European Public Sector Information Platform

Topic Report No. 2015/09

Linked Data Validation and Quality

Author: Jose Emilio Labra Gayo

Published: November 2015

Table of Contents

Table of Contents 2
Keywords: 3
Abstract/ Executive Summary: 3
1 Introduction 4
2 Linked Data 6
3 Quality of Linked Data Portals 9
4 RDF Validation 11
5 Conclusions and recommendations 15
References 16
About the Author 17
Copyright information 17

Keywords:

Linked Data, Quality, Validation, RDF, SHACL, Shape Expressions

Abstract/ Executive Summary:

This report contains an overview of the linked data principles and the importance of linked data quality. It describes several initiatives with special emphasis on government and public sector data. Given the fact that RDF plays a central role in linked data, the report also includes a specific section on RDF validation, describing the new W3c proposal for data shapes description and validation.

1 Introduction

Linked data was proposed in 2006 as a set of principles and best practices for data publishing on the Web (Heath, 2011). Since its conception, the number of linked data initiatives has been increasing and a large number of datasets have been added to the so-called linked data cloud[1]. As can be seen in (Miller, 2010), some of those initiatives have been related to the government and public sector information domain. Following Schmachtenberg et al (2014), there were 199 datasets related to government, which represents an 18% of all the linked data cloud and an increase of 306% since 2011.

However, the general adoption of linked open data has not yet arrived. The conclusions of a previous ePSI topic report (Dietrich, 2012) were:

"Uptake by the public and private sector and real-world implementations, however clearly fall behind this academic enthusiasm for the technology. Although more and more private companies and Public Sector Bodies start embracing Linked Data Principles and Technologies it appears that real or anticipated barriers of adapting to these technologies remain as too big."

In practice, there have been lots of prototypes and pioneer projects that have been proposed in academic settings and were abandoned later on. Some reasons were already hinted by (Dietrich, 2012):

- The technology appears to be complicated
- The initial investment too big
- The expected benefits too vague to convince both stakeholders in the private and public sector

Another possibility is that linked data projects still lack some common tools and methodologies that are available in more conventional settings to assess and validate data quality. The underlying technology based on RDF was not intended for safe information exchange or application integration using linked data based services. Some techniques that are popular in relational databases or XML that enable to define data schemas and validate data according to them are not available in RDF, which is more inclined towards Open World Assumption where any system can assert anything about any topic. Although that vision is very interesting for the Web of Data, it is necessary to develop techniques that allow heterogeneous data

[1] Linked data cloud: http://lod-cloud.net/

producers/consumers to coexist while at the same time they can also safely validate the data that they are producing and consuming.

In this report, we will survey the main approaches that have been proposed for linked data quality assessment with special emphasis on RDF validation, which is a core part of this process.

The report is structured as follows: **in section 2 we review the linked data principles and give some examples** of linked data initiatives related with **eProcurement and public contracts**. **Section 3 contains some references and a justification of the importance of linked data quality**. We cover the specific problem of **linked data and RDF validation in section 4** and we finally present some conclusions in section 5.

2 Linked Data

The original linked data principles were proposed as[2]:

- Use URIs as names for things.
- Use HTTP URIs so that people can look up those names.
- When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL).
- Include links to other URIs so that they can discover more things.

The first principle promotes the use of **URIs (Uniform Resource Identifiers) or IRIs (Internationalized Resource Identifiers) to denote all the things** that appear in the problem domain that a linked data application is modelling. It is important that those URIs are unambiguous in the sense that those URIs identify one specific thing and not another depending on some context. Also, if there is some declaration that two URIs represent the same thing, it is important to check that everything asserted about the first URI must be consistently asserted for the second. These two properties can improve the trustworthiness of the linked data portal (Ulicny, 2015).

The second principle states that those **URIs should be dereferenceable**, which means that clients can look up the URI using the HTTP protocol and retrieve a description of the resource that is identified by the URI. An important quality metric for linked datasets is the percentage of dereferenceable URIs as well as the quality of the descriptions retrieved. **It is important to differentiate between URIs that represent web documents and URIs that represent concepts or real-world objects.**

The third principle proposes to return useful information using web standards and specifically mentions RDF and SPARQL. The reason is that in the web of data, **clients cannot only be humans but also machines** that have to process the contents in an automatic way. It is important to provide useful information for both types of agents. In the case of humans that access the information using a web browser, the preferred data format is HTML. However, in the case of machines it is necessary to provide other formats like RDF that can automatically and unambiguously be processed. The RDF graph data model, which is based on the use of URIs to represent properties, allows new data to be seamlessly aggregated and integrated into what has been called knowledge graphs or the web of data.

[2] See Design Issues: Linked data: http://www.w3.org/DesignIssues/LinkedData.html

An important aspect of RDF is the promotion of the **Open World Assumption which allows data to be aggregated without a fixed schema** of allowed relations as in relational databases. New relations (represented by URIs) can be added at any time without changing anything. RDF graphs can be combined freely, since the use of URIs guarantee that connections are only made between the same entities. This freedom on the use of RDF to represent anything can have its cost. Linked data producers may not describe properly the data they are publishing and linked data consumers usually have difficulties to know how to integrate that data in their applications. In practice, although the RDF toolset is growing, **RDF has not yet been established as a popular technology** for web developers and engineers. In the last decade, most of the data integration solutions opted for XML as the lingua franca for that purpose. A lot of technologies emerged around XML for validation, transformation and exchange. Nowadays, JSON is gaining in popularity in the web development community which considers it easier to manipulate and process by the existing tools and programming languages. The linked data principles do not depend on any particular data format as long as it is machine-processable. Furthermore, although the RDF format was originally based on XML, there are other RDF formats like Turtle, which is more human friendly, and JSON-LD[3], which is based on JSON.

The fourth principle of **linked data promotes the use of links from resources to other resources** so linked data consumers can discover new information. Those links are essential for the linked data project as they represent the glue between different datasets.

As can be seem, the linked data principles are quite intuitive and easy to understand, and there have been a lot of projects and initiatives that successfully embraced the linked data project[4]. As a running example of linked data initiatives of special interest for the Public Sector Information, we will consider the **eProcurement and public contracts domain** (Ordóñez et al, 2012; Álvarez et al, 2014).

Some pioneer initiatives to represent procurement notices as linked data were the MOLDEAS project (Álvarez et al, 2012) which was a prototype developed in collaboration with the EuroAlert service[5] and the LOTED (Linked Open Tenders Electronic Daily) project[6], which collects tenders in the European Union coming from the Tenders Electronic Daily portal. In the

[3] JSON-LD: http://json-ld.org/

[4] State of the LOD cloud (2014): http://linkeddatacatalog.dws.informatik.uni-mannheim.de/state/

[5] Euroalert service: http://euroalert.net/

[6] LOTED project: http://loted.eu/

context of the LOD2 European Project, a Public Contract Filing Application was proposed aimed both for contract authorities issuing calls for tenders and for bidders responding to those calls. The use of linked data can help in the matchmaking process of finding similar contracts (Necaský, 2014). A good example of a linked data portal is the Italian public contracts service provided by Nexa[7], which translates to Linked Open Data the XML data released by the Italian Public Sector bodies following the "anti-corruption" Act (law no. 190/2012). The project offers a linked dataset with an SPARQL endpoint and a browseable interface.

[7] Nexa public contracts project: http://nexa.polito.it/public-contracts

3 Quality of Linked Data Portals

A very simple and clear attempt to assess open data quality was the 5-star model proposed by Tim Berners-Lee in 2010 as a way to encourage governments to adopt the linked data principles[8]. **The 5-star model** classifies open data initiatives according to the type of open data that they publish:

- One star: Available on the web (whatever format) but with an open licence, to be Open Data
- Two stars: Available as machine-readable structured data (e.g. excel instead of image scan of a table)
- Three stars: as (2) plus non-proprietary format (e.g. CSV instead of excel)
- 4 stars: All the above plus use open standards from W3C (RDF and SPARQL) to identify things, so that people can point at your data.
- 5 stars. All the above, plus: Link one's data to other people's data to provide context.

The previous classification was very useful to motivate the adoption of linked data in different projects. However, once a project adopts the linked data model, it doesn't go into further details to assess the quality of its linked data.

Assessing the quality of linked data quality portals must take into account different aspects like maintainability, sustainability, usability, etc. that can be measured for data and web quality in general. As expressed in (Heath, 2011):

"Linked Data might be outdated, imprecise, or simply wrong. Therefore, Linked Data applications should consider all RDF statements that they discover on the Web as claims by a specific source rather than as facts. Applications should contain a module to filter RDF spam and prefer data from sources that are known for good quality to data from others. "

The LOD2 project proposed the Linked Data Stack[9] as a set of tools to manage the life cycle of Linked Data. The life cycle was divided in several stages like: authoring, interlinking, enrichment, etc. One of those stages is called Quality Analysis and several tools are proposed like RDFUnit for RDF validation and Sieve for quality assessment and fusion.

There is a growing interest to find metrics and methodologies to assess the quality of linked

[8] 5 stars model: http://5stardata.info

[9] Linked Data Stack: http://stack.linkeddata.org/

data portals which can be seen in the two international workshops organized about linked data quality[10]. Zaperi et al (2012) include a systematic survey on literature related with Linked data quality and identify a **set of data quality dimensions** that can be applied to assess the quality of linked data. The dimensions are classified in 4 groups and each dimension is accompanied by several metrics. The dimensions are:

- **Accessibility**: Availability, licensing, interlinking, security and performance.
- **Intrinsic**: syntactic validity, semantic accuracy, consistency, conciseness and completeness.
- **Contextual**: relevancy, trustworthiness, understandability and timeliness.
- **Representational** dimensions: representational conciseness, interoperability, interpretability and versatility.

There have appeared some recent initiatives to inspect and clean linked datasets. For example, Loupe[11] is a tool which can be used to inspect which vocabularies (classes and properties) are used including statistics and frequent triple patterns and LOD Laundromat[12] provides access to all Linked Open Data (LOD) in the world by crawling the LOD cloud and converting all its contents in a standards-compliant way, removing all data stains such as syntax errors, duplicates, and blank nodes.

[10] Workshops on Linked Data Quality: http://ldq.semanticmultimedia.org/ and

[11] Loupe: http://loupe.linkeddata.es/loupe/

[12] LOD Laundromat: http://lodlaundromat.org/

4 RDF Validation

RDF is a central part of any linked data project to provide information that can automatically be processed by machines. It is based on **simple statements of the form "subject – predicate – object"** where the predicates are uniquely identified by an IRI. An RDF dataset is comprised of a set of statements that can describe some information on a given domain. There have been several notations for RDF like Turtle, RDF/XML, N-Triples, etc. Using Turtle[13], we could describe, for example, some information about a public contract could be:

```
:c23 rdfs:label      "Maintenance service" ;
     time:year       2015;
     pc:agreedPrice 259870;
     pc:tender       :e45 ;
     pc:tender       :e47 .
:e45 rdf:type        gr:BusinessEntity;
     rdfs:label      "Company ABC" ;
:e47 rdf:type        gr:BusinessEntity;
     rdfs:label      "Company XYZ" .
```

Figure 1. Example of RDF data represented in Turtle

Figure 1 represents in RDF a public contract `:c23` with a property `rdfs:label` with value `"Maintenance service"` that has been awarded in `2015` by a price of 259870 and has two tenders: the entity `:e45` and the entity `:e47` which are both of type `gr:BusinessEntity`. The previous information can be represented using the graph in figure 2.

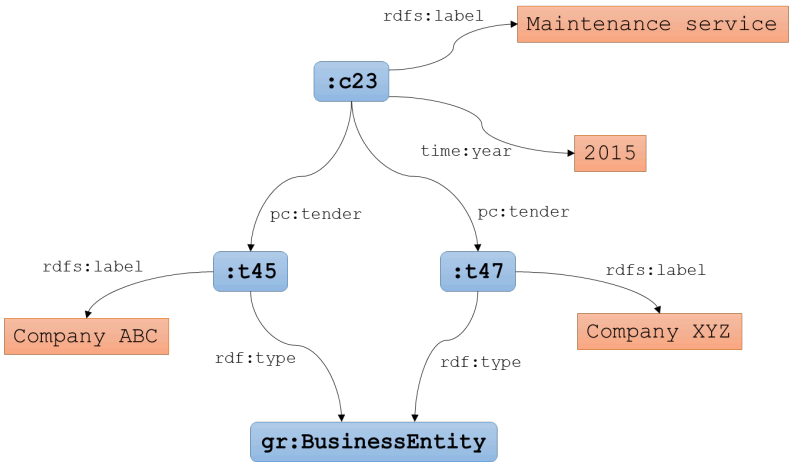

Figure 2: Example of RDF data that represents a public contract

[13] Turtle notation is intended for human readability. It enables the replacement of full IRIs by qualified names preceded by an alias and a colon. The aliases employed in this example have been taken from http://prefix.cc.

One of the main advantages of RDF is that **it is possible to automatically merge data from different RDF data graphs based on the universality of using URIs to represent entities**. At the same time, the flexibility of the graph model enables different systems to easily reuse data from heterogeneous sources. Although the benefits of RDF for data representation and integration are undisputable, its adoption by everyday programmers and system architects who care more about by creating and accessing well-structured data in databases than about inference has not yet taken off.

In 2013, an RDF validation workshop[14] was organized by W3C to gather the requirements of the different stakeholders. A conclusion of the workshop was that, although SPARQL could be used to validate RDF, there was a need for a more high level and concise language. **Shape Expressions** (Prud'hommeaux et al, 2014) emerged as such a language. As an example, figure 3 contains a description of the previous RDF data using Shape Expressions[15].

```
<PublicContract> {
  rdfs:label      xsd:string ,
  time:year       xsd:year,
  pc:agreedPrice xsd:integer,
  pc:tender       @<BusinessEntity> +
}
<BusinessEntity> {
 rdf:type    ( gr:BusinessEntity ),
 rdfs:label xsd:string
}
```

Figure 3. Simplified Public contracts schema represented in ShEx

The previous definition declares the shape of a `<PublicContract>` as having a property `rdfs:label` whose value must be of type `xsd:string`, two other properties `time:year` and `pc:agreedPrice` with values of type `xsd:year` and `xsd:integer`, a one or more properties `pc:tender` whose value must be a node with shape `<BusinessEntity>`. Finally, a `<BusinessEntity>` has type `gr:BusinessEntity` and `rdfs:label` of type `xsd:string`.

The **Shape Expressions language has been designed as an intuitive and human-friendly high level language for RDF validation**. There are several implementations available[16] and some

[14] RDF validation workshop: https://www.w3.org/2012/12/rdf-val/

[15] The example can be tested online using the RDFShape validator available at: http://goo.gl/3pahF0

[16] More information about Shape Expressions and implementations is available at: http://shex.io/

online validators[17]. Shape Expressions can even be represented using data model diagrams as in figure 4.

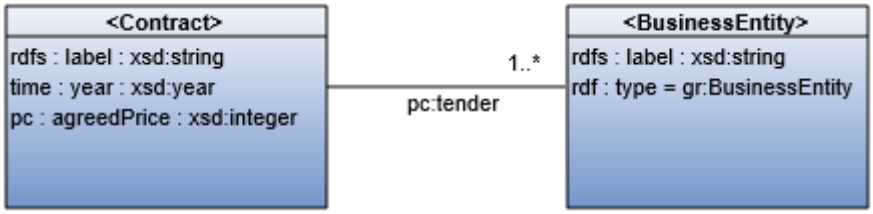

Figure 4. Data model diagram of a simplified Public Contract

In 2014, W3C chartered a working group called **RDF Data Shapes** to produce a language for defining structural constraints on RDF graphs. The language has been called **SHACL** and in October 2015, a first public working draft has been published[18]. Figure 5 contains a description of the simplified public contracts data model using SHACL —notice that as SHACL is based on RDF, the example uses Turtle notation. The Working Group is currently considering the use of a more human friendly syntax for SHACL inspired by Shape Expressions.

```
<PublicContract> a sh:Shape ;
  sh:property [
    sh:predicate rdfs:label ;
    sh:minCount 1 ; sh:maxCount 1 ;
    sh:dataType xsd:string
 ] ;
 sh:property [
    sh:predicate time:year ;
    sh:minCount 1 ; sh:maxCount 1 ;
    sh:dataType xsd:year
 ] ;
 sh:property [
    sh:predicate pc:agreedPrice ;
    sh:minCount 1 ; sh:maxCount 1 ;
    sh:dataType xsd:integer
 ] ;
 sh:property [
    sh:predicate pc:tender ;
    sh:minCount 1 ;
    sh:valueShape <BusinessEntity>
 ].

<BusinessEntity> a sh:Shape ;
 sh:property [
    sh:predicate rdf:type ;
    sh:minCount 1 ; sh:maxCount 1 ;
    sh:hasValue gr:BusinessEntity
 ] ;
 sh:property [
    sh:predicate rdfs:label ;
    sh:minCount 1 ; sh:maxCount 1 ;
```

[17] RDFShape: online RDF validator available at: http://rdfshape.herokuapp.com

[18] SHACL First Public Working Draft: http://www.w3.org/TR/shacl/

```
    sh:dataType xsd:string
  ].
```

Figure 5. Simplified Public contracts schema represented in SHACL

5 Conclusions and recommendations

Although the history of the linked data movement that emerged in 2007 has yet to be written and it is yet too early to assess its global impact, in these years, there are several lessons that can be learnt with a lot of linked data initiatives that have developed and have later been abandoned while other initiatives seem to have been established and maintain their datasets active for academic and industrial reuse.

The tools and techniques needed for linked data publishing are gradually maturing. However, there is yet a lack of tools to measure and guarantee the quality of linked data solutions.
In fact, the main piece of any linked data portal, RDF, still lacks a standard way to be described and validated. The current work developed by the **W3C Data Shapes Working Group** and the **Shape Expressions** community may help to improve RDF adoption in industrial scenarios where there is a real need to ensure the structure of RDF data, both to produce and to consume it. **These initiatives can be seen as a sign of the increased maturity of RDF and the linked data project.**

References

Álvarez R., J. M., Labra G., J. E., Ordóñez de Pablos, P. New trends on e-Procurement applying semantic technologies. Computers in Industry 65(5): 797-799 (2014)

Álvarez R., J. M., Labra G., J. E., Cifuentes S., F., Alor-Hernández, G., Sánchez-Ramírez, C., Guzmán L., J. A.: Towards a Pan-European E-Procurement Platform to Aggregate, Publish and Search Public Procurement Notices Powered by Linked Open Data: the Moldeas Approach. International Journal of Software Engineering and Knowledge Engineering 22(3): 365-384 (2012)

Dietrich, D. (2012) Linked Data, European Public Sector Information Topic Report No 2012 / 11.

Heath, T. and Bizer, C. (2011) Linked Data: Evolving the Web into a Global Data Space (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1-136. Morgan & Claypool.

Labra G., Prud'hommeaux E., J. E., Solbrig, H., Álvarez R., J. M. (2014) Validating and describing linked data portals using Shape Expressions, 1st Workshop on Linked Data Quality, Leipzig, Germany, Sept. 2014.

Miller, P. (2010), Linked Data and Government. ePSI Platform Topic Report No 2010 / 7.

Necaský, M., Klímek, J., Mynarz, J., Knap, T., Svátek, V., Stárka, J.: Linked data support for filing public contracts. Computers in Industry 65(5): 862-877 (2014)

Ordóñez de Pablos, P., Cueva L., J. M., Labra G., J. E., Tennyson, R. (2012), E-Procurement management for Successful Electronic Government Systems, IGI Global

Prud'hommeaux E., Labra G., J. E., Solbrig, H. (2014) Shape Expressions: An RDF validation and transformation language, 10th International Conference on Semantic Systems, Leipzig, Germany, Sept. 2014.

Prud'hommeaux E., Labra G., J. E., (2015) RDF ventures to boldly meet your most pedestrian needs, Bulletin of the American Society for Information Science and Technology, Volume 41, Issue 4, pages 18–22, April/May 2015

Schmachtenberg M, Bizer C., Paulheim H. (2014) Adoption of the Linked Data Best Practices in Different Topical Domains, The Semantic Web – ISWC 2014, Lecture Notes in Computer Science, vol. 8796, pp. 245-260.

Ulicny B. (2015). Constructing Knowledge Graphs with Trust, 4th International Workshop on Methods for Establishing Trust of (Open) Data, Bentlehem, USA.

Zaveri A., Rula A., Maurino A., Pietrobon R., Lehmann J., Auer S. (2014) Quality Assessment for Linked Data: A Survey, Semantic Web Journal

About the Author

Dr. Jose Emilio Labra Gayo is an Associate Professor from the University of Oviedo, Spain. He has been the Dean of the School of Computer Science Engineering at the University of Oviedo from 2004 until 2012. He founded and is the main researcher of the WESO (Web Semantics Oviedo) research group. The group collaborates on practical applications of semantic web and linked open data and has been involved in several projects with industrial partners and public administrations. His research interests are Semantic Web technologies, Declarative Programming Languages and Web Engineering. He is also chair of the W3c Best practices on Multilingual Linked Open Data Community Group and is member of the W3c RDF Data Shapes Working Group.

Copyright information