Semantic Web 1 (0) 1–5
IOS Press

LinkedDataOps:Quality Oriented End-to-end Geospatial Linked Data Production Pipelines

Beyza Yaman [a,*], Kevin Thompson [b], Fergus Fahey [b] and Rob Brennan [c]
[a] *ADAPT Centre, Trinity College Dublin, Dublin, Ireland*
E-mail: beyza.yaman@adaptcentre.ie
[b] *Ordnance Survey Ireland, Dublin, Ireland*
E-mails: kevin.thompson@osi.ie, fergus.fahey@osi.ie
[c] *ADAPT Centre, School of Computing, Dublin City University, Dublin, Ireland*
E-mail: rob.brennan@adaptcentre.ie

Editors: First Editor, University or Company name, Country; Second Editor, University or Company name, Country
Solicited reviews: First Solicited Reviewer, University or Company name, Country; Second Solicited Reviewer, University or Company name, Country
Open reviews: First Open Reviewer, University or Company name, Country; Second Open Reviewer, University or Company name, Country

Abstract. This work describes the application of semantic web standards to data quality governance in the architectural, engineering, and construction (AEC) domain for Ordnance Survey Ireland (OSi). It illustrates an approach based on establishing a unified knowledge graph for data quality measurements across a complex, quality-centric data production pipeline. It provides a series of new mappings between semantic models of the heterogeneous data quality standards applied by different tools and business units. 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 through enabling an end-to-end data governance approach. Current industrial practice tends towards stove-piped, vendor-specific and domain-dependent tools to process data quality observations however there is a lack of open techniques and methodologies for combining quality measurements derived from different data quality standards to provide end-to-end data quality reporting, root cause analysis or visualization. This work demonstrated that it is effective to use a knowledge graph and semantic web standards to unify distributed data quality monitoring in an organization and present the results in an end-to-end data dashboard in a data quality standards agnostic fashion for the Ordnance Survey Ireland data publishing pipeline. This paper provides the first comprehensive mapping of standardized generic information systems data quality dimensions and geospatial data quality dimensions into a unified semantic model of data quality.

Keywords: Geospatial Linked Data, Data Quality, Data Governance

1. Introduction

Architectural, engineering, and construction (AEC) industries has been rising recently with a high number of impact areas such as Building Information Modelling (BIM), smart construction, smart cities and digital twin applications. Digital technologies has played a significant role the way the products are designed, modelled and maintained due to its benefits such as ease of usage, powerful design, sustainability and data sharing within different domains.

With the advancements of the technology and requirements from the industry, AEC systems are evolving to a more automated and interchangeable management of data, such as, Industry 4.0 communications among heterogeneous industrial assets [32] sustainable buildings for environment-friendly construction structures [15], sensors embedded smart city applications [16]. There is a common feature of all these systems that these applications need unification of high quality

*Corresponding author. E-mail: beyza.yaman@adaptcentre.ie.

geospatial data, computer methods and domain knowledge to provide high quality results [16].

Given these circumstances, structured and interlinked characteristics of Semantic Web technology lay the foundations for seamless integration of different knowledge domains into AEC domain such as geospatial information systems (GIS), built systems, energy performance related systems which provides highly connected structure [25]. In addition, current standardization efforts have promoted interoperability among Linked Open Data (LOD), and community initiatives have focused decentralized data lakes among datasets. This allowed location-based AEC applications to gain more prominence in the domain by incorporating geospatial semantics into the data.

Geospatial information systems has long been considered a high value resource for different domains -as well as AEC domain- due to its rich semantics. However, Geospatial Linked Data (GLD) has been even more crucial with the rise of the knowledge graphs. The process of producing and transforming geospatial linked data is prone to errors and high demand is required on quality. Thus, data governance is required for the effective and efficient use of data, as well as the management and tracking of data, in order to provide high quality. However, due to multi-standards (regional or international) proposed for various data formats, a lack of metadata, or diverse and siloed data storage in the organizations, achieving efficient data governance is not always trivial. Those cases lead to a situation where data governance is hard to administer.

Effective data management is needed for high quality geospatial data but when the data is ingested and transformed from several sources, the manual process becomes harder to maintain. Data has a life-cycle which is subject to analysis. In case these processes are performed manually, provenance of this data would be even more critical to track. Thus, tracking data requires an automatic approach to manage the data in a pipeline.

Taking into account the above challenges, this paper answers the question “To what extent can semantic web-based methods and tools provide effective data quality governance for end-to-end production of AEC-ready geospatial linked data?".

In order to solve this problem, we propose a data governance approach to ensure the consistent operations of data production pipelines and monitoring quality. LinkedDataOps approach [35] is employed to achieve this goal as merging data from multiple perspectives of data by uplifting or transforming data from various formats by leveraging the flexibility of semantic web tools. A knowledge graph that integrates geospatial data aspects with a standards-based data quality and data lineage models into data catalogs was consolidated to generate a uniform data quality knowledge graph.

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 supporting good decision-making through end-to-end data governance. The following approach is employed in order to achieve this goal: *i)* Quality assessment results are gathered from relational geospatial data by uplifting to Linked Data automatically. The tool is integrated with the Luzzu framework (Fig.1, Step 1). *ii)* Geospatial data quality metrics are implemented. 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 and standards conformance. Existing quality metadata definition of the Luzzu are extended by those metrics in both dataset and triple levels via standard vocabularies (Fig.1, Step 2). *iii)* The integration of the data was performed using a unified data quality knowledge graph for interoperability of the data (Fig.1, Step 3). *iv)* 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.1, Step 4).

The contributions of this paper are *i)* creating a data lineage model to track the workflow of the data based on unified data quality knowledge graph[1] *ii)* identification of the geospatial data quality standards *iii)* a set of semantic data quality models using the W3C DQV standard for ISO 25012 and ISO 8000[2] *iv)* a set of comprehensive inter-dimensional mappings between data quality dimensions defined in all relevant standards[3], *v)* a set of data quality metrics to measure geospatial data quality[4] *vi)* semantic uplift mappings using R2RML for the 1Spatial 1Integrate tools[5] *vii)*

[1] https://opengogs.adaptcentre.ie/OrdnanceSurveyIreland-OSi/DataCatalog.git

[2] https://opengogs.adaptcentre.ie/OrdnanceSurveyIreland-OSi/StandardsMappings.git

[3] https://opengogs.adaptcentre.ie/OrdnanceSurveyIreland-OSi/StandardsMappings.git

[4] https://opengogs.adaptcentre.ie/OrdnanceSurveyIreland-OSi/StandardGeospatialQualityMetrics.git

[5] https://opengogs.adaptcentre.ie/OrdnanceSurveyIreland-OSi/R2RMLmappings.git

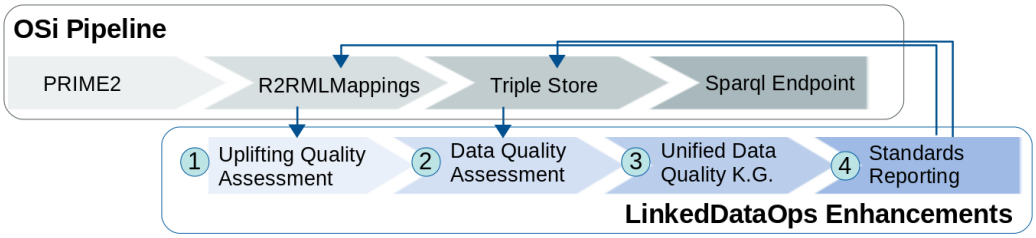

Fig. 1. LinkedDataOps Workflow (arrows indicate data input)

an open source dashboard for e2e data quality management based on a unified quality graph[6] and *viii)* a case study describing our deployment of this system in Ordnance Survey Ireland and lessons learned from this process.

The remainder of this paper is structured as follows: Section 2 describes the OSi use case, section 3 summarizes employed quality standards and tools as well as the R2RML mapping language. Section 4 discusses the unified data quality knowledge graph approach including architecture, data quality metrics, the mappings between standards, semantic uplifting and data lineage. We present the evaluation based on case study and results in Section 5 followed by lessons learned Section 6. Finally, conclusions and future work are discussed in Section 7.

2. Use Case : OSi Data Production Pipeline

National mapping agencies such as Ordnance Survey Ireland (OSi) are now geospatial data publishers more than cartographic institutions. The data in the OSi use case comes from numerous sources and is heterogeneous in terms of types, transformations, and versions. It was also discovered that data requires multi-dimensional, diverse quality measures in order to meet the needs of stakeholders and internal departments, making the process of tracking data flow and data management in a dynamic environment more difficult. Each stakeholder requires a set of different quality metrics for their specific data. On the other side, the lack of a real, high-quality pipeline and various, non-standardized data sources prevents the organization from effectively seeing the entire pipeline. Concrete data analysis and, as a result, data-driven data decisions are required for business plans to succeed. These requirements need an automatic end-to-end data pipeline solution which will allow the reproducibility of the processes with standardized approaches and methodologies. Thus, in the scope of this work, a data pipeline was established throughout the data lifecycle, allowing for a series of data processing steps and the flow of data operations from the data source to the last task.

Ordnance Survey Ireland (OSi) is the national mapping agency of Ireland and it manages the national geospatial digital infrastructure. OSi is producing maps for planning, construction and engineering purposes which provides a detailed database of roads, rivers, buildings and various features which might be found in a map. These maps are used for different occasions including emergency situations. In a fire alarm situation, fire services i) require to know the exact location of the incident, ii) need to arrive to the location in the shortest possible time. Quality of the data in the database is highly important to save lives of people and to stop the fire as soon as possible. On the other hand, necessary assets such as hydrants should be recorded on the database following an intelligent engineering and construction process along with a good planning on a common geographic dataset. Government departments and public-sector bodies under the National Mapping Agreement (NMA) (an Irish agreement) have unrestricted access to the most of OSi's geospatial data. With the NMA, one can request access to other datasets such as buildings and infrastructure [22]. On the other hand, this scenario is relevant in every country [34].

The OSi dataset 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 database for printing as cartographic products at a wide range of scales or onto other data sales and distribution channels such as Irish Geospatial Linked Data available

[6] https://opengogs.adaptcentre.ie/OrdnanceSurveyIreland-OSi/OSiDashboard.git

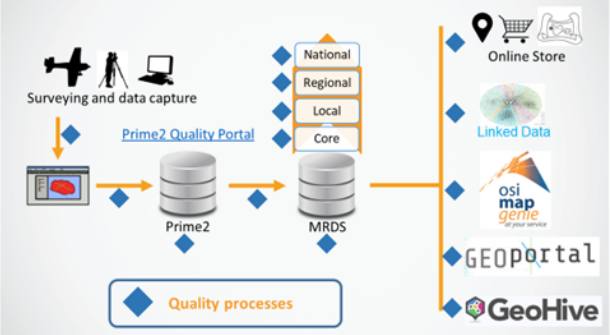

Fig. 2. OSi Geospatial Information Publishing Pipeline with Quality Control Points

through data.geohive.ie [13, 22]. All of these services run on a state of the art Oracle Spatial and Graph installation that supports both relational and RDF models using dedicated exadata hardware.

Data is collected, maintained and consumed by various levels of an organisation which results as data being distributed over disparate departments. Distributed data introduces a challenge of discovering the occurring data quality problems. Moreover, the data is often stored in different formats in different departments which makes the challenge even more difficult. Therefore, it is utmost importance to have an end-to-end data quality portal.

Managing data quality throughout the data pipeline and lifecycle is at the core of OSi operating practices (Fig. 3) and there are already quality checks at various stages that focus on the data quality dimensions. Current data quality assessment within OSi depends on i) two automated tools: the rules-based 1Spatial 1Integrate and Luzzu for linked data and ii) manual or semi-automated techniques by domain experts.

In addition, standardization approaches are more significant more than ever to create an interoperable and standard tool for the organizations. Frameworks such as the UN-GGIM (United Nations Secretariat: Global Geospatial Information Management) publishes a set of advises managing data quality and developing Integrated Geospatial Information systems at the national and international level [33]. It is required to conform with such standards for monitoring and reporting of the data at different levels. This will provide assurances for OSi's customers, help inform appropriate uses for their data; enable upward reporting to the Irish government, European Commission and UN; enable more sophisticated data quality monitoring within the organisation and provide feedback to managers within OSi for teams involved in data collection, modelling and transformation. Over 600 staff will be impacted by the new system and 10% of those staff will interact directly with the system.

Through a series of internal workshops with stakeholders the following requirements were identified:

- **Req 1:** Monitoring, analyzing and reporting the data quality using an unified end-to-end data quality knowledge graph.
- **Req 2:** Ability to report quality in arbitrary format for different stakeholders.
- **Req 3:** Aligning diverse standards to provide a uniform view for data quality results.
- **Req 4:** Ability to combine, query, visualise and report on quality results of diverse tools at many stages of the data production pipeline.
- **Req 5:** Tracking data to provide the back tracing for spotting the location of the errors occurring in the data.
- **Req 6:** Classification of the data to provide contextualization for statistical purposes.

3. Background

This study especially aims at providing a unified solution for the enterprise quality pipelines which is easily solved by a semantic approach using e2e knowledge graph. To the best of our knowledge this is not performed prior to this study.

3.1. International Data Quality Standards for Geospatial Data

Quality models are important for providing consistent terminology and guidance for quality assessment and are the basis for the evaluation of any product or service [27]. All the above mentioned standards aim at filling the gap for a specific area e.g. software quality, geospatial data quality. Thus, a standard might not be able to meet the all the requirements needed by a data pipeline.

This section identifies, evaluates and compares a set of relevant standards and recommendations for GLD quality proposed by the OGC, ISO and W3C. The ISO/TC 211 Geographic information/Geomatics committee defines geographic technology standards in the ISO 19000 series [1] as well as the OGC creates open geospatial standards. The both organizations have close connections such that some documents prepared by OGC are adopted by ISO or implemented

by the collaboration of both parties. The standards are evaluated in 3 main groups:

Geospatial datasets: ISO 19103, 19107, 19108, 19109, 19112, 19123, 19156[1] are published to describe the data, in particular the schema, spatial referencing by geospatial data, and methods for representing geographical data and measurements. OGC equivalence of the documents can be seen on the right hand side of the table. Old ISO 19113/19114/19138 are combined to 19157 data quality standards. Thus, while ISO 8000 defines data quality concepts and processes for generic information systems, ISO 19157 and ISO 19158 provide more detailed guidance on data quality practices for geospatial data. ISO 19158 specifies metrics and measurements for evaluation of data quality elements at different stages of the geospatial data lifecycle. It also defines quality metric evaluation by using aggregation methods and thresholds. ISO 19157 defines a set of data quality measures when evaluating and reporting data quality of geospatial data.

Geospatial metadata: ISO 19111 and 19115 describe the metadata standards for geospatial data. While ISO 19115 focuses on metadata for cataloging and profiling purposes with the extensions for imagery and gridded data; ISO 19111 describes appropriate metadata for a Coordinate Reference System.

Geospatial Linked Data: There are three relevant types of documents for data quality. *i)* ISO 19150 which guides high level ontology schema appropriate for geospatial data and rules for using OWL-DL. *ii)* OGC's GeoSPARQL standard that define a set of SPARQL extension functions for geospatial data, a set of RIF rules and a core RDF/OWL vocabulary for geographic information based on the General Feature Model, Simple Features, Feature Geometry and SQL MM [26]. *iii)* W3C has two documents, first the Data on the Web Best Practices recommendation for improving the consistency of data management and secondly the Spatial Data On the Web working group note which complements the earlier recommendation but is specialized for geospatial data.

There are many standard ways to represent quality metadata proposed for managing quality data. This papers focuses on the 3 main quality standards as well as W3C Best Practices to present quality reports:

ISO 8000 [7] defines characteristics of information and data quality applicable to all types of data. The document also provides methods to manage, measure and improve the quality of information and data which can be used in conjunction with quality management systems. The standard has 3 main categories namely semantic, syntactic and pragmatic quality including 16 dimensions.

ISO 19157 [8] is published to understand the concepts of data quality related to geographic data including data quality conformance levels in data product specifications, schemas, evaluating and reporting data quality with geospatial focus. The standard describes 6 dimensions to define the quality of geospatial data.

ISO 25012 [9] is one of the SQuaRE (Software product Quality Requirements and Evaluation) series of International Standards, which defines a general data quality model for data retained in a structured format within a computer system. In this study we consider this standard as our main standard due to its high coverage of the wide range of dimensions. The standard includes 17 dimensions to describe generic data quality.

W3C Best Practices DQV [2] to publish and usage of high quality data on the web. The practice has 14 recommendations and one include to provide data quality information with published datasets. Amrapali *etal.* [40] proposes 18 quality dimensions spread into 4 categories for the Linked Data environment thus in the scope of this work we use these categories and dimensions to sketch middleware standard mappings.

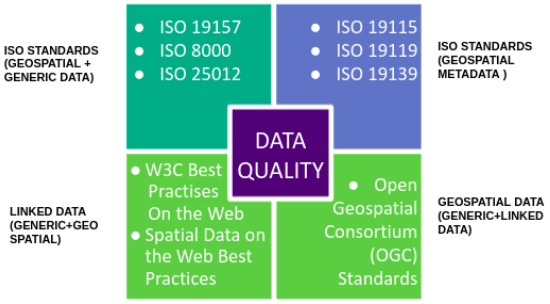

Fig. 3. OSi Geospatial Information Publishing Pipeline with Quality Control Points

3.2. Data Quality Tools for Geospatial Data

Several quality assessments of GLD have previously been conducted [19, 21, 24] but one of them relies on

[7] https://www.iso.org/standard/50798.html

[8] https://www.iso.org/standard/32575.html

[9] https://www.iso.org/standard/35736.html

crowdsourced evaluations rather than automated metrics [19], another one provides a generic Linked Data quality assessments of the data that is not specific to geospatial concerns [21] and the other is tied to a custom ontology predating GLD standardisation [24]. In contrast, there are not a large amount of dedicated geospatial data quality tools implemented per se, especially for Linked Data. Existing tools are focused on the traditional data and business products such as ArcGis[10], GeoToolkit[11]. The tools which are employed in OSi data pipeline are 1Spatial 1Integrate and Luzzu tools.

1Spatial 1Integrate[12] is a tool which automates the correction of invalid data by applying rules-based data re-engineering tasks. Compliance of the data is achieved by creating and managing multiple rule sets for the datasets. Using rules-based automation, the tool aims at ensuring the accuracy, inviolability and validity of the data and it is in the publishable state. The 1Integrate system performs over 200 rules on the relational data to ensure the compliance of the data with model prerequisites and to maintain the consistency of the data. The system produces statistical summaries, map view of the results or GIS files for the analysis on the data.

Luzzu [8] is an open-source Java based Linked Data quality assessment framework which allows user to use custom quality metrics to produce quality based statistics about the data. This is an interoperable tool allowing ontology driven backend to produce machine readable quality reports and metadata about the assessment results. After the processor streams all the triples quality metadata is produced by provenance information and problematic triples are described in the problem report. The quality metadata is represented by domain independent daQ (Dataset Quality) core ontology based on W3C RDF Data Cube and PROV-O vocabularies [11]. The data can be processed either from bulk data or SPARQL endpoints.

3.3. R2RML

R2RML[13][7] is a language to define mapping rules from relational data to RDF data so that they can be processed by a compliant mapping engine. It is a W3C recommendation. The mappings and any metadata are expressed in RDF. An R2RML mapping is written for a particular database schema and target vocabulary e.g. DQV, the W3C standard data quality vocabulary. A set of mapping rules and a relational database or tabular data in CSV (comma-separated value) format is used as an input to produce RDF data with the corresponding schema. R2RML mappings refer to logical tables to convert data from the given database, hence database views or actual tables can be mapped to RDF. The result of the R2RML process is a graph representation to the input database. Once a set of mapping rules is written, data can be rapidly and reliably transformed between relational and RDF formats. For example, the Oracle Spatial and Graph database product can naively load a set of R2RML rules into the database to dynamically create an RDF view of the underlying data.

3.4. Data Lineage for Geospatial Data

Data lineage can be used for data validation and verification as well as data auditing. These features are proven to be practical to store and track an enterprise metadata repository for data governance and data quality monitoring[14]. This subsection investigates the data linage approaches for geospatial data.

Chen et al. [3] define a domain-specific provenance model and a tracking approach to represent and track provenance information for remote sensing observations in a Sensor Web enabled environment. Closa et al. [4] analyses the potential for representing geospatial provenance in a distributed environment at the three levels of granularity (dataset, feature and attribute levels) using ISO 19115 and W3C PROV models. Another work of Closa et al. [5] present a provenance engine (PE) that captures and represents provenance information using a combination of the Web Processing Service (WPS) standard and the ISO 19115 geospatial lineage model. Di et al. [14] capture the provenance information in a standard lineage model defined in ISO 19115:2003 and ISO 19115-2:2009 standards (geographic metadata). Also, the authors extend both workflow language and service interface between provenance and geo-processing workflow by making it possible for the automatic capture of prove-

[10] https://www.esri.com/en-us/arcgis/products/arcgis-data-reviewer/overview

[11] https://www.sinergise.com/en/solutions/gis-tools/geo-toolkit-data-quality-tools

[12] https://1spatial.com/products/1integrate/

[13] https://www.w3.org/TR/r2rml/

[14] Hoang,Natalie"DataLineageHelpsDrivesBusinessValue". Trifacta.Retrieved2020-11-16.

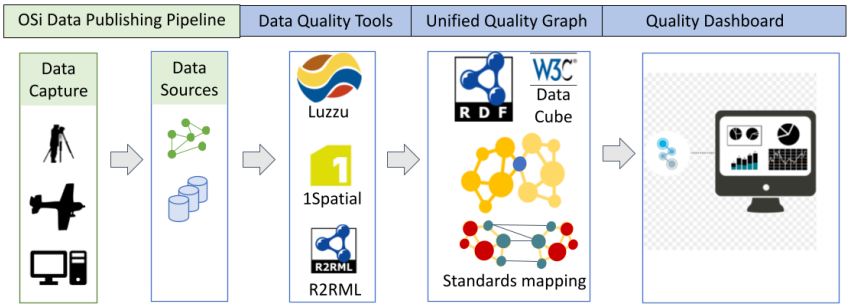

Fig. 4. Technical Architecture for a Unified Quality Graph Supporting End to End Data Quality Views

nance information in the geospatial web service environment.

Sadiq et al. [29] present ontologies for land administration workflows in the spatial information life cycle to determine record and allow access to provenance information. Sun et al. [30] present an ontological framework for geospatial data integration and sharing called GeoDataOnt which is divided into three compound modules: essential ontology, morphology ontology, and provenance ontology. Yuan et al. [39] propose to publish geospatial data provenance into the Web of Data extending the Provenir ontology.

To the best of our knowledge there are not any proposals to catalogue the quality of data in an e2e pipeline providing comparative results w.r.t. the different standards.

4. A Unified Data Quality Knowledge Graph

This section describes the knowledge graph to present and visualize the data quality in different steps of the end-to-end pipeline.

4.1. End-to-End Data Quality Views for Data Governance

A unified quality graph is a knowledge graph which is implemented to enable data governance and to manage all the data controls from one place. It has been designed to handle data quality outcomes from multiple sources with heterogeneous data formats, transformations, and versions.

Figure 4 illustrates the layered technical architecture needed to support end to end data quality for an e2e data pipeline. Starting from the OSi data publishing pipeline in Fig. 4, layers are created incrementally to build an end-to-end management of the existing pipeline. The first step was to enable dataset monitoring through the lifecycle of the data. Since data format changes throughout the pipeline, it is not a trivial process to detect the step where a problem occurs in the data. Thus, uplifting existing data quality results to the RDF format enables tracking the evaluation of the dataset in different formats along the pipeline. For tools that do not naively publish as RDF quality assessment data, this requires the creation of an uplift or data transformation workflow using R2RML.

In the second layer, quality assessment was performed for each dataset presented in various formats (e.g., rdf,rdb) to obtain an integrated quality assessment result for different sources. While the assessment of Linked Data was performed using Luzzu tool, relational data was assessed using 1Spatial tool. Following quality assessment, the existing 1Spatial quality results in relational tables were uplifted using R2RML-F tool[15]. This tool is specifically chosen due its extension of R2RML's vocabulary with predicates for declaring specific functions (functions, function calls and parameter bindings) during conversion.

Third layer presents a unified quality graph with measurement results employing standardized tools such as ISO standards, OGC recommendation, and W3C best practises. As it was mentioned in Section 2 different standardization approaches are required from various OSi departments and its stakeholders. Geospatial data quality standards are mapped into one another to present measurement results according to the requested standardization tool and the results were presented using RDF Data Cube approach. The metadata model was also added to provide the data lineage in this layer.

[15] https://github.com/chrdebru/r2rml

Finally, in the upper layer the results are visualized using time series and bars which are enabled by posing Sparql queries.

4.2. Data Uplifting for Traditional Data

This task was performed to map the computation results of the relational data into Linked Data format by uplifting the quality rules defined for the PRIME2 dataset. There are two types of data quality results for the PRIME2 data which are uplifted: i) 1Spatial data assessment results ii) OGC data validation results.

Listing 1: Example R2RML Mapping for Schema Uplifting

```
<#TriplesMapForMetricClass>
rr:logicalTable <#Class-ValidationRule-View> ;
    rr:subjectMap [
    rr:template
"http://data.example.com/metric/{ORA_ERROR_ID}";
    rr:class rdfs:Class ;
] ;
rr:predicateObjectMap [
   rr:predicate rdfs:label ;
   rr:objectMap [ rr:column "ORA_ERROR_ID" ];
] ;
rr:predicateObjectMap [
   rr:predicate rdfs:subClassOf ;
   rr:objectMap [ rr:constant daq:Metric ] ;
] ;
rr:predicateObjectMap [
   rr:predicate rdfs:comment ;
   rr:objectMap
   [ rr:column "ERROR_DESCRIPTION" ] ;
] ;
rr:predicateObjectMap [
   rr:predicate daq:expectedDataType ;
   rr:objectMap [ rr:constant xsd:double ] ;
] .
```

Listing 2: Produced triples for R2RML Mapping

```
<http://data.example.com/metric/13356ERROR>
a
   <http://www.w3.org/2000/01/rdf-schema#Class>;
<http://www.w3.org/2000/01/rdf-schema#comment>
   "Adjacent points in a geometry are redundant"
;
<http://www.w3.org/2000/01/rdf-schema#label>
   "13356ERROR" ;
<http://www.w3.org/2000/01/rdf-schema#subClassOf>
   <http://purl.org/eis/vocab/daq#Metric> ;
<http://purl.org/eis/vocab/daq#expectedDataType>
   <http://www.w3.org/2001/XMLSchema#double> .
```

Listing 3
R2RML Mappings for Data Uplifting

```
<#CalculateValue>
        rrf:functionName "calculateValue" ;
        rrf:functionBody """
function calculateValue
(numInstances,totalInstances)
{
return 1-(numInstances/totalInstances);
}""" ;
```

Uplifting 1Spatial Results: In order to achieve this goal, 1Spatial data quality results were analyzed and then R2RML mappings were created to enable creation of a Linked Data representation of the quality results and schema. Each 1Spatial rule is defined as a quality metric which are defined by description. These metrics are classified into a dimension and a category by using Linked Data and ISO Standards.

The PRIME2 data stored in Oracle tables which are extracted from 1Spatial 1Integrate observations into daQ-specific instances [11] are two fold: *i)* A schema highlighting the category, dimension and metric, extending the daQ meta-schema was created and 186 different metric descriptions are extracted from the tables. Since 1Spatial 1Integrate tool did not have dimensions and categories, these metrics are mapped to the 7 different dimensions and 2 categories based on the ISO 19157 Standard using previous classification. *ii)* The quality observations are produced using the metadata extracted in the first step i.e. the quality metadata. The observations are produced for each PRIME2 sub-dataset by extracting and inferring quality values from both tables. In the 1Spatial 1Integrate database, the results are given in terms of the number of failing instances and the number of total instances. These are used to define a comparable [0.0 - 1.0] value across all metrics. More formally, value m_v is calculated as follows: $m_v = 1.0 - \frac{numberoffailinginstances}{totalnumberofinstances}$. Nonetheless, the raw data provided during uplifting is stored in the various provenance and profiling properties defined by the daQ meta-model.

Uplifting OGC Validation Results: OSi uses Oracle's Spatial & Graph Database to manage geospatial data, perform spatial analytic operations with the spatial features. OSi creates links between the national authoritative geospatial platform PRIME2 and semantic web using this tool. The

Listing 1 demonstrates the R2RML snippet for mapping an OGC validation metric into RDF. The metric describes an Oracle validation error numbered as

Listing 4
R2RML Mappings for Data Uplifting

```
<http://data.example.com/1spatialassessment/observation/13356ERROR-1-c>
      a       <http://purl.org/eis/vocab/daq#Observation> ;
      <http://purl.org/eis/vocab/daq#computedOn>
              <http://ontologies.adaptcentre.ie/dataset-hierarchy#BUILDING> ;
      <http://purl.org/eis/vocab/daq#isEstimate>
               false ;
      <http://purl.org/eis/vocab/daq#metric>
              <http://data.example.com/1spatialassessment/metric/13356ERROR-instance> ;
      <http://purl.org/eis/vocab/daq#value>
              0.9999997209017775" ;
      <http://purl.org/linked-data/cube#dataSet>
              <http://data.example.com/1spatialassessment/quality-graph/> ;
      <http://purl.org/linked-data/sdmx/2009/dimension#timePeriod>
              "31-JAN-20 00:00:00" ;
      <http://www.w3.org/ns/prov#generated>
              <http://data.example.com/1spatialassessment/observation/13356ERROR-1-c-profiling> ;
      <http://www.w3.org/ns/prov#wasGeneratedBy>
              <http://data.example.com/1spatialassessment/r2rmlconverter/> .
```

ORA13349 which identifies crossed polygon boundaries among spatial objects. As a result the the relational schema is materialized as RDF views demonstrated in Listing 2. On the other hand, materialized data is presented in Listing 3. This is a "BUILDING" dataset in PRIME2 which is assessed on the day of "31-JAN-20 00:00:00". The quality measurement is computed using R2RML function demonstrated in Listing 4. As a result of this computation the score was found as "0.9999997209017775".

4.3. Geospatial Data Quality Metrics for Linked Data

This section describes the geospatial data quality metrics in order to assess a dataset in terms of standards conformance including metadata, spatial reference systems and geometry classes. In order to create these metrics, a list of requirements were determined with the help and feedback from the OSi data quality team. Following the identification of these requirements the most important metrics are created for OSi. The details are described in our previous papers [36, 37]. We summarize the metric description and formula in Table 1. Metric computations are given as follows:

Geometry Extension Property Check (CS-M1): If the entity in the dataset is a member of class `geo:Geometry` then this metric checks the rate of employed `geo:asWKT` or `geo:asGML` properties in the dataset. This is evaluated using functions as *hasWKT(e)* or *hasGML(e)* which return a boolean value. The metric is computed as a rate over the whole dataset as follows (Note that the following metrics also compute their rate over the whole dataset and thus Equation 1 will not be repeated in each metric definition):

$$\sum_{i=1}^{e} \frac{\overline{e}(i)}{size(e)} \tag{1}$$

Geometry Extension Object Consistency Check (CS-M2): This metric checks the conformance of the dataset to the serialization requirement of OGC GeoSPARQL by checking the conformance of objects in terms of the order of use of coordinate system URI, spatial dimension and literal URI. Geometry data should consist of an optional URI identifying the coordinate reference system (e.g., CRS84, WGS 84) followed by WKT describing a geometric value. Spatial dimension may include polygon, multipolygon, line, point, or multilinestring shapes. Finally, the syntax should include the `geo:wktLiteral` URI declaring the object is a literal.

Geometry Classes and Properties Check (CS-M3, CS-M4): This metric checks the rate of declaration of geometry classes and properties in the datasets. The *hasGeometry(e)* and *hasDefaultGeometry(e)* functions check each entity and return a boolean value for property existence. The metric checks each entity which is an individual of the `geo:Geometry` class.

Spatial Dimension Existence Check(CS-M5): This metric assesses the rate of spatial dimension properties related to each entity in the dataset. It compares the to-

Table 1
New Geospatial Standards Conformance Quality Metrics

ID	Metric Name	Dimension	Formula
CS-M1	Geometry Extension Property Check	Completeness	$\overline{e} := \{e \mid \forall e \in class(geo : Geometry) \cdot hasWKT(e) \vee hasGML(e)\}$
CS-M2	Geometry Extension Object Consistency Check	Completeness	$\overline{e} := \{e \mid \forall e \in class(geo : Geometry) \cdot hasCRSURI(e) \wedge hasSpatialDimension(e) \wedge hasWKTLiteral(e))\}$
CS-M3	Geometry Classes and Properties Check	Completeness	$\overline{e} := \{e \mid \forall e \in class(geo : Geometry) \cdot hasGeometry(e))\}$
CS-M4	Geometry Classes and Properties Check	Completeness	$\overline{e} := \{e \mid \forall e \in class(geo : Geometry) \cdot hasDefaultGeometry(e))\}$
CS-M5	Spatial Dimensions Existence Check	Completeness	$\overline{e} := \{e \mid \forall e \in class(geo : Geometry) \cdot (isMultipolygon(e) \vee isPolygon(e) \vee isLine(e) \vee isPoint(e) \vee isMultilinestring(e))\}$
I-M6	Links to Spatial Things (internal&external)	Interlinking	$\overline{e} := \{e \mid \forall e \in class(geo : Geometry) \cdot hasST(e))\}$
I-M7	Links to Spatial Things from popular repositories	Interlinking	$\overline{e} := \{e \mid \forall e \in class(geo : Geometry) \cdot (isDBpedia(e) \vee isWikidata(e) \vee isGeonames(e)))\}$
CY-M8	Polygon and Multipolygon Check	Consistency	$\overline{e} := \{e \mid \forall e \in class(geo : Geometry) \cdot (hasClosedPolygon(e))\}$
T-M9	Freshness Check	Timeliness	f = (max(1 - c/v, 0))

tal number of spatial dimensions (multipolygon, polygon, line, point, multilinestring) described for each entity in the dataset to the overall number of entities.

Links to Spatial Things Check (I-M6, I-M7): W3C SDOTW suggests two types of links for Spatial things: i) links to other spatial things using an object with its own URI within dataset or to other datasets decreasing the computational complexity and enriching the data semantically ii) links to spatial things from popular repositories which increases the discoverability of the dataset. However, the challenge in this metric is that it is not possible to understand if a link has spatial extent without visiting the other resource. Thus, first a set of different pay-level-domains are detected manually and according to the used schema, the rate of the links are computed as an efficient approximation.

I-M6 Metric Computation: First the metric detects the rate of entities having links to external spatial things in other datasets and internal spatial links within dataset. In I-M6, the *hasST(e)* function checks the entities with these links and later this number is divided into the overall number of entities.

I-M7 Metric Computation: This metric detects the rate of entities having links to external spatial links in popular and highly referenced datasets. In this work, we specifically looked at the usage of DBpedia, Wikidata and Geonames datasets. We counted the entities with these links and divided to the overall entity number.

Consistent Polygon and Multipolygon Usage Check (CY-M8): This metric checks the equality of the starting and end points of polygons. Each polygon in a multipolygon must be checked. We measure the rate of correctly described polygons and multipolygons in a dataset. In metric CY-M8 the function `hasClosedPolygon(e)` detects the correct usage for each entity in the dataset.

Freshness Check (T-M9): This metric checks the age of the data (f) by looking at the creation time and when it was last updated to the recent version. This metric was used as an updated version from [12]. In this formula, Volatility (v) is "the length of time the data remains valid" which is analogous to the shelf life of perishable products; Currency (c) is "the age of the data when it is delivered to the user" [17]. This metric is computed at the dataset and not instance level level due to lack of information in the entity level.

4.4. Uniform view for diverse standards using Mapping

This section introduces the mapping of all relevant data quality ontologies by defining semantic links according to the W3C OWL recommendation among standard ontologies namely those defined by ISO/TC 211, ISO/TC 184,ISO/IEC JTC 1/SC 7 and the W3C Data on the Web Best Practices recommendation data quality vocabulary with the goal of creating a uniform quality graph view. This approach is an extension of the correspondences between quality dimensions in ISO/IEC 25012 and Zaveri et al. [2, 31].

The followed steps are: i) discovering and investigating the quality standards relevant to the data requirements ii) comparing the data quality dimensions employed in different standards to discover the intersections or the similarities between them. iii) creating missing RDF models for the quality standards. It was seen that some standardization bodies already imple-

Table 2
Example Standards Dimension Mapping

```
iso25012dqi:Completeness owl:sameAs dqm:Completeness.
iso25012dqi:Completeness owl:sameAs iso8000dqi:Complete.
iso25012dqi:Completeness owl:sameAs iso8000dqi:Completeness.
iso25012dqi:Completeness owl:sameAs iso19157:DQ_Completeness.
dqm:Completeness owl:sameAs iso8000dqi:Completeness.
dqm:Completeness owl:sameAs iso19157:DQ_Completeness.
iso8000dqi:Complete rdfs:subClassOf dqm:Completeness.
iso8000dqi:Complete rdfs:subClassOf iso8000dqi:Completeness.
iso8000dqi:Complete rdfs:subClassOf iso19157:DQ_Completeness.
iso8000dqi:Completeness owl:sameAs iso19157:DQ_Completeness.
```

mented the RDF models of the their standards such as ISO 19157[16] or W3C Data Quality Vocabulary[17]. Thus non-existing models are implemented by us based on the daQ model of Luzzu framework to be integrated into the end-to-end knowledge graph. iv) creating mappings between the standards.

Table 3 presents the standards and dimensions which are corresponding to each other[18]. ISO 25012 standard was employed as the main object of mapping due to its broad inclusiveness of the quality dimensions and, thus, other standards were mapped to this standard. 3 types of relations were accommodated for the quality dimensions: equality (concept unification) relationship `owl:sameAs`[19], inclusion relationship (broader/narrower concept) `rdfs:subClassOf`[20] and similarity relationship `ov:similarTo` [21]. A part of the mapping can be seen for *Completeness* dimension in Table 2. The predicates are shortened for the sake of space.

In many cases, standards use the same term in subtly differing ways, leading to more complex mappings. For example, the standards descriptions of the Completeness dimensions are given below:

Completeness (ISO 25012) The degree to which subject data associated with an entity has values for all expected attributes and related entity instances in a specific context of use.

Complete (ISO 8000) Information is perceived to be mapped completely to entities in the domain of interest in a reliable 1:1 mapping.

[16] https://def.isotc211.org/ontologies/iso19157/

[17] https://www.w3.org/TR/vocab-dqv/

[18] Following dimensions are omitted from the table due to not having mappings with other standards Flexible content, Flexible layout (ISO 8000), Recoverability, Precision (ISO 25012), Relevancy, Interpretability (Linked Data)

[19] @prefix http://www.w3.org/2002/07/owl#

[20] @prefix http://www.w3.org/2000/01/rdf-schema#

[21] @prefixhttp://open.vocab.org/terms#

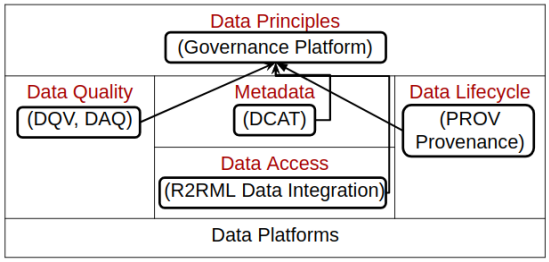

Fig. 5. Data Governance Platform [20]

Given example shows that two different types of system properties are described in the descriptions even though as a word they seem to be describing the same notion. In this specific example, while *Complete* has a narrow definition, *Completeness* has a more general definition, thus, `rdfs:subClassOf` logical relation was created between these dimensions.

4.5. Data Lineage Model

A data lineage model was needed to structure the metadata repository and provide effective data quality monitoring. The conceptual model was created based on Khatri *et al.*'s data governance model including data principles, data quality, lifecycle, access and metadata [20] as presented in Fig. 5. The created model demonstrates high-level structure of the data origin and the evolution of data over time, as well as, describing the datasets and their relationships in the end-to-end data quality pipeline. Each dataset is described using the metadata model of the Data Catalog Vocabulary (DCAT)[22]. The data lineage model consists of several components:

– **Provenance:** This W3C standard is used to create the links between datasets through DCAT de-

[22] https://www.w3.org/TR/vocab-dcat-2/

Table 3

Comparison of Standards Conformance Quality Dimensions. Bold dimension is the subject of the mapping triple (usually an ISO 25012 dimension as it has the widest coverage), the predicate is defined in the mapping column, the object is defined in the non-bold column.

ISO 25012	ISO 19157	ISO 8000			Linked Data	Suggested Mapping
		Semantic	Syntactic	Pragmatic		
Completeness	Completeness	Completeness	-	-	Completeness	owl:sameAs
Completeness	-	-	-	**Complete**	-	rdfs:subClassOf
Consistency	-	Consistency	-	-	Consistency	owl:sameAs
Consistency	**Logical consistency**	-	**Entity integrity**	-	-	rdfs:subClassOf
Accuracy	-	Accuracy	-	-	-	owl:sameAs
Accuracy	**Positional accuracy** **Thematic accuracy**	-	-	-	**Semantic Accuracy**	rdfs:subClassOf
Currentness		-	-	-	Timeliness	owl:sameAs
Currentness	Temporal quality	-	-	-		rdfs:subClassOf
Compliance	-	Compliance	-	-	-	owl:sameAs
Compliance	-	-	**Domain integrity** **Referential int.** **User defined int.**	-	**Representational Conciseness**	rdfs:subClassOf
Confidentiality	-	-	-	-	Security	rdfs:subClassOf
		-	-	Secure	Security	owl:sameAs
Traceability	-	-	-	-	Security	ov:similarTo
Traceability	-	-	-	-	Trustworthiness	ov:similarTo
Credibility	-	-	-	-	Trustworthiness	rdfs:subClassOf
Efficiency	-	-	-	-	Performance	owl:sameAs
Understandability	-	-	-	-	Understandability	owl:sameAs
Understandability		-			Versatility	ov:similarTo
Availability	-	-	-	-	Availability	owl:sameAs
Accessible	-	-	-	Accessibility		owl:sameAs
Accessible	-	-	-	-	Interlinking	ov:similarTo
Accessible	-	-	-	-	Licensing	ov:similarTo
Portability	-	-	-	-	Interoperability	ov:similarTo
-	**Usability element**	-	-	Useful	-	ov:similarTo

scriptions. When the metadata is provided for end to end lineage, a reference point provides a complete audit trail of that data point of interest from sources to its final destinations. This links the OSi datasets into a pipeline. The links between the datasets allow applications to trace errors back to the root cause in a data analytic process wherever it happens.

– **Data Quality:** Describing data quality at different points of the end-to-end data pipeline allows users to track the quality of subdatasets during their historical evolution. The quality of the data is tracked through that specific data point in the data lineage which allows discovering the source of the problems. We present data quality with different aspects: i) Spatial / temporal resolution (data granularity) ii) Quality assessments expressed with quantitative test results.
– **Geospatial information:** This type of metadata is necessary to have a high level summary of the geospatial information covered by the datasets.
– **UN-GGIM Data Themes:** OSi data governance and quality reporting must be according to priority national data themes, which are aligned to the globally endorsed fundamental geospatial data themes. Including this metadata in the model allows OSi to analyze the data according to the main geospatial themes and the stakehold-

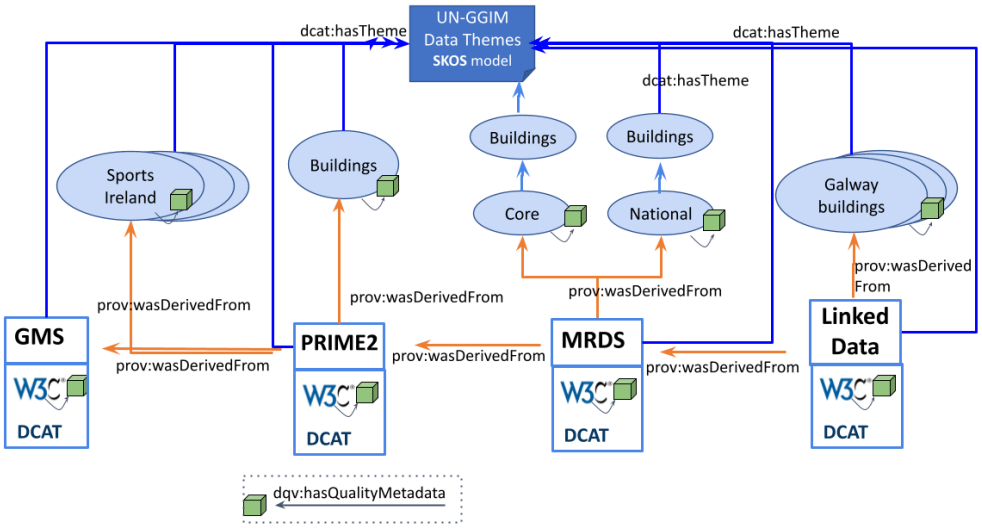

Fig. 6. Conceptual Data Lineage Model

ers to visualize it according to their requirements. This vocabulary is created using Simple Knowledge Organization System (SKOS) concepts[23] [23]. The details of the vocabulary was described in our previous paper [38].

The OSi business data lineage view of the OSi data pipeline model 6 defines the structure to describe the datasets (GMS [Sensor Data], PRIME2, MRDS, and Linked Data) and subdatasets (Buildings, Core, etc.), of the OSi Geospatial Data Publishing Pipeline through metadata vocabularies. The metadata information is described based on the W3C standards DCAT, PROV ontology, DQV and UN-GGIM data themes. The DCAT properties `dcat:hasTheme` defines the data theme(s), from the list provided in the UN-GGIM, for each dataset and subdataset (Fig. 6 purple arrows), respectively. The `prov:wasDerivedFrom` property from the PROV ontology defines a dataset or subdataset as the result of a derivation or transformation from a pre-existing source dataset or subdataset (Fig. 6 orange arrow).

Listing 5 presents a piece of data catalog created for Data Quality Pipeline. This code snippet shows the provenance data (`prov:wasDerivedFrom`, `dct:created`, `dct:modified`), quality data (`dqv:hasQualityMetadata`), standardization data (`dct:conformsTo`), data themes (`dcat:theme`) and spatial aspect (`dct:spatial`) of the data in one place using DCAT[6]. The metadata is created for various granularity and hierarchical layers to analyze the query results in various levels. The query results performed on this piece of data is presented in Figure 8.

4.6. *A Unified Quality Knowledge Graph Interface for Data Pipeline*

In order to to deliver meaningful insights into data by displaying noteworthy correlations, a dashboard is implemented consisting of four main parts:

Pipeline: The pipeline page serves as the dashboard's home screen. The pipeline page displays a high-level depiction of OSi's data flow pipe and graphs are connected to these tiles when the dashboard is first opened. After connecting a graph to a node, the overall data quality of that data is displayed, and the details of that data can be viewed by clicking into the node.

Pipeline Node: After clicking into a pipeline node, you may see a high-level summary of the dataset's content. This page displays the overall dataset quality, as well as a bar chart depicting past assessment results and a list of dimensions with aggregate quality results compared to a threshold. These attributes are divided into categories. When you click on any of the dimensions, you'll be taken to a new screen that displays the dimension's assessment quality as well as a list of the

[23] https://www.w3.org/TR/skos-primer/

Listing 5
Data Catalog for the Data Quality Pipeline

```
<http://ontologies.adaptcentre.ie/dataset-hierarchy#galway-building-linked-data>
    a dcat:Dataset ;
    a prov:Entity;
    prov:wasDerivedFrom <http://ontologies.adaptcentre.ie/dataset-hierarchy#prime2>;
    dc:title "Linked Dataset for Galway Building"@en ;
    dc:description "This is a subset of Linked Dataset
         covering the building dataset from PRIME2."@en ;
    dct:created "2019-10-26"^^xsd:date ;
    dct:modified "2020-09-10"^^xsd:date ;
    dcat:theme <http://purl.org/eis/vocab/unggim-data-themes#Buildings-Settlements>,
    <http://purl.org/eis/vocab/unggim-data-themes#Addresses>;
    dqv:hasQualityMetadata <https://w3id.org/lodquator/resource/232a-440a-b483-2fcbcf652d5b>;
    dcat:distribution <http://data.geohive.ie/dumps/building/GALWAY_BUILDING_DATA_ITM.n3>;
    dct:conformsTo <http://www.opengis.net/def/crs/EPSG/0/2157>;
    dct:conformsTo std:geoDCAT-AP ;
    dct:spatial [
        dcat:bbox "<gml:Envelope srsName=\"http://www.opengis.net/def/crs/OGC/1.3/CRS84\">
        <gml:lowerCorner>-62.9951 -21.378367</gml:lowerCorner>
        <gml:upperCorner>55.813367 70.620781</gml:upperCorner></gml:Envelope>"^^gsp:gmlLiteral ,
    "POLYGON((-62.9951 70.620781,55.813367 70.620781,55.813367 -21.378367,-62.9951
    -21.378367,-62.9951 70.620781))"^^gsp:wktLiteral ].


<http://data.geohive.ie/dumps/building/GALWAY_BUILDING_DATA_ITM.n3>
    a dcat:Distribution ;
    dc:description "An n3 serialization of feed of Galway Building Data"@en ;
    dc:mediaType "text/n3" ;
    dc:license <http://purl.oclc.org/NET/rdflicense/cc-by4.0> ;
    dct:conformsTo std:geoDCAT-AP ;
    dct:conformsTo <http://www.opengis.net/def/crs/EPSG/0/2157>.
```

metrics that have been applied to it. These metrics also include a quality assessment result, a success threshold, and a statement that explains the metric's purpose.

Data Quality Analysis: The reporting page enables tracking the data quality of an OSi data pipeline node over time, as well as see a unified view of quality dimensions: ISO 19157 vs Linked Data, and see how these quality dimensions have changed over time. The data quality analysis page in Fig. 7 is divided into three parts: a bar chart depicting the data quality over time of the pipeline nodes, a second bar chart depicting the quality dimensions of the pipeline nodes, and a navigation bar on the left with a slider for changing standards and checkboxes for selecting/deselecting dimensions. The toggle can be clicked to swap between quality standards and the quality dimension bar chart reflects these changes. By using the options in the navigation bar, you may adjust the quality over time bar chart to show individual dimensions.

Reporting: The reporting options on this page are based on the OSi data catalog, which is linked to data theme, provenance and data quality metadata. The dashboard page has several filters on the left side of the page which allows users to click interactively. By clicking on the filters, the user can pose numerous questions with different views and semantically display the dataset relations. The dashboard is an attempt to present a comprehensive perspective of the data, including the flow and connections of the many datasets. The end-to-end view allows users to see the various quality findings and where data quality issues occur across the process. Blue filters show data quality dimensions whereas yellow ones present the data standards and red filters present the data lineage. The selection can be performed with different variations from the existing filter types. The datasets are displayed on the right of the dashboard panel based on the left-hand selections.

5. Evaluation

This section describes a first study showing our new metrics in operation with experimental set-up in Sec-

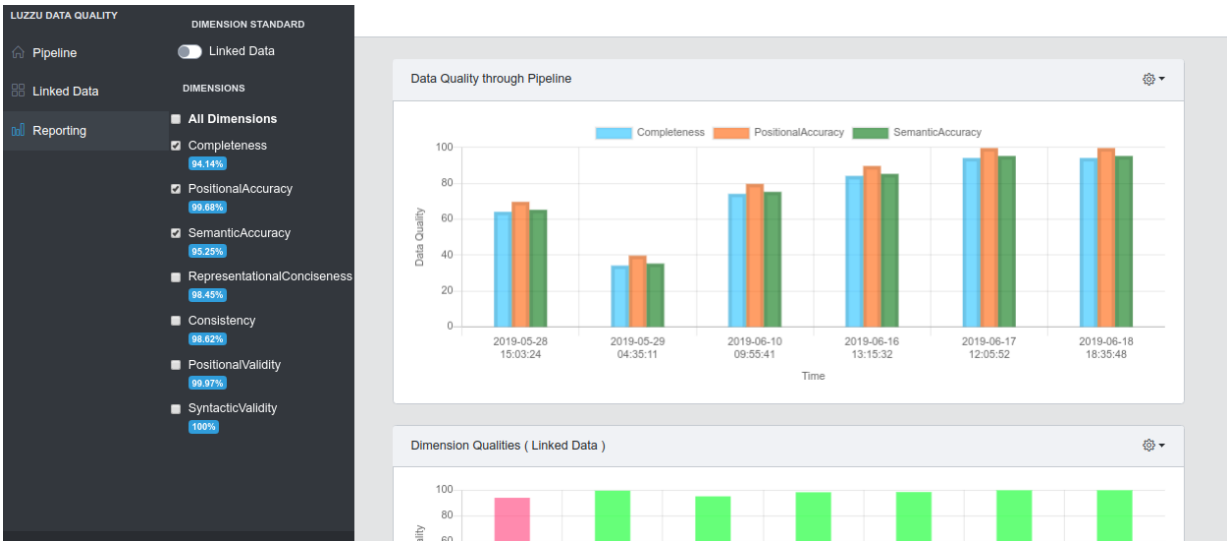

Fig. 7. End to end Dashboard for Data Quality Analysis. Showing data lineage information over time according to ISO 25012 Linked Data Quality Model. An ISO 19157 view of data quality can be displayed by changing the "dimension Standard" toggle switch.

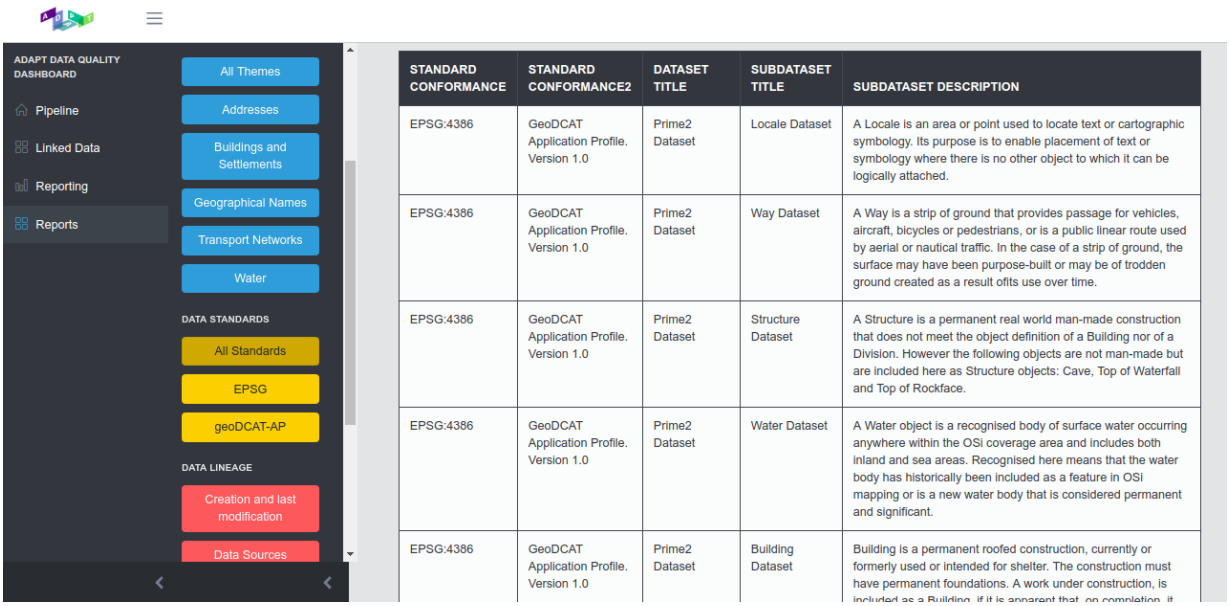

Fig. 8. End to end Dashboard Reporting.

tion 5.1 followed by the usability evaluation of the defined metrics in Section 5.3 and the lessons learned in Section 6.

5.1. Experimental Setup

Experiments were executed to measure the metrics' ability to detect the standards compliance of GLD datasets, as well as, the extent of standards compliance of published Open GLD to meet OSi's requirements. Investigation was performed by implementing new metrics as scalable Luzzu plug-ins in Java and assessing a set of four open GLD datasets. We used a computer with Intel i7 8th generation processor and 8GB memory.

Datasets: Major open topographical geospatial datasets describing political or administrative boundaries were chosen to ensure geometrical features were repre-

sented in each dataset. Despite this selection, there is considerable variation in the datasets in number of triples, size, languages and used coordinate reference systems (CRS) as depicted in Table 4. Ordnance Survey Ireland (OSi) is the national mapping agency of Ireland and they publish a subset of their data as Linked Open Data. The OSi boundaries dataset describes political and administrative boundaries in Ireland. Ordnance Survey UK is the national mapping agency of the United Kingdom and they also publish their data partially as Linked Data. LinkedGeoData is provided by the University of Leipzig by converting OpenStreetMap data to Linked Data. Greece LD is provided by the University of Athens as part of the TELEIOS project.

Method: Assessments were performed on each dataset using the Luzzu framework. In addition to assessing the full datasets, subset were also assessed to provide a common baseline for comparison between datasets. Observations for the nine metrics presented in in our previous paper [37] were collected as quality metadata using the daQ vocabulary[10][24]. This paper further presents the evaluation of the metrics for the OSi use case.

By applying a suite of over 500 quality rules to the PRIME2 topographic dataset it is possible to assure very high levels of compliance with those rules. However execution of the explicit rules over 50 million spatial objects can take days, even on custom high end hardware like a state of the art Oracle exadata platform. This does not pose a problem when a regular flow of localised transactions is used to update the PRIME2 model but when large-scale data transformations must be carried out (for example for schema updates or to fix systematic errors identified in older releases) then the time required is unsustainable. In this case the use of probabilistic (sampling-based) metrics as deployed in Luzzu for computationally expensive metrics is an advantage.

5.2. A Hierarchical Data Quality Knowledge Graph

Aggregated Results for Completeness: Average aggregated results shows around 50% completeness for the overall data at most for these datasets. It should be noted that the Greek LGD and OS UK datasets were created before the standardization efforts so they have lower rates. However, an important question to ask is how publishers could be encouraged to migrate their data to new approaches and standards, especially, when conformance is a straightforward change in the prefixes in the data. Given that datasets are not updated for a long period, indeed, LOD cloud is about to become a museum for the datasets [12].

*Aggregated Results for Interlinking:*This metric is very similar to the Debattista etal. [12] external link data providers metric which calculates all the datasets in the LOD cloud where LOD cloud has average 27% external links to other datasets. Our results show that compared to the LOD cloud, these datasets have a higher rate of external spatial links but a much lower rate of links to popular datasets. If we consider the aggregated result for the Interlinking dimension, the rate is similar to LOD cloud rate with a mean of 27%.

5.3. Metric Design Evaluation

Heinrich et al. [18] have defined a set of five design requirements for effective data quality metrics for both decision making under uncertainty and economically oriented data quality management. This section evaluates our new metrics against these five requirements (summarised in Table 5).

Existence of minimum and maximum metric values (MR1): As defined by Heinrich et al., data quality metrics should take values only within a specified range. The minimum values should represent poorest data quality and the maximum representing highest data quality. Each value within this range should represent different data quality levels. All our metrics are defined over the bounded interval [0-1] representing gradually increasing quality levels. Thus all the metrics fulfill this requirement.

Interval-scaled metric values (MR2): The data quality metric must represent the computation results as interval-scaled or ratio-scaled values. This avoids metrics with arbitrary scales such as poor, good, or best. Metric values (except T-M9) are interval scaled, the impact of a data quality improvement measure can thus be assessed precisely.

Quality of the Configuration Parameters and the Determination of the Metric Values (MR3): The scientific quality criteria (i.e., objectivity, reliability, and validity) must be satisfied by any metric configuration parameters. The provided metrics have formal, mathematical formulae for calculating the scores that allow for an objective and reliable determination based on defined data quality dimensions (completeness, consistency, interlinking). All metrics fulfil this

[24]https://opengogs.adaptcentre.ie/OrdnanceSurveyIreland-OSi

Table 4
Dataset Summary

Dataset	#Triple	Size	Languages	Coordinate System	CRS
OSi	1936763	274M	EN,GA	GEOsparql, Open Vocab, RDF, RDFS, OSi	IRENET95 / ITM
OS UK	64641	224.1M	EN	RDF, RDFS, OS UK	WGS 84
LinkedGeoData	464193	1.5G	EN,Various	NeoGeo, RDF, RDFS, LinkedGeoData	WGS 84
Greece LD	24583	183M	EN,GR	RDF, RDFS,Greece LD	WGS 84

Table 5
Heinrich et al. Metric Requirement Testing Results (Y=yes, P=partial)

Requirement Metric	CS-M1	CS-M2	CS-M3	CS-M4	CS-M5	I-M6	I-M7	CY-M8	T-M9
Minimum and maximum values (MR1)	Y	Y	Y	Y	Y	Y	Y	Y	Y
Interval-scale (MR2)	Y	Y	Y	Y	Y	Y	Y	Y	Y
Parameters and determination (MR3)	Y	Y	Y	Y	Y	Y	Y	Y	P
Sound aggregation (MR4)	Y	Y	Y	Y	Y	Y	Y	Y	Y
Economic efficiency (MR5)	Y	Y	Y	Y	Y	Y	Y	Y	P

except for T-M9 as it is not possible to determine a fixed value for the configuration parameter shelf life of the metric.

Sound Aggregation of the Metric Values (MR4): A data quality metric must be applicable to single data values as well as to sets of data values. The metric should be performed in different levels of data with consistent aggregation values. In all cases, we propose normalised metrics that are scaled to the number of triples or geospatial terms they assess. Thus this requirement is satisfied.

Economic Efficiency of the Metric (MR5): This requirement addresses the metric's utility from an economic perspective. Application of the data quality metric should provide a cost beneficial effect on the business, thus computation time should not be excessive. The metric should support effective decision making. All of the metrics can be calculated with mathematical formulations automatizing the computations in an effective way at low cost. They have proved effective for decision making in OSi. All the metrics fulfill this requirement except perhaps T-M9 since it depends on knowledge of the dataset creation date, which is not always available.

6. Lessons learned

The ADAPT Centre developed this work over two years collaboration with the Geospatial Services, Data Governance & Quality department in OSi and knowledge exchange was a key outcome. This was facilitated by quarterly workshops with senior stakeholders as well as regular weekly meetings between the design and implementation teams. Key lessons learned from the deployment of semantic web technologies and standards for unified data governance are described below.

It is essential to ensure the quality of the data to make informed and effective decisions. This depends on selecting, designing data quality metrics and thresholds. despite the rapid advances on general purpose linked data metrics in the last decade[9, 28, 40], domain-specific metrics needed to developed for the AEC domain. If they are not defined effectively, these metrics can lead to poor decisions and economic losses. An open, standards-based quality metric definition language such as is found in the semantic knowledge model underlying the Luzzu framework and R2RML uplifting is useful as it generates self-describing plug and play metrics and quality observation metadata that can be combined or transformed in applications and manipulated in views for presentation to specific audiences. Furthermore, employing metric requirement testing to evaluate these metrics allowed us to assess if they were appropriate for the organization.

Despite the broad adoption of Linked Data standards are still transitioning to Linked Data implementation e.g. ISO/TC 211 (specifies methods, tools and services for geographic data management) a continuous effort to create the Linked Data ontologies required, whereas some standards do not have any initiative for this. On the other hand, there is a great poten-

tial for semantic modeling of data quality standards. Another challenge was to decide the correct type of relation between the dimensions. Despite the fact that some quality dimensions have the same name in different standards, it was necessary to analyze the definitions in depth in order to prevent making mistakes in the mapping. Getting validation of the mapping results by OSi domain experts was also important to the success of our approach.

We've learned that putting together a useful data quality dashboard isn't easy. Because the dashboard should be able to efficiently and clearly manage and visualize various data concepts. The appropriate level of abstraction should be determined and provided to the user, and it should be tailored to the requirements of various stakeholders who may think about quality in terms of different standards depending on the domain of application (e.g. surveying vs sales vs economic and statistical modelling) but also in terms of different goals (e.g. process improvement vs customer-facing data quality certification) . However, Linked Data offers a great level of flexibility, which aids in the unification of quality assessments and the flexible processing and presentation of data.

We combined numerous concepts, such as data themes to classify data based on the needs of the company and provenance to monitor data evaluation. Furthermore, the mix of diverse standards and concepts allowed us to create rich queries which combine quality results with provenance metadata and standards compliance, allowing us to display them at various levels of granularity visualizing on the dashboard. The provenance information was especially important to show data flows within the organisation and support temporal queries that highlighted the evolution of the datasets. The Semantic Web approach enabled rich data fusion across different organisational contexts into a unified data governance system without requiring any loss to the underlying data which connects multiple data collection systems. The use of data quality dimensions as a unifying concept was particularly important for the final system as it allowed us to combine different metrics, collected in different systems to get a global view of the evolution of user-centric perspectives on data quality that were previously not correlated. It also provided an important mechanism for abstracting the reports of quality rule failures into single numbers that were easy for staff to interpret or drill down into.

Overall, we received very positive feedback from the OSi staff and stakeholders. They especially liked the ability to dynamically present the same quality data from multiple quality standards' perspectives. This was especially important as at the beginning of this work it was unknown which standards were the most important ones to comply with and this will vary as more stakeholders and use cases are introduced to the system. OSi gained an advantage by creating and classifying metrics based on 1Spatial rule-based data validation into the ISO 19157 data quality framework, since they were previously limited to reviewing the raw outputs of validation rules, which was difficult to track over time for trends. This work was also necessary to ensure quality traceability along the data pipeline, in addition to our mappings.

7. Conclusions and Future Work

This research looked into how an uniform semantic information space for data quality measures may be used to give end-to-end views of data quality from disparate quality assessment instruments by providing a set of mappings between standard semantic models of data quality. In the OSi use case (Section 2), this unified quality graph allowed us to present the heterogeneous data with different formats and assessment of different tools (Luzzu and 1Spatial 1Integrate) to be presented in a homogeneous way (Section 4). A new web-based dashboard was designed and implemented to visualize the quality analysis and changes through time (Section 4.6). It was seen even though there is a lot of standards representation in Linked Data that it is still in the transforming phase and they are mostly disconnected (Section 4.4). However, it is hoped that this study could be a starting point for researchers who would like to interlink their data and present in a homogeneous way and perform a standards-based assessment for its own datasets. On the other hand, this work might have an impact on the standardization approaches and evolve the way they are implemented for instance with the RDF data model.

In future work, we intend to add more standards to our compliance governance dashboard and more features that are functional to the users. The next steps will expand the data quality model to include FAIR principles, data value dimensions, include R2RML mappings support for more quality tools.

Acknowledgements

This research received funding from the 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_2] and Ordnance Survey Ireland. We would like to thank Jeremy Debattista for his contribution to the work and architecture.

References

[1] International standardization organization. https://ec.europa.eu/eip/ageing/standards/ict-and-communication/data/iso-19000-series_en. Access date:15.09.2020.

[2] R. Albertoni and A. Isaac. Data on the web best practices: Data quality vocabulary. *W3C Working Draft*, 19, 2016.

[3] Z. Chen and N. Chen. Provenance information representation and tracking for remote sensing observations in a sensor web enabled environment. *Remote Sensing*, 7(6):7646–7670, 2015.

[4] G. Closa, J. Masó-Pau, B. Proß, and X. Pons. W3C PROV to describe provenance at the dataset, feature and attribute levels in a distributed environment. *Comput. Environ. Urban Syst.*, 64:103–117, 2017.

[5] G. Closa, J. Masó-Pau, A. Zabala, L. Pesquer, and X. Pons. A provenance metadata model integrating ISO geospatial lineage and the OGC WPS: conceptual model and implementation. *Trans. GIS*, 23(5):1102–1124, 2019.

[6] W. W. W. Consortium et al. Data catalog vocabulary (dcat). 2014.

[7] S. Das, S. Sundara, and R. Cyganiak. R2RML: RDB to RDF mapping language. *W3C Recommendation 27 September 2012*, 2012.

[8] 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.

[9] J. Debattista, E. Clinton, and R. Brennan. Assessing the quality of geospatial linked data–experiences from ordnance survey ireland (osi). 2018.

[10] J. Debattista, C. Lange, and S. Auer. daq, an ontology for dataset quality information. In C. Bizer, T. Heath, S. Auer, and T. Berners-Lee, editors, *Proceedings of the Workshop on Linked Data on the Web co-located with the 23rd International World Wide Web Conference (WWW 2014), Seoul, Korea, April 8, 2014*, volume 1184 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2014.

[11] J. Debattista, C. Lange, and S. Auer. Representing dataset quality metadata using multi-dimensional views. In *Proceedings of the 10th International Conference on Semantic Systems, SEMANTICS 2014, Leipzig, Germany, September 4-5, 2014*, pages 92–99, 2014.

[12] J. Debattista, C. Lange, S. Auer, and D. Cortis. Evaluating the quality of the lod cloud: An empirical investigation. *Semantic Web*, (Preprint):1–43, 2018.

[13] 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. Springer, 2017.

[14] L. Di, Y. Shao, and L. Kang. Implementation of geospatial data provenance in a web service workflow environment with ISO 19115 and ISO 19115-2 lineage model. *IEEE Trans. Geosci. Remote. Sens.*, 51(11):5082–5089, 2013.

[15] J. F. T. Djuedja, F. H. Abanda, B. Kamsu-Foguem, P. Pauwels, C. Magniont, and M. Karray. An integrated linked building data system: AEC industry case. *Adv. Eng. Softw.*, 152:102930, 2021.

[16] P. Geyer, C. Koch, and P. Pauwels. Fusing data, engineering knowledge and artificial intelligence for the built environment. *Adv. Eng. Informatics*, 48:101242, 2021.

[17] O. Hartig and J. Zhao. Using web data provenance for quality assessment. In *Proceedings of the International Workshop on Semantic Web and Provenance Management, Washington DC, USA*, 2009.

[18] B. Heinrich, D. Hristova, M. Klier, A. Schiller, and M. Szubartowicz. Requirements for data quality metrics. *Journal of Data and Information Quality (JDIQ)*, 9(2):1–32, 2018.

[19] 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.

[20] V. Khatri and C. V. Brown. Designing data governance. *Communications of the ACM*, 53(1):148–152, 2010.

[21] 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.

[22] K. McGlinn, R. Brennan, C. Debruyne, A. Meehan, L. McNerney, E. Clinton, P. Kelly, and D. O'Sullivan. Publishing authoritative geospatial data to support interlinking of building information models. *Automation in Construction*, 124:103534, 2021.

[23] A. Miles, B. Matthews, M. Wilson, and D. Brickley. Skos core: simple knowledge organisation for the web. In *International conference on dublin core and metadata applications*, pages 3–10, 2005.

[24] M.-A. Mostafavi, G. Edwards, and R. Jeansoulin. An ontology-based method for quality assessment of spatial data bases. 2004.

[25] P. Pauwels and W. Terkaj. Express to owl for construction industry: Towards a recommendable and usable ifcowl ontology. *Automation in construction*, 63:100–133, 2016.

[26] M. Perry and J. Herring. Ogc geosparql-a geographic query language for rdf data. *OGC implementation standard*, 40, 2012.

[27] F. Radulovic, N. Mihindukulasooriya, R. García-Castro, and A. Gómez-Pérez. A comprehensive quality model for linked data. *Semantic Web*, 9(1):3–24, 2018.

[28] F. Radulovic, N. Mihindukulasooriya, R. García-Castro, and A. Gómez-Pérez. A comprehensive quality model for linked data. *Semantic Web*, 9(1):3–24, 2018.

[29] M. A. Sadiq, G. West, D. A. McMeekin, L. Arnold, and S. Moncrieff. Provenance ontology model for land administration spatial data supply chains. In *2015 11th International Conference on Innovations in Information Technology (IIT)*, pages 184–189. IEEE, 2015.

[30] K. Sun, Y. Zhu, P. Pan, Z. Hou, D. Wang, W. Li, and J. Song. Geospatial data ontology: the semantic foundation of geospatial data integration and sharing. *Big Earth Data*, 3(3):269–296, 2019.

[31] J. Tandy, L. van den Brink, and P. Barnaghi. Spatial data on the web best practices. *W3C Working Group Note*, 2017.

[32] D. Thakker, P. Patel, M. I. Ali, and T. Shah. Semantic web of things for industry 4.0. *Semantic Web*, 11(6):885–886, 2020.

[33] W. B. United Nations Statistics Division, Global Geospatial Information Management. A Strategic Guide To Develop And Strengthen National Geospatial Information Management Part 1: Overarching Strategic Framework. http://ggim.un.org/meetings/GGIM-committee/8th-Session/documents/Part%201-IGIF-Overarching-Strategic-Framework-24July2018.pdf, 2018. [Online; accessed 11-Dec-2018].

[34] N. van Oorschot and B. van Leeuwen. Intelligent fire risk monitor based on linked open data. In *ISCRAM*, 2017.

[35] B. Yaman and R. Brennan. Linkeddataops: Linked data operations based on quality process cycle. In D. Garijo and A. Lawrynowicz, editors, *Proceedings of the EKAW 2020 Posters and Demonstrations Session co-located with 22nd International Conference on Knowledge Engineering and Knowledge Management (EKAW 2020), Globally online & Bozen-Bolzano, Italy, September 17, 2020*, volume 2751 of *CEUR Workshop Proceedings*, pages 31–35. CEUR-WS.org, 2020.

[36] B. Yaman, K. Thompson, and R. Brennan. Quality metrics to measure the standards conformance of geospatial linked data. In K. L. Taylor, R. S. Gonçalves, F. Lécué, and J. Yan, editors, *Proceedings of the ISWC 2020 Demos and Industry Tracks: From Novel Ideas to Industrial Practice co-located with 19th International Semantic Web Conference (ISWC 2020), Globally online, November 1-6, 2020 (UTC)*, volume 2721 of *CEUR Workshop Proceedings*, pages 109–114. CEUR-WS.org, 2020.

[37] B. Yaman, K. Thompson, and R. Brennan. Standards conformance metrics for geospatial linked data. In *Iberoamerican Knowledge Graphs and Semantic Web Conference*, pages 113–129. Springer, 2020.

[38] B. Yaman, K. Thompson, and R. Brennan. A SKOS taxonomy of the UN global geospatial information management data themes (short paper). In B. Yaman, M. A. Sherif, A. N. Ngomo, and A. Haller, editors, *Proceedings of the 4th International Workshop on Geospatial Linked Data (GeoLD) Co-located with the 18th Extended Semantic Web Conference (ESWC 2021), Virtual event (instead of Hersonissos, Greece), June 7th, 2021*, volume 2977 of *CEUR Workshop Proceedings*, pages 89–96. CEUR-WS.org, 2021.

[39] J. Yuan, P. Yue, J. Gong, and M. Zhang. A linked data approach for geospatial data provenance. *IEEE Trans. Geosci. Remote. Sens.*, 51(11):5105–5112, 2013.

[40] A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, and S. Auer. Quality assessment for linked data: A survey. *Semantic Web*, 7(1):63–93, 2016.