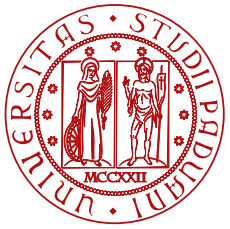

Università degli Studi di Padova

Dipartimento di Ingegneria dell'Informazione

Master Thesis in Computer Engineering

Open Data for Italian Municipalities: Ontology, Data and WebApps

Master Candidate

Luca Martinelli

Student ID 2005837

Supervisor

Prof. Gianmaria Silvello

University of Padova

Academic Year
2021/2022

To my parents
and friends

Abstract

Sommario

Contents

List of Figures

List of Tables

List of Algorithms

List of Code Snippets

List of Acronyms

CSV Comma Separated Values

TSV Tab Separated Values

W3C World Wide Web Consortium

URI Uniform Resource Identifier

URL Uniform Resource Locator

HTTP HyperText Transfer Protocol

RDF Resource Description Framework

LOD Linked Open Data

SPARQL SPARQL Protocol and RDF Query Language

IRI International Resource Identifier

ASCII American Standard Code for Information Interchange

RDFS RDF Schema

OWL Web Ontology Language

XML eXtensible Markup Language

JSON JavaScript Object Notation

RDFa RDF in Attributes

HTML HyperText Markup Language

1 Introduction

1.1 The OntoIM Ontology

1.2 Scope and organization of the thesis

2 Background

2.1 The Web of Data

The World Wide Web was originally designed to be a space where documents are connected by links without semantic value, and most of these documents are designed for humans to read, not for machines to process. For this reason, Tim Berners-Lee in 2001 introduced the idea of the Semantic Web. In particular, the Semantic Web is an enhancement of the current Web that aims to create a web of data, in which information has a well-defined meaning and can be easily read and processed by programs [BHL01].

Due to this machine-comprehensible capacity, the Semantic Web has enormous potential to automate daily tasks in our lives and is helping to advance scientific and health care fields [Fei+07], such as drug discovery and clinical research, but also in the automotive industry, in the enhancement of cultural heritage, etc. . . .[1] In this context, ontologies play a fundamental role in supporting interoperability and common understanding between different web applications and services, solving the problem of semantic heterogeneity [Tay10].

Although there are different definitions of "ontology" [Tay10], in computer science, an ontology is defined as an "explicit and formal specification of a shared conceptualization" [Gru95], where conceptualization means a simplified view of

[1] https://www.w3.org/2001/sw/sweo/public/UseCases/

the world we wish to represent. An ontology is made up of four main types of components, which are (1) *classes* (or *concepts*), which describe concepts in the domain; (2) *instances* of classes, which represent specific objects or elements of a class; (3) *properties* (or *slots*), which are used to express relationships between a first concept in the domain and a second concept in the range; (4) *axioms* (or *role restrictions*), which are used to impose constraints on the values of instances and classes [Tay10; NM+01]. In addition to the interoperability problem, ontologies are also used to satisfy the following needs:

- To share common understanding of the structure of information among people or software agents;
- To enable reuse of domain knowledge;
- To make domain assumptions explicit;
- To separate domain knowledge from the operational knowledge
- To analyze domain knowledge [NM+01].

Along with ontologies, controlled vocabularies, taxonomies, and thesauri are other resources used in different domains, in particular the medical one [IB14]. A controlled vocabulary is a closed list of named subjects, called *terms*, which is usually used for classification. A taxonomy is a subject-based classification that organizes terms in a controlled vocabulary into a hierarchy. Finally, a thesaurus extends taxonomies, allowing making other statements about the subjects and providing a much richer vocabulary [IB14].

In 2006, Tim Berners-Lee used for the first time the term Linked Data to describe the structured and interlinked data that populate the Semantic Web [Ber06]. He also introduced a set of rules, also known as the Linked Data Principles, to provide some best practices for publishing and connecting data on the Web [BHB11]. These principles, published by W3C, are the following:

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

The two main fundamental technologies for Linked Data are Uniform Resource Identifiers (URIs) and HyperText Transfer Protocol (HTTP). In particular, URIs are used to identify any entity that exists in the world, while HTTP provides a simple and universal mechanism for retrieving the resources to which they refer. These two technologies are integrated in RDF, which provides a graph-based data model to structure and link data that describe entities in the world [BHB11]. Using HTTPs, URIs, and RDF, Linked Data builds on the architecture of the Web, called the Web of Data. This means that the Web of Data shares many properties with the traditional Web, which are:

- Web of Data can contain any type of data;
- Anyone can publish data on the Web of Data;
- Publishers are not restricted in the choice of vocabularies used to represent the data;
- Entities are connected by RDF links [BHB11].

However, in addition to those of the traditional Web, the Web of Data also has the following characteristics:

- Data are separated from formatting and presentational aspects;
- Data is self-describing;
- Data access is simplified by the use of the HTTP and RDF standards;
- Web of Data is open, and new data sources can be discovered at run-time by following RDF links [BHB11].

Semantic Web and Linked Data are empowered by technologies developed by the World Wide Web Consortium such as RDF, OWL, serialization formats (Section 2.3), and SPARQL (Section 2.4).[2]

2.2 The five stars of Open Data

Linked Data does not have to be open and can be used internally, such as for personal data. When Linked Data is released under an open license that does not

[2] https://www.w3.org/2001/sw/wiki/Main_Page

impede its reuse for free, such as Creative Commons CC-BY[3] or the Italian Open Data License[4], we can use the term Linked Open Data (LOD) [Ber06]. In 2010 Tim Berners-Lee developed a star rating system to define and classify Linked Open Data, "in order to encourage people, especially government data owners, along the road to good linked data" [Ber06]. The star rating system assigns a star if the information is publicly available under an open license, even if the information is a photo or an image scan of a table. The more stars the information gets, the easier it will be for people (and machines) to use it [Ber06].

★ Available on the Web (any format) but with an open license to be Open Data

★★ Available as machine-readable structured data (e.g., Excel instead of an image scan of a table)

★★★ Available in a non-proprietary format (e.g., CSV instead of Excel)

★★★★ Use URIs to identify things, so that people can point at your stuff

★★★★★ Data are linked to other people's data to provide context

However, as the information receives a greater number of stars, both the benefits for consumers and the costs for the publisher increase. In particular, a five-stars data let consumers discover new data of interest, access to the data schema, reuse parts of the data, and link it to other places. They also do not have to pay for tools in order to read the data (e.g., Excel), and they can download and export the data into other formats and process them. On the other hand, to make these data available, publishers must invest time and resources in slicing and organizing the data, assigning URIs to the data items, thinking about how to represent them, linking the data with other data on the Web and making them discoverable [BK11].

2.3 RDF, OWL, AND SERIALIZATION FORMATS

The Resource Description Framework (RDF) is a W3C graph-based standard model to represent information about resources on the Web (including documents,

[3] https://creativecommons.org/
[4] https://www.dati.gov.it/content/italian-open-data-license-v20

people, physical objects, and abstract concepts). Using RDF, machines can process information on the Web using common parsers and processing tools, and information can be exchanged between different applications without losing meaning [Con+14b]. In particular, in recent years RDF has become the *de-facto* standard for publishing Linked Data on the Web. The core structure of the RDF syntax is a set of statements, called *triples*, because they consist of three elements: a *subject*, a *predicate*, and an *object*, following the structure `<subject> <predicate> <object>`, which can be visually represented in Figure 2.1 [Con+14a].

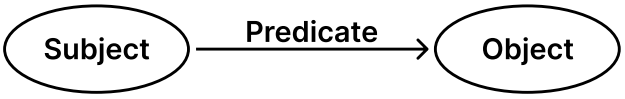

Figure 2.1: The structure of a triple, with two nodes and a predicate connecting them.

The subject and the object represent the two resources being related. The relationship that goes from the subject to the object is called *property*, and its nature is represented by the predicate. A set of statements generate a direct graph, called RDF graph, where subjects and objects are the nodes of the graph, and the predicates form the arcs. For example, the set of triples below produces the graph shown in Figure 2.2 [Con+14b].

```
<Bob> <is a> <person>.
<Bob> <is a friend of> <Alice>.
<Bob> <is born on> <the 4th of July 1990>.
<Bob> <is interested in> <the Mona Lisa>.
<the Mona Lisa> <was created by> <Leonardo da Vinci>.
<the video 'La Joconde à Washington'> <is about> <the Mona Lisa>
```

In an RDF graph, resources may be represented using an International Resource Identifier (IRI), a *literal value* or a *blank node*. An IRI is a generalization of URI, where non-ASCII characters are allowed in the IRI character string. IRIs identify resources, and can appear in all three positions of a triple. In the example above, the IRI for Leonardo Da Vinci in DBpedia[5] is `http://dbpedia.org/resource/Leonardo_da_Vinci`.

[5] `https://www.dbpedia.org/`

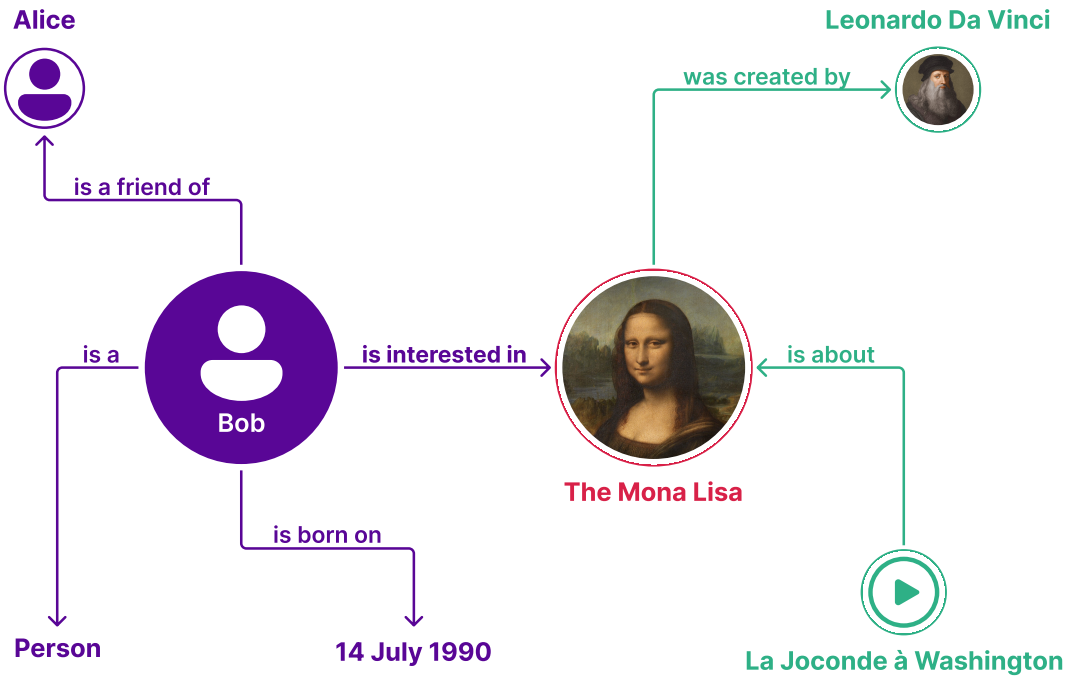

Figure 2.2: The example of the RDF graph presented by W3C.

Literals are basic values such as strings, dates, and numbers. In the RDF graph literals can only be used as objects, and consists of two or three elements, which are: (1) the value itself; (2) an IRI that identifies the *datatype* (string, number, date, etc.); (3) if and only if the datatype is a `rdf:langString`,[6] a *language tag* (such as `en`, `it`, `fr`, etc.) [Con+14a].

Finally, blank nodes can appear in the subject and object position of a triple and are used to represent resources without using a IRI [Con+14b].

In Section 2.1 ontologies and vocabularies are presented as a core element for creating the Semantic Web. The RDF data model does not provide semantic information about the resources. For this reason, RDF provides the RDF Schema (RDFS) language, that allows to define semantic characteristics of data. RDF Schema uses the notion of *class* to classify resources, while uses the *type* property to define a relation between an instance and its class. RDF Schema also allows defining type restrictions on subject and objects of particular triples through *domain* and *range* restrictions. Finally, with RDF Schema it is also possible to define hierarchies of classes and properties, using *subClassOf* and *subPropertyOf*

[6] `http://www.w3.org/1999/02/22-rdf-syntax-ns#langString`

predicates [Con+14b]. All of these modeling constructs provided by RDF Schema are summarized in Table 2.1.

Construct	Syntactic form	Description
Class	C rdf:type rdfs:Class	C is an RDF class
Property	P rdf:type rdf:Property	P is an RDF property
type	I rdf:type C	I is an instance of C
subClassOf	C1 rdfs:subClassOf C2	C1 is a subclass of C2
subPropertyOf	P1 rdfs:subPropertyOf P2	P1 is a sub-property of P2
domain	P rdfs:domain C	domain of P is C
range	P rdfs:range C	range of P is C

Table 2.1: The main modeling constructs provided by RDF Schema.

However, in 2004 the World Wide Web Consortium presented Web Ontology Language (OWL), a more complete language for publishing and sharing ontologies on the Web [Bec+04], and replaced in 2009 and then in 2012 by OWL 2. OWL 2 is a Semantic Web language to represent rich and complex knowledge about things, groups of things, and relations between things. In addition, since OWL is a computational logic-based language, the knowledge expressed in OWL can be reasoned with by computer programs either to verify the consistency of that knowledge or to make implicit knowledge explicit. A OWL document, called *ontology*, can be published in the World Wide Web and may refer to or be referred from other OWL ontologies [Hit+09]. In OWL 2 knowledge is represented by statements, called *axioms*. Axioms normally refer to objects of the world and describe them by putting them into categories or saying something about their relation. In OWL 2 objects, categories and relations are called *entities*, and in particular objects are denoted as *individuals*, categories as *classes* and relations as *properties*. Moreover, properties are further subdivided into (1) *object properties* that relate objects to objects; (2) *datatype properties* that assign data values to objects; (3) *annotation properties* that encode information about the ontology itself. Finally, names of entities can be combined into *expressions* using *constructors* to form complex descriptions from basic ones [Hit+09].

In order to publish RDF data on the Web, the RDF graphs need to be serialized. Today there are several serialization formats, but the most famous one are: N-Triples, Turtle, RDF/XML, RDFa, and JSON-LD. These formats are briefly described below, reporting as example of small excerpt of DBpedia[7] is reported.

N-Triples[8] It's one of the simplest formats, formed by sequences of RDF triples. Each statement is formed by the subject, predicate, object, and a ".", that are separated by white space.

```
<http://dbpedia.org/page/Jotaro_Kujo>
  <http://dbpedia.org/ontology/relative>
    <http://dbpedia.org/page/Joseph_Joestar> .
```

Turtle[9] It's a common data format for serializing RDF graphs that introduces some features to N-Triples language. In particular, it introduces the use of `@base` IRI and relative IRIs, `@prefix` and prefixed names, predicate lists separated by ";", object lists separated by ",", and the representation of `rdfs:type` with the token `a`.

```
@prefix dbr: <http://dbpedia.org/page/> .
@prefix dbo: <http://dbpedia.org/ontology/> .

dbr:Jotaro_Kujo dbo:relative dbr:Joseph_Joestar .
```

RDF/XML[10] Expresses RDF graphs as an XML document. The nodes and predicates are represented in XML terms: element names, attribute names, element contents and attribute values.

```
<rdf:RDF xmlns:dbr="http://dbpedia.org/page/"
  xmlns:dbo="http://dbpedia.org/ontology/"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xml:base="http://www.ldf.fi/service/rdf-serializer/">
  <rdf:Description
    rdf:about="http://dbpedia.org/page/Jotaro_Kujo">
```

[7] https://www.dbpedia.org/
[8] https://www.w3.org/TR/n-triples/
[9] https://www.w3.org/TR/turtle/
[10] https://www.w3.org/TR/rdf-syntax-grammar/

```
    <dbo:relative
      rdf:resource="http://dbpedia.org/page/Joseph_Joestar"/>
  </rdf:Description>
</rdf:RDF>
```

RDF in Attributes (RDFa)[11] Provides a set of markup attributes to HTML pages to augment the visual information on the Web with machine-readable hints.

```
<body
  prefix="dbr: http://dbpedia.org/page/
    dbo: http://dbpedia.org/ontology/">
  <div about="dbr:Jotaro_Kujo">
    <div
      rel="dbo:relative"
      resource="dbr:Joseph_Joestar">
    </div>
  </div>
</body>
```

JSON-LD[12] Serializes RDF graphs into JavaScript Object Notation (JSON). The syntax is designed to easily integrate into deployed systems that already use JSON. It's intended to be a way to use Lined Data in Web-based programming environments, to build interoperable Web services, and to store Linked Data in JSON-based storage engines.

```
[
  {
    "@id": "http://dbpedia.org/page/Joseph_Joestar"
  },
  {
    "@id": "http://dbpedia.org/page/Jotaro_Kujo",
    "http://dbpedia.org/ontology/relative": [
      {
        "@id": "http://dbpedia.org/page/Joseph_Joestar"
```

[11] https://www.w3.org/TR/rdfa-primer/
[12] https://www.w3.org/TR/json-ld/

```
      }
    ]
  }
]
```

2.4 SPARQL

SPARQL Protocol and RDF Query Language (SPARQL) is a query language developed by W3C retrieve and manipulate RDF graph content on the Web or in a RDF store. A SPARQL query contains a set of triple patterns called *basic graph pattern*. These patterns are like RDF triples except that subjects, predicates and objects may be replaced by variables. The basic graph pattern matches a sub-graph of the RDF data and returns a new RDF graph in which the variables are replaced with the matched data. Queries are usually processed by an HTTP service, called *SPARQL endpoint*. [Con+13]. The example below shows a SPARQL query on DBpedia SPARQL endpoint,[13] while Table 2.2 shows its result.

```
PREFIX dbr: <http://dbpedia.org/page/>
PREFIX dbo: <http://dbpedia.org/ontology/>
PREFIX dbp: <http://dbpedia.org/property/name/>

SELECT ?relative ?name WHERE {
    dbr:Jotaro_Kujo dbo:relative ?relative .
    ?relative dbp:name ?name .
}
```

relative	name
<http://dbpedia.org/resource/Dio_Brando>	"Dio Brando"@en
<http://dbpedia.org/resource/Joseph_Joestar>	"Joseph Joestar"@en
<http://dbpedia.org/resource/Jonathan_Joestar>	"Jonathan Joestar"@en

Table 2.2: A query result example from DBpedia.

SPARQL queries supports features like union of patterns, nesting queries, optional patterns or filtering values. Once the RDF sub-graph is computed, it's also possible to modify it by ordering, limiting and grouping the values.

[13] `https://dbpedia.org/sparql`

Another important feature of SPARQL is the possibility to perform federated queries, which explicitly delegates certain sub-queries to different SPARQL endpoints, allowing to navigate through the Web of Data.

Finally, to return a more machine-readable form, SPARQL supports four common exchange formats, which are: eXtensible Markup Language (XML), JSON, Comma Separated Values (CSV), and Tab Separated Values (TSV) [Con+13].

2.5 Protégé

2.6 Virtuoso

2.7 CKAN

2.8 OntoPiA

3

Related works

4

Requirements analysis

5

Description of the OntoIM Ontology

6

Ontology Development and Data Mapper

7 Web Applications

8

Conclusions and Future Works

References

[Gru95] Thomas R Gruber. "Toward principles for the design of ontologies used for knowledge sharing?" In: *International journal of human-computer studies* 43.5-6 (1995), pp. 907–928.

[BHL01] Tim Berners-Lee, James Hendler, and Ora Lassila. "The semantic web". In: *Scientific american* 284.5 (2001), pp. 34–43.

[NM+01] Natalya F Noy, Deborah L McGuinness, et al. *Ontology development 101: A guide to creating your first ontology*. 2001.

[Bec+04] Sean Bechhofer et al. "OWL web ontology language reference". In: *W3C recommendation* 10.2 (2004), pp. 1–53.

[Ber06] Tim Berners-Lee. "Linked Data - Design Issues". In: (July 2006). URL: `http://www.w3.org/DesignIssues/LinkedData.html`.

[Fei+07] Lee Feigenbaum et al. "The semantic web in action". In: *Scientific American* 297.6 (2007), pp. 90–97.

[Hit+09] Pascal Hitzler et al. "OWL 2 web ontology language primer". In: *W3C recommendation* 27.1 (2009), p. 123.

[Tay10] Mohammad Mustafa Taye. "Understanding semantic web and ontologies: Theory and applications". In: *arXiv preprint arXiv:1006.4567* (2010).

[BK11] Florian Bauer and Martin Kaltenböck. "Linked open data: The essentials". In: *Edition mono/monochrom, Vienna* 710 (2011).

[BHB11] Christian Bizer, Tom Heath, and Tim Berners-Lee. "Linked data: The story so far". In: *Semantic services, interoperability and web applications: emerging concepts*. IGI global, 2011, pp. 205–227.

[Con+13] World Wide Web Consortium et al. "SPARQL 1.1 Overview". In: (Mar. 2013). URL: `https://www.w3.org/TR/sparql11-overview/`.

[Con+14a] World Wide Web Consortium et al. *RDF 1.1 Concepts and Abstract Syntax*. Feb. 2014. URL: https://www.w3.org/TR/rdf11-concepts/.

[Con+14b] World Wide Web Consortium et al. "RDF 1.1 Primer". In: (June 2014). URL: https://www.w3.org/TR/rdf11-primer/.

[IB14] Mirjana Ivanović and Zoran Budimac. "An overview of ontologies and data resources in medical domains". In: *Expert Systems with Applications* 41.11 (2014), pp. 5158–5166.

Acknowledgments

No thanks