LODNav – An Interactive Visualization of the Linking Open Data Cloud

Christopher Neal

A Thesis
in
The Department
of
Computer Science and Software Engineering

Presented in Partial Fulfillment of the Requirements

For the Degree of Master of Applied Sciences (Software Engineering) at

Concordia University

Montreal, Quebec, Canada

March 2014

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: Christopher Neal

Entitled: LODNav – An interactive visualization of the Linking Open Data Cloud

and submitted in partial fulfillment of the requirements for the degree of

Master of Applied Sciences (Software Engineering)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final Examining Committee:

_Dr. Lata Narayana__________________ Chair

_Dr. Rene Witte____________________ Examiner

_Dr. Todd Eavis__ __________________ Examiner

_Dr. Juergen Rilling_________________ Supervisor

Approved by ____

Chair of Department of Graduate Program Director

_______________2014 ______________________________

Dean of Faculty

Abstract

LODNav: An Interactive Visualization of the Linking Open Data Cloud

Christopher Neal

The emergence of the Linking Open Data Cloud (LODC) is an example of the adoption of Linked Data principles and the creation of a Web of Data. There is an increasing amount of information linked across member datasets of the LODC by means of RDF links, yet there is little support for a human to understand which datasets are connected to one another. This research presents a novel approach for understanding these interconnections with the publicly accessible tool LODNav – Linking Open Data Navigator. LODNav provides a visualization metaphor of the LODC by positioning member datasets of the LODC on a world map based on the geographical location of the dataset. This interactive tool aims to provide a dynamic up-to-date visualization of the LODC and allows the extraction of information about the datasets as well as their interconnections as RDF data.

Table of contents

List of Figures

List of Tables

List of Abbreviations

HTTP	Hypertext Transfer Protocol
LOD	Linked Open Data
LODC	Linked Open Data Cloud
OWL	Web Ontology Language
RDF	Resource Description Framework
RDFS	RDF Schema
SPARQL	SPARQL Protocol and RDF Query Language
URI	Uniform Resource Identifier
W3C	World Wide Web Consortium

1. Introduction

With the advent of the World Wide Web, the way in which information is published and accessed has vastly changed. With an Internet connection one is now able to contribute and access a global information space [7]. People looking for information are no longer limited to static texts restricted to a geographic location, such as books in libraries or circulated newspapers. A web of information has rapidly grown by means of a network of hypertext links, which can be conceived as a web of documents [5]. Internet users can traverse this web of documents by means of a web browser and view the text within particular web documents. This traversal is often aided by search engines which provide suggested web pages for a user based on an inputted search query. The search engine accomplishes this task by indexing web documents and evaluating the arrangement of links between them [10]. This principal of connecting documents with links has provided a great tool for humans to access and read documents. On the other hand, using this approach does not allow for the connection of data itself in a machine-readable format [7] [33].

However, the idea of 'Linked Data' provides a solution to this problem and has recently begun to take shape. Linked Data is a manner in which data within a dataset is described using a particular standardized format, known as RDF (Resource Description Framework), which can then be connected with other datasets within the same format [5]. The result is a graph-based structure of data. With connections in place between datasets data can now be retrieved using a query as if it were a single source [7]. At present, the most noticeable implementation of the principles behind Linked Data has been the

‘Linking Open Data Project’ [7]. This community driven effort, founded in 2007, has the goal of extending the capabilities of the Web by means of using the concept of ‘Linked Data’. That is, by making various open datasets available as RDF data, these datasets can now be connected due to their common representation format. Information no longer needs to be held in distinct “silos” rather it can form a Web of Data [2]. When a similar concept is present in another dataset links can be established between these datasets. More explicitly, RDF links can be made in this manner between items from disparate data sources [43]. See Figure 1-1 for an overview of this process.

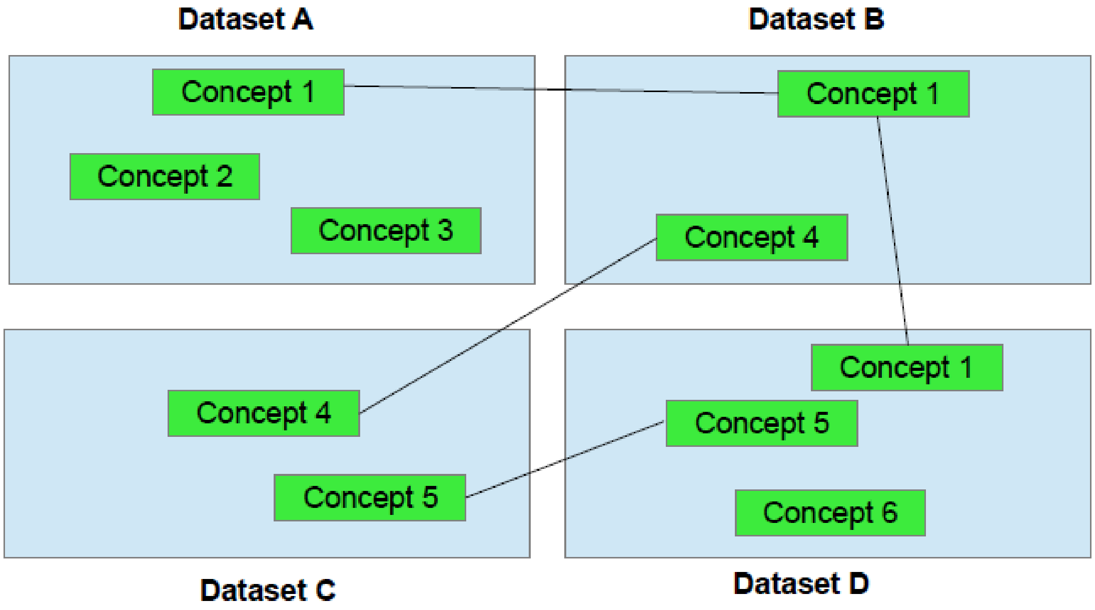


Figure 1-1 : Example of linking Concepts between Datasets

Over time the amount of connected datasets as part of the Linking Open Data Project has steadily increased:

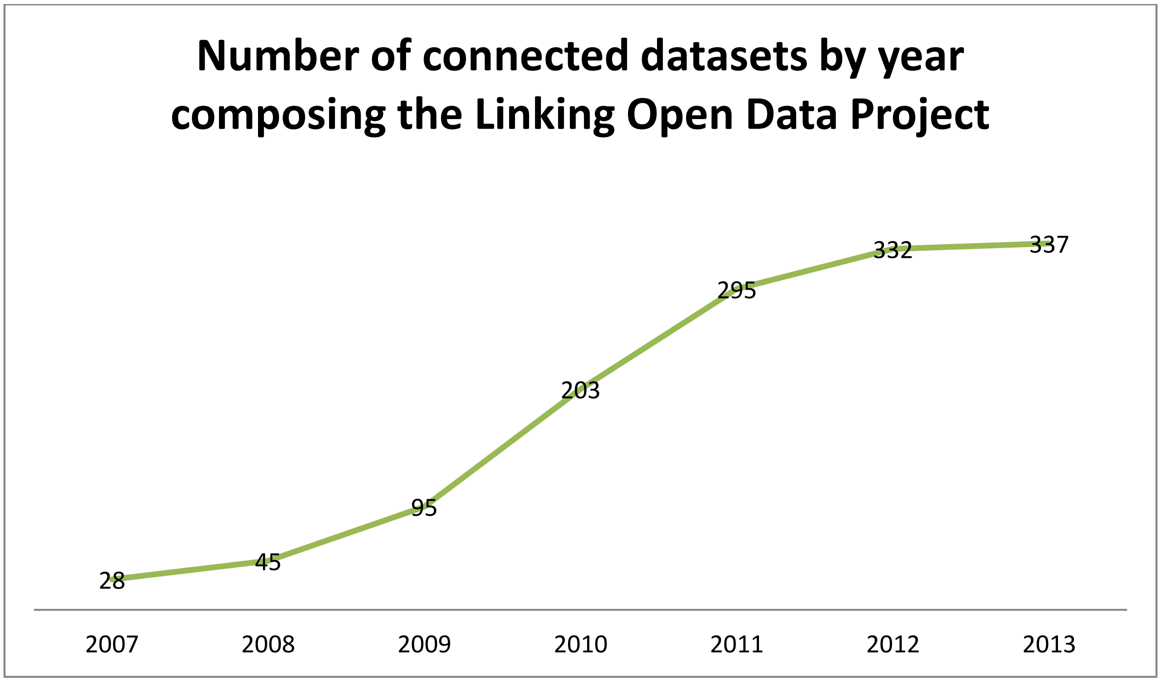


Table 1-1 : LOD datasets by year

As of September 2011 the Linking Open Data Project consisted of 295 connected datasets, which further increased to 332 datasets by October 2012 and 337 connected datasets as of October 2013 [43]. The Linking Open Data Project provides the following diagram seen in Figure 1-2 for illustrating how its member datasets are connected (for a larger view consult the webpage[1]).

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

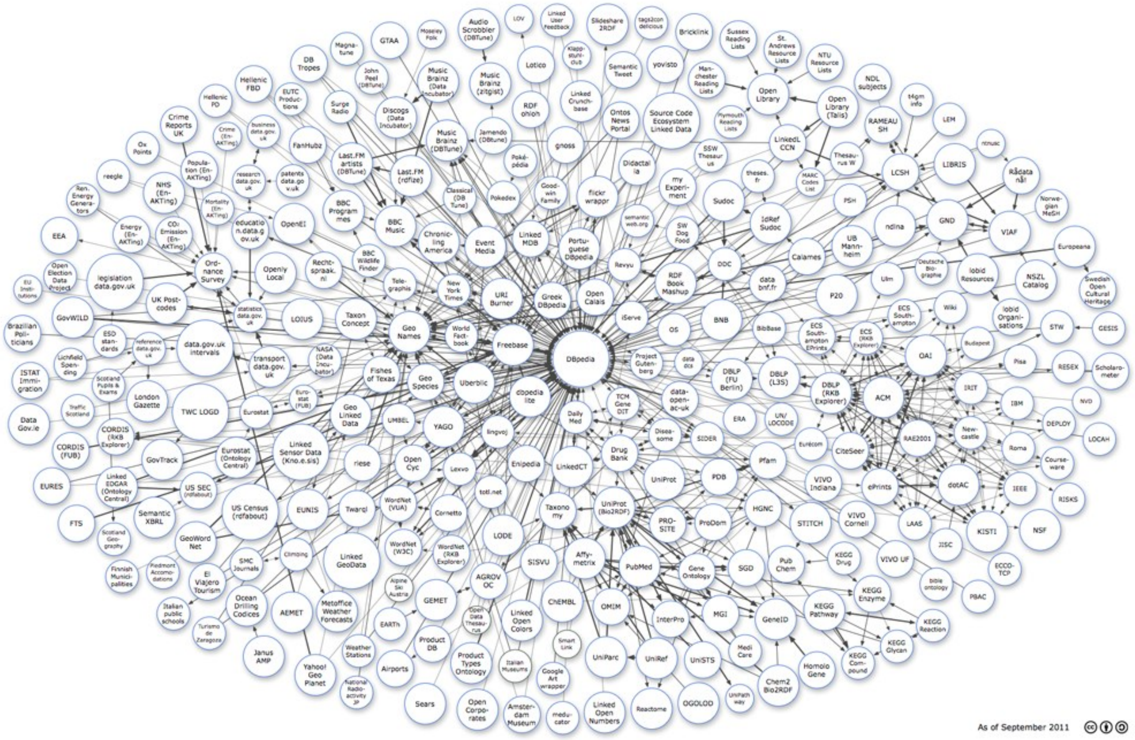


Figure 1-2 : Linking Open Data Cloud diagram [50]

Although the number of connected datasets continues to grow, implying a richer degree of data available for use, there are some fundamental limitations with this visual representation when it comes time to actually trying to understand the presented data.

Some of the problems associated with this visualization include:

- The graph representation used to visualize the size and interconnectivity of nodes in the LOD is manually hand curated, resulting in very few graph updates and therefore being often outdated. At the time of writing, the graph visualization has been last updated on 2011-09-19.

- It is difficult to see which datasets are connected, since only the complete graph is shown without a filtering option.

- Given the static nature of the graph, no information is included mentioning which datasets are currently ‘on-line’ in order to be queried.

Another limitation of using the datasets within the Linking Open Data Project is that the project lacks a comprehensive and up-to-date list of the member datasets with their respective meta-data, reflecting the status of these datasets (e.g. actual size, availability of the datasets).

As a result, the main goal of the research presented here is to provide a visualization tool for improved interaction with the datasets, which comprise the Linking Open Data Cloud (LODC), allowing for a better understanding how these datasets are interconnected while providing additional insights about the link types and the status of these datasets.

1.1 Contributions

The motivation of this research is to provide a novel visualization approach to improve the navigation and understanding of the Linking Open Data (LOD) member datasets and their interconnections. The research introduces a tool called ‘LODNav – Linking Open Data Navigator’ and is accessible via its webpage[2]. LODNav provides an interface for a human user to view the Linking Open Data Cloud datasets. Additionally, human users can use the provided functionality for interacting with the presented information and extract facts as RDF data. Figure 1-3 provides an overview of the complete approach.

[2] http ://www.lodnav.com

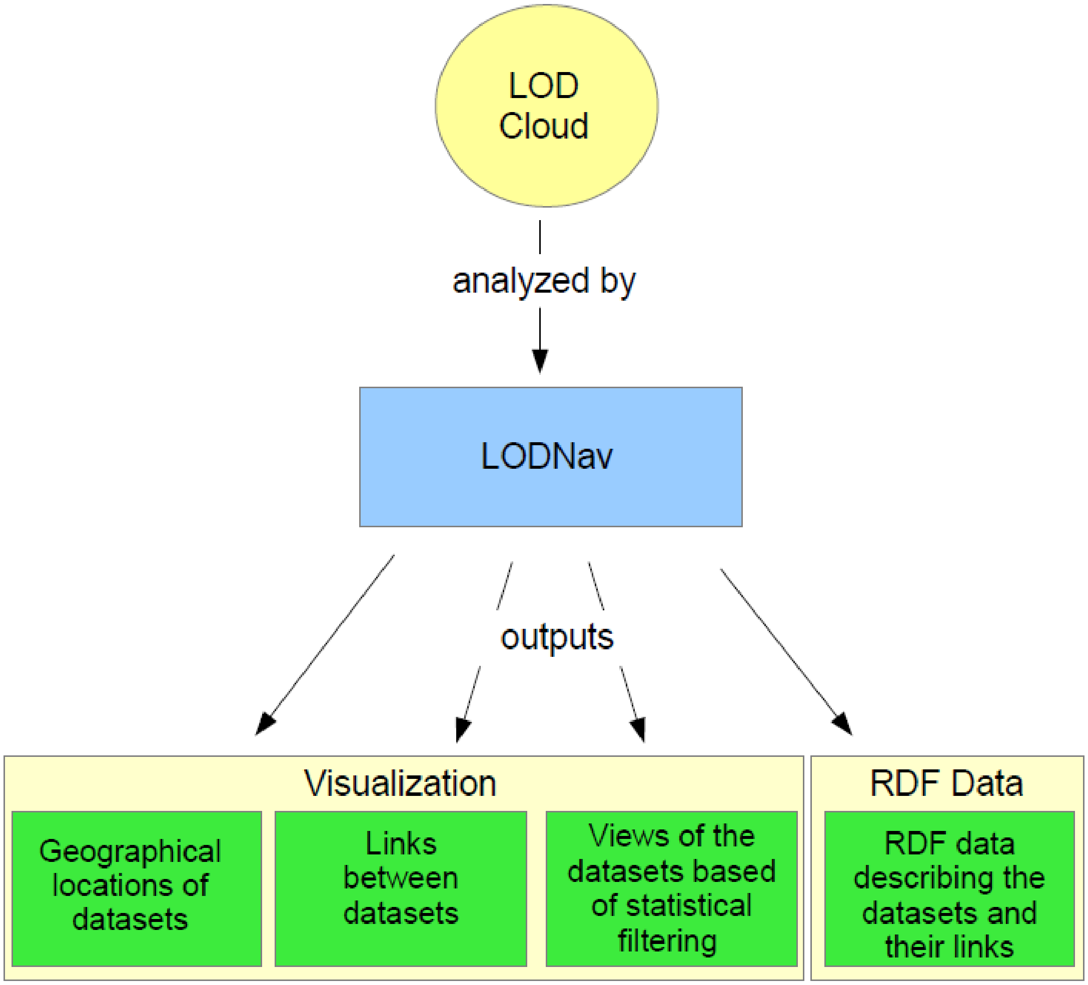

LOD Cloud
analyzed by
LODNav
outputs
Visualization
Geographical locations of datasets
Links between datasets
Views of the datasets based of statistical filtering
RDF Data
RDF data describing the datasets and their links

Figure 1-3 : Overview of LODNav features

The main contributions of LODNav include:

- a visualization of the LOD Cloud by providing a geographical visualizing of the member datasets using their geographical position and map visualization metaphor

- crawling of the LOD cloud data to create up-to-data visualization of datasets and their links

- provides various methods for filtering and viewing datasets depending on different configuration settings

- LODNav offers an interface for extracting RDF data describing the datasets and the relationships between them.

1.2 Thesis Outline

This remainder of this document will comprise of the following sections. Chapter 2 presents a more in-depth look at the motivation and objects of this research. A background of the supporting technologies is described in Chapter 3. In Chapter 4 the tool LODNav is explained in greater depth, followed by use cases demonstrating the tool's utility in Chapter 5. A discussion is presented in Chapter 6 with mention of the threats to validity, related work, and finishes with some conclusions as well as ideas for future work.

2. Motivation, Questions, & Objective

2.1 Motivation

Increasingly, organizations and individuals are embracing the concept of Linked Data as a method for distributing knowledge and embedding it *in* the Web [52] [33]. As a result we can refer to this as a *Web of Data* as part of a *Giant Global Graph* [7] [22]. The Linking Open Data (LOD) Project was introduced in 2007 as a grassroots effort to bootstrap the Web of Data [33]. This project encourages anyone to convert openly available datasets to RDF data, according to Linked Data principles, then publish and link these datasets to the Web of Data as part of the Linking Open Data Cloud (LODC) [33]. The LODC covers a large variety of themes including geography, people, companies, books, scientific publications, films, music, television, radio, genes, proteins, drugs, clinical trials, statistical data, census results, online communities, reviews, and source code facts [33][39].

The Web of Data on the LODC has the following properties [33]:

- Data can be published by anyone.
- Any kind of data can be published since the Web of Data is generic.
- Disagreeing and contradictory information about an entity can be represented.
- There are no fixed vocabularies in which to represent data.
- By connecting entities with RDF links, the data graph spans all connected data sources allowing the possibility of discovering other data sources at run-time.

- The contained data is self-describing.

- The standardized data access mechanism is HTTP and the standardized data model is RDF. This allows for simplified data access not relying on assorted data models or access interfaces.

While this infrastructure provides the theoretical backbone for the Web of Data, for somebody interested in utilizing or contributing to this network of information there exists some challenges, mainly, (1) determining where datasets are geographically located and discovering where centres of expertise are emerging in the world, (2) the ability to determine which datasets are linked, and (3) obtaining meta-data about the datasets, which is useful to determine the importance and relevance of a particular dataset. This meta-data being information such as the dataset's:

- Website URL
- SPARQL endpoint
- current state of the SPARQL endpoint ('online' or 'offline')
- the availability of the SPARQL endpoint over the last 24 hours
- number of data assertions
- number of links to other datasets
- data license
- description
- location

Contributing towards the LOD requires that any new dataset can be linked to another dataset that is already part of the LODC. However, the LOD portal does not currently provide any support to (dynamically) visualize existing node dependencies (links), which complicates the integration of new nodes within the current LOD scope. Additionally for an individual trying to create an application, which uses information within the LODC it is difficult to determine which datasets can be used for a particular application context. Furthermore, since the linking of data is a relatively new endeavour, it would be useful to see throughout the globe where these concepts are presently being adopted in order to seek individuals/groups that have expertise in this domain.

The research presented in this thesis introduces LODNav, a tool for better understanding the current state of the LODC. The tool makes use of different metaphors to represent the structural information of the LODC. This visualization tool provides a succinct representation of the entire LODC structure and meta-information in one source. The underlying feature is that the datasets of the LODC are displayed as nodes on a geographical map referencing where the dataset is located. As a result clusters or nodes emerge around locations that have begun the initial adoption of Linked Open Principles. In support of this, meta-information is displayed to include the current state of the SPARQL endpoint for the datasets being displayed (whether they are 'online' or 'offline') as well as a visual depiction demonstrating which datasets are linked to one another. Additionally, meta-data describing the datasets is also available. This information can be extracted and exported by the user as RDF data.

A key motivation for this approach is to provide an entry point for parties, who wish to reuse or contribute to the LODC. LODNav aims to provide a novel geographic perspective of the LODC, for better understanding of the current state of the Linking Open Data Project. This can guide users in aligning their own datasets within this global graph of information, help them to improve the use of information captured within the LODC, or discover geographic centres of Linked Data expertise. With a better understanding of how the projects are interconnected and where they are located, the barrier for contributing and reusing data on the LODC may be reduced. Ultimately, this may lead to a larger adoption of the Linking Open Data paradigm while fostering a continued growth and use of the Web of Data.

2.2 Goals and Requirements

The main goal of this research is to provide a novel visualization approach of the LODC. The visualization approach proposed by LODNav uses a geographic metaphor which displays the datasets of the LODC at the location where the dataset is located with the purpose to see where centres of expertise in Linking Open Data are emerging.

Requirement #1: Visualizing centres of expertise in Linking Open Data

The adoption of Linked Data is a collaborative community-based effort which requires the linking datasets from various different groups. The principles of Linked Data are presently not being adopted uniformly across the globe and certain geographic regions have laid some of the initial groundwork. By understanding where datasets of the LODC are located, interested parties can understand where centres of expertise are emerging and get an idea of which groups can be sought after to further continue collaboration or knowledge-sharing activities.

Requirement #2: Displaying an overview of all the datasets of the LODC and their respective links

In order for a visualization of the LODC to be complete an entire overview of all the datasets must be presented in a single view to the user.

Requirement #3: Filtering mechanisms to provide different views of the data held within the visualization

Capabilities must be provided in order to provide filtering of the presented data in order to focus on particular regions of interest in the data.

Requirement #4: Exporting the data held within the LODC visualization

In order to reuse the presented information there must be a manner to extract the information held within the LODC visualization. Additionally, there is presently a lack of support for a data dump of the information held within LODC by other tools which provides users with an overview of the datasets and their interconnections.

2.3 Research Questions

When constructing the tool LODNav several research questions were considered, which were derived from limitations of the current status of the LODC (see Motivation section).

Main Research Question

> **RQ1:** What is a novel visualization approach for viewing the datasets of the Linking Open Data Cloud in order to see the regions of the adoption of Linked Open Data principles across the globe?

Research Sub-Questions

Given a particular visualization approach of the Linking Open Data Cloud:

RQ2.1: can the meta-data and some additional status information associated with each dataset be visualized?

RQ2.2: can a different approach be introduced for visualizing the datasets and their links?

RQ2.3: can support for extracting the information describing the datasets and their links be provided?

2.4 Objective

The research is driven by the fact that visualizing the LODC can benefit various stakeholders. The objective of this research is to create a visualization tool, accessible to the public, visualizing the dataset interconnections of the Linking Open Data Project, but also taking advantage of available meta-information to improve the information content being visualized (e.g., use of geographical, size, and linking information). By taking into account the geographic location of datasets, there is the possibility to see where the adoption of Linked Open Data is taking place, in order to provide insights as to where to locate individuals/groups with knowledge and expertise.

3. Background

This chapter provides an overview of related technologies and work that form the basis of this research. The topics included in this background section are: Semantic Web Technologies, Linked Data, and some example datasets from the Linking Open Data Project.

3.1 Semantic Web Technologies

This section introduces the concept of the 'Semantic Web' in which the idea of 'Linked Data' is built upon. The technologies at the heart of the Semantic Web allow for shared data to be published, linked, queried, and reasoned upon.

3.1.1 The Semantic Web

The term 'Semantic Web' was originally conceived by Tim-Berners Lee, inventor of the World Wide Web. This term is used to describe an interlinked 'Web of Data' which can be traversed and reasoned upon by machines [68] [4]. The data held within the Web of Data will also contain Semantic information describing itself.

The main idea of the 'Semantic Web' is to provide a framework which enables data, described using the Resource Description Framework (RDF), to be shared and reused [67]. This sharing and reutilization of data should occur across community boundaries and not be limited to particular applications [67]. The reason for promoting the adoption of RDF as a common data format on the Web has been developed in hopes to evolve the current document-centric web paradigm into a 'Web of Data' [67].

The organization leading the collaborative effort of furthering the Semantic Web is The World Wide Web Consortium (W3C). The W3C is an international community which develops standards for the Web in hopes to allow the Web to meet its maximum potential [59].

3.1.2 RDF

The core facility of the Semantic Web is that all data be described in RDF format. Once data is in the same format (RDF), different datasets can be linked when identical concepts are represented amongst the datasets. This section will describe what RDF data is and how it can be used to create a Web of Data.

This framework is a set of W3C specifications initially proposed in 1999 [9] and officially published in 2004 [66]. This list of W3C standards continues to evolve. At the time of writing the most current accepted standard dates at 2013-10-29 [63].

RDF is a language intended for representing information that can be identified as resources in the World Wide Web [64]. RDF is intended to be used when information about these resources requires to be handled by an application/machine instead of a human [64].

An RDF statement is represented as a triple [33]. A triple consists of three parts; a *subject*, *predicate*, and *object*. The *predicate* is a piece of information that describes the relationship between the *subject* and the *object*. Thus RDF statements can be conceptualized as node and arc labeled directed graph [33]. See Figure 3-1 for an example.

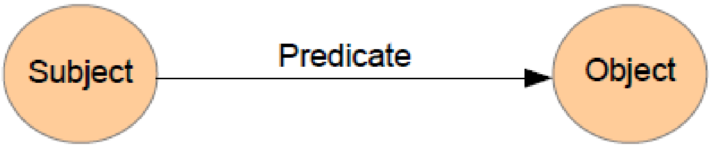


Figure 3-1 : RDF Model

When using RDF all the resources and their relationships are identified using a unique Web-based identifier called a 'Uniform Resource Identifier' (URI) [64]. For now consider a URI as a key to represent something (more information on URIs will appear later).

There are two main types of RDF statements; *Literal Triples* and *RDF Links* [33]. In both cases the *subject* is a URI which describes a resource. The *predicate* is also always a URI in both cases and it describes the relationship between the *subject* and *object*.

In the case of a *Literal Triple* the *object* is a literal value (i.e. a string, integer, date, etc.) which describes something about the *subject*. In the case of an *RDF Link* the *object* is a URI of another resource which has some sort of relation to the *subject.*

To illustrate these concepts an example is given below. Table 3-1 shows a set of RDF triple statements. Figure 3-2 shows how this information is linked together in a graph format.

Subject	Predicate	Object
<http://example.com/Professor1>	<http://example.com/hasName>	<Prof Smith>.
<http://example.com/Professor1>	<http://example.com/teaches>	<http://example.com/Student1>.
<http://example.com/Professor1>	<http://example.com/teaches>	<http://example.com/Student2>.
<http://example.com/Student1>	<http://example.com/hasName>	<Susan>.
<http://example.com/Student2>	<http://example.com/hasName>	<Ahmed>.
<http://example.com/Student1>	<http://example.com/hasGpa>	<4.0>.
<http://example.com/Student2>	<http://example.com/hasGpa>	<3.0>.
<http://example.com/Professor1>	<http://example.com/teachesAt>	<http://example.com/University1>.
<http://example.com/Student1>	<http://example.com/attendsClassAt>	<http://example.com/University1>.
<http://example.com/Student2>	<http://example.com/attendsClassAt>	<http://example.com/University1>.

Table 3-1 : Example set of RDF statements

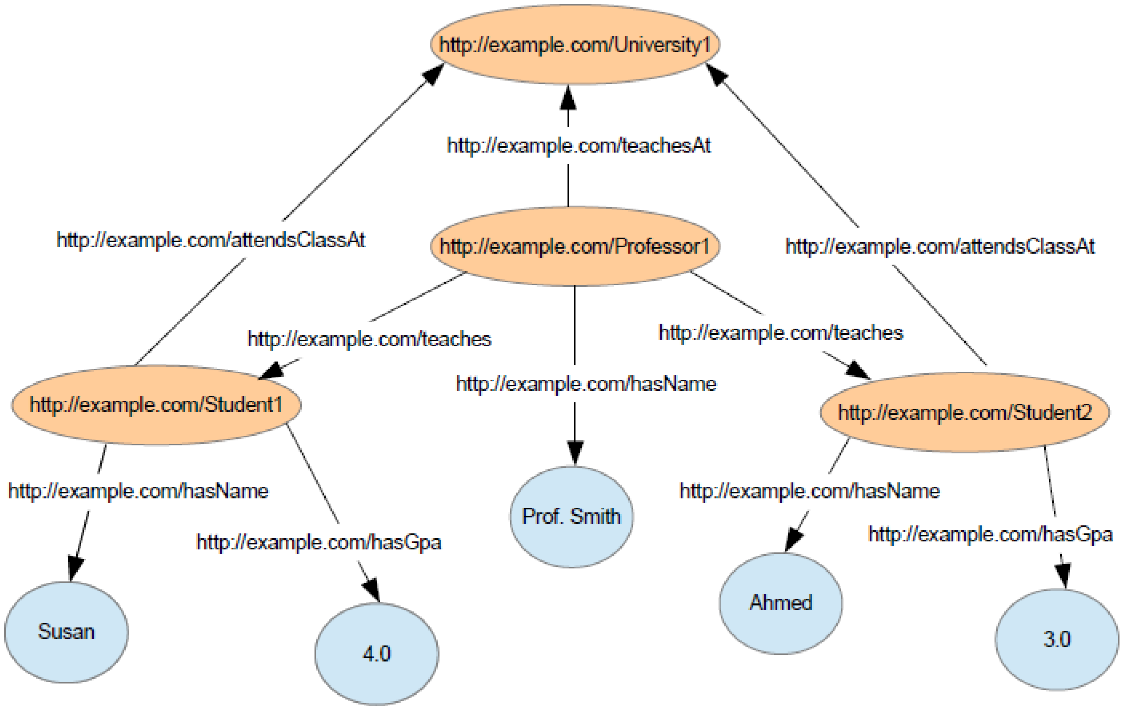

http://example.com/University1
http://example.com/teachesAt
http://example.com/attendsClassAt
http://example.com/Professor1
http://example.com/attendsClassAt
http://example.com/teaches
http://example.com/teaches
http://example.com/hasName
http://example.com/Student1
http://example.com/Student2
http://example.com/hasName
http://example.com/hasGpa
Prof. Smith
http://example.com/hasName
http://example.com/hasGpa
Susan
4.0
Ahmed
3.0

Figure 3-2 : Graph view of the example RDF statements

3.1.3 Ontologies and the Semantic Web

As discussed in the previous section, describing relationships between information with RDF has certain benefits. However, to offer more richness to information described with RDF it is useful to have standardized vocabularies for particular domains. To achieve this with RDF, for information on the Semantic Web, *ontologies* are used [55].

Ontologies can be thought as being formal specifications for shared concepts [31]. Ontologies can be built by using the RDF language to help enhance how data is stored as RDF triples. By using ontologies concepts held within RDF data can be denoted with additional type, property, and relationship information.

With the introduction of ontologies, the URI naming scheme can also be more explicitly described. The following is an example URI with its distinct sections highlighted:

- http://www.example.com/exampleOntology#exampleObject

  1. http://www.example.com – The HTTP portion describing who created the RDF assertion. Additionally, since ideally RDF data should be made accessible on the Web, this part of the URI serves as providing a method for getting this information from a location on the Web.

  2. exampleOntology – This portion mentions which ontology is being used to describe the object.

  3. exampleObject – The object/resource which is being used.

A shortened version to describe an object's URI is as follows:

- exampleOntology:exampleObject

The HTTP portion is often removed to save space and the ‘#’ character is replaced with the ‘:’ character.

Ontologies can be created by anyone and can be extremely specialized for particular fields or tasks. In order to describe ontologies and how they interact with one another the languages RDFS and OWL are used [33].

3.1.4 RDFS – RDF Schema

The Resource Description Framework Schema (RDFS) is a W3C standard, initially published in 1998 with its final recommendation published in 2004 [74] [65]. RDFS extends RDF by providing the ability to define application-specific *classes* and *properties* [33]. This is done with its two basic classes; *rdfs:Class* (class of resources which are RDF classes) and *rdfs:Property* (class of all RDF properties) [33].

Other constructs also exist in RDFS. For annotating resources *rdfs:label* and *rdfs:comment* are used to help provide human-readable information about a resource. To describe relationships between classes and properties the following can be used; *rdfs:subClassOf*, *rdfs:subPropertyOf*, *rdfs:domain*, and *rdfs:range* [33].

RDFS introduces a more distinct separation between the two main types of elements found within an ontology; the *t-box* and the *a-box*. The *t-box* is the information describing the domain concepts and relationships. The *a-box* is comprised of the tangible individuals that are asserted within the data. To describe this in Object-Oriented programming terms; the *t-box* can be considered as being the set of *Classes*, while the *a-box* can be considered the set of asserted *Objects* of those classes.

An example of RDF triples using RDFS are below:

```
<http://www.example.com/foods#food>
     <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
          <http://www.w3.org/2000/01/rdf-schema#Class>.

<http://www.example.com/foods#apple>
     <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
          <http://www.w3.org/2000/01/rdf-schema#Class>.

<http://www.example.com/foods#apple>
     <http://www.w3.org/2000/01/rdf-schema#subClassOf>
          <http://www.example.com/foods#food>.
```

These triple statements can be more succinctly represented using the *prefix* command as follows:

```
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix foods: <http://www.example.com/foods#>.

foods:food rdf:type rdfs:Class.
foods:apple rdf:type rdfs:Class.
foods:apple rdfs:subClassOf foods:food.
```

In this example we see that the resource **`apple`** is a subclass of the class **`food`**.

3.1.5 OWL - Web Ontology Language

The Web Ontology Language (OWL) is a Semantic Web language with increased modeling primitives which enhances the expressivity seen in RDFS [70] [33]. OWL became a W3C recommendation in 2004 and was revised in 2009 as OWL2 [70]. Currently, OWL is the most popular language for ontology creation [74].

The goal of OWL is the same as that of RDFS; defining ontologies which include classes, properties, and their relationships for particular application domains as RDF data [74].

However, OWL can model more complex relationships and as a result applications can be built with enhanced reasoning capabilities [74].

One of the key features of OWL in terms of Linked Data is its ability to create links between RDF resources in separate datasets [33]. There are many modeling primitives for doing so, however the most evident is the primitive **owl:sameAs**. This statement can be used to state that two distinct URIs refer to the same resource [60]. When a link is in place between URIs a machine is able to traverse the link and extract richer information about that resource. For an example of this consult Figure 3-3.

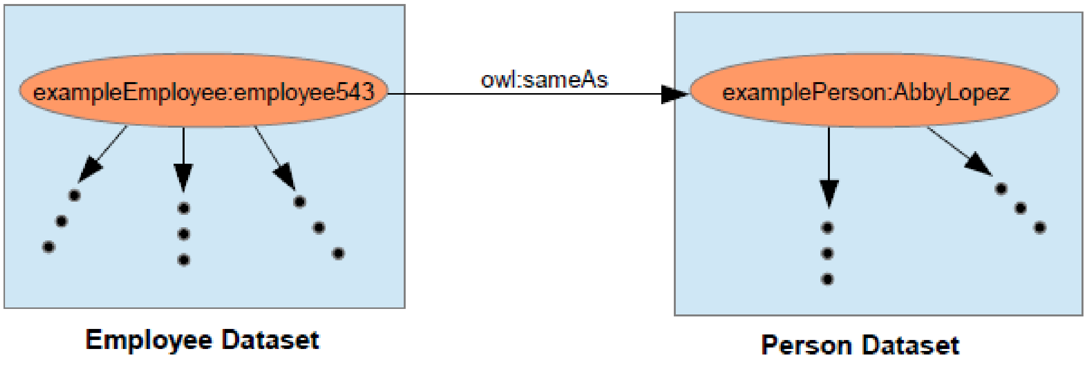


Figure 3-3 : Example of linking URIs with owl:sameAs

In Figure 3-3 we see that `employee543` is the same individual as `AbbyLopez`. If an application using the **Employee Dataset** was compiling information about `employee543` it could also use any information connected to `AbbyLopez` in the **Person Dataset**. Although this is an extremely simple example, it shows how an application can traverse linked data to make richer descriptions of the contained resources.

3.1.6 SPARQL – SPARQL Protocol and RDF Query Language

SPARQL is an RDF query language and a protocol for accessing RDF data over the Semantic Web [74]. SPARQL 1.0 became a W3C standard in 2008 and has been revised to SPARQL 1.1 in 2013 [62].

SPARQL 1.0 provided four different forms of query [74];

- **SELECT** – Extracts raw data for a given query.
- **CONSTRUCT** – Extracts a valid RDF structure for a given query.
- **ASK** – Returns a True/False result for a given query.
- **DESCRIBE** – Extracts RDF from a source, however the content of the graph is the decision of the query processor, not the actual query. This is used when the query performer does not know much about the data graph and needs more information.

SPARQL 1.1 included the ability to modify the RDF data inside a data source. This includes addition/removal of triples and the creation/deletion of graphs [61].

- **INSERT DATA** – Adds triples into an RDF graph. Creates a graph if the graph doesn’t already exist.
- **DELETE DATA** – Removes triples from an RDF graph.

It is out of the scope of this thesis to provide a detailed explanation of SPARQL syntax, however an example query is provided below. The following SELECT query finds airports within 50 miles of London [25]. Note that in SPARQL the '#' character is used to comment out a line of code.

```
# Find airports near London
# List of prefixes
PREFIX geo-pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX omgeo: <http://www.ontotext.com/owlim/geo#>
PREFIX dbpedia: <http://dbpedia.org/resource/>
PREFIX dbp-ont: <http://dbpedia.org/ontology/>
PREFIX ff: <http://factforge.net/>
PREFIX om: <http://www.ontotext.com/owlim/>

# Result description
SELECT distinct ?airport ?label ?RR

# Query pattern
WHERE {
        dbpedia:London geo-pos:lat ?latBase ;
                       geo-pos:long ?longBase .
        ?airport omgeo:nearby(?latBase ?longBase "50mi");
                 a dbp-ont:Airport ;
                 ff:preferredLabel ?label ;
                 om:hasRDFRank ?RR .
}

# Query modifier
ORDER BY DESC(?RR)
```

SPARQL is necessary for querying RDF graphs from a data source. There are three types of data sources that can be queried; RDF files, Triple Stores, and SPARQL endpoints.

- **RDF file** – A file containing RDF can be loaded by an application into its corresponding graph and be queried against. RDF can be held in several different file formats (i.e. .rdf, .xml, .nt, .owl, .ttl).

- **Triple Store** – A specialized database used for the storing and retrieving of RDF statements. Every record in the database must follow the triple format of *subject-*

predicate-object. This is in contrast to a Relational Database System with modifiable tables and schemas. A triple store can also be referred to as a 'RDF Data Store' as well as a 'RDF Database' [74].

- **SPARQL Endpoint –** A web service interface for human or machine users to provide SPARQL queries and receive results. SPARQL endpoints are important for furthering the discussion of Linked Data because they are the means in which open datasets are made available for public consumption [69].

3.2 Linked Data

The previous section introduced RDF data being a core aspect of the Semantic Web technologies with RDF being the initial building block on which the ideas of Ontologies, RDFS, OWL, and SPARQL are all built upon.

In what follows, a more detailed review of the concepts of Linked Data will be presented, including coverage of the basic rules for linking data and an example of linked data use from the Linking Open Data Project.

3.2.1 Linking Data Principles

The term Linked Data describes data that is published on the Web in such a way that the data is machine-readable, the meaning behind the data must be explicitly defined, the data must contain links to externally located datasets, and the data must in turn be open in so that it can be linked to/from other datasets [7].

This concept of Linked Data was proposed initially by Tim-Berners Lee in 2006 as part of a Web architecture note he made available, entitled 'Linked Data' [71] [33] [74]. He introduces the following principles one should follow when publishing and linking Linked Data on the Web. These principles are:

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

Each of these four principles will be reviewed in more detail to illustrate their significance to Linked Data.

1) Use URIs as names for things

In order for something to be represented or published on the web it must have a unique identifier. This first principle promotes the idea that anything can be a resource with a URI, not just Web documents or digital content [33]. URIs can be used to represent things such as people, locations, food, films, colours, and even more significantly, the relationships between resources. A *resource* can be considered as a *thing of interest* which is a URI. With relationship information between resources being described as URIs, networks of information can be created. This principle enhances the scope of what the Web encompasses from an interconnecting network of digital content to something that can represent any concept in the physical world [33].

2) Use HTTP URIs, so that people can look up those names

A description of using URIs within RDF was given in Section 3.1, however it was not explained why these URIs should begin with '`http://`'. That will be explained in this section here. The idea is that a URI representing Linked Data should be dereferenceable over the Web. That is, it can be looked up by HTTP clients using the HTTP protocol [33]. By using HTTP URIs this is all made possible. There are many different URI schemes for different applications, but for Linked data HTTP URIs are used.

HTTP URIs serve another advantage when naming Linked Data resources. It allows for a decentralized manner for naming things since every owner of a domain name can create new URI with their particular domain [33].

3) When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL)

To allow a generalized method for accessing resources on the Web to be used by a broad range of applications, it is paramount that there is an agreed standard. With the adoption of HTML as a standardized format for representing documents on the Web, the Web was enabled to become the scale that it is [33]. Similarly for Linked Data, by using RDF as the standardized format, the scope of the information held within Linked Data can be scaled up significantly. The advantage of RDF is that it provides a simple data model that works specifically within the architecture of the Web [33]. RDF data has the ability to represent data from any domain and allows for the linking of different datasets.

4) Include links to other URIs so that they can discover more things

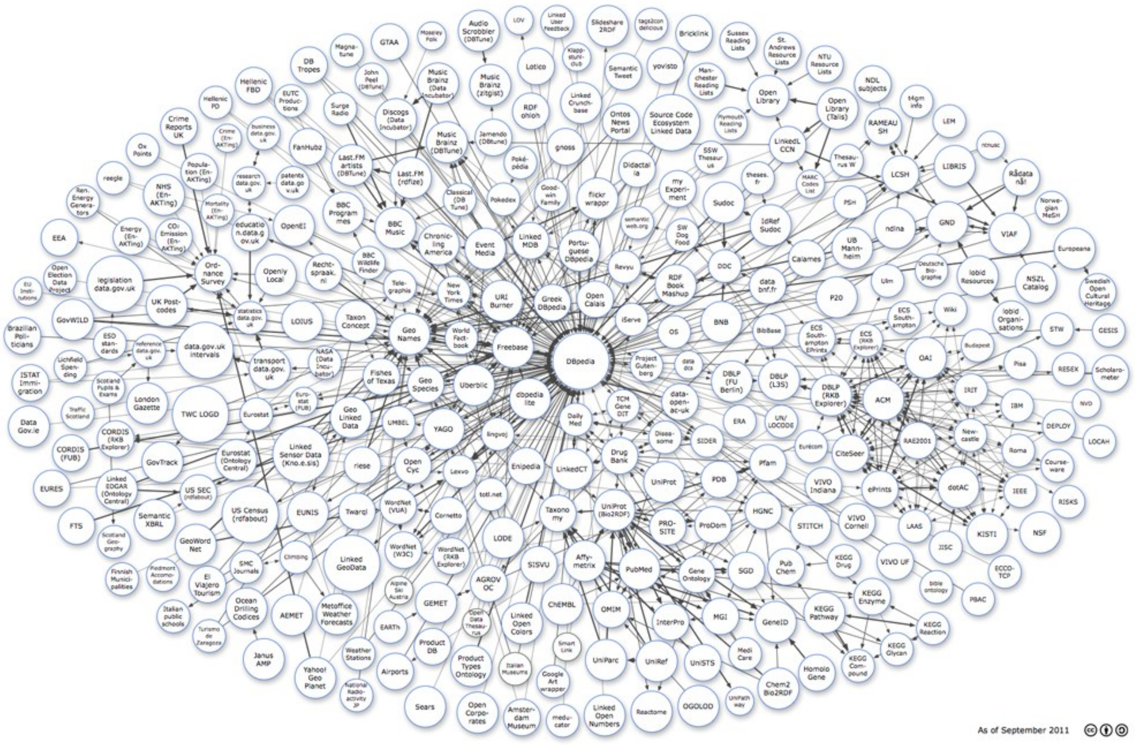
This principal states the significance of creating RDF links to other RDF data sources contained on the Web. Links with the context of Linked Data have type information associated with them and allows for the description of relationships between concepts. When RDF links are in place between resources a machine can traverse these links and discover richer information about concepts. This allows for a semantically rich Web of Data to emerge that is machine-readable.

3.2.2 Linking Open Data Project (LOD)

Linked Data is the idea of connecting information by following the principles which were described in Section 3.1.1. Currently the most visible implementation of these principles is the ‘Linking Open Data Project’ [7]. This project is a community driven effort, founded in 2007 and is supported by the ‘W3C Semantic Web Education and Outreach Group’ [6]. The purpose of this project is to extend the Web by providing a data commons space [43]. This is accomplished by publishing different open data sets in RDF format and then creating RDF links between data items describing similar concepts between the different data sources [43].

Figure 3-4 shows an overview of the member datasets of the Linking Open Data Project as of 2011[3].

[3] http://lod-cloud.net



As of September 2011

Figure 3-4 : Linking Open Data Cloud diagram [50]

A total of 295 open datasets have been interlinked as part of the LODC (as of September 2011 [43]), containing roughly 31 billion RDF triples and 504 million RDF Links [43]. Each dataset provides a SPARQL endpoint for extracting RDF information of HTTP.

While it is apparent that the LODC captures a large amount of data, it is not clear how useful this information may actually be. Research by [11] has illustrated the difficulties when working with the LODC. The work mentions the following limitations associated with the Linking Open Data Cloud project related to the *discoverability*, *interoperability*, *efficiency*, and *availability* of the data [11]:

- *Discoverability* – About one-third of the endpoints provide descriptive meta-data which severely limits the ability to discover and reuse the data.

- *Interoperability* – There is inconsistent support for certain SPARQL commands (i.e `ORDER BY`) and the adoption of SPARQL 1.1, making it difficult to access this data in a reliable manner.

- *Efficiency* – The performance of the endpoints vary greatly for generic queries by up to 3-4 orders of magnitude.

- *Availability* – Based on an experiment of 27 months it was found that only 32.2% of the datasets had monthly up-times of 99-100%.

While the Linking Open Data Project has its limitations, the fact remains that this initiative has managed to link a large amount of data using RDF technology spanning a large variety of domains. In the upcoming section (Section 3.3) some of the datasets which comprise this Web of Data will be examined more closely.

3.3 Linking Open Data Project Datasets

In this section we take a closer look at how datasets can be integrated in the Linking Open Data Project, as well as a description of a couple of the linked datasets and their role within the Web of Data.

The datasets present in the Linking Open Data Project should be open data. That is, they should be freely available to use without any copyright restriction or other forms of control. However, [50] notes that not all data the data publishers explicitly describe the license with their data.

3.3.1 Linking to the Linking Open Data Project

In order for a project to be part of the Linking Open Data Project Web of Data certain criteria must be met. The following points are taken from [50]:

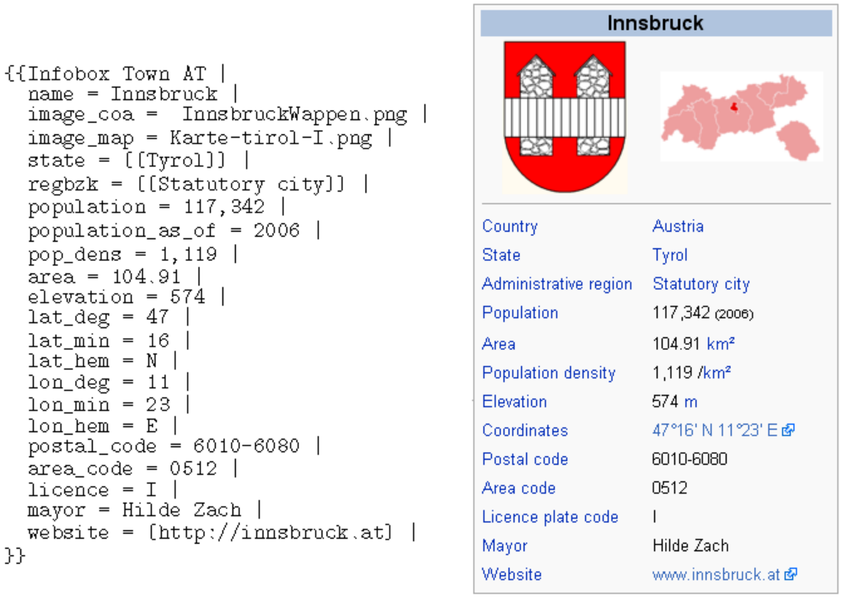
- The URIs must be resolvable over *http://* (or *https://*)
- They must resolve, with or without content negotiation, to *RDF* data in one of the popular RDF formats (RDFa, RDF/XML, Turtle, N-Triples)
- The dataset must contain at least 1000 triples.
- The dataset must be connected via RDF links to a dataset that is already in the diagram. A dataset must use URIs from another dataset, or vice versa. The minimum required number of links is arbitrarily set at 50.
- The entire dataset must be accessible by means of either RDF crawling, an RDF dump, or through a SPARQL endpoint.

If these criteria are not met, then a dataset cannot be officially become part of the Linking Open Data Cloud Project.

3.3.2 DBpedia

Wikipedia is the largest online encyclopedia containing over 3.1 million articles [74]. These articles are created and maintained by authors across the globe using a wiki format. This provides a great wealth of information however; it is text based and is not machine-readable. The DBpedia project was initially founded by researchers from University of Leipzig, Freie Universität Berlin, and OpenLink Software; the initial release of the dataset occurred in January 2007 [74]. At the time of writing the most current version of DBpedia is *DBpedia 3.9* released in September 2013 [20]. The English version contains descriptions of 4.0 million “things” with 471 million “facts” [21]. In total the full DBpedia dataset has information in 120 different languages describing 12.6 million “things” and contains 2.46 billion RDF triples [21]. There exists 45 million RDF links to external RDF datasets [21].

Since most of the data in Wikipedia is in unstructured free text, the DBPedia project extracts information contained in the infoboxes of Wikipedia pages (generally located on the right side of a Wikipedia article). This information follows a structure which can be converted into RDF. See Figure 3-5 below for an example infobox from Wikipedia for the municipality of Innsbruck, Austria [21]. The DBPedia project team created and made the DBpedia ontology publicly available. The ontology is a mapping of the structured information contained in the structured infoboxes to classes and properties for the classes.



```
{{Infobox Town AT |
  name = Innsbruck |
  image_coa =  InnsbruckWappen.png |
  image_map = Karte-tirol-I.png |
  state = [[Tyrol]] |
  regbzk = [[Statutory city]] |
  population = 117,342 |
  population_as_of = 2006 |
  pop_dens = 1,119 |
  area = 104.91 |
  elevation = 574 |
  lat_deg = 47 |
  lat_min = 16 |
  lat_hem = N |
  lon_deg = 11 |
  lon_min = 23 |
  lon_hem = E |
  postal_code = 6010-6080 |
  area_code = 0512 |
  licence = I |
  mayor = Hilde Zach |
  website = [http://innsbruck.at] |
}}
```

Innsbruck	
Country	Austria
State	Tyrol
Administrative region	Statutory city
Population	117,342 (2006)
Area	104.91 km²
Population density	1,119 /km²
Elevation	574 m
Coordinates	47°16′ N 11°23′ E
Postal code	6010-6080
Area code	0512
Licence plate code	I
Mayor	Hilde Zach
Website	www.innsbruck.at

Figure 3-5 : Example of a Wikipedia infobox and its structured data

The DBpedia project works to extract much of the information contained in these articles and turns it into RDF data [1]. DBpedia is a crowd-sourced community initiative which extracts information from Wikipedia and makes this information accessible on the Web [1] [19]. DBpedia has become a core dataset within the Linking Open Data Project, with 185 nodes linking to DBPedia (other nodes linking to DBPedia) as well as containing links being instantiated from DBPedia to 30 other dataset nodes in the LOD (outgoing links). From the Linking Open Data Project Cloud Diagram in Figure 3-4 it can be noted that DBPedia is the central node and has a large number interconnecting links.

Accessing DBpedia content

DBpedia data can be accessed via three different methods: *Linked Data*, a *SPARQL Endpoint*, and downloadable *RDF dumps* [21].

1) Linked Data:

The data within DBpedia is linked to other external datasets of the Linking Open Data Project. This is the access method most pertinent to this thesis. An example of such links can be shown for the municipality of Busan, South Korea below:

```
<http://dbpedia.org/resource/Busan> owl:sameAs <http://sws.geonames.org/1838524/> .
```

This link connects the DBpedia URI for Busan with the Geonames URI for the same city [1]. An application agent can then traverse this link and retrieve additional information about Busan published within the Geonames dataset.

2) SPARQL Endpoint:

DBpedia provides a SPARQL endpoint at `http://dbpedia.org/sparql`, which client-applications can use to send SPARQL queries to retrieve data from DBPedia. This type of access is appropriate when a client application developer has prior knowledge of what information they wish to retrieve [1]. In order to protect the SPARQL endpoint service from being overloaded with processes there is a limit in place to the amount of information which can be retrieved for a given query [1]. The *Virtuoso Universal Server* is used to host this SPARQL endpoint [1].

3) RDF Dumps

For organizations interested in using larger RDF graphs than which can be retrieved with the previous two methods or to download a complete dataset, N-Triple serializations can be used for downloading the complete dataset from the DBpedia website [1].

3.3.3 SECOLD

SECOLD is an abbreviation for *Source code ECOsystem Linked Data* and is another member dataset of the Linking Open Data Project Cloud. SECOLD is the first online Linked Data repository of RDF triples containing facts about open source software development projects [54] [39]. The creation of SECOLD is part of the research conducted by the *Ambient Software Evolution Group* at Concordia University in Montreal, Canada. The first version of SECOLD was published in January 2011 and the current version, SECOLD 2.0, was published in May 2012 [54].

SECOLD contains 1.5 billion RDF triples extracted from crawling 18,000 open source Java projects from different internet sources [40]. These Java projects were taken from open source project repositories located on SourceForge and Apache.

SECOLD is a member of the Linking Open Data Project and has RDF links to the datasets DBpedia, OpenCyc, and Freebase. These RDF connections in the Linking Open Data cloud can be seen below in Figure 3-6.

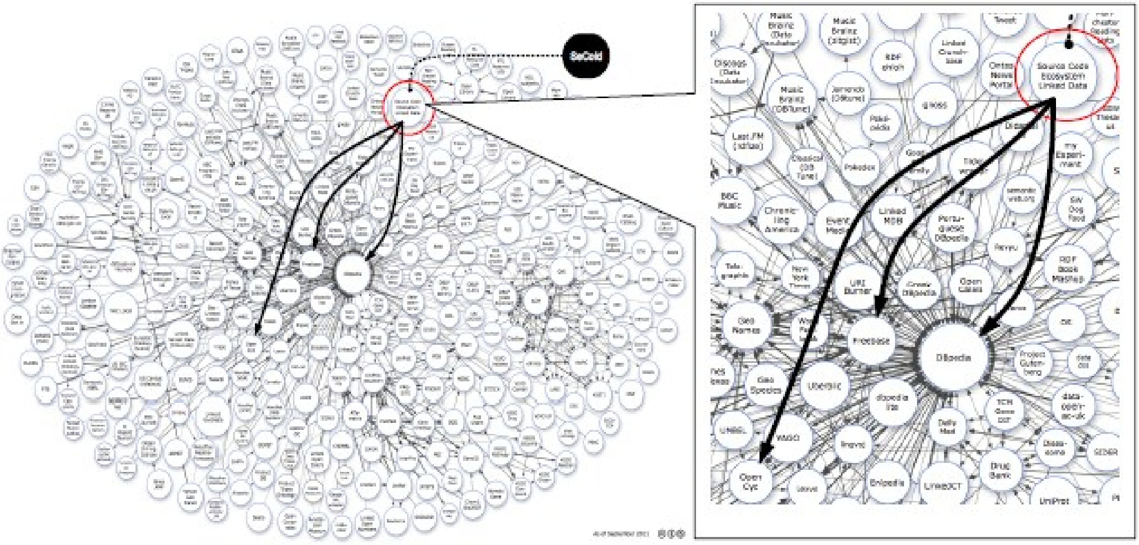


Figure 3-6 : SECOLD on the Linking Open Data Cloud [40]

The information extracted from the crawled Java projects includes; source code, bug/issue tracker information, and versioning systems. All the extracted information is converted into RDF and is interlinked [54]. In order to publish information as RDF a common vocabulary (ontology) must be used along with the generation of unique identifiers. This was done by the SECOLD project by creating a family of vocabularies called SECON – Source code Ecosystem ONtology family [39]. Within this family are the ontologies SOCON and VERON. SOCON – Source Code Ontology Model – is used for publishing RDF facts extracted from source code. VERON – Versioning Ontology Model – is used for publishing RDF facts extracted from version control information.

The purpose of creating this dataset is to provide a resource for the Software Engineering community to perform research upon and to have a standardized representation of source code facts [39]. An example of research which has been conducted on this dataset has worked towards detecting clones of source code across all the projects within SECOLD [40].

3.4 Summary

The chapter provides a background on the underlying principles which the research was built upon. What this research offers is a new metaphor for visualizing and interacting with the Linking Open Data Project Cloud of datasets. Since the Linking Open Data Project is an interconnected Web of RDF data this section began by explaining the technologies of the Semantic Web which are used to create Linked Data. Subsequently this section provided the principles of creating Linked Data. This went on to explain that the Linking Open Data Project is an example of Linked Data and described the characteristics of the project. To conclude, this section highlighted two example datasets which comprise the Linking Open Data Project. These examples illustrate the type of information contained within the network of Linked Data and how this information can be linked together.

4. LODNav: Interacting with the Linking Open Data Cloud

In this chapter, the tool LODNav is presented. This chapter will first specify what LODNav is in Section 4.1. Section 4.2 goes on to explain the architecture of LODNav's two main elements; (1) the extraction of information about the LOD datasets and (2) the visualization using the Google Maps API. The chapter then proceeds to illustrate what LODNav accomplishes and how LODNav contributes to the Software Engineering community. Lastly, this chapter ends with some concluding remarks highlighting how LODNav seeks to fulfill the *Research Objective* set out in Section 2.3.

4.1 Linking Open Data Navigator

4.1.1 LODNav Overview

LODNav – Linking Open Data Navigator - is a visualization tool for providing a better understanding of the current state of the Linking Open Data Project. This tool is a located on its own publicly available website[4].

To continue with specifying what LODNav is, the contributions of LODNav will be restated. The main contributions of LODNav include:

- a visualization of the LOD Cloud by providing a geographical visualizing of the member datasets using their geographical position and a map visualization metaphor

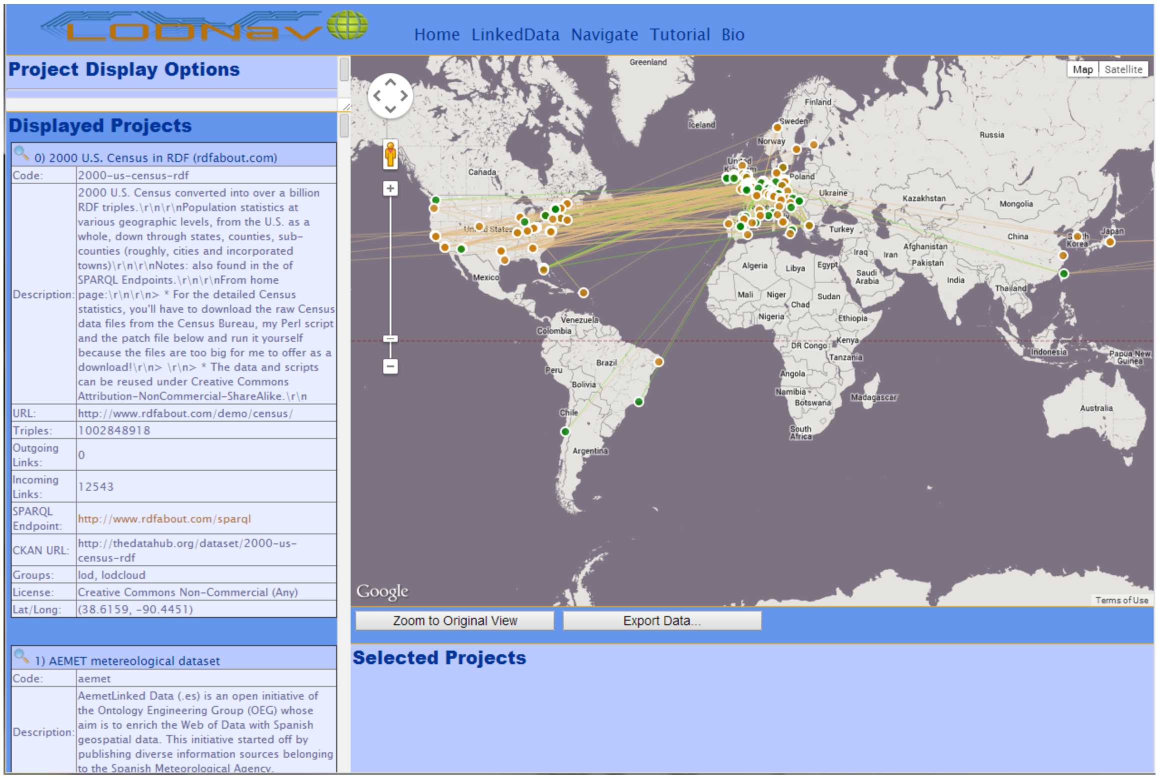
- crawling of the LOD cloud data to create up-to-data visualization of datasets and their links

[4] http://www.lodnav.com/

- providing various methods for filtering and viewing datasets depending on different configuration settings

- an interface for extracting RDF data describing the datasets and the relationships between them.

Figure 4-1 shows the initial view of the LODNav tool presented to a user when accessing the Navigate[5] page of the LODNav website



Project Display Options

Displayed Projects

0) 2000 U.S. Census in RDF (rdfabout.com)	
Code:	2000-us-census-rdf
Description:	2000 U.S. Census converted into over a billion RDF triples.\r\n\r\nPopulation statistics at various geographic levels, from the U.S. as a whole, down through states, counties, sub-counties (roughly, cities and incorporated towns)\r\n\r\nNotes: also found in the of SPARQL Endpoints.\r\n\r\nFrom home page:\r\n\r\n> * For the detailed Census statistics, you'll have to download the raw Census data files from the Census Bureau, my Perl script and the patch file below and run it yourself because the files are too big for me to offer as a download!\r\n> \r\n> * The data and scripts can be reused under Creative Commons Attribution-NonCommercial-ShareAlike.\r\n
URL:	http://www.rdfabout.com/demo/census/
Triples:	1002848918
Outgoing Links:	0
Incoming Links:	12543
SPARQL Endpoint:	http://www.rdfabout.com/sparql
CKAN URL:	http://thedatahub.org/dataset/2000-us-census-rdf
Groups:	lod, lodcloud
License:	Creative Commons Non-Commercial (Any)
Lat/Long:	(38.6159, -90.4451)

1) AEMET metereological dataset	
Code:	aemet
Description:	AemetLinked Data (.es) is an open initiative of the Ontology Engineering Group (OEG) whose aim is to enrich the Web of Data with Spanish geospatial data. This initiative started off by publishing diverse information sources belonging to the Spanish Meteorological Agency.

Selected Projects

Figure 4-1 : Screenshot of initial view in LODNav

[5] http://www.lodnav.com/navigate.html

4.1.2 Data Visualization Requirements and LODNav

In [13] the author summarizes some high level requirements for providing a user interface for data visualization tools as described in [56]. Among the identified requirements for a data visualization user interface are [13]:

1) The system needs the capacity to produce an overview of the underlying data.

2) The system must allow for the filtering of data to allow the user to focus on desired regions of interest.

3) It is essential that the system provides the ability to provide greater details for regions of interest.

These requirements were a driving force when developing LODNav as a data visualization tool. The objective of the implementation is to adhere to these high-level requirements for data visualization. The following sections describe in more detail how LODNav meets these requirements.

The implementation of LODNav as a data visualization tool adheres to these high-level requirements for data visualizations. The functionality of LODNav is described in further sections and it is shown how these requirements are met.

4.2 Architecture of LODNav

This section provides a general overview of the architecture of LODNav. There are two main components for displaying the datasets of the Linking Open Data Project with LODNav.

1. The data used to populate LODNav

2. The program LODNav itself which a human user can interact with.

LODNav is hosted at its website and is accessible over the internet to the general public. In broad terms, when a user accesses the *Navigate* page of LODNav, the precompiled data regarding the Linking Open Data Project is loaded into the application (an HTML page) and the user can interact with LODNav. This overview can be seen in Figure 4-2.

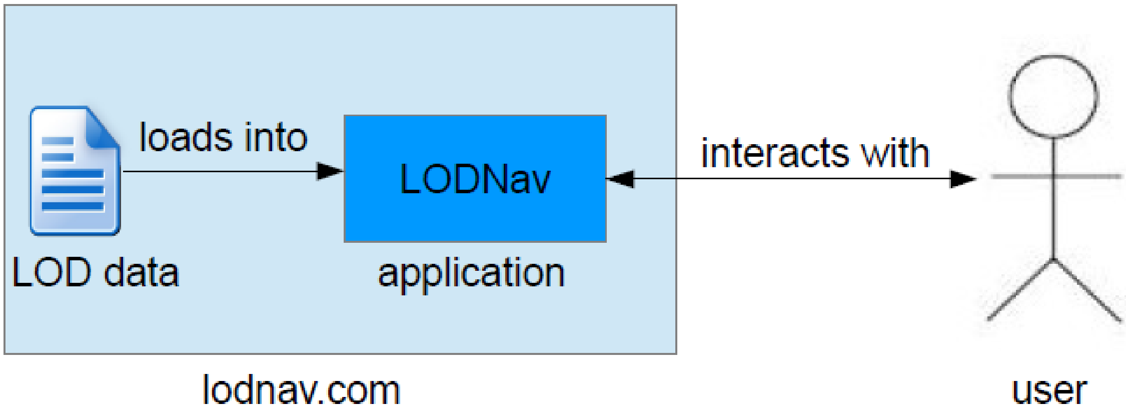


Figure 4-2 : Basic overview of LODNav architecture

The following sections will go into more detail in describing what each of the two parts is composed of.

4.2.1 Extraction of LOD data

The information, which is used to populate LODNav, is obtained by crawling the Datahub's Linking Open Data Cloud organization page [16]. Datahub is a free-to-use data management platform provided by the *Open Knowledge Foundation* and is based on the *CKAN data management system* (also from the *Open Knowledge Foundation)* [14]. CKAN is a free open-source system for managing and publishing data collections [14].

Datahub enables access to several of CKAN's features enabling the publication and management of groups of datasets [14]. The Linking Open Data Cloud is one of these groups of datasets and its information can be accessed via a provided API. Datahub's API can be used to access the number of datasets within the LODC as well as a String of information containing the meta-data for each dataset. This String of data contains some superfluous information and must be parsed in order to retrieve the metadata for the project as well as the statements describing which datasets link to on another.

The meta-data has information for the dataset which is displayed to the user; such as the dataset's URL, number of contained triples, description, SPARQL endpoint, etc. Some important things which are displayed to the user in LODNav for each dataset cannot be retrieved from this String of information and need to be completed in additional processing steps.

The first of these unattainable pieces of information using Datahub's API is the current status of the dataset's SPARQL endpoint (whether *Active* or *Inactive*). The other piece of information relates to LODNav's positioning of the LOD datasets on a global map based on their geographical position and is the task of determining the latitude & longitude for each dataset.

In order to determine the status of a SPARQL endpoint, the following SPARQL query is used against a dataset's SPARQL endpoint every hour:

```
select ?s ?p ?o where {?s ?p ?o} LIMIT 10
```

This query will return the 10 first RDF triples located within the dataset. If results are returned for this query, then it is deemed that the SPARQL endpoint is *Active* for that given hour period. LODNav uses this hourly data to determine the availability of the endpoint for the last 24 hours and reflects this information as part of the metadata provided to the user (shown in Section 4.4).

The first step in obtaining the latitude and longitude of the dataset is to obtain the IP address based on the dataset's URL. This can be done using the Java classes `java.net.InetAddress` and `java.net.URL`:

The next step is to obtain the latitude and longitude and is done using an API from the service *IPInfoDB* [36]. With *IPInfoDB* one can register for an API key and use their provided API. The following example command shows how the API works using the IP address '`109.104.92.188`':

```
http://api.ipinfodb.com/v3/ip-
city/?key=private_key&ip=109.104.92.188
```

The above command provides the following result:

```
OK;;109.104.92.188;GB;UNITED KINGDOM;ENGLAND;DERBY;DE1
3AE;52.9228;-1.47663;+00:00
```

From this result the longitude and latitude can be parsed. In the case of the above example the longitude is `52.9228` and the latitude is `-1.47663`.

Once the longitude and latitude are determined for each project, the meta-data for the datasets are complete and can be serialized into a structured XML document. This document is then hosted at the LODNav website to be used by the LODNav application.

The entire process of compiling the information for each dataset into the XML document has been automated. The developed application crawls Datahub to obtain information about each of the datasets of the Linking Open Data Cloud. Then the additional processing steps are performed to gain a complete view of the meta-data for each dataset in a structured format. Afterwards all the structured information for each dataset is compiled into a single XML document. By programmatically automating these steps, it is a simple process to update the content presented by LODNav.

This program allows the possibility to update the information used by LODNav on a regular basis and provides a more dynamic representation of the Linking Open Data Cloud.

4.2.2 LODNav

The LODNav application is at its core an HTML page hosted at the LODNav website. The HTML page is built predominantly using JavaScript and jQuery. When the page is opened, the XML document (mentioned in Section 4.2.1) is analyzed and each LOD dataset has a JavaScript object created from the loaded information. Each of these JavaScript objects contains the dataset's meta-data, a marker to be placed on the visualization map, and link information for each dataset. The visualization map is implemented using the Google Maps API [29]. This API supports functionalities which allow for placing nodes on a map of Earth (based on geographical positions), the creation of links between the nodes, and the ability to interact with the nodes (mouse press events and mouse hovering over events). A more in depth look at the Google Maps API follows.

4.2.3 Google Maps API

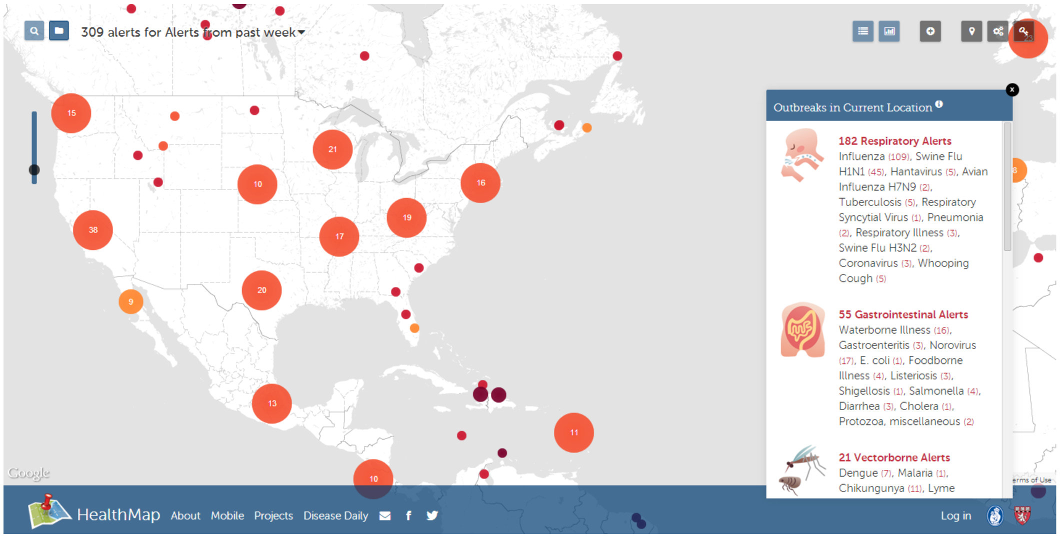
A key objective of this research was to advance the current LODC visualization by providing a more intuitive and richer visualization approach. A subsequent decision was made to plot the datasets on the LOD cloud on a global map based on their geographic locations, which is in contrast to the original LODC graph visualization, which focused only on node links and their crossing. The Google Maps API was selected as the visualization approach due to the following reasons:

- it provides an interface which many internet users are familiar with
- it eliminates the need to develop an entirely new system
- it offers many customizable visualization features
- it provides an interactive user interface
- it has a mature and well supported API managed by Google
- it is free to implement into a webpage

Increasingly the Google Maps API is being used in academic, commercial, and government applications to help provide data visualizations [57]. Some example applications are highlighted in order to highlight the versatility of the Google Map API when applying it to visualizing data from different domains.

One example is HealthMap[6], which visualizes outbreak data of infectious diseases on a global map [27].

[6] http://www.healthmap.org/en/



Outbreaks in Current Location

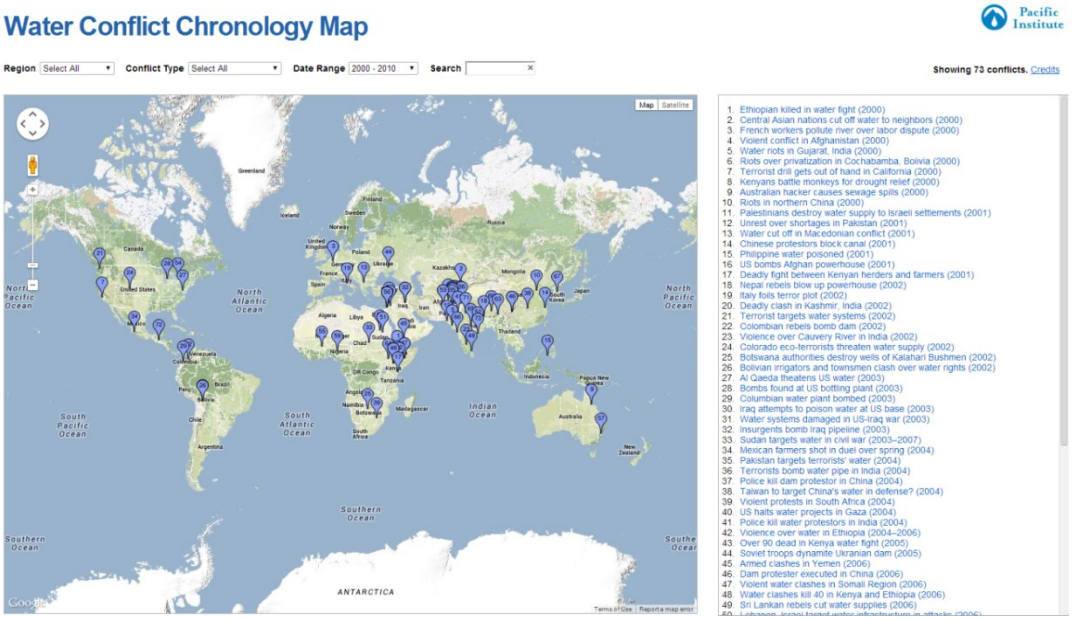
182 Respiratory Alerts
Influenza (109), Swine Flu H1N1 (45), Hantavirus (5), Avian Influenza H7N9 (2), Tuberculosis (5), Respiratory Syncytial Virus (1), Pneumonia (2), Respiratory Illness (3), Swine Flu H3N2 (2), Coronavirus (3), Whooping Cough (5)

55 Gastrointestinal Alerts
Waterborne Illness (16), Gastroenteritis (3), Norovirus (17), E. coli (1), Foodborne Illness (4), Listeriosis (3), Shigellosis (1), Salmonella (4), Diarrhea (3), Cholera (1), Protozoa, miscellaneous (2)

21 Vectorborne Alerts
Dengue (7), Malaria (1), Chikungunya (11), Lyme

Figure 4-3 : Screenshot of the HealthMap application [32]

In [28], a chronology of conflicts over water is presented and a companion website[7] to the book offers a geographic visualization of the locations described in the book using Google Maps.



Showing 73 conflicts. Credits

1. Ethiopian killed in water fight (2000)
2. Central Asian nations cut off water to neighbors (2000)
3. French workers pollute river over labor dispute (2000)
4. Violent conflict in Afghanistan (2000)
5. Water riots in Gujarat, India (2000)
6. Riots over privatization in Cochabamba, Bolivia (2000)
7. Terrorist drill gets out of hand in California (2000)
8. Kenyans battle monkeys for drought relief (2000)
9. Australian hacker causes sewage spills (2000)
10. Riots in northern China (2000)
11. Palestinians destroy water supply to Israeli settlements (2001)
12. Unrest over shortages in Pakistan (2001)
13. Water cut off in Macedonian conflict (2001)
14. Chinese protestors block canal (2001)
15. Philippine water poisoned (2001)
16. US bombs Afghan powerhouse (2001)
17. Deadly fight between Kenyan herders and farmers (2001)
18. Nepal rebels blow up powerhouse (2002)
19. Italy foils terror plot (2002)
20. Deadly clash in Kashmir, India (2002)
21. Terrorist targets water systems (2002)
22. Colombian rebels bomb dam (2002)
23. Violence over Cauvery River in India (2002)
24. Colorado eco-terrorists threaten water supply (2002)
25. Botswana authorities destroy wells of Kalahari Bushmen (2002)
26. Bolivian irrigators and townsmen clash over water rights (2002)
27. Al Qaeda theatens US water (2003)
28. Bombs found at US bottling plant (2003)
29. Columbian water plant bombed (2003)
30. Iraq attempts to poison water at US base (2003)
31. Water systems damaged in US-Iraq war (2003)
32. Insurgents bomb Iraq pipeline (2003)
33. Sudan targets water in civil war (2003–2007)
34. Mexican farmers shot in duel over spring (2004)
35. Pakistan targets terrorists' water (2004)
36. Terrorists bomb water pipe in India (2004)
37. Police kill dam protestor in China (2004)
38. Taiwan to target China's water in defense? (2004)
39. Violent protests in South Africa (2004)
40. US halts water projects in Gaza (2004)
41. Police kill water protestors in India (2004)
42. Violence over water in Ethiopia (2004–2006)
43. Over 90 dead in Kenya water fight (2005)
44. Soviet troops dynamite Ukranian dam (2005)
45. Armed clashes in Yemen (2006)
46. Dam protester executed in China (2006)
47. Violent water clashes in Somali Region (2006)
48. Water clashes kill 40 in Kenya and Ethiopia (2006)
49. Sri Lankan rebels cut water supplies (2006)
50. Lebanon, Israel target water infrastructure in attacks (2006)

Figure 4-4 : Screenshot of Water Conflict Chronology Map application [72]

[7] http://www2.worldwater.org/conflict/map/

The United States Census Bureau created an application[8] to visualize some of its census data from 2010 using the Google Maps API.

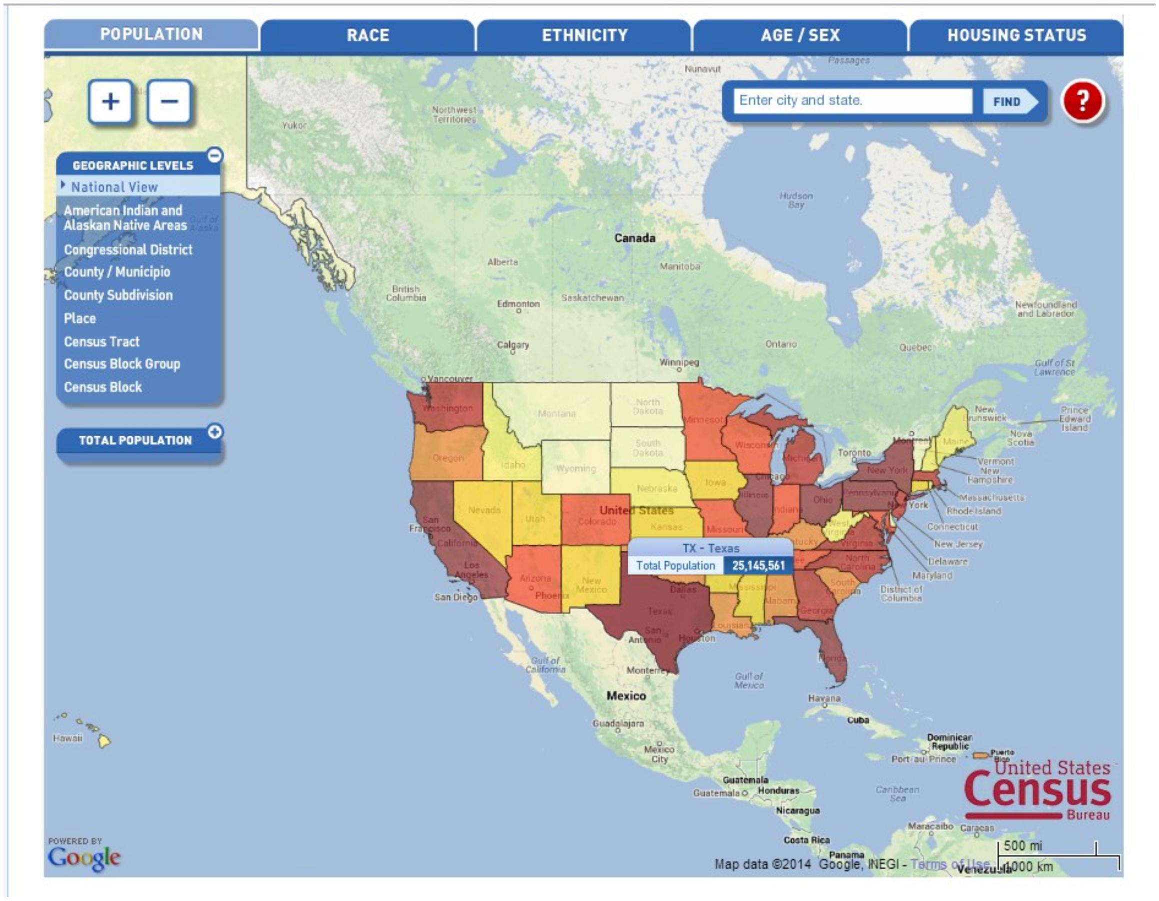


Figure 4-5 : Screenshot of U.S. Census Bureau application [58]

These three examples give a small glimpse into the different application domains in which the Google Maps API has been deployed. Given the flexibility and maturity of the Google Maps API, it was adopted for plotting the geographical locations of the datasets of the LOD cloud.

[8] http://www.census.gov/2010census/popmap/

Overview of using the Google Maps API

The application LODNav is built using HTML and JavaScript/jQuery. In order to embed Google Maps into an HTML page the Google Maps API had to be first included as a JavaScript script. Google offers developers to use a free API key if less than 25,000 map loads per day are used, otherwise a Maps API for Business license can be purchased [30]. In the case of this research the free option was used with an API key provided by Google. The more important elements of the Google Maps API used by the LODNav implementation are `LatLng`, `Marker, Listener`, and `Polyline`. Google provides a tutorial[9] as means to provide a more in-depth introduction of Google Maps to developers.

4.3 Geographical Location of datasets as a Visualization Metaphor

The main research question being addressed in this thesis is how to use a novel visualization approach for viewing the datasets of the Linking Open Data Cloud in order to see geographic regions of expertise and adoption of Linked Open Data (RQ1).

In order to address this research objective, a novel visualization metaphor for the representing both structural as well as meta-information for the LODC portal is introduced. LODNav specifically focuses on providing an enriched visualization of the LODC information by:

Plotting dataset information onto a geographical map based on geographical information associated with each dataset.

[9] https://developers.google.com/maps/documentation/javascript/tutorial

Taking advantage of the available geographical information allows LODNav to place the LOD datasets onto a world map based on their origins. The important part is that the topological space of the interlinked network remains the same despite the transformation. The new coordinate system has no impact on the qualitative properties of the graph [12]. To illustrate this concept a small example will be given. The linking between the following four datasets will be demonstrated; *dbpedia*, *freebase*, *opencyc*, and *secold*.

Note the following links which exist:

- *dbpedia* linksTo *freebase*
- *dbpedia* linksTo *opencyc*
- *freebase* linksTo *dbpedia*
- *secold* linksTo *dbpedia*
- *secold* linksTo *freebase*
- *secold* linksTo *opencyc*

Below in Figure 4-6 these links can be seen between the datasets.

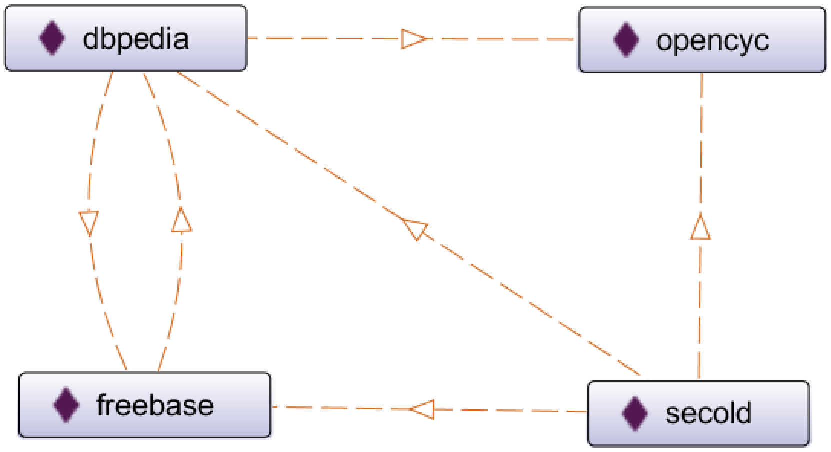


Figure 4-6 : Linking of 4 example datasets

Figure 4-7 highlights where these four datasets are located in the Linking Open Data Cloud diagram and shows which projects are linked. It should be noted that directional lines are not present in this diagram and bidirectional links are shown as a single line.

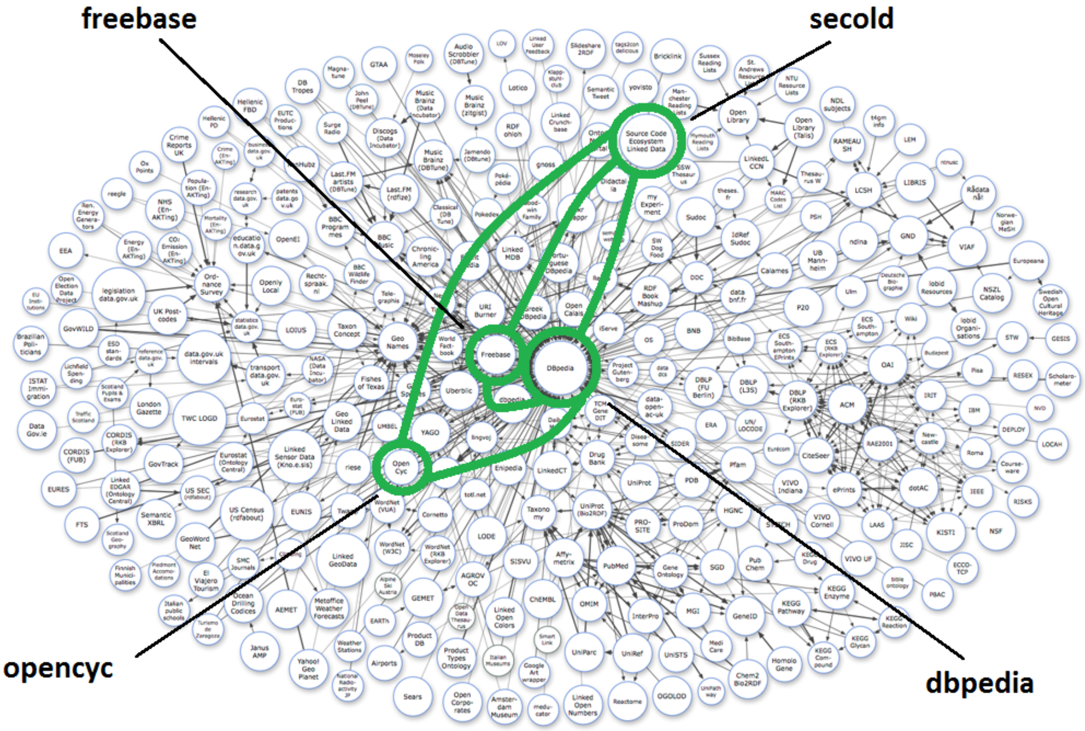


Figure 4-7 : Linking of 4 example datasets in LOD diagram

Figure 4-8 uses LODNav to illustrate where these four datasets are located geographically, as well as their respective interconnections. It should be noted that directional lines are not present in the diagram and the bidirectional links are shown as single lines.

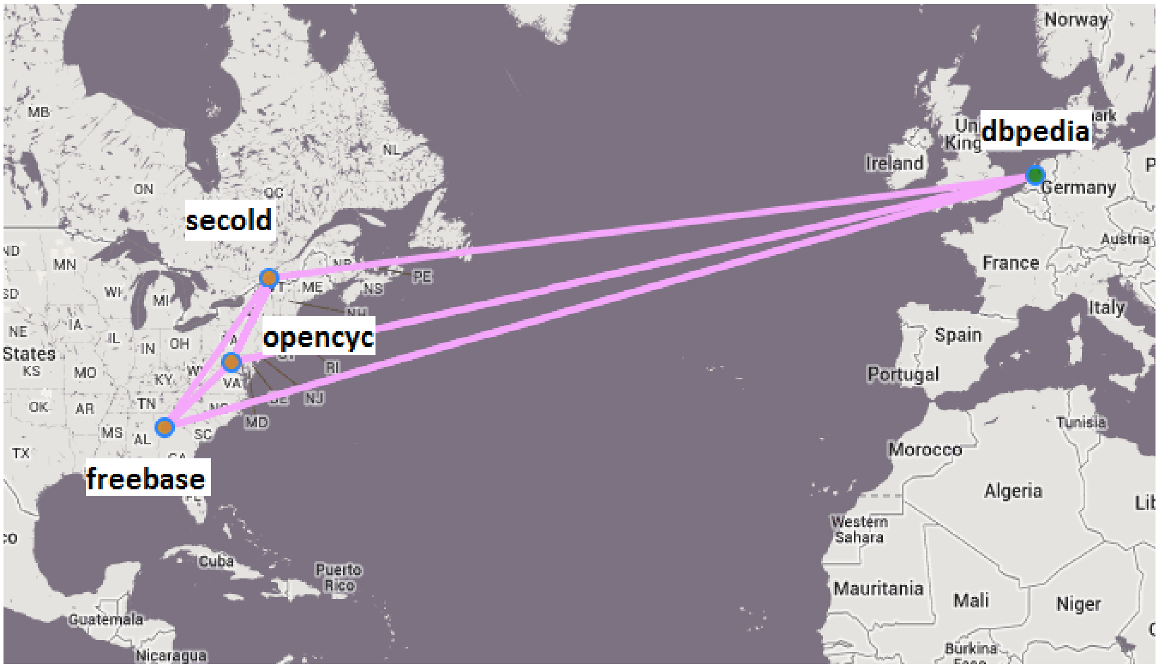


Figure 4-8 : Linking of 4 example datasets using LODNav

This example illustrates, while the metaphor has changed, the actual topology (nodes and their links) conveyed remains the same. This example also highlights that when plotting the datasets geographically, in addition to the structural information also the geographical location of the publisher of the RDF dataset can be shown. Using this visualization approach can provide some meaningful insight on the distribution of LOD datasets across the globe. Figure 4-9 for example shows that the majority of the LODC datasets are located in Europe and then in the United States.

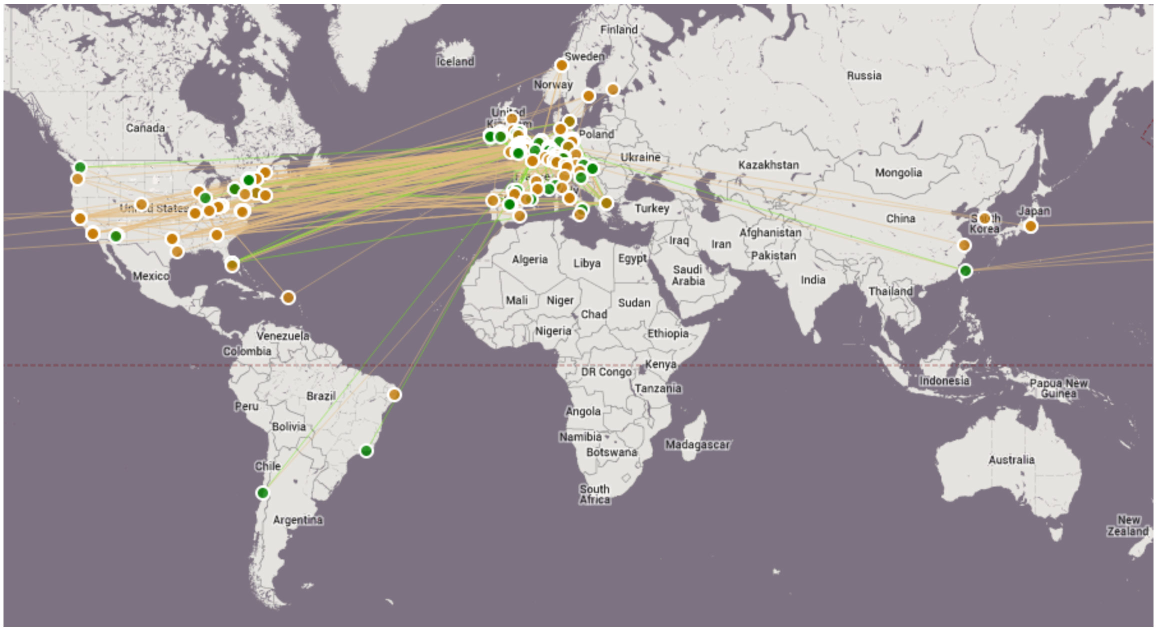
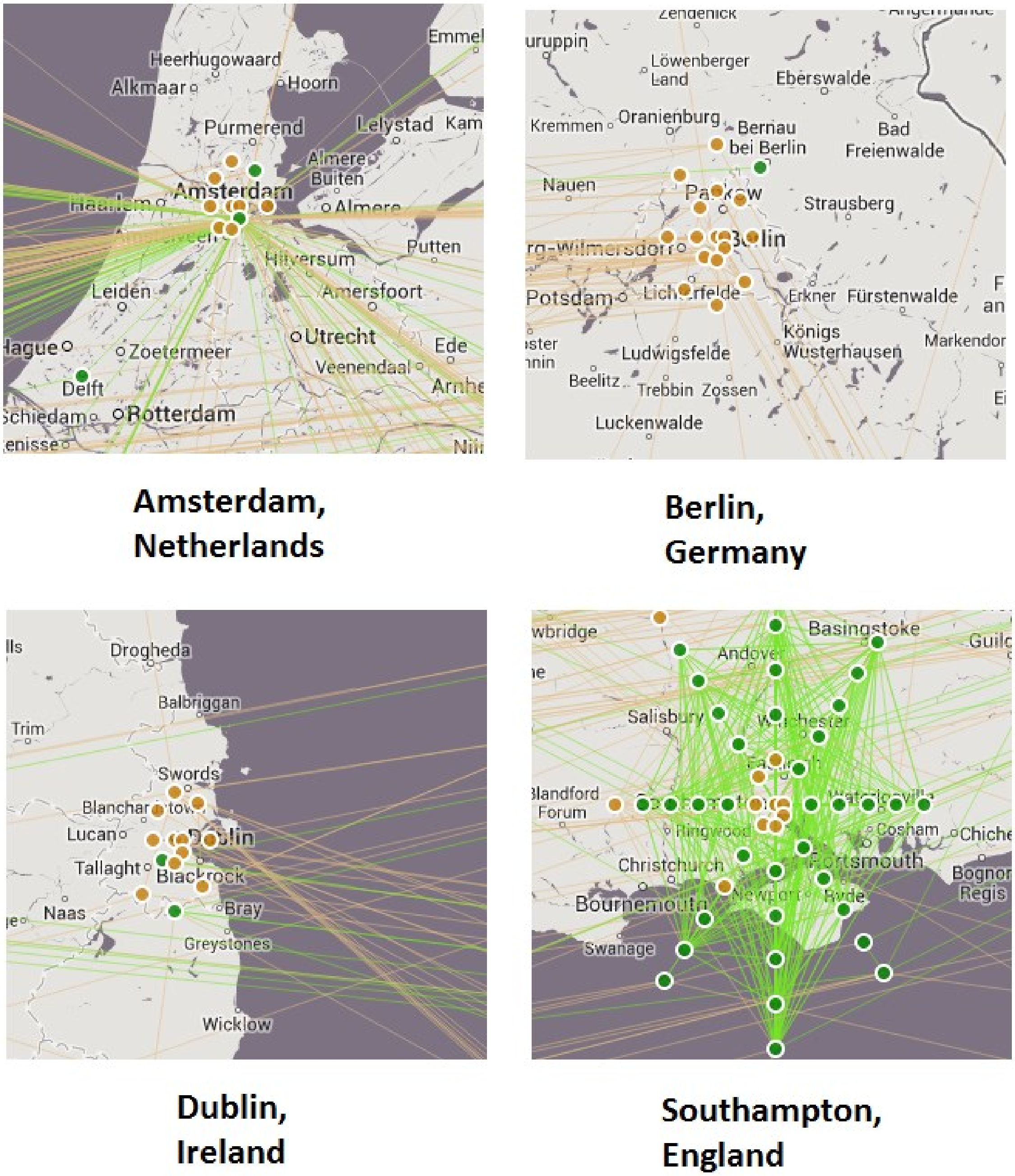


Figure 4-9 : Distribution of LOD datasets on the globe using LODNav

However, using such a geographical layout also introduces some new challenges, when visualizing these datasets. One problem arises when datasets happen to be located at or near the same location. LODNav determines the geographical position of a dataset based on the latitude and longitude of the dataset's IP address, which is derived from the URL address provided with dataset registration for the LODC. This conversion might lead to situations that a number of datasets will have an identical latitude and longitude since they are located in the same city. In order to circumvent this problem LODNav displays these datasets spiralling outward from the contested location. As a result, while the displayed position of the dataset might no longer be accurate from a geographic perspective, clusters will emerge around certain cities demonstrating multiple datasets at that location. See Figure 4-10 for some examples of such spiralling clusters.



Amsterdam, Netherlands

Berlin, Germany

Dublin, Ireland

Southampton, England

Figure 4-10 : Example locations of LOD dataset clusters

This section described in more detail how LODNav takes advantage of a map metaphor to visualize member datasets of the Linking Open Data Project by plotting them onto a geographical map based on where the IP address of the dataset is located. The section also shows how the visualization metaphor used by LODNav preserves the topology of the networked datasets and can offer a different perspective for viewing the datasets. Lastly a closer look at the geographical distribution of the datasets across the globe was given.

4.4 Representing meta-data related to datasets

As part of the registration process of a new dataset with the LODC, different types of meta-information have to be provided. This section focuses on how the visualization approach presented in this research can address (RQ2.1), the visualization of meta-data related to each dataset

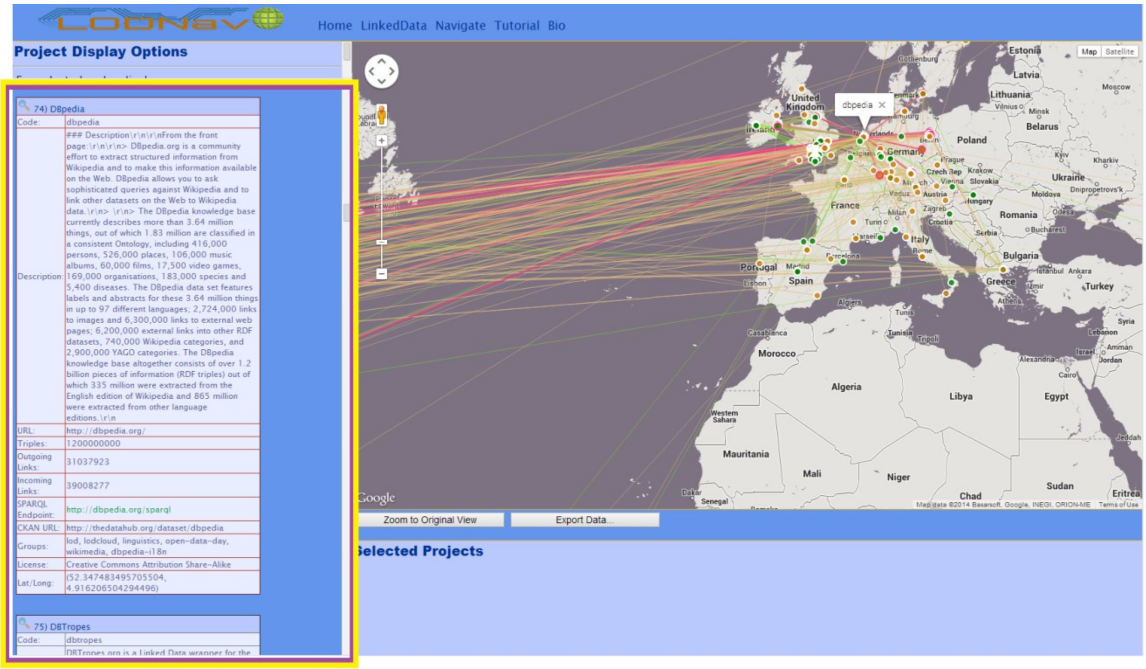
In Section 4.3 it was shown that the visualization approach used by LODNav can be used to plot the geographical location of datasets of the LODC onto a global map. In what follows, the discussion will focus on how LODNav's visualization approach is not only capable to visualize structural information mined from the LODC website, but also meta-information.

While the previous map metaphor allowed for the visualization of detailed node information (geographical location and links to other nodes), no meta-information was conveyed to the user. This meta-information has to be provided at the time of registration of a dataset with the LODC in order for a new dataset to be included in the LODC. The meta-information includes:

- full and abbreviated name of the dataset
- website URL
- SPARQL endpoint
- current state of the SPARQL endpoint (online of offline)

- the availability of the SPARQL endpoint for the last 24 hours
- number of data assertions
- number of outgoing links to other datasets
- number of incoming links from other datasets
- data license
- description
- location

In LODNav this meta-data is displayed as part of a panel on the left side of the screen. The actual meta-information for each dataset in the LODC is displayed in a table within this panel. The meta-information is ordered alphabetically based on the abbreviated names of the datasets (code name). Figure 4-11 shows the section of LODNav, which contains the meta-data for the datasets.



Project Display Options

Selected Projects

74) DBpedia	
Code:	dbpedia
Description:	### Description\r\n\r\nFrom the front page:\r\n\r\n> DBpedia.org is a community effort to extract structured information from Wikipedia and to make this information available on the Web. DBpedia allows you to ask sophisticated queries against Wikipedia and to link other datasets on the Web to Wikipedia data.\r\n> \r\n> The DBpedia knowledge base currently describes more than 3.64 million things, out of which 1.83 million are classified in a consistent Ontology, including 416,000 persons, 526,000 places, 106,000 music albums, 60,000 films, 17,500 video games, 169,000 organisations, 183,000 species and 5,400 diseases. The DBpedia data set features labels and abstracts for these 3.64 million things in up to 97 different languages; 2,724,000 links to images and 6,300,000 links to external web pages; 6,200,000 external links into other RDF datasets, 740,000 Wikipedia categories, and 2,900,000 YAGO categories. The DBpedia knowledge base altogether consists of over 1.2 billion pieces of information (RDF triples) out of which 335 million were extracted from the English edition of Wikipedia and 865 million were extracted from other language editions.\r\n
URL:	http://dbpedia.org/
Triples:	1200000000
Outgoing Links:	31037923
Incoming Links:	39008277
SPARQL Endpoint:	http://dbpedia.org/sparql
CKAN URL:	http://thedatahub.org/dataset/dbpedia
Groups:	lod, lodcloud, linguistics, open-data-day, wikimedia, dbpedia-i18n
License:	Creative Commons Attribution Share-Alike
Lat/Long:	(52.347483495705504, 4.916206504294496)

75) DBTropes	
Code:	dbtropes
	DBTropes.org is a Linked Data wrapper for the

Figure 4-11 : LODNav's section containing dataset meta-data

Table 4-1 provides an example of the meta-data, which is captured for each dataset. The example dataset shown is dataset *114 - Hellenic Police.*

144) Hellenic Police	
Code:	hellenic-police
Description:	The Hellenic Police project encompasses efforts to extract valuable information from Greek Open Data originating from the Ministry of Public Order & Citizen Protection and in particular from the Hellenic Police Department. It involves mainly crime incidents and aims to exploit these in the best possible manner so as to form meaningful scenarios. The primary goal is to provide applications and services that would reveal potentials for the department to improve upon its management procedures, have economic benefits from cost reductions and improvements in its crime prevention efficiency. A secondary but equally important goal is to encourage additional contributions of Greek Open Data as well as of innovative applications and services based on the latter
URL:	http://greek-lod.math.auth.gr/police/
Triples:	145368
Outgoing Links:	64587
Incoming Links:	0
SPARQL Endpoint:	http://greek-lod.auth.gr/police/sparql
Endpoint Availability (last 24 hours):	(100%) 111111111111111111111111
CKAN URL:	http://thedatahub.org/dataset/hellenic-police
Groups:	lodcloud, country-gr
License:	Creative Commons Attribution Share-Alike
Lat/Long:	(40.613783495705505, 22.970416504294494)

Table 4-1 : Example meta-data of a dataset

In table 4-1, the green coloured *SPARQL Endpoint* URL indicates that at the time of writing, this particular SPARQL endpoint was online/active and accepts SPARQL query requests. In the case where the SPARQL endpoint is ‘offline’ the SPARQL URL text will be in orange.

The row with *Endpoint Availability (last 24 hours)* demonstrates how available the dataset’s SPARQL endpoint was over the last 24 hour period. Each hour every SPARQL endpoint is queried and if a result is returned then the endpoint is deemed active for that hour. For the example provided in Table 4-1 we see a String value of:

(100%) 111111111111111111111111

There is a percentage followed by a series of 24 characters of either ‘1’ or ‘0’. Each of the 24 characters correspond to a given hour period. The leftmost character is the most recent hourly period (**Time t**), the second character represents the second most recent time period (**Time t-1**), and so on for all 24 characters. A ‘1’ is used to signify that the SPARQL endpoint was *Active* for the given hour, while a ‘0’ is used to signify the endpoint was *Inactive*. The percentage that is shown before the series of Active/Inactive characters represents the overall availability of the SPARQL endpoint for the given 24 hour period (calculated by the total number of 1s over 24). In the provided example the endpoint was available for 100% of the last 24 hours and is denoted with 24 ‘1’ characters (Appendix A provides the availability of all the LODC SPARQL Endpoints for the time period of 2014-04-12 14:00 to 2014-04-13 14:00).

The meta-data visualization of LODNav enhances the interactive experience with the LODC visualization and provides users with additional insights about the datasets without the need to switch between the LODC website documentation and the graph visualization provided on the LODC site.

4.5 Enriched visualization of links between datasets

While the previous section focused on the basic visualizations provided by LODNav this section will focus in particular on RQ2.2 – and the discussion on how LODNav takes advantage of its map visualization metaphor. As shown in Figure 4-9 (which includes all LODC datasets and their geographical location), a colour scheme of green and orange are used to indicate the status of the individual datasets. Status refers to the fact whether a SPARQL endpoint for a dataset is active (accepts/processes incoming SPARQL queries) or inactive (does not accept/process SPARQL queries).

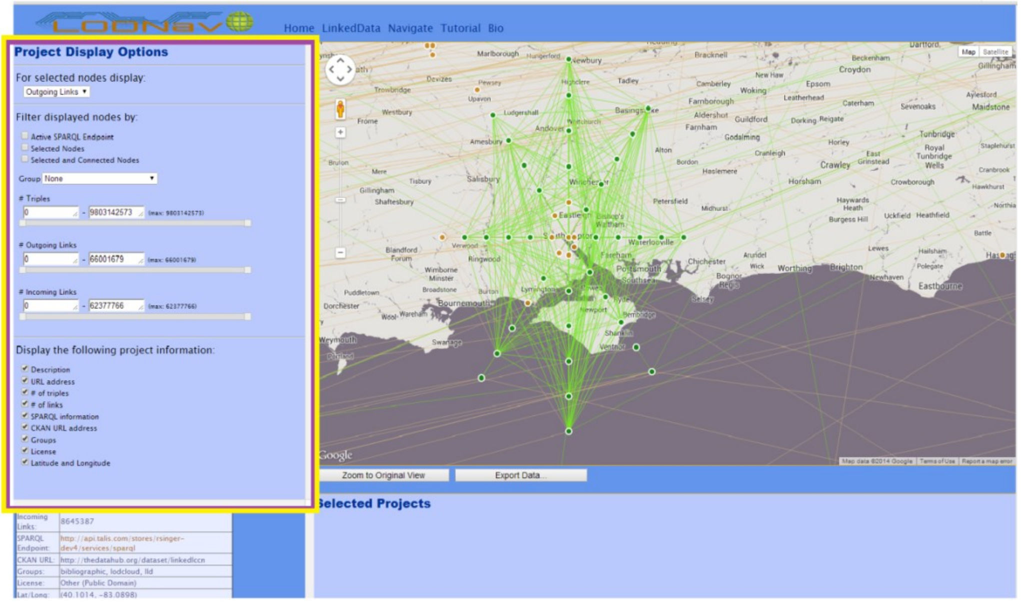
- **Green** – represents a dataset whose SPARQL endpoint is active ('on')
- **Orange** – represents a dataset whose SPARQL endpoint is inactive ('off')

A similar colour scheme is used to indicate if a link among two datasets is active or not (Figure 4-9).

- **Green line** – indicates both connected datasets have active SPARQL endpoints, implying the link can be used

- **Orange line** – indicates at least one of the connected datasets does not have an active SPARQL endpoint, implying the link cannot be used/is unavailable currently

One of the challenges with the current line visualization for representing links among connected datasets is there is no support for directional arrow indicating the direction of a connection. This is due to the fact, when LODNav was being developed, the Google Maps API did not support this feature natively. As a result, a different scheme was introduced to represent the direction of connectivity. In LODNav the datasets can be highlighted and selected to provide the user with additional information. Two different modes are supported to perform this action; *Outgoing Links* mode and *Incoming Links* mode. In both modes there is also the display of bi-directional links. Figure 4-12 provides an overview of the various visualization options provided by LODNav. It should be noted that this menu also contains other filtering options for changing what datasets will be displayed to the user.



Project Display Options

For selected nodes display:
Outgoing Links ▾

Filter displayed nodes by:
□ Active SPARQL Endpoint
□ Selected Nodes
□ Selected and Connected Nodes

Group: None ▾

Triples
0 - 9803142573 (max: 9803142573)

Outgoing Links
0 - 66001679 (max: 66001679)

Incoming Links
0 - 62377766 (max: 62377766)

Display the following project information:
☑ Description
☑ URL address
☑ # of triples
☑ # of links
☑ SPARQL information
☑ CKAN URL address
☑ Groups
☑ License
☑ Latitude and Longitude

Incoming Links:	8645387
SPARQL Endpoint:	http://api.talis.com/stores/rsinger-dev4/services/sparql
CKAN URL:	http://thedatahub.org/dataset/linkedlccn
Groups:	bibliographic, lodcloud, lld
License:	Other (Public Domain)
Lat/Long:	(40.1014, -83.0898)

Figure 4-12 : LODNav’s Display Options section

4.5.1 Outgoing Links Mode

An *Outgoing Link* from a dataset *A* pointing to another dataset *B* implies that there exists at least one actual RDF linked triple from dataset *A* to dataset *B*.

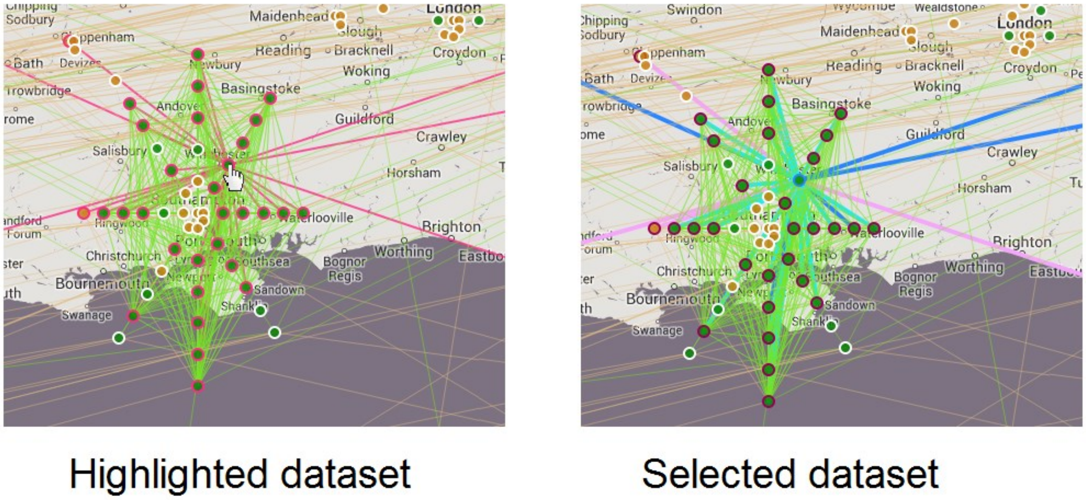
For the Outgoing Link option, LODNav will display all outgoing links of a particular dataset, when the mouse is hovering over a dataset.

- all the outgoing links are displayed in Light Red
- the connected dataset nodes are highlighted in Light Red

When a dataset is selected in the Outgoing Links mode:

- the selected dataset node is highlighted in Dark Blue (signifying outgoing links)
- the connected dataset nodes are highlighted in Dark Red (signifying an incoming link)
- a link between datasets with *Active* SPARQL Endpoints is displayed in Dark Blue
- a link between datasets with at least one *Inactive* SPARQL Endpoint is displayed in Pink
- a bi-directional link between *Active* SPARQL Endpoints is displayed in Fluorescent Blue
- a bi-directional link between datasets where at least one of the SPARQL endpoints is *Inactive* is displayed in Yellow

Figure 4-13 shows sample screen captures using the option for the dataset *DBLP Computer Science Bibliography (RKBExplorer)*.


Highlighted dataset

Selected dataset

Figure 4-13 : LODNav links in *Outgoing Links* mode

4.5.2 Incoming Links Mode

An *Incoming Link* to a dataset *A* from another dataset *B* implies that there exists at least one actual RDF linked triple from dataset *B* which points to dataset *A*.

For the Incoming Link option, LODNav will display all incoming links into a particular dataset, when the mouse is hovering over the dataset:

- all the incoming links are displayed in Purple

- the connected dataset nodes are highlighted in Light Red

When a dataset is selected in the Incoming Links mode:

- the selected dataset node is highlighted in Dark Red (signifying incoming links)

- the connected dataset nodes are highlighted in Blue (signifying an outgoing link)

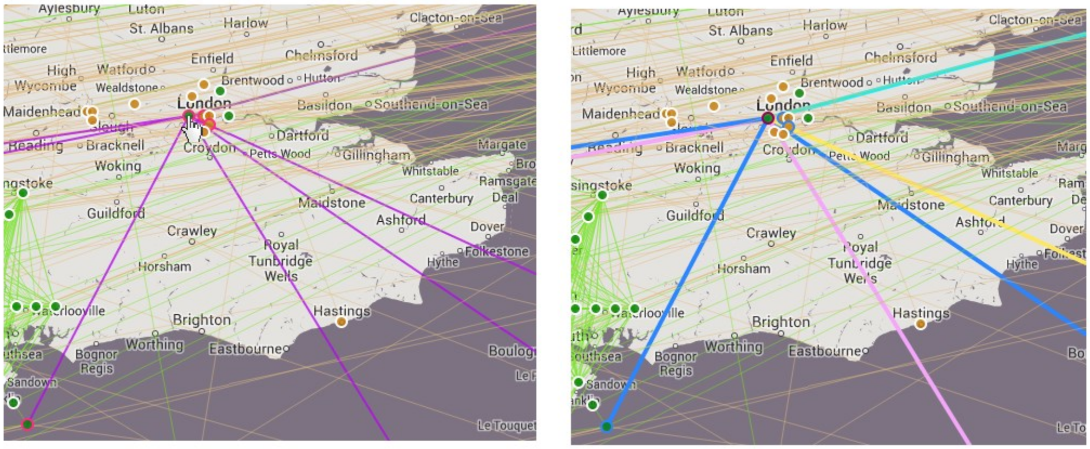
- a link between datasets with *Active* SPARQL Endpoints is displayed in Dark Blue

- a link between datasets with at least one *Inactive* SPARQL Endpoint is displayed in Pink

- a bi-directional link between *Active* SPARQL Endpoints is displayed in Fluorescent Blue

- a bi-directional link between datasets where at least one of the SPARQL endpoints is *Inactive* is displayed in Yellow

Figure 4-14 shows sample screen captures using the option for the dataset *DBTune.org MusicBrainz D2R Server*.



Highlighted dataset

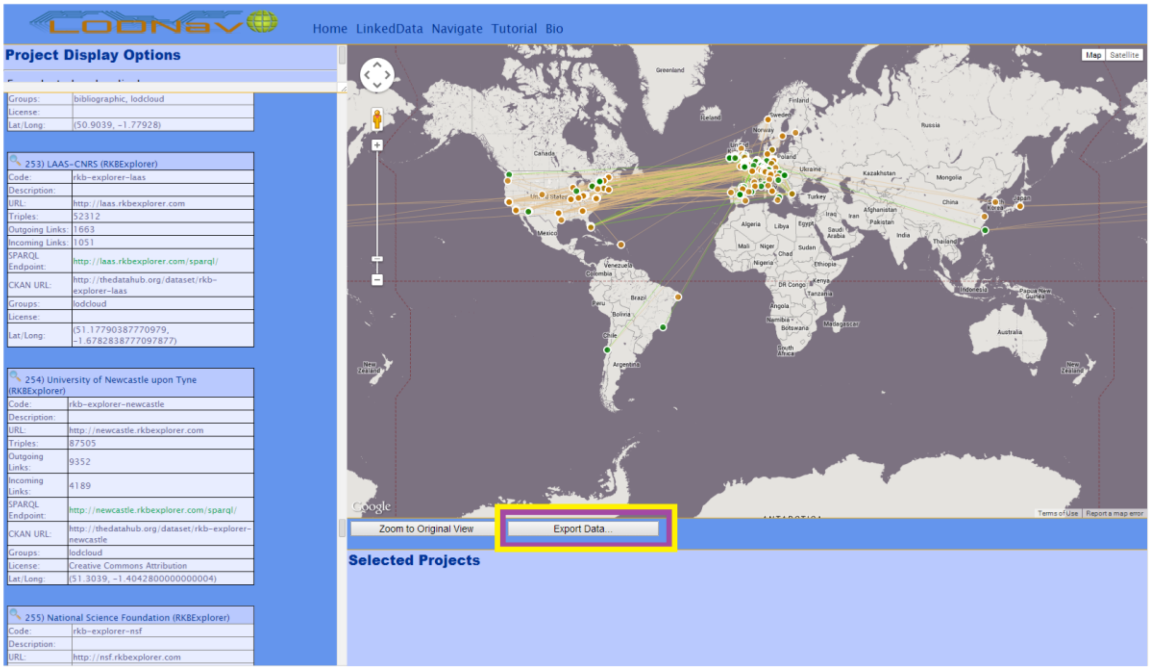
Selected dataset

Figure 4-144 : LODNav links in *Incoming Links* mode

4.6 Extracting dataset information

This section will discuss in more detail, how LODNav addresses the RQ2.3, referring to its ability to provide support for extracting the information describing the datasets and their links for further reuse. This is in particular interest, since no complete dump is available for the LODC, containing all the meta-data.

This section discusses how LODNav supports the export (Figure 4-15 shows the export button highlighted) of the LODC meta-information and structural information as RDF data.



Groups:	bibliographic, lodcloud
License:	
Lat/Long:	(50.9039, -1.77928)

253) LAAS-CNRS (RKBExplorer)	
Code:	rkb-explorer-laas
Description:	
URL:	http://laas.rkbexplorer.com
Triples:	52312
Outgoing Links:	1663
Incoming Links:	1051
SPARQL Endpoint:	http://laas.rkbexplorer.com/sparql/
CKAN URL:	http://thedatahub.org/dataset/rkb-explorer-laas
Groups:	lodcloud
License:	
Lat/Long:	(51.17790387770979, -1.6782838777097877)

254) University of Newcastle upon Tyne (RKBExplorer)	
Code:	rkb-explorer-newcastle
Description:	
URL:	http://newcastle.rkbexplorer.com
Triples:	87505
Outgoing Links:	9352
Incoming Links:	4189
SPARQL Endpoint:	http://newcastle.rkbexplorer.com/sparql/
CKAN URL:	http://thedatahub.org/dataset/rkb-explorer-newcastle
Groups:	lodcloud
License:	Creative Commons Attribution
Lat/Long:	(51.3039, -1.4042800000000004)

255) National Science Foundation (RKBExplorer)	
Code:	rkb-explorer-nsf
Description:	
URL:	http://nsf.rkbexplorer.com

Figure 4-15 : LODNav's export button

The LODNav export function offers users two options for exporting some RDF data. The user first specifies the scope of the export either; (1) all datasets which are currently shown on the map, or (2) all the datasets which have been selected by a user.

Secondly, the user selects the export format used for the RDF, which can either be RDF/XML or N-Triple format.

The output of the LODNav export function will create a small ontology file in the specified RDF format and can be reused for further processing or analysis.

4.6.1 Selecting datasets for export

As mentioned earlier different export options are available when specifying the scope of the export; *all visible projects on the map* OR *all selected projects*.

For the ***all visible projects on the map*** option, every dataset being displayed (regardless of the zoom level) will have its information exported to RDF. When LODNav is initially loaded every dataset of the Linking Open Data Cloud is displayed. However, some of these datasets can be filtered out by using the *LODNav Display Options* in Figure 4-12. Different combinations of filter criteria can be applied:

- **Active SPARQL Endpoints** – displays only the datasets which have an Active SPARQL endpoint
- **Selected nodes** – displays only the datasets which have been selected by the user
- **Selected and Connected Nodes** – displays only the datasets which are selected by the user as well as the datasets connected to the selected datasets
- **Group** – displays only the datasets which belong to a user selected group
- **A range of RDF triples within the datasets** – displays only the datasets which contain a number of RDF triples which falls within a user specified range
- **A range of Outgoing Links from datasets** – displays only the datasets which contain the total number of outgoing RDF links within a user specified range

- **A range of Incoming Links from datasets** – displays only the datasets which contain the total number of incoming RDF links within a user specified range

When a user modifies these criteria, only datasets containing the selected properties will be displayed on the map. For example in Figure 4-16 all datasets with an *Active* SPARQL Endpoint are shown.

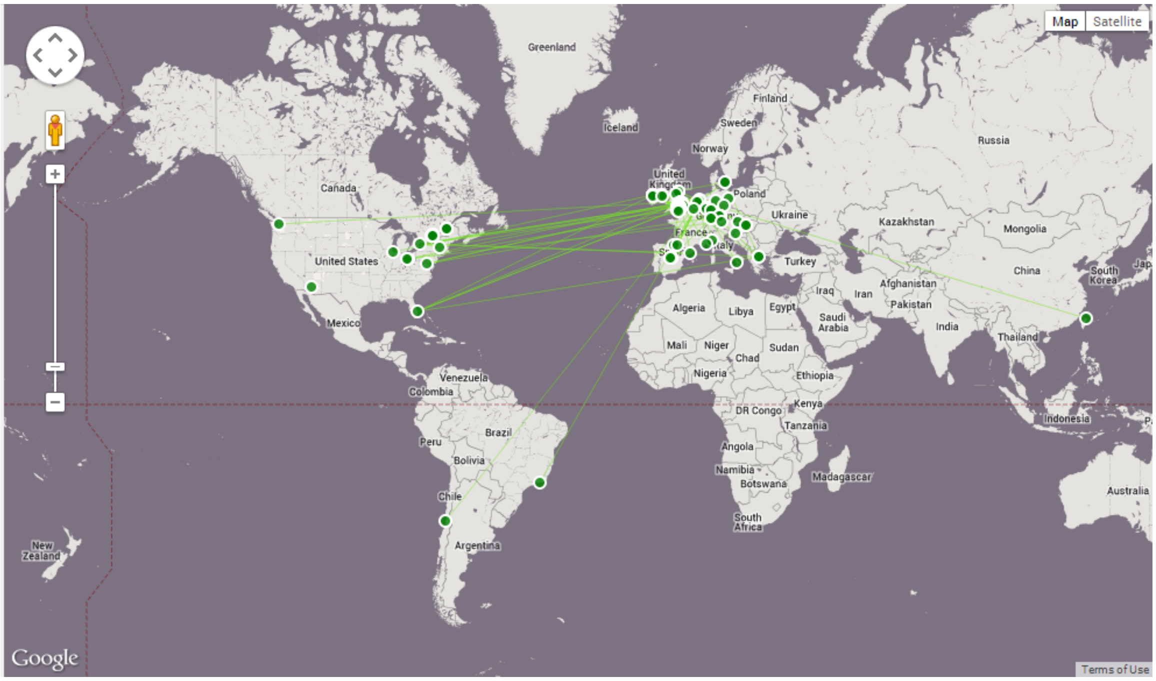


Figure 4-16 : Datasets with *Active* SPARQL Endpoints

If a user were to export information for the visible datasets then only these datasets shown on the map would be exported.

For the option, ***all selected projects*** only the datasets, which are selected by the user will have its information exported in RDF. An example for such a selection is shown in Figure 4-17, which includes the selection of four different datasets.

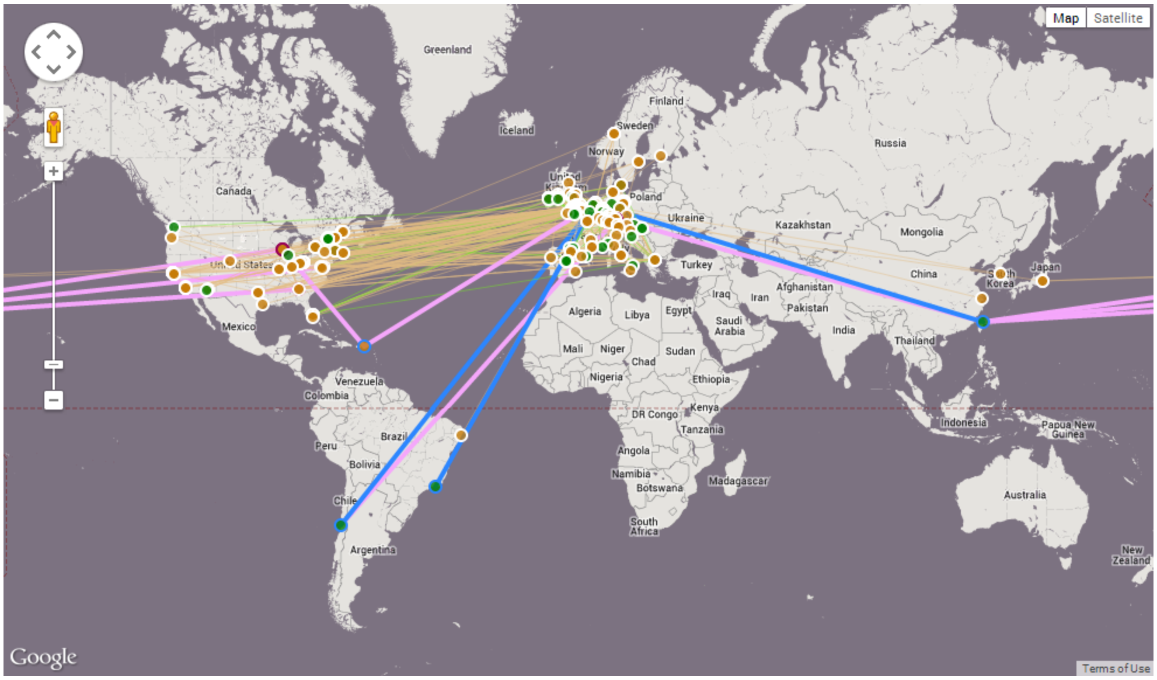


Figure 4-17 : Four selected datasets

4.6.2 The exported RDF data

In what follows a more detailed discussion on the information, being exported to an ontology is provided.

The ontology file is called:

```
http://www.lodnav.com/lodnav
```

The triple representation is:

```
<http://www.lodnav.com/lodnav>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://www.w3.org/2002/07/owl#Ontology> .
```

The class `lodnav:LodNode` is used to assign the datasets as entity resources in the ontology.

The entities in the exported ontology have the following Datatype Properties:

- lodnav:title
- lodnav:codename
- lodnav:url
- lodnav:sparqlEndpoint
- lodnav:numTriples
- lodnav:latitude
- lodnav:longitude

For capturing the outgoing links to the other datasets, the following Object Property is used:

- lodnav:linksTo

Resources get URIs based on the following format:

- lodnav:***datasetCodename***

The following is an example of a LODNav ontology created in (N-Triple format), for the Linked Move DataBase dataset:

```
<http://www.lodnav.com/lodnav#linkedmdb>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://www.w3.org/2002/07/owl#NamedIndividual> .

<http://www.lodnav.com/lodnav#linkedmdb>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://www.lodnav.com/lodnav#LodNode> .

<http://www.lodnav.com/lodnav#linkedmdb>
<http://www.lodnav.com/lodnav#latitude>
"43.627683495705504"^^<http://www.w3.org/2001/XMLSchema#str
ing> .
```

```
<http://www.lodnav.com/lodnav#linkedmdb>
<http://www.lodnav.com/lodnav#longitude> "-
79.5405834957055"^^<http://www.w3.org/2001/XMLSchema#string
> .

<http://www.lodnav.com/lodnav#linkedmdb>
<http://www.lodnav.com/lodnav#numTriples>
"6148121"^^<http://www.w3.org/2001/XMLSchema#int> .

<http://www.lodnav.com/lodnav#linkedmdb>
<http://www.lodnav.com/lodnav#title> "Linked Movie
DataBase"^^<http://www.w3.org/2001/XMLSchema#string> .

<http://www.lodnav.com/lodnav#linkedmdb>
<http://www.lodnav.com/lodnav#codename>
"linkedmdb"^^<http://www.w3.org/2001/XMLSchema#string> .

<http://www.lodnav.com/lodnav#linkedmdb>
<http://www.lodnav.com/lodnav#url>
"http://linkedmdb.org/"^^<http://www.w3.org/2001/XMLSchema#
string> .

<http://www.lodnav.com/lodnav#linkedmdb>
<http://www.lodnav.com/lodnav#sparqlEndpoint>
"http://data.linkedmdb.org/sparql"^^<http://www.w3.org/2001
/XMLSchema#string> .

<http://www.lodnav.com/lodnav#linkedmdb>
<http://www.lodnav.com/lodnav#linksTo>
<http://www.lodnav.com/lodnav#yago> .

<http://www.lodnav.com/lodnav#linkedmdb>
<http://www.lodnav.com/lodnav#linksTo>
<http://www.lodnav.com/lodnav#dbtune-musicbrainz> .

<http://www.lodnav.com/lodnav#linkedmdb>
<http://www.lodnav.com/lodnav#linksTo>
<http://www.lodnav.com/lodnav#lingvoj> .

<http://www.lodnav.com/lodnav#linkedmdb>
<http://www.lodnav.com/lodnav#linksTo>
<http://www.lodnav.com/lodnav#geonames-semantic-web> .

<http://www.lodnav.com/lodnav#linkedmdb>
<http://www.lodnav.com/lodnav#linksTo>
<http://www.lodnav.com/lodnav#flickr-wrappr> .
```

```
<http://www.lodnav.com/lodnav#linkedmdb>
<http://www.lodnav.com/lodnav#linksTo>
<http://www.lodnav.com/lodnav#dbpedia> .

<http://www.lodnav.com/lodnav#linkedmdb>
<http://www.lodnav.com/lodnav#linksTo>
<http://www.lodnav.com/lodnav#rdf-book-mashup> .
```

4.7 Summary

In this chapter, LODNav the tool implementation for visualizing the LODC data was presented. A brief overview of LODNav was given explaining its main visualization objective and metaphor, the ability to provide a more human user centric approach to visualizing the Liking Open Data Cloud. Next, a brief overview of the actual architecture of LODNav and further details about the Google Maps API, the main visualization metaphor used by LODNav was presented. The section also discusses how LODNav's geographic map based visualization and representation of the meta-data extracted from the LODC address our research objectives introduced in Section 2.2 of the thesis. Both the ability to provide an interactive visualization for the LODC as well as the ability to filter and export selected data, which are some of the key features of LODNav, were discussed in detail.

Given the interactive visualization of LODC data, not only the comprehensibility of the data can be improved but also the participation in publishing and consuming of Linked Open Data will be further facilitated.

5. Use Cases

As part of this research, we performed several use cases to illustrate the applicability of LODNav in visualizing information from the LODC. The data being visualized was gathered from *The Linking Open Data Cloud* group located on *Datahub* (as explained in Section 4.2.1) in October 2013 [16]. For the use cases data from a total of 337 datasets were crawled and extracted.

5.1 Use Case #1: Viewing centres of expertise/adoption of Linking Open Data

In this use case, it is shown how LODNav can be used to locate regions which have clusters of datasets located at particular locations. When clusters of datasets are present at a region we see that some expertise in Linked Data is emerging at that location and groups/individuals have begun to adopt the concepts of Linking Open Data.

To begin this use case, LODNav is used to zoom into Europe prior to using any filtering mechanisms (Figure 5-1).

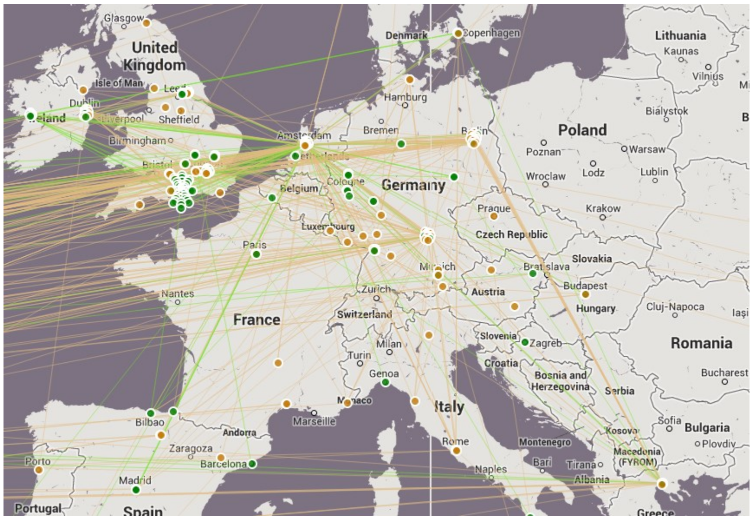
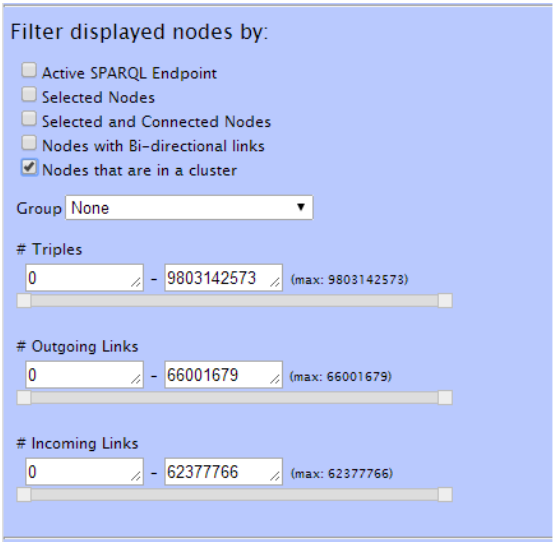


Figure 5-1 : View of Europe

The next step in this use case is to apply the filtering option to only display *Nodes that are in a cluster* (Figure 5-2), implying only datasets which have the same location as at least one other dataset are displayed.



Filter displayed nodes by:

☐ Active SPARQL Endpoint
☐ Selected Nodes
☐ Selected and Connected Nodes
☐ Nodes with Bi-directional links
☑ Nodes that are in a cluster

Group None ▾

Triples
0 - 9803142573 (max: 9803142573)

Outgoing Links
0 - 66001679 (max: 66001679)

Incoming Links
0 - 62377766 (max: 62377766)

Figure 5-2 : Filtering to see Nodes that are clustered

In Figure 5-3 we see the datasets which remain after the filtering operation in Europe. The datasets which do not share a location with any other dataset are filtered out, leaving only datasets which cluster around a particular location. Some of the more pronounced locations of clusters of datasets (as previously shown in Figure 4-10) are; Amsterdam, Netherlands, Berlin, Germany, Dublin, Ireland, and Southampton, England.

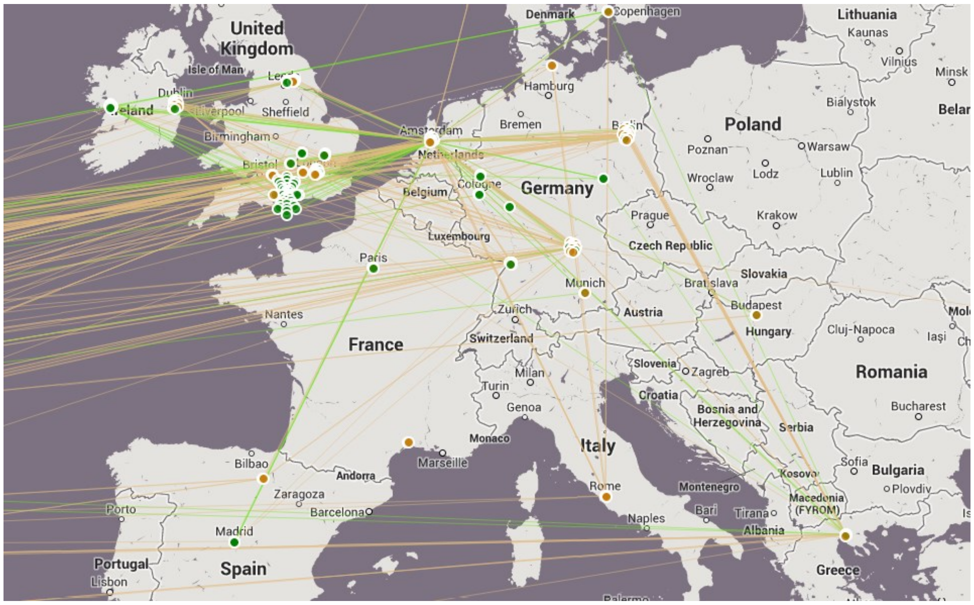


Figure 5-3 : Datasets that are part of clusters in Europe

By navigating across the globe using LODNav other clusters can be identified. For instance, in Eastern Canada some notable clusters are located in Ottawa and Quebec City.

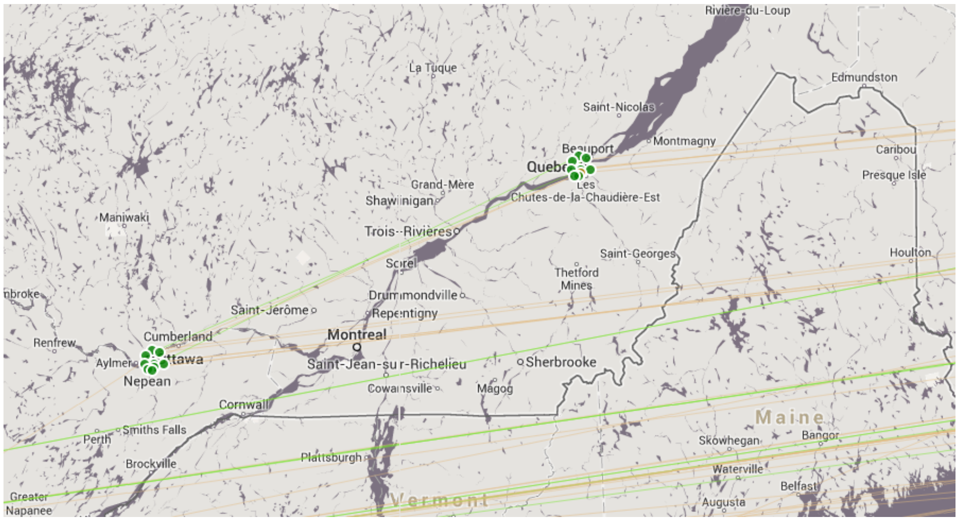


Figure 5-4 : Clusters of datasets in Eastern Canada

5.2 Use Case #2: All datasets with an *Active* SPARQL Endpoint

For this use case, all datasets with an *Active* SPARQL endpoint are extracted and visualized. Figure 5-5 shows the datasets with an active SPARQL endpoint.

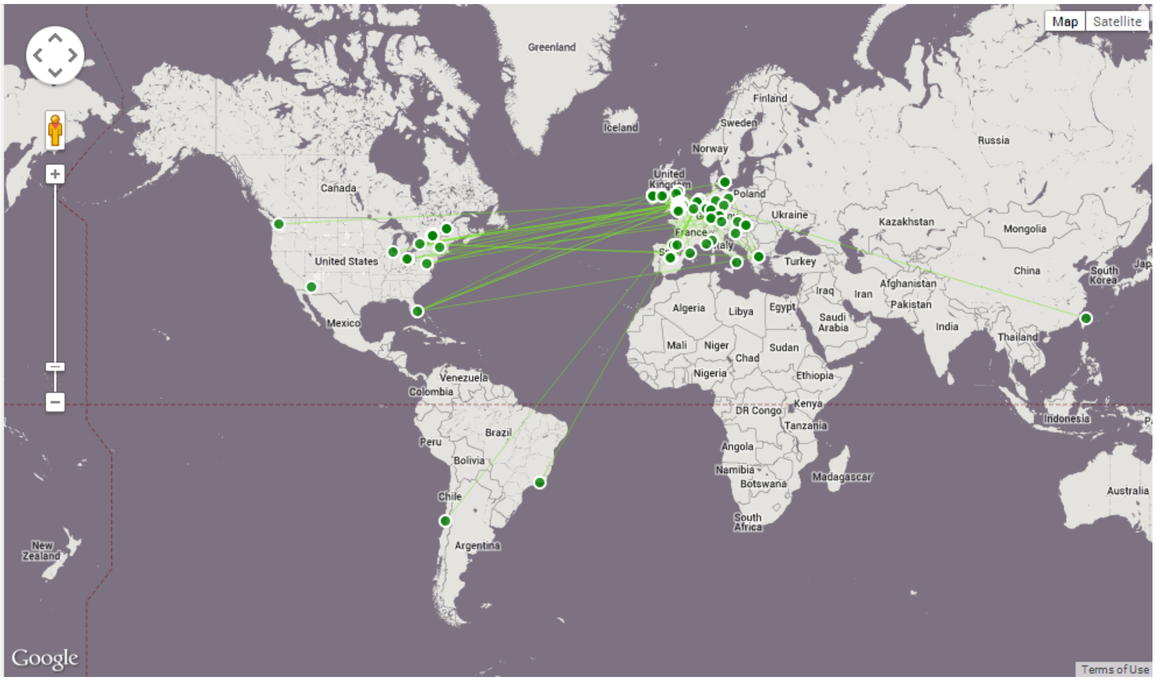


Figure 5-5 : All datasets with an *Active* SPARQL Endpoint

Nodes displayed by filter option:
Active SPARQL Endpoint

Table 5-1 : Case Study #2 filter options

Table 5-1 shows the filter option, which was used within LODNav to visualize all the Active SPARQL endpoint datasets. A complete listing (*codename* and *title*) of these datasets with an active SPARQL endpoint is included in Appendix B.

5.3 Use Case #3 All *Active* datasets which link to DBpedia

The objective of this use case is to visualize all datasets with an Active endpoint and link to the DBpedia dataset. DBpedia dataset was selected, since it was at the time of writing, a key dataset within the LODC. DBpedia is one of the largest datasets, with respect to the number of links to other datasets and its size (in number of triples). Figure 5-6 shows the visualization produced by LODNav.

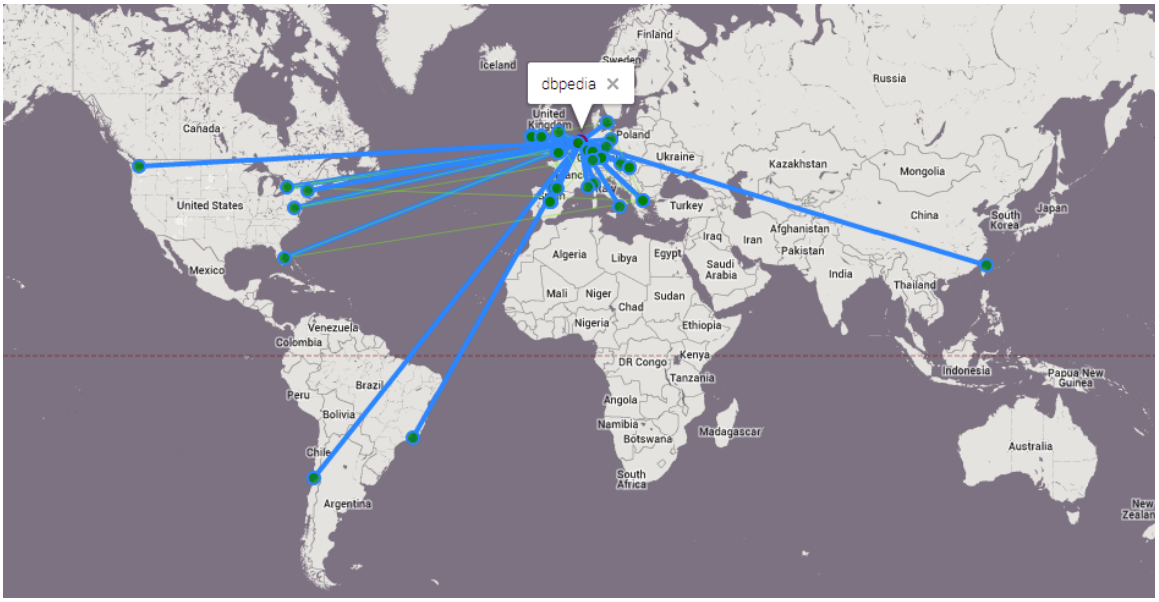


Figure 5-6 : All datasets which link to DBPedia

For the visualization, we configured LODNav to display *Incoming Links* for selected nodes. For the use case DBpedia was selected as the dataset for which all incoming nodes are displayed. The filter option was set to view the *Selected and Connected Nodes*, which causes the nodes that are not connected to DBPedia to be filtered out from the visualization. An additional filter option was used to view only *Active SPARQL Endpoints*. The complete filter options are shown in Table 5-2, with Appendix C showing the complete list of the datasets with their respective *codename* and *title*.

Nodes displayed by filter options:
Selected and Connected Nodes
Active SPARQL Endpoint

Table 5-2 : Use Case #3 filter options

5.4 Use Case #4: All datasets which DBPedia links to

In this use case, we were interested to show all datasets which DBpedia links to (Figure 5-7) This is in contrast to the previous use case (use case #2), which shows all the nodes that link to DBpedia.

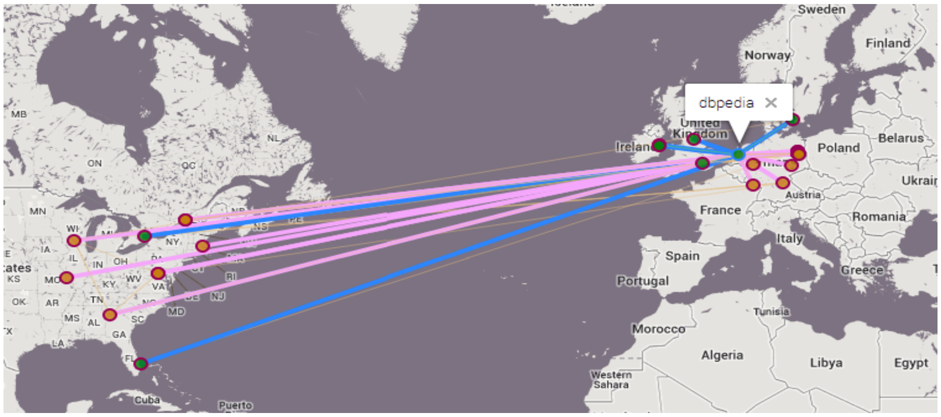


Figure 5-7 : All datasets which DBPedia links to

LODNav was configured to display *Outgoing Links* for selected nodes as the filtering criteria for the visualization. For this case study, DBpedia was selected as the focus dataset. The resulting visualization (Figure 5-3) shows all links, which initiate from DBpedia to other datasets within the LODC. Using the filter option to view only *Selected and Connected Nodes*, omits all links and nodes not connected to DBpedia from the visualization. It should be noted that for this use case, no restriction was made regarding if a dataset is active or not and therefore links from DBpedia to both active and inactive SPARQL endpoints are included

Nodes displayed by filter option:
Selected and Connected Nodes

Table 5-3 : Use Case #4 filter option

Table 5-4 shows the list of datasets (*codename* and *title*) which DBPedia connects to displaying their respective *codename* and *title*.

#	Dataset Codename	Dataset Title
1	2000-us-census-rdf	2000 U.S. Census in RDF (rdfabout.com)
2	dbpedia	DBpedia
3	dbtune-musicbrainz	DBTune.org Musicbrainz D2R Server
4	education-data-gov-uk	education.data.gov.uk
5	eunis	European Nature Information System
6	flickr-wrappr	flickr™ wrappr
7	freebase	Freebase
8	fu-berlin-dailymed	DailyMed
9	fu-berlin-dblp	DBLP Bibliography Database in RDF (FU Berlin)
10	fu-berlin-diseasome	Diseasome
11	fu-berlin-drugbank	DrugBank
12	fu-berlin-eurostat	Eurostat in RDF (FU Berlin)
13	fu-berlin-project-gutenberg	Project Gutenberg in RDF (FU Berlin)
14	fu-berlin-sider	SIDER: Side Effect Resource

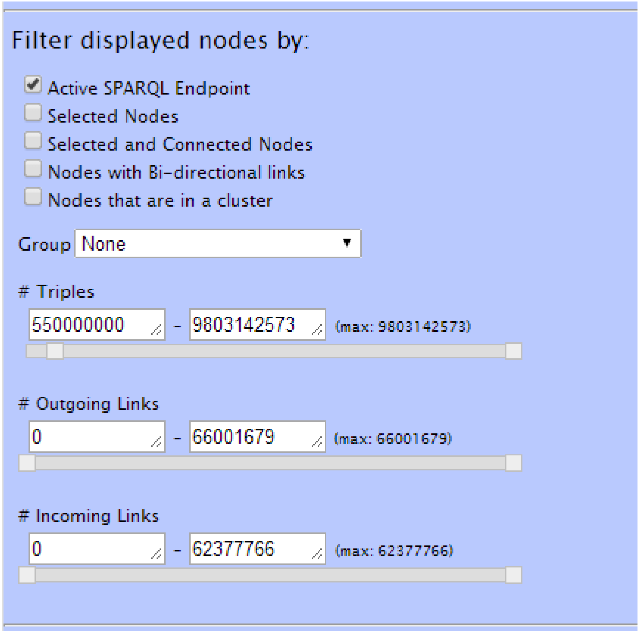
15	geonames-semantic-web	GeoNames Semantic Web
16	geospecies	GeoSpecies Knowledge Base
17	italian-public-schools-linkedopendata-it	Italian public schools (LinkedOpenData.it)
18	linkedgeodata	LinkedGeoData
19	linkedmdb	Linked Movie DataBase
20	nytimes-linked-open-data	New York Times - Linked Open Data
21	opencyc	OpenCyc
22	rdf-book-mashup	RDF Book Mashup
23	reference-data-gov-uk	reference.data.gov.uk
24	revyu	Revyu.com - Review Anything
25	tcmgenedit_dataset	TCMGeneDIT Dataset
26	transport-data-gov-uk	transport.data.gov.uk
27	uk-legislation-api	UK Legislation
28	w3c-wordnet	WordNet 2.0 (W3C)
29	world-factbook-fu-berlin	World Factbook (FU Berlin)
30	yago	YAGO

Table 5-4 : Use Case #4 – All datasets which DBpedia links to

5.5 Use Case #5: Filtering datasets based on number of RDF triples

The objective of this use case was to show LODNav's ability to filter the nodes displayed in the visualization with the number of triples inside the member datasets of the LODC. This was done by identifying the five largest datasets with an *Active* SPARQL endpoint within the LODC. It should be noted that we consider the size of a dataset as the number of RDF triples that are contained within it.

For the use case, LODNav was configured to show only datasets with *Active* datasets and within a certain range of triples. The largest five datasets are within the range of 550,000,000 and 9,803,142,573 RDF triples, with the range manually determined in LODNav by adjusting the filtering option to only display datasets until only 5 datasets remain in the visualization (Figure 5-4).



Filter displayed nodes by:

☑ Active SPARQL Endpoint
☐ Selected Nodes
☐ Selected and Connected Nodes
☐ Nodes with Bi-directional links
☐ Nodes that are in a cluster

Group: None ▼

Triples
550000000 - 9803142573 (max: 9803142573)

Outgoing Links
0 - 66001679 (max: 66001679)

Incoming Links
0 - 62377766 (max: 62377766)

Figure 5-7 : Filtering by # *of Triples* and *Active SPARQL Endpoint*

The value of 9,803,142,573 is used as a maximum value, this corresponding to the number of triples for the largest dataset. It should be noted that at point of writing the largest dataset happens to have an *Inactive* SPARQL endpoint. This dataset is *TWC: Linking Open Government Data* and is not present in this use case. The five biggest datasets (based on number of triples) with an active SPARQL endpoint are shown below in Figure 5-8.

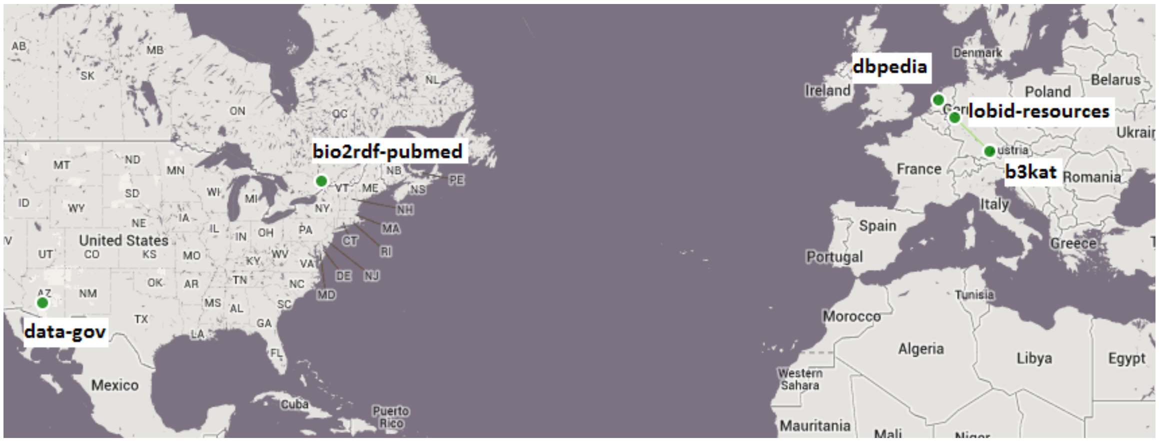


Figure 5-8 : Five largest Active datasets in LODNav

Table 5-5 shows the results of this case study by displaying the list of the five largest *Active* datasets with their respecting *codename*, *title*, and *number of RDF triples*.

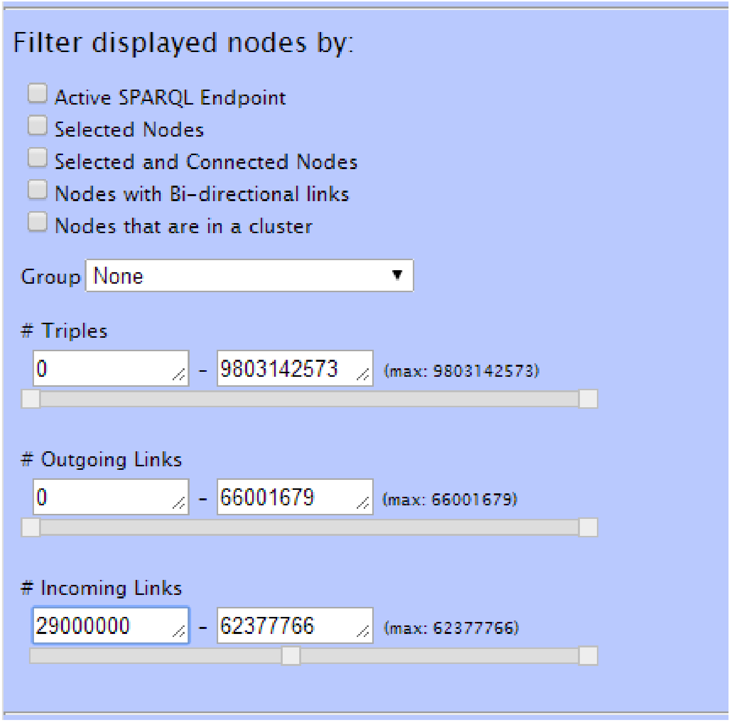
#	Dataset Codename	Dataset Title	# RDF triples
1	b3kat	B3Kat – Library Union Catalogues of Bavari	570,000,000
2	bio2rdf-pubmed	PubMed	797,000,000
3	data-gov	Data.gov	6,400,000,000
4	dbpedia	DBPedia	1,200,000,000
5	lobid-resources	Lobid, Bibliographic Resources	667,675,574

Table 5-5 : Case Study #5 – Five largest *Active* datasets

5.6 Use Case #6: Filtering datasets by the number of incoming RDF links

For this use case, the objective was to illustrate how LODNav can filter the displayed nodes in the visualization based on the number of incoming links into datasets. This is exemplified by identifying and visualizing the five datasets in the LODC with the largest number of incoming RDF links. This is of particular interest, since incoming links highlight how the LODC can facilitate the sharing of data.

In this visualization the datasets that have a range of incoming RDF links between 29,000,000 and 62,377,766 are displayed. The range was again determined within LODNav, by manually adjusting the range of incoming links until only 5 datasets remain in the visualization (Figure 5-9). The value of 62,377,766 is used as a maximum value because this the largest number of incoming RDF links for all the datasets.



Filter displayed nodes by:

□ Active SPARQL Endpoint
□ Selected Nodes
□ Selected and Connected Nodes
□ Nodes with Bi-directional links
□ Nodes that are in a cluster

Group None ▼

Triples

0 - 9803142573 (max: 9803142573)

Outgoing Links

0 - 66001679 (max: 66001679)

Incoming Links

29000000 - 62377766 (max: 62377766)

Figure 5-9 : Filtering by # *of Incoming Links*

The five datasets with the most incoming links are shown in Figure 5-10.

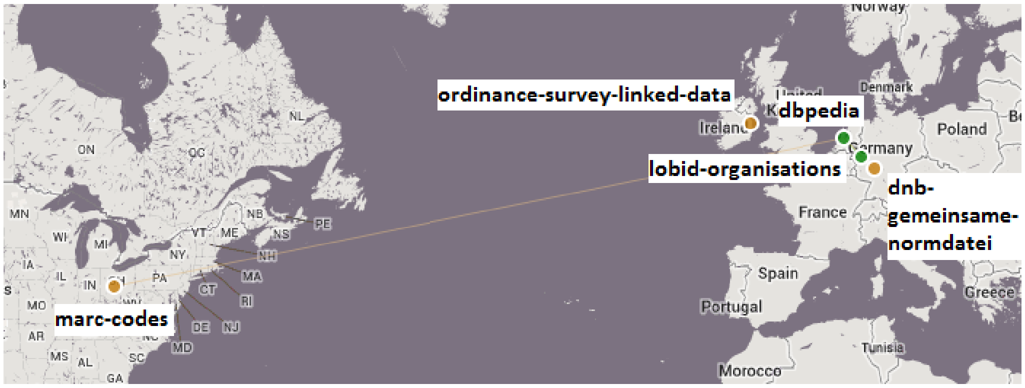


Figure 5-10 : Five datasets with the most incoming links

Table 5-6 shows the list with the five most linked to datasets. The table includes the *codename*, *title*, *number of incoming RDF links*, and the *number of datasets which link to the dataset*.

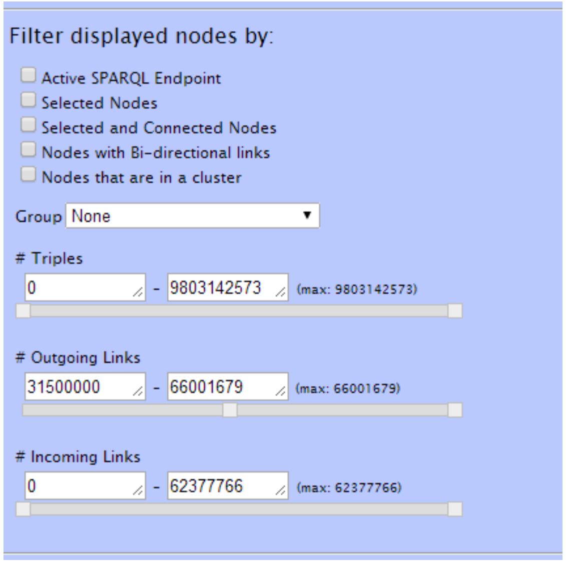
#	Dataset Codename	Dataset Title	# of incoming RDF links	# of datasets which link to this dataset
1	dbpedia	DBPedia	39,008,277	185
2	dnb-gemeinsame-normdatei	Gemeinsame Normdatei (GND)	62,377,766	11
3	lobid-organisations	Lobid. Index of libraries and related organisations	32,286,583	1
4	marc-codes	MARC Codes List	30,370,350	2
5	ordnance-survey-linked-data	Ordnance Survery Linked Data	29,717,902	16

Table 5-6 : Use Case #6 – Datasets with the five most incoming number of RDF links

5.7 Use Case #7: Filtering datasets by number of outgoing links

The objective of this use case is to highlight how LODNav can display nodes in the visualization by adjusting the filter option to display only the datasets within a certain range of outgoing RDF links. In this use case we show the datasets which take the most advantage of the LODC, by displaying the five datasets with the most outgoing RDF links.

In this use case the five datasets with the largest number of outgoing links are within the range of 31,500,000 and 66,001,679 RDF links. By manually adjusting the filtering option to display datasets within a range of outgoing RDF links, it was found that the five largest lie within the specified range. The value of 66,001,679 is used as a maximum value by LODNav because this is the highest value of outgoing RDF links of all the LODC datasets. The options used within LODNav for this use case are shown in Figure 5-11.



Filter displayed nodes by:

- □ Active SPARQL Endpoint
- □ Selected Nodes
- □ Selected and Connected Nodes
- □ Nodes with Bi-directional links
- □ Nodes that are in a cluster

Group None ▼

# Triples

0 - 9803142573 (max: 9803142573)

# Outgoing Links

31500000 - 66001679 (max: 66001679)

# Incoming Links

0 - 62377766 (max: 62377766)

Figure 5-11 : Filtering by # *of Outgoing Link*

The datasets that remain in the visualization after applying the filtering options can are shown in Figure 5-12.

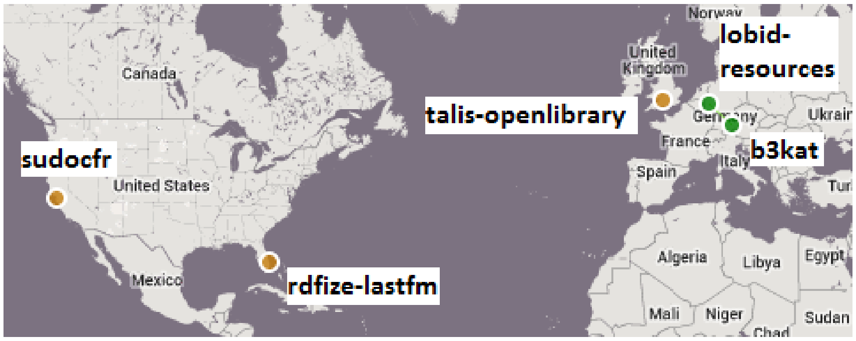


Figure 5-12 : Five datasets with largest number of outgoing links

Table 5-10 lists the five datasets with the most number of outgoing RDF links.

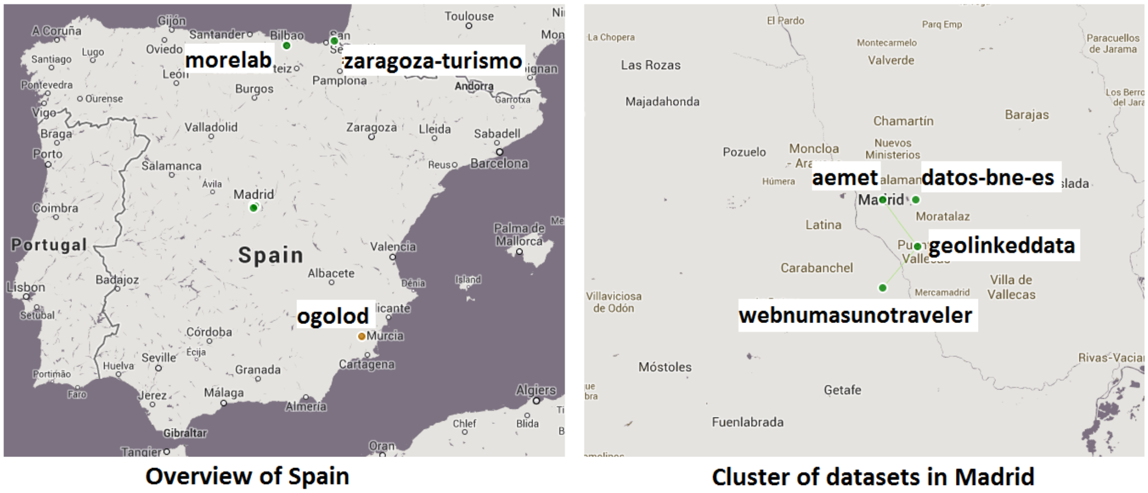
#	Dataset Codename	Dataset Title	# of outgoing RDF links	# of datasets which the dataset links to
1	b3kat	B3Kat – Library Union Catalogues of Bavari	33,924,158	3
2	lobid-resources	lobid. Bibliographic Resources	66,001,679	12
3	rdfize-lastfm	Last.FM RDFization of Event	50,023,000	4
4	sudocfr	Sudoc bibliographic data	31,500,000	3
5	talis-openlibrary	Open Library data mirror in the Talis Platform	51,937,095	4

Table 5-7 : Use case #7 – Datasets with the five most outgoing number of RDF links

5.8 Use Case #8: Datasets which are members of the *Spain* group

Some of the LODC datasets have included in their meta-data that they belong to a particular *Group*, allowing datasets with similar properties to be grouped together. A total of 44 different groups are registered with the LODC (Appendix D shows the complete list of all LODC dataset groups).

The objective of this use case is, to show how the *Group* filter option can be used to create logical groupings in the visual representation. For this example, only nodes in the LODC which have been registered as being part of the *Spain* group are shown (Figure 5-7) with the filter option being shown in table 5-8.



Overview of Spain

Cluster of datasets in Madrid

Figure 5-13 : Datasets which are members of the *Spain* group

Nodes displayed by filter option:
Group: spain

Table 5-8 : Use Case #8 filter option

Table 5-9 presents the list of the datasets which are members of the *Spain* group of the Linking Open Data Cloud.

#	**Dataset Codename**	**Dataset Title**
1	aemet	AEMET metereological dataset
2	datos-bne-es	datos.bne.es
3	geolinkeddata	GeoLinkedData
4	morelab	morelab
5	ogolod	Orthology and Diseases Information – OGO
6	webnmasunotraveler	El Viajero’s tourism dataset
7	zaragoza-turismo	Turismo de Zaragoza

Table 5-9 : Use Case #8 – Datasets which are members of the *spain* group

5.9 Use Case #9: Transitivity in LODC

In this use case, LODNav is used to discover transitivity connections between datasets. For the use case LODNav was used to iteratively identify all datasets which the SECOLD dataset can potentially connect to. This includes not only direct but also potential transitive links. Identifying these transitive links, can guide users during a visual exploration of potential traceability links within the LODC and therefore help identify sources (nodes) that can be used to obtain an even richer knowledgebase for a given dataset.

The default visualization option in LODNav is that a user can select a dataset with having all direct RDF links to the dataset being displayed. Another filtering option within

LODNav is to display only *Selected and Connected Nodes*. Selecting *secold* and applying the option to present only the *Selected and Connected Nodes* displays only *secold* as well as all nodes that are directly connected to *secold*. These datasets being *freebase*, *opencyc*, and *dbpedia* as shown in Figure 5-10.

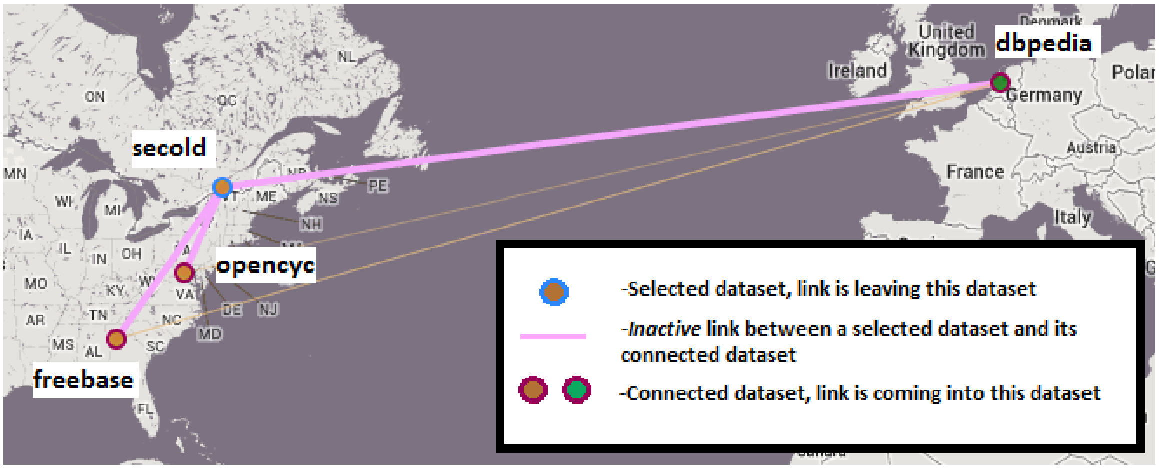

secold
opencyc
freebase
dbpedia
United Kingdom
Ireland
Denmark
Germany
Polar
France
Austria
Italy
ON
QC
NL
MN
WI
MI
IA
IL
IN
OH
MO
KY
AR
TN
MS
AL
NC
SC
VA
DE
NJ
MD
ME
NS
NH
PE
FL
-Selected dataset, link is leaving this dataset
-*Inactive* link between a selected dataset and its connected dataset
-Connected dataset, link is coming into this dataset

Figure 5-14 : Datasets which SECOLD is directly connected with

In addition, LODNav was used for this use case to also include the outgoing links from *freebase*, *opencyc*, and *dbpedia,* providing a first level transitive view. The resulting view provides all the potential datasets which SECOLD can connect to through following first level transitive links.

This resulted in the datasets whom *freebase*, *opencyc*, and *dbpedia* have connections with to appear in the visualization. This is displayed in Figure 5-11 (From this figure it is difficult to see everything that is going on in the visualization and will be explained further. From the dataset *freebase*, 5 new outgoing links appear. Out of *dbpedia* there are 30 connected datasets. A call-out box is provided to get a closer view of the *dbpedia* node; however it is still difficult to view all the connections in the figure since many of the connected datasets are located close together. The dataset *opencyc* does not have any new outgoing links.)

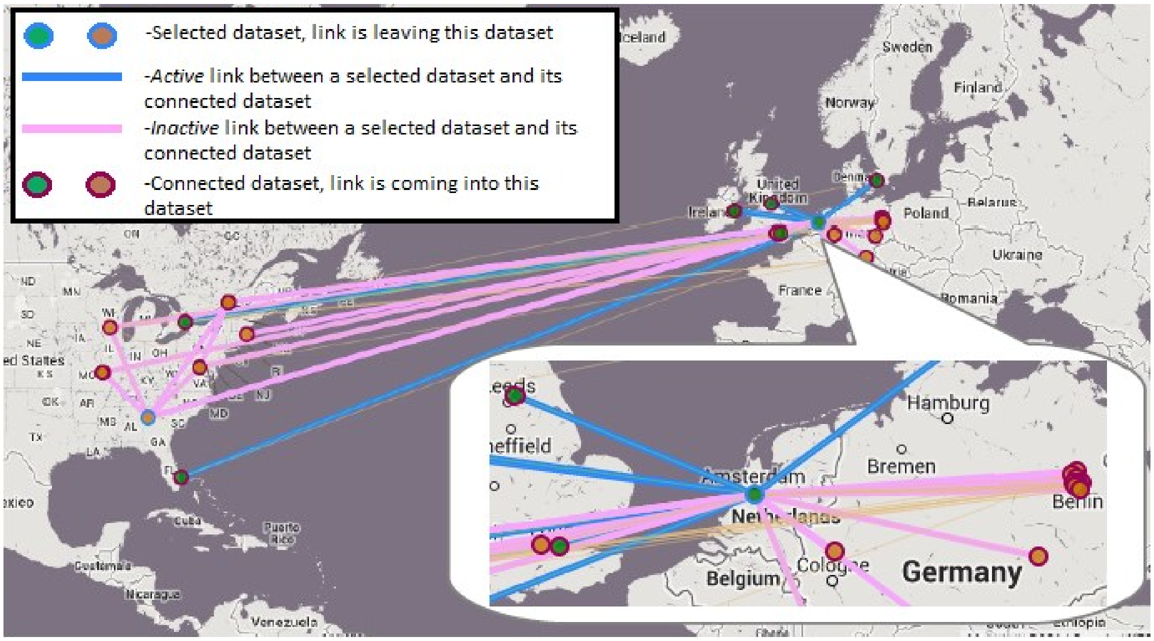

-Selected dataset, link is leaving this dataset
-*Active* link between a selected dataset and its connected dataset
-*Inactive* link between a selected dataset and its connected dataset
-Connected dataset, link is coming into this dataset
Iceland
Sweden
Finland
Norway
United Kingdom
Ireland
Poland
Belarus
Ukraine
Romania
France
Hamburg
Bremen
Amsterdam
Netherlands
Berlin
Belgium
Cologne
Germany
Leeds
Sheffield
Cuba
Puerto Rico
Guatemala
Nicaragua
Venezuela

Figure 5-15 : Transitive datasets connected to SECOLD

The datasets shown in Figure 5-9 are therefore all datasets SECOLD can potentially connect to. By exporting the data for all visible datasets in the visualization, a list of all datasets transitively (first-level) linked to SECOLD can be extracted and are listed in Table 5-10.

#	**Dataset Codename**	**Dataset Title**
1	2000-us-census-rdf	2000 U.S. Census in RDF (rdfabout.com)
2	bbc-music	BBC Music
3	dbpedia	DBpedia
4	dbtune-musicbrainz	DBTune.org Musicbrainz D2R Server

5	education-data-gov-uk	education.data.gov.uk
6	eunis	European Nature Information System
7	flickr-wrappr	flickr™ wrappr
8	freebase	Freebase
9	fu-berlin-dailymed	DailyMed
10	fu-berlin-dblp	DBLP Bibliography Database in RDF (FU Berlin)
11	fu-berlin-diseasome	Diseasome
12	fu-berlin-drugbank	DrugBank
13	fu-berlin-eurostat	Eurostat in RDF (FU Berlin)
14	fu-berlin-project-gutenberg	Project Gutenberg in RDF (FU Berlin)
15	fu-berlin-sider	SIDER: Side Effect Resource
16	geonames-semantic-web	GeoNames Semantic Web
17	geospecies	GeoSpecies Knowledge Base
18	italian-public-schools-linkedopendata-it	Italian public schools (LinkedOpenData.it)
19	linkedgeodata	LinkedGeoData
20	linkedmdb	Linked Movie DataBase
21	nytimes-linked-open-data	New York Times - Linked Open Data
22	opencyc	OpenCyc
23	rdf-book-mashup	RDF Book Mashup
24	reference-data-gov-uk	reference.data.gov.uk
25	revyu	Revyu.com - Review Anything
26	sec-rdfabout	U.S. Securities and Exchange Commission Corporate Ownership RDF Data (rdfabout)

27	secold	Source Code Ecosystem Linked Data
28	tcmgenedit_dataset	TCMGeneDIT Dataset
29	transport-data-gov-uk	transport.data.gov.uk
30	uk-legislation-api	UK Legislation
31	w3c-wordnet	WordNet 2.0 (W3C)
32	world-factbook-fu-berlin	World Factbook (FU Berlin)
33	yago	YAGO

Table 5-10 : Use case #9 – Datasets transitively linked to from SECOLD

5.10 Summary – Use Cases

In this chapter several use cases have been presented to illustrate the applicability of LODNav to visualize different aspects of the LODC. Among the usage scenarios described are filtering nodes based on:

- Clusters of datasets
- Active SPARQL Endpoints
- Outgoing links from a dataset
- Incoming links into a dataset
- Number of RDF triples held within a dataset
- Number of incoming RDF links into a dataset
- Number of outgoing RDF links from a dataset
- The group a dataset is a member of

While these use cases were described independently, they also can be combined for further customization of the views generated by LODNav. The examples also illustrate how LODNav can provide additional abstractions and insights by iteratively exploring the LODC datasets.

6. Discussion

This chapter provides some points of discussion covering first some threats to validity (Section 6.1) followed by an overview of some related work (Section 6.2), and finishing with some conclusions as well as future work (Section 6.3).

6.1 Threats to Validity

6.1.1 Lack of automation

With the concept of interoperable Open Data becoming increasingly popular [44] the number of datasets as part of the Linking Open Data Project is constantly increasing. One of the initial goals of LODNav was to create tool support that allows for up-to-data representation of the Linking Open Data Cloud. However, given the current architecture of LODNav, with the data being extracted from Datahub's Linking Open Data Cloud organization group (as described in Section 4.2.1) is not yet a fully automated process. In its current implementation LODNav requires the LODC data acquisition to be triggered. In order to provide actual up-to-date information, these updates should be fully automated. This can be achieved by performing these updates either during a regular (frequent) time interval (e.g. daily) or whenever a change of the LODC occurs. Similarly, the availability of the SPARQL endpoints should be polled automatically at time of creation of visualizations.

A full automation of the information extraction of the LODC data is feasible and will be implemented as part of the future work. As a result, while the data currently being visualized by LODNav is correct (as of October 2013), it does not ensure currently that the data is the most up-to-date information available on the LODC.

6.1.2 Problems with information extraction

As discussed in Section 4.2.1, the information from the Linking Open Data Project is extracted through the Datahub Linking Open Data Cloud group API to receive a String of information for each dataset. This String of information must then be parsed in order to obtain the meta-data for the datasets to be displayed within LODNav.

However, within LODNav not all the datasets have a fully complete meta-data section. Some of the datasets are missing their SPARQL endpoint and some lack information related to their outgoing/incoming links.

Several reasons are possible for this lacking information. (1) Member datasets of Datahub's Open Data Cloud group might not have provided all the required information. (2) The format of the returned String provided by Datahub's API does not follow a parseable standard format. If this is the case, then the algorithm used in this research for parsing the meta-data might result in erroneous meta-information.

As a result, some of the meta-information extracted from the datasets might be incomplete or information within LODNav might not be completely accurate. These issues can be addressed by additional data cleaning or by excluding datasets with incomplete or inconsistent data.

6.1.3 Bugs in the program

The development of LODNav is suspect to potential bugs and issues, which is a common problem to any software. While different forms of system testing were performed during the development and testing phase, the software has known and unknown issues. This threat to system stability and quality can be addressed through additional testing and code reviews during future software development iterations.

6.1.4 Datahub has been changed to use *Organizations*

The data contained within LODNav as presented in this thesis uses data extracted from Datahub's Linking Open Data Cloud group from early October 2013. This collected data consists of 337 datasets which are members of the Linking Open Data Project.

Datahub has since that time undergone a change on how groups are managed within its repository. On 2013-10-11 Datahub has migrated its structural scheme to no longer display 'groups' but instead to display 'organizations' [18]. All datasets which were part of a 'group' are now part of an 'organization'.

The purpose of this migration was to remove spam that was present in Datahub and also to prevent further spam to be published. As part of this change, administrators of 'organizations' on Datahub must now authorize the publishing of datasets to their respective organization [17].

At the time of writing there appears to be two different organizations on Datahub for the datasets of the LOD Project being; *Linking Open Data Cloud* and *Linking Open Data*. The organization name *Linking Open Data Cloud* has 228 member datasets and the organization named *Linking Open Data* has 47 member datasets [15]. While conducting this research it is not clear why the number has drastically changed, nor which organization should be considered as the actual organization to reflect the current status of the LOD Project. It should further be noted that certain datasets appear in one of the organizations but not the other. For instance, DBpedia is present in the organization *Linking Open Data* but not in *Linking Open Data Cloud,* while SECOLD is not present in *Linking Open Data* but it is present in *Linking Open Data Cloud*.

The Linking Open Data Project website does not yet reflect this data migration therefore it is unclear what the actual state of the project is. Perhaps some remaining datasets still need to be migrated or perhaps the number of datasets has in fact been reduced (removing spam data). For the purpose of this research, it has been decided to display the information extracted from Datahub for the LODC prior to the data migration.

6.1.5 Scalability issues

LODNav has been designed to load automatically all the meta-data for each of the LODC datasets into the user's web browser, when a user selects the 'Navigate' page of the LODNav website. This was a design decision during the initial development stages of the application. The decision was supported by the fact that it was unclear where the website was going to be hosted online or made publicly accessible. As a result, LODNav was not developed to retrieve information as needed from a hosted database, instead all the information is retrieved from an XML file upon loading.

This design decision leads to a noticeable delay during the start-up and the internal implementation of the Internet browser being used might limit the number of datasets being loaded. While the current 337 datasets do not create any problems, hard and software limits used by LODNav potentially create a bottleneck for a very large number of datasets.

6.1.6 Logical Hosting Issues

LODNav determines the Latitude & Longitude of a dataset based on the IP address from its URL; however, this approach will not always work, such as when hosting is conducted by an external provider. Using the approach set out in LODNav, the physical location of the dataset will be established where the dataset is being hosted, although it may be more appropriate to place the dataset at the physical location where the individuals who compile the dataset are located.

In order to get an idea of the reliability of LODNav when displaying the location of the dataset with its current approach, a subset of the 337 datasets has been manually inspected with their displayed location and actual location being compared. The subset comprises 20% of the original 337 datasets (68 in total) and has been randomly selected. For each of these 68 datasets, the provided URL which was used to determine the Latitude & Longitude in LODNav has been visited and given the information in the webpage it has been determined whether the location in LODNav is the correct geographical location (for a number of the datasets, it could not be determined whether the location in LODNav was in fact correct or incorrect due to a lack of information in resources provided by the URL).

Results for inspecting the location of 68 datasets	
Location is correct:	41
Location is incorrect:	15
Location could not be determined	12

Table 6-1 : Dataset inspection results

If the datasets whose location could not be determined are omitted from the results, we see that 41 of 56 datasets have the correct location, implying a success rate of 73% for LODNav's method of placing datasets on a world map. For a complete table of the results refer to Appendix E.

6.2 Related Work

In this section, other representations of the Linking Open Data Project will be reviewed and differences be discussed, including their merits as well as shortcomings.

6.2.1 LOD Cloud diagram

Throughout the thesis it was shown that the Linking Open Data Cloud is an interconnection of RDF datasets [50]; this graph visualization has been previously shown in Figure 3-4. Figure 6-1 shows a zoomed-in view of the Linking Open Data Cloud diagram. What follows is an explanation of what type of information is conveyed in this existing graph visualization.

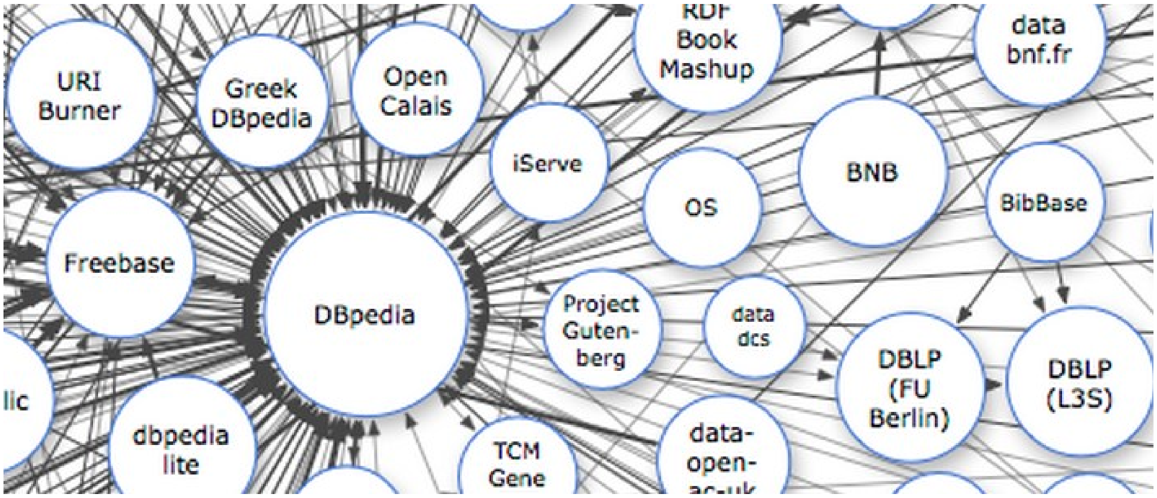


Figure 6-1 : A closer look at the LOD diagram

In the LODC diagram, the size of the datasets has a relation to the number of triples present in the actual dataset as explained Table 6-2.

Circle Size	**Triple Count**
Very large	> 1,000,000,000
Large	1,000,000,000 – 10,000,000
Medium	10,000,000 – 500,000
Small	500,000 – 10,000
Very small	< 10,000

Table 6-2 : LOD diagram circle size in relation to triple count [50]

Links are also presented in the diagram as directional edges when resources in one dataset are linked to resources in another dataset. Bidirectional links are present when links exist in both directions between datasets. The thickness of the line has a relation to the number of RDF links between the datasets as explained in Table 6-3.

Arrow Thickness	**Triple Count**
Thick	> 100,000
Medium	100,000 – 1,000
Thin	< 1,000

Table 6-3 : LOD diagram line thickness in relation to triple count [50]

The location of the datasets in the diagram reflects also the fact that datasets with a larger number of incoming links from external datasets will be displayed closer to the center of the diagram. In addition, the layout of the dataset nodes attempts to reduce the number of crossed edges.

Figure 6-2 shows an alternative layout of the dataset nodes in the diagram. This version hosted at [51] applies a predefined colour scheme to further group the dataset nodes based on their application/domain. Figure 6-2 shows this diagram. In order to improve its readability, the legend indicating the different types of groups has been manually annotated to help identify the different dataset domain groups.

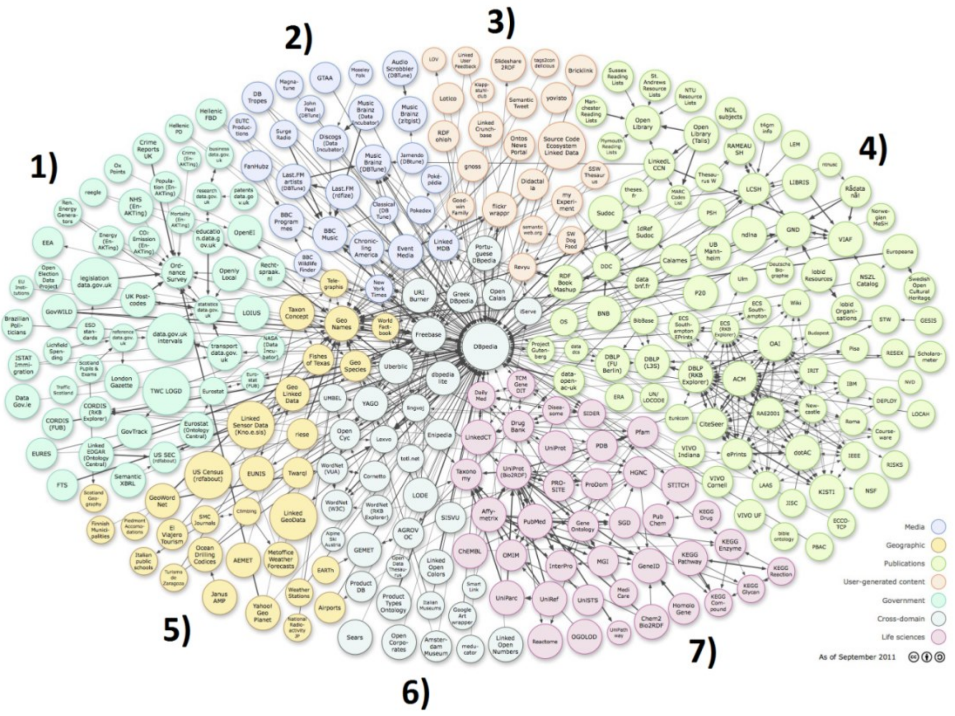


Figure 6-2 : Coloured groupings of LODC diagram

The objective of this diagram is to group and colour code datasets from a similar domain together. Table 6-4 summarizes the 7 groups currently used in this diagram to group the datasets.

Group	Domain
1	Government
2	Media
3	User-generated content
4	Publications
5	Geographic
6	Cross-domain
7	Life sciences

Table 6-4 : Group descriptions from LODC coloured diagram

As a result, the colour-coded graph visualization from Figure 6-2 attempts to meet the following three main criteria:

1. Datasets with more incoming links are placed towards the center.

2. Datasets are placed to reduce the number of overlapping links.

3. Similar types of datasets are grouped together.

These guide the layout of the actual diagram. The final step in creating this visualization is the neatly placing the datasets on the diagram with a manual hand-curated approach providing a well-organized looking shape.

A final feature of this diagram when viewing it at [50] or the colour version at [51], a user is able to click on the individual nodes. By performing this action, the web browser loads the Datahub page for that particular dataset. This interactive feature allows users to view and explore more detailed information for each of these datasets. The disadvantage of this approach is however, that users have to deal with a context change in the data representation, when switching from the graphical notation to a complete new HTML view. Furthermore, no filtering or update mechanisms are provided for the graph visualization, making the information captured in the diagram static in nature. In addition, only limited information is provided with respect to the linking of the datasets (incoming/outgoing/bidirectional links) and no information about the availability or status of the SPARQL endpoints is provided.

6.2.2 Datahub LOD Organization pages

Datahub [14] can be considered as a representation of the Linking Open Data Project since it contains the registered information for each of the member datasets of the LOD Project. While the HTML pages contain all available information, the textual representation does not allow representing all datasets in a single view. In addition no grouping or filtering mechanism is available, as well as the interconnections and transitive dependencies among datasets are difficult to determine.

6.2.3 Mondeca – SPARQL endpoint availability

Mondeca Labs has developed an application which monitors the SPARQL endpoint availability of all publicly accessible SPARQL endpoints and is accessible online[10] (at the time of writing the tool was not accessible, however it was accessible in January 2014) [45] [73].

This tool verifies the availability of all publicly available SPARQL Endpoints referenced by CKAN. The availability of endpoints is determined by conducting the following tests on an hourly basis [45]:

1) Checking the access to the server offering the SPARQL service over the HTTP protocol.
2) The accessibility of the SPARQL service is tested with the following “ASK” query:

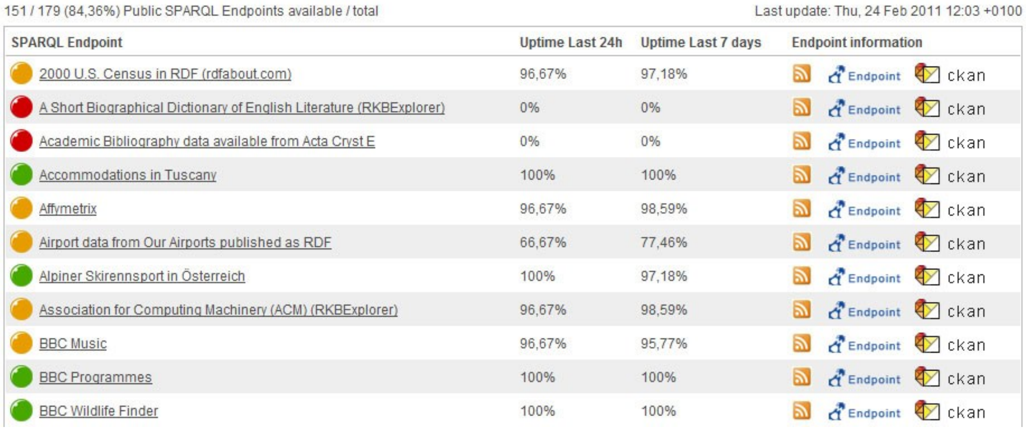
```
ASK WHERE {?s ?p ?o.}
```

[10] http://sparqles.okfn.org/availability

3) Certain endpoints do not accept the “ASK” query, in those cases a “SELECT” query is then used:

```
SELECT WHERE {?s ?p ?o}LIMIT 1
```

If results are returned from these steps then the SPARQL endpoint is deemed as being available for that hourly period. This information is used to display the availability of the SPARQL endpoint over a period of the last 24 hours as well as for the last 7 days. A screenshot of the tool can be seen in Figure 6-3.



151 / 179 (84,36%) Public SPARQL Endpoints available / total

Last update: Thu, 24 Feb 2011 12:03 +0100

SPARQL Endpoint	Uptime Last 24h	Uptime Last 7 days	Endpoint information
🟠 2000 U.S. Census in RDF (rdfabout.com)	96,67%	97,18%	Endpoint ckan
🔴 A Short Biographical Dictionary of English Literature (RKBExplorer)	0%	0%	Endpoint ckan
🔴 Academic Bibliography data available from Acta Cryst E	0%	0%	Endpoint ckan
🟢 Accommodations in Tuscany	100%	100%	Endpoint ckan
🟠 Affymetrix	96,67%	98,59%	Endpoint ckan
🟠 Airport data from Our Airports published as RDF	66,67%	77,46%	Endpoint ckan
🟢 Alpiner Skirennsport in Österreich	100%	97,18%	Endpoint ckan
🟠 Association for Computing Machinery (ACM) (RKBExplorer)	96,67%	98,59%	Endpoint ckan
🟠 BBC Music	96,67%	95,77%	Endpoint ckan
🟢 BBC Programmes	100%	100%	Endpoint ckan
🟢 BBC Wildlife Finder	100%	100%	Endpoint ckan

Figure 6-3 : SPARQL Endpoint Status Example [46]

Since this tool checks for the availability of all the public SPARQL endpoints registered on CKAN, also the datasets, which are members of the LOD Project are represented within this tool. LODNav uses a similar representation by offering the availability of the SPARQL endpoints within the visualization as described in Section 4.4. Mondeca’s Endpoint Availability application performs its checks on a more regular basis and offers more statistical information about the availability of the endpoint.

6.2.4 inkdroid – Linked Open Data Graph

A website known as *inkdroid* offers a visualization of the Linking Open Data Project[11] [13]. This is a visualization which is built with *Protovis*, a JavaScript visualization library [49] [8]. This tool extracts data from the CKAN API (similar to LODNav) in order to populate data for the visualization. There is a colour scheme for the nodes on the graph reflecting their respective CKAN rating. Green is used to represent average ratings, while low averages are in red [35]. Further, a strong colour signifies many ratings, near-white colours represent few ratings, and white represents unrated datasets [35]. The number of triples located within the dataset of the node also plays a factor. Nodes containing more than 5,000,000 triples have their name presented next to the node [35]. There are links between the nodes but their significance is not explicitly stated. The links also do not have directional arrows to signify the direction of linking between the datasets. When a user clicks on a node within this visualization they are taken to the dataset's page located on Datahub. Figure 6-4 provides an example of a visualization created by *inkdroid*.

[11] http://inkdroid.org/lod-graph/

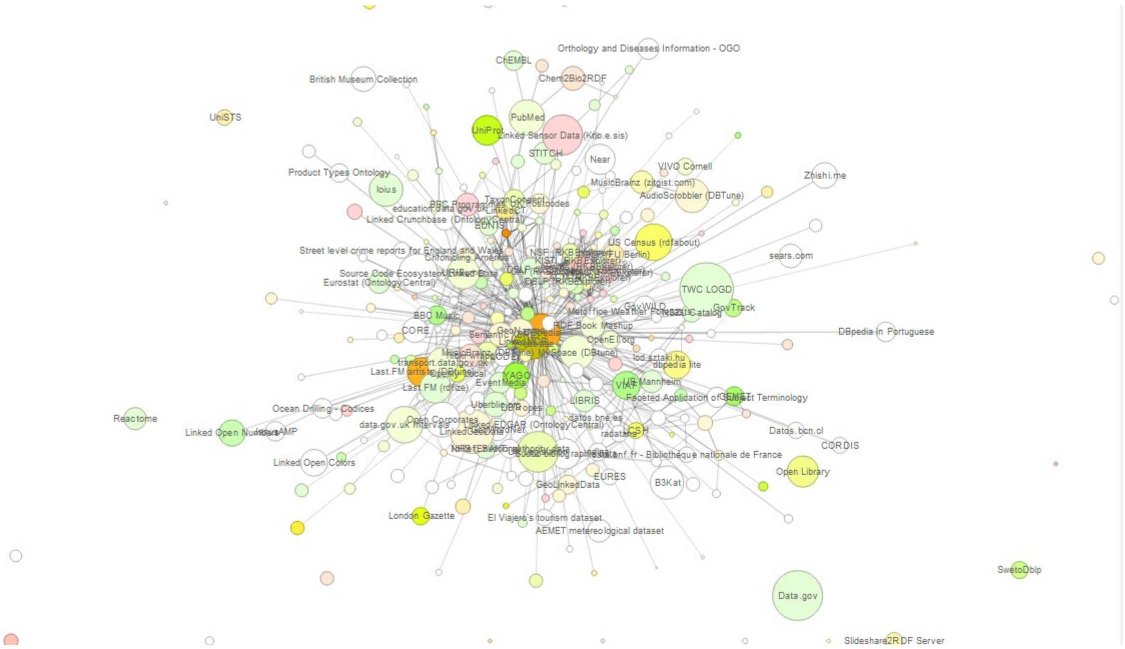


Figure 6-4 : inkdroid Linked Open Data Graph [35]

This tool offers an overview of the Linking Open Data Project however, the nodes tend to be clumped into tightly packed groups. Furthermore, due to information overload, it is often difficult to effectively understand how datasets are actually linked together. The size of the nodes presumably signifies the number of triples within the dataset, which is something that LODNav does not visualize in its current form. The visualization not only also lacks the geographical information included with LODNav, but also LODNav's detailed filtering mechanisms.

6.2.5 Linked Data Browsers

Another approach for understanding Linked Data is by using a Linked Data Browser (LD Browser) [13]. A LD Browser has the general goal of allowing a user to traverse RDF links between resources and provide the user with information about the resources they encounter [7]. This type of navigation would be similar to how users can traverse HTML pages by following hypertext links [7]. Ultimately, a user of a LD browser should be able to traverse links between different Linked Data datasets. This type of navigation poses a number of opportunities and challenges for presenting useful aggregations of information by means of a user interface [7].

In [13] a distinction is made between LD Browsers that provide text-based user functionality and LD Browsers, which provide visual presentations of information. Some text based LD Browsers are; *Dipper*, *Disco*, *Marbles*, *Piggy Bank*, *Sig.ma*, *URI Burner*, and *Zitgist DataViewer* [13]. Some browsers with visualization capabilities include; *DBpedia Mobile*, *Fenfire*, *IsaViz*, *LESS*, *OpenLink Data Explorer*, *RDF Gravity*, *RelFinder*, and *Tabulator* [13].

These tools provide functionality, which is out of the scope of LODNav. LD Browsers provide functionality to navigate RDF links between the Web of Data as part of the LODC. These tools however lack the general abstraction of the LODC, capturing the linking of datasets. In contrast to LODNav, they also provide only limited grouping and filtering mechanisms to represent the LODC dataset information.

6.2.6 RDF Visualizations

The LODC at its most fundamental underpinning is a distributed network of datasets based on RDF data. As a result, the LODC can be considered as a single RDF graph and visualizing the LODC is inherently a RDF visualization problem.

RDF visualizations exist in a variety of different forms including; graphs, charts, dashboards, maps, etc., [42]. One of the main challenges when visualizing RDF data is the ability to cope with the sheer size of the information while still being able to provide a clear succinct overview of the information to a human user [26].

Some approaches for displaying RDF information are based on the use of ontology editors, such as Protégé, OntoEdit, RDF Instance Creator (RIC), and TopBraid, to name a few [26]. These are text-based environments, which support the browsing and editing of RDF data [26]. This visualization is however limited in its scalability. With increasing RDF size it becomes increasingly difficult to present information in a manner which facilitates the comprehension and navigation of the data [26]. For this reason it is often advocated to use visual tools for the display and browsing of RDF data to maximize the effectiveness screen real-estate [26] [23]. Some tools which provide RDF graph visualizations are IsaViz, OntoRAMA, RDF Gravity, and Protégé plugins such as OntoViz or Jambalaya [26] [38]. Again, this approach has issues when the size of data increases and therefore resulting in an incomprehensible Great Big Graph (GBG) or, even more negatively, a Big Fat Graph (BFG) [38].

An alternative to presenting entire graphs of RDF data is the idea of displaying nodes on the graph to users in an iterative exploratory method [53] [23]. The idea is to only visualize relevant/interesting/hot-spot portions of an RDF graph and allow the user to expand the graph only where they find appropriate [23] [53]. A challenge with these approaches is determining which portion of the graph to initially present to the user and understanding what type of information is actually useful for a specific user scenario [23] [53].

In contrast, LODNav is not visualizing the actual RDF data at the instance level and instead treats the individual nodes of the LODC as black-box instances of nodes of RDF data. Although the actual RDF triples linking the datasets cannot be viewed, LODNav focuses on the visualization of the overall structure of the data. Furthermore, LODNav also allows visualizing multiple datasets and their interconnections, where as many of the existing RDF visualization approaches rather focus on individual datasets and their representation.

6.3 Conclusions and Future Work

With the adoption of Linked Data principles, a different way to publish and consume data using the internet has emerged. By publishing information in RDF format, relationships between data can be described with meaningful links and collections of RDF data are used to create machine-readable directed graphs of information. Linking similar concepts between datasets described in RDF data, allows for sharing, reusing, and extending information captured in individual datasets. Being at the core of creating a Web of Data, the most notable implementation of a structure of Linked Data is the LOD Project, which aims create a network of open datasets. At the time of writing 337 linked datasets are part of the Linking Open Data Cloud, but there is little support in understanding how the datasets are connected.

The research presented in this thesis describes LODNav – Linking Open Data Navigator – as a visualization tool for users to better understand how LODC datasets are interconnected. LODNav offers a novel visualization metaphor of the Linking Open Data Cloud by positioning member datasets on a global map based on their geographical location and displaying the links between datasets. This interactive visualization approach, aims to offer customizable views using different filtering and grouping options. LODNav further provides an interface for extracting RDF data describing the datasets and the relationships between them.

By taking a look at each of the initial requirements we can see how the functionalities provided by LODNav provide a viable solution to each requirement.

Requirement #1: Visualizing centres of expertise in Linking Open Data

By placing datasets of the LODC as nodes on a geographic map where the datasets are located, we can see clusters emerge around geographic locations where multiple datasets are positioned. When multiple datasets are situated at a particular place we can infer that a centre of expertise is emerging at that location, and groups in that region can be a good source for collaboration as well as knowledge-sharing.

Requirement #2: Displaying an overview of all the datasets of the LODC and their respective links

All of the datasets are placed in a single view by placing them on a geographical map using the GoogleMaps API. This allows for a unique perspective and overview of the LODC datasets.

Requirement #3: Filtering mechanisms to provide different views of the data held within the visualization

A number of filtering options are provided to the user to filter which datasets are presented at a given time. These filtering options take into account the different statistical properties of the datasets and allow the user to configure the data presented depending on which properties they want to emphasize.

Requirement #4: Exporting the data held within the LODC visualization

LODNav provides an export feature which allows a user to extract information about the depicted LODC visualization. The extracted information is outputted in RDF format and describes how the LODC datasets are interlinked while providing some of the attributes of the datasets. This provides a user with the ability to reuse the information held within LODNav and gain more insights into the LODC.

This thesis describes the architecture and core features of LODNav as well as several use cases are provided to illustrate the applicability of LODNav. As part of the future work different visualization metaphors can be explored, as well as user studies will be conducted to evaluate the applicability and usefulness of the tool. Automated discovering and establishing links between datasets could be addressed as part of the future work. A possible approach which might be explored is the use of a social networking interface which provides users incentive to discover links and post findings in a collaborative manner.

7. References

1. Auer, Sören, et al. "Dbpedia: A nucleus for a web of open data." *The semantic web*. Springer Berlin Heidelberg, 2007. 722-735.

2. Bauer, Florian, and Martin Kaltenböck. "Linked Open Data: The Essentials."*Edition mono/monochrom, Vienna* (2011).

3. Berners-Lee, Tim, et al. "Tabulator: Exploring and analyzing linked data on the semantic web." *Proceedings of the 3rd International Semantic Web User Interaction Workshop*. Vol. 2006. 2006.

4. Berners-Lee, Tim, James Hendler, and Ora Lassila. "The semantic web."*Scientific American* 284.5 (2001): 28-37.

5. Bizer, Christian. "The emerging web of linked data." *Intelligent Systems, IEEE*24.5 (2009): 87-92.

6. Bizer, Christian, et al. "Linked data on the web (LDOW2008)." *Proceedings of the 17th international conference on World Wide Web*. ACM, 2008.

7. Bizer, Christian, Tom Heath, and Tim Berners-Lee. "Linked data-the story so far." *International journal on semantic web and information systems* 5.3 (2009): 1-22.

8. M. Bostock and J. Heer, "Protovis: A graphical toolkit for visualization". *IEEE Transactions on Visualization and Computer Graphics*. 15(6), pp. 1121-1128.

9. D. Brickley and R.V. Guha, "Resource Description Framework (RDF) Schema Specification". *World Wide Web Consortium* (1999).

10. Brin, Sergey, and Lawrence Page. "The anatomy of a large-scale hypertextual Web search engine." *Computer networks and ISDN systems* 30.1 (1998): 107-117.

11. Buil-Aranda, Carlos, et al. "SPARQL Web-Querying Infrastructure: Ready for Action?." *The Semantic Web–ISWC 2013*. Springer Berlin Heidelberg, 2013. 277-293.

12. Carlsson, Gunnar. "Topology and data." *Bulletin of the American Mathematical Society* 46.2 (2009): 255-308.

13. Dadzie, Aba-Sah, and Matthew Rowe. "Approaches to visualising linked data: A survey." *Semantic Web* 2.2 (2011): 89-124.

14. Datahub (http://datahub.io/) – accessed 03/06/2014

15. Datahub – Organizations – Linking Open Data (http://datahub.io/organization/about/lod) accessed 02/11/2014

16. Datahub – Organizations – Linking Open Data Cloud (http://datahub.io/organization/about/lodcloud) accessed 02/11/2014

17. Datahub Blog – Archive for Community Updates (http://blog.datahub.io/category/community-updates/) accessed 02/10/2014

18. Datahub Blog – Organization Migration Complete (http://blog.datahub.io/2013/10/11/organization-migration-complete/) accessed 02/10/2014

19. DBpedia - About (http://dbpedia.org/About) accessed 01/22/2014

20. DBpedia Wiki – Changelog (http://wiki.dbpedia.org/Changelog) accessed 01/22/2014

21. DBpedia Wiki – The DBpedia Data Set (3.9) (http://wiki.dbpedia.org/Datasets) accessed 01/22/2014

22. Decentralized Information Group - Giant Global Graph, 2007 (http://dig.csail.mit.edu/breadcrumbs/node/215) accessed 01/11/2014

23. Deligiannidis, Leonidas, Krys J. Kochut, and Amit P. Sheth. "RDF data exploration and visualization." *Proceedings of the ACM first workshop on CyberInfrastructure: information management in eScience*. ACM, 2007.

24. Euzenat, Jérôme, et al. "Ontology alignment evaluation initiative: six years of experience." *Journal on data semantics XV*. Springer Berlin Heidelberg, 2011. 158-192.

25. FactForge – SPARQL Query (http://factforge.net/sparql) accessed 01/21/2014

26. Frasincar, Flavius, Alexandru Telea, and Geert-Jan Houben. "Adapting graph visualization techniques for the visualization of RDF data." *Visualizing the semantic web*. Springer London, 2006. 154-171.

27. Freifeld, Clark C., et al. "HealthMap: global infectious disease monitoring through automated classification and visualization of Internet media reports."*Journal of the American Medical Informatics Association* 15.2 (2008): 150-157.

28. Gleick, Peter H., and Matthew Heberger. "Water conflict chronology." *The world's water*. Island Press/Center for Resource Economics, 2011. 175-214.

29. Google Developers – Google Maps API (https://developers.google.com/maps/) accessed 01/02/2014

30. Google Developers – Google Maps JavaScript API v3 – Usage Limits and Billing (https://developers.google.com/maps/documentation/javascript/usage#usage_limits) accessed 02/19/2014

31. Gruber, Tom. "Ontology. Encyclopedia of Database Systems, Ling Liu and M. Tamer Özsu." (2009).

32. HealthMap (http://www.healthmap.org/en/) accessed 02/18/2014

33. Heath, Tom, and Christian Bizer. "Linked data: Evolving the web into a global data space." *Synthesis lectures on the semantic web: theory and technology*1.1 (2011): 1-136.

34. Hendler, Jim. "Web 3.0 Emerging." *Computer* 42.1 (2009): 111-113.

35. inkdroid – Linked Open Data Graph (http://inkdroid.org/lod-graph) accessed 02/11/2014

36. IPInfoDB – IP Location API (http://www.ipinfodb.com/ip_location_api.php) accessed 02/01/2014

37. Jain, Prateek, et al. "Ontology alignment for linked open data." *The Semantic Web–ISWC 2010*. Springer Berlin Heidelberg, 2010. 402-417.

38. Karger, David. "The pathetic fallacy of RDF." (2006). (http://eprints.soton.ac.uk/262911/1/the_pathetic_fallacy_of_rdf-33.html) accessed 03/06/2014

39. Keivanloo, Iman, et al. "Towards sharing source code facts using linked data."*Proceedings of the 3rd International Workshop on Search-Driven Development: Users, Infrastructure, Tools, and Evaluation*. ACM, 2011.

40. Keivanloo, Iman, Juergen Rilling, and Philippe Charlan. "Semantic Web-The Missing Link in Global Source Code Analysis?" *Computer Software and Applications Conference (COMPSAC), 2012 IEEE 36th Annual.* IEEE, 2012.

41. Khatchadourian, Shahan, and Mariano P. Consens. "Explod: Summary-based exploration of interlinking and rdf usage in the linked open data cloud." *The Semantic Web: Research and Applications*. Springer Berlin Heidelberg, 2010. 272-287.

42. Leida, Marcello, Ali Afzal, and Basim Majeed. "Outlines for dynamic visualization of semantic web data." *On the Move to Meaningful Internet Systems: OTM 2010 Workshops*. Springer Berlin Heidelberg, 2010.

43. Linking Open Data W3C SWEO Community Project Homepage (http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData) accessed 1/22/2014

44. Miller, Paul, Rob Styles, and Tom Heath. "Open Data Commons, a License for Open Data." *LDOW* 369 (2008).

45. Mondeca – SPARQL Endpoint Status (http://labs.mondeca.com/sparqlEndpointsStatus.html) accessed 02/11/2014

46. Mondeca – SPARQL Endpoint Status – image (http://labs.mondeca.com/img/sparqlEndpointsStatus/sparqlEndpointsStatus_01.jpg) accessed 02/11/2014

47. Ontology Alignment Evaluation Initiative (http://oaei.ontologymatching.org/) accessed 02/12/2014

48. Protégé (http://protege.stanford.edu/) accessed 01/31/2014

49. Protovis (http://mbostock.github.io/protovis/) accessed 02/11/2014

50. The Linking Open Data cloud diagram (http://lod-cloud.net/) accessed 01/15/2014

51. The Linking Open Data colored cloud diagram (http://lod-cloud.net/versions/2011-09-19/lod-cloud_colored.html) accessed 02/01/2014

52. The Self-Describing Web: RDF Section (http://www.w3.org/2001/tag/doc/selfDescribingDocuments.html#RDFSection) accessed 01/15/2014

53. Sayers, Craig. "Node-centric rdf graph visualization." *Mobile and Media Systems Laboratory, HP Labs* (2004).

54. SECOLD (http://secold.org/) accessed 01/23/2014

55. Semantic Web – Ontology (http://semanticweb.org/wiki/Ontology) accessed 01/16/2014

56. Shneiderman, Ben. "The eyes have it: A task by data type taxonomy for information visualizations." *Visual Languages, 1996. Proceedings., IEEE Symposium on*. IEEE, 1996.

57. Shrestha, Shailesh, and Franz-Josef Behr. "Data Visualization Using Google Maps: the Hard Way and the Easy Way." *AGSE 2010*: 125.

58. U.S. Census 2010 Interactive Population Map (http://www.census.gov/2010census/popmap/) accessed 02/18/2014

59. W3C – About W3C (http://www.w3.org/Consortium/) accessed 01/15/2014

60. W3C – OWL Web Ontology Language Guide (http://www.w3.org/TR/owl-guide) accessed 01/20/2014

61. W3C – SPARQL 1.1 Update (http://www.w3.org/TR/2013/REC-sparql11-update-20130321/) accessed 01/21/2014

62. W3C – SPARQL Query Language for RDF (http://www.w3.org/TR/rdf-sparql-query/) accessed 01/20/2014

63. W3C – RDF Current Status (http://www.w3.org/standards/techs/rdf#w3c_all) accessed 01/15/2014

64. W3C – RDF Primer (http://www.w3.org/TR/rdf-primer/) accessed 01/15/2014

65. W3C – RDF Vocabulary Description Language 1.0: RDF Schema (http://www.w3.org/TR/rdf-schema/) accessed 01/20/2014

66. W3C - Resource Description Framework (RDF) (http://www.w3.org/2001/sw/wiki/RDF) accessed 01/15/2014

67. W3C Semantic Web Activity (http://www.w3.org/2001/sw/) accessed 01/15/2014

68. W3C – Tim Berners-Lee (http://www.w3.org/People/Berners-Lee/) accessed 01/15/2014

69. W3C - Wiki – SparqlEndpoints (http://www.w3.org/wiki/SparqlEndpoints) accessed 01/21/2014

70. W3C – Wiki – Web Ontology Language (OWL) (http://www.w3.org/2001/sw/wiki/OWL) accessed 01/20/2014

71. W3C – Design Issues – Linked Data (http://www.w3.org/DesignIssues/LinkedData.html) accessed 01/21/2014

72. Water Conflict Chronology Map (http://www2.worldwater.org/conflict/map/) accessed 02/18/2014

73. Vandenbussche Pierre-Yves. “Qualité et Robustesse dans le Web de Données: SPARQL Endpoints Status,” White paper, Mondeca, (2012).

74. Yu, Liyang. *A developer's guide to the semantic Web*. Heidelberg: Springer, 2011.

8. Appendix

A) SPARQL Endpoint availability over 24 hour period

This provides the availability of all the LODC SPARQL Endpoints for the time period of 2014-04-12 14:00 to 2014-04-13 14:00.

ID	Dataset Codename	Endpoint Availability Percentage	Endpoint Availability by hour
0	2000-us-census-rdf	0%	000000000000000000000000
1	aemet	100%	111111111111111111111111
2	agrovoc-skos	0%	000000000000000000000000
3	amsterdam-museum-as-edm-lod	0%	000000000000000000000000
4	archiveshub-linkeddata	100%	111111111111111111111111
5	austrian_ski_racers	100%	111111111111111111111111
6	b3kat	100%	111111111111111111111111
7	bbc-music	0%	000000000000000000000000
8	bbc-programmes	0%	000000000000000000000000
9	bbc-wildlife-finder	0%	000000000000000000000000
10	beneficiaries-of-the-european-commission	100%	111111111111111111111111
11	bfs-linked-data	100%	111111111111111111111111
12	bibbase	0%	000000000000000000000000
13	bible-ontology	0%	000000000000000000000000
14	bio2rdf-affymetrix	100%	111111111111111111111111
15	bio2rdf-cas	0%	000000000000000000000000
16	bio2rdf-chebi	100%	111111111111111111111111
17	bio2rdf-genbank	100%	111111111111111111111111
18	bio2rdf-goa	100%	111111111111111111111111
19	bio2rdf-hgnc	100%	111111111111111111111111
20	bio2rdf-homologene	100%	111111111111111111111111
21	bio2rdf-interpro	100%	111111111111111111111111
22	bio2rdf-kegg-compound	0%	000000000000000000000000
23	bio2rdf-kegg-drug	0%	000000000000000000000000
24	bio2rdf-kegg-enzyme	0%	000000000000000000000000
25	bio2rdf-kegg-glycan	0%	000000000000000000000000
26	bio2rdf-kegg-pathway	100%	111111111111111111111111
27	bio2rdf-kegg-reaction	0%	000000000000000000000000
28	bio2rdf-mgi	100%	111111111111111111111111
29	bio2rdf-ncbigene	0%	000000000000000000000000
30	bio2rdf-obo	95%	111101111111111111111111
31	bio2rdf-omim	100%	111111111111111111111111
32	bio2rdf-pdb	100%	111111111111111111111111
33	bio2rdf-pfam	100%	111111111111111111111111
34	bio2rdf-prodom	100%	111111111111111111111111
35	bio2rdf-prosite	100%	111111111111111111111111

36	bio2rdf-pubchem	0%	000000000000000000000000
37	bio2rdf-pubmed	100%	111111111111111111111111
38	bio2rdf-reactome	100%	111111111111111111111111
39	bio2rdf-sgd	100%	111111111111111111111111
40	bio2rdf-unists	0%	000000000000000000000000
41	bluk-bnb	100%	111111111111111111111111
42	brazilian-politicians	0%	000000000000000000000000
43	bricklink	0%	000000000000000000000000
44	british-museum-collection	0%	000000000000000000000000
45	business-data-gov-uk	0%	000000000000000000000000
46	calames	0%	000000000000000000000000
47	chem2bio2rdf	0%	000000000000000000000000
48	chronicling-america	0%	000000000000000000000000
49	clean-energy-data-reegle	100%	111111111111111111111111
50	colinda	0%	000000000000000000000000
51	core	0%	000000000000000000000000
52	cornetto	0%	000000000000000000000000
53	courts-thesaurus	100%	111111111111111111111111
54	data-bnf-fr	0%	000000000000000000000000
55	data-cnr-it	100%	111111111111111111111111
56	data-gov-ie	0%	000000000000000000000000
57	data-gov-uk-time-intervals	0%	000000000000000000000000
58	data-gov	0%	000000000000000000000000
59	data-incubator-climb	0%	000000000000000000000000
60	data-incubator-discogs	0%	000000000000000000000000
61	data-incubator-metoffice	0%	000000000000000000000000
62	data-incubator-moseley	0%	000000000000000000000000
63	data-incubator-musicbrainz	0%	000000000000000000000000
64	data-incubator-nasa	0%	000000000000000000000000
65	data-incubator-our-airports	0%	000000000000000000000000
66	data-incubator-pokedex	0%	000000000000000000000000
67	data-incubator-smcjournals	0%	000000000000000000000000
68	data-open-ac-uk	100%	111111111111111111111111
69	datos-bcn-cl	62%	000000000111111111111111
70	datos-bne-es	100%	111111111111111111111111
71	dbpedia-el	100%	111111111111111111111111
72	dbpedia-lite	0%	000000000000000000000000
73	dbpedia-pt	100%	111111111111111111111111
74	dbpedia	100%	111111111111111111111111
75	dbtropes	0%	000000000000000000000000
76	dbtune-artists-last-fm	0%	000000000000000000000000
77	dbtune-audioscrobbler	0%	000000000000000000000000
78	dbtune-classical	0%	000000000000000000000000
79	dbtune-john-peel-sessions	0%	000000000000000000000000
80	dbtune-magnatune	0%	000000000000000000000000
81	dbtune-musicbrainz	83%	000011111111111111111111
82	dbtune-myspace	0%	000000000000000000000000
83	dcs-sheffield	0%	000000000000000000000000
84	deutsche-biographie	0%	000000000000000000000000
85	dewey_decimal_classification	100%	111111111111111111111111
86	didactalia	0%	000000000000000000000000
87	dnb-gemeinsame-normdatei	0%	000000000000000000000000
88	ecco-tcp-linked-data	0%	000000000000000000000000

89	ecs	0%	000000000000000000000000
90	educationalprograms_sisvu	100%	111111111111111111111111
91	education-data-gov-uk	100%	111111111111111111111111
92	eea	100%	111111111111111111111111
93	enakting-co2emission	0%	000000000000000000000000
94	enakting-crime	0%	000000000000000000000000
95	enakting-energy	0%	000000000000000000000000
96	enakting-mortality	0%	000000000000000000000000
97	enakting-nhs	0%	000000000000000000000000
98	enakting-population	0%	000000000000000000000000
99	enipedia	100%	111111111111111111111111
100	environmental-applications-reference-thesaurus	100%	111111111111111111111111
101	esd-standards	0%	000000000000000000000000
102	eu-institutions	0%	000000000000000000000000
103	eumida-linked-data	0%	000000000000000000000000
104	eunis	100%	111111111111111111111111
105	europeana-lod	0%	000000000000000000000000
106	eurostat-rdf	0%	000000000000000000000000
107	eutc-productions	0%	000000000000000000000000
108	event-media	100%	111111111111111111111111
109	fanhubz	0%	000000000000000000000000
110	fao-geopolitical-ontology	0%	000000000000000000000000
111	fao-linked-data	100%	111111111111111111111111
112	farmbio-chembl	62%	111100000000011111111111
113	farmers-markets-geographic-data-united-states	100%	111111111111111111111111
114	finnish-municipalities	0%	000000000000000000000000
115	fishes-of-texas	0%	000000000000000000000000
116	flickr-wrappr	0%	000000000000000000000000
117	freebase	0%	000000000000000000000000
118	fu-berlin-cordis	0%	000000000000000000000000
119	fu-berlin-dailymed	0%	000000000000000000000000
120	fu-berlin-dblp	0%	000000000000000000000000
121	fu-berlin-diseasome	0%	000000000000000000000000
122	fu-berlin-drugbank	0%	000000000000000000000000
123	fu-berlin-eures	0%	000000000000000000000000
124	fu-berlin-eurostat	0%	000000000000000000000000
125	fu-berlin-medicare	0%	000000000000000000000000
126	fu-berlin-project-gutenberg	0%	000000000000000000000000
127	fu-berlin-sider	0%	000000000000000000000000
128	fu-berlin-stitch	0%	000000000000000000000000
129	gemeenschappelijke-thesaurus-audiovisuele-archieven	0%	000000000000000000000000
130	gemet	0%	000000000000000000000000
131	geolinkeddata	100%	111111111111111111111111
132	geological-survey-of-austria-thesaurus	100%	111111111111111111111111
133	geonames-semantic-web	0%	000000000000000000000000
134	geospecies	0%	000000000000000000000000
135	geowordnet	0%	000000000000000000000000
136	german-labor-law-thesaurus	100%	111111111111111111111111
137	gesis-thesoz	100%	111111111111111111111111

138	gnoss	0%	000000000000000000000000
139	googleart-wrapper	0%	000000000000000000000000
140	government-web-integration-for-linked-data	0%	000000000000000000000000
141	govtrack	0%	000000000000000000000000
142	grrp	100%	111111111111111111111111
143	hellenic-fire-brigade	0%	000000000000000000000000
144	hellenic-police	100%	111111111111111111111111
145	hungarian-national-library-catalog	0%	000000000000000000000000
146	idreffr	0%	000000000000000000000000
147	iserve	100%	111111111111111111111111
148	istat-immigration	0%	000000000000000000000000
149	italian-public-schools-linkedopendata-it	100%	111111111111111111111111
150	jamendo-dbtune	100%	111111111111111111111111
151	japan-radioactivity-stat	0%	000000000000000000000000
152	john-goodwins-family-tree	0%	000000000000000000000000
153	klappstuhlclub	0%	000000000000000000000000
154	knoesis-linked-sensor-data	0%	000000000000000000000000
155	l3s-dblp	100%	111111111111111111111111
156	lcsh	0%	000000000000000000000000
157	lexvo	0%	000000000000000000000000
158	libris	0%	000000000000000000000000
159	libver	0%	000000000000000000000000
160	lichfield-spending	0%	000000000000000000000000
161	lingvoj	0%	000000000000000000000000
162	linked-crunchbase	0%	000000000000000000000000
163	linkedct	0%	000000000000000000000000
164	linked-edgar	0%	000000000000000000000000
165	linked-eurostat	0%	000000000000000000000000
166	linkedgeodata	100%	111111111111111111111111
167	linkedlccn	0%	000000000000000000000000
168	linkedmdb	95%	011111111111111111111111
169	linked-open-data-of-ecology	100%	111111111111111111111111
170	linked-open-numbers	0%	000000000000000000000000
171	linked-open-vocabularies-lov	100%	111111111111111111111111
172	linked-user-feedback	100%	111111111111111111111111
173	lista-encabezamientos-materia	100%	111111111111111111111111
174	lobid-organisations	95%	111101111111111111111111
175	lobid-resources	95%	111101111111111111111111
176	loc	0%	000000000000000000000000
177	loius	0%	000000000000000000000000
178	london-gazette	0%	000000000000000000000000
179	los_metar	0%	000000000000000000000000
180	lotico	0%	000000000000000000000000
181	manchester-university-reading-lists	0%	000000000000000000000000
182	marc-codes	0%	000000000000000000000000
183	meducator	0%	000000000000000000000000
184	morelab	95%	111110111111111111111111
185	museums-in-italy	100%	111111111111111111111111
186	my-experiment	0%	000000000000000000000000

187	my-family-lineage	0%	000000000000000000000000
188	national-diet-library-authorities	0%	000000000000000000000000
189	nomenclator-asturias	0%	000000000000000000000000
190	normesh	0%	000000000000000000000000
191	nottingham-trent-university-resource-lists	0%	000000000000000000000000
192	ntnusc	0%	000000000000000000000000
193	nvd	0%	000000000000000000000000
194	nytimes-linked-open-data	0%	000000000000000000000000
195	oceandrilling-codices	79%	000001111111111111111111
196	oceandrilling-janusamp	79%	000001111111111111111111
197	oclc-fast	0%	000000000000000000000000
198	oecd-linked-data	100%	111111111111111111111111
199	ogolod	0%	000000000000000000000000
200	ontos-news-portal	0%	000000000000000000000000
201	opencalais	0%	000000000000000000000000
202	opencorporates	0%	000000000000000000000000
203	opencyc	0%	000000000000000000000000
204	open-data-thesaurus	100%	111111111111111111111111
205	open-election-data-project	0%	000000000000000000000000
206	open-energy-info-wiki	100%	111111111111111111111111
207	open-library	0%	000000000000000000000000
208	openlylocal	0%	000000000000000000000000
209	ordnance-survey-linked-data	0%	000000000000000000000000
210	osm-semantic-network	100%	111111111111111111111111
211	oxpoints	100%	111111111111111111111111
212	patents-data-gov-uk	0%	000000000000000000000000
213	pleiades	0%	000000000000000000000000
214	pokepedia-fr	0%	000000000000000000000000
215	prefix-cc	0%	000000000000000000000000
216	printed-book-auction-catalogues	0%	000000000000000000000000
217	productdb	0%	000000000000000000000000
218	productontology	0%	000000000000000000000000
219	pscs-catalogue	0%	000000000000000000000000
220	psh-subject-headings	0%	000000000000000000000000
221	radatana	0%	000000000000000000000000
222	rdf-book-mashup	0%	000000000000000000000000
223	rdfize-lastfm	0%	000000000000000000000000
224	rdfohloh	0%	000000000000000000000000
225	rechtspraak	0%	000000000000000000000000
226	reference-data-gov-uk	100%	111111111111111111111111
227	renewable_energy_generators	0%	000000000000000000000000
228	research-data-gov-uk	0%	000000000000000000000000
229	revyu	0%	000000000000000000000000
230	riese	0%	000000000000000000000000
231	rkb-explorer-acm	100%	111111111111111111111111
232	rkb-explorer-budapest	100%	111111111111111111111111
233	rkb-explorer-citeseer	100%	111111111111111111111111
234	rkb-explorer-cordis	100%	111111111111111111111111
235	rkb-explorer-courseware	100%	111111111111111111111111
236	rkb-explorer-crime	100%	111111111111111111111111

237	rkb-explorer-curriculum	100%	111111111111111111111111
238	rkb-explorer-darmstadt	100%	111111111111111111111111
239	rkb-explorer-dblp	100%	111111111111111111111111
240	rkb-explorer-deepblue	100%	111111111111111111111111
241	rkb-explorer-deploy	100%	111111111111111111111111
242	rkb-explorer-dotac	100%	111111111111111111111111
243	rkb-explorer-ecs	0%	000000000000000000000000
244	rkb-explorer-eprints	100%	111111111111111111111111
245	rkb-explorer-era	100%	111111111111111111111111
246	rkb-explorer-eurecom	100%	111111111111111111111111
247	rkb-explorer-ft	100%	111111111111111111111111
248	rkb-explorer-ibm	100%	111111111111111111111111
249	rkb-explorer-ieee	100%	111111111111111111111111
250	rkb-explorer-irit	100%	111111111111111111111111
251	rkb-explorer-jisc	100%	111111111111111111111111
252	rkb-explorer-kisti	100%	111111111111111111111111
253	rkb-explorer-laas	100%	111111111111111111111111
254	rkb-explorer-newcastle	100%	111111111111111111111111
255	rkb-explorer-nsf	100%	111111111111111111111111
256	rkb-explorer-oai	100%	111111111111111111111111
257	rkb-explorer-os	100%	111111111111111111111111
258	rkb-explorer-pisa	100%	111111111111111111111111
259	rkb-explorer-rae2001	100%	111111111111111111111111
260	rkb-explorer-resex	100%	111111111111111111111111
261	rkb-explorer-risks	100%	111111111111111111111111
262	rkb-explorer-roma	100%	111111111111111111111111
263	rkb-explorer-southampton	100%	111111111111111111111111
264	rkb-explorer-ulm	100%	111111111111111111111111
265	rkb-explorer-unlocode	100%	111111111111111111111111
266	rkb-explorer-wiki	100%	111111111111111111111111
267	rkb-explorer-wordnet	100%	111111111111111111111111
268	schemapedia	0%	000000000000000000000000
269	scholarometer	0%	000000000000000000000000
270	sears	0%	000000000000000000000000
271	secold	0%	000000000000000000000000
272	sec-rdfabout	0%	000000000000000000000000
273	semantictweet	0%	000000000000000000000000
274	semantic-web-dog-food	100%	111111111111111111111111
275	semanticweb-org	0%	000000000000000000000000
276	semantic-xbrl	0%	000000000000000000000000
277	semsol-crunchbase	0%	000000000000000000000000
278	slideshare2rdf	0%	000000000000000000000000
279	smartlink	100%	111111111111111111111111
280	socialsemweb-thesaurus	100%	111111111111111111111111
281	southampton-ecs-eprints	0%	000000000000000000000000
282	st-andrews-resource-lists	0%	000000000000000000000000
283	statistics-data-gov-uk	100%	111111111111111111111111
284	stitch-rameau	0%	000000000000000000000000
285	stw-thesaurus-for-economics	0%	000000000000000000000000
286	sudocfr	0%	000000000000000000000000
287	surge-radio	0%	000000000000000000000000
288	swedish-open-cultural-heritage	0%	000000000000000000000000

289	sweto-dblp	0%	000000000000000000000000
290	sztaki-lod	100%	111111111111111111111111
291	t4gm-info	0%	000000000000000000000000
292	tags2con-delicious	0%	000000000000000000000000
293	talis-openlibrary	0%	000000000000000000000000
294	taxonconcept	0%	000000000000000000000000
295	tcmgenedit_dataset	0%	000000000000000000000000
296	telegraphis	0%	000000000000000000000000
297	temple-ov-thee-lemur-datasets	0%	000000000000000000000000
298	thesaurus-w	0%	000000000000000000000000
299	thesesfr	0%	000000000000000000000000
300	the-view-from	0%	000000000000000000000000
301	traffic-scotland	0%	000000000000000000000000
302	transparency-linked-data	100%	111111111111111111111111
303	transport-data-gov-uk	100%	111111111111111111111111
304	twarql	0%	000000000000000000000000
305	twc-ieeevis	100%	111111111111111111111111
306	twc-logd	0%	000000000000000000000000
307	uberblic	0%	000000000000000000000000
308	ub-mannheim-linked-data	0%	000000000000000000000000
309	uk-legislation-api	0%	000000000000000000000000
310	uk-postcodes	0%	000000000000000000000000
311	umbel	0%	000000000000000000000000
312	uniprotkb	100%	111111111111111111111111
313	uniprot-taxonomy	100%	111111111111111111111111
314	uniprot-uniparc	100%	111111111111111111111111
315	uniprot-unipathway	100%	111111111111111111111111
316	uniprot	0%	000000000000000000000000
317	uniref	100%	111111111111111111111111
318	university-plymouth-reading-lists	0%	000000000000000000000000
319	university-sussex-reading-lists	0%	000000000000000000000000
320	uriburner	100%	111111111111111111111111
321	viaf	0%	000000000000000000000000
322	vivo-cornell-university	0%	000000000000000000000000
323	vivo-indiana-university	0%	000000000000000000000000
324	vivo-university-of-florida	0%	000000000000000000000000
325	vu-wordnet	0%	000000000000000000000000
326	w3c-wordnet	0%	000000000000000000000000
327	webnmasunotraveler	100%	111111111111111111111111
328	world-bank-linked-data	83%	000011111111111111111111
329	world-factbook-fu-berlin	0%	000000000000000000000000
330	yago	0%	000000000000000000000000
331	yahoo_geoplanet	0%	000000000000000000000000
332	yovisto	100%	111111111111111111111111
333	zaragoza-turismo	100%	111111111111111111111111
334	zbw-pressemappe20	0%	000000000000000000000000
335	zhishi-me	0%	000000000000000000000000
336	zitgist-musicbrainz	0%	000000000000000000000000

Table 8-1 : SPARQL endpoint availability over 24 hours

B) Use Case #2: All datasets with an Active SPARQL endpoint

#	Dataset Codename	Dataset Title
1	aemet	AEMET metereological dataset
2	archiveshub-linkeddata	Archives Hub Linked Data
3	austrian_ski_racers	Alpine Ski Racers of Austria
4	b3kat	B3Kat - Library Union Catalogues of Bavari
5	beneficiaries-of-the-european-commission	EU: fintrans.publicdata.eu
6	bfs-linked-data	Bundesamt für Statistik (BFS) - Swiss Federal Statistical Office (FSO) Linked Data
7	bio2rdf-affymetrix	Bio2RDF::Affymetrix
8	bio2rdf-genbank	Bio2RDF - GenBank
9	bio2rdf-goa	Bio2RDF::Gene Ontology Annotations
10	bio2rdf-hgnc	Bio2RDF::HGNC
11	bio2rdf-homologene	Bio2RDF::HomoloGene
12	bio2rdf-interpro	Bio2RDF::InterPro
13	bio2rdf-kegg-pathway	KEGG Pathway
14	bio2rdf-mgi	Mouse Genome Database (MGD) from Mouse Genome Informatics (MGI)
15	bio2rdf-ncbigene	Bio2RDF::NCBI Gene

16	bio2rdf-obo	OBO
17	bio2rdf-omim	Bio2RDF::OMIM
18	bio2rdf-pubmed	PubMed
19	bio2rdf-sgd	Saccharomyces Genome Database
20	bluk-bnb	British National Bibliography (BNB) - Linked Open Data
21	british-museum-collection	British Museum Collection
22	clean-energy-data-reegle	Linked Clean Energy Data (reegle.info)
23	courts-thesaurus	Courts thesaurus
24	data-cnr-it	Open Data from the Italian National Research Council
25	data-gov	Data.gov
26	data-open-ac-uk	data.open.ac.u
27	datos-bcn-cl	Datos.bcn.cl
28	datos-bne-es	datos.bne.es
29	dbpedia	DBpedia
30	dbpedia-el	DBpedia in Greek
31	dbpedia-pt	DBpedia in Portuguese
32	dbtune-musicbrainz	DBTune.org Musicbrainz D2R Server

33	dewey_decimal_classification	Dewey Decimal Classification (DDC)
34	education-data-gov-uk	education.data.gov.uk
35	educationalprograms_sisvu	Educational programs - SISVU
36	eea	European Environment Agency Published Products
37	enipedia	Enipedia - Energy Industry Data
38	environmental-applications-reference-thesaurus	EARTh
39	eunis	European Nature Information System
40	event-media	EventMedia
41	fao-linked-data	Food and Agriculture Organization of the United Nations (FAO) Linked Data
42	farmers-markets-geographic-data-united-states	Farmers Markets Geographic Data (United States)
43	geolinkeddata	GeoLinkedData
44	geological-survey-of-austria-thesaurus	Geological Survey of Austria (GBA) - Thesaurus
45	german-labor-law-thesaurus	German labor law thesaurus
46	gesis-thesoz	TheSoz Thesaurus for the Social Sciences (GESIS)
47	hellenic-police	Hellenic Police
48	hungarian-national-library-catalog	Hungarian National Library (NSZL) catalog

49	iserve	iServe: Linked Services Registry
50	jamendo-dbtune	DBTune.org Jamendo RDF Server
51	l3s-dblp	DBLP in RDF (L3S)
52	linked-open-data-of-ecology	Linked Open Data of Ecology
53	linked-open-vocabularies-lov	Linked Open Vocabularies (LOV)
54	linked-user-feedback	Linked User Feedback
55	linkedmdb	Linked Movie DataBase
56	lista-encabezamientos-materia	Lista de Encabezamientos de Materia as Linked Open Data
57	lobid-organisations	lobid. Index of libraries and related organisations
58	lobid-resources	lobid. Bibliographic Resources
59	morelab	morelab
60	nomenclator-asturias	Nomenclator Asturias
61	oecd-linked-data	Organisation for Economic Co-operation and Development (OECD) Linked Data
62	open-data-thesaurus	Open Data Thesaurus
63	open-energy-info-wiki	OpenEI - Open Energy Info
64	osm-semantic-network	OSM Semantic Network
65	oxpoints	OxPoints (University of Oxford)

66	pscs-catalogue	Product Scheme Classifications Catalogue
67	reference-data-gov-uk	reference.data.gov.uk
68	revyu	Revyu.com - Review Anything
69	rkb-explorer-acm	Association for Computing Machinery (ACM) (RKBExplorer)
70	rkb-explorer-budapest	Budapest University of Technology and Economics (RKBExplorer)
71	rkb-explorer-citeseer	CiteSeer (Research Index) (RKBExplorer)
72	rkb-explorer-cordis	Community R&D Information Service (CORDIS) (RKBExplorer)
73	rkb-explorer-courseware	Resilient Computing Courseware (RKBExplorer)
74	rkb-explorer-crime	Street level crime reports for England and Wales
75	rkb-explorer-curriculum	ReSIST MSc in Resilient Computing Curriculum (RKBExplorer)
76	rkb-explorer-darmstadt	Technische Universität Darmstadt (RKBExplorer)
77	rkb-explorer-dblp	DBLP Computer Science Bibliography (RKBExplorer)
78	rkb-explorer-deepblue	Deep Blue (RKBExplorer)
79	rkb-explorer-deploy	DEPLOY (RKBExplorer)
80	rkb-explorer-dotac	dotAC (RKBExplorer)
81	rkb-explorer-eprints	ePrints3 Institutional Archive Collection

		(RKBExplorer)
82	rkb-explorer-era	ERA - Australian Research Council publication ratings (RKBExplorer)
83	rkb-explorer-eurecom	Institut Eurécom (RKBExplorer)
84	rkb-explorer-ft	France Telecom Recherche et Développement (RKBExplorer)
85	rkb-explorer-ibm	IBM Research GmbH (RKBExplorer)
86	rkb-explorer-ieee	IEEE Papers (RKBExplorer)
87	rkb-explorer-irit	Université Paul Sabatier - Toulouse 3 (RKB Explorer)
88	rkb-explorer-jisc	UK JISC (RKBExplorer)
89	rkb-explorer-kisti	Korean Institute of Science Technology and Information (RKBExplorer)
90	rkb-explorer-laas	LAAS-CNRS (RKBExplorer)
91	rkb-explorer-newcastle	University of Newcastle upon Tyne (RKBExplorer)
92	rkb-explorer-nsf	National Science Foundation (RKBExplorer)
93	rkb-explorer-oai	Open Archive Initiative Harvest over OAI-PMH (RKBExplorer)
94	rkb-explorer-os	Ordnance Survey (RKBExplorer)
95	rkb-explorer-pisa	Università di Pisa (RKBExplorer)
96	rkb-explorer-rae2001	Research Assessment Exercise 2001 (RKBExplorer)

97	rkb-explorer-resex	ReSIST Resilience Mechanisms (RKBExplorer.com)
98	rkb-explorer-risks	RISKS Digest (RKBExplorer)
99	rkb-explorer-roma	Università degli studi di Roma "La Sapienza" (RKBExplorer)
100	rkb-explorer-southampton	School of Electronics and Computer Scienc
101	rkb-explorer-ulm	Universität Ulm (RKBExplorer)
102	rkb-explorer-unlocode	UN/LOCODE (RKBExplorer)
103	rkb-explorer-wiki	ReSIST Project Wiki (RKBExplorer)
104	rkb-explorer-wordnet	WordNet (RKBExplorer)
105	semantic-web-dog-food	Semantic Web Dog Food Corpus
106	smartlink	SmartLink: Linked Services Non-Functional Properties
107	socialsemweb-thesaurus	Social Semantic Web Thesaurus
108	statistics-data-gov-uk	statistics.data.gov.uk
109	sztaki-lod	National Digital Data Archive of Hungary (partial)
110	transparency-linked-data	Transparency International Linked Data
111	transport-data-gov-uk	transport.data.gov.uk
112	twc-ieeevis	IEEE VIS Source Data
113	uriburner	URIBurner

114	webnmasunotraveler	El Viajero's tourism dataset
115	world-bank-linked-data	World Bank Linked Data
116	yovisto	Yovisto - academic video search
117	zaragoza-turismo	Turismo de Zaragoza

Table 8-2 : Use Case #2 – All datasets with an Active SPARQL Endpoint

C) Use Case #3: All Active endpoints linked to DBpedia

#	Dataset Codename	Dataset Title
1	aemet	AEMET metereological dataset
2	austrian_ski_racers	Alpine Ski Racers of Austria
3	beneficiaries-of-the-european-commission	EU: fintrans.publicdata.eu
4	bfs-linked-data	Bundesamt für Statistik (BFS) - Swiss Federal Statistical Office (FSO) Linked Data
5	clean-energy-data-reegle	Linked Clean Energy Data (reegle.info)
6	courts-thesaurus	Courts thesaurus
7	data-cnr-it	Open Data from the Italian National Research Council
8	datos-bcn-cl	Datos.bcn.cl
9	datos-bne-es	datos.bne.es
10	dbpedia	DBpedia
11	dbpedia-el	DBpedia in Greek
12	dbpedia-pt	DBpedia in Portuguese
13	dbtune-musicbrainz	DBTune.org Musicbrainz D2R Server
14	education-data-gov-uk	education.data.gov.uk
15	enipedia	Enipedia - Energy Industry Data
16	environmental-applications-reference-thesaurus	EARTh
17	eunis	European Nature Information System
18	event-media	EventMedia
19	fao-linked-data	Food and Agriculture Organization of the United Nations (FAO) Linked Data
20	farmers-markets-geographic-data-united-states	Farmers Markets Geographic Data (United States)
21	geolinkeddata	GeoLinkedData
22	geological-survey-of-austria-thesaurus	Geological Survey of Austria (GBA) - Thesaurus
23	german-labor-law-thesaurus	German labor law thesaurus
24	gesis-thesoz	TheSoz Thesaurus for the Social Sciences

		(GESIS)
25	hellenic-police	Hellenic Police
26	hungarian-national-library-catalog	Hungarian National Library (NSZL) catalog
27	linked-open-data-of-ecology	Linked Open Data of Ecology
28	linkedmdb	Linked Movie DataBase
29	lobid-organisations	lobid. Index of libraries and related organisations
30	lobid-resources	lobid. Bibliographic Resources
31	morelab	morelab
32	nomenclator-asturias	Nomenclator Asturias
33	oecd-linked-data	Organisation for Economic Co-operation and Development (OECD) Linked Data
34	open-data-thesaurus	Open Data Thesaurus
35	open-energy-info-wiki	OpenEI - Open Energy Info
36	reference-data-gov-uk	reference.data.gov.uk
37	revyu	Revyu.com - Review Anything
38	rkb-explorer-acm	Association for Computing Machinery (ACM) (RKBExplorer)
39	rkb-explorer-citeseer	CiteSeer (Research Index) (RKBExplorer)
40	rkb-explorer-cordis	Community R&D Information Service (CORDIS) (RKBExplorer)
41	rkb-explorer-courseware	Resilient Computing Courseware (RKBExplorer)
42	rkb-explorer-dblp	DBLP Computer Science Bibliography (RKBExplorer)
43	rkb-explorer-era	ERA - Australian Research Council publication ratings (RKBExplorer)
44	rkb-explorer-nsf	National Science Foundation (RKBExplorer)
45	rkb-explorer-os	Ordnance Survey (RKBExplorer)
46	rkb-explorer-rae2001	Research Assessment Exercise 2001 (RKBExplorer)
47	rkb-explorer-resex	ReSIST Resilience Mechanisms (RKBExplorer.com)
48	rkb-explorer-southampton	School of Electronics and Computer Scienc
49	rkb-explorer-unlocode	UN/LOCODE (RKBExplorer)
50	rkb-explorer-wiki	ReSIST Project Wiki (RKBExplorer)

51	rkb-explorer-wordnet	WordNet (RKBExplorer)
52	socialsemweb-thesaurus	Social Semantic Web Thesaurus
53	sztaki-lod	National Digital Data Archive of Hungary (partial)
54	transparency-linked-data	Transparency International Linked Data
55	transport-data-gov-uk	transport.data.gov.uk
56	twc-ieeevis	IEEE VIS Source Data
57	uriburner	URIBurner
58	webnmasunotraveler	El Viajero's tourism dataset
59	world-bank-linked-data	World Bank Linked Data
60	yovisto	Yovisto - academic video search
61	zaragoza-turismo	Turismo de Zaragoza

Table 8-3 : Use Case #3 – All active endpoints linked to DBpedia

D) Complete list of groups registered with the LODC

#	Group Name
1	archaeology
2	art
3	bibliographic
4	bibliohack
5	bibsoup
6	bio2rdf
7	civil-society
8	country-gr
9	data-explorer-examples
10	datafqs
11	dbpedia-i18n
12	energy-data
13	eu-linked-data
14	funded-research-projects
15	gnoss
16	gnoss-education
17	history
18	latc
19	legislation
20	linguistics
21	linked-building-data
22	linked-education
23	lld
24	lod
25	lod2-eu-project
26	lodcloud

27	nederland
28	nhs
29	oers
30	ontologies
31	open-data-day
32	open-glam
33	openspending
34	planet-data
35	prizms
36	publicdomain
37	religion
38	spain
39	tetherless-world
40	constellation
41	ukdiscovery
42	university-of-oxford
43	visualizing-org
44	wikimedia

Table 8-4 : List of groups in LODNav

E) Results for inspecting the location of 68 datasets

#	ID	Dataset codename	Dataset URL	Correct Location in LODNav
1	1	aemet	http://aemet.linkeddata.es/	TRUE
2	23	bio2rdf-kegg-drug	http://dr.bio2rdf.org	FALSE
3	24	bio2rdf-kegg-enzyme	http://ec.bio2rdf.org	FALSE
4	25	bio2rdf-kegg-glycan	http://gl.bio2rdf.org	TRUE
5	27	bio2rdf-kegg-reaction	http://rn.rkbexplorer.com/	TRUE
6	29	bio2rdf-ncbigene	http://geneid.bio2rdf.org	TRUE
7	41	bluk-bnb	http://bnb.data.bl.uk	TRUE
8	42	brazilian-politicians	http://www.ligadonospoliticos.com.br	TRUE
9	61	data-incubator-metoffice	http://metoffice.dataincubator.org	FALSE
10	62	data-incubator-moseley	http://moseley.dataincubator.org/	FALSE
11	68	data-open-ac-uk	http://data.open.ac.uk/	TRUE
12	72	dbpedia-el	http://wiki.el.dbpedia.org	TRUE
13	73	dbpedia-lite	http://dbpedialite.org/	TRUE
14	77	dbtune-audioscrobbler	http://dbtune.org/last-fm/	TRUE
15	79	dbtune-john-peel-sessions	http://dbtune.org/bbc/peel/	TRUE
16	86	didactalia	http://didactalia.net	TRUE
17	91	educationalprograms_sisvu	http://education.data.gov.uk/	FALSE
18	95	enakting-energy	http://energy.psi.enakting.org/	TRUE
19	100	environmental-applications-reference-thesaurus	http://thesaurus.iia.cnr.it/index.php/vocabularies/earth	TRUE
20	106	eurostat-rdf	http://ec.europa.eu/eurostat/ramon/rdfdata/	TRUE
21	108	event-media	http://eventmedia.eurecom.fr/	TRUE

22	109	fanhubz	http://fanhu.bz	Can't tell
23	121	fu-berlin-diseasome	http://www4.wiwiss.fu-berlin.de/diseasome/	TRUE
24	133	geonames-semantic-web	http://www.geonames.org/ontology/	Can't tell
25	135	geowordnet	http://geowordnet.semanticmatching.org/	FALSE
26	137	gesis-thesoz	http://lod.gesis.org/thesoz/	Can't tell
27	138	gnoss	http://gnoss.com/en/home	TRUE
28	150	jamendo-dbtune	http://dbtune.org/jamendo/	FALSE
29	153	klappstuhlclub	http://klappstuhlclub.de	FALSE
30	155	l3s-dblp	http://dblp.l3s.de/d2r/	TRUE
31	156	lcsh	http://id.loc.gov/authorities/	TRUE
32	168	linked-user-feedback	http://linkedmdb.org/	TRUE
33	170	linkedgeodata	http://km.aifb.kit.edu/projects/numbers/	TRUE
34	171	linkedlccn	http://lov.okfn.org/dataset/lov/index.html	FALSE
35	173	lista-encabezamientos-materia	http://id.sgcb.mcu.es	FALSE
36	175	lobid-resources	http://lobid.org/resource	Can't tell
37	176	loc	http://linkedopencolors.appspot.com/	Can't tell
38	179	los_metar	http://www.linkedopenservices.org/services/geo/SpatialResources/point/ICAO/	FALSE
39	186	my-experiment	http://www.myexperiment.org	TRUE
40	195	oceandrilling-codices	http://data.oceandrilling.org/codices/	FALSE
41	202	open-election-data-project	http://opencorporates.com/	TRUE
42	210	osm-semantic-network	http://wiki.openstreetmap.org/wiki	FALSE

			/OSM_Semantic_Network	
43	218	productontology	http://www.productontology.org	FALSE
44	220	psh-subject-headings	http://psh.ntkcz.cz/skos/home/html/en	Can't tell
45	222	rdf-book-mashup	http://www4.wiwiss.fu-berlin.de/bizer/bookmashup/	TRUE
46	225	rechtspraak	http://www.best-project.nl	Can't tell
47	231	rkb-explorer-acm	http://acm.rkbexplorer.com/	TRUE
48	232	rkb-explorer-budapest	http://budapest.rkbexplorer.com/	TRUE
49	239	rkb-explorer-dblp	http://dblp.rkbexplorer.com	TRUE
50	244	rkb-explorer-eprints	http://eprints.rkbexplorer.com/	TRUE
51	254	rkb-explorer-newcastle	http://newcastle.rkbexplorer.com	TRUE
52	256	rkb-explorer-oai	http://oai.rkbexplorer.com	TRUE
53	257	rkb-explorer-os	http://os.rkbexplorer.com	TRUE
54	266	rkb-explorer-wiki	http://wiki.rkbexplorer.com	TRUE
55	267	rkb-explorer-wordnet	http://wordnet.rkbexplorer.com	TRUE
56	272	secold	http://www.rdfabout.com/demo/sec/	FALSE
57	296	telegraphis	http://telegraphis.net/data/	Can't tell
58	297	temple-ov-thee-lemur-datasets	http://data.totl.net/	FALSE
59	304	twarql	http://wiki.knoesis.org/index.php/Twarql	TRUE
60	305	twc-ieeevis	http://ieeevis.tw.rpi.edu	Can't tell
61	311	umbel	http://umbel.org	Can't tell
62	320	uriburner	http://uriburner.com/	Can't tell
63	321	viaf	http://viaf.org/viaf/data/	TRUE
64	323	vivo-indiana-university	http://vivo.iu.edu	TRUE
65	324	vivo-university-of-florida	http://vivo.ufl.edu	FALSE
66	328	world-bank-linked-data	http://worldbank.270a.info/	TRUE
67	329	world-factbook-fu-berlin	http://www4.wiwiss.fu-	TRUE

			berlin.de/factbook/	
68	332	yovisto	http://www.yovisto.com/ontology/	Can't tell

Table 8-5 : Location Inspection Results