LSD Dimensions*†: Use and Reuse of Linked Statistical Data

Albert Meroño-Peñuela[1,2]

[1] Department of Computer Science, VU University Amsterdam, NL
albert.merono@vu.nl
[2] Data Archiving and Networked Services, KNAW, NL

Abstract. RDF Data Cube (QB) has boosted the publication of Linked Statistical Data (LSD) on the Web, making them linkable to other related datasets and concepts following the Linked Data paradigm. In this demo we present `LSD Dimensions`, a web based application that monitors the usage of *dimensions* and *codes* (variables and values in QB jargon) over five hundred public SPARQL endpoints. We plan to extend the system to retrieve more in-use QB data, serve the dimension and code data through SPARQL and an API, and provide analytics on the (re)use of statistical properties in LSD over time.

Keywords: Semantic Web, Statistics, Linked Statistical Data

1 Motivation

RDF Data Cube (QB) has boosted the publication of Linked Statistical Data (LSD) as Linked Open Data (LOD) by providing a means "to publish multi-dimensional data, such as statistics, on the web in such a way that they can be linked to related data sets and concepts" [4]. QB defines *cubes* as sets of *observations* consisting of *dimensions*, *measures* and *attributes*. For example, the observation "the measured life expectancy of males in Newport in the period 2004-2006 is 76.7 years" has three dimensions (*time period*, with value *2004-2006*; *region*, with value *Newport*; and *sex*, with value *male*), one measure (*population life expectancy*) and two attributes (the units of measure, *years*; and the metadata status, *measured*[3]). In some cases, it is useful to also define *codes*, a closed set of values that a dimension can get (e.g. sensible codes for the dimension *sex* could be *male* and *female*).

There is a vast diversity of domains to publish LSD about, and lots of dimensions and codes can be heterogeneous, domain specific and hardly comparable [2,3,5,6]. To this end, QB allows users to mint their own URIs to create arbitrary dimensions and associated codes. Conversely, some other dimensions and codes are quite common in statistics, and could be easily reused. However, publishers

*Web application at `http://lsd-dimensions.org`

†Source code at `https://github.com/albertmeronyo/LSD-Dimensions/`

[3] Other metadata statuses could be e.g. *estimated* or *provisional*

```
PREFIX qb: <http://purl.org/linked-data/cube#>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT DISTINCT ?dimensionu ?dimension ?codeu ?code
WHERE {
?dimensionu a qb:DimensionProperty ;
rdfs:label ?dimension .
OPTIONAL {?dimensionu qb:codeList ?codelist .
?codelist skos:hasTopConcept ?codeu .
?codeu skos:prefLabel ?code . }
} GROUP BY ?dimensionu ?dimension ?codeu ?code ORDER BY ?dimension
```

Listing 1.1: SPARQL sent to all endpoints to retrieve LSD dimensions and codes.

of LSD have no means to monitor the dimensions and codes currently used in other datasets published in QB as LOD, and consequently they cannot (a) link to them; nor (b) reuse them.

This is the motivation behind our demoed tool. `LSD Dimensions` monitors the usage of existing dimensions and codes in LSD. It allows users to browse, search and gain insight into these dimensions and codes. We depict the diversity of statistical variables in LOD, and we improve their reusability.

In Section 2 we describe the current `LSD Dimensions` system, available online at `http://lsd-dimensions.org/`. Source code is available at `https://github.com/albertmeronyo/LSD-Dimensions/`. In Section 3 we discuss extensions that will be added to `LSD Dimensions` in the future.

2 LSD Dimensions

Retrieving LSD dimensions. The system queries automatically the CKAN API of Datahub.io and retrieves the most up-to-date list of publicly available SPARQL endpoints (582 at the moment of writing this paper) in the LOD cloud. Once every hour, these endpoints are sent the SPARQL query shown in listing 1.1. The query retrieves all defined `qb:DimensionProperty` (dimensions) in each endpoint, and optionally all resources belonging to a `qb:CodeList` (codes) and associated to each `qb:DimensionProperty`, together with their labels (if available). The system stores all data in a NoSQL (MongoDB) database.

User interface. `LSD Dimensions` provides two different views to allow users browse dimensions (3098 at the moment of writing this paper).

The main view is shown in Figure 1. It shows the full list of retrieved dimensions, listing their URIs, labels (if available) and *references* (a count of how many times that dimension is referred by the endpoints). By default, the list is sorted in descending order of references, but it can be sorted by any other field. To enhance the browsing experience, the main view provides two functionalities: (a) pagination customization (page browsing and number of results per page); and (b) a search feature that looks up the given string in the dimension labels and URIs, filtering the list accordingly.

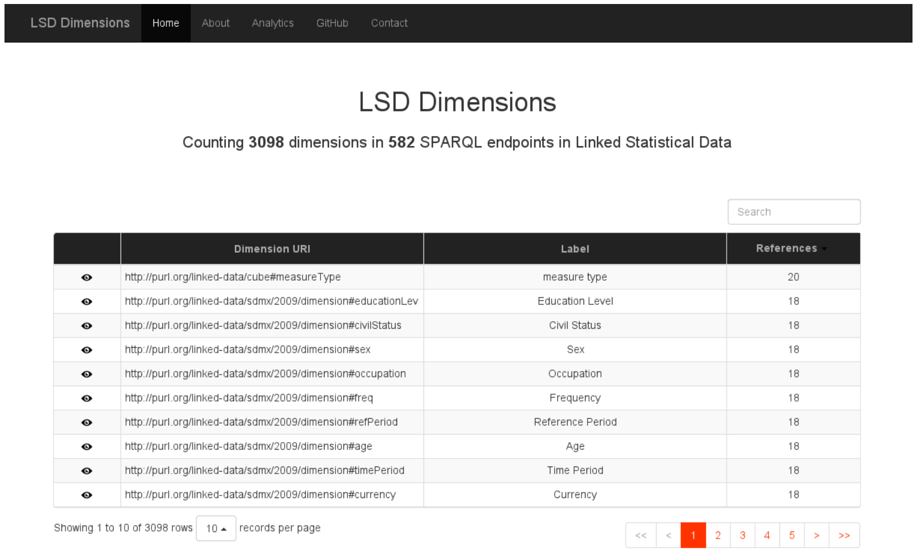

Fig. 1: Screenshot of the main view of the LSD Dimensions user interface, listing retrieved dimensions from the LOD cloud in the last hour. Users can browse and search through the results.

Users can get insight into a chosen dimension by clicking on the corresponding eye icon, leading to the dimension details view, shown in Figure 2. This view shows the list of SPARQL endpoints that use the selected dimension, together with a list of assigned codes to that dimension (if any). All URIs are clickable and users can browse through their dereferenced representations.

By clicking on the *Analytics* tab, users can get an overview of the current usage of dimensions and codes in LSD: (1) a line chart shows the *logarithmic law* followed by the usage of dimensions; (2) a pie chart shows the ratio of SPARQL endpoints defining at least one QB dimension (48, 8.2% over all endpoints); and (3) another pie chart shows the ratio of dimensions defining at least one code (26, 0.8% over all dimensions). Defined codes are thus very scarce.

3 Future Extensions

The current implementation of `LSD Dimensions` monitors the hourly usage of dimensions and codes in Linked Statistical Data. We wish to further extend `LSD Dimensions` in several ways. First, we will retrieve dimensions and their associated values directly from dataset `qb:Observation` observations, in addition to the current definitions. Second, we will share the collected data by (a) modelling it conveniently in RDF and using standard vocabularies; (b) making it available via a SPARQL endpoint; and (c) offering an API so that client applications can get suggested dimensions to link to, improving reusability of dimensions in LSD. Third, we will monitor additional dimension metadata, like associated

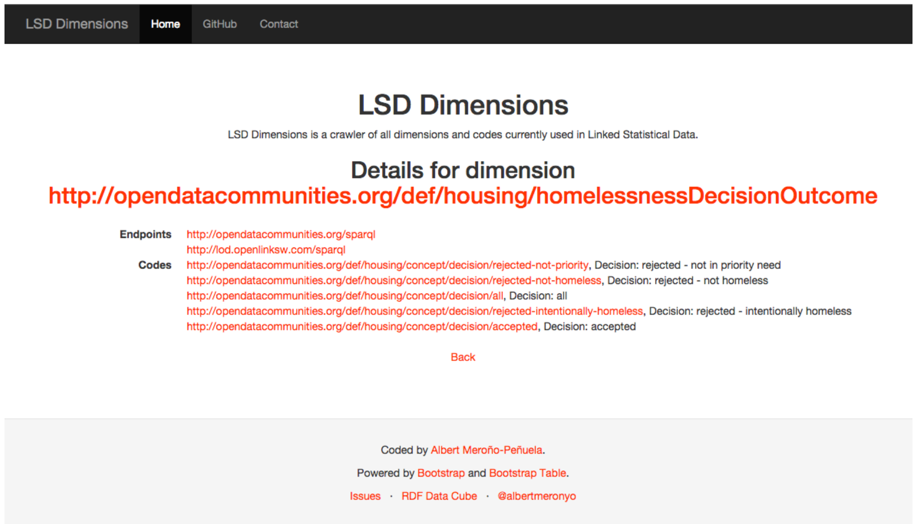

Fig. 2: Screenshot of the dimension details for a chosen dimension, including endpoints using the dimension and popular codes assigned to it.

`rdfs:subPropertyOf` or `rdfs:range` values. Fourth, we will leverage `owl:sameAs` links to match identical dimensions. Fifth, we will run further analyses on the retrieved data (e.g. distribution on the evolution of dimension usage) to better understand practice in the LSD community. Finally, we plan to make use of [1] to study how much LSD is left out of the SPARQL endpoints in Datahub.io.

Acknowledgements. This work was supported by the Computational Humanities Programme of the KNAW (see http://ehumanities.nl) and the Dutch national program COMMIT.

References

1. Beek, W., Rietveld, L., Bazoobandi, H.R., Wielemaker, J., Schlobach, S.: LOD Laundromat: A Uniform Way of Publishing Other People's Dirty Data. In: ISWC 2014 (2014)
2. Capadisli, S., Auer, S., Riedl, R.: Linked Statistical Data Analysis. In: Proceedings of the 1st International Workshop on Semantic Statistics (SemStats 2013), ISWC. CEUR (2013)
3. Capadisli, S., Meroño-Peñuela, A., Auer, S., Riedl, R.: Semantic Similarity and Correlation of Linked Statistical Data Analysis. In: Proceedings of the 2nd International Workshop on Semantic Statistics (SemStats 2014), ISWC. CEUR (2014)
4. Cyganiak, R., Reynolds, D., Tennison, J.: The RDF Data Cube Vocabulary. Tech. rep., World Wide Web Consortium (2013), http://www.w3.org/TR/vocab-data-cube/
5. Meroño-Peñuela, A.: Semantic Web for the Humanities. In: The Semantic Web: Semantics and Big Data, 10th European Semantic Web Conference. LNCS 7882. pp. 645–649. Springer (2013)
6. Meroño-Peñuela, A., Guéret, C., Hoekstra, R., Schlobach, S.: Detecting and Reporting Extensional Concept Drift in Statistical Linked Data. In: 1st International Workshop on Semantic Statistics (SemStats 2013), ISWC. CEUR (2013)