Relationships in semantic data cubes

FACULTY OF MATHEMATICS AND PHYSICS
Charles University

Alexandr Mansurov and Olga Majlingova Charles University and Czech Technical University, Prague

Introduction

Linked data is a method for publishing **structured data while preserving its semantics.** Nowadays, there are many linked data sources available, either within enterprise knowledge graphs or as linked open data publicly available on the WebThere are many situations where it would be useful to be able to publish multi-dimensional data, such as statistics or measurements.

The RDF Data Cube Vocabulary [1] provides a way to publish such data on the Web in semantic form. It is a general (OLAP) data cube expressed using RDF that is compatible with the cube model underlying the SDMX. Briefly summarized, the cube is a collection of observations ("table cells"), each containing information about the values on dimensions of the cube and the actual measurements.

When describing such multi-dimensional data, one often needs to reference an existing taxonomy or similar knowledge organization system. For this use case, the Simple Knowledge Organization System (SKOS) [3] defines a standard data model for sharing and linking such knowledge organization systems on the Web.

Current data available

An example of such taxonomy is a description of **geographical entities**. Nowadays, there are several such taxonomies available as **linked open data**. Some of them, such as

- the Czech Registry of Territorial Identification, Addresses and Real Estate [4] or
- the Linked Data Service of the Swiss federal geoportal [5]

come from **the official sources**.

Other, like

- Geonames [6],
- Wikidata [7] or
- open street maps [8],

are **community projects**.

Acknowledgement

This study was supported by the EU funds as part of the project CZ.03.4.74/0.0/0.0/15_025/0013983.

Observation

Common features of those taxonomies include

a **hierarchy of geographical entities** (e.g. countries – regions – cities)

represented either using SKOS or custom ontology. Some of them are already interlinked together.

On a side note, one could ask why to focus on RDF when there is GML available. GML originated in RDF; however, it is very domain focused, whereas linked data in RDF provides us with a more generic toolbox, in this case, an intermediary between "the world of maps" and "the world of tables".

A collection of data published or curated by a single agent, available for access or download in one or more serializations or formats represents a dataset. An open data catalogue can aggregate metadata of such datasets and data services [2]. Our further research focuses on

discovering relations between datasets available as linked data in such catalogues.

Our approach

The RDF Data Cube Vocabulary allows for the following **definition** of how two datasets could be related: **we consider a pair of datasets to be related if they share related resources on their dimensions**.

When such two resources are related then differs according to particular data.

In our case,

if two data cubes reference geographical properties using some well-known hierarchical geographical taxonomies and if the referenced elements are related,

then measurements (and thus the datasets) are being geographically related.

Finally,

the referenced elements are related if they represent the same entities, or one is a sub-entity of the other.

Evaluation

From a practical point of view, it is essential to keep in mind that there are two kinds of data needed to discover a relationship. On the one hand, there are the actual data cubes referencing elements in some code lists. Next to it are the hierarchical taxonomies with relations between individual concepts.

For non-trivial results, the two need to be connected. However, it is more economical

to **find the references into such taxonomies first** and only then

start **connecting a smaller number of elements** of the large graph of geographical entities.

Analyzing the Czech national open data catalogue, NODC [9] also shows that:

- There are "implicit code lists" – elements that form a code list a human analyst can see from the IRI structure – e.g. https://data.cssz.cz/ontology/sdmx/code/sex-F. However, no dataset in the catalogue contains a definition of such concepts nor their collection so that an automated tool [10] could mark them as a code list purely based on the data.
- Sometimes the public sector does not like to directly reference data outside of their control, resulting in only IRIs like https://data.cssz.cz/resource/reference.data.gov.uk/id/gregorian-year/2019 listed in the distributions.

We can **dereference** both examples above

and **get a valid RDF with its semantic definition and a proper linkage**. However,

no dataset in NODC has that extra data as part of their referenced distributions and thus such IRIs:

- cannot be attributed to any explicit dataset in NODC that could be then marked as related with the dataset referencing it,
- could be considered as an external resource and skipped from further processing completely.

Conclusion

We demonstrated there are rich taxonomies available for semantic identification of the location. One can use these taxonomies for labelling of scientific measurements, now expressed as an RDF data cube. **Such form of data publication will enable discovery of related data based on semantic similarity.**

References

[1] The RDF Data Cube Vocabulary, W3C Recommendation, 16 January 2014, Editors: Richard Cyganiak, DERI, NUI Galway, Dave Reynolds, Epimorphics Ltd, Contributors Jeni Tennison , Available from: http://www.w3.org/TR/2014/REC-vocab-data-cube-20140116/

[2] Data Catalog Vocabulary (DCAT) - Version 2, W3C Recommendation, 04 February 2020, Editors: Riccardo Albertoni (CNR - Consiglio Nazionale delle Ricerche, Italy), & al.Available from: https://www.w3.org/TR/2020/REC-vocab-dcat-2-20200204/

[3] SKOS Simple Knowledge Organization System Reference, W3C Recommendation, 18 August 2009, Editors: Alistair Miles, STFC Rutherford Appleton Laboratory / University of Oxford, Sean Bechhofer, University of Manchester, Available from: http://www.w3.org/TR/2009/REC-skos-reference-20090818/

[4] https://www.cuzk.cz/ruian/RUIAN.aspx, https://linked.cuzk.cz.opendata.cz/sparql

[5] https://www.geo.admin.ch/en/geo-services/geo-services/linkeddata.html

[6] http://www.geonames.org

[7] https://www.wikidata.org/wiki/Wikidata:SPARQL

[8] https://wiki.openstreetmap.org/wiki/Sophox

[9] https://data.gov.cz/datasets

[10] https://github.com/eghuro/dcat-dry

Licence