Assessing the Quality of Geospatial Linked Data – Experiences from Ordnance Survey Ireland (OSi)

Jeremy Debattista[1], Eamon Clinton[2] and Rob Brennan[1]

[1] Knowledge and Data Engineering Group, ADAPT Centre, School of Computer Science and Statistics, Trinity College Dublin, Dublin 2, Ireland
{debattij, rob.brennan}@scss.tcd.ie
[2] Ordnance Survey Ireland, Phoenix Park, Dublin 8, Ireland
eamonn.clinton@osi.ie

Abstract. Ordnance Survey Ireland (OSi) is Ireland's national mapping agency that is responsible for the digitisation of the island's infrastructure in terms of mapping. Generating data from various sensors (e.g. spatial sensors), OSi build its knowledge in the Prime2 framework, a subset of which is transformed into geo-Linked Data. In this paper we discuss how the quality of the generated sematic data fares against datasets in the LOD cloud. We set up Luzzu, a scalable Linked Data quality assessment framework, in the OSi pipeline to continuously assess produced data in order to tackle any quality problems prior to publishing.

Keywords: Data Quality, Geospatial Data, Linked Data

1 Introduction

Ordnance Survey Ireland (OSi) has been publishing authoritative open Linked Data since 2016 through its national geo-data portal data.geohive.ie [2]. The semantically-uplifted data being analysed in this paper is a subset of OSi's Prime2 national data infrastructure (see §2). A key feature for authorative data publication is the assurances provided by the use of extensive quality processes, human curation and rule-based quality checking within Prime2. However Linked Data has its own quality measures and challenges [1] and so it was important to investigate the relative quality of the Linked Data being produced. Also, it is hoped that the semantic enrichment of the data provided by uplifting to RDF could help to provide further semantic quality assurances and feedback for the underlying Prime2 data.

2 Generating and Publishing OSi Linked Geospatial Data

Currently OSi detects and tracks nearly 50 million spatial objects (e.g. buildings, roads, legal boundaries) through time by combining sensor data and state datasets. These are stored in OSI's Prime2 object model in an Oracle Spatial and Graph database using state of the art geo-modelling standards and metadata. The OSi Linked Data is a subset of Prime2 generated via R2RML mappings to the OSi geohive ontology. The

geohive ontology is an extension of the geoSPARQL ontology for describing geographical features and their geometries. The current published data focuses on describing Ireland's administrative entries and their boundaries. However in 2018 it is planned to have a significant expansion of the coverage of Prime2 spatial entities published as Linked Data. Geohive currently supports browsing via Pubby and its boundary entities are interlinked to equivalent DBpedia concepts where available. Typical entity properties are multi-lingual labelling in both English and Irish, WKT (well known text) representations of the 2D polygons of their geometry at multiple resolutions and administrative boundary classifications e.g. county, barony, or city council.

3 Assessing the Quality of Linked Geospatial Data

As part of the OSi Linked Data publishing pipeline, the quality of the generated Linked Data is assessed in order to maintain the high standards expected of a national spatial data infrastructure. This is of utmost importance as OSi's data is used for the infrastructural planning and development of Ireland. The objectives of this quality assessment in this pipeline is to help OSi, as Linked Data producers, to **(O1)** identify if there are any errors in the R2RML mappings that are generating incorrect RDF data; and **(O2)** check if Linked Data best practices are being followed. Based on these two objectives we first identified 19 suitable Linked Data quality metrics (see Table 1) as defined in [1]. In order to achieve these objectives deployed Luzzu [3] a Linked Data quality assessment framework, on the OSi release server to assess the quality of the last snapshot of OSi boundary data[1].

In this work, the chosen metrics and quality results reflect a data producer view of quality, rather than the data consumer view that is more common in the literature. Furthermore, we assume that geographic boundary data (polygons) provided by OSi is accurate[2] and thus no geo-specific quality measures are required to be assessed. Therefore, our quality assessment scope was narrowed to objective domain-independent Linked Data quality metrics. [1] formally define 27 metrics and assess them over the 2015 version of the LOD cloud, which includes geographic datasets. Nonetheless, not all of these metrics were relevant to the OSi datasets and our generation pipeline. With regard to our first objective (O1) we looked at the *intrinsic category* metrics. These metrics enable us to understand if there are any consistency, syntactic and conciseness issues in the generated RDF datasets. Metrics from the other three categories were considered for our second objective (O2).

The following results are based on the data dump that was available on December 10th 2017 and newer versions of this data is available on data.geohive.ie. In this section, we will present and discuss the quality results and compare them to the mean values observed in the LOD cloud as in [3].

[1] https://www.osi.ie/education/third-level-and-academic/boundaries/

[2] https://www.osi.ie/about/osi-positional-accuracy/

3.1 Quality Results and Discussion of OSi RDF Datasets.

Nineteen quality metrics were assessed by Luzzu. Seven of these metrics fell significantly below 100% or a maximum score and are worthy of further investigation. However, overall it must be observed that the generated OSi Boundary dataset displays high quality characteristics by these objective measures that are accepted within the Linked Data community as state of the art ways to measure Linked Data quality. A more detailed breakdown is shown in Table 1. The second column (Value) shows the exact value recorded for each metric and the overall quality picture is good with 13 of the 19 metrics equalling or exceeding the mean quality levels seen in LOD (as per [3]). Three metrics in which we vastly exceed the quality seen in common LOD practice are U1, CS9 and A3. U1 (Human Readable Labelling) shows that our dataset is well annotated for human consumption and our result is more than double that of the average. For CS9 (correct usage of domain and range datatypes), at c. 76% we exceed the average of 60% for LOD but we aim at 100% in this metric so the R2RML mappings were investigated. Similarly, in A3 (Dereferenceability) at 64% we are nearly double the average of 37% since OSi hosts definitions and resources as Linked Data in addition to the dump files.

Table 1 also allows us to more clearly identify the 6 metrics that are of most concern: RC1, IN4, V2, I1, L1, L2. There is a secondary set of metrics that must also be investigated based on their power as objective validators of the dataset (i.e. they can easily detect defects that should not be present, despite the OSi data exceeding the LOD average for these metrics): IN3, CN2, CS2, CS9, A3. These 5 represent metrics for which the OSi should be achieving 100% scores and hence are worth investigating, even when the score is very high like U1 (Labelling) at 99.94%.

Table 1 - OSi Boundary Dataset Quality Metrics 10th Dec 2017

(Abbr) Metric Name	Value	Mean LOD Value[3]
(RC1) Keeping URIs Short	65.87%	84.07%
(RC2) Usage of RDF Data Structures	100%	99.44%
(IN3) Usage of Undefined Classes and Properties	56.01%	54.48%
(IN4) Usage of Blank Nodes	29.41%	96.01%
(V2) Usage of Multiple Languages	1 lang	1.72 languages
(U1) Human Readable Labelling and Comments	99.94%	43.76%
(CN2) Extensional Conciseness (Approximate)	99.35%	92.04%
(CS1) Correct use of Entities as Members of Disjoint Classes	100%	100%
(CS2) Misplaced Classes or Properties	99.99%	99.99%
(CS3) Misused OWL Datatype or Object Properties	100%	98.88%
(CS4) No Use of Deprecated Classes or Properties	100%	99.97%
(CS5) Valid Usage of the Inverse Functional Properties	100%	96.98%
(CS6) No Ontology Hijacking	100%	93.64%
(CS9) Correct use of Domain or Range Datatypes	75.61%	60.11%

[3] As identified in [1]

(SV3) Compatible Datatypes	100%	96.80%
(A3) Dereferenceability of URIs (Approximate)	64.1%	36.86%
(I1) Links to External *Linked Data* Providers	1 LD Prov	27.01 LD Prov
(L1) Presence of a Machine-Readable Licence	0	14.4 licences
(L2) Presence of a Human-Readable Licence	0[4]	8.8 licences

3.2 Root cause analysis of under-performing metrics

RC1 - Approximately 33% of the OSi Boundary data URIs exceed the recommended length for short URIs (60 chars). Part of this is due to the reuse of long internal OSi GUIDs to identify spatial thing. Therefore, techniques should be explored in order to map the internal OSi identifiers with more meaningful URIs.
IN3 - Luzzu identified some 11 typos that have crept into the mappings e.g. ElectoralDistrict vs ElectoralDistricts.
IN4 - While blank nodes are generally discouraged in LOD, it is part of the design of the OSi Boundary dataset that feature geometries cannot directly be addressed and users need to go through the associated feature concept e.g. Co. Dublin to access a geometry.
V2 - The average number of languages used per resource is less than 2, i.e. less than 50% of all resources include a label in Irish as well as English. The way this metrics is calculated is not really representative since all geometry resources are totally unlabelled and count for this metric. Hence it suggests a potential revision of the metric.
A3 - Despite having a high score here we expect to achieve 100%. It is possible that some of the errors were due to load throttling at the data.geohive.ie server as opposed to the resources not being available.
I1 - Our score is low here because we have only interlinked to DBpedia.

4 Final Remarks

In this paper, we discuss and compare the quality of the generated OSi geo-linked datasets. Our assessment was done in light of two objectives: problems in mappings that is generating noisy data, and if best practices are followed. We assessed 19 different quality metrics and compared them with the LOD cloud's mean quality value as assessed in [1]. We show that for most metrics assessed our published data has a better quality value than the assessed mean value of the LOD cloud datasets.

References

1. Debattista, J., Lange, C., Auer, S., Cortis, D.: Evaluating the Quality of the LOD Cloud: An Empirical Investigation. Semantic Web Journal (2018, preprint)
2. Debruyne, C., Meehan, A., Clinton, E., McNerney, Nautiyal, Lavin, P., O'Sullivan, D., Ireland's Authoritative Geospatial Linked Data, In Proceedings 16th International Semantic Web Conference (ISWC 2017), vol. 10588, pp. 66 – 74, Springer, Vienna, Austria (2017)
3. Debattista, J., Auer, S., Lange, C.:. Luzzu —A Methodology and Framework for Linked Data Quality Assessment. *J. Data and Information Quality* 8, 1, Article 4 (2016).

[4] The OSi data has a human-readable license specified in the Geohive web-page