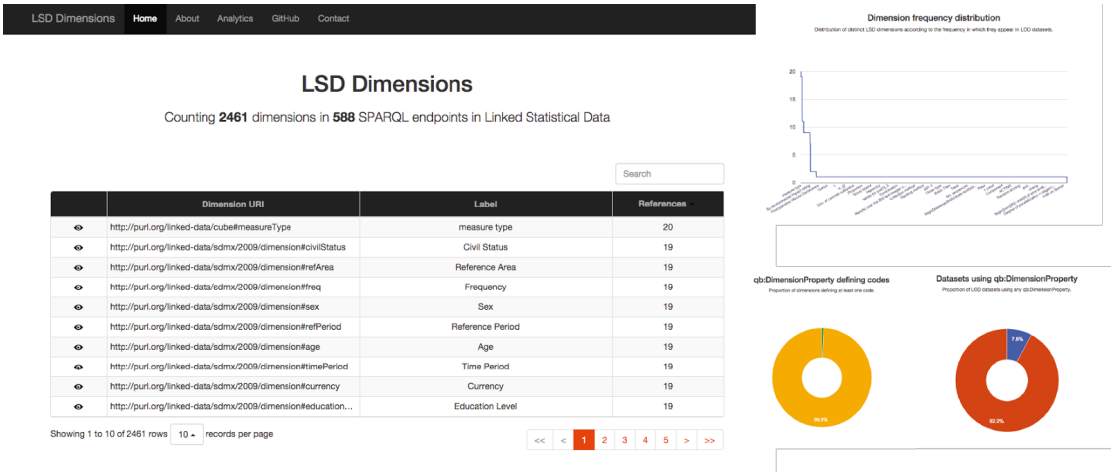
In designing the current research landscape, there seems to be little attention to one very basic inquiry prior to any research design, so basic that it almost seems to be a shame to mention is, and that is a *baseline statistics* on the size of the problem we encounter to tackle, its occurrence, and the the actual need to tackle it. For our own research project [2] it was important to know all these baseline statistics, for instance, how much datasets, variables and values we have, and of which nature [12]. In the context of harmonization we looked for other existing vocabularies to avoid reinventing the wheel. The effort on trying to reuse these vocabularies was the immediate trigger for this exercise.

In this paper we pose the following questions: What kind of dimensions are used in LSD? How often do they occur on the Web? Hereby, dimensions stand for the variables (and the categories used to encode their values) used when describing statistical data inquired to monitor society. Most of those statistical datasets published on the Web use the RDF Data Cube vocabulary [4] to describe the observations, dimensions, measures and attributes they contain. It is the beauty of semantic web technologies which allow us to perform, once every hour, an on-the-flight collection of data from those machine-readable publications of statistical information from almost 600 endpoints all across the Web. Web visualization technologies allow us to build a user interface for this meta study almost without any efforts - for all of those capable of playing with those tools of course. We call it LSD Dimensions, and it is available online [6,7]. The availability of baseline statistics about the usage of multidimensional data on the Web opens up multiple questions, especially regarding comparability and reusability of statistical data, metadata and concepts, and how these can be combined or mixed in a semantically meaningful way [8].

But its implications are much bigger. Insights into the amount of statistical data available allows to better place the need for efforts to build research infrastructure for them. Baseline statistics as shown in this paper also highlight where research efforts are still needed, e.g. on coding schemes that do not exist yet and for which automatic tools that assist knowledge experts on building and publishing them are highly demanded [13].

At the end the innovativeness of this paper lies neither in the visualizations not in the data analytics, but in the smart use of existing technologies to gain overview about the problem space prior to entering into it. This way of thinking about visualizing *baseline statistics* in very much in line with the goals of KnoweScape combining data analytics with visualizations to enhance the

access to large information spaces.
This exercise has been inspired by numerous works from colleagues, which either use visualizations to advocate their own profound efforts of data integration and semantical referenceable data publishing [9,11], or build pipelines of data harvest, analysis and visualization for all kind of other information [10].



References

[1] The Iceland Historical Censuses
http://www.worldcat.org/title/manntalid-1703-population-census-1703/oclc/20410076
[2] Ashkan Ashkpour, Albert Meroño-Peñuela and Kees Mandemakers. The Dutch Historical Censuses: Harmonization and RDF. Historical Methods: A Journal of Quantitative and Interdisciplinary History, 2014 (to appear).
[3] 2.000 US Census in RDF http://datahub.io/dataset/2000-us-census-rdf
[4] Richard Cyganiak, Dave Reynolds, and Jeni Tennison. The RDF Data Cube Vocabulary. Technical report, W3C, 2014. http://www.w3.org/TR/vocab-data-cube/
[5] The Norwegian Historical Censuses http://www.rhd.uit.no/folketellinger/folketellinger_avansert_e.aspx
[6] LSD Dimensions http://lsd-dimensions.org/
[7] LSD Dimensions source code repository at Github https://github.com/albertmeronyo/LSD-Dimensions
[8] Sarven Capadisli, Albert Meroño-Peñuela, Sören Auer, and Reinhard Riedl. Semantic Similarity and Correlation of Linked Statistical Data Analysis. In Proceedings of the 2nd International Workshop on Semantic Statistics (SemStats 2014), ISWC. CEUR Workshop Proceedings, 2014.
[9] Sarven Capadisli, Sören Auer, and Reinhard Riedl. Linked Statistical Data Analysis. http://csarven.ca/linked-statistical-data-analysis
[10] Prov-O-Viz, the PROV-O provenance visualizer https://github.com/Data2Semantics/provoviz
[11] linkitup, the Web-based dashboard for Figshare research output enrichment http://linkitup.data2semantics.org/
[12] Statistics over the CEDAR data integration http://lod.cedar-project.nl/cedar/stats.html
[13] Albert Meroño-Peñuela, Ashkan Ashkpour, Christophe Guéret. “From Flat Lists to Taxonomies: Bottom-up Concept Scheme Generation in Linked Statistical Data”. Proceedings of the 2nd International Workshop on Semantic Statistics (SemStats 2014), ISWC 2014, Riva del Garda, Italy (2014).