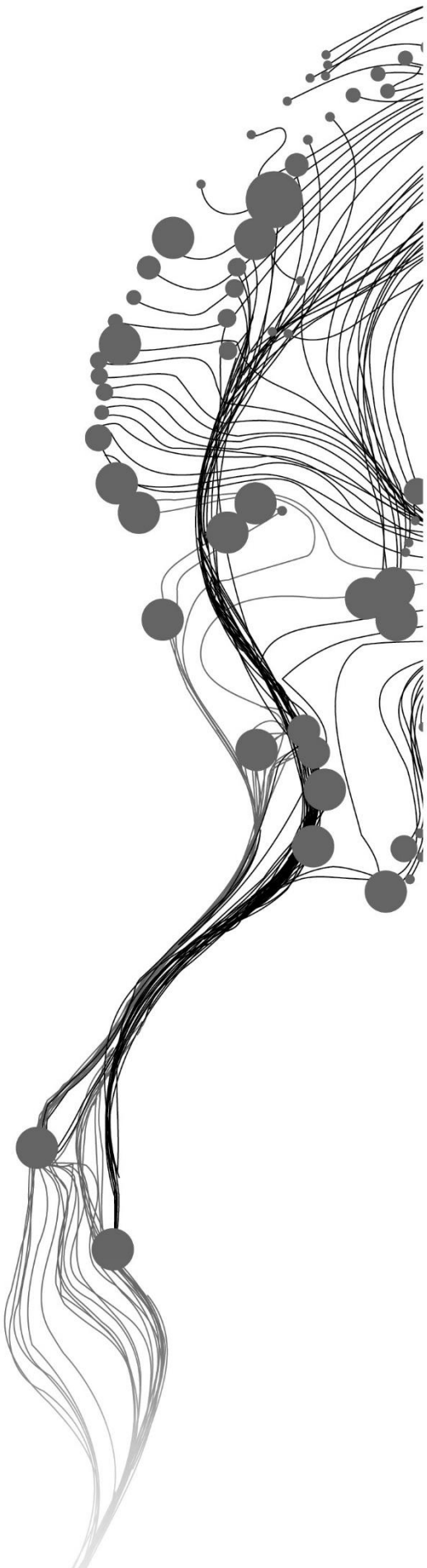
LINKED SPATIAL DATA: BEYOND THE LINKED OPEN DATA CLOUD

CHAIDIR ARSYAN ADLAN
February 2018

SUPERVISORS:
dr.ir. R.L.G. Lemmens
dr. E. Drakou



LINKED SPATIAL DATA: BEYOND THE LINKED OPEN DATA CLOUD

CHAIDIR ARSYAN ADLAN
Enschede, The Netherlands, February 2018

Thesis submitted to the Faculty of Geo-Information Science and Earth Observation of the University of Twente in partial fulfilment of the requirements for the degree of Master of Science in Geo-information Science and Earth Observation.
Specialization: Geoinformatics

SUPERVISORS:
dr.ir. R.L.G. Lemmens
dr. E. Drakou

THESIS ASSESSMENT BOARD:
prof. dr. M.J. Kraak (Chair)
dr. C. Stasch (External Examiner, 52°North Initiative for Geospatial Open Source Software GmbH)

ABSTRACT

The Linked Open Data Cloud (LOD Cloud) is the constellation of available interlinked open datasets which has become one of the biggest repositories on the web. An increasing number of spatial semantically annotated datasets provide a huge potential source of knowledge for data enrichment in a spatial context. Yet, there is lack of information about the structure of the spatial datasets in the LOD Cloud which can discourage the integration efforts. In addition, most of the existing studies of link discovery have yet to exploit spatial information richness (topology and geometry). Thus, a structured way to assess spatial datasets and to integrate linked spatial data is required.

This study aims to evaluate the LOD Cloud by assessing the data structure and the representation of linked spatial data, in order to support exploration and integration purposes. To achieve this objective, this study proposes: (i) a workflow for analyzing linked spatial data resources in the LOD Cloud, which consists of the identification of the linked spatial data sources, strategies for dataset retrieval, pipeline design for data processing, and linked data quality principles and metrics analysis; (ii) a review of linked data visualization systems, which includes an assessment of the current LOD Cloud Diagram based on expert opinion with respect to key requirements for visual representation and analytics for linked data consumption; and (iii) a workflow for linked spatial data integration. The main contribution of this thesis is the provision of case studies of integrating various spatial data sources. We presented two case studies, geometry-based integration using the spatial extension of Silk Link Discovery, and toponym-based integration using Similarity Measure. The datasets of *Basisregistratie Topografie* (BRT) Kadaster, Natura2000, and Geonames were used for the data integration.

The results of the study include: (i) a structured way to consume and extract spatial information from linked data resources. In this thesis, we proposed one metric to assess linked spatial data, namely the existence of geospatial ontology – vocabulary in the linked data resources; (ii) identification of suitable visualization element for exploration and discovery, especially for spatial data. The top-level relationship (overview) visualization is potentially facilitating an effective datasets discovery and also able to expose the spatial content and relationship in a sensible way. This study discovered that the linkset concept in the level of the dataset, subset, and distribution could be used as basis information for overview visualization; and finally, (iii) findings of spatial components (geometry and toponym) that can be used as important "hook" for integrating different datasets. The commonly used geospatial ontology and vocabulary also enable semantic interoperability to support data integration.

***Keywords**: Linked Spatial Data, Geospatial Ontology, Link Discovery, GeoSPARQL*

ACKNOWLEDGEMENTS

Foremost, I thank the Indonesia Endowment Fund for Education (LPDP) for providing me full funding to support my MSc in the Netherlands.

I would like to thank my first supervisor dr.ir. R.L.G. (Rob) Lemmens for their invaluable support and guidance throughout my thesis. My gratitude also goes to my second supervisor, dr. Evangelia (Valia) Drakou, for her patience to support me in the academics writing aspect. This thesis would not have been possible without their scientific knowledge and constructive advice. I would also appreciate my advisor Stanislav Ronzhin, M.Sc. who brought me to linked data world in the Netherlands.

I thank all my fellow Geoinformatics classmates. It has been a roller-coaster ride since the modules until thesis-making, especially thanks for Aldino Rizaldy, Ahmed El-Seicy, Noé Landaverde Cortes, and Joseph Frimpong who help me a lot during the study period.

Sincere thanks to all my Indonesian colleague, all of you have become my family in the Netherlands, thanks for sharing the joy, laugh, and food. Thanks especially to Dewi Ratna Sari, Rini Hartati, Arya Lahassa, and Aji Perdana who help me on proofread and give valuable advice to my thesis.

Lastly, nobody has been more important to me in pursuing MSc than my family. For my parents and my sisters, thanks for love and prayer. "*Kita adalah apa yang kita terus yakini"*

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

INDEX OF LISTING

1. INTRODUCTION

1.1. Motivation and Problem Statement

Accessing, retrieving, integrating, and sharing information are the important activities of exploiting the web as global information space (Konstantinou & Spanos, 2015). To yield robust Information Retrieval (IR), two main issues should be taken into account; first, to provide the meaning of content, and second, to integrate the information. The common IR methods which are based on keyword-based searching are insufficient to capture the conceptualization related to content meaning (Fernández et al., 2011). Concerning the first issue, keyword-based IR suffers from ability to extract the meaning from literal string content of web pages resources. Concerning the second issue, web pages merely rely on hyperlinks whose functionality does not fulfil the intended goal of information integration (Bizer, Heath, & Berners-Lee, 2009). Meanwhile, utilizing the web as a tool for information integration, searching, and querying is mentioned as the biggest challenge in the study area of intelligent information management (Ngomo, Auer, Lehmann, & Zaveri, 2014). To cover both issues, the implementation of a concept that can provide: 1) searching by meaning, and 2) easy integration mechanism, data on the web is required.

To that extent, the Semantic Web is designed to structure data on the web in order to generate insight, value, and meaning of the data (Heath & Bizer, 2011). The semantic web allows the annotation of contextual meaning to the data so that it can be easily understood and searched. However, the semantic web can only be established if the data follow a standard structure so that data from various sources can be integrated in order to generate a new knowledge. Hence, the development of methods for data structuring is needed to solve the data integration problem. To overcome this problem, Linked Data principles introduce a standardization method of structuring, publishing and linking data on the web in machine-readable format (Becker & Furness, 2010). The essential element of linked data is structured data regarding the standard for data representation, identification, and retrieval (Bizer, 2009). This standard data structure allows establishment of semantic link between data. By providing meaningful links to related information from different data sources, linked data offers an endless discovery of information on the web (Hart & Dolbear, 2013). Although this functionality is approaching the ideal IR, link establishment between different data sources still remains a challenge.

An increasing number of semantically annotated datasets on the web led the World Wide Web Consortium (W3C) to organize an initiative called Linking Open Data Community Project (Konstantinou & Spanos, 2015). The goal of this initiative is to present different data sources on the web as Resource Description Framework (RDF) and to create a linkage among them (W3C SWEO, 2017). This initiative encourages the communities as data owner to enrich their data by integrating them to existing data on the LOD Cloud (see Section 2.1). The encouragement is aligned with Berners-Lee (2009), who asserted that the five-star quality data can be achieved by data integration. The data integration has a purpose to enrich data through Semantic Web (Stadler, Lehmann, Höffner, & Auer, 2012), and its aim is defined as *"linking data across the web using controlled semantics"* (Kuhn, Kauppinen, & Janowicz, 2014). Currently, the LOD Cloud contains 1146 datasets and 150 billions of triples (Ermilov, Lehmann, Martin, & Auer, 2016). Undoubtedly, it provides a huge potential source of knowledge for communities to enrich their datasets. Nevertheless, there is lack of information about the datasets structure of the LOD Cloud (Arturo et al., 2016) which can discourage the integration efforts. According to a study from Assaf, Troncy, & Senart (2015), some datasets are deteriorated

which are indicated by the low quality of metadata. Furthermore, the following study from Assaf, Senellart-Telecom ParisTech, Stefan Dietze, and Troncy (2015) stated that most of the datasets have problems with bad quality of access information and poor maintainability. These kinds of problems could potentially hinder the integration effort. Considering these conditions, information about dataset in the LOD Cloud is desirable. To this end, the state of the LOD Cloud is needed to improve the understanding of the communities about the structures and inconsistencies of the datasets. Thus, it is important to assess the data structure and representation and to understand potential use of LOD Cloud interface to support the exploration and integration purposes.

Spatial data integration on the web which covers discoverability and linkability issues remain a challenge (Knibbe, 2016). These issues becomes important because 21 of 1091 datasets in LOD Cloud are spatial datasets (Schmachtenberg, Bizer, & Paulheim, 2014) and still growing until now. This fact making it worth to study how spatial data can be integrated to an interoperability environment on the web. One main problem of data integration mentioned by Smeros & Koubarakis (2016) is that most of the existing studies of link discovery were not exploiting spatial information richness (topology and geometry). Link discovery is activity to discover the existence of relevant datasets and resources in the LOD Cloud. Spatial-enabled link discovery issue is mentioned by the Open Geospatial Consortium (OGC) & the W3C Join Working Group as the key problem that is yet to be solved (W3C & OGC, 2017). It becomes more important since the amount of linked spatial data is getting larger. The development of GeoSPARQL (Geographic Query Language for RDF Data) has facilitated link discovery based on spatial relationships. However, a complex query to discover spatial relationships among heterogeneous data is not suitable for real-time purposes due to the high computation time (Smeros & Koubarakis 2016). As a consequence, the link materialization between resources is needed (Smeros, 2014). Taking all these needs and issues into account, this study focuses on how to design workflow of spatial link discovery and spatial data integration to the LOD Cloud.

1.2. Research Identification

This research is divided into three major tasks: to evaluate the current state of LOD Cloud, to determine the potential usage of the LOD Cloud Diagram (http://lod-cloud.net/), and to develop workflow for spatial data integration into LOD Cloud. The first task deals with analysing datasets of the LOD Cloud which metadata sits at http://datahub.io. This study focuses on the assessment of the resource level since the links between data in the LOD Cloud only exist at resource level, not in the set level. The LOD Cloud has various dataset domains, one of them composed by spatial datasets which are categorized as geography domain. This research is specifically targeting at spatial datasets to be examined. The characterization of the spatial datasets in the LOD Cloud done by assessing the resources and links using a designed workflow of data processing. Since this study focuses on data integration, the analysis will only be implemented on links quality. Links quality refers to the level of integration that represents the coherence of two linked data resources. The outcome of this task is the profile of spatial datasets in the LOD Cloud.

The second activity focuses on assessing LOD Cloud Diagram. This assessment is conducted to explore the potential usage of the LOD Cloud Diagram from a user perspective. To get this information, an expert opinion is conducted to measure the extent to which the LOD Cloud interface can be operationalized. The outcome of this task is the identification of potential use and user requirements. The third activity focuses on the identification of linked spatial data integration procedures based on a review of standards, guidelines, studies, and tools. The identified methodology will be implemented on study cases datasets. This final activity consists of finding relevant datasets in the LOD Cloud, discovering potential link, and establishing

the links. The goal is to explore possibilities and limitations of integrating spatial data. The outcomes of the third activity are: 1) the workflow of linked spatal data integration and 2) the linked spatial data which will be integrated into LOD Cloud.

1.2.1. Research Objectives

The main objective of this research is to evaluate the LOD Cloud by assessing the data structure and the representation of linked spatial data, in order to support exploration and integration purposes. To achieve this main objective, three sub-objectives are set:

1. To evaluate the current state of spatial datasets in the LOD Cloud.
2. To determine the potential use of the LOD Cloud for the exploration and integration of spatial data.
3. To determine the conditions, and to design a workflow, for adding and maintaining datasets in the LOD Cloud.

1.2.2. Research Questions

1. To evaluate the current state of spatial datasets in the LOD Cloud.
   a. What are the elements that can be used to characterize linked data in the LOD Cloud?
   b. What are the principles of linked data quality frameworks?
   c. What are the dimensions and metrics of linked data quality frameworks that can measure the quality of links?
   d. How to use the evaluation result to find the potential links between datasets in the LOD Cloud?

2. To determine the potential use of the LOD Cloud for the exploration and integration of spatial data.
   a. What kind of activities can be supported by LOD Cloud?
   b. How should linked spatial data be represented in the LOD Cloud in order to support the potential use?
   c. What are the options to represent spatial relations?
   d. How can the LOD Cloud user interface be improved for exploration and integration purposes?

3. To determine the conditions, and to design a workflow, for adding and maintaining datasets in the LOD cloud.
   a. To what extent standards can be used for representing spatial data in a linked data format?
   b. How can a dataset be added to the LOD Cloud? What are the restrictions?
   c. How to use relevant GeoSPARQL queries to discover potential links among LOD Cloud datasets? To what resources the link should be established?

1.3. Innovation

Introducing assessment metric of spatial dataset in LOD cloud is the novelty of this research. The latest study on the state of LOD Cloud did not provide sufficient information for supporting the exploration and integration purposes, especially for spatial data as it only provided general statistics of datasets and aggregated the information based on dataset domain (Schmachtenberg, Bizer, & Paulheim, 2014). To fill this gap, this research proposes the provision of detailed information per data provider or pay-load domain (PLD). This research focuses on examining how the assessment of the LOD Cloud data structure and visualization can assist the exploration and integration purposes. This study provides an analysis on LOD Cloud Diagram, to better accommodate the potential usage. The output of this study also represents the innovation: the workflow for linked spatial data integration. The main contribution of this thesis is the study cases provision of integrating various spatial data sources. By integrating spatial data, this study contributes to systematically build richer relationships among resources using proper spatial vocabularies which will go beyond the *SameAs* relation.

1.4. Research Methodology

This study consists of three major sections based on the sub-objectives. The first is to evaluate the current state of spatial datasets in the LOD Cloud. This objective is explained by analyzing the linked spatial data in the LOD Cloud as discussed in Chapter 2. It includes the identification of the linked spatial data sources, strategies for dataset retrieval, pipeline design for data processing, and anaysis of linked data quality principles and metrics. The second is to determine the potential use of the LOD Cloud Diagram. This is discussed in Chapter 3, which includes literature review of linked data visualization, identification of suitable visualization for dataset and linkset exploration and discovery, especially for spatial data. The aim of this chapter is to analyse how well the LOD Cloud represents spatial datasets and the links between them in order to support exploration and integration purposes. Finally, the third objective, discussed in Chapter 4, is to determine the conditions and to design a workflow for adding and maintaining datasets in the LOD Cloud. Chapter 4 also provides an analysis of the standard for spatial data on the web and workflow design of spatial data integration to LOD Cloud. Figure 1-1 depicts the work phases of the research based on sub-objectives.

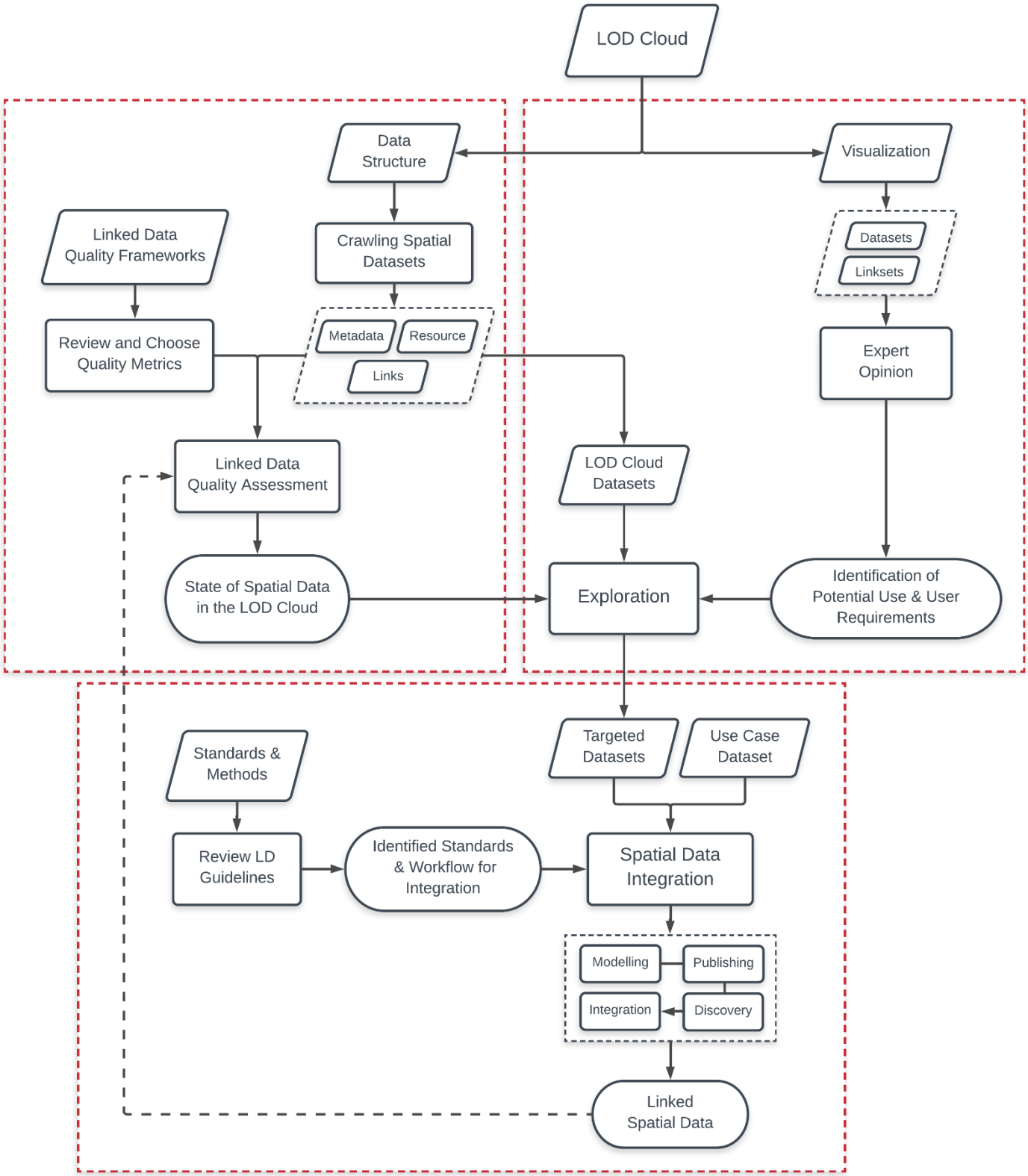


Figure 1-1. Flowchart of methodology

2. ANALYZING LINKED SPATIAL DATA IN THE LINKED OPEN DATA CLOUD

In this section, the state of linked spatial data is described and investigated using the workflow of linked data analysis. Section 2.1 explains the data architecture of published linked data in the data catalogue and explains how to deal with data retrieval with respect to a certain data architecture. Subsequently, Section 2.2 and 2.3 elaborate on linked data quality assessment that focuses on the link quality. Section 2.4 and 2.5 focus to answer the question of "*what makes linked data linked spatial data?*". The discussion includes the combination of geospatial ontologies and vocabularies with linked data. Finally, Section 2.6 presents the design of a data analysis workflow to investigate and assess the linked spatial data.

2.1. Linked Data in the LOD Cloud

The open data movement advocates the idea that data should be open and freely available for public to be reused and republished under Open License. The growth of semantically annotated open data leads to a continuation initiative of open data movement called the Linking Open Data. The initiative is started by SWEO community project from W3C which aims to build a data common by interlinking open data (set) on the web. LOD Cloud, or Linked Open Data Cloud, is the constellation of available interlinked datasets on the web which has become one of the biggest repositories of interlinked data on the web (Assaf, Troncy, & Senart, 2015). As mentioned in the Linked Data principles (Berners-Lee, 2006), the value of a data will increase when it is re-used and interlinked to another source. Therefore, linked data publication is one of the most important phase to allow the public to discover the datasets on the web and interlink them. There are three options to publish linked data (Rietveld, 2016). First, hosting a serialized RDF dump file in web-server. Second, using Internationalized Resource Identifier (IRIs) to denote unique resources and allow public to retrieve (or dereference) the resource via HTTP GET request. Third, providing a SPARQL endpoint to query specific resources. At least one of these three access information of linked data should be advertised in the data catalogue.

One of the foremost data catalogues is datahub.io (see Section 2.1.1) that is supported by Comprehensive Knowledge Archive Network (CKAN) from Open Knowledge International (see Section 2.1.1). This data catalogue provides a rich repository of metadata that can be used for further steps in linked data life-cycle. Amongst many available data catalogues by CKAN, this research only considers datahub.io as a source for the data collection because it contains cross-domain datasets from multiple organizations around the world. Hence, it gives abundant information to get insight on the current condition of linked data implementation. The data catalogue provides both a sensible way to discover the dataset and access information to the published linked data resource.

2.1.1. CKAN Dataset Model

Datahub.io is only one of many CKAN data portal implementations. CKAN also supports open data portal platform such catalog.data.gov and data.gov.uk. CKAN as Data Management System (DMS) define their own data model to present the data in the platform. Data model in a data catalogue refers to metadata model, this information includes a set of entities of datasets metadata. Metadata in datahub.io adopts CKAN Domain Model as their data model (see Figure 2-1). CKAN Domain Model consists of several elements of CKAN object or so-called entity, i.e.: datasets, resource, group, dataset relationship, tag, vocabulary, etc. Assaf et al., (2015) classified metadata model information into eight main types of metadata information,

i.e.: (1) General Information, (2) Access Information, (3) Ownership Information, (4) Provenance Information, (5) Geospatial Information, (6) Temporal Information, (7) Statistical Information, and (8) Quality Information. The eight main types of metadata information are classified into CKAN entity.

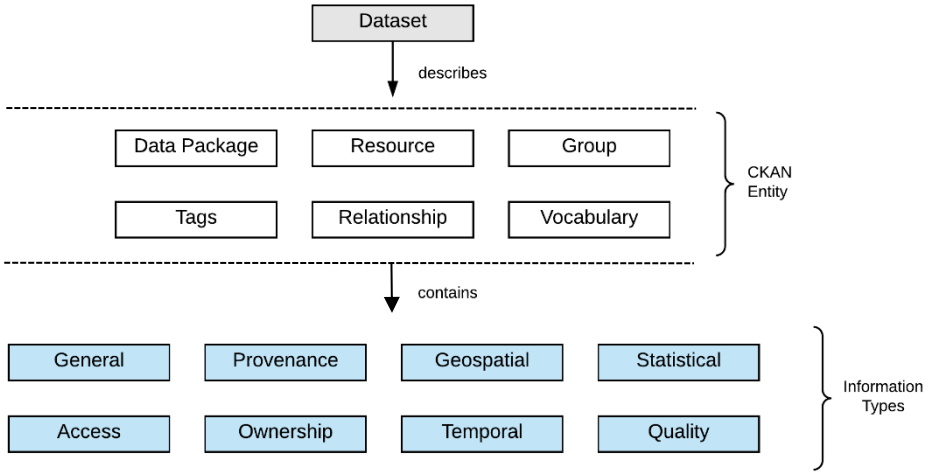


Figure 2-1. CKAN Domain Model

The CKAN entity is highly significant for tech savvy users to search and discover datasets. The CKAN entity can be used as an argument to retrieve the information by passing the API requests. For layman-users, the interface of datahub.io can facilitates a query and filter to search the datasets, for instance string-based query, tags filter, format filter, etc. The dataset searching leads to the dataset page (see Figure 2-2), which contains two main elements: data package and resources. The data package element contains core metadata information of dataset (CKAN entity), for instance license, tags, relationship, etc. While the resources element contains a set of extended dataset attributes, such as URL of resources (RDF dump, SPARQL endpoint, example RDF resources), mime type & format, timeliness, etc. Therefore, users can choose a convenient way to discover and retrieve the dataset catalogue by using either API or user interface.

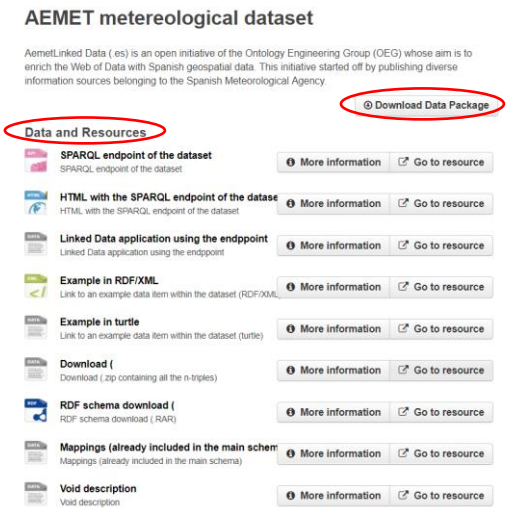


Figure 2-2. Example of datasets page

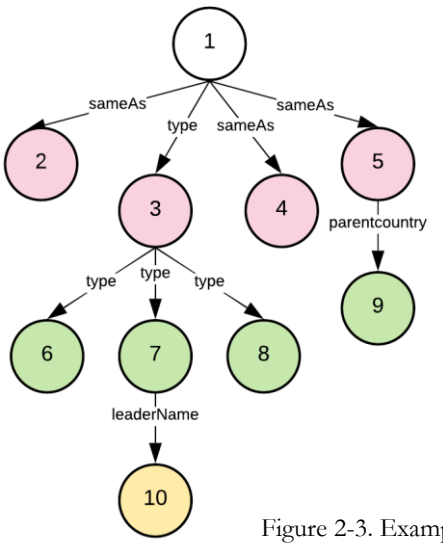
2.1.2. Approach for Strategies of LD Dataset Retrieval in the LOD Cloud

One fact that makes a LD dataset in the LOD Cloud hard to be found and retrieved is the nature of Linked Data itself. As asserted by Rietveld (2016), Linked Data is distributed and not centralized. CKAN - datahub.io is only a data catalogue (data portal) that stores metadata of LD datasets. The original resources are hosted in each data provider repository. Until now, the CKAN API does not have the capability to harvest in scalable way the resources of LD datasets that are listed in datahub.io catalogue. Therefore, determining various strategies to retrieve or crawl the actual resources of LD datasets is required. We assume in this experiment LD datasets discovery and retrieval activities start from datahub.io without prior knowledge of any available LD datasets on the web. The expected result of the retrieval activity is the LD resources, in various RDF serializations.

Before discovering datahub.io, an effort called LOD Cloud Diagram (http://lod-cloud.net/) visualizes the constellation of available LD datasets on the web for which metadata is hosted in datahub.io. This interactive visualization could be a very useful entry point to discover LD datasets. It categorizes the content of LD datasets based on tags of datahub.io. As the metadata is hosted in datahub.io, this visualization also sets a hyperlink of each node to the dataset page in datahub.io (see Figure 2-2). Afterwards, from datahub.io dataset page, the following strategies can be applied to retrieve LD resources:

A. Semantic Crawler

The first approach is to use semantic crawler, in this case using LD Spider (Isele, Umbrich, Bizer, & Harth, 2010). LD Spider has an ability to crawl the resources using RDF link between resources. The crawling activities starts with one seed IRI and the LD Spider will follow and crawl the deferenceable IRIs through RDF link. The shortcoming of this approach is not all linked data resources are published in deferenceable format via HTTP GET request. Therefore, only a limited number of LD resources can be crawled using this approach. The advantages using LD Spider is the crawling activity does not restrict to certain URL's domain, hence it possible to crawl LD resources more than one data provider. LD Spider is a dedicated linked data crawler that has special parameters in managing crawling strategies. There are two crawling strategies, breadth-first, and depth-first. The expected result of this semantic crawler is an RDF file containing a set of triples that begin from a seed IRI. Seed IRI can be obtained from example RDF in dataset page.



1. http://dbpedia.org/resource/Enschede
2. http://linkedgeodata.org/triplify/node45753522
3. http://dbpedia.org/resource/List_of_municipalities_of_NL
4. http://yago-knowledge.org/resource/Enschede
5. http://sws.geonames.org/2756071/
6. http://dbpedia.org/resource/Rotterdam
7. http://dbpedia.org/resource/Almelo
8. http://dbpedia.org/resource/Alkmaar
9. http://sws.geonames.org/Netherlands
10. http://dbpedia.org/resource/Jon_Hermans-Vloedbeld

Figure 2-3. Example of breadth-first crawling strategies, numbering indicate order of crawling and IRIs

Figure 2-3 illustrate the example of crawling process using LD Spider, we aimed to retrieve LD resources start from a seed IRI of Enschede in DBPedia dataset (http://dbpedia.org/resource/Enschede). Breadth-first strategy was implemented to get certain number of LD resources that has RDF link to Enschede resource. For this experiment we limited to 15 resources. LD Spider crawled resources based on breadth-first search algorithm which made an order in extending graph. The numbering inside the nodes indicate the order of crawling and also refers to IRIs. The node's colour indicates the equal deep level from seed IRI. It started at the tree root and extended the graph through neighbour nodes (pink nodes), if the neighbour nodes are completed explored then it extended to the next deep levels of neighbour nodes (green and yellow nodes).

B. RDF Dump

The simplest approach to retrieve the LD resources is to download the whole RDF file of a datasets using resource access information from data catalogue. A scalable application is developed by others that can fetch the access information of metadata element from the Vocabulary of Interlinked Datasets - VoID (`void:dataDump`) and Data Catalog Vocabulary - DCAT (`dcat:downloadURL`). Explanation of VoID and DCAT are provided in Section 2.1.4. This approach is suitable for users who are interested in inspecting the whole LD resources within a dataset. This approach also has a number of shortcomings. First, the data dump tends to be outdated because the updating process of data dump is rather infrequent. In addition, data dump is separated with triple store where the live updating resource committed. Second, the client cost is rather high, since downloading data dump requires high bandwidth. The third is LD resources validity, in which the majority of data dump does not follow the standard. Most cases involve wrong syntax and incorrect serialization. This makes the data unable to proceed to RDF parser for further purposes. These problems commonly require manual data cleaning from users.

C. CKAN – API Extension of SPARQL

Recent development of CKAN plugin has extended into linked data. The effort begins with CKAN DCAT plugins that can retrieve metadata catalogue in RDF serialization. This is done by mapping the CKAN dataset model to DCAT model. However, this effort is still unable to meet the needs of crawling the LD resources via data catalogue because it only retrieves the metadata, and not the actual LD resources (triples). The effort is continued by ODW Project (Lee, Chuang, & Huang, 2016) that aimed to integrate SPARQL endpoint with data catalogue. This project has upgraded the previous CKAN DCAT plugins by extending the harvesting mechanism and the RDF Profile. The ODW Project is still in its early stages of development and is only implemented in a prototype data portal (http://data.odw.tw). The core idea is that LD resources can be transformed to CKAN instances and can be queried by SPARQL endpoint capabilities of Openlink Virtuoso (see Figure 2-4). The project was proposed to provide an alternative way to retrieve datasets by providing native SPARQL queries in CKAN which is independent from a triple store. However, this development contradicts with the storage in triple stores that is recommended as de-facto linked data lifecycle.

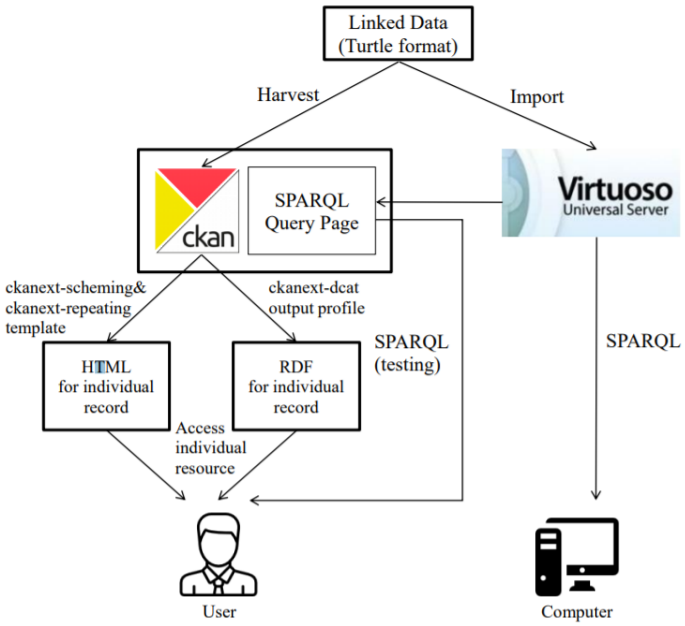


Figure 2-4. System architecture of CKAN-SPARQL extension

2.1.3. LOD Cloud Diagram and LD Dataset Domain

As briefly mentioned in the previous subsection, we involved LOD Cloud Diagram as an entry point for dataset discovery and retrieval. This diagram is considered as the most-up-to-date visualization of available LD datasets that implemented linked data principles (last updated August 22nd, 2017) and widely-used dataset domain categorization (Abele, Mccrae, & Buitelaar, 2017). The diagram was created based on dataset metadata which was curated by contributors in the datahub.io. All datasets were added to the diagram if it established or materialized link to the LOD Cloud Diagram datasets. In this diagram, we might not find important linked data provider such as Ordnance Survey UK or Kadaster Netherlands. Even though these datasets are advanced regarding the application and quality of LD resources, but their resources are not referring to LOD Cloud diagram resource. Therefore, these datasets were not included yet in the diagram.

To categorize the datasets in datahub.io, the CKAN entity of tags (see Figure 2-1) are used as the attribute. The datahub.io tag are crowdsourced or curated by the contributors, thus one dataset might have more than one tag depending of the data owner's judgment on the resources content of the dataset. This tag heterogeneity led to the creation of datasets domain categorization in the LOD Cloud 2017 version by Abele, Buitelaar, Mccrae, & Bordea (2016). The categories are determined using datahub.io tags as features for Support Vector Machine Classifier. This classification used the 2014 version of LOD Cloud domain categorization as training data. The result of datasets domain categorization presents in Figure 2-5.

Besides datahub.io tags, there are other elements that can be used as attribute to define dataset domain. **The first is the VoID** (Vocabulary of Interlinked Datasets), there is one VoID vocabulary that can be used for categorizing datasets by subject which is `dcterms:subject`. It can be used to denote datasets topic or subject. However, VoID is suffering from low existence, and those which have, the `dcterms:subject` property usually does not exist. The VoID will be elaborated in Section 2.1.4. **Second, CKAN entity of Vocabularies** (see Figure 2-1), this entity grouped related tags into one vocabulary to facilitate the high variation of datasets content. Based on observation, this element is not accurate. Considering the limitations of these two pieces of information, datahub.io tags still give the most reliable option of dataset domain

categorization. Furthermore, to concur with Abele et al. (2016), crowdsourced tag attributes give more accurate information compare to other elements. Considering these facts, CKAN entity of tags will be used in dataset discovery (see Section 2.4)

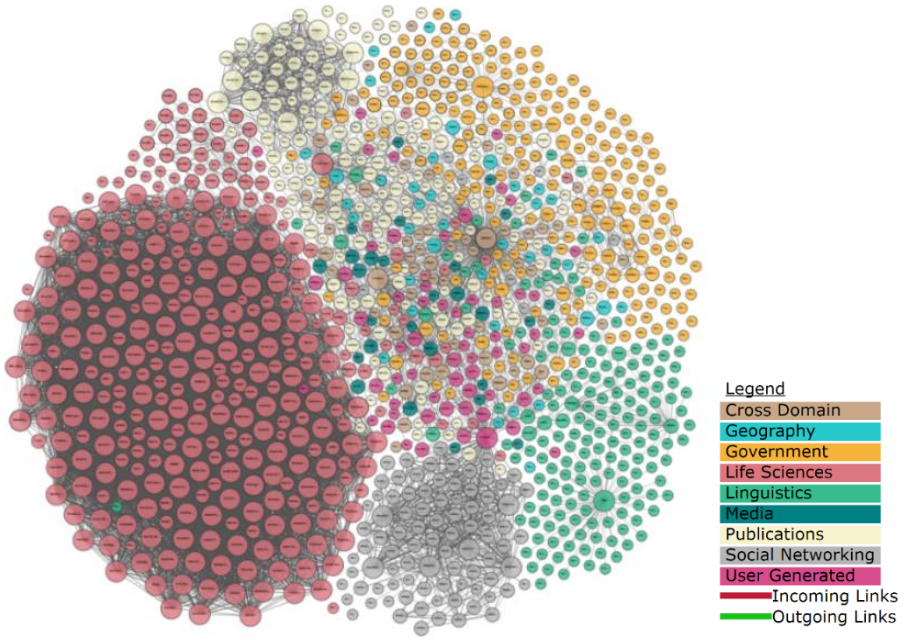


Figure 2-5. Linking Open Data Cloud diagram 2017 (cited from Andrejs Abele, John P. McCrae, Paul Buitelaar, Anja Jentzsch and Richard Cyganiak (2017) http://lod-cloud.net/)

2.1.4. Metadata of Linked Data

Linked data in LOD Cloud has two main metadata, namely dataset metadata and catalogue metadata. The dataset metadata is standardized in the form of a VoID vocabulary which is categorized into four types: general, structural, access and linkset descriptions (W3C, 2011). Catalogue metadata is standardized in the form of DCAT vocabulary to facilitate interoperability between data catalogues on the web (W3C, 2014). Between these two-metadata, the VoID is more suitable to characterize the linked data because it has sufficient vocabulary to represent the summary of linked data resources within one dataset. The VoID components that are able to describe linked data in the set level include: 1) Basic information related to categories of data, 2) Vocabulary usage, 3) Basic statistics about the dataset, 4) The external dataset that linked, and 5) Linkset description. The following information shows the use of VoID vocabulary to describe linked data in set level.

1. Basic information related to categories of data

```
Geonames a void: Dataset;
dcterms: subject <http://dbpedia.org/resource/Location>
```

http://dbpedia.org/resource/Location state the subject or category of the data. This resource describes the dataset and classifies it into the general category, for instance computer science, books, location etc.

2. Vocabulary usage

```
void: vocabulary <http://www.w3.org/2003/01/geo/wgs84_pos#>;
void: vocabulary <http://purl.org/uF/hCard/terms/>;
void: vocabulary <http://www.w3.org/2006/vcard/ns#>;
void: vocabulary <http://www.geonames.org/ontology#>;
void: vocabulary <http://purl.org/dc/terms/>;
void: vocabulary <http://lobid.org/vocab/lobid#>;
void: vocabulary <http://www.w3.org/2001/XMLSchema#>;
void: vocabulary <http://purl.org/ontology/mo/>;
```

This list is very useful for identifying certain data category. For instance, spatial data usually described by vocabulary W3C Basic Geo, NeoGeo, etc. The examples of vocabulary listed above indicate that this dataset has spatial data as it uses vocabulary of http://www.w3.org/2003/01/geo/wgs84_pos# .

3. Basic statistics about the dataset

Several vocabularies could explain basic statistics of the dataset, for instance:

- `void:triples` = the total number of triples in a dataset
- `void:entities` = number of URI in the dataset
- `void:properties` = number of distinct properties in the dataset

4. The external dataset that linked

The vocabularies of `void:Linkset`, `void:subjectsTarget`, and `void:objectsTarget` can explain the involvement of external resources in dataset. The `void:Linkset` explains the existence of relation or link between two datasets. The `void:subjectsTarget` indicates which dataset provides the subject of the triples and `void:objectsTarget` indicates which dataset provides the object of the triples.

5. Linkset description

The `void:linkPredicate` and `void:triples` completes link information between two datasets. This snippet of metadata explains that there is a linkset between FAO Geopolitical Ontology and Geonames datasets, where FAO as a subject and Geonames as an object. The predicate of this linkset is explained by `owl:sameAs` of 2000 triples.

```
FAO Geopolitical Ontology_Geonames a void:Linkset;
   void:subjectsTarget: Geopolitical Ontology;
   void:objectsTarget: Geonames;
   void:linkPredicate owl: sameAs;
   void:triples 2000;
```

2.2. Linked Data Quality Framework

Quality Domains and Metrics

Several studies related to the linked data quality assessment have been conducted. These studies proposed a variety of dimensions, elements, and metrics to assess the quality of dataset, resource, and links. Hogan et al. (2012) carried out an empirical survey to assess linked data quality based on conformance with respects to linked data guidelines. This study covered four domain issues, i.e., naming resources, link, data, and deference. These domains are divided into 14 metrics that generate a comprehensive report of quality metrics. In other research, Assaf et al. (2015) structured the objective of linked data quality indicators which are based on four quality categories, namely entity, dataset, semantic model, and linking process. These categories characterize 10 identified quality attributes: completeness, availability, licensing, freshness, correctness, comprehensibility, provenance, coherence, consistency, and security. Quality attributes are divided into 64 concrete quality indicator metrics which cover the quality indicator assessment of Comprehensive Knowledge Archive Network (CKAN) based model.

Another study by Zaveri et al. (2014) reviewed quality assessment of linked data and summarized it based on four domains of assessment. This study covers accessibility, intrinsic, contextual, and representational issues. From a systematic review, this author extracted 18 data quality dimensions, gave a clear definition of it and divided it into 69 quality metrics. Furthermore, their study distinguished quantitative and qualitative measured metrics. This bottom-up framework gave a clear understanding of quality assessment based on the different dimensions and metrics. All these studies demonstrated a wide-range of linked data quality assessment. Nevertheless, these three studies assigned the metrics that are related to link quality in a different hierarchy and did not categorize them in a single domain.

Provision of Statistics Information

Schmachtenberg, Bizer, and Paulheim (2014) demonstrated insight of the development of linked data in LOD Cloud. This study generated basic statistics about resource, metadata, and links. These statistics report is aggregated into the topical domain. The author related the result with best practices in various domains. Regarding the quality metrics, it only presents 11 metrics. Additionally, Auer, Demter, Martin, and Lehmann (2012) and Auer et al. (2012) developed profiling tools to present statistics of datasets. The analysis is based on 32 different statistics analytical criteria which are aggregated into four domains: quality analysis, coverage analysis, privacy analysis, and link target identification. These analytical criteria statistics also cover the statistical criteria of The Statistical Core Vocabulary (SCOVO), which is defined by the Vocabulary of Interlinked Datasets (VoID), for instance, property and vocabulary usage. However, these two studies cannot be considered as a linked data quality assessment framework as they only aimed to present dataset profile and description of statistical dataset characteristics.

Linkset Quality

Several studies have also been conducted on the topics of profiling and linkset quality. Arturo et al. (2016) aimed to define LOD Cloud datasets clusterization based on its metadata. It created dataset profiles by relating label of datasets to the ontology of Wikipedia. This study also examined the extracted linkset and assessed its cross-domain linkage using three chosen algorithms, i.e. Edge Betweenness Methods (EBM), Greedy Clique Expansion (CGE), and Community Overlap Propagation Algorithm (COPRA). The result of the study is the candidate of the targeted datasets are interlinked based on domain similarity or clusterization. On a side note, other studies focused more on the assessment of the linkset quality. Ruckhaus & Vidal (2012) used a bayesian network to assess the incompleteness of the links and ambiguities of label between links. This study mainly used the occurrence of linkset among datasets. It employed five metrics as an approach to assess link quality.

Another study provided contribution of proposing linkset quality measurement (Albertoni & Gómez Pérez, 2013). The measurement used three dimensions of quality: quality indicator, scoring function, and aggregate metrics. This study aimed to relate the measurement of link quality with dataset integration issue, in which publisher can use it to improve the quality of linkset. The last related work which covered linkset quality was done by Guéret, Groth, Stadler, & Lehmann (2012). It used five metrics: SameAs chain, description richness, in and outdegree, centrality, and clustering coefficient. These five metrics are summarized to define good and the bad links. Furthermore, Assaf and Senart (2012) proposed data quality principle of semantic web which adopts Linked Open Data guidelines. The authors mentioned the quality of linking principle which covers connectedness, isomorphism, and directionality. This principle is only one of the five principles, while the other four principles comprise of the quality of data source, raw data, semantic conversion and global quality. These five principles are divided into 20 attributes of assessment.

2.3. Domain and Metrics Assessment

2.3.1. Domain Assessment

Based on the examination of linked data principle, there are four key issues that should be achieved which are 1) Assign correct URIs to identify entities, 2) Use HTTP URIs to make data in machine-readable format, 3) Use RDF standard, and 4) Link to external data. Until now, there is no formal metrics to assess the quality of linked data, because the quality is defined as fitness for use. As discussed in Section 2.2, several linked data quality frameworks developed metrices to be used to assess the linked data quality principle, each of which are developed based on those four key issues.

Six studies related to linked data quality have been examined, each of which has different ways to structure the linked data quality elements, from abstract concepts to the measurement assessment. Table 2-1 shows the comparison of hierarchies of linked data quality element based on six chosen studies. The first hierarchy of linked data quality elements is the broader concept of data quality assessment. It groups the specific elements into general categories. The second hierarchy is the dimension which explains narrower concept of linked data quality. Basically, this dimension groups metrics elements that can be used to measure qualities which are relevant to user criteria (Ngomo et al., 2014). The third hierarchy has a function to operationalize these linked data quality dimensions. Metrics and indicators are the procedure for measuring and assessing a data quality principle.

Table 2-1. Comparison of linked data quality elements hierarchies based on six chosen studies

Studies	Linked Data Quality Elements		
	1st hierarchy (Concept)	**2nd hierarchy (Dimension)**	**3rd hierarchy (Metrics)**
Assaf & Senart (2012)	Data Quality Principles (4)	Attribute (20)	-
Guéret, Groth, Stadler, & Lehmann (2012),	-	-	Metrics (5)
Hogan et al. (2012)	-	Attribute (4)	Metrics (14)
Albertoni & Gómez Pérez (2013).	-	Quality Measure (3)	Metrics (6)
Zaveri et al. (2014)	Dimension (4)	Dimensions (18)	Metrics (69)
Assaf et al. (2015)	Quality Category (4)	Quality Attribute (10)	Quality Indicator (64)

Based on review on six chosen studies, Zaveri et al. (2014) elaborate on the linked data quality data elements comprehensively. The Table 2-2 shows the list of the second hierarchy (dimension) of linked data quality elements from this study.

Table 2-2. Linked Data Quality Dimensions (Zaveri et al. 2014):

Availability	Licensing	Interlinking	Security
Consistency	Conciseness	Completeness	Versatility
Relevancy	Timeliness	Trustworthiness	Understandability
Performance	Interoperability	Interpretability	
Syntactic Validity	Semantic Accuracy	Representational Conciseness	

From user's perspective, data quality information is essential information to support exploration in order to choose right dataset based on their application. Thus, this linked data quality elements are needed to be described to assist users in identifying the quality of data. In this case, the Dataset Quality Ontology (daQ) can be used to describe the linked data quality (Debattista, Lange, & Auer, 2014). The provide vocabularies to describe quality in category (concept), dimension, and metrics. The daQ is adopted and extended by W3C as Data Quality Vocabulary (DQV) for Linked Data (W3C, 2016b). DQV is not only provide vocabulary to express the quality of dataset, but also expression statement about the quality of metadata. Furthermore, Linked Data Quality Model (Radulovic, Mihindukulasooriya, García-Castro, & Gómez-Pérez, 2018) extended the W3C DQV to describe the particular linked data quality elements which are not covered yet by the existing ontology.

There are several implementation tools of linked data quality assessment, i.e.: 1) Luzzu (Debattista, Auer, & Lange, 2016). This tool assesses 22 metrics from nine different linked data quality dimensions. The current version of Luzzu provides the result still in the daQ vocabulary but targeted to serialize the assessment result in DQV (W3C, 2016c). 2) LD Sniffer (Mihindukulasooriya, García-Castro, & Gómez-Pérez, 2017). The current version of this tool provides assessment of accessibility metrics of Linked Data Quality Model. The development is in progress to extend in order to cover other metrics. Linked data quality report from both tools will be described DQV as additional vocabulary in the DCAT to describe data quality information.

2.3.2. Metrics Assessment for Link Quality

This thesis specifically aimed to elaborate on how data quality metrics deal with link elements in the LD resources. Based on the literature review, we identified several metrics that can quantitatively assess the link quality. These metrics comes not only from interlinking dimension but also completeness dimension. These findings prove the importance of the literature review, to study thoroughly every linked data quality elements and find the relevant metrics that can assess specific goals. **The first group of metrics relate to a concept that assumes linked data as the web of data** (Guéret et al., 2012). Based on this assumption, the network-measure based concept can be used for assessing link quality. This assessment is based on the network topology of LD resources. The network topology refers to connected nodes by the edges. In linked data, directed graph uses as a conceptualization of one-way relationship between two nodes. The network topology of LD resources will be tested with metrics such as the **link degree, clustering coefficient, and centrality**. Since we interested on information per PLD, thus it only considers local network of LD resources to be assessed. This means the metrics are only implemented to each dataset instead of cross-datasets.

The link degree metric refers to the number of links on a network. The links include the total number of outgoing and incoming links. To assess link degree, we can refer to the number of predicates vocabulary on each triple. Clustering coefficient metric indicates the comparison between the number of links from one node to the direct neighbourhood node and the number of potential links that may exist. The value of the clustering coefficient will be different for each resource on a network. Lastly, the centrality metric is ratio between incoming and outgoing link to a specific resource.

The second group of metrics relate to the completeness of linkset, which can be assessed based on two metrics. The first is **interlinking completeness**, which represents the ratio between number of resources to a dataset whose link is already established and the total amount of resources in a dataset. The second is the **complementation** of two datasets using a linkset. We assume that a link can enrich information from one dataset to another dataset. The linkset has a complementary role that provides new information to a resource. This concept can be explained in detail by examining the two functions, linkset coverage and linkset completeness. These two functions are examined based on the application of ontology alignment to resource vocabulary between two datasets.

The third group of the metrics relate to the content of LD resources in this thesis, linked spatial data. We assume that a linked spatial data must have appropriate content, which is spatial resource. The spatial resource can be identified using geospatial vocabulary. Therefore, **in this thesis we proposed one metrics that is not included in any previous literature: the existence of geospatial ontology – vocabulary in the LD resources**. This metrics will be further elaborated in the Section 2.5. We developed certain workflow to assess this metrics.

2.4. Identification and Analysis of Linked Spatial Data Sources

The meaning of "spatial" in the terms of linked spatial data can be very diverse. It could contain description about geometry, spatial thing or feature, toponyms or place name, geo web service, geo data format and representation or only an abstract knowledge about geography. That is one issue in how a linked data resources categorizes as linked spatial data. The answer of *"What makes a linked data into linked spatial data?"* will definitely refer to the content and ontology used in that resource. Even though a formal categorization of linked spatial data is not necessary in linked data principle but in practices this categorization is required for dataset discovery activities. Assuming that a data provider wants to enrich their non-spatial LD resources using existing LOD Cloud spatial datasets, *how do they find the proper spatial LD resources on the web?* LD dataset categorization or profiling is the way to solve that problem.

Data catalogue has big role in organizing datasets information to facilitate LD datasets discovery. A data catalogue must record the various datasets using proper identifier or profiles. As LD datasets discovery and retrieval using datahub.io was discussed in Section 2.1.2, in this section we will conduct a more in-depth analysis on how to implement data retrieval. Both datahub.io interfaces and LOD Cloud diagrams are very useful as entry point to discover LD dataset. Here, we used datahub.io tags to find linked spatial data. Datahub.io tags itself is the attribute to define dataset domain (See section 2.1.1). We used CKAN API for listing the CKAN entity of tags to find the relevant tags which may be used by data owner to tag their spatial datasets. Based on the examination, the use of tags by data providers is diverse, for instance, geographic, geography, geo-format, geodata, and others. Data providers might use all these tags to make their dataset labelled as spatial data and become easy to be discovered. However, in most cases, data providers tend to use only one or two tags. We found seven (7) relevant tags for linked spatial data and list of the datasets that used certain tag.

Table 2-3. Datasets that used geo-related tags in datahub.io

Tags	Number of Datasets Found
"geographic"	77
"geography"	22
"format-geo"	42
"geodata"	76
"geo"	81
"spatial-data"	4
"format-spatial"	2

Datahub.io is not a dedicated data catalogue for linked data. Therefore, a comprehensive checking should be performed to the list of datasets. A thorough observation has been done to the datasets that has geo-related tags from Table 2-3. We found that not all datasets have geo-related tags containing RDF data, some of them only contain GeoJSON, KML, or other formats. Another approach that can be used to find LD dataset in datahub.io is to use LOD Cloud Diagram and choose geography domain. The disadvantage of using LOD Cloud Diagram is it only refers to the datasets that contains links to existing datasets that are part of LOD Cloud Diagram. Therefore, this approach did not address the datasets discovery to the whole available spatial linked data on the web. This is evident in the difference in numbers of datasets between geography domain datasets in the LOD Cloud Diagram and geo-related tags in datahub.io. The diagram shows only 38 geographic datasets while datahub.io provides more numbers of spatial datasets. Thus, we prefer to discover LD datasets using CKAN API and conduct manual content checking.

To ensure the datasets as intended, we set three criteria to filtered out some of the datasets. The criteria are:

1. Datasets must have geo-related tags
2. Datasets must have one of the RDF Serialization data format.
3. Datasets must have either RDF Dump or SPARQL endpoint to access whole datasets.

Differences in data format, access and storage

During observation, we also found the variety in data format, access and storage (see Figure 2-6). Only a few of the data owners provide SPARQL endpoints for querying data. This happens due to the high cost of server and maintainability. To establish a SPARQL endpoint requires a query engine and a SPARQL server. Most data owners only provide RDF data in the form of webpage (RDFa) and RDF dump for public to access their data. Data storage through RDF dump also varies; some data owners store all RDF data on single files, and other data owners store in subsets. Variations on RDF dump data also occur in the serialization format. Most of data owners use rdf/xml format (14 dataset), and some use n-triples (4 dataset) and turtles (6 dataset).

After checking the 304 datasets in datahub.io, we selected 26 datasets as candidate datasets to be used in this study (see Table 2-4). During observation, we found that there are several variations of access and storage of linked data. From selected 26 datasets, three datasets have SPARQL endpoints, and 23 have RDF dump. From 26 datasets that have RDF Dump, seven datasets choose to subset their dataset based on certain use categories, and the rest use single storage. All these differences will certainly determine the method to retrieve the datasets. A data retrieval workflow was developed to deal with data format, access, and storage differences. This implementation will be further elaborated in data processing (Section 2.6).

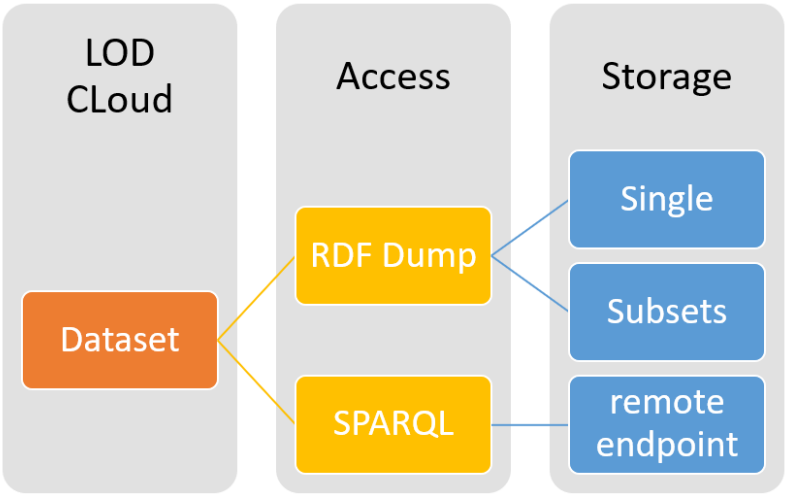


Figure 2-6. Data Architecture of LOD Cloud

Table 2-4. List of Candidate Datasets

No	Dataset	Format of RDF Dump Format / SPARQL Endpoint	RDF Storage	Appear in LOD Cloud Diagram?
1	AEMET meteorological dataset	endpoint sparql (http://aemet.linkeddata.es/sparql/)	No	
2	GeoLinkedData	endpoint sparql (http://geo.linkeddata.es/sparql)	No	Yes
3	Linked NUTS (ONS)	endpoint sparql (http://statistics.data.gov.uk/sparql)	No	
4	Administrative Unit Germany	n-triples	Subsets	
5	Ordnance Survey Linked Data	n-triples	Subsets	Yes*
6	GADM	n-triples	Subsets	
7	Accommodations in Piedmont (LinkedOpenData.it)	rdf/xml	Single	Yes
8	Australian Climate Observations Reference Network	rdf/xml	Subsets	
9	CAP Grids	rdf/xml	Single	Yes*
10	EARTh	rdf/xml	Single	Yes
11	education.data.gov.uk	rdf/xml	Single	
12	European Nature Information System	rdf/xml	Subsets	Yes*
13	FAO geopolitical ontology	rdf/xml	Single	Yes
14	Geological Survey of Austria (GBA) - Thesaurus	rdf/xml	Single	Yes
15	GeoNames Semantic Web	rdf/xml	Single	Yes*
16	GeoSpecies Knowledge Base	rdf/xml	Single	
17	Hellenic Police	rdf/xml	Single	
18	Postal codes Italy (LinkedOpenData.it)	rdf/xml	Single	Yes*
19	Telegraphis Linked Data	rdf/xml	Single	Yes
20	Linked Sensor Data (Kno.e.sis)	tar	Single	Yes
21	DataGovIE - Irish Government Data	turtle	Single	
22	Geo Names Information System (GNIS)	turtle	Subsets	Yes
23	Lower Layer Super Output Areas	turtle	Single	
24	NUTS (GeoVocab)	turtle	Single	Yes
25	Pleiades	turtle	Single	Yes
26	transport.data.gov.uk	turtle	Single	Yes*

2.5. Geospatial Ontologies - Vocabularies

As briefly mentioned in Section 1.1, semantic web aimed to facilitate data integration. In geospatial field, the main issue is to discover how spatial data can be integrated to an interoperability environment on the web. Hu (2017) mentioned that ontologies development is the major approach to facilitate semantic interoperability. Therefore, geospatial ontologies development is essential to realize the geospatial semantic web. Geospatial ontologies are considered as domain ontologies as they specifically aim for interoperability within geo-information science field. Di & Zhao (2017) stated that interoperability in geospatial semantic web is the ability to conduct sharing cross domain resources and knowledge between geo-information science specific domain fields in the semantic web environment. Technically, the authors added, it must support the ability for cross-domain discovery and various resource queries. This ability can only be implemented if geospatial concepts and relationship are declared.

In the linked data context, the role of the ontologies is to provide the classes and individual (instance) definitions. In terms of triples that contain subject predicate object, the ontology act as the predicate relations that capture the relationship between two LD resources, subject, and object. The predicate statement can be represented by object property or data property of ontology depending on the level of data. The relationship between instance data can be explained with the data property, while relationship between classes can be explained using the object property.

In terms of geospatial ontologies, there is not any single ontology that fits all data and services. Hence, every domain-specific community and dataset provider puts an effort to build their own ontologies (see Table 1 in the Appendix.). Each ontology is developed on conformity to geospatial semantic web context. They can be categorized into seven (7) groups based on the role in geospatial semantic web (Di & Zhao, 2017), namely: 1) General Ontology, 2) Geospatial Feature Ontology, 3) Geospatial Factor Ontology, 4) Geospatial Relationship Ontology, 5) Geospatial Domain-Specific Ontology, 6) Geospatial Data Ontology, and 7) Geospatial Service Ontology. The hierarchy of geospatial ontology can be seen in Figure 2-7.

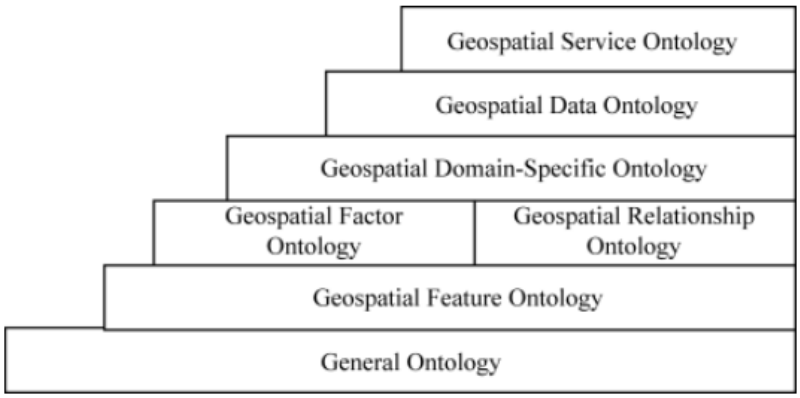


Figure 2-7. Hierarchy of Geospatial Ontology, obtained from Hu (2017)

In this thesis, as described in Section 2.3.2, we want to identify the existence of geospatial vocabulary in the LD resources. We restricted the analysis to only two types of geospatial ontologies, i.e. Feature and Relationship Ontology. First, Geospatial Feature Ontology represents the geospatial entities, it aim to provide representation that align with the OGC and ISO standard for general feature model (W3C, 2007). Example for geospatial feature ontology is GeoRSS, Second, Geospatial Relationship Ontology signifies the logical relationship between geospatial features. The examples for this type are NeoGeo and Ordnance Survey Spatial Relations. This ontology is very useful in linked data because it enables the topological

relationship as well as qualitative reasoning which is widely used in GeoSPARQL query to retrieve spatial resource. 26 chosen datasets (see Table 2-5) will be processed using workflow for data analysis. This implementation is described further in Section 2.6

2.6. Workflow for Linked Spatial Data Analysis

As we discussed previously, the candidate of LD datasets has differences in access, storage, and data formats. First, regarding the data access, data processing should be able to access two types of data access: local endpoint and remote endpoint (SPARQL endpoint). Although the remote endpoint can directly access via SPARQL endpoint in the browser, this will not be effective for further data analysis. Therefore, in this study Python Integrated Development Environment (IDEs) is used for the data processing. The process includes: 1) Data fetching, 2) Data parsing to get a graph, and 3) Graph traversing to get intended information. Several python modules are required for the workflow development. The general workflow has been developed to process linked data in the LOD cloud into valuable information (see Figure 2-8).

2.6.1. Designing Workflow for Data Analysis

Figure 2-8 depict the workflow for data processing. Gastrodon, RDFlib, SPARQLWrapper modules were used to implement workflow of data access of both local and remote endpoint, data parsing, and post a relevant SPARQL queries for the data. Data processing required thorough checking especially datasets that only has RDF dump as data access. We discovered that several RDF dump datasets, for instance Linked Sensor Data (Kno.e.sis), GeoNames, and NUTS did not follow the standards of linked data. The problems include serialization error, syntax error, HTTP header error, and namespace prefix error. We fixed these problems using function in python script and applying loop to all the triples (see listing 1, 2, 3 in the Appendix). Rietveld (2016) also stated that many data dumps are not fully standard-compliance. Rigorous data processing needs to be applied on RDF dump to ensure they follow RDF writing standard. In addition, as we understand that RDF dump is a static repository that is only updated in certain frequency, the resources tend to be out-dated for real-time applications. Moreover, we also must deal with subsets of data storage (RDF dump). Subsets means that the data provider store the LD resources based on certain category and dumped them in different repositories instead of in a single RDF dump. These reasons make linked data consumption from a local deployment of RDF dump is more difficult than SPARQL endpoint

Data processing from SPARQL endpoint is relatively easier since the data already followed the standards. The endpoint allows the user to pass a query to the certain triple store. Therefore, a syntax error never found from triples. Furthermore, the datasets are guaranteed to be updated because that live query is connected to the active database instead of static repository. Yet, there are also a few shortcomings of using SPARQL endpoint as data access. They include: 1) the low availability of SPARQL endpoint, only couples of datasets have this functionality, and 2) Usage limitations; as a public service the server usually sets restrictions for instance maximum query execution time and maximum query cost estimation time.

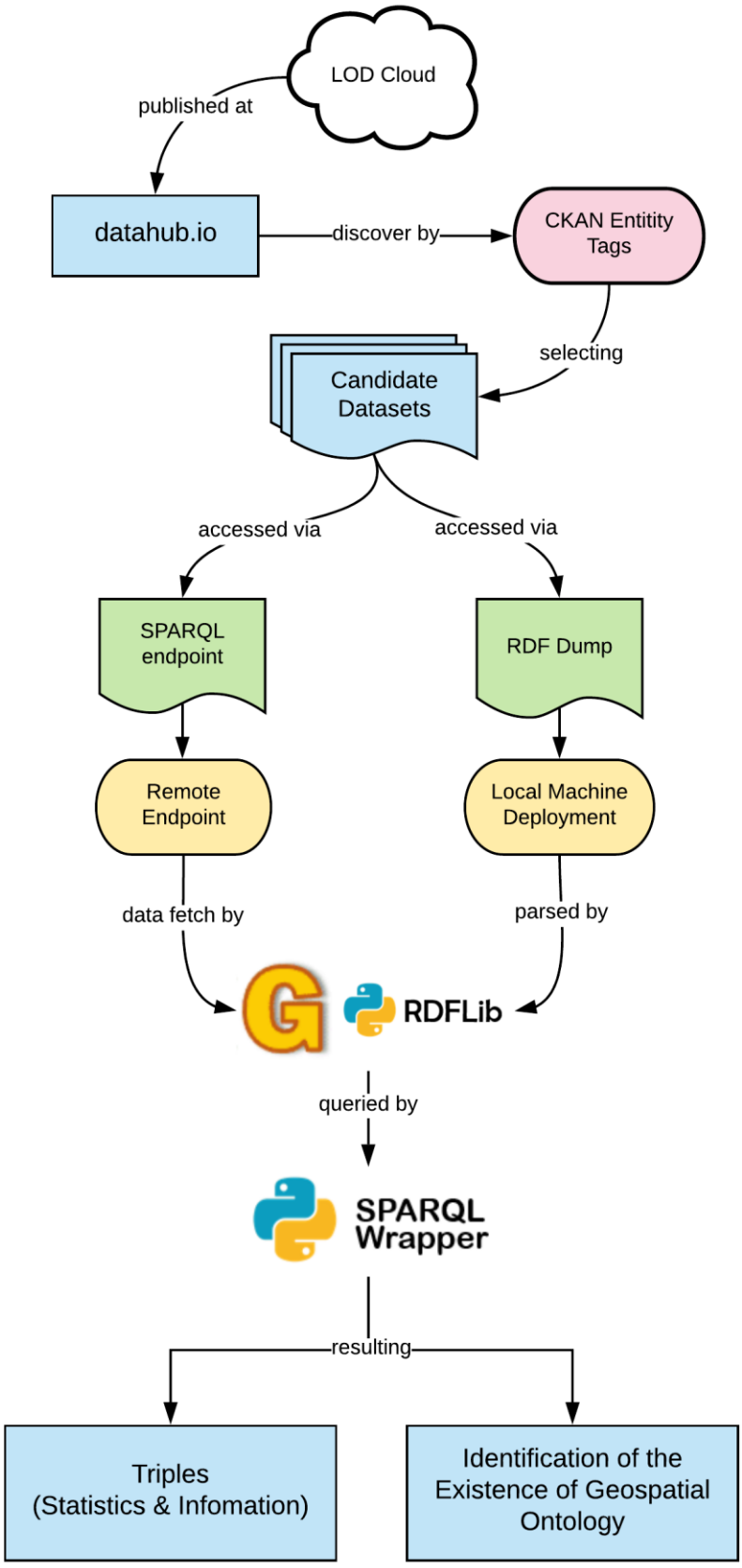


Figure 2-8. Workflow 1: Data Analysis

The explanation of Triples (Statistic & Information) is provided in Section 2.6.2
The result of Identification of Existence of Geospatial Ontology provided in Table 2-6

The existence of geospatial vocabulary and content of LD resources will be checked on each chosen dataset. These checking process is conducted based on linked data principle which encourage the integration to other separate (geospatial-related) data attributes which are not stored in the same storage. Thus, we set two criteria for assessing the LD resources to fulfil the principles of linked data and to identify the spatial content. The criteria are:

1. LD Resource must have geospatial feature or relationship ontology or geometry vocabularies on its predicate statement.
2. LD Resource must refer to external resource (URI) on its object statement.

Geospatial Ontology Checking

The Geospatial Ontology existence checking starts by posing SPARQL queries to local or remote endpoint. It aims to identify the used vocabularies in the LD resources within certain datasets (see Listing 4 in the Appendix). The result of queries depends on RDF serialization format of the data. N3, n-quads, and n-triples return a complete URI of vocabularies, while rdf/xml only returns a namespace prefix. To understand what namespace prefixes means, we should check those prefixes to the header of the RDF. A detailed check has been done to the 26 chosen datasets, and we listed the vocabularies that are used in each dataset. The ontologies found are very diverse, in total 31 ontologies are used in the 26 datasets (see Table 1 in the Appendix). However, in this thesis we are only interested in geospatial feature and relationship ontologies. To ensure this, we did content analysis to each vocabulary in 26 datasets to learn and analyse the role of those vocabularies in certain ontology. We also use Linked Open Vocabularies (Vandenbussche, Atemezing, Poveda-Villalón, & Vatant, 2016) as reference to search geo-related vocabularies.

External URI Checking

The next step after checking the existence of geospatial vocabularies is examining the object statement of each LD resources. The examination was conducted to find out whether those spatial vocabularies are used to describe the spatial resources. Next checking, object statement of a LD resource must refer to external resources. This task can be done by traversing the graph and posing the relevant SPARQL queries. SPARQL queries that are used in this section provided in the Listing 4 in the Appendix. This process consists of two levels which only triples that contain geospatial ontologies – vocabularies will be checked for its object statement. Workflow 2 (see Figure 2-9) has been designed to depicts the order of implementation process of these two criteria.

Based on examination of 26 datasets, 6 datasets do not contain any geospatial ontologies. Therefore, we only proceeded with 20 datasets to examine their object statement. We observed that these 20 datasets do not fully implement the linked data best practices. While linked data practices encourage a LD resource to establish link to the external resource (URI), we found that most of the data do not have any external resources in their dataset. After a rigorous examination, we found that 17 datasets only refer to their own resources. At last, only found 3 datasets that has geospatial ontology and refer to external resources as their object statement, as follows:

1. Linked Sensor Data (Kno.e.sis) resources linked to GeoNames resources
2. Lower Layer Super Output Areas resources linked to Statistical UK resources
3. NUTS (GeoVocab) resources linked to many of external resources.

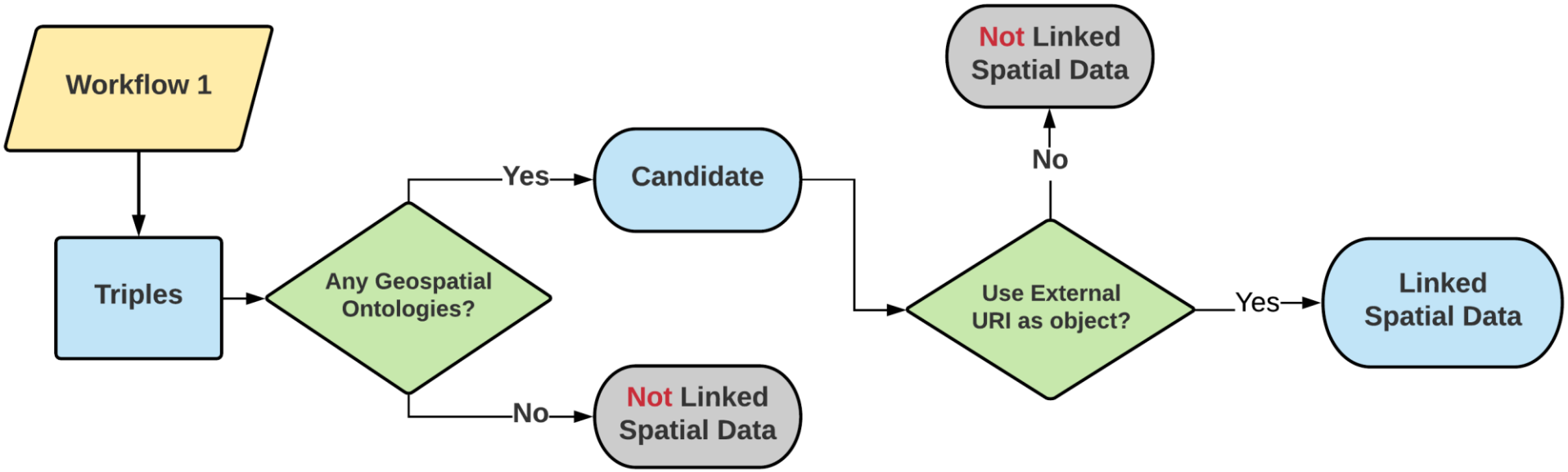


Figure 2-9. Workflow 2: Identification of Geospatial Feature & Relationship Vocabularies

Relation between analysis result and LOD Cloud Diagram

Table 2-6 provides a list of all used geospatial feature and relationship vocabularies and the object statement of each dataset. Based on this result, we found interesting relation between our analysis result and LOD Cloud Diagram. Most of 17 datasets, which are linked only to their own resources. were not appear on LOD Cloud Diagram. For instance, the dataset of Ordnance Survey only established link between their own resources, and none of the triples refers to external URI as object. It seems that they assume their own data is more than enough to provide information and there is no need to enrich their datasets by establish back link to the external URIs. This also happens to big data providers such as European Nature Information System (ENIS) and GeoNames Semantic Web. The data completeness and trustworthiness from official data provider make their dataset as reference to the other data. For this reason, it returned nothing when we check the existence of the external URIs in their object statement.

Redundant Ontologies

Another finding is each data provider tends to develop their own ontologies - vocabularies to describe their datasets. This is rather redundant because the new developed vocabularies have the same function with the existing vocabularies. As an example, there is more than one vocabulary to describe longitude and latitude. A similar problem also occurs for topological relationships vocabularies; instead of promoting the reuse of existing vocabularies, the data provider tends to develop other topological relationship vocabularies. Vocabulary development itself is divided into two categories. The first, a specific-application ontology. The ontology that develop all the vocabularies to fully describe their dataset. Datasets like FAO geopolitical, Ordnance Survey, and transport.data.gov.uk developed their own ontology – vocabularies to describe their datasets. The second is the ontology that uses existing vocabularies as baseline then develop detail vocabularies that have not been covered yet by existing vocabularies. The example is IGN France vocabulary which re-use vocabulary of GeoSPARQL and NeoGeo to describe geometry and Geonames vocabulary for define a feature class. Pattuelli, Provo, & Thorsen (2015) stated that the openness of the linked open data should give flexibility for ontologies development to combine vocabularies form different sources.

Table 2-5. Result of checking the existence of geospatial ontology	Any Geometry Vocab?	Linked to External Object?	What are the geo Vocab?	Namespace
AEMET Meteorological	Yes	No	geo: lat, geo: lang, geo: location	geo: http://www.w3.org/2003/01/geo/wgs84_pos#
GeoLinkedData	Yes	No	geo: lat, geo: long, geo: geometry, geosparql: hasGeometry, geosparql: asWKT, geosparql: asGML,	geo: http://www.w3.org/2003/01/geo/wgs84_pos#, geosparql: http://www.opengis.net/ont/geosparql/
Administrative Unit Germany	Yes	No	geosparql: spatialDimension, geosparql: hasGeometry, geosparql: asWKT, geosparql: asGML, geosparql: isSimple, geosparql: is3D, geosparql: hasSerialization, geosparql: Dimension, geosparql: coordinateDimension	geosparql: http://www.opengis.net/ont/geosparql/
Linked NUTS (ONS) - NUTS UK	Yes	No	spatialuk: withincountry, spatialuk: region, spatialuk: localauthoritydistrict, publish: within	spatialuk: http://statistics.data.gov.uk/def/spatialrelations/, publish: http://publishmydata.com/def/ontology/spatial/
Ordnance Survey Linked Data	Yes	No	osgeom: asGML, osspr: touch, osspatial: contains, osspatial: contains, osspatial: within,	osgeom: http://data.ordnancesurvey.co.uk/ontology/geometry/, osadm: http://data.ordnancesurvey.co.uk/ontology/admingeo/, osspr:

			osadm: countyElectoralDivision, osadm: inDistrict, osadm: hasUnitID	http://data.ordnancesurvey.co.uk/ontology/spatialrelations/
GADM	Yes	No	spatial: PP, spatial: Ppi, ngeo: geometryOf, ngeo: geometry, gadm: contains, gadm: inregion, gadm: incountry	spatial: http://geovocab.org/spatial, gadm: http://gadm.geovocab.org/ontology, ngeo: http://geovocab.org/geometry#
Accommodations in Piedmont (LinkedOpenData.it)	Yes	No	vcard: geo, vcard: latitude, vcard: longitude	vcard: http://www.w3.org/2006/vcard/ns#
Australian Climate Observations Reference Network	No	No		
CAP Grids	Yes	No	osgeo: extent, osgeo: asGeoJSON, osgeo: asWKT,	osgeo: http://data.ordnancesurvey.co.uk/ontology/geometry/, osspatial: http://data.ordnancesurvey.co.uk/ontology/spatialrelations/
EARTh	No	No		
education.data.gov.uk	No	No		
European Nature Information System	No	No		
FAO geopolitical ontology	Yes	No	geopol: hasMaxLongitude, geopol: hasMinLongitude,	geopol: http://www.fao.org/countryprofiles/geoinfo/geopolitical/resource/

			geopol: hasMaxLatitude, geopol: hasMinLatitude	
Geological Survey of Austria (GBA) - Thesaurus	Yes	No	geo: lat, geo: lang	geo: http://www.w3.org/2003/01/geo/wgs84_pos#
GeoNames Semantic Web	Yes	No	geo: long, geo: lat, gn: locationMap	geo: http://www.w3.org/2003/01/geo/wgs84_pos#, gn: http://www.geonames.org/ontology#
GeoSpecies Knowledge Base	Yes	No	geo: long, geo: lat	geo: http://www.w3.org/2003/01/geo/wgs84_pos#
Hellenic Police	Yes	No	geo: long, geo: lat	geo: http://www.w3.org/2003/01/geo/wgs84_pos#
Postal codes Italy (LinkedOpenData.it)	No	No		
Telegraphis Linked Data	Yes	No	geographis: onContinent	geographis: http://telegraphis.net/ontology/geography/geography#
Linked Sensor Data (Kno.e.sis)	Yes	Yes (to geonames)	om-owl: hasLocation, geo: long, geo: lat, geo: alt	om-owl: http://knoesis.wright.edu/ssw/ont/sensor-observation.owl#, geo: http://www.w3.org/2003/01/geo/wgs84_pos#
DataGovIE - Irish Government Data	No	No		
Geo Names Information System (GNIS)	Yes	No	ago: geometry, geosparql: asWKT, gnis: state, gnis: county	ago: http://awesemantic-geo.link/ontology/, geosparql: http://www.opengis.net/ont/geosparql/, gnis: http://data.usgs.gov/lod/gnis/ontology/
Lower Layer Super Output Areas	Yes	Yes (to Stats UK)	spatial: easting, spatial: northing, stats: district, geo: lat, geo: long	spatial: http://data.ordnancesurvey.co.uk/ontology/spatialrelations/, stats: http://statistics.data.gov.uk/def/administrative-geography/,

				geo: http://www.w3.org/2003/01/geo/wgs84_pos#
NUTS -Europe-(GeoVocab)	Yes	Yes (a lot)	spatial: PP, spatial: PPi, spatial: EC, ngeo: posList, ngeo: exterior, ngeo: geometry, ngeo: interior, ngeo: polygonMember, geo: lat, geo: long,	geo: http://www.w3.org/2003/01/geo/wgs84_pos#, ngeo: http://geovocab.org/geometry#, spatial: http://geovocab.org/spatial#
Pleiades	Yes	No	pleiades: hasLocation, geo: long, geo: lat, spatial: C, osspatial: within, osspatial: partiallyOverlaps, osgeo: extent, osgeo: asWKT, osgeo: asGeoJSON	geo: http://www.w3.org/2003/01/geo/wgs84_pos#, osgeo: http://data.ordnancesurvey.co.uk/ontology/geometry/, osspatial: http://data.ordnancesurvey.co.uk/ontology/spatialrelations/, pleiades: https://pleiades.stoa.org/places/vocab#, spatial: http://geovocab.org/spatial#
transport.data.gov.uk	Yes	No	geouk: point, geouk: hasArea	geouk: http://geo.data.gov.uk/0/ontology/geo#

2.6.2. SPARQL Implementation, Result, and Visualization

Besides checking the existence of geospatial vocabularies and external resources in LD resources, we can also post some questions to find out general information or statistics on the triples within datasets and also information about relationship between resources. Some sample questions about general information include *"How many triples in dataset X?"*, *"What is the most used ontology – vocabularies in dataset Y?"* or any other questions that are related to the triples. Additionally, questions about relationship between resources include *"What kind of resources that defined by topological relationship vocabularies?", "What are the class hierarchy of the resources?",* or even *"Which properties are inherited by super class to subclass?"*. They can be answered by traversing the graph using SPARQL query to get the intended information. The knowledge of SPARQL and the ontology or data model of certain datasets are required to perform the queries. In this attempt, we thoroughly explore the spatial element in the LD resources, such as geometry resources and geospatial vocabularies. We also visualized the query result to better understand the state of spatial datasets in the LOD Cloud. In this study, we implemented the SPARQL queries into two types of data access: local endpoint and remote endpoint in order to find out the shortcomings of each options. The queries to local endpoint are discussed in this section while the queries to remote endpoint (SPARQL endpoint) is presented in Chapter 4.

We used Gastrodon module in python environment to deploy RDF dump, which has previously been downloaded via datahub.io, as local endpoint. The function in the module made possible to post query to the local collection of triples. The process on this local endpoint includes parsing the dataset into a set of graphs, each of which contains one subject, one predicate, and one object. Based on this structure, we can query to traverse among the graph. We implemented to two datasets that have different data storages: single and subsets storage.

1. **GADM - The World's spatial database of administrative areas (countries and subdivisions)**
   A RDF dump of GADM dataset is stored in a single storage file of n-triples format. However, this file cannot be processed directly because some statements that do not fit the standard RDF writing. To overcome this, data cleaning is required (see Listing 2 in the Appendix). After the data is clean, this RDF dump can be processed for parse stage. Once this stage is completed, we found this RDF dump file contains 10,196,504 triples. We also identified the type and percentage of each used predicate in the GADM datasets (see Figure 2-10).

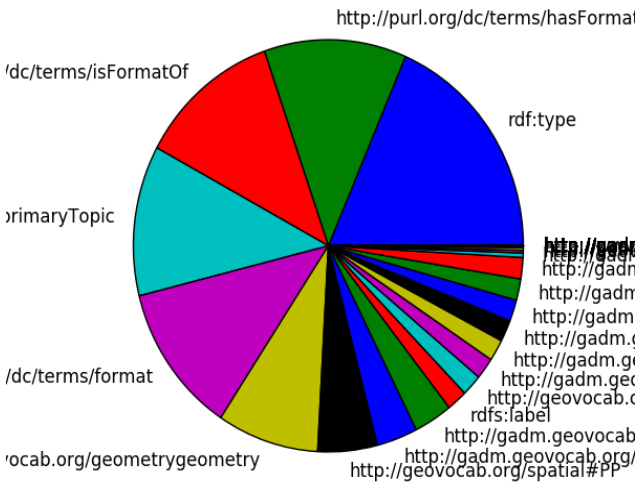


Figure 2-10. Percentage of used vocabularies in GADM Dataset

2. **Ordnance Survey (OS) linked data**

RDF dump from OS linked data are stored in subset files based on categories of data. Ordnance surveys provide several categories such as gazetteer, boundary, and geographic area, which are stored in different dump files. In this section, we tried to capture the relationship between resources. In this case, we used the example OS linked data resource (http://data.ordnancesurvey.co.uk/doc/7000000000037256 - City of Southampton) to implement the code. The goal is to list geospatial vocabularies and to observe what kind of resource that linked to those geospatial vocabularies.

Table 2-6. List of geospatial vocabularies in LD resource of City of Southampton.

Prefixes	Namespace	Triples
j.2:	http://data.ordnancesurvey.co.uk/ontology/spatialrelations/	
	within	1
	contains	2062
	northing	1
	easting	1
j.3:	http://data.ordnancesurvey.co.uk/ontology/geometry/	
	extent	1
j.4:	http://www.w3.org/2003/01/geo/wgs84_pos	
	lat	1
	long	1
j.5:	http://www.georss.org/georss/	
	point	1

Using the data above, we can find out the most used vocabulary in one OS linked data resource is `spatialrelation:contains`. Furthermore, we can examine the use of `spatialrelation:contains`, whether this vocabulary is used to describe geometry data. It can be done by writing the correct SPARQL query to the resource by using SPARQLWrapper in python (see Listing 1).

Listing 1. Query to extract object from triple

```
properties=e.select("""
   SELECT ?o {
      ?s j.2:contains ?o .
   } """)
properties
```

The query above generates object resource of a triple that uses `spatialrelation:contains` as its predicate. From the examination of this result, the object resource that uses `spatialrelation: contains` does not represent the actual geometry (see Table 2-8). Geometry data is found on others triples that use predicate such as `wgs84_pos:lat, wgs84_pos:long` and `georss:point` (see Table 2-9). This examination is important to understand which vocabularies that carries geometry resources.

Table 2-7. List of objects in a triple that use `spatialrelation:contains` as predicate

No	Predicate	Object
1	spatialrelation:contains	http://data.ordnancesurvey.co.uk/id/4000000023503841
2	spatialrelation:contains	http://data.ordnancesurvey.co.uk/id/4000000023491917
…		…….
2062	spatialrelation:contains	http://data.ordnancesurvey.co.uk/id/4000000023491661

Table 2-8. List of objects in a triple that use geometry vocabulary as predicate.

No	Predicate	Object
1	`wgs84_pos: lat`	"50.9169971687"
2	`wgs84_pos: long`	"-1.39872335574"
3	`georss: point`	"50.9169971687 -1.39872335574"

2.7. Summary

This chapter summarised the essential element to understand the state of linked open data cloud. We discussed the current condition of linked data in LOD Cloud by investigating strategies to retrieve dataset from cloud, examining the structure of LD resources, and describing previous research on linked data quality framework. We developed a workflow that provides a structured way to analyse the LD resources from different data access, storage, and format files. Overall, this chapter provided a structured way to consume and extract information of the LD resources. Data analysis includes processes of data fetching, data cleaning, parsing, and graph traversing. The result provides information about LD resource, for instance the type of ontologies - vocabularies that are used, or the type of data that is contained in certain LD resources. In investigating linked data, we have to fully consider the ontology and vocabularies that used in certain LD resources. The ontology could indicate the domain or topic of the datasets. The workflow can be applied to identify data domain, including geospatial domain. The whole workflow provides a pipeline that can be used by other users to analyse the spatial content from LD resources in the LOD Cloud for their own applications. In this chapter, we also defined the terms of "Linked Spatial Data" by setting several criteria.

3. LINKED SPATIAL DATA VISUALIZATION FOR DISCOVERY AND EXPLORATION

The landing page of LOD Cloud (http://lod-cloud.net/) has a significant role in promoting the reuse of linked data to the wider community. This can be achieved by providing intuitive datasets representation to expose the available datasets of linked data across sources. As the constellation of available interlinked datasets on the web, LOD Cloud Diagram interface may potentially provide several insights: first, to support the users in utilizing the LD datasets according to their needs; second, to support the data owner in finding potential links to create new links between datasets.

LOD Cloud Diagram also has the potential role to fill the gap in the current linked data visualization, which currently is only able to depict the relationship between the resource in instance level and not in set level. Previous studies on linked data visualization at large only focused on visualizing the RDF graphs and finding suitable interface for resources in the **instance levels** within specific datasets. Hence, linked data exploration and discovery are restricted to only a small part of the linked data that exists on the web. While LOD Cloud Diagram can help to get the *"big picture"* of available linked data to support application that requires unbounded resources. This section elaborates recent research about the role of intuitive visualization for linked data discovery, exploration, and consumption task through link across datasets in **set levels**, called **linkset.**

3.1. Linked Data Exploration and Visualization Systems

The fast growth of linked data resources on the web makes LOD Cloud one of the biggest repository of interlinked data on the web (Assaf, Troncy, & Senart, 2015). As argued by Graziosi, Di Iorio, Poggi, & Peroni (2017), one major question is how to visualize large semantic dataset across sources in comprehensive manner. Apart from storage and query solution, this section will discuss the role of visualization to support the data exploration and discovery. Idreos, Papaemmanouil, & Chaudhuri (2015) characterize visualization as visual data exploration, arguing that the function of visualization has shifted from only information perception provision to knowledge extraction. The main goal of visualization is to transform the data to visual representation in order to provide better understanding for the user (Brunetti, Auer, García, Klímek, & Nečaský, 2013). This is supported by Dadzie & Rowe (2011), who argue that the visualization must guide the user into further process of knowledge discovery and information retrieval (IR). One of basic principles of linked data is reusability, and the visualization is the frontier in promoting and encouraging the usability of linked data by the wider community.

The visualization framework should be able to create a sensible technique to expose the content of knowledge. An effective knowledge discovery starts with sufficient insight about datasets that leads to deeper exploration. An insight can be obtained from the overview visualization which provides mechanisms for information abstraction and summarization (Bikakis, Papastefanatos, Skourla, & Sellis, 2017). Indisputably, an intuitive visualization is necessary to support the exploration. The data exploration itself is diverse based on the preferences and requirement of various users and applications (Bikakis, Greece, & Sellis, 2016). The important part of linked data visualization is to provide sufficient understanding of linked data structure and relationship for both tech-savvy and lay-user. This

understanding could assist the users to investigate the existing relationship between datasets, to identify link between LD resources, and to find relevant knowledge for their application. Those exploration activities are applied not only to the famous and centred resources like DBPedia but also to several hidden linked data resources that exist beyond the spotlight. The expected result of dataset exploration is to encourage the creation of new linked data. The creation of new resources by extending the existing resource to the external data certainly will make LOD Cloud bigger, richer, and more comprehensive from across domain knowledge. The growth of LOD Cloud in recent years indicates the usability and usefulness the Web of Data.

Several works have been developed to face the challenges of an effective linked data visualization. Bikakis et al. (2016) conducted a thorough survey and categorized linked data visualization tools. The type of exploration and visualization system are as follows:

1. **Browser and Exploratory systems**
   Browsers provide link navigation (hyperlink) and representation of linked data resources in the format of RDFa. The browser present LD resource in the textual representation on the HTML page. Human-readable aspect is one of the advantages of using browser as exploration system. The example linked data browser is URI Burner.

2. **Generic Visualization systems**
   This type of system is defined as a wide-range options of visualization within one standalone visualization tool. For instance, Linked Data Visualization Model (Brunetti et al., 2013) provides various types of visualization across datasets and several techniques to depict the dataset relationship. The other example is Rhizomer (García, Brunetti, Gil, & Gimeno, 2012), which offers various types of visualization to support linked data exploration, adopting the visualization seeking mantra by Shneiderman (1996), *"overview first, zoom and detail in demand"*.

3. **Domain, Vocabulary & device-specific visualization system**
   This system includes various tools to describe certain domains, for instance statistical data or geospatial data. Geospatial data visualization tools will be further elaborated in Section 3.4. The sample of device-specific visualization tool is DBPedia Mobile, a dedicated application to visualize DBPedia location-related resources.

4. **Graph-based visualization system**
   Graph is the most-popular technique to describe linked data resources since graph can depict relationships and hierarchy. It aligns with linked data structure that suitable with visualization of a node-link – two resources that connected by a relationship. Therefore, there are many researches that used graph as a baseline for visualization application, for instance RDF-Gravity, RelFinder, graphvizdb, Gephi, and many others. Most visualization tools describe resources in instance level, and one dataset might consist of thousands of interlinked instances. Therefore, to avoid clutter visualization, these tools apply several strategies, for instance, data filtering, data sampling, and aggregation mechanism. Chawuthai & Takeda (2016) proposed strategies like graph simplification, property selection, and triple ranking. Several visualization tools allow the presentation of selected ROIs of graph or even the whole graph. However, considering the memory requirement to load a whole graph layout, the visualization tools are generally limited or restricted presentation into a piece of sampling graph.

5. **Ontology visualization systems**

Ontology is the reference for SPARQL query construction, thus it has an important role in data retrieval process. As several ontologies are very complex, the ontology visualization can assist users to get a better understanding of the data type and structure within a dataset. An example of ontology visualization system is LD-VOWL (Weise, Lohmann, & Haag, 2016), which uses a formal Visual Notation for OWL Ontologies (VOWL) to visualize extracted schema information. The VOWL determines the visual language of ontologies for user-oriented representation. For instance, it uses circle to represent class of ontology, line to represent property relations, and rectangle for property labels and datatype, which can be very useful for comprehensive ontology. For a simple ontology, it might use a simple hierarchical of ontology, for instance ontology of Kadaster BAG dataset (Kadaster, 2017) that explains classes and properties.

6. **Visualization libraries**

There are several libraries that can be used to embedded linked data visualization in the web page. The visualization commonly includes JavaScript, CSS, and HTML. For instance, Sgvizler (Skjaeveland, 2012) is a JavaScript wrapper that can visualize the result of SPARQL query. A variety of visualization options are provided by Sgvizler, for instance tree map, timeline, and pie chart. Recent works conducted by Rietveld & Hoekstra (2015), they developed a comprehensive SPARQL clients called YASGUI (Yet Another SPARQL Graphical User Interface). YASGUI includes visualization library called YASR (Yet Another SPARQL Result-set visualizer). This JavaScript library is able to parse and visualize the result or response from SPARQL query. It supports line, bar, maps, scatter plots, and scatterplots. YASGUI is described further in the Section 4.3.6

Dadzie & Rowe (2011) also conducted comprehensive work to review existing visualization tools, specifically Linked Data browser, and found that there is no tool that provides the overview aspect. It contradicts to the discussion in the beginning of Section 3.1 about the fact that users need overview as an entry point for the further data exploration. The overview functionality could support users to understand the datasets structure comprehensively, which initiates efficient exploration. The visualization should lead to user's awareness about available content and the links between datasets (De Vocht et al., 2014). Beek & Folmer (2017) developed an integrated linked data browsing for Kadaster Netherlands that aimed to support users from multiple use cases. They stated that the users might need to obtain the overview from available datasets, which are information about data structure or link between resources. This information may become the starting point to go to the intended resources. It helps to retrieve chosen ROIs resources for their application and guiding users to deeper exploration towards knowledge discovery. The ultimate end-goal is new linked data creation.

This section also discusses the theoretical framework of the key requirement of Linked Data visualization in order to make linked data more accessible for consumption. Dadzie & Rowe (2011) developed a design guideline for visual representation and analytics which is applied to the linked data consumption. The guideline is the result from exhaustive survey based on summarization of previous studies on linked data visualization. The key requirements are described in Table 3-1:

Table 3-1. Key Requirements for Visual Representation & Analytics for Linked Data Consumption, obtained from Dadzie & Rowe (2011)

No	Requirements	Explanation
1	Visual presentation	To achieve an intuitive discovery and analysis, various visualization element and feature are implemented to support user understand the data structure and content.
2	Data overview	Provenance of global view of the datasets that include information abstraction and summarization. It has role as entry point for further exploration
3	Detail on demand	Functionality to get a clear view in the targeted part of the dataset (ROIs resource) in order to support detail discovery and exploration.
4	Highlight links in data	The ability to investigate the type (used vocabulary) and number of links within a dataset or between datasets. It gives important information about dataset(s) since the core of linked data is the links or relationship itself.
5	Support for scalability	The ability to manage complex, distributed, heterogeneous large LD datasets.
6	Support for querying	In the context of linked data, a formal query syntax of SPARQL used as information retrieval procedure. To support non-technical users, it also possible to develop a simple query of string-based matching, keyword-in-context, interactive user interface (point & click), and etc.
7	Filtering	A customization of the visualization, possibility to apply dynamic filter of selected ROIs to the user interface. Also include possibility to generate preview from ROIs based attribute or property selection of the resources.
8	History	Simple functionality to return from recent discovery navigation to the previous page (navigation or analysis)

Based on the requirements above, we conducted an assessment on the LOD Cloud Diagram, using current LOD Cloud Diagram (version 2017-08-22) as the baseline. LOD Cloud Diagram could not be classified as one of five categories in the visualization system because the diagram itself are not designed to be a dedicated linked data exploratory system. Initially, LOD Cloud Diagram was developed to give insight into the growing initiatives of linked data at that time (2007) which started with only 12 datasets. Previously, it provided most recent updates by giving notifications of new datasets to the users. At that period of time, there was no formal syntax query to retrieve the linked data resource on the web. The storage and resources information were not systemically linked yet. These reasons instigated the maintenance of LOD Cloud Diagram until now, for it gives valuable contribution to depicting constellation of available datasets of linked data on the web. This diagram is considered as the most up-to-date visualization of available LD dataset that implements linked data principles and widely-used dataset domain categorization (Abele et al., 2017). Based on this condition, it is worthwhile to understand to what extent the current feature in LOD Cloud Diagram supports the exploration and discovery activities and how LOD Cloud user interface can be improved for exploration and data integration. The following section (3.2) will further elaborate this discussion.

3.2. Expert Opinion

3.2.1. Interview Setup

In this chapter, we focused on elaborating how LOD Cloud Diagram could be extended in order to support discovery and exploration purposes. For this purpose, we assessed current LOD Cloud Diagram visualization with respect to key requirements for visual representation & analytics for linked data consumption. Since LOD Cloud diagram is not a visualization system, typical usability test which evaluates certain visualization dashboard, tools, or interface was not used in this analysis. Instead, this analysis focuses on understanding the potential use of LOD Cloud from a user's perspective. This test was conducted to assess to what extent LOD Cloud can be operationalized based on user requirements and to know whether LOD Cloud accommodates those requirements.

Data collection on users' perspectives were conducted through semi-structured interviews. The respondents were selected based on their knowledge about linked data visualization and their ability to give critical opinion on the topic. Therefore, the respondents are targeted to linked-data researchers. The aim of this step is to analyse how well the LOD Cloud represents datasets and the links between them in order to support exploration and integration purposes. In addition to the interview, review of previous studies and projects on linked data visualization were used to complete the analysis. It was analysed to understand which part of LOD Cloud should be improved to accommodate a better usage. 14 questions were formulated and grouped into five components. Five respondents that consisted of PhD students/ staff and linked data researchers were interviewed. The following questions were asked to the experts:

Obtaining an overview

1. What kind of information about the linked data can be obtained from this diagram? To what extent it gives you overview of linked data?
2. From your opinion, what additional functionality can be applied to improve the overview aspect?

Navigation and Exploratory Discovery

3. Can the lod-cloud interface help you navigate, observe, and explore the interconnections between datasets?
4. Is the node-link in the lod diagram biased due to the unsuitable representation to describe the amount of links? Do you think this issue will interfere the navigation and exploratory discovery aspects?
5. From your opinion, what additional functionality can be applied to improve the navigation and exploratory discovery aspect?

Presentation and Visualization

6. In the current LOD Cloud Diagram, the thickness lines of the diagram do not represent the amount of the links, there is no different representation between one and one hundred links. Do you think applying proportional line is suitable to represent the amount of link? Do you think it makes cluttered or not?

7. Line in LOD Diagram represent linkset. Linkset is the aggregation of several type of predicate on the instance level. Should the representation of each predicate type be separated instead of merged as a single linkset?
8. From your opinion, what additional functionality can be applied to improve the visualization aspect?
9. How should linked spatial data be represented in the LOD Cloud? Considering web map application as benchmark, for instance browser.linkedgeodata.org map browser or DBPedia lodlive map browser.

Information Retrieval

10. Does the datahub.io dashboard have sufficient element to support the process of finding the important part of the dataset that you are interested in?
11. Based on your experience in exploring the LOD dataset diagram that refers to the datahub.io portal, is CKAN catalog metadata sufficient to describe the linked spatial data? *comparing to GeoDCAT-AP (for further information open Annex.1 page.39)
12. In practice, the data owner usually does not complete the metadata vocabulary fields that leads to unclear data overview. From my observation it potentially discourages the consumption of the linked spatial data. For instance, there is no bounding box information or spatial resolution, hence the user does not know which location is covered by this linked spatial data. From your opinion, what is the kind of approach to overcome this problem?
13. What additional functionality can be applied to improve the Information Retrieval aspect?

Data Verification

14. Dataset quality (metadata and data structure) information have not yet provided either in diagram or data portal datahub.io. In your opinion, what is the best option to visualize or provide information about dataset quality? *Consider using https://www.w3.org/TR/vocab-dqv/#DimensionsofZaveri as the standard.

3.2.2. Interview feedback

The interview results revealed several limitations as well as the advantages of LOD Cloud Diagram. The opinions from expert tend to be domain dependent meaning the focus of assessment and suggestion differs from each expert based on their specific expertise in the linked data. For example, linked data developers tend to discuss linked data structure, information retrieval, SPARQL, and data analysis. Meanwhile, experts of graph visualization or linked data profiling prefer discussions on other approach that can be used to better visualize the large LD datasets. On the other hand, respondents who identified themselves as linked data scientists are more critical to the data verification and data quality. Nevertheless, all the opinions on each component are related to the knowledge discovery and exploration.

The first component: obtaining overview. Most of the respondents agreed that LOD Cloud Diagram visualizations succeed to provide insight into how current available LD datasets are linked to each other. The respondent argued that it depicts pattern of linked datasets and the availability of open data from each knowledge domain. Based on the diagram, the life-sciences domain is observed as the leading knowledge domain that successfully promotes and encourages the data provider to publish their resources as linked data. Regarding the overview depiction and datasets summarization, the experts argued that there is no need to put thousands of datasets in one interface, even more in the

landing page (homepage). One of respondents said that *"It's a nice idea to start with couples of datasets categorization and provide functionality to expand"*. Other respondent mentioned that the linked data profiling provides sensibility to understand the high-level categorization of datasets and assist in harnessing the datasets. The respondents also deliver substantial suggestion to improve overview component. For instance, the provision of summarization of available datasets and some of ideas to provide filtering functionality to refine the datasets, for instance based on data domain or other parameters.

Regarding **the second component, navigation and exploratory discovery**, the most critical option was asserted by one respondent who mentioned that LOD Cloud Diagram could not be used for real analysis and exploration. This is contrary to other respondents who argued that the diagram could assist user to explore and find new connections based on existing relationship between datasets. In addition, the respondent mentioned that *"without a starting point it won't provide much assistant in finding a new dataset related to a topic"*. The respondent stated that they will start the exploration from the centroid, which is presumed to contain a complete resource or knowledge. Centroid, possibly the biggest nodes of the graphs, attracts user attention and prompts user to do exploration. A famous dataset like DBPedia and Freebase commonly start off as the starting point of the data exploration. This diagram only provides visualization of top-level links with static snapshot. In such a way, the navigation component does not really apply to this diagram. A line in the diagram is the linkset that represent the aggregation of several type and number of links. It is not sufficient to give the detailed information and provide a navigation to the ROIs datasets. The limitation also exists in showing relationship, which is restricted to the first degree of relationship only. Several users are seemingly interested to observe the second degree of relationship which explains links between two datasets through other intermediaries. These intermediaries are important in the web of data which act as data hubs. These hubs have a role to bridge the of unconnected knowledge from cross-domain. However, the current network (linked nodes in the diagram) did not show extensive interconnection of multi-domain. For instance, government datasets are not well-connected either to the social networking or linguistics.

The third component: Presentation & Visualization. In this component the expert particularly discussed how the visualization should improve the description of link relationship. Considering the various number of predicates types, *"… visualize all types of predicate won't become a good approach"* was stated by one respondent. The segregation of linkset into predicates would increase the degree of clutter in the visualization. Furthermore, not all the predicate types are important for the user. Commonly, only a few of predicates are considered by users. Therefore, it is a useful approach to visualize the most-used predicates that link two datasets. There is a strong relationship between the second and third component, an efficient discovery and exploration comes with sufficient and intuitive visualization. Thus, the feedbacks tend to discuss both these two components simultaneously.

The substantial suggestions are particularly directed to link or relationship discovery. One of the respondents mentioned *"Filtering to the intended tree relationship and the list of datasets that are included in 1st degree relationship, for further analysis"*. In the same way, other respondents expected the functionality to redraw the overview based on any definitions of linkage. Apart from relationship issue, the suggestions from experts are quite diverse, for instance functionality to support multilevel exploration that bridge discovery from overview to detail. One of respondent suggest a precise solution that is *"Grouping the datasets and visualizing using collapsible graphs so a user can drill down to a dataset"*. The others suggest to provides subgraph selection, search and find query. The issue of data scope was also discussed, the respondent who has a background in geospatial data argue that the information about the scope of data, whether global, regional, or national, are needed for a spatial discovery and to simplify the exploration.

The fourth component: the information retrieval. Each node in the diagram are interlinked with the datasets metadata in the datahub.io. Regarding information retrieval, the expert argued it is case-dependant. For providing basic information to obtain datasets insight, the data & resources elements that are displayed in the datahub.io interface are sufficient. For instance, address of SPARQL endpoint, RDF dump URL, and example resource. Regarding the metadata availability, datahub.io adopts the vocabularies of DCAT to define the metadata. Only a small number of metadata are displayed in the datahub.io interface, for example the author, the datasets timeliness (created and updated), and number of links. This metadata is filled by the data provider manually, therefore it contains a considerable number of inconsistencies and availability issues. While high-quality datasets such as European Nature Information System or FAO Geopolitical Ontology provide sufficient metadata, other datasets that have lower quality do not provide sufficient metadata. In DCAT document there are several recommendations to fill the metadata elements, mandatory (M), conditional (C), and optional (O). One of the respondents that identified himself as a linked data developer mentioned *"The challenges is on how the data catalogue (datahub.io) can build a system that can automatically examine and report the metadata without any involvement of data owner or at least don't require a manual work to fill the metadata field"*. The minimum subset for metadata submission would make the datasets more reusable, at least the mandatory metadata. The challenge is to implement those recommendations into practices. Apart from metadata, tags features in the data catalogue must be utilized to support discovery. The following issue discussed the quality of tags which derives either from crowdsourcing or formal (authoritative) process.

The fifth component: data verification. This component is closely related to the data quality. As discussed in the Section 2.2, high quality of linked data will encourage other parties to reuse the data. However, the low availability of quality information discourages further linked data consumption. One of the respondents argued that *"The big problem is that the dataset publishers often don't have that information. We would need to provide a tool that collects this information and dataset publisher can publish the results"*. Apart from the mechanism to provide automatic quality information, the data quality framework itself is subject to debate. *"Data quality is case-dependent, fitness for use, and you should know the application upfront"* was mentioned by one respondent. As there are various linked data quality frameworks, the linked data community should agree on which standard to be implemented. Since the elements of linked data quality framework are very diverse, it is challenging to interpret the quality element for use-case. One respondent argued that vocabulary of VoID and DCAT is more than sufficient to denote simple quality element, for instance to support mainstream goal like linking resources. Data Quality Vocabulary (W3C, 2016a) also provides complete vocabularies which describe various dimension and metrics. Additionally, one respondent argued that specific data quality vocabularies will attract user to exploit the linked data resources. For data owners, this encourages them to improve the datasets quality continuously.

In the interview session, we also put several questions on how to better depict spatial datasets. Most of the respondents agreed that there is a need to separate interface of visualization for spatial data. Map becomes the first-place to describe extent of spatial data. However, the current datahub.io – CKAN functionalities does not provide spatial data description functionality. The reason is the datahub.io has not implemented the GeoDCAT metadata (see Section 4.1.1), therefore spatial datasets is more difficult to be discovered and explored. One of the respondents mentioned that a snippet bounding box is required to describe extent of spatial data. Regardless of not all data has spatial properties and spatial does not necessarily mean geographical, map is not the only solution. Several substantial suggestions asserted by the respondents regarding spatial data description, for instance, decoding resources content string into a plotted place name. Other solution could be applied is to provides functionality to describe scope of data (global, regional, national). Other facet could be

defined and applied in appropriate visual encoding or visualization techniques to better visualize spatial-related datasets. The opinion summary are listed in the Table 3-2.

Table 3-2. Summary of Expert Opinion regarding LOD Cloud Diagram with respect to Linked Data Consumption Requirement

Component	Opinion
Obtaining an overview	- Providing the summary of available datasets - Filtering functionality to refine the datasets based on data domain or other parameters.
Navigation & Exploratory Discovery	- Multilevel exploration that bridge discovery from overview to detail - Subgraph selection - Search and find query. - Information about scope of data (global, regional, or national)
Presentation and Visualization	- Use collapsible graphs to group the datasets, for instance: (http://mbostock.github.io/d3/talk/20111116/force-collapsible.html) - List datasets in certain tree relationship
Information Retrieval	- Provide a system that can automatically examine and report the metadata without any involvement of data owner or at least do not require a manual input of the metadata. - Encourage to fill mandatory metadata by pushing the recommendation into practices. - Utilized the tags functionality by making it formal or authoritative
Data Verification	- Agreed which standard will be implemented. - Provide mechanism to provide automatic quality information

3.3. Dataset and Linkset Exploration and Discovery

As we discussed in Section 3.1 and 3.2 on shortcomings of available visualization system and tools which lack of overview insight, only a small number of visualizations provide relationship between datasets. This sub-section elaborates recent research about linked data visualization in the set level, both datasets and linkset. Firstly, we should understand the definition of dataset and linkset. Vocabularies of VoID and DCAT can be used as reference for dataset and linkset definition. Dataset is a collection of triples which is commonly published and maintained by a single provider (W3C, 2011) who ensures the use of common URI. Dataset also can be defined as a set of descriptions of certain entities, which often share a common URI prefix. In the VoID vocabulary, dataset is explained by `void:Dataset` that is used to state the whole triples within single dataset.

A single dataset comprises several parts which are termed as subset. The properties `void:Subset` is used to state the part of the dataset that has a significant number of differences. For instance, some part of the data can be accessed through RDF dump, while the others are accessed through SPARQL endpoint. Moreover, there might be different versions (publication dates) or different sources of data. In addition to dataset and subset concept, there is distribution concept which, hierarchically, can be mapped below the subset concept because it defines the specific condition of certain resources of a dataset. It can be the license, format, language, or resource access. The vocabulary of `dcat:distribution` is used to explain this concept. The last concept is linkset, the vocabulary `void:Linkset` commonly used to define number of predicate properties within two datasets. Linkset

is an aggregation of links in the dataset, subset or distribution level which depicts the relation between them.

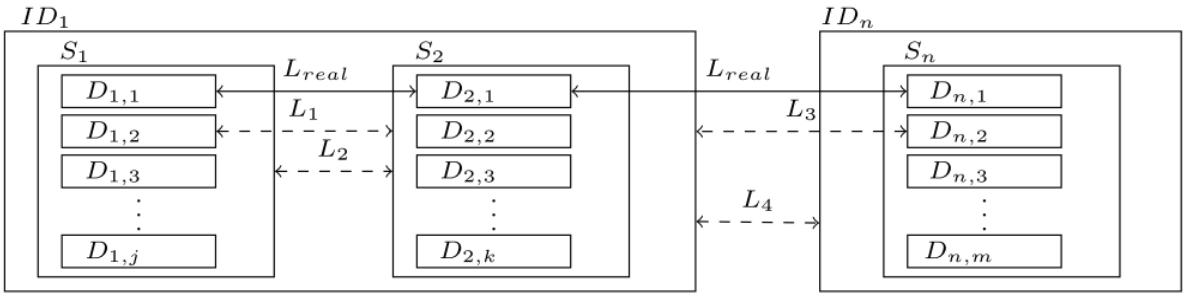


Figure 3-1. Different granularity of the linkset, obtained from Neto et al. (2016)

Figure 3-1 depict the abstraction concepts of linkset in level of dataset, subset, and distribution. These three concepts could assist to identify the link between resources. They could be aggregated to the upper hierarchy using a structured way (Neto, Kontokostas, Hellmann, Müller, & Brümmer, 2016). This figure describes the different of the links granularity, with respect to respect to resource representation. IDn, Sn, and Dn represent datasets, subsets, and distributions respectively. The linkset is represented by Lreal, which count links between two distributions. The evidence of linked distribution is the intersection of subjects and objects from two corresponding distributions (Neto, Kontokostas, Hellmann, et al., 2016), for instance the link between distribution of $D_{1,1}$ and $D_{2,1}$. The distribution of $D_{1,1}$ consists of (s_a, p_a, o_a) and distribution of $D_{2,1}$ involves (s_b, p_b, o_b). The intersection exists if $o_a = s_b$, which means the subject from resource of distribution of $D_{1,1}$ is equal to the object from resource of the distribution of $D_{2,1}$. Based on this conceptualization, the linkset between subset or dataset (L_1 until L_4) can be easily calculated by aggregating the linkset of distribution within subset and datasets.

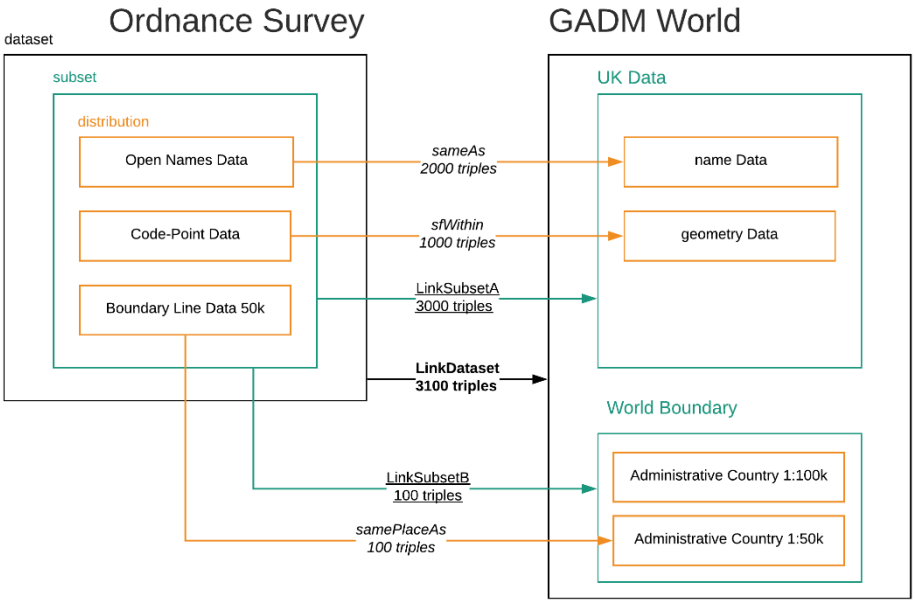


Figure 3-2. Different granularity of linkset between Ordnance Survey and GADM World dataset

As an example of illustration, Figure 3-2 depicts different granularity of linkset between Ordnance Survey (OS) and GADM World dataset. It is assumed GADM has two subset data and OS has one subset. Each subset comprises several distributions. Linkset in distribution level indicates by orange colour, linkset in subset level indicates by green colour, linkset in subset level indicates by black colour. Linkset in subset level aggregates linksets in distribution level and linkset in dataset level aggregates linkset subset A and B. This information can be used as basis to visualize the top-level relationship by extracting links information.

To assess the quantity of links between LD datasets, Neto et al. (2016) provided a workflow that consists of tuple splitting, bloom-filters fetching, and link extraction. To execute link extraction, an index that lists the subjects, predicates, object of a datasets are needed. LD-Lex, developed by Neto, Kontokostas, Publio, et al. (2016), is an architecture that has the functionality to stream datasets and make an index based on Bloom-Filter methods. In addition, this index can be used as a comparison for other datasets.

In another study, Milić, Veljković, & Stoimenov (2015) designed a workflow of an architecture called Linked Relation Architecture (LIRE) which aimed to assist user to discover, manage, integrate the relations between datasets. LIRE also encouraged linkset consumption of linked data on the web. The semantic relation between datasets were defined by considering 13 rules which elaborated components of datasets tags, datasets timeliness (creation, updated dates, etc.), resources number, number of view and the level of datasets index. Based on those 13 rules, one of the modules in the architecture can generate relationship of parent_of, child_of, links_from, and link_to. Even though there is a limitation in defining relationship option (predicates types), this architecture presented a solution to the undefined relation of a huge amount of distributed LD datasets on the web. Now the architecture is available as CKAN plugins.

Regarding linkset visualization, one of the comprehensive implementation is LODVader (Neto, Müller, Brümmer, Kontokostas, & Hellmann, 2016). LODVader combines the aspect of discovery, analytics, and visualization. It also gives more flexibility in the linked data visualization compared to LOD Cloud Diagram. The LOD Cloud Diagram defines Pay-Level Domain (PLD) as the basis for a dataset. Hence, LOD Cloud Diagram assumes that every distribution sits in the same PLD. In contrast, LODVader compares every single resource (subject and object) for certain datasets. As it explores the property of `void:Linkset`, it allows describing and visualizing the predicate types. The example of visualization generated from LODVader is provided in Figure 3-3. The figure depicts the relationship between datasets, and the coloured-background of nodes represents the subsets and distribution within datasets. Each link of subsets or distribution is represented by different colour edges.

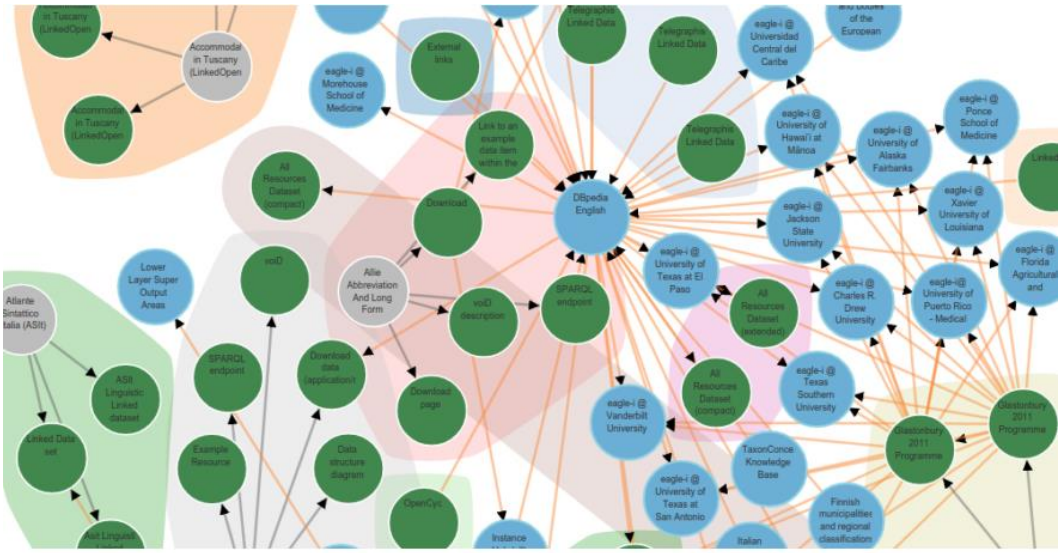


Figure 3-3. Visualization generated from LODVader architecture, obtained from Neto, Müller, et al. (2016)

3.4. Visualization for Linked Spatial Data

In this sub-section, we further elaborate the domain-specific visualization system discussion which previously introduced in Section 3.1. Linked spatial data is a domain-specific data in linked data which is characterized by the existence of geospatial ontologies – vocabularies in the resources. This specific type of vocabularies should be visualized in a meaningful way. Although map is still the most appropriate method to describe the extent of spatial data, several systems have been developed to explore and discover linked spatial data by providing intuitive interface.

Cesium (Potnis & Durbha, 2016) is a javaScript library based on 3D globes. To visualize RDF data, the Cesium was integrated with jOWL to parse the ontology and the instances. JOWL is a javaScript semantic library used to navigate and visualize semantic web documents. The geometry or geospatial content that is indicated by geospatial ontology will be extracted, parsed, and rendered to Cesium. The geospatial vocabularies such as latitude, longitude, extent, bounding box will be mapped to the 3D globes. This extent can be filled by other values from various predicate mapped using colour intensity. In addition to span and zoom functionalities, Cesium has live user interaction respect to SPARQL-DL query.

Map4rdf (De León, Wisniewki, Villazón-Terrazas, & Corcho, 2012) is a faceted navigation interface that allows user to browse multiple explicit dimension (facets) and put selected facet in the visualization. The facets can be retrieved by passing SPARQL query to the triplestore. The map-based visualization is generated based on the instance of the selected facet. Map4rdf includes geometry and geospatial visualization through OpenStreetMap or Google Maps. It has been used for visualizing Spanish geospatial linked data, for instance GeoLinkedData and AEMET meteorological dataset.

Facete (Stadler, Martin, & Auer, 2014) is a client-side javaScript application that is connected to the server-side SPARQL endpoint. Facete provides faceted browsing for geospatial content of RDF data. However, exploring the geospatial content in the large RDF data is a challenge because in most cases the geospatial content is not directed to all resources or instances. Therefore, Facete provides workflow

to find related geospatial content of certain resources. To ensure a rigorous exploration, Facete provide three (panel) i.e.: Selection, Data, and Geographical. This step-by-step exploration assists users to select the value of facet and visualize in the map-based visualization.

SexTant (Nikolaou et al., 2015) is a comprehensive visualization system for linked spatial data. which also exhibits the capabilities to visualize temporal data. Sextant is a tool based on the web technologies and, therefore, can be applied to cross-platform. The front-end is developed based on javaScript Timemap library. To reach the full potentiality of SexTant to visualize spatio-temporal information, the architecture is simply integrated with Strabon (Kyzirakos, Karpathiotakis, & Koubarakis, 2012) which support storage and query of linked spatio-temporal data. The Sextant front-end renders data from StSPARQL and GeoSPARQL query. The main strength of Sextant is the capability to explore and visualize multi-layer spatial data from different SPARQL endpoints. Based on the spatial extent called by the queries, it is possible to visualize the temporal dimension of data using a combination of map and timeline simultaneously. Sextant is also built based on OGC standards. Thus, it has interoperable functionality to another GIS platform. Furthermore, Sextant allows creation, sharing, and collaborative activities to edit and combine linked spatial data and other format of spatial data (Nikolaou et al., 2015). Specifically, to support collaboration activities SexTant provides functionality to import other formats for instance GeoJSON or KML. Ontology manager was developed to transform other representation to OGC standards, and thus enables a query based on GeoSPARQL or StSPARQL. Figure 3-4 shows a visualization of spatio-temporal data using a combination of map and timeline. The datasets were queried from three different SPARQL endpoints.

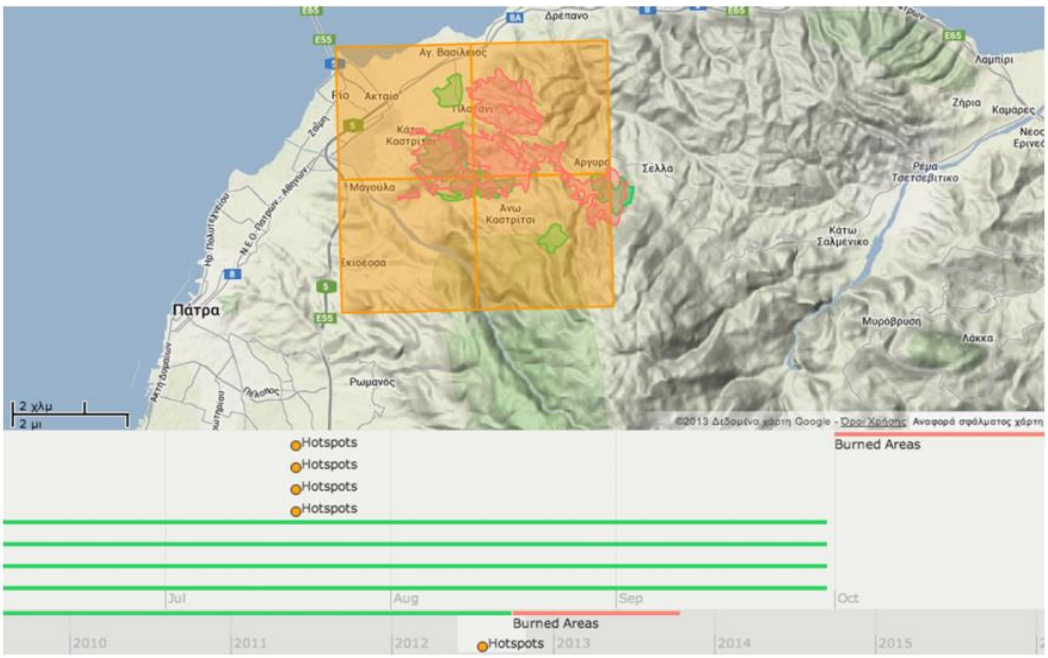


Figure 3-4. Linked spatio-temporal data visualization, retrieved from http://test.strabon.di.uoa.gr/sexTant/?map=mulrcpb74onu1smi_

Spacetime (Valsecchi & Ronchetti, 2014) is a web-based application which is particularly designed for DBPedia dataset. Spacetime exploits the rich-information of location and time from DBPedia dataset. It tried to solve the complexity of SPARQL query by designing user-friendly interface. The DBPedia Knowledge Base (DBKB) is presented in a graphical form. However, users can only write a simple string-query based on the targeted class, while SPARQL query is generated and compared to DBKB resource. To ensure that the resource includes spatial and temporal components, the system only lists the resource that contains both spatial and temporal information in the discovery form. This selection causes a decrease in the amount and scope of the data as well as the possibility for complex SPARQL query. On the other hand, this system compensates the shortcomings with the simplicity of knowledge discovery process.

DBpedia Atlas (Valsecchi, Abrate, Bacciu, Tesconi, & Marchetti, 2015) is one of the visualization systems that intends to tackle the overview issue of LD dataset. The "big picture" of dataset is often needed by users to understand dataset structure and the relationship between resources within a dataset. This insight functions as an entry point before detailed exploration. To achieve the overview, a map-like visualization has been designed to render instances of a dataset. It is obvious that not all resources contain spatial information, so a method called *"spatialization"* is required. This method aims to assign a position and shape for non-geometrical data to the 2D maps. DBPedia Atlas implements the Gosper-treemaps methods, and it generates 2D maps that visualizes the whole entities (resources) within one dataset and categorizes it based on hierarchical ontology class. One hexagonal tiling represents single resource within one ontology class. The visualization between resource is possible by connecting the tiling via edges. The ability to assist users to understand both overview and details is the main strength of DBPedia. Figure 3-5 presents the relationship between "Pink Floyd" resource of "Agent" Class to other resources intra-class and inter-class within one dataset. The info box on the right side supports the detailed information on the predicate type and the object resources.

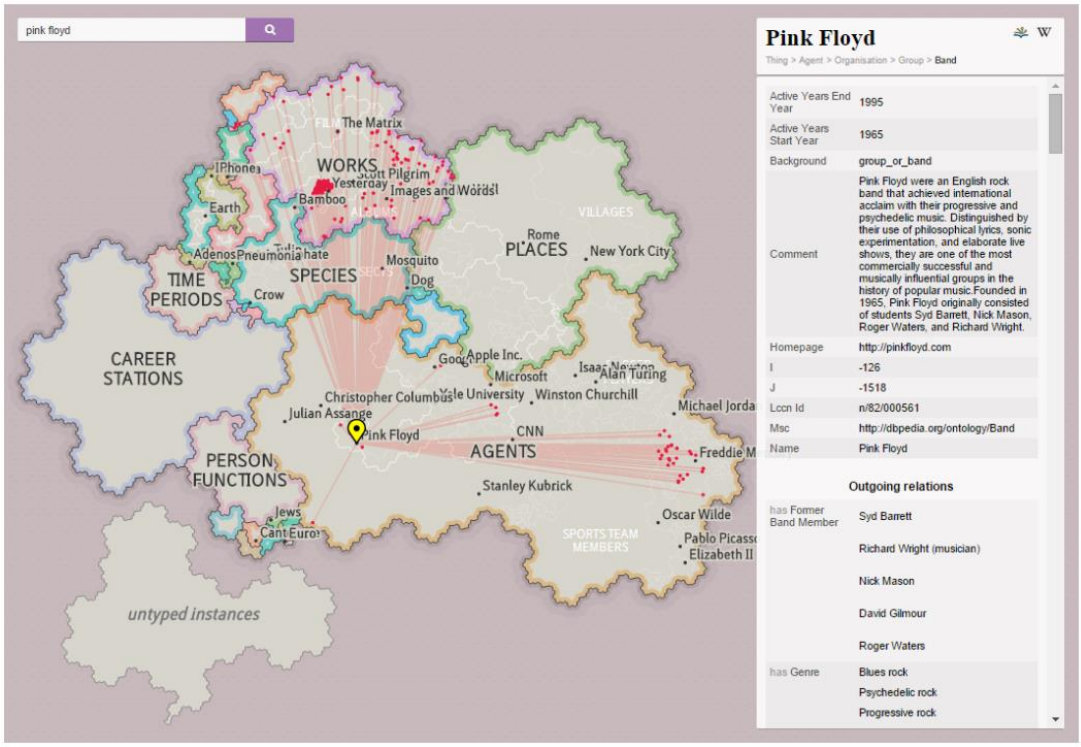


Figure 3-5. Relationship between instances of intra and inter ontology class in DBPedia Atlas, obtained from Valsecchi et al. (2015)

LinkedGeoData Browser (LGD Browser) (Stadler et al., 2012) is a dedicated web-based visualization tools for OSM linked data (LinkedGeoData). It aimed to show the structure of LinkedGeoData resources. The class and instances will be displayed to the interface based on selected region of the map. The browser provides filtering functionality by allowing user to select certain classes and render corresponding instance. The class and instance will update as selected region in the interface changes.

GeoYASGUI (GeoSPARQL Query Editor and Result Set Visualizer) (Beek, Folmer, Rietveld, & Walker, 2017) is a web-based SPARQL client, a development tools of YASGUI (Yet Another SPARQL Graphical User Interface) introduced by Rietveld & Hoekstra (2013, 2014, and 2017). Initially, it aimed to visualize the linked spatial data of Kadaster Netherlands. It combines SPARQL query writing functionality and various types of data visualization. YASGUI is a mature framework that is already widely-used, specifically by a number of publisher such as US Health Department (http://www.healthdata.gov/sparql), LOD Laundromat (http://lodlaundromat.org/sparql/), and Linked Open Vocabularies (http://lov.okfn.org/dataset/lov/sparql). GeoYASGUI extended previous library of YASR (see Section 3.1) to support visualization of parsed Well-Known Text (WKT) on the map. While previous tools do not support the GeoSPARQL query writing and evaluation, GeoYASGUI enable GeoSPARQL query component of core, topology vocabulary, geometry topology, RDFS entailment, and Query rewrite (see Section 4.3.2 for further explanation).

3.5. Summary

This chapter reviewed the concepts of linked data exploration and visualization systems for discovery purpose. We collected expert opinions on the current LOD Cloud diagram, which was assessed based on the key requirements for visual representation & analytics for linked data consumption. *How the LOD Cloud user interface can be improved for better support exploration activity* was one of the main topic being discussed, and we found that obtaining the overview is important element since it provides entry point to LD dataset discovery. Regarding the overview, we showed that the concept of linkset between dataset, subset, and distribution can be used to capture the concept of top-level relationship. Finally, we reviewed several spatial domain-specific visualization systems. The review showed different approaches of tools to visualize spatial component and relationship. The advantages and limitations of each visualization solution were also provided.

4. DESIGNING A WORKFLOW FOR LINKED SPATIAL DATA INTEGRATION

This chapter describes the linked data integration workflow. An integration means materializing or establishing links between two resources. The most important component in the data integration process is discoverability, which focuses on how to relate two resources based on a certain similarity concept or relationship. Geospatial content in the resources can be an important "*hook*" to relate two different resources. However, the heterogeneity of geospatial content causes difficulty in the discovery process.

Section 4.1 aims to discuss to what extent available metadata can be used for representing spatial data on the web and what the role of data model for query purposes is. Section 4.2 discusses the importance of data integration and the state of the art technology in the field of spatial data integration. Subsequently, Section 4.3 describes the experiment of linked spatial data integration to the LOD Cloud. The main contribution of this thesis is the case studies provision of integrating various spatial data sources.

4.1. Standards for Spatial Data on the Web

Before going through to the spatial data integration, firstly, we discuss an overview of spatial data on the web. Historically, World Wide Web and Spatial Data (Geographic Information System) were developed separately, in which each domain defines their own standards. These two domains must be integrated in a sensible way to ensure the ease in publishing and using spatial data on the web. The pioneer effort was coined by W3C and OGC, who formed a working group named The Spatial Data on the Web Working Group (SDWWG), which was responsible for clarifying and improving standards related to issues of spatial data on the web (W3C & OGC, 2015). Common standards on spatial data encoding, spatial metadata, and spatial relations facilitate spatial data discoverability on the web. The end-goal of standards creation is a spatial data integration which allowed spatial data on the web to be linked to each other. For data integration purposes, linked data play a major role since it provides easy mechanism for integration. Taylor & Parsons (2015) stated that the Working Group responsible for assessing Linked Data practices both on the web and OGC community then proposed best practice recommendations to support the goal of Geospatial Semantic Web.

4.1.1. Spatial Aspect for Metadata

From the list of Spatial Data on the Web Best Practices Recommendations (W3C & OGC, 2017), the issues related to geospatial vocabulary of (linked) spatial data in instance level have been discussed in (see Section 2.5). In this section, we aim to elaborate the spatial aspect of dataset metadata. As mentioned earlier, the focus of this thesis is to discuss the discovery of linked spatial data, and thus, spatial aspect of metadata plays a main role in discovery issue since it simplifies the spatial discovery and data reuse. The spatial aspect of metadata enables spatial query in the data catalogue, in which can be an important basis to discover and integrate different datasets. This helps users to understand the data extent and assist them to decide which dataset should be used based on their application.

Reyna, Simoes, & Genuchten (2016) identified four domain which are likely to have spatial data on the web. Each domain has own their specifications and vocabularies of metadata to describe the spatial data (see Table 4-1). Due to this condition, these domains tend to be disintegrated with each other in describing geospatial datasets, dataset series, and services. As previously discussed, this heterogeneity

complicates the discovery and consumption of spatial data on the web. Interoperability of metadata is needed to support spatial dataset discovery. Commonly, spatial datasets are discovered through data portal (data catalogue). To facilitate interoperability between data catalogue, a standard for publishing dataset metadata was developed. DCAT is a RDF vocabulary to describe datasets in data catalogues (W3C, 2014) which aims to enable metadata sharing across domains and catalogue platforms.

Table 4-1. Each specification from related domain for describing spatial data on the web.

Domain	Data Specification
Spatial	ISO 19115
Open Data	DCAT
Linked Data	VoID
Search Engines	Schema.org

Regarding the spatial content, GeoDCAT-AP has been developed to overcome the limitations of DCAT capabilities in describing certain characteristics of spatial datasets (ISA GeoDCAT-AP Working Group, 2016). GeoDCAT as a geospatial extension of DCAT aims to provide additional RDF syntax binding for INSPIRE metadata schema and the core profile of ISO 19115 (Van Den Brink et al., 2017). Several important metadata elements covered by GeoDCAT-AP are spatial coverage (bounding box), spatial representation, spatial resolution, and coordinate reference system. The following issue of exposing spatial data on the web is on how to make metadata index-able by search engine, which can be resolved by mapping and structuring the spatial data to the schema.org mark-up (Van Den Brink & De Visser, 2016).

The spatial aspect of metadata provides significant benefits in discovery activity by making spatial datasets, dataset series, and service more searchable across data platform. It also assists user to obtain sufficient spatial overview for their intended application. In the linked data context, finding candidate datasets to be linked is challenging. However, it can be overcome with proper organization of spatial aspect of metadata, at least the mandatory element which can improve the discoverability of dataset and reusability of the datasets. The next step after finding candidate datasets is establishment of semantic links, which can be done by materializing the relationship in the resource level. Up to this level, it is imperative to explain the data model of geospatial information to be able to represent and inquire into integration purposes. The next subsection elaborates the data model for spatial data.

4.1.2. Data Model for Spatial Data in the Semantic Web

Spatial data could be represented in various data types, such as toponym and geometry. In the context of semantic web, it has to deal with different parties which developed vocabularies to represent spatial data, for instance NeoGeo Vocabulary, ISA Programme Location Core Vocabulary, W3C Geospatial Vocabulary, OGC GeoSPARQL, stRDF, and schema.org. The role of vocabulary is to define basic semantics for various spatial concept, for instance class and subclasses, properties, basic datatypes for geometry, and geometry representation. These spatial vocabularies are adopted by data providers to describe their resources.

We examined these vocabularies differences by first comparing the top-level classes of spatial data model (see Figure 4-1). The comparison is applied to W3C Geospatial Vocabulary (`prefix:w3cgeo`) and OGC GeoSPARQL (`prefix:geosparql`). The top-level class in W3C Geospatial Vocabulary is *"Spatial Thing"*. This `w3cgeo:SpatialThing` class has properties of `w3cgeo:long`, `w3cgeo:lat`, and `w3cgeo:alt`. Commonly, geometry component is defined as property of Spatial Thing like W3C

Geospatial Vocabulary describes longitude and latitude. On the other hand, OGC GeoSPARQL define their top-level class by Spatial Object. This `geosparql:SpatialObject` has two subclasses i.e. `geosparql:Feature` and `geosparql:geometry`. It means that geometry is not the property of the top-level class but comes as a disjoint class.

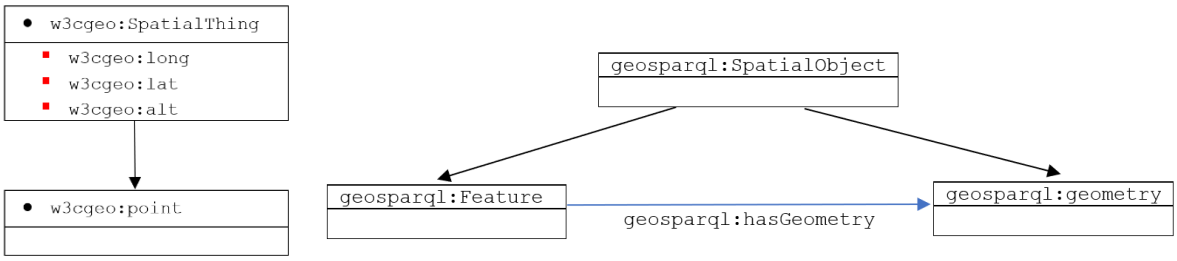


Figure 4-1. The top-level class from W3C Geospatial Vocabulary and OGC GeoSPARQL

These differences lead to confusion and question in data exploration, for instance, *"Which data class carries geometry information?"*. This is a typical question that usually arises when they want to find literal resource of geometry for integration process. There are other potential problems, for instance difference in coordinate reference system and the geometry serialization, which could hinder the interoperability. If we come across the datasets, each data provider builds their own ontology to describe the resources. This is affected by the absence of single vocabulary that can describe a whole spatial data scenario. Contrarily, the linked data principle recommends that the existing spatial vocabulary be re-used as much as possible. A commonly used spatial vocabulary is important as reference since there is no agreement of standard vocabularies to describe spatial data in RDF format as of yet.

To overcome this issue, Van Den Brink et al. (2017) argued that the possible approach is to update OGC GeoSPARQL vocabularies which should have the following capabilities: 1) bridging geometry and non-geometry spatial data, 2) bridging W3C and OGC standards, 3) aligning with the ISO abstract model, and 4) reusing the existing vocabularies like GeoRSS, NeoGeo, ISA Core Location Vocabularies. This update is required as GeoSPARQL only defines core set of classes, properties, and datatypes for query purposes, and does not provide a comprehensive vocabulary for representing spatial information (OGC, 2012). The development of other vocabularies to describe spatial data are both encouraged and recommended.

However, apart from vocabulary development issue, since 2012 GeoSPARQL vocabulary has been proposed and accepted as OGC standard for SPARQL-based query for spatial data in the semantic web (RDF). It has been widely-used in alignment with the application and system development to query data. It has several components that allow qualitative reasoning of spatial data. Based on OGC (2012b) and Koubarakis, Karpathiotakis, Kyzirakos, Nikolaou, & Sioutis (2012), the components are:

1) **Core**. This component defines top-level spatial vocabularies. Two classes are defined: `geosparql:SpatialObject` to define any spatial representation instances, and `geosparql:Feature` to define superclass of feature classes.

2) **Geometry Extension.** This component defines vocabularies for geometry data query and assertion. The superclass for all geometry classes is `geosparql:geometry`, which defines

properties such as spatial dimension (`geosparql:spatialDimension`), feature with geometry association (`geosparql:spatialDimension`), geometry with literal representation association. WKT (`geosparql:asWKT`), and GML (`geosparql:asGML`) are introduced as the serialization standards for the literal representation of geometry data. This component also provides non-topological operation for geometry data such as `geof:intersection`, `geof:buffer`, etc.

3) **Topology Vocabulary Extension.** This component defines vocabularies for querying topological relation between geometry object. It supports topological relations of Simple Feature of ISO 119125 RRC-8, and Egenhofer Framework, for instance, `geosparql:sfWithin`, and `geosparql:sfContains`. These vocabularies can be asserted in the RDF graph and used in the query.

4) **Geometry Topology Extension.** This component provides Boolean functions of topological relationship with respect to topological vocabulary extension (Simple Feature of ISO 119125 RRC-8, and Egenhofer Framework). These functions are used to check whether certain topological relationships hold between given spatial objects. For instance, Egenhofer query functions to check whether two spatial objects disjoint or not by using `geof:ehDisjoint`.

5) **The RDFS Entailment Extension.** This component enables RDFS standard reasoning for GeoSPARQL classes and properties. It also provides a mechanism for realizing the RDFS entailments that follow geometry class hierarchies based on WKT and GML standards.

6) **Query Rewrite Extension.** This component allows the transformation or translation of qualitative topological information into quantitative data. Based on the translation, the topological relation of topological vocabulary extension can be derived by Boolean function of geometry topology extension (Koubarakis et al., 2012). RIF-rules is the standard for query rewriting.

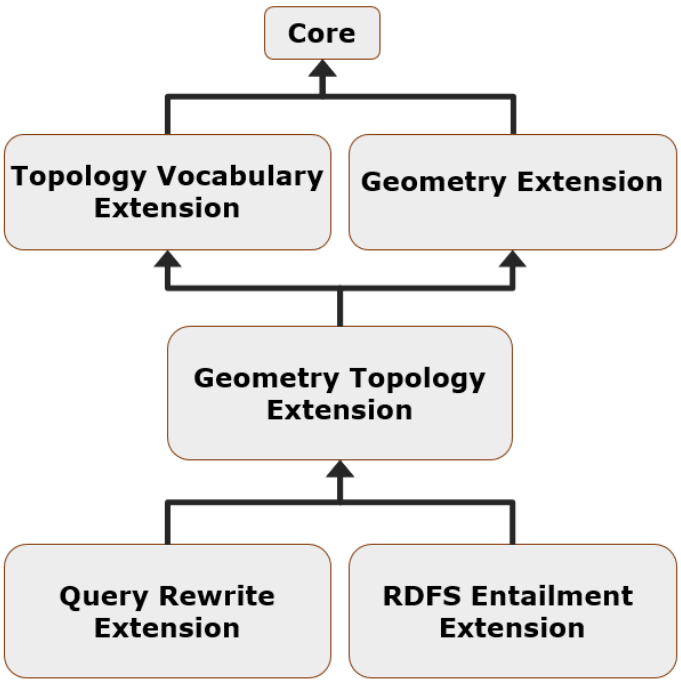


Figure 4-2. The components of OGC GeoSPARQL, adapted from OGC (2012b)

The OGC GeoSPARQL vocabulary has been used by multiple LD datasets (see Table 2-6). Hence, GeoSPARQL query is essential in finding intended LD resources for data integration purposes. The use of GeoSPARQL for spatial data integration in the Link Discovery activity will be elaborated in Section 4.3.3.

4.2. Linked Spatial Data Integration

Integration means to establish the link between resources. The integration in this section is the follow-up phase based on previous results and findings in this thesis**. First**, in Chapter 2, we found that most datasets did not integrate with other datasets (see Table 2-5), and instead only refer to their own resources. Therefore, it is essential to establish link to align with the fourth Linked Data principle, which *"Include links to other URIs, so that they can discover more things"*. In the linked spatial data context, storing the whole literal geometry data is rather costly and, therefore, it is better to link to the referenced or authorized spatial data on the web. Link to another spatial resource also provides wider data context and enriches the information.

Second, from the explanation of GeoSPARQL component in Section 4.1.2, we understand that topological vocabularies can be asserted into RDF triples. This is aligned with the findings of Van Den Brink et al. (2017), who stated that assertion of topological relationships is more preferable than performing geometry calculations for each query due to the high computational cost to perform geometry calculations. Similarly, Smeros & Koubarakis (2016) argued that a complex query to discover spatial relationships among heterogeneous data is not suitable for real-time purposes because of the high computation time. Regarding query writing, the complexity and expressivity of GeoSPARQL will make the formulation of the query difficult for lay-users. Considering these two reasons, designing workflow for establishment of link between linked spatial data is necessary.

4.2.1. Linked Spatial Data Integration Tools

Several tools have been developed by others to support linked spatial data integration. They focus on how to deal with spatial contents in the RDF data. Linked spatial data have been applied in the wider domain such as Earth Observation (EO) data or raster EO data. In this thesis, we focus on using GeoSPARQL data which only support literal geometry. For that reason, we limit the spatial content for vector data only. In this section, we present several solutions for certain parts of linked data lifecycle which are related to geospatial contents.

Conversion

Recently, two conversion tools that can handle geospatial content were developed to improve the shortcomings of the previous tools. First, **GeoTriples** (Kyzirakos, Vlachopoulos, Savva, Manegold, & Koubarakis, 2014) extends the R2RML mapping to take into account spatial content input data. This functionality was designed to deal with geospatial data specification. It allows the mapping to different geospatial vocabulary. GeoTriples presents a combination between R2RML and spatial RDBMS for RDF transformation purposes (R2RML discussed further in Section 4.3.1) and allows the processing of the input from raw vector file (ESRI shapefiles and XML, GML, KML, GeoJSON and CSV).

Second, **TripleGeo** (Patroumpas, Giannopoulos, & Athanasiou, 2014) is developed based on direct mapping of geometry2rdf library. It is developed to improve the specifications of geospatial domain. It has several distinctive features, for instance, recognizing multiple geometry data types (point, multi-linestrings, multipolygon, etc.), extracting thematic attribute that associated with each feature, allowing projection on-the-fly between several coordinate reference system (CRS), exporting to various serialization of triples (rdf/xml, rdf/xml-abbrev, n-triples, turtle, and n3), and allowing export to various geometry vocabularies (GeoSPARQL vocabulary, W3C Basic Geo Vocabulary and Virtuoso). Generally, it has the function to extract spatial contents from resources, transform them into triples serialization, and load the triples into intended RDF store. It accepts various geometries format as an input (Shapefile, KML, GML) and spatial RDBMS (MySQL, Oracle Spatial and Graph, PostGIS, and IBM DB2 with Spatial Extender).

Two linked spatial data conversion tools mentioned above (relational to database) have certain limitations. First, the conversion process that uses spatial RDBMS as input will convert each unique row that exist in the table to be a single subject of the triples. The geometry column of the table will be the object of the triples. Basically, the tools convert the whole resources from tables without any way to select intended resources. It might not be applicable for certain case. If the databases are sizable in volume and the data source is updated frequently, converting the whole data source from spatial RDBMS is not the best option. It could be a repetitive and costly process that requires large storage. In order to tackle these issues, Bereta & Koubarakis (2016) developed **Ontop-spatial**, which aims to create a virtual RDF view on top of spatial RDBMS by posing GeoSPARQL queries without explicitly converting the data. Ontop-spatial is a geospatial Ontology-Based Data Access (OBDA) system that has the ability to perform on-the-fly translation of GeoSPARQL-to-SQL queries using R2RML mappings. The advantages of this tool include the possibility to act as SPARQL endpoint and retrieve subset of LD resources that match with user's needs.

Storage

Strabon is a RDF store for linked spatio-temporal data (Kyzirakos et al., 2012). It supports GeoSPARQL component of core, geometry extension, and geometry topology extension to query static spatial data. In addition to GeoSPARQL, stSPARQL functionalities allow query on the temporal data. Strabon uses spatial RDBMS as back-end (PostGIS & MonetDB) which allows spatial and

temporal selection, join, and also selection of spatial functions and operators. Moreover, it also supports multiple Coordinate Reference System (CRS). Strabon is an open-source framework.

Discovery

Silk is a linked data integration framework (Volz, Bizer, Gaedke, & Kobilarov, 2009) which aims to establish explicit RDF links by discovering relationship between LD resources. It introduces Silk-LSL (Silk Link Specification Language) as a declarative language to specify link condition. Link condition refers to certain conditions and types of link that are intended to be established. Link condition applies several similarity metrics for discovering semantic relationship. However, the previous version did not consider spatial data. This limitation was responded by Smeros & Koubarakis (2016) by extending the framework with spatial and temporal functionality.

4.3. Workflow for Integration to LOD Cloud

In this section, we present two case studies to perform integration between LOD Cloud dataset and non-LOD Cloud dataset. LOD Cloud datasets was analysed by Workflow 1 and Workflow 2 in Section 2.6.1, while the analysis of non-LOD Cloud dataset (raw data) explained in the Workflow 3. The integration was done by determining the conditions and designing a workflow for adding and maintaining datasets in the LOD cloud. The first case study is an integration between spatial data that contain geometry or geometry-based integration. The second case study, we elaborated the integration based on the place name (toponym). The question that we intended to answer for second case study was how to exploit spatial non-geometry data that are uniquely identified by toponym and how to deal with different data models and vocabularies of toponym across datasets. The toponym-based integration becomes an alternative for linking data for spatial data that does not contain geometry. For data integration, we used Kadaster, Natura2000, and Geonames datasets. Figure 4-3 shows the linked spatial data integration workflow.

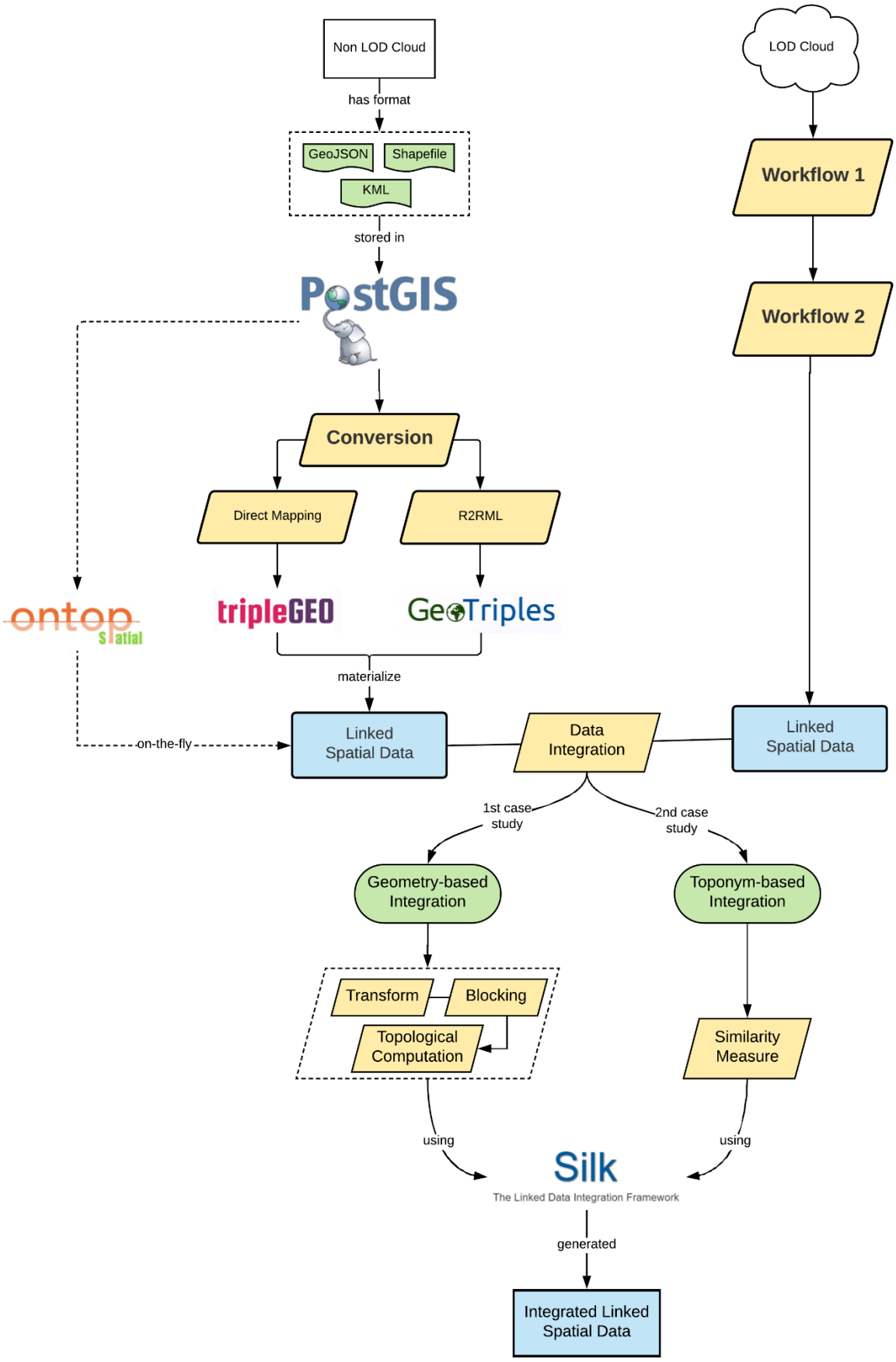


Figure 4-3. Workflow 3: Linked Spatial Data Integration

4.3.1. 1st case study: Geometry-based Integration

In this case study, we integrated two sources of spatial data. The first source is already available as linked data, while the second source is still in the raw format of spatial data. We explained step-by-step how to deal with these differences, including examining the ontology of dataset, transforming data to linked data format, identifying geometry vocabularies, posing relevant SPARQL query, and integrating resource based on geometry.

Dataset Sources.

Kadaster. Kadaster is an authoritative body for land registry data in the Netherlands. In June 2016, Kadaster released the *Basisregistratie Topografie* (BRT) and *Basisregistratie Kadaster* (BRK) datasets as linked data. The following year, they offered *Basisregistratie Adressen en Gebouwen* (BAG) as linked data. Kadaster provides SPARQL endpoint to exploit the dataset through https://data.pdok.nl/sparql. In this case, we were interested in BRT dataset specifically TOP10NL that is served in the linked data format. TOP10NL is the digital topographic base file of the Land Registry which covers the whole country of Netherlands. The TOP10NL object is indicated on Table 4-2. Each TOP10NL object owns several properties and subclasses. These subclasses inherit all the properties from the superclass.

Table 4-2. The TOP10NL objects

FuntioneelGebied (Functional Area)	*Gebouw* (Building)	*GeografischGebied* (Geographical Area)	*Hoogte* (Height)
Inrichtingselement (Fixture Element)	*Plaats* (Place)	*PlanTopografie* (Topographic Plan)	*RegistratiefGebied* (Administrative Area)
Relief (Relief)	*Spoorbaandeel* (Railway Section)	*Terrein* (Terrain)	*Waterdeel* (Water part)
Wegdeel (Road Section)			

Natura2000. The Natura 2000 dataset is a network of nature protection areas in European Union. It contains data on Habitat Directive and Birds Directive. We used Natura 2000 as a case study to transform raw spatial data (shapefile) to linked data format. The most recent version used as dataset was obtained from https://www.eea.europa.eu/data-and-maps/data/natura-8#tab-gis-data.

Data Story

Natura2000 dataset contains information about habitat directive of certain species. Based on EU habitat directive, one species of dragonfly called *Leucorrhinia pectoralis* is mentioned as a protected species. The statement is aligned with the environmental NGOS IUCN Red List 's finding (Kalkman, 2014), which asserts that the *Leucorrhinia pectoralis* is a threatened species in 2014. According to the law, the local authorities must monitor this species. The fact is *Leucorrhinia pectoralis* inhabits water bodies, specifically lakes. Water Directive Framework stated that *"the 'sense of urgency' areas, where water quality must be restored quickly, before 2016, otherwise natural values will be irretrievably lost"*. Based on this condition, information about species and water bodies should be integrated since Kadaster and Natura2000 are separated datasets.

Modelling

Data Preparation. The Natura2000 shapefile data covers the whole EU countries. As our focus was the Netherlands area, filtering is required to acquire ROI data. The attribute of Natura2000 data is comprehensive, including bioregion, designation status, directive species, habitat class, habitat, impact, management, natura2000 site, other species, and species. The selected attributes for integration purposes are natura2000 site, species, and the geometry data which is inherited from shapefile representation. This data is loaded into PostGIS database using shp2pgsql-gui. The variety of feature types in Natura2000 dataset needs to be described properly. Thus, we used OGC GeoSPARQL vocabularies to describe geometry data as it follows the OGC standard to describe basic feature of geometries. Since we applied OGC standard, thus we used the EPSG:4326 as CRS.

Ontology Selection. Based on data attribute selection, we determined vocabularies to describe those attributes. To describe general relation between resources we used RDF Schema vocabularies. Regarding spatial feature, we utilised OGC GeoSPARQL vocabularies because most linked data conversion tools support the OGC GeoSPARQL vocabulary (see Section 4.2.1). The Natura2000 site attribute is the subclass of spatial feature, for instance NL9803006 belongs to site upper class. NL9803006 has the geometry of multipolygon feature, which has Well-Known-Text (WKT) representation. The model of spatial data is indicated in Figure 4-4:

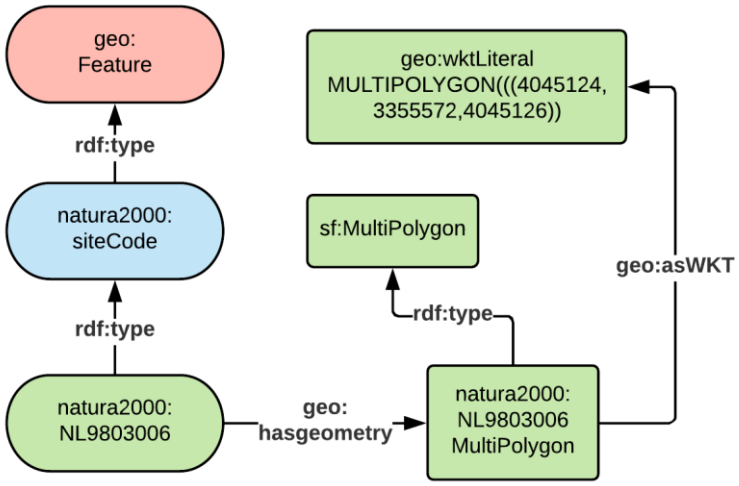


Figure 4-4. Spatial data modelled based on OGC GeoSPARQL vocabularies

On the other hand, *Basisregistratie Topografie* (BRT) data is available in linked data format and has been modelled with BRT ontology (https://brt.basisregistraties.overheid.nl/query/model - `prefix:brt`). Kadaster describe their geometry resource in two ways: first, in local coordinate EPSG:28992 (RD-New Projection) using PDOK ontology (asWKT-RD vocabulary), and second, in the WGS 84 (EPSG:4326) using OGC GeoSPARQL vocabulary (asWKT). The subset of spatial component in the BRT model is indicated in the Figure 4-5:

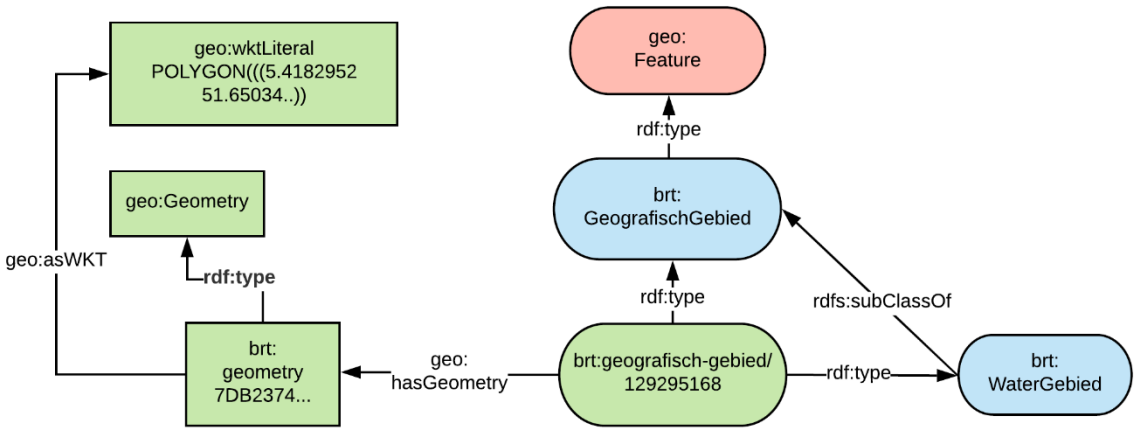


Figure 4-5. Spatial component in the BRT Model

Conversion

The conversion phase was aimed to transform the Natura2000 dataset from RDBMS to RDF triples. There are two recommendations from W3C to do the conversion from relational database to RDF or known as RDB2RDF.

1. Direct Mapping of Relational Data to RDF (https://www.w3.org/TR/rdb-direct-mapping/)
   A straightforward mapping that transforms the relational tables to classes. The fields are mapped to RDF properties or predicate. The primary key of the table is transformed to be a unique IRI of subject's resources and the values of the fields (foreign key) are transformed to be object resources. The mapping depends on the database schema. Thus, the output RDF represents the structure schema.

2. R2RML: RDB to RDF Mapping Language (https://www.w3.org/TR/r2rml/)
   R2RML is a language that expresses customization of mapping rule from RDB to RDF. It allows user to explicitly adjust the mapping by managing the expected output structure and choosing vocabulary to be used. R2RML consists of three components, i.e.: Logical Table to indicate the resources that intended to be mapped, SubjectMap to describe how to generate the subject, and PredicateObject Map to describe how to generate predicate and object. The output RDF represents the structure and ontology of user choice.

The conversion method selection depends on the expected output from user. In essence, the R2RML is quite useful for datasets that sit on complicated attribute from multi-tables. However, in this case, we already conducted data preparation from Natura2000 raw data and filtered the important attribute for integration purpose. In addition, there is only one attribute that would proceed to the geometry-based integration, which is geometry literal resources. Considering those reasons, we chose Direct Mapping methods for conversion. Regarding the automation tools of RDB2RDF, only few of the existing conversion tools support geospatial resources conversion. These tools was expected to be able to perform geometric extraction and coordinate transformation. The tools functionality comparison of geospatial features to RDF is presented in the Table 4-3:

Table 4-3. The functionality comparison of tools that support geospatial features

	Direct Mapping	R2RML	GeoSPARQL Compliance	RDBMS	Shapefile	GML	KML	Geo JSON
Geometry2RDF	√			√	√			
TripleGeo	√		√	√	√	√	√	
GeoTriples		√	√	√	√	√	√	√

Considering the format of datasets and direct mapping options, we chose TripleGeo to convert the Natura2000 dataset. We present the conversion process from input data, mapping rule, and the output in Figures 4-6 until Figure 4-9.

sitecode character varying(80)	sitename character varying(203)	speciesnam character varying(50)	geom geometry(MultiPolygon,4326)
NL2003029	Lonnekermeer	Leucorrhinia_pectoralis	
NL2003036	Oostelijke Vechtplassen	Leucorrhinia_pectoralis	
NL2003064	De Wieden	Leucorrhinia_pectoralis	
NL9801013	Weerribben	Leucorrhinia_pectoralis	
NL9801023	Veluwe	Leucorrhinia_pectoralis	
NL9801036	Leenderbos, Groote Heide & De Plateaux	Leucorrhinia_pectoralis	
NL9801071	Holtingerveld	Leucorrhinia_pectoralis	
NL9801080	Noordhollands Duinreservaat	Leucorrhinia_pectoralis	
NL9803006	Rottige Meenthe & Brandemeer	Leucorrhinia_pectoralis	

Figure 4-6. Spatial RDMS as input of mapping

The selected Natura2000 database consists of three fields: Site Code, Site Name, Species, and Geometry. Extract, Transform, and Load have been implemented to the spatial RDBMS and generated the triples in the rdf/xml serialization.

Subject	IRI formed from the concatenation of the base IRI (@prefix natura), table name (natura2000), primary key column name (sitecode) and primary key value (NL_2003029)
Predicate	IRI formed from the concatenation of the base IRI (@prefix natura), table name and the column name (sitename)
Object	RDF literals formed from the lexical form of the column value (Lonnekermeer)

Figure 4-7. Direct Mapping Rule

TripleGeo implemented direct mapping rule to convert spatial RDBMS to triples. The mapping result depends on the database structure, such as table name, column name, and primary keys.

```
@prefix rr:      <http://www.w3.org/ns/r2rml#> .
@prefix geo:     <http://www.opengis.net/ont/geosparql#> .
@prefix natura:    <http://www.http://natura2000.eea.europa.eu/Natura2000#

<RDB2RDFMapping>
a rr:TriplesMap;

rr:logicalTable [rr:tableName "Natura2000"];

rr:SubjectMap [
  rr:template "http://www.http://natura2000.eea.europa.eu/Natura2000/sitecode={sitecode}";
  rr:class < natura:natura2000>
];

rr:predicateObjectMap [
  rr:predicate <natura:sitename >
  rr:objectMap [rr:column "sitename"]
].
```

Figure 4-8. Mapping rule as R2RML

Even though TripleGeo does not require R2RML mapping since it adopts direct mapping methods, it is still important to understand the logic behind the automatic conversion. As explained before, R2RML is the customization expressivity of mapping rule with direct mapping rules as the basis of the customization.

```
@prefix geo:     <http://www.opengis.net/ont/geosparql#> .
@prefix natura:  <http://www.http://natura2000.eea.europa.eu/Natura2000#> .
@prefix rdf:     http://www.w3.org/1999/02/22-rdf-syntax-ns# .

natura:natura2000_NL2003029
    rdf:type   natura:natura2000 ;
    http://www.w3.org/2000/01/rdf-schema#label "NL2003029"@en ;
    natura:sitename "Lonnekermeer"@en ;
```

Figure 4-9. RDF triple as mapping output

The result of the conversion is the RDF file. The conversion tools can generate several formats such as Turtle, N-Triples, N3, and rdf/xml.

Discovery and Integration

We used Silk (see Section 4.2.1), an open-source linked data framework to discover links between resources from different data sources. Silk offers Silk Workbench (web application), Silk Single Machine, Silk MapReduce, and Silk Server. We used Silk Single Machine for the implementation of first case study. The discovery and integration process in Silk consist of several components that must be set, namely:

- **Data Access.**
  Silk allows data access both from local and remote endpoints by using ARQ, a SPARQL Processor for Apache Jena (open source Java framework for Semantic Web). Using this query engine then we can call the intended subset of resources. A proper SPARQL query writing is crucial to select targeted resource in order to reduce the computational cost. As explained in

the data story section, we intended to retrieve water bodies (lake) geometries from BRT Kadaster data through remote SPARQL endpoint and the habitat directive geometries of *Leucorrhinia pectoralis* dragonflies through local endpoint. We present the SPARQL query for two datasets in the Listing 2 and 3.

Listing 2. Natura2000 Query

```
?s a
<http://www.http://natura2000.eea.europa.eu/
Natura2000#natura2000>; a
<http://www.http://natura2000.eea.europa.eu/
Natura2000#Leucorrhinia_pectoralis>.
 ?s
<http://www.opengis.net/ont/geosparql#hasGeo
metry> / geom geo:asWKT ?geom .
```

Listing 3. BRT Query

```
PREFIX rdf: <http://www.w3.org/1999/02/22-
rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-
schema#>
PREFIX brt:
<http://brt.basisregistraties.overheid.nl/de
f/top10nl#>
PREFIX geo:
<http://www.opengis.net/ont/geosparql#>
prefix n2k: <http://data.pdok.nl/natura-
2000/def/natura-2000#>

select *
WHERE {
  ?a a brt:Watergebied.
  ?a brt:geometrie ?geom .
  ?geom geo:asWKT ?o .
    }
```

- **Transformation and Blocking**

  The heterogeneity of spatial data makes spatial transformation important in order to standardize the data format and serialization. Silk allows transformations from several vocabularies to OGC GeoSPARQL vocabulary, literal serialization to WKT, and coordinate reference system to WGS 84. The other transformations are developed to handle the complexity of geometries, for instance simplification transformation and envelope transformation, or points-to-centroid transformation. In the conversion and modelling section, we managed to proceed the data into OGC GeoSPARQL standards. Thus, here we only used simplification transformation to make spatial relations computation more efficient.

  The blocking is a technique to divide earth surface into several blocks area. These blocks are inserted by given resource geometries indicated by minimum bounding box coordinate. These selected blocks area were compared. It is useful as pre-computation rather than comparing the whole surface. The block area is determined by the formula (Smeros, 2014) as follow:

$$block\ area = \left(\frac{1}{(sbf)^2}\right)^{deg^2}$$

where *sbf* is spatial blocking factor. The block area stated in square degree (deg^2). The number of blocks were obtained from dividing the whole surface area (41,253 deg^2) with given block area (deg^2). Based on Smeros & Koubarakis (2016) experiment, the optimum *sbf* is 10 since it gives the optimum result with respect to computation time and number of discovered link.

- **Link Condition**
  Since we focused on geometry data in this study case, only the spatial link condition was considered. Silk has the capability to compute spatial distance and to check topological relationship between two geometries.

All components mentioned above must be stated in the declarative language of Silk – Link Specification Language (Silk – LSL). It specifies the whole of linkage rule, for instance define data access, define metrics to compare between resources, specify the link type, specify the threshold, and specify the link limits. There are two ways to obtain this: first, by writing the specification in XML format, and second, by using Silk - graphical user interface (GUI) by drag-and-drop components then the workbench will generate the linkage rule for users. We used the first way to define linkage rule and Figure 4-10 illustrate the linkage rules on Silk - GUI. The full Silk – LSL for geometry-integration presented in the Listing 5 in the Appendix.

Figure 4-10 shows the intended resources from both datasets filtered by SPARQL query, we targeted the literal representation of geometry which is indicated by `geo:asWKT` vocabulary. We applied simplify transformer to complex geometry feature of Natura2000. Next, we set topological relationship of within to check whether the point feature of *Watergebied* class of Kadaster within the polygon feature of *Leucorrhinia pectoralis* species. Successively, we defined links of `geo:sfWithin` vocabulary to discover the relationship. Table 4-4 indicates the number of queried resources and discovered link between these two sources.

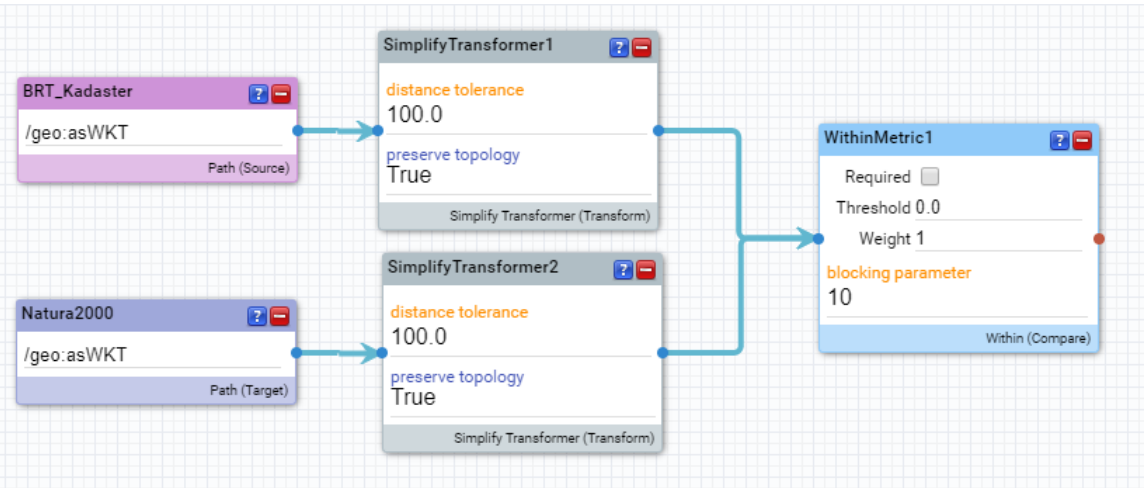


Figure 4-10. Linkage rule setting by Silk-GUI

The computation process based on geometry is time and memory consuming. Even after posing relevant SPARQL query to retrieve subset of resources and also setting the transformation and blocking, the computation time to execute the geometry comparison is still high. The process is prone to failure because the it takes excessive computation time, as indicated by the message of *"OutOfMemoryError: GC overhead limit exceeded"*. Apart from computation limitation, Workbench and Single Machine version only can retrieve string maximum 65 kilobytes per instance. Thus, the instance which contain complex geometry and encoded it into long string, it could not be processed directly. Thus, pre-processing is needed by applying simplified geometry function (ST_Simplify or ST_SimplifyPreserveTopology) in PostGIS. Therefore, based on the experiments, the Silk Workbench and Single Machine is not sufficient to perform comparison of large amount of geometry data. Silk version of MapReduce and Server should give better performance since they employ multi-thread and multi-machine parallelization for computation (Smeros & Koubarakis, 2016).

Table 4-4. Number of queried resources and discovered link between these two sources

Dataset	Number
Kadaster resources (*Watergebied* class)	15
Natura2000 resources (*Leucorrhinia pectoralis* habitat directive)	9
Links discovered	2

4.3.2. 2nd case study: Toponym-based Integration

We aimed to integrate two sources based on place name. The purpose of the integration based on string similarity was to deal with information incompleteness of certain dataset. Several conditions require toponym-based integration, for instance integration between national geospatial data and thematic data which does not contain any coordinates. Information trade-off can be achieved if the semantic relationship is established. Information complementation will be achieved by interlinking these datasets. The most important part in toponym-based integration is the ontology analysis. There are huge number of names duplication in different places and different hierarchy of administrative. In this section we explained how to deal with these challenges.

Dataset Sources.

Geonames. Geonames is open database of geographical data. It contains 11 millions of place name around the world. Geonames has an ontology to describe features in the datasets. It also offers the exonyms of toponyms. The linked data is served through both of RDF dump and SPARQL endpoint by third parties (http://factforge.net/sparql).

Kadaster. Apart from spatial data (see Section 4.3.1), Kadaster also offers toponym attribute that relates to spatial feature. The place name is the properties of certain class in the Kadaster dataset.

Data Stories

Since the early period of linked data, Geonames has offered the useful resources of place names around the world. Since then, Geonames became important data hubs and central of linked data network. The Geonames dataset is widely-used by users from across-countries since it offers exonym data (external name of geographic places from different language). Based on the huge number of incoming links, this dataset has become a reference for many linked data resources. Recently, Kadaster published their

resources as linked data format. Kadaster as an authoritative body ensures the quality and trustworthiness of the data, including place name. Kadaster also offers official ontology based on national division and cataloguing system. Hence, it is important to interlink the existing high-demand and widely-used dataset on the web with the official and most-updated data from national authorities.

Modelling

In the Kadaster BRT dataset, the toponym resides on the several properties, for instance *naamOfficieel, naamNL, naamFries,* and *naam*. Those properties belong to several classes in BRT datasets, i.e.: *Plaats, RegistratiefGebied, Gebouw, Inrichtingselement, PlanTopografie, Terrein, FuntioneelGebied, GeografischGebied* and *Waterdeel*. These classes have subclasses which inherit all the properties. In Table 4-5 below we present the available place name properties for each class.

Table 4-5. Availability of toponym properties in the BRT Kadaster class

Class	naamOfficieel	naamNL	naamFries	naam
Plaats	√	√	√	
RegistratiefGebied	√	√	√	
Gebouw				√
Inrichtingselement				√
PlanTopografie				√
Terrein				√
FuntioneelGebied		√	√	
GeografischGebied		√	√	
Waterdeel	√	√	√	

We executed simple SPARQL query to check the availability of the resources with respect to certain properties and class. We also checked the resources of *Plaats* class that have place name properties. The result showed 653 resources of *naamFries*, 9,898 resources of *naamNL,* and 2,604 resources of *naamOfficieel*. The other queries for different class showed similar trends. Generally, we observed that not all the *naamNL* properties in the resources class of *Plaats, RegistratiefGebied*, and *GeografischGebied* are transformed into exonyms or official toponyms that are recognized by wider international community *(naamOfficieel)*.

On the other hand, Geonames dataset consists of several properties that indicate toponym. Alternative names are available in several different languages. Considering the unavailability of other language, for instance Netherlands and Western Frisian, we only relied on English standard names. Geonames organize their resources based on feature class ontology, which include administrative boundary feature, hydrographic feature, area feature, populated place feature, road/railroad feature, spot feature, hypsographic feature, and vegetation feature. These super feature classes have several sub feature classes. The specific-countries ontology describes the Netherlands geographic features concept in detail (Geonames, 2017).

These two different ontologies from Kadaster and Geonames needed to be examined and analysed for integration purposes. The huge amount of toponym from these two resources should be queried properly in order to decrease the computational cost of link discovery process. Instead of querying the whole toponyms in one single query, we could specify the class and subclasses of targeted toponym properties. To achieve this, aligning the concept was required. We examined the definition of each

class in both ontology and present the example of simple alignment in the Figure 4-11 and Figure 4-12.

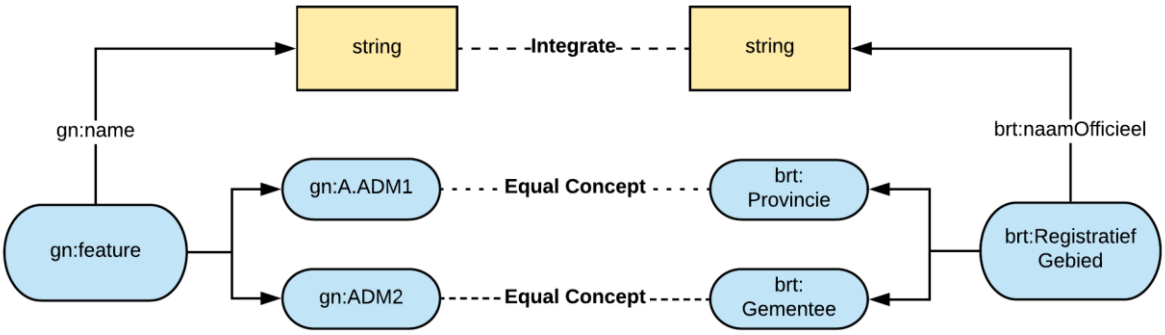


Figure 4-11. Equal concept of administrative area for toponym integration

The activity to make concept equal does not aim for ontology matching and mapping, but only to ease the SPARQL query writing by indicating certain class or properties of intended resources. Figure 4-11 compares the concept of administrative area in a rather straightforward manner, by defining equal concept between 1st and 2nd level of administration in the Netherlands from two different datasets. Figure 4-12 shows the concept of living area between both sources. The living area or settlement area concept can be found in the *Plaats* class in BRT Kadaster and in the data properties of *P.PPL feature code*. Class of *Plaats* is defined as *"Geographical area characterized by a concentration of building that used for living and work"*. While the *P.PPL feature code* is the abbreviation for populated place which is defined as *"a city, town, village, or other agglomeration of buildings where people live and work"*. Based on this definition, we can use targeted class and data properties to assist the correct SPARQL query construction

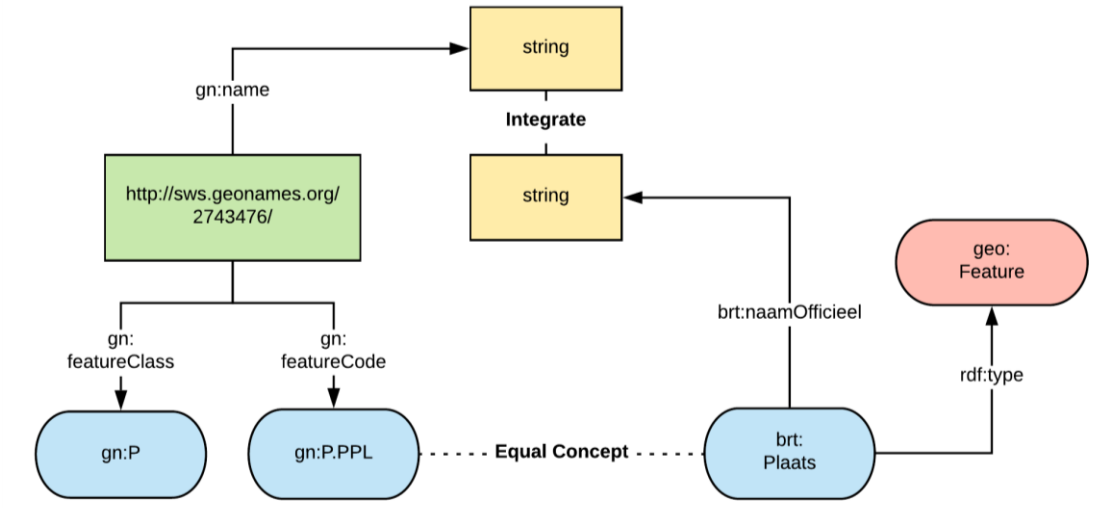


Figure 4-12. Equal concept of living area for toponym integration

To verify the query, the statistics of available feature in Geonames NL datasets (http://www.geonames.org/statistics/netherlands.html) can be used as the reference of comparison to the SPARQL query result. The correctness of SPARQL query writing and result is very fundamental because there are a lot of name duplication that exist in different level of class. Thus, we must ensure about targeted resources, otherwise the establishment of links might be inaccurate.

We presented one example of toponym duplication, a toponym named "Witteveen". The examination was conducted to check the existence of toponym "Witteveen" in both datasets. We posed simple SPARQL query in Kadaster data and Geonames data. Five different resources were found in the Kadaster data (see Table 4-6) and four in the Geonames data (see Table 4-7), all of them using "Witteveen" as toponym. This duplication might lead into ambiguities. Based on this fact, defining similar or equal concept between different ontologies is required in order to retrieve the exact targeted resources in certain class and establish true semantic relationship between them.

Table 4-6. Multiple existence of "Witteveen" toponym in Geonames dataset

Toponym	Feature Class	Location (Gementee)
Witteveen	Populated Place (PPL)	Midden-Drenthe
Witteveen	Populated Place (PPL)	Westerveld
Witteveen	Locality (LCTY)	Aa en Hunze
Witteveen	Locality (LCTY)	Covorden

Table 4-7. Multiple existence of "Witteveen" toponym in Kadaster data

Toponym	Class
Witteveen	FuntioneelGebied
Witteveen	Plaats
Witteveen	GeografischGebied
Witteveen	Waterdeel
Witteveen	Plaats

Discovery and Integration

We used Silk Workbench for the implementation of link discovery. The component of string-based integration are as follows:

- **Data Access.**
  Both datasets of Kadaster and Geonames were accessed through Silk Apache query engines which are connected to https://data.pdok.nl/sparql and http://factforge.net/sparql. As explained in the modelling section, SPARQL query is constructed based on identified equal concept between two ontologies. We aimed to establish link of two concepts, administrative area and living area as explained in the modelling section. Listing 4 – 7 show the SPARQL queries of two datasets.

Listing 4. Living Area Query on Geonames

```
PREFIX gn:      <http://www.geonames.org/ontology#>

CONSTRUCT {?a gn:name ?x}
WHERE {
    ?a gn:featureCode gn:P.PPL ;
       gn:countryCode "NL" ;
       gn:name ?x .
    MINUS { ?a gn:featureClass gn:A .}
FILTER(strStarts(str(?a), "http://sws.geonames.org/"))
FILTER (!regex(?x, "Gemeente"))}
```

Listing 5. Living Area Query on Kadaster

```
PREFIX brt:
       <http://brt.basisregistraties.overheid.n
l/def/top10nl#>


SELECT *
WHERE  { ?a a brt:Plaats.
         ?a brt:naamNL ?b  }
```

Listing 6. Administrative Area Query on Geonames

```
PREFIX gn: <http://www.geonames.org/ontology#>

SELECT ?a ?x
WHERE {
       ?a gn:name ?x .
    ?a gn:featureCode gn:A.ADM2 .
    ?a  gn:countryCode "NL" .
    FILTER(strStarts(str(?a),
"http://sws.geonames.org/"))
      }
```

Listing 7. Administrative Area on BRT

```
PREFIX brt:
       <http://brt.basisregistraties.overheid.n
l/def/top10nl#>


SELECT *
WHERE  { ?a a brt:Gementee .
         ?a brt:naamNL ?b }
```

- **Link Conditions**

  Toponym component is string and, thus, all the link discovery was performed based on string similarity. Silk implemented similarity measure concept of character-based and token-based. The character-based measures make string comparison on the character level. Several measures are implemented by Silk, i.e.: jaro, jaroWinkler, levenshtein, levenshteinDistance. The token-based measure was designed to deal with character sequence issue and separate character by separator, for instance punctuation or apostrophes. Silk implements token-based measures of jaccard, dice, qGram (n-grams), cosine, and softjaccard. A considerable number of studies focusing on the best-performing techniques for identify string similarity, especially toponym matching, but they presented different results and recommendations since string

similarity is dataset-dependent (Recchia & Louwerse, 2013). Hence, the uniqueness of the languages across countries and the nature of the string of the toponyms should be considered.

Paris, Abadie, & Brando (2017) studied the link between spatial named entities of historical text and the toponyms in the LOD Cloud. The study was performed on French language based dataset and they stated that cosine measure is the most suitable approach to find the similarity. Recchia & Louwerse (2013) conducted exhaustive works to find the best-suite measures by evaluating different countries dataset. They argued that for United Kingdom dataset, the skip-grams or bigrams are the most suitable. Another research by Santos, Murrieta-Flores, & Martins (2017) studied the combination of similarity measure to find effective toponym matching based on Geonames dataset. Based on the findings, they asserted that the best result measures as individual metrics is dice. Taking these studies into consideration, we chose dice as similarity measure since it provides the best result for Geonames dataset. Furthermore, token-based measure is also suitable to the Netherlands toponyms which contain apostrophe, for instance 't Schoor.

Besides defining the suitable techniques for toponym matching, we also determined the threshold value for optimal matching quality. In the matching process, we could define two values, namely accept and verify threshold values. Accept threshold means that the discovered links are acceptable and correct. The value below accepts threshold value until verify threshold value should be verified by the experts. Volz et al. (2009) used 0.7 and 0.9 for threshold value. As we intended to establish high-quality toponym link, we set high two threshold values. First, 100% equality, in this case we used equality measures in order to only materialize the perfectly similar string. Second, considering that we compare toponym in Dutch from Kadaster with the English version of Geonames, we slightly decreased the threshold to 0.9 using dice measures.

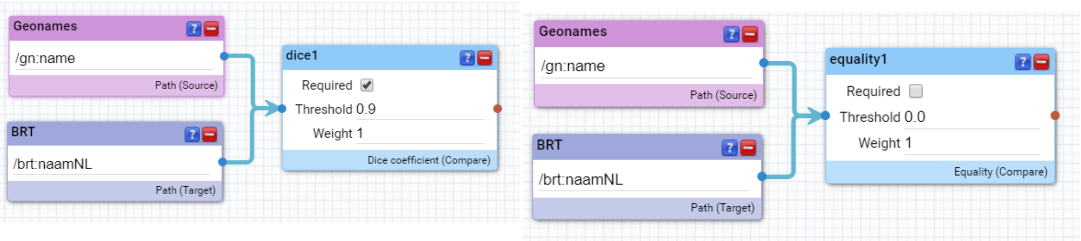


Figure 4-13. Dice measure (left) and Equality measure (right)

We declared these two components (data access and link condition) through Silk – Link Specification Language (Silk – LSL) by writing the XML file. The Link – LSL file presented in Listing 6 in the Appendix. Figure 4-13 and 4-14 illustrate the linkage rule. We also present the result of applying two different string-based similarity measures, Equality, and Dice, in the Table 4-8. Based on this result, these two measures give same result.

Table 4-8. Number of queried resources and discovered link between these two sources

Dataset	Number
Kadaster resources (*naam NL of Plaats class*)	9898
Geonames resources (*name of P.PPL feature code*)	6571
Links discovered by equality measure	8780
Links discovered by dice measure (threshold 0.9)	8780

We already presented the experiment of integrating linked spatial data based on geometry and toponyms. Geometry-based integration has to deal with high-cost computation, while string-based integration should handle the heterogeneity of ontology. Apart from individual testing of each option, the combination of these two options is also possible to more accurate links. Martins (2011) identified several ways to integrate spatial data, i.e.: using only toponym for string-based similarity, using only geospatial footprint for geometry-based calculation, and using combination of toponyms and geometry. In this thesis, we did not implement the combination methods due to the separate resources of geometry and toponym in the certain ontology dataset. For instance, Kadaster BRT ontology puts geospatial information in a separate resource. An example is http://brt.basisregistraties.overheid.nl/top10nl/id/geografisch-gebied/103059275 resource. This resource contains toponym but not the geometry information. The geometry literal is stored different resources(https://brt.basisregistraties.overheid.nl/top10nl/doc/geometry/8DB868524DA513F225806076C7B5AB05). The combination methods can only be implemented if both toponyms and geometry are stated in the one single RDF file.

4.4. Summary

This chapter summarises the linked spatial data integration. The workflow has been designed to deal with the heterogeneity of spatial (linked) data in terms of data format, access, and model. TripleGeo, a direct mapping implementation tool, was used to convert data from RDBMS to RDF format. Regarding the integration purposes, we used Silk for linking between resources of LOD Cloud. We presented two case studies of integration, geometry-based and toponym-based. In the geometry-based case, we showed that the standard vocabulary is needed to describe the spatial component and function. The experiment demonstrated that spatial transformation of the resources is the key to implement the link discovery. Data model is also important to represent data structure and assist to construct queries for integration purposes. In toponym-based case, we found that ontology or data model analysis have an important role to create accurate semantic relationship since we aimed to establish link of equal concept between two different data. We also discussed how the spatial component in the GeoDCAT dataset metadata provides significant benefits in discovery activity by making spatial datasets, dataset series, and service more searchable across data platform.

5. DISCUSSION, CONCLUSIONS AND RECOMMENDATIONS

5.1. Discussion

Metadata provision for better data (set) discovery and exploration

Due to lack of provenance of geo-enabled metadata in the current data catalogue, finding candidate spatial datasets and resources to be integrated was a rather challenging task. It was proven by our experiment accounted in Chapter 2, where we encountered difficulty in filtering datasets that fit our application. To solve the task at the resource level, we set the metrics of the existence of geospatial vocabularies in the dataset and examine all the resources in the candidate dataset using the data processing workflow that has been designed (Figure 2-8). Regarding the tasks at dataset level, GeoDCAT-AP provides spatial coverage (bounding box), spatial representation, spatial resolution, and coordinate reference system description. Moreover, for those spatial elements in the DCAT metadata can be queried, this leads to an ease of the dataset exploration and discovery. The European Data Portal (https://www.europeandataportal.eu/sparql-manager/en/) is an example of the query implementation platform on DCAT. This platform enables SPARQL query functionality to discover datasets across data catalogues in Europe.

Currently, the availability and completeness of Vocabulary of Interlinked Datasets (VoID) is very low. In fact, VoID as summary of resources information potentially provides a useful approach for the better data discovery, for instance, `void:Linkset` that describe relationship between datasets. It indicates the source and target dataset, number of incoming and outgoing link. The importance of VoID brought in the discussion in Chapter 3 on linkset concept in the level of dataset, subset, and distribution. Implementing this concept will provide a better metadata organization which subsequently delivers a structured way to create a graph of relationships between datasets (the overview of the dataset network). This information can be extracted and used for top-level relationship visualization. Since the VoID is stated in RDF vocabularies, it can be easily queried. The example implementation of the VoID query platform is RKB explorer (http://void.rkbexplorer.com/sparql/). Currently, the RKB explorer only allows to query the VoID within their repository. To deal with huge number of LD datasets on the web, the development of this kind of platform is required, for instance, to harvest the distributed VoID across data catalogues and query it.

Lastly, linked data quality is an issue in data exploration. Low availability even absence of data quality information in the data catalogue and metadata makes users difficult to judge the quality of the datasets. The automatic assessment tools of linked data quality, for instance Luzzu or LD Sniffer (see Section 2.3.1), should be integrated into the data catalogue. Thus, the data owner can publish and assess the quality of their resources simultaneously. Both implementation tools are developed to aligned with W3C DQV as explained in Chapter 2. The DQV provides a complete vocabulary to represent the linked data quality and the quality of the metadata. This effort is very useful and should be taken into consideration in practice. It certainly could assist users to find the high-quality dataset based on their application.

Ontology as the basis of data retrieval, consumption, and integration

We explored several datasets through their SPARQL endpoint and deployed a local machine with a SPARQL endpoint of RDF dump using python libraries. The SPARQL queries were constructed based on the data models of specific datasets, but the majority of them only provided the proper ontology documentation in the machine-readable format instead of a human-readable format. The documentation should depict information about class, data properties, object properties, relationship between class or relationship, and inherited properties in a meaningful way. In this study we can observe that Kadaster Netherlands provides a complete and easily comprehensible. In their ontology description, https://brt.basisregistraties.overheid.nl/query/model, they provide an interactive hierarchical ontology (intra-hyperlink-page) which allows users to explore the structure in a sensible way. In addition, graph visualization also illustrates the relationship between classes and properties.

An ontology is the main reference for SPARQL query construction in particular dataset. Most data integration processes start with SPARQL query construction to get the subset of the data. Regarding integration, defining equal concepts between different ontology is important as explained in Figure 4-11 and 4-12. A correct and efficient SPARQL query leads to a semantically correct integration. To sum up, an accurate linked data integration, consumption, and retrieval are the implication of a well-described ontology from the data provider.

Spatial component as the important "hook" for link discovery.

Both geometry and toponym are important elements in data integration. It is possible if those elements are described by one commonly used vocabulary. Based on the examination of numbers of LD datasets, we found different geometry vocabularies to describe the geometry resources. These differences lead to variety in data format, coordinate, and serialization. Furthermore, the non-standard representation could hinder the link discovery phase. As best practice, the vocabulary of `geosparql:asWKT` should be reused by data owners to describe geometry. Regarding the toponym, there is no commonly used vocabulary to describe the place name. Based on the dataset examination, mostly the toponym information is merged into general name vocabulary whereas the toponym is the unique name that represents certain space entities. Hence, the development of a toponym vocabulary is required to better describe place names and to differentiate it from other name attributes. The existence of official place names also can lead to better link discovery. The official place name will provide relatively comparable resource, so that string-based similarity measures can be used in a sensible way.

5.2. Conclusions

This thesis is aimed to evaluate the LOD Cloud by assessing the data structure and the representation of linked spatial data, in order to support exploration and integration purposes. This can be achieved by designing: 1) the workflow for analysing linked spatial data resource in the linked open data cloud, and 2) the workflow for integrating linked spatial data. Based on a thorough analysis of 26 datasets, we found 32 ontologies related to general spatial data description and seven geometry vocabularies for specific geometry description. This number reflected the heterogeneity of the current available linked spatial data on the web.

Regarding spatial data integration, a workflow was designed to deal with different data access (SPARQL endpoint and RDF dump), data storage, and data format. It aimed establishing the link

between non-LOD and LOD Cloud dataset. The case study datasets are Kadaster, Geonames, and Natura2000 data. The main contribution of this thesis is the provision of the study cases about integrating various spatial data sources.

The result of this thesis proved that spatial component, geometry and toponym, can be used as an important "hook" for integrating different datasets. We also found that the commonly used geospatial ontology and vocabulary enable semantic interoperability to support data integration. In addition, we found that the existence of 1) spatial-aspect of metadata, 2) well-described ontology, 3) overview visualization of relationship between dataset, assist users for better linked data discovery, retrieval, and consumption.

We presented the thesis result with respect to each research questions per sub-objectives:

1a. What are the elements that can be used to characterize linked data in the LOD Cloud?

Metadata is the reference that summarizes the resources of certain LD datasets. There are two vocabularies are used to describe resources, Vocabulary of Interlinked Datasets (VoID) and Data Catalog Vocabulary (DCAT). In Section 2.1.4 we presented several important elements of VoID that can characterize the linked data, such as 1) Basic information related to categories of data, 2) Vocabulary usage, 3) The linked external dataset, 4) Predicates used, 5) Number of triples linked, and 6) Basic statistics about the dataset. DCAT provides elements to describe dataset in order to enable across data catalogue exploration. Even though it mostly describes the data access, several elements are useful especially for spatial data such as: 1) Geographic bounding, 2) Spatial representation type, 3) Spatial resolution, and 4) CRS. In addition, DQV is developed to describe the quality of dataset with respect to the numbers of dimension and metrics.

1b. What are the principles of linked data quality frameworks?

Section 2.2 presented the review of the previous linked data quality studies. Seven studies related to linked data quality have been examined. Each of these studies has different ways to structure the linked data quality elements, from abstract concepts to the measurement assessment. In Section 2.3.1, we mapped linked data quality elements from these seven studies to three level of hierarchies. The first hierarchy is the broader concept of data quality. The second hierarchy is the dimension which explain narrower concept of linked data quality. Finally, the third hierarchy has a function to operationalize these linked data quality dimension elements. Metrics and indicators are the components for measuring and assessing data quality principle.

These seven studies developed a variety of dimensions and metrics to assess the quality of dataset, metadata, resource, and links. Most dimensions and metrics were developed to assess the four key issues of linked data, namely: 1) Assign correct URIs to identify entities, 2) Use HTTP URIs to make data in machine-readable format, 3) Use RDF standard, and 4) Link to external data. The examples of linked data quality dimensions are consistency, syntactic validity, completeness, interlinking, etc. These dimensions and metrics assist the users to judge the quality of the LD datasets based on their applications.

1c. What are the dimensions and metrics of linked data quality frameworks that can measure the quality of links?

Section 2.3.2 presented metrics that can be used for assessing the quality of link. We categorized the related metrics into three groups. The first group is related to the concept of network-measure, which assumed linked data as web of data. This group includes link degree metric, clustering coefficient metric, and centrality metric. The second group is related to completeness concept, which describes how the link between datasets can enrich, trade-off, and complement information. The metrics includes the interlinking completeness metric and the complementation metric. The above-mentioned metrics are only related to measuring the link quality of linked data in general instead of topical data. To align with this thesis objective, we presented one new metric namely the existence of geospatial ontology and vocabularies. This metric was proposed based on the result of Chapter 2. In order to assess the link quality of linked spatial data, the resources must contain geospatial ontology – vocabulary to describe the spatial component properly. Evaluation results are mentioned in the next research question.

1d. How to use the evaluation result to find the potential links between datasets in the LOD Cloud?

In Section 2.6 we presented workflow to assess the linked data. The workflow is developed to implement the metrics of the existence of geospatial ontology – vocabulary. The assessment result showed which datasets were described by geospatial ontology – vocabulary. This can assist user to find spatial-related resources, such as geometry and toponym. For instance, vocabulary of `geosparql:asWKT` indicates literal geometry representation. Finding the potential link starts from finding the right candidate resources. Thus, the result presented in Table 2-6 can lead users to find potential links between dataset.

2a. What kind of activities can be supported by LOD Cloud?

LOD Cloud term here refers to two elements: 1) the LOD Cloud diagram visualization and 2) the hyperlinks that connected to datahub.io data catalogue. First, the visualization showed the network of available linked data on the web based on domain categorization. Based on the discussion of linked data visualization requirement in Section 3.1, the diagram provides several elements for exploration activity such as visual presentation, data overview, detail on demand, and highlights links in data. Based on interview feedback on the diagram (Section 3.2.2), the overview insight obtained from graph visualization of relationships between datasets is highlighted as important entry points to explore datasets. This is also supported by dataset domain categorization which effectively depicts the overview for further and detail data exploration. Second, the hyperlinks that connected to datahub.io data catalogue. The datahub.io is CKAN-based data portal, thus it implements CKAN data model. CKAN data model provides CKAN entity to facilitate data discovery. (Section 2.1.1).

2b. How should linked spatial data be represented in the LOD Cloud in order to support the potential use?

Analysis of recent studies of linked spatial data visualization (Section 3.4) and the interview feedback (Section 3.2.2), were used to answer this question. Potential use refers to the opinions of experts who have the experience of using LOD Cloud. The functionality which is demanded by the users is to understand the spatial extent of the dataset which can ease the datasets selection for user's application. As we discussed in Section 3.1, to achieve an intuitive discovery and analysis, various visualization

elements and features are implemented to support user's understanding of the data structure and content. Currently, LOD Cloud Diagram provides a static representation of linked data and aggregates several types of links between datasets into a single line representation. Nevertheless, other approaches also can be tested as summarized in Table 3-2, for instance: (i) providing information about scope of data (global, regional, or national), (ii) providing functionality to support multilevel exploration that bridges discovery from overview to detail, (iii) using collapsible graphs to group the datasets, or (iv) visualizing the spatial extent of dataset using inset map using bounding box information from metadata. Apart from aligning representation solution to LOD Cloud Diagram, a separate interface of visualization for spatial data can also be tested. Several options are mentioned in the next research question.

2c. What are the options to represent spatial relations?

The linked spatial data visualization tools were presented in Section 3.4. All the tools are map-based visualization but offered several alternatives regarding data exploration. Map4rdf tool and Facete tool provide faceted browsing interface to discover the related geospatial content of certain resources. Next, Cesium tool is able to parse and extract geospatial vocabularies and project the geometry to the map. While Sextant tool and Spacetime tool allow visualization of linked spatio-temporal data using combination of map and timeline simultaneously. DBPedia Atlas implemented spatialization to show intra-class and inter-class relation within dataset, whereas LinkedGeoData browser displays the classes and instances to the interface based on selected region of the map. Lastly, GeoYASGUI tool is developed to detect the `geosparql:asWKT` and plot it in the map.

All the tools described above visualize the links in the instance level which commonly provide SPARQL query functionality to filter the data. Apart from this approach, link visualization in the set level can also be considered. Regarding the linkset, the concept of linkset in different granularity is presented in Section 3.3. It provides the conceptualization and organization of link in the dataset, subset, and distribution level. Using this concept, the spatial relation between subset of dataset can be visualized in a more organized way.

2d. How can the LOD Cloud user interface be improved for exploration and integration purposes?

The points of improvement are obtained from interview feedback of the expert's opinions of LOD Cloud user interface. Section 3.2.2 presented the summary of the feedback with respect to linked data consumption components. It includes components of obtaining an overview, navigation and exploration discovery, presentation and visualization, information retrieval, and data verification. The summary of the feedback is provided in Table 3-2. For exploration and integration purposes, information of spatial extent is required. This information can be retrieved from bounding box element of metadata. However, the availability of this bounding box element is extremely rare. This problem can be overcome by providing a system that can automatically generate the metadata based on resources without any involvement of data owner or at least does not require manual input of the metadata. If this information is available, it can be visualized in the inset map and support exploration and integration purposes.

3a. To what extent standards can be used for representing spatial data in a linked data format?

There are two elements that are used to describe spatial element of the linked data format. The first element is the spatial-enabled metadata. Section 4.1.1 discussed how GeoDCAT facilitate the description of geospatial datasets, dataset series, and services. The spatial aspect of metadata provides significant benefits in discovery activity by making spatial datasets, dataset series, and service more searchable across data platforms. The second element is geospatial ontology and vocabulary, which defines basic semantics for various spatial concept, for instance, class and subclasses, properties, basic datatypes for geometry, geometry representation, etc. In this study, we analysed OGC GeoSPARQL vocabulary which was described in Section 4.1.2. It has five components, namely: 1) Core, 2) Geometry Extension, 3) Topology Vocabulary Extension, 4) Geometry Topology Extension, 5) RDFS entailment, and 6) Query rewrite. These components not only provide spatial data description but also enable qualitative reasoning through query.

3b. How can a dataset be added to the LOD Cloud? What are the restrictions?

The workflow of LOD Cloud integration was designed to provide a way for adding dataset to LOD Cloud. Section 4.3 presented a step-by-step explanation of integration process. To add dataset to the LOD Cloud, several processes must be done, i.e.: finding relevant datasets to be linked in the data catalogue using proper identifier (tags, metadata, etc), conversion to RDF (if dataset not available as linked data) and finding relevant resources (link discovery). Link discovery require data model analysis, query construction, link condition definition, and transformation & blocking. The process in this study is restricted to the integration between the existing dataset in the LOD Cloud and non-LOD Cloud dataset. Regarding spatial data, geometry and toponym can be used as component to integrate different dataset. We presented the explanation of these two case studies in Section 4.3.1 and 4.3.2.

3c. How to use relevant GeoSPARQL queries to discover potential links among LOD Cloud datasets? To what resources the link should be established?

For the geometry-based integration case study, the link discovery is quite straightforward. The resources that was intended to be discovered is based on the topological relationship condition. The queries are constructed based on Topology Vocabulary Extension component of GeoSPARQL vocabulary. For the toponym-based integration case study, an understanding of both ontology datasets that targeted to be linked is necessary. Finding similar or equal concept between different ontologies is required in order to retrieve the exact targeted resources and establish true semantic relationship between them.

5.3. General Conclusion

Based on this study, we extracted high-level information regarding LOD Cloud. LOD Cloud provides an overview of available datasets to support exploration and discovery. It has valuable role in promoting and encouraging the usability of linked data by the wider community but still must consider some points to be considered. We present three points in this section, which are:

1. The current version of LOD Cloud Diagram depicts well the datasets in nodes representation and connect it to the datahub.io with all the dataset information behind it. However, the diagram did not include the several LD datasets because they did not refer to existing LOD Cloud Diagram datasets even though LOD Cloud Diagram datasets refer to them. Considering this limitation, a back-link mechanism is required. This mechanism could be an automation tools that has ability to notify the dataset owner if their dataset has been referred by other datasets (incoming link), transform the incoming link into outgoing link, generate it and store in their triple store. This information would be essential as basic information involve more datasets in the visualization.

2. Discover and explore link between resources in datasets was a rather challenging task since the diagram only able to depicts one aggregation line of linkset between datasets. The implementation of linkset concept in the different level (dataset, subset, and distribution) will provide a better metadata organization which subsequently delivers more structured and sensible way to create a graph of relationships between datasets (the overview of dataset network) in different granularities.

3. Metadata in datahub.io as the basis information of visualization suffers from availability. Thus, provision of a system that can automatically generate metadata without any (or with minimum) involvement of data owner is required. In addition, the integration between data catalogue automatic assessment tools of linked data quality is also essential.

5.4. Recommendations

We identify the parts of this thesis that still need further development, as follows:

- The identification and analysis of linked spatial data sources in this thesis restricted is to certain data catalogue (CKAN datahub.io) as data pool and used CKAN entity of tags for dataset exploration. Meanwhile there are still other options besides datahub.io to find linked spatial data. For further study, other approaches are recommended, for instance utilize the VoID and GeoDCAT to find relevant dataset. Utilization of GeoDCAT can expand the exploration of the dataset since it enables metadata sharing across domains and catalogues platforms.
- Integration process using Silk Workbench and Single Machine still has a limitation in computing capabilities. To deal with big data integration, Silk MapReduce or Silk Server version are recommended.
- Spatial data is not always related to geometry. Due to this condition, toponym becomes an important element for spatial data integration. Defining equal concept amongst different ontology is essential in toponym-based integration. Thus, for further study on big data integration, the implementation of rigorous methodology is required and there are solutions in the semantic field namely ontology mapping. Ontology matching takes ontologies as input then define an alignment as output. Ontology matching is the key in enabling interoperability in Linked Data by merging the ontology and translating the data to match semantically related entities of the ontologies (Shvaiko et al., 2016). The Ontology Alignment Evaluation Initiative (OAEI) provides studies on ontology matching and instance matching or link discovery (OAEI, 2017).
- The conversion pipeline was designed for transforming raw spatial data format to linked data format. It can only proceed the conventional spatial data format such as shapefile, JSON, spatial RDBMS and GML. On the other hand, there is still a huge volume of spatial data on the web which does not have those formats. For further study, another approach can be applied to discover more spatial data. For instance, using web scraping approach by employing Natural Language Processing (NLP) methods to find relevant information and convert it into linked data format.

LIST OF REFERENCES

Abele, A., Buitelaar, P., Mccrae, J., & Bordea, G. (2016). Linked Data Profiling Identifying the Domain of Datasets Based on Data Content and Metadata. In *Proceedings of the 25th International Conference Companion on World Wide Web - WWW '16 Companion* (pp. 287–291). New York, New York, USA: ACM Press. https://doi.org/10.1145/2872518.2888603

Abele, A., Mccrae, J., & Buitelaar, P. (2017). An Evaluation Dataset for Linked Data Profiling. In J. Gracia, F. Bond, J. P. McCrae, P. Buitelaar, C. Chiarcos, & S. Hellmann (Eds.), *Language, Data, and Knowledge. LDK 2017. Lecture Notes in Computer Science, vol 10318* (pp. 1–9). Springer, Cham. https://doi.org/10.1007/978-3-319-59888-8_1

Albertoni, R., & Gómez Pérez, A. (2013). Assessing Linkset Quality for Complementing Third-Party Datasets. In *Proceedings of the Joint EDBT/ICDT 2013 Workshops on - EDBT '13* (p. 52). New York, New York, USA: ACM Press. https://doi.org/10.1145/2457317.2457327

Arturo, A., Caraballo, M., Nunes, B. P., Lopes, G. R., André, L., Paes Leme, P., & Casanova, M. A. (2016). Automatic Creation and Analysis of a Linked Data Cloud Diagram. In W. Cellary, M. F. Mokbel, J. Wang, H. Wang, R. Zhou, & Y. Zhang (Eds.), *Web Information Systems Engineering – WISE 2016. WISE 2016. Lecture Notes in Computer Science, vol 10041* (pp. 417–432). Springer, Cham. https://doi.org/10.1007/978-3-319-48740-3_31

Assaf, A., & Senart, A. (2012). Data Quality Principles in the Semantic Web. In *2012 IEEE Sixth International Conference on Semantic Computing* (pp. 226–229). Palermo, Italy: IEEE Computer Society. https://doi.org/10.1109/ICSC.2012.39

Assaf, A., Senellart -Telecom ParisTech, P., Stefan Dietze, F., & Troncy, R. (2015). *Enabling Self-Service Data Provisioning Through Semantic Enrichment of Data*. TELECOM ParisTech. Retrieved from http://www.eurecom.fr/en/publication/4739/detail/enabling-self-service-data-provisioning-through-semantic-enrichment-of-data

Assaf, A., Troncy, R., & Senart, A. (2015a). Roomba: An Extensible Framework to Validate and Build Dataset Profiles. In F. Gandon, C. Guéret, S. Villata, J. Breslin, C. Faron-Zucker, & A. Zimmermann (Eds.), *The Semantic Web: ESWC 2015 Satellite Events. ESWC 2015. Lecture Notes in Computer Science, vol 9341* (pp. 325–339). Springer, Cham. https://doi.org/10.1007/978-3-319-25639-9_46

Assaf, A., Troncy, R., & Senart, A. (2015b). What's up LOD Cloud? Observing the State of Linked Open Data Cloud Metadata. In F. Gandon, C. Guéret, S. Villata, J. Breslin, C. Faron-Zucker, & A. Zimmermann (Eds.), *The Semantic Web: ESWC 2015 Satellite Events. ESWC 2015. Lecture Notes in Computer Science, vol 9341* (pp. 247–254). Springer, Cham. https://doi.org/10.1007/978-3-319-25639-9_40

Auer, S., Demter, J., Martin, M., & Lehmann, J. (2012). LODStats – An Extensible Framework for High-Performance Dataset Analytics. In A. ten Teije, J. Völker, S. Handschuh, H. Stuckenschmidt, M. D'Acquin, A. Nikolov, … N. Hernandez (Eds.), *Knowledge Engineering and Knowledge Management. EKAW 2012. Lecture Notes in Computer Science, vol 7603* (pp. 353–362). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-33876-2_31

Becker, C., & Furness, P. (2010). Linking spatial data from the Web. *Journal of Direct, Data and Digital Marketing Practice*, *11*(4), 317–323. https://doi.org/10.1057/dddmp.2010.10

Beek, W., & Folmer, E. (2017). An Integrated Approach for Linked Data Browsing. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *XLII-4/W2*, 35–38. https://doi.org/10.5194/isprs-archives-XLII-4-W2-35-2017

Beek, W., Folmer, E., Rietveld, L., & Walker, J. (2017). GeoYASGUI: The GeoSPARQL Query Editor and Result Set Visualizer. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *Volume XLI*, 39–42. https://doi.org/10.5194/isprs-archives-XLII-4-W2-39-2017

Bereta, K., & Koubarakis, M. (2016). Ontop of Geospatial Databases. In P. Groth, E. Simper, La. Gray, M. Sabou, M. Krötzsch, F. Lecue, … Y. Gil (Eds.), *The Semantic Web – ISWC 2016* (pp. 37–52). Springer, Cham. https://doi.org/10.1007/978-3-319-46523-4_3

Berners-Lee, T. (2006). Linked Data - Design Issues. Retrieved August 4, 2017, from https://www.w3.org/DesignIssues/LinkedData.html

Berners-Lee, T. (2009). Linked Data - Design Issues. Retrieved August 18, 2017, from https://www.w3.org/DesignIssues/LinkedData.html

Bikakis, N., Greece, A. R. C., & Sellis, T. (2016). Exploration and Visualization in the Web of Big Linked Data: A Survey of the State of the Art and Challenges Ahead. *6th International Workshop on Linked Web Data Management (LWDM 2016)*. Retrieved from https://arxiv.org/pdf/1601.08059.pdf

Bikakis, N., Papastefanatos, G., Skourla, M., & Sellis, T. (2017). A Hierarchical Framework for Efficient Multilevel Visual Exploration and Analysis. *Semantic Web*, *8*(1), 139–179. https://doi.org/10.3233/SW-160226

Bizer, C. (2009). The Emerging Web of Linked Data. *IEEE Intelligent Systems*, *24*(5), 87–92. https://doi.org/10.1109/MIS.2009.102

Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked Data -The Story So Far. *International Journal on Semantic Web and Information Systems*, *5*(3), 1–22. Retrieved from http://linkeddata.org/docs/ijswis-special-issue

Brunetti, J. M., Auer, S., García, R., Klímek, J., & Nečaský, M. (2013). Formal Linked Data Visualization Model. In *Proceedings of International Conference on Information Integration and Web-based Applications & Services - IIWAS '13* (pp. 309–318). New York, New York, USA: ACM Press. https://doi.org/10.1145/2539150.2539162

Chawuthai, R., & Takeda, H. (2016). RDF Graph Visualization by Interpreting Linked Data as Knowledge. In G. Qi, K. Kozaki, J. Z. Pan, & S. Yu (Eds.), *JIST 2015: Semantic Technology* (pp. 23–39). Springer, Cham. https://doi.org/10.1007/978-3-319-31676-5_2

Dadzie, A.-S., & Rowe, M. (2011). Approaches to visualising Linked Data: A survey. *Semantic Web*, *2*(2), 89–124. https://doi.org/10.3233/SW-2011-0037

De León, A., Wisniewki, F., Villazón-Terrazas, B., & Corcho, O. (2012). Map4rdf -Faceted Browser for Geospatial Datasets. In *Using Open Data: Policy Modeling, Citizen Empowerment, Data Journalism*. Retrieved from https://www.w3.org/2012/06/pmod/pmod2012_submission_33.pdf

De Vocht, L., Dimou, A., Breuer, J., Compernolle, M. Van, Verborgh, R., Mannens, E., … Van De Walle, R. (2014). A Visual Exploration Workflow as Enabler for the Exploitation of Linked Open Data. In *IESD'14 Proceedings of the 3rd International Conference on Intelligent Exploration of Semantic Data - Volume 1279* (pp. 30–41). Riva del Garda, Italy: CEUR-WS.org. Retrieved from https://dl.acm.org/citation.cfm?id=2877803

Debattista, J., Auer, S., & Lange, C. (2016). Luzzu -- A Framework for Linked Data Quality Assessment. In *2016 IEEE Tenth International Conference on Semantic Computing (ICSC)* (pp. 124–131). IEEE. https://doi.org/10.1109/ICSC.2016.48

Debattista, J., Lange, C., & Auer, S. (2014). daQ, an Ontology for Dataset Quality Information. In *International World Wide Web Conference, WWW 2014* (p. 8). ACM Press.

Di, L., & Zhao, P. (2017). Geospatial Semantic Web, Interoperability. In S. Shekhar, H. Xiong, & X. Zhou (Eds.), *Encyclopedia of GIS* (pp. 746–746). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-17885-1_516

Ermilov, I., Lehmann, J., Martin, M., & Auer, S. (2016). LODStats: The Data Web Census Dataset. In *The Semantic Web – ISWC 2016* (pp. 38–46). Springer, Cham. https://doi.org/10.1007/978-3-319-46547-0_5

Fernández, M., Cantador, I., López, V., Vallet, D., Castells, P., & Motta, E. (2011). Semantically enhanced Information Retrieval: An ontology-based approach. *Web Semantics: Science, Services and Agents on the World Wide Web*, *9*(4), 434–452. https://doi.org/10.1016/j.websem.2010.11.003

García, R., Brunetti, J. M., Gil, R., & Gimeno, J. M. (2012). Rhizomer: Overview, Facets and Pivoting for Semantic Data Exploration. Retrieved from http://imash.leeds.ac.uk/event/2013/assets/proceeding/Rhizomer.pdf

Geonames. (2017). Geonames Feature Statistic Netherlands. Retrieved January 28, 2018, from http://www.geonames.org/statistics/netherlands.html

Graziosi, A., Di Iorio, A., Poggi, F., & Peroni, S. (2017). Customised Visualisations of Linked Open Data. In *Proceedings of the Third International Workshop on Visualization and Interaction for Ontologies and Linked Data co-located with the 16th International Semantic Web Conference (ISWC 2017)* (pp. 20–33). Retrieved from http://ceur-ws.org/Vol-1947/paper03.pdf

Guéret, C., Groth, P., Stadler, C., & Lehmann, J. (2012). Assessing Linked Data Mappings Using

Network Measures. In E. Simperl, P. Cimiano, A. Polleres, O. Corcho, & V. Presutti (Eds.), *9th Extended Semantic Web Conference, ESWC 2012* (pp. 87–102). eraklion, Crete, Greece: Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-30284-8_13

Hart, G., & Dolbear, C. (2013). *Linked data : a geographic perspective*. Taylor & Francis. https://doi.org/10.1201/b13877

Heath, T., & Christian Bizer, T. (2011). *Linked Data: Evolving the Web into a Global Data Space*. *Synthesis Lectures on the Semantic Web: Theory and Technology* (Vol. 1). Morgan & Claypool. https://doi.org/10.2200/S00334ED1V01Y201102WBE001

Hogan, A., Umbrich, J., Harth, A., Cyganiak, R., Polleres, A., & Decker, S. (2012). An empirical survey of Linked Data conformance. *Web Semantics: Science, Services and Agents on the World Wide Web*, *14*, 14–44. https://doi.org/10.1016/j.websem.2012.02.001

Hu, Y. (2018). Geospatial Semantics. In *Comprehensive Geographic Information Systems* (pp. 80–94). Elsevier. https://doi.org/10.1016/B978-0-12-409548-9.09597-X

Idreos, S., Papaemmanouil, O., & Chaudhuri, S. (2015). Overview of Data Exploration Techniques. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data - SIGMOD '15* (pp. 277–281). New York, New York, USA: ACM Press. https://doi.org/10.1145/2723372.2731084

ISA GeoDCAT-APWorking Group. (2016). GeoDCAT-AP: A geospatial extension for the DCAT application profile for data portals in Europe, *Version 1*. Retrieved from https://joinup.ec.europa.eu/node/139283#Distributions

Isele, R., Umbrich, J., Bizer, C., & Harth, A. (2010). LDspider: an open-source crawling framework for the web of linked data. In *Proceedings of the 2010 International Conference on Posters & Demonstrations Track - Volume 658* (pp. 29–32). CEUR-WS.org. Retrieved from https://dl.acm.org/citation.cfm?id=2878407

Kadaster. (2017). Basisregistraties adressen en gebouwen vocabulaire. Retrieved from https://bag.basisregistraties.overheid.nl/query/model

Kalkman, V. . (2014). Leucorrhinia pectoralis. Retrieved January 12, 2018, from http://dx.doi.org/10.2305/IUCN.UK.2014-1.RLTS.T165486A19167032.en

Knibbe, F. (2016). Spatial data on the Web: how should it work? - Geodan. Retrieved August 7, 2017, from https://www.geodan.com/spatial-data-web-work/

Konstantinou, N., & Spanos, D.-E. (2015). Introduction: Linked Data and the Semantic Web. In *Materializing the Web of Linked Data* (pp. 1–16). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-16074-0_1

Koubarakis, M., Karpathiotakis, M., Kyzirakos, K., Nikolaou, C., & Sioutis, M. (2012). Data Models and Query Languages for Linked Geospatial Data. In T. Eiter & T. Krennwallner (Eds.), *Reasoning Web. Semantic Technologies for Advanced Query Answering. Reasoning Web 2012. Lecture Notes in Computer Science* (Vol. 7487, pp. 290–328). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-33158-9_8

Kuhn, W., Kauppinen, T., & Janowicz, K. (2014). Linked Data - A Paradigm Shift for Geographic Information Science. In M. Duckham, E. Pebesma, K. Stewart, & A. U. Frank (Eds.), *Geographic Information Science* (pp. 173–186). Springer, Cham. https://doi.org/10.1007/978-3-319-11593-1_12

Kyzirakos, K., Karpathiotakis, M., & Koubarakis, M. (2012). Strabon: A Semantic Geospatial DBMS. In P. Cudré-Mauroux, J. Heflin, E. Sirin, T. Tudorache, J. Euzenat, M. Hauswirth, … E. Blomqvist (Eds.), *The Semantic Web – ISWC 2012* (pp. 295–311). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-35176-1_19

Kyzirakos, K., Vlachopoulos, I., Savva, D., Manegold, S., & Koubarakis, M. (2014). GeoTriples: a Tool for Publishing Geospatial Data as RDF Graphs Using R2RML Mappings. In *Proceedings of the ISCW 2014 Posters & Demonstrations Track* (pp. 393–396). Riva del Garda, Italy. Retrieved from http://ceur-ws.org/Vol-1401/paper-03.pdf

Lee, C.-J., Chuang, T.-R., & Huang, A. (2016). Open Data Web - A Linked Open Data Repository Built with CKAN. In *CKANCon 2016,*. Madrid, Spain. https://doi.org/10.13140/RG.2.2.32957.67043

Martins, B. (2011). A Supervised Machine Learning Approach for Duplicate Detection over Gazetteer Records (pp. 34–51). https://doi.org/10.1007/978-3-642-20630-6_3

Mihindukulasooriya, N., García-Castro, R., & Gómez-Pérez, A. (2017). LD Sniffer: A Quality Assessment Tool for Measuring the Accessibility of Linked Data. In P. Ciancarini, F. Poggi, Matthew Horridge, J. Zhao, T. Groza, M. C. S.-F. Presutti, … V. Presutti (Eds.), *EKAW 2016: Knowledge Engineering and Knowledge Management* (pp. 149–152). Springer, Cham. https://doi.org/10.1007/978-3-319-58694-6

Milić, P., Veljković, N., & Stoimenov, L. (2015). Linked Relations Architecture for Production and Consumption of Linksets in Open Government Data. In Marijn Janssen, M. Mäntymäki, J. Hidders, B. Klievink, W. Lamersdorf, B. van Loenen, & A. Zuiderwijk (Eds.), *14th IFIP WG 6.11 Conference on e-Business, e-Services, and e-Society, I3E 2015* (pp. 212–222). Springer, Cham. https://doi.org/10.1007/978-3-319-25013-7_17

Neto, C. B., Kontokostas, D., Hellmann, S., Müller, K., & Brümmer, M. (2016). Assessing Quantity and Quality of Links Between Linked Data Datasets. Retrieved from http://aksw.org/Groups/KILT

Neto, C. B., Kontokostas, D., Publio, G., Müller, K., Hellmann, S., & Moletta, E. (2016). LD-LEx: Linked Dataset Link Extractor (Short Paper). https://doi.org/10.1007/978-3-319-48472-3

Neto, C. B., Müller, K., Brümmer, M., Kontokostas, D., & Hellmann, S. (2016). LODVader: An Interface to LOD Visualization, Analytics and DiscovERy in Real-time, (16). https://doi.org/10.1145/2872518.2890545

Ngomo, A.-C. N., Auer, S., Lehmann, J., & Zaveri, A. (2014). Introduction to Linked Data and Its Lifecycle on the Web. In *Reasoning Web* (pp. 1–99). Springer International Publishing Switzerland. https://doi.org/10.1007/978-3-319-10587-1_1

Nikolaou, C., Dogani, K., Bereta, K., Garbis, G., Karpathiotakis, M., Kyzirakos, K., & Koubarakis, M. (2015). Sextant: Visualizing time-evolving linked geospatial data. *Web Semantics: Science, Services and Agents on the World Wide Web*, *35*, 35–52. https://doi.org/10.1016/j.websem.2015.09.004

OAEI. (2017). Ontology Alignment Evaluation Initiative. Retrieved from http://oaei.ontologymatching.org/

OGC. (2012). GeoSPARQL - A Geographic Query Language for RDF Data | OGC. Retrieved August 8, 2017, from http://www.opengeospatial.org/standards/geosparql

Paris, P.-H., Abadie, N., & Brando, C. (2017). Linking Spatial Named Entities to the Web of Data for Geographical Analysis of Historical Texts. *Journal of Map & Geography Libraries*, *13*(1), 82–110. https://doi.org/10.1080/15420353.2017.1307306

Patroumpas, K., Giannopoulos, G., & Athanasiou, S. (2014). TripleGeo: an ETL Tool for Transforming Geospatial Data into RDF Triples. Retrieved from http://ceur-ws.org/Vol-1133/paper-44.pdf

Pattuelli, M. C., Provo, A., & Thorsen, H. (2015). Ontology Building for Linked Open Data: A Pragmatic Perspective. *Journal of Library Metadata*, *15*, 265–294. https://doi.org/10.1080/19386389.2015.1099979

Potnis, A. V, & Durbha, S. S. (2016). Exploring Visualization of Geospatial Ontologies Using Cesium. Retrieved from http://ceur-ws.org/Vol-1704/paper14.pdf

Radulovic, F., Mihindukulasooriya, N., García-Castro, R., & Gómez-Pérez, A. (2018). A comprehensive quality model for Linked Data. *Semantic Web*, *9*, 3–24. https://doi.org/10.3233/SW-170267

Recchia, G., & Louwerse, M. (2013). A Comparison of String Similarity Measures for Toponym Matching. Retrieved from http://stko.geog.ucsb.edu/comp2013/comp2013_submission_2.pdf

Reyna, M. A. De, Simoes, J., & Genuchten, P. Van. (2016). Integrating the spatial web with linked open data using GeoDCAT-AP. FOSS4G,OSGeo. https://doi.org/https://doi.org/10.5446/20321

Rietveld, L. (2016). *Publishing and Consuming Linked Data Optimizing for the Unknown*. Vrije Universiteit. https://doi.org/10.3233/978-1-61499-623-1-i

Rietveld, L., & Hoekstra, R. (2013). YASGUI: Not Just Another SPARQL Client. *Extended Semantic Web Conference*, 78–86. Retrieved from http://laurensrietveld.nl/pdf/Yasgui.pdf

Rietveld, L., & Hoekstra, R. (2014). YASGUI: Feeling the Pulse of Linked Data. *EKAW 2014: Knowledge Engineering and Knowledge Management*, 441–452. Retrieved from https://link.springer.com/content/pdf/10.1007%2F978-3-319-13704-9_34.pdf

Rietveld, L., & Hoekstra, R. (2017). The YASGUI Family of SPARQL Clients. *Semantic Web*, *SWJ 8(3)*, 373–383. Retrieved from http://www.semantic-web-journal.net/system/files/swj1001.pdf

Ruckhaus, E., & Vidal, M.-E. (2012). LNCS 8194 - LiQuate-Estimating the Quality of Links in the Linking Open Data Cloud. Retrieved from https://ezproxy.utwente.nl:3351/content/pdf/10.1007%2F978-3-642-45263-5_4.pdf

Santos, R., Murrieta-Flores, P., & Martins, B. (2017). Learning to combine multiple string similarity metrics for effective toponym matching. *International Journal of Digital Earth*, 1–26. https://doi.org/10.1080/17538947.2017.1371253

Schmachtenberg, M., Bizer, C., & Paulheim, H. (2014). Adoption of the Linked Data Best Practices in Different Topical Domains. Retrieved from https://www.planet-data.eu/sites/default/files/publications/SchmachtenbergBizerPaulheim-AdoptionOfLinkedDataBestPractices.pdf

Shneiderman, B. (1996). The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. *Proceedings 1996 IEEE Symposium on Visual Languages*, 336–343. Retrieved from https://www.cs.umd.edu/~ben/papers/Shneiderman1996eyes.pdf

Shvaiko, P., Euzenat, J. ome, Jimenez-Ruiz, E., Cheatham, M., Hassanzadeh, O., & Ichise, R. (2016). OM-2016. In *Proceedings of the 11th International Workshop on Ontology Matching co-located with the 15th International Semantic Web Conference (ISWC 2016)*. CEUR-WS.org. Retrieved from http://ceur-ws.org/Vol-1766/om2016_preface.pdf

Skjaeveland, M. G. (2012). Sgvizler: A JavaScript Wrapper for Easy Visualization of SPARQL Result Sets. *ESWC, 2012*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.370.1482&rep=rep1&type=pdf

Smeros, P. (2014). *Discovering Spatial and Temporal Links Among RDF Data*. Retrieved from http://efessos.lib.uoa.gr/applications/disserts.nsf/0f1ab5fee83fbb88c225770c0042ce4f/a25cf38fc1cf2000c2257db2004758ec/$FILE/document.pdf

Smeros, P., & Koubarakis, M. (2016). Discovering Spatial and Temporal Links among RDF Data. In *Proceedings of the Workshop on Linked Data on the Web, LDOW 2016, co-located with 25th International World Wide Web Conference (WWW 2016)*. CEUR-WS.org. Retrieved from ceur-ws.org/Vol-1593/article-06.pdf

Stadler, C., Lehmann, J., Höffner, K., & Auer, S. (2012). LinkedGeoData: A core for a web of spatial open data. *Semantic Web*, *3*, 333–354. https://doi.org/10.3233/SW-2011-0052

Stadler, C., Martin, M., & Auer, S. (2014). Exploring the Web of Spatial Data with Facete. https://doi.org/10.1145/2567948.2577022

Taylor, K., & Parsons, E. (2015). Where Is Everywhere: Bringing Location to the Web. *IEEE Internet Computing*, *19*(2), 83–87. https://doi.org/10.1109/MIC.2015.50

Valsecchi, F., Abrate, M., Bacciu, C., Tesconi, M., & Marchetti, A. (2015). DBpedia Atlas: Mapping the Uncharted Lands of Linked Data. In *Proceedings of the Workshop on Linked Data on the Web co-located with the 24th International World Wide Web Conference (WWW 2015)*. CEUR-WS.org. Retrieved from http://ceur-ws.org/Vol-1409/paper-05.pdf

Valsecchi, F., & Ronchetti, M. (2014). Spacetime: a Two Dimensions Search and Visualisation Engine Based on Linked Data. In *SEMAPRO 2014 : The Eighth International Conference on Advances in Semantic Processing* (pp. 8–12). IARIA. Retrieved from http://www.iit.cnr.it/sites/default/files/semapro2014.pdf

Van Den Brink, L., Barnaghi, P., Tandy, J., Atemezing, G., Atkinson, R., Cochrane, B., … Troncy, R. (2017). Best Practices for Publishing, Retrieving, and Using Spatial Data on the Web. *IOS Press*. Retrieved from http://www.semantic-web-journal.net/system/files/swj1785.pdf

Van Den Brink, L., & De Visser, I. (2016). *Spatial data on the web*. Amsterdam. Retrieved from https://www.w3.org/2016/11/sdsvoc/SDSVoc16_paper_9

Vandenbussche, P.-Y., Atemezing, G. A., Poveda-Villalón, M., & Vatant, B. (2016). Linked Open Vocabularies (LOV): A gateway to reusable semantic vocabularies on the Web. *Semantic Web*, *8*(3), 437–452. https://doi.org/10.3233/SW-160213

Volz, J., Bizer, C., Gaedke, M., & Kobilarov, G. (2009). Silk – A Link Discovery Framework for the Web of Data. In *Proceedings of the Linked Data on the Web Workshop (LDOW2009)*. Madrid, Spain: CEUR-WS.org. Retrieved from http://ceur-ws.org/Vol-538/ldow2009_paper13.pdf

W3C. (2007). W3C Geospatial Ontologies. Retrieved August 7, 2017, from https://www.w3.org/2005/Incubator/geo/XGR-geo-ont/
W3C. (2011). Describing Linked Datasets with the VoID Vocabulary. Retrieved August 2, 2017, from https://www.w3.org/TR/void/
W3C. (2014). Data Catalog Vocabulary (DCAT). Retrieved August 1, 2017, from https://www.w3.org/TR/vocab-dcat/
W3C. (2016a). Data on the Web Best Practices: Data Quality Vocabulary. Retrieved October 7, 2017, from https://www.w3.org/TR/vocab-dqv/
W3C. (2016b). Data Quality Vocabulary for Linked Data. Retrieved October 7, 2017, from https://www.w3.org/TR/vocab-dqv/#DimensionsofZaveri
W3C. (2016c). List of DQV implementations.
W3C, & OGC. (2015). Spatial Data on the Web Working Group. Retrieved January 18, 2018, from https://www.w3.org/2015/spatial/wiki/Main_Page
W3C, & OGC. (2017). Spatial Data on the Web Best Practices. Retrieved August 7, 2017, from https://w3c.github.io/sdw/bp/
W3C SWEO. (2017). Linking Open Data Community Project. Retrieved April 9, 2017, from https://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData
Weise, M., Lohmann, S., & Haag, F. (2016). LD-VOWL: Extracting and Visualizing Schema Information for Linked Data. In *Proceedings of the Second International Workshop on Visualization and Interaction for Ontologies and Linked Data (VOILA '16),*. Kobe, Japan: CEUR-WS.org. Retrieved from http://ceur-ws.org/Vol-1704/paper11.pdf
Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., & Auer, S. (2015). Quality assessment for Linked Data: A Survey. *Semantic Web*, *7*(1), 63–93. https://doi.org/10.3233/SW-150175

APPENDIX

Table 1. The list of geo-ontologies that used in the 26 datasets

Name	Namespace
NeoGeo Spatial Ontology	http://geovocab.org/spatial#
NeoGeo Geometry Ontology	http://geovocab.org/geometry#
Simplified Features Geometry	http://www.opengis.net/ont/sf#
Ordnance Survey Spatial Relations Ontology	http://data.ordnancesurvey.co.uk/ontology/spatialrelations/
Ordnance Survey Geometry Ontology	http://data.ordnancesurvey.co.uk/ontology/geometry/
Ordnance Survey Geography Ontology	http://data.ordnancesurvey.co.uk/ontology/admingeo/
ISA Programme Location Core Vocabulary	http://www.w3.org/ns/locn#
OGC GeoSPARQL	http://www.opengis.net/ont/geosparql#
OGC Geometry	http://www.opengis.net/ont/gml#
OWL representation of ISO 19107 (Geographic Information)	http://def.seegrid.csiro.au/isotc211/iso19107/2003/geometry#
Ontology for geometry	http://data.ign.fr/def/geometrie#
OWL representation of ISO 19115 (Geographic Information – Metadata)	http://def.seegrid.csiro.au/isotc211/iso19115/2003/extent#
OWL representation of ISO 19107 (Geographic Information – Spatial)	http://def.seegrid.csiro.au/isotc211/iso19107/2003/geometry#
Ontology for representation of cartesian co-ordinates	http://purl.org/net/cartCoord#
FAO Geopolitical Ontology	http://aims.fao.org/aos/geopolitical.owl#
WGS84 Geo Positioning	http://www.w3.org/2003/01/geo/wgs84_pos#
ESRI Geo Features	http://www.mindswap.org/2003/owl/geo/geoFeatures20040307.owl#
Frappe - FraPPE: Frame, Pixel, Place, Event vocabulary	http://streamreasoning.org/ontologies/frappe#
The Geonames Ontology	http://www.geonames.org/ontology#
LinkedGeoData Ontology	http://linkedgeodata.org/ontology/
Vocabulario de Localizaciones	http://purl.org/ctic/infraestructuras/localizacion#
Geographically Encoded Objects for RSS feeds	http://www.georss.org/georss/
USGS	http://data.usgs.gov/lod/geometry/
USGS GNIS	http://data.usgs.gov/lod/gnis/ontology/
W3 vcard	https://www.w3.org/2006/vcard/ns#
Geopolitical Ontology	http://www.fao.org/countryprofiles/geoinfo/geopolitical/resource/
UK GeoCode	http://geo.data.gov.uk/0/ontology/geo#
Statistics UK	http://statistics.data.gov.uk/def/spatialrelations/
Publish My Data	http://publishmydata.com/def/ontology/spatial
Telegraphis Geography Ontology	http://telegraphis.net/ontology/geography/geography#
Awesemantic	http://awesemantic-geo.link/ontology/
Pleiades	https://pleiades.stoa.org/places/vocab#

Listing 1. Data Cleaning in Linked Sensor Data (Kno.e.sis) Dataset.

```
wrong_syntax = ['rdf/', 'rdf//']

with open('D:/git/data/Choosen_RDF_DUMP/Linked_Sensor_Data(Kno.e.sis)/knoesis_linkeddata.txt') as oldfile, \
        open('D:/git/data/Choosen_RDF_DUMP/Linked_Sensor_Data(Kno.e.sis)/revise.txt', 'w') as newfile:
    for line in oldfile:
        if not any(wrong_statement in line for wrong_statement in wrong_syntax):
            newfile.write(line)
```

Listing 2. Data Cleaning in Geonames Dataset.

```
import rdflib
fo = open("D:/git/data/Choosen_RDF_DUMP/GeoNames/all-geonames-rdf/all-geonames-rdf.nt", "wb")
totaltriples = 0
with open("D:/git/data/Choosen_RDF_DUMP/GeoNames/all-geonames-rdf/all-geonames-rdf.txt", encoding="utf8") as file:

    count = 0
    for line in file:
        if count/10000 == int(count/10000):
            print(count)

        if count%2 != 0:
            g=rdflib.ConjunctiveGraph()
            result=g.parse(data=line,format='xml')

            totaltriples += len(g)
            s=g.serialize(format='nt')
            fo.write(s)

        count = count + 1

print ("Total triples: ", totaltriples)
fo.close()
```

Listing 3. Data Cleaning in NUTS Dataset.

```
import string
import re

a=open("D:/git/data/Choosen_RDF_DUMP/NUTS/geography_test.txt",'r+')

for line in a:
        b=line.replace('^^<decimal>','')
        print(b)
```

Listing 4. Processing Data for Checking the Geospatial Ontology Existence (Example apply for GADM Dataset)

```
import rdflib
from rdflib.graph import Graph, URIRef
import matplotlib.pyplot as plt
import gzip
from __init__1 import LocalEndpoint,one,QName

g = Graph()

g.parse("D:/git/data/Choosen_RDF_DUMP/GADM/gadm_dump2.nt",format="nt")
len(g)
print(len(g))

e=LocalEndpoint(g)

##Type of predicate and count the number
properties1=e.select("""
   SELECT ?p (COUNT(*) AS ?cnt) {
      ?s ?p ?o .
   } GROUP BY ?p ORDER BY DESC(?cnt)
""")
properties1
print(properties1)

##Investigate object that used certain predicate
properties1=e.select("""
   SELECT ?s ?o  { ?s <http://geovocab.org/spatial#PP> ?o } LIMIT 200
""")
properties1
print(properties1)



fig1=properties1["cnt"].plot.pie(figsize=(6,6)).set_ylabel('')
plt.show()
```

Listing 5. Link – LSL for geometry-based integration between Natura2000 and BRT Kadaster

```
<Silk>
  <Prefixes>
    <Prefix id="rdf" namespace="http://www.w3.org/1999/02/22-rdf-syntax-ns#"/>
    <Prefix id="bag" namespace="http://bag.basisregistraties.overheid.nl/def/bag#"/>
    <Prefix id="geo" namespace="http://www.opengis.net/ont/geosparql#"/>
    <Prefix id="owl" namespace="http://www.w3.org/2002/07/owl#"/>
    <Prefix id="rdfs" namespace="http://www.w3.org/2000/01/rdf-schema#"/>
    <Prefix id="brt" namespace="http://brt.basisregistraties.overheid.nl/def/top10nl#"/>
  </Prefixes>
  <DataSources>
    <Dataset id="Natura2000" type="file">
      <Param name="file" value="SQL.nt"/>
      <Param name="format" value="N-TRIPLE"/>
    </Dataset>
    <Dataset id="BRT" type="sparqlEndpoint">
      <Param name="endpointURI" value="https://data.pdok.nl/sparql"/>
      <Param name="pageSize" value="10000"/>
    </Dataset>
  </DataSources>
  <Interlinks>
    <Interlink id="Natura-BRT">
      <SourceDataset dataSource="Natura2000" var="a">
        <RestrictTo>
        </RestrictTo>
      </SourceDataset>
      <TargetDataset dataSource="BRT" var="b" typeUri="http://brt.basisregistraties.overheid.nl/def/top10nl#Watergebied">
        <RestrictTo>
        </RestrictTo>
      </TargetDataset>
      <LinkageRule linkType="http://www.opengis.net/ont/geosparql#sfWithin">
        <Compare id="WithinMetrics" required="false" weight="1" metric="WithinMetric" threshold="0.9" indexing="true">
            <TransformInput id="SimplifyTransformer1" function="SimplifyTransformer">
                <Input id="Natura2000" path="?a/geo:asWKT"/>
            </TransformInput>
            <TransformInput id="EnvelopeTransformer2" function="EnvelopeTransformer">
                <Input id="BRT" path="?b/brt:geometrie/geo:asWKT"/>
            </TransformInput>
            <Param name="blockingParameter" value="10"/>
        </Compare>
        <Filter/>
      </LinkageRule>
    </Interlink>
  </Interlinks>
  <Transforms>
  </Transforms>
  <Outputs>
    <Dataset id="Natura-BRT" type="file">
      <Param name="file" value="SQLresult123.nt"/>
      <Param name="format" value="N-TRIPLE"/>
    </Dataset>
  </Outputs>
</Silk>
```

Listing 6. Link – LSL for toponym-based integration between Geonames and BRT Kadaster

```
<Silk>
  <Prefixes>
    <Prefix id="rdf" namespace="http://www.w3.org/1999/02/22-rdf-syntax-ns#"/>
    <Prefix id="bag" namespace="http://bag.basisregistraties.overheid.nl/def/bag#"/>
    <Prefix id="geo" namespace="http://www.opengis.net/ont/geosparql#"/>
    <Prefix id="gn" namespace="http://www.geonames.org/ontology#"/>
    <Prefix id="owl" namespace="http://www.w3.org/2002/07/owl#"/>
    <Prefix id="rdfs" namespace="http://www.w3.org/2000/01/rdf-schema#"/>
    <Prefix id="brt" namespace="http://brt.basisregistraties.overheid.nl/def/top10nl#"/>
    <Prefix id="n2k" namespace="http://data.pdok.nl/natura-2000/def/natura-2000#"/>
  </Prefixes>
  <DataSources>
    <Dataset id="Geonames" type="file">
      <Param name="file" value="geonamesPPL.nt"/>
      <Param name="format" value="N-TRIPLE"/>
    </Dataset>
    <Dataset id="BRT" type="sparqlEndpoint">
      <Param name="endpointURI" value="https://data.pdok.nl/sparql"/>
      <Param name="pageSize" value="10000"/>
    </Dataset>
  </DataSources>
  <Interlinks>
    <Interlink id="Geonames-BRT">
      <SourceDataset dataSource="Geonames" var="a">
        <RestrictTo>
        </RestrictTo>
      </SourceDataset>
      <TargetDataset dataSource="BRT" var="b" typeUri="http://brt.basisregistraties.overheid.nl/def/top10nl#Plaats">
        <RestrictTo>
        </RestrictTo>
      </TargetDataset>
      <LinkageRule linkType="http://www.w3.org/2002/07/owl#sameAs">
        <Compare id="equality1" required="false" weight="1" metric="equality" threshold="0.0" indexing="true">
              <Input id="Geonames" path="?a/gn:name"/>
              <Input id="BRT" path="?b/brt:naamNL"/>
        </Compare>
        <Filter/>
      </LinkageRule>
    </Interlink>
  </Interlinks>
  <Transforms>
  </Transforms>
  <Outputs>
    <Dataset id="Geonames-BRT" type="file">
      <Param name="file" value="onlyPPL.nt"/>
      <Param name="format" value="N-TRIPLE"/>
    </Dataset>
  </Outputs>
</Silk>
```