LinkedDataOps:Linked Data Operations based on Quality Process Cycle*

Beyza Yaman and Rob Brennan

ADAPT Centre, Dublin City University, Dublin, Ireland
{beyza.yaman,rob.brennan}@adaptcentre.ie

Abstract. Quality assessment is extremely relevant to measure the utility of data and it is especially critical for geospatial data due to its importance in daily life, e.g. navigation, and self-piloted vehicles. This paper describes a new end-to-end framework for quality-oriented continuous development and improvement of data based on standards compliance. The implemented methods build upon the open-source Luzzu framework with an open-source standards-agnostic dashboard to visualize and analyze quality metric observations in a data production pipeline.

1 Introduction

Linked Data has a life-cycle and data quality issues in Linked Open Data (LOD) are the result of a combination of data and process-related factors in this life-cycle. This dynamic process requires continuous improvement, in contrast to the static releases that typify most LOD datasets. In particular, geospatial data suffers from high demands on quality and if not met, these can cause major problems in real life, such as the navigation problems leading to the Irish coast guard helicopter crash in 2017[1].

While there is an extensive number of studies on the quality of traditional data, the topic of Geospatial Linked Data (GLD) has received little attention. GLD problems include *i)* a lack of quality metrics for GLD, e.g., geo-political boundaries datasets are required to define the extent of town or county boundaries (spatial things) as polygons. However, there are no well-known LOD quality metrics to check the conformance of the polygon shapes (e.g. if they're closed or not). *ii)* a lack of the end to end (e2e) data quality cycle considering data transformation, objective assessment metrics or root causes of problems [3–5]. Whereas, many quality aspects can be achieved by assuring standards conformance, e.g., Findable in FAIR principles implies the availability of appropriate catalog metadata like W3C DCAT.

Thus, the research question we are tackling is "To what extent can we implement effective methods and tools for quality-oriented continuous development and improvement of Linked Data deployments taking into account the e2e data

[1] Irish coast guard helicopter crash, t.ly/COas

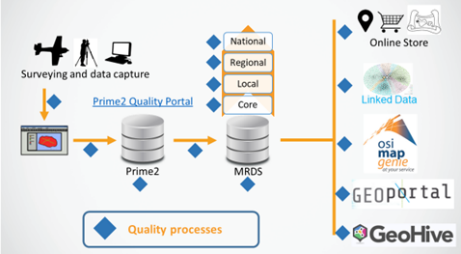


Fig. 1. OSi Geospatial Data Publishing Pipeline with Quality Control Points

lifecycle". The objective of the Elite-S Marie Sklodowska-Curie Cofund Action LinkedDataOps project[2] is to meet dynamic quality needs by providing new continuous data quality tools and methodologies.

This project will be realized in the Ordnance Survey Ireland (OSi) data publication use-case. OSi's national geospatial digital infrastructure (Fig. 1) encompasses surveying and data capture, image processing, translation to the Prime2 object-oriented spatial model of over 50 million spatial objects tracked in time and provenance, conversion to the multi-resolution data source (MRDS) database for printing as cartographic products or data sales and distribution at data.geohive.ie [2]. Thus, managing data quality throughout the data pipeline and lifecycle is key to OSi. Moreover, the United Nations Global Geospatial Information Management (UN-GGIM) framework highlights the importance of standards conformance of data for quality. Thus there is a need for monitoring and reporting on the standards conformance of OSi GLD. For example, to provide continuous upward reporting to the Irish government, European Commission and UN. Inspired by the *DevOps* methodology, *LinkedDataOps* will enable sophisticated data quality monitoring within the organization.

LinkedDataOps implements an e2e quality assessment framework based on the Luzzu framework [1]. The project defines roles and responsibilities to ensure liability for data quality with policies and procedures. It supports the process by means of the proposed standards while maintaining the performance for good decision-making. Continuous validation of quality and standards conformance will be performed by data experts and engineers in the OSi data production pipeline using an e2e standards reporting tool developed in this project (Fig. 3).

Contributions of this project are as follows: *i)* Defining new standards compliant quality and FAIR metrics according to existing data governance standards *ii)* Implementing a novel tool based on the existing state of the art approaches on data quality which will be integrated to OSi Linked Data cycle to improve the quality of data *iii)* Publishing data and quality metadata reusing standard vocabularies *iv)* Implementing an open-source dashboard for e2e data quality management based on a unified quality graph v) Deploying a system based on the case study in OSi.

[2] linkeddataops.adaptcentre.ie

2 LinkedDataOps Approach

The overall scope of this work is to improve the quality and service outcomes of an organization while conforming to the standards and support good decision-making. The following approach is employed in order to achieve this goal: *i)* Quality assessment is performed for the transformation phase from relational data to Linked Data automatically. Quality constraints are defined for R2RML mappings for high quality transformation of data. The tool is integrated with the Luzzu framework (Fig.2, Step 1). *ii)* Implementation of the geospatial data quality metrics and FAIR assessment metrics is performed. Aligned with the OSi's standard compliance objectives, relevant metrics are defined for the geospatial data at hand and then they are integrated with Luzzu framework to measure the quality, standards conformance and FAIRness of the dataset. Existing quality metadata definition of the Luzzu are extended by those metrics in both dataset and triple levels via standard vocabularies (Fig.2, Step 2).

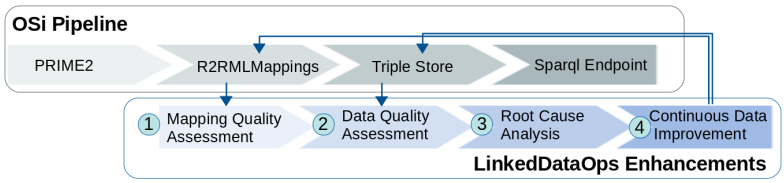


Fig. 2. LinkedDataOps Workflow (arrows indicate data input)

iii) Quality problems are detected while consuming the data, and errors should be fixed to publish the OSi dataset with high characteristics. For this reason, feedback on the data is gathered from the experts using an automatic standards compliance reporting portal. The given feedback is used to improve the data. Moreover, logs are analyzed for incompatibilities between software versions and data versions to verify the efficiency of the tool. Validation of the tool is realized by the integrity checking of the input and output of the tool (Fig.2, Step 3). *iv)* Continuous monitoring of the data for inconsistencies is performed, thus, automation of the steps is realized for data profiling. Different quality assessment results are saved as a W3C data cube with different versions of the assessment and quality metadata along with their assessment date and time (Fig.2, Step 4). The detected errors are fixed via using an extension for the Luzzu tool.

3 Results and Conclusions

The purpose of a dashboard is to provide meaningful insights to the user by depicting significant correlations for the given data. A proof of concept Linked-DataOps user dashboard was implemented (Fig. 3) to visualize and analyze quality metric observations. The interactive dashboard allows users to visualize the *i)* data quality along the e2e pipeline *ii)* data quality metrics and their scores

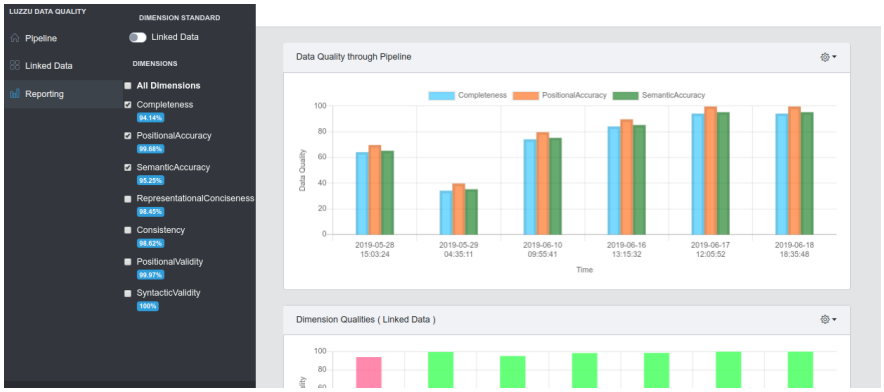


Fig. 3. OSi End-to-End Dashboard Reporting

for specific datasets (associated with stages or systems within the pipeline) *iii)* historical quality changes over time (illustrated in Fig. 3) iv) pass/fail quality status of a specific dataset for a given quality threshold, for example *w.r.t* the different standardization approaches.

Conclusions This paper presents the *LinkedDataOps* end-to-end geospatial data standards compliance framework for quality-oriented continuous data development and improvement. Although it is still a proof of concept, however, there is already an open source e2e dashboard available. The research is being carried out with OSi where the framework is planned to be deployed.

Acknowledgement This research received funding from European Union's Horizon 2020 research and innovation programme under Marie Sklodowska-Curie grant agreement No. 801522, by Science Foundation Ireland and co-funded by the European Regional Development Fund through the ADAPT Centre for Digital Content Technology [grant number 13/RC/2106] and Ordnance Survey Ireland.

References

1. J. Debattista, S. Auer, and C. Lange. Luzzu—a methodology and framework for linked data quality assessment. *Journal of Data and Information Quality (JDIQ)*, 8(1):1–32, 2016.
2. C. Debruyne, A. Meehan, É. Clinton, L. McNerney, A. Nautiyal, P. Lavin, and D. O'Sullivan. Ireland's authoritative geospatial linked data. In *International Semantic Web Conference*, pages 66–74, 2017.
3. R. Karam and M. Melchiori. Improving geo-spatial linked data with the wisdom of the crowds. In *Proceedings of the joint EDBT/ICDT 2013 workshops*, pages 68–74. ACM, 2013.
4. J. Lehmann, S. Athanasiou, A. Both, A. García-Rojas, G. Giannopoulos, D. Hladky, J. J. Le Grange, A.-C. N. Ngomo, M. A. Sherif, C. Stadler, et al. Managing geospatial linked data in the geoknow project., 2015.
5. M.-A. Mostafavi, G. Edwards, and R. Jeansoulin. An ontology-based method for quality assessment of spatial data bases. 2004.

LinkedDataOps:Linked Data Operations based on Quality Process Cycle

Beyza Yaman and Rob Brennan
ADAPT Centre, Dublin City University, Ireland
beyza.yaman,rob.brennan@adaptcentre.ie

Problem Statement

- Data available from poorly controlled sources
- Dynamic sources causing inconsistencies
- Geospatial data suffers from high demands on quality
- e.g. transportation, navigation, GIS guidance, and self-piloted vehicles

Proposal

- An end-to-end (e2e) quality-oriented continuous development framework
- Improvement of data based on standards compliance
- Implemented methods build upon the Luzzu framework
- An open-source dashboard to

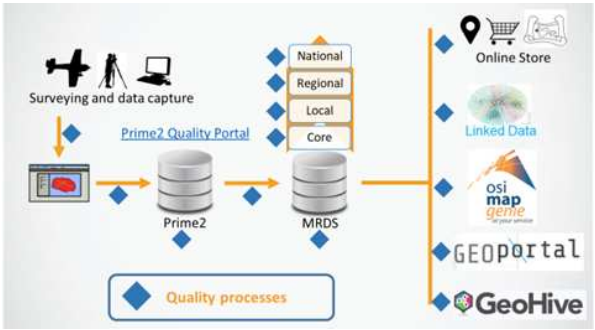
→ visualize quality metric observations

→ analyze quality scores in a data production pipeline

Contributions

- Implementing a novel tool to be integrated to OSi Linked Data cycle
- Defining new standards compliant quality and FAIR metrics
- Publishing data and quality metadata reusing standard vocabularies
- Implementing a data governance dashboard for e2e data quality management
- Deploying a system based on the case study in OSi

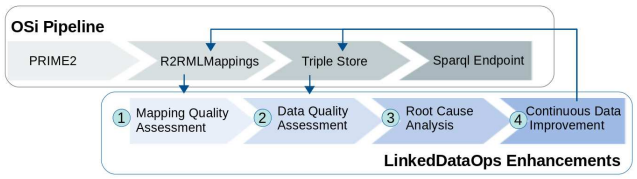
OSi Data Publishing Pipeline



LinkedDataOps Approach

- LinkedDataOps [2, 3] : Inspired by the DevOps methodology
- Quality assessment for the transformation phase from relational data to Linked Data
- Implementation of the geospatial data quality metrics and FAIR assessment metrics
- Detecting the root causes of problems by analysing the errors occurring in the pipeline
- Monitoring the data for inconsistencies continuously

LinkedDataOps Workflow



Data Governance Dashboard

The interactive dashboard allows users to visualize the

- data quality along the e2e pipeline
- data quality metrics and their scores for specific datasets
- historical quality changes over time
- pass/fail quality status of a dataset for a given quality threshold

OSi End to End Dashboard Reporting

References

[1] Debattista, Jeremy *et al.* , Luzzu—a methodology and framework for linked data quality assessment, *Journal of Data and Information Quality 8.1 (2016): 1-32*

[2] This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under the ELITE-S COFUND Marie Skłodowska-Curie grant agreement No. 801522.

[3] Project Website: linkeddataops.adaptcentre.ie