Chapter 2
Methodological Guidelines for Publishing Government Linked Data

Boris Villazón-Terrazas, Luis. M. Vilches-Blázquez,
Oscar Corcho, and Asunción Gómez-Pérez

Abstract Publishing Government Linked Data (and Linked Data in general) is a process that involves a high number of steps, design decisions and technologies. Although some initial guidelines have been already provided by Linked Data publishers, these are still far from covering all the steps that are necessary (from data source selection to publication) or giving enough details about all these steps, technologies, intermediate products, etc. In this chapter we propose a set of methodological guidelines for the activities involved within this process. These guidelines are the result of our experience in the production of Linked Data in several Governmental contexts. We validate these guidelines with the GeoLinkedData and AEMETLinkedData use cases.

2.1 Introduction

Electronic Government (e-Gov) is an important application field [17] for the transformations that governments are undergoing and will continue to undergo in the following decades. Moreover, currently there is a trend to transform the e-Gov into the e-Governance[1], by means of opening government data to the public.

initiatives across the world are making large amounts of raw governmental data available to the public on the Web. Opening this data to citizens enables transparency, delivers more public services, and encourages greater public and commercial use and re-use of governmental information. Some governments have even

Correspondance author: Boris Villazón-Terrazas, Ontology Engineering Group, Departamento de Inteligencia Artificial, Facultad de Informática, Universidad Politécnica de Madrid, Madrid, España, e-mail: bvillazon@fi.upm.es. See the List of Contributors for full contact details.

[1] e-Governance is the application of Information and Communication Technology (ICT) for delivering government Services, exchange of information communication transactions, integration various stand-one systems and services between Government-to-citizens (G2C), Government-to-Business (G2B), Government-to-Government (G2G) as well as back office processes and interactions within the entire government framework [17].

D. Wood (ed.), *Linking Government Data*, DOI 10.1007/978-1-4614-1767-5_2,

created catalogs or portals, such as the United States[2] and the United Kingdom[3] governments, to make it easy for the public to find and use this data [23], which are available in a range of formats, e.g., spreadsheets, relational database dumps, RDF; and span through a wide range of domains, e.g., geospatial, statistics, transport.

The application of Linked Data principles to government datasets brings enormous potential [7]. However, this potential is currently untapped mostly because of the lack of resources required to transform raw data to high-quality Linked Data on a large scale [4].

Linked Data generation and publication does not follow a set of common and clear guidelines to scale out the generation and publication of Linked Data. Moreover, there is a lack of detailed guidelines and software catalogs to support the whole life cycle of publishing government Linked Data, and most of existing guidelines are intended for software developers, not for governments.

In this chapter we take the first step to formalize our experience gained in the development of government Linked Data, into a preliminary set of methodological guidelines for generating, publishing and exploiting Linked Government Data. This chapter is addressed to developers who pertain to public administrations, but governments may find the guidelines useful because these guidelines are based and have been applied in real case government scenarios. Therefore, the guidelines are very good starting point for local or national public administrations when they want to publish their data as Linked Data. The rest of the chapter is organized as follows: Section 2.2 presents a summary of the initiatives for helping governments to open and share their data. Section 2.3 explains the guidelines for the generation of government Linked Data. Then, Section 2.4 describes the application of these guidelines to particular use cases. Finally, Section 2.5 presents the conclusions and future work.

2.2 Open Government Initiatives

During the last years several initiatives emerged to improve the interface between citizens and government through effective use of Information and Communication Technology (ICT), and specifically through use standards-base of the Web. In this section, we present a summary of those efforts that help governments in the use of technology and the Web to implement the full promise of electronic government, by managing their data in a transparent and efficient way.

- Since 2008 The W3C eGovernment Activity[4] is promoting several charters for helping goverments to follow best practices and approaches to improve the use of the Web. Currently, this activity includes the eGovernment Interest Group[5]

[2] http://www.data.gov/

[3] http://data.gov.uk/

[4] http://www.w3.org/egov/Activity.html

[5] http://www.w3.org/2007/eGov/IG/

and the Government Linked Data Working Group[6]. Some of the results of this activity are described next.

  - *Improving Access to Government through Better Use of the Web*[7], a W3C Interest Group Note that attempts to describe the challenges and issues faced by governments and their efforts to apply technologies of the 21*st* century. Moreover, the document introduces the definition of Open Government Data, describes its benefits and how to achieve Open Government Data. However, the document does not include a detailed set of guidelines.
  - *Publishing Open Government Data*[8], a W3C Working draft that proposes a set of preliminary guidelines to help governments to open and share their data. This document enumerates the following straightforward steps to publish government data (1) publish well-structured data in its raw form, e.g., an XML file; (2) create an online catalog of the raw data; and (3) make the data machine and human readable. This document also introduces the four Linked Data principles, but does not provide detailed guidelines.

- Since 2004 the Open Knowledge Foundation[9], a not-for-profit organization is promoting open knowledge[10]. The Open Knowledge Foundation has released the Open Data Manual[11], which is a report that includes discussions about the legal, social and technical aspects of open data, and its target audience are those who are seeking to open up data. Although the report is focused on data from the public sector, the target audience of the report are not governments.
- Finally, it is worth mentioning the suggestion given by Tim Berners-Lee about the 5-star deployment scheme for Linked Open Data that are described with examples in http://lab.linkeddata.deri.ie/2010/star-scheme-by-example/.

After having reviewed the available efforts to help governments for managing their data in a transparent and efficient way, by means of Open Data initiatives, we can conclude that those efforts are not based in real case government scenarios, neither there is a report of having applied them into real case scenarios.

2.3 Methodological Guidelines

In this section we present our preliminary set of guidelines that are based on our experience in the production of Linked Data in several Governmental contexts. Moreover, the guidelines have been applied in real case government scenarios and include

[6] http://www.w3.org/2011/gld/

[7] http://www.w3.org/TR/2009/NOTE-egov-improving-20090512/

[8] http://www.w3.org/TR/gov-data/

[9] http://okfn.org/

[10] Any kind of data, which can be freely used, reused and redistributed.

[11] http://opendatamanual.org/

methods, techniques and tools for carrying out the activities and tasks involved in the Government Linked Data publishing process.

The process of publishing Government Linked Data must have a life cycle, in the same way of Software Engineering, in which every development project has a life cycle [20]. According to our experience this process has an iterative incremental life cycle model, which is based on the continuous improvement and extension of the Government Linked Data resulted from performing several iterations.

The guidelines, for the process of publication Government Linked Data, consist of the following main activities: (1) specification, (2) modelling, (3) generation, (4) publication, and (5) exploitation. Each activity is decomposed in one or more tasks, and some techniques and tools are provided for carrying out them. It is worth mentioning that the order of the activities and tasks might be changed base on particular needs of the government bodies. Moreover, we are continuously getting feedback about these guidelines, and therefore, we are improving them constantly. Figure 2.1 deptics the main activities that are described next.

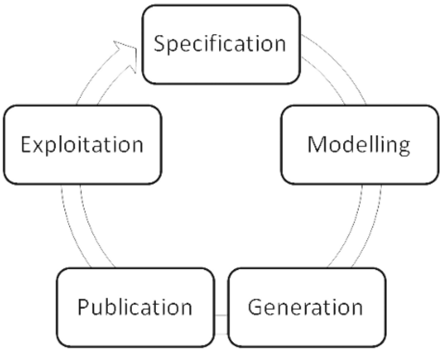

Fig. 2.1 Main Activities for Publishing Government Linked Data

2.3.1 Specification

As any other eGovernment project, aimed at the implementation and further development of e-administration and other IT solutions, the first activity is the drawing up of a detailed specification of requirements. It has been proved that detailed requirements provides several benefits [16], such as (a) the establishment of the basis for agreement between customers and suppliers on what the government application is supposed to do, (b) the reduction of the development effort, (c) the provision of a basis for estimating costs and schedules, and (d) the offer of a baseline for validation and verification.

When a government Linked Data application is being developed, government Linked Data requirements should be identified in addition to the application ones. Our experience in the publication of Linked Data in several Governmental contexts has showed that more critical than capturing software/application requirements was the efficient and precise identification of the government Linked Data requirements.

At this stage, the description of this activity is not intended to be exhaustive but it just introduces the most important points. The preliminary set of tasks identified for this activity are: (1) identification and analysis of the government data sources, (2) URI design, and (3) definition of license.

2.3.1.1 Identification and analysis of the government data sources

Within this task we identify and select the government data that we want to publish. In this task we have to distinguish between (i) open and publish data that government agencies have not yet opened up and published, and (ii) reuse and leverage on data already opened up and published by government agencies. Next, we describe briefly both alternatives.

- In the case of open and publish data that government agencies have not yet opened up and published, we will face, most of the time, a costly and tedious task that may require contacting to specific government data owners to get access to their legacy data.
- In the other case, when we want to reuse and leverage on data already opened up and published by government agencies, we should look for these data in public government catalogs, such as the Open Government Data[12], datacatalogs.org[13], and Open Government Data Catalog[14].

After we have identified and selected the government data sources, we have to (i) search and compile all the available data and documentation about those resources, including purpose, components, data model and implementation details; (ii) identify the schema of those resources including the conceptual components and their relationships; and (iii) identify the items in the domain, i.e., things whose properties and relations are described in the data sources, according to [7] the Web architecture term *resource* is used to refer to these *things of interest*.

2.3.1.2 URI design

The goal of the Linked Data initiative is to promote a vision of the Web as a global database, and interlink data the same way that Web documents. In this global database it is necessary to identify a resource on the Internet, and precisely URIs are thought for that. According to [15] URIs should be designed with simplicity, stability and manageability in mind, thinking about them as identifiers rather than as names for Web resources.

There are some existing guidelines for URI design, for example (1) *Cool URIs for the Semantic Web W3C Interest Group Note* [15], which introduces a useful guidance on how to use URIs to describe things that are not Web documents; (2)

[12] http://opengovernmentdata.org/data/catalogues/

[13] http://datacatalogs.org/

[14] http://datos.fundacionctic.org/sandbox/catalog/

Designing URI Sets for the UK Public Sector[15], a document from the UK Cabinet Office that defines the design considerations on how to URIs can be used to publish public sector reference data; and (3) *Sytle Guidelines for Naming and Labelling Ontologies in the Multilingual Web* [11], which proposes guidelines for designing URIs in a multilingual scenario.

Based on the aforementioned guidelines and on our experience we propose the following design decisions regarding the assignment of URIs to the elements of the dataset.

- Use meaningful URIs, instead of opaque URIs, when possible. Since one of the goals is to publish data to citizens, it is recommended to put into the URI as many information as possible.
- Use slash (303) URIs, instead of hash URIs, when possible. In spite of the fact that there is some criticism of the slash URIs because using them requires two HTTP requests to retrieve a single description of an object, they are appropriate for dealing with resource descriptions that are part of very large datasets [7].
- Separate the TBox (ontology model) from the ABox (instances) URIs. Therefore, we have to manage the following URI elements
  - Base URI structure. Here we need to choose the right domain for URIs, and this domain will expect to be maintained in perpetually, support a direct response to agency servers. Governments can follow the UK cabinet Office guides for choosing the right domain for URIs, for example for the Bolivian Government[16] http://data.gov.bo, and for a particular government sector, in this case health, http://health.data.gov.bo.
  - TBox URIs. We recommend to append the word *ontology* to the base URI structure, following our previous example we would have http://data.gov.bo/ontology/.
    Then, we would append all the ontology elements, classes and properties.
  - ABox URIs. We recommend to append the word *resource* to the base URI structure, again following our previous example we would have http://data.gov.bo/resource/.
    Additionally, we recommend to use *Patterned URIs*[17] by adding the class name to the ABox base URI. For example we want to identify a particular province, we would have http://data.gov.bo/resource/province/Tiraque.
- Use the main official language of the government, when possible. In some cases we will deal with some special characters depending on the language. Following our previous example, within the Bolivian Government we should use Spanish, therefore we would have for identifying the Tiraque Province http://data.gov.bo/resource/Provincia/Tiraque

[15] http://www.cabinetoffice.gov.uk/resource-library/designing-uri-sets-uk-public-sector

[16] The URI examples for the Bolivian Government are fictitious.

[17] http://patterns.dataincubator.org/book/patterned-uris.html

2.3.1.3 Definition of the license

Within the government context it is important to define the license of the data that governments are publishing. Currently, there are several licenses that can be used for government data. Next, we list a few of them.

- The UK Open Government License[18] was created to enable any public sector information holder to make their information available for use and reuse under its terms.
- The Open Database License[19] (ODbL) is an open license for databases and data that includes explicit attribution and share-alike requirements.
- Public Domain Dedication and License[20] (PDDL) is a document intended to allow you to freely share, modify, and use a particular data for any purpose and without any restrictions.
- Open Data Commons Attribution License[21] is a database specific license requiring attribution for databases.
- The Creative Commons Licenses[22] are several copyright licenses that allow the distribution of copyrighted works.

It is also possible to reuse and apply an existing license of the government data sources.

2.3.2 *Modelling*

After the specification activity, in which the government data sources were identified, selected and analysed, we need to determine the ontology to be used for modelling the domain of those data sources. The most important recommendation in this context is to reuse as much as possible available vocabularies[23] [2]. This reuse-based approach speeds up the ontology development, and therefore, governments will save time, effort and resources. This activity consists of the following tasks:

- Search for suitable vocabularies to reuse. Currently there are some useful repositories to find available vocabularies, such as, SchemaWeb[24], SchemaCache[25], Swoogle[26], and LOV[27]. For choosing the most suitable vocabularies we recom-

[18] http://www.nationalarchives.gov.uk/doc/open-government-licence/

[19] http://opendatacommons.org/licenses/odbl/

[20] http://opendatacommons.org/licenses/pddl/

[21] http://opendatacommons.org/licenses/by/

[22] http://creativecommons.org/

[23] Along this chapter we use vocabulary or ontology without distinction.

[24] http://schemaweb.info/

[25] http://schemacache.com/

[26] http://swoogle.umbc.edu/

[27] Linked Open Vocabularies http://labs.mondeca.com/dataset/lov/index.html

mend to follow the guidelines proposed in [19] that detail how to reuse vocabularies at different levels of granularity, i.e., reusing general ontologies, domain ontologies, and ontology statements.
- In case that we did not find any vocabulary that is suitable for our purposes, we should create them, trying to reuse as much as possible existing resources, e.g., government catalogues, vocabularies available at sites like http://semic.eu/, etc. Within this task, we recommend to follow the guidelines proposed in [22] that state how to (1) search government resources from highly reliable Web Sites, domain-related sites and government catalogs; (2) select the most appropriate government resources; and (3) transform them into ontologies.
- Finally, if we did not find available vocabularies nor resources for building the ontology, we have to create the ontology from scratch. To this end, we can follow the first scenario proposed in the NeOn Methodology [18].

There are several tools that provide technological support to this activity and some of them are Neologism[28], Protégé[29], NeOn Toolkit[30], TopBraid Composer[31], and Altova Semantic Works[32].

2.3.3 *Generation*

The Resource Description Framework, RDF[33], is the standard data model in which the government information has to be made available, according to the Linked Data principles. Therefore, in this activity we have to take the data sources selected in the specification activity (see Section 2.3.1), and transform them to RDF according to the vocabulary created in the modelling activity (see Section 7.3). The generation activity consists of the following tasks: (1) transformation, (2) data cleansing, and (3) linking.

2.3.3.1 Transformation

The preliminary guidelines proposed in this chapter consider only the transformation of the whole data source content into RDF, i.e., following an Extract, Transform, and Load ETL-like[34] process, by using a set of RDF-izers, i.e., RDF converters.

[28] http://neologism.deri.ie/

[29] http://protege.stanford.edu/

[30] http://www.neon-toolkit.org

[31] http://www.topquadrant.com/products/TB_Composer.html

[32] http://www.altova.com/semanticworks.html

[33] http://www.w3.org/RDF/

[34] Extract, transform, and load (ETL) of legacy data sources, is a process that involves: (1) extracting data from the outside resources, (2) transforming data to fit operational needs, and (3) loading data into the end target resources [9]

Guidelines for this task are based on the method proposed in [22] that provides guide for transforming the content of a given resource into RDF instances. The requirements of the transformation are (1) full conversion, which implies that all queries that are possible on the original source should also be possible on the RDF version; and (2) the RDF instances generated should reflect the target ontology structure as closely as possible, in other words, the RDF instances must conform to the already available ontology/vocabulary.

There are several tools that provide technological support to this task, and the format of the government data source is relevant for the selection of a particular tool. Next, we provide a list of some those tools[35] grouped by the common formats of government data.

- For CSV and spreadsheets: RDF Extension of Google Refine[36], XLWrap[37], RDF123[38], and NOR$_2$O[39].
- For relational databases: D2R Server[40], ODEMapster[41], Triplify[42], Virtuoso RDF View[43], and Ultrawrap[44]. It is worth mentioning that the RDB2RDF Working Group[45] is working on R2RML[46], a standard language to express mappings between relational databases and RDF.
- For XML: GRDDL[47] through XSLT, TopBraid Composer, and ReDeFer[48].
- For other formats any23[49], and Stats2RDF[50].

2.3.3.2 Data cleansing

The paradigm of generating, publishing and exploiting government linked data (and linked data in general) has inevitably led to several problems. There are a lot of noise which inhibits applications from effectively exploiting the structured information

[35] For a complete list see http://www.w3.org/wiki/ConverterToRdf

[36] http://lab.linkeddata.deri.ie/2010/grefine-rdf-extension/

[37] http://xlwrap.sourceforge.net/

[38] http://rdf123.umbc.edu/

[39] http://www.oeg-upm.net/index.php/en/downloads/57-nor2o

[40] http://sites.wiwiss.fu-berlin.de/suhl/bizer/d2r-server/

[41] http://www.oeg-upm.net/index.php/en/downloads/9-r2o-odemapster

[42] http://triplify.org/

[43] http://virtuoso.openlinksw.com/whitepapers/relational%20rdf%20views%20mapping.html

[44] http://www.cs.utexas.edu/~miranker/studentWeb/UltrawrapHomePage.html

[45] http://www.w3.org/2001/sw/rdb2rdf/

[46] http://www.w3.org/TR/r2rml/

[47] http://www.w3.org/TR/grddl/

[48] http://rhizomik.net/redefer/

[49] http://any23.org/

[50] http://aksw.org/Projects/Stats2RDF

that underlies linked data [8]. This activity focuses on cleaning this noise, e.g., the linked broken data. It consists of two steps

- To identify and find possible mistakes. To this end, Hogan et al. [8] identified a set of common errors
  - http-level issues such as accessibility and derefencability, e.g., HTTP URIs return 40x/50x errors.
  - reasoning issues such as namespace without vocabulary, e.g., `rss:item`; term invented in related namespace, e.g., `foaf:tagLine` invented by LiveJournal; term is misspelt version of term defined in namespace, e.g., `foaf:image` vs. `foaf:img`.
  - malformed/incompatible datatypes, e.g., "true" as `xsd:int`.
- To fix the identified errors. For this purpose Hogan et al. [8] also propose some solutions at the (1) application side, i.e., all the issues have a suitable antidote once we are aware of them; (2) publishing side, by means of all-in-one validation service such as the RDF Alerts[51].

One outstanding initiative within the context of data cleansing is driven by "The Pedantic Web Group"[52] that aims to engage with publishers and help them improve the quality of their data.

2.3.3.3 Linking

Following the fourth Linked Data Principle, *Include links to other URIs, so that they can discover more things*, the next task is to create links between the government dataset and external datasets. This task involves the discovery of relationships between data items. We can create these links manually, which is a time consuming task, or we can rely on automatic or supervised tools. The task consists of the following steps:

- To identify datasets that may be suitable as linking targets. For this purpose we can look for data sets of similar topics on the Linked Data repositories like CKAN[53]. Currently there is no tool support for this, so we have to perform the search in the repositories manually. However, there are approaches to peform this step, such as [13] and [10].
- To discover relationships between data items of government dataset and the items of the identified datasets in the previous step. There are several tools for creating links between data items of different datasets, for example the SILK framework [3], or LIMES [12].
- To validate the relationships that have been discovered in the previous step. This usually is performed by government domain experts. In this step we can use

[51] http://swse.deri.org/RDFAlerts/

[52] http://pedantic-web.org/

[53] http://ckan.net/group/lodcloud

tools like *sameAs Link Validator*[54] that aims to provide a user friendly interface for validating *sameAs* links.

2.3.4 Publication

In this section we review the publication of RDF data. In a nutshell, this activity consists in the following task (1) dataset publication, (2) metadata publication, and (3) enable effective discovery. These activities are described next.

2.3.4.1 Dataset publication

Once we have the legacy data transformed into RDF, we need to store and publish that data in a triplestore[55]. There are several tools for storing RDF datasets, for example Virtuoso Universal Server[56], Jena[57], Sesame[58], 4Store[59], YARS[60], and OWLIM[61]. Some of them already include a SPARQL endpoint and Linked Data frontend. However, there are some tools like Pubby[62], Joseki[63], and Talis Platform[64] that provide these functionalities. A good overview of the recipes for publishing RDF data can be found in [7].

2.3.4.2 Metadata Publication

Once our dataset is published we have to include metadata information about it. For this purpose there are vocabularies like (1) VoID[65] that allows to express metadata about RDF datasets, and it covers general metadata, access metadata, structural metadata, and description of links between datasets; and (2) Open Provenance Model[66] that is a domain independent provenance model result of the Provenance

[54] http://oegdev.dia.fi.upm.es:8080/sameAs/

[55] A triplestore is a purpose-built database for the storage and retrieval of RDF.

[56] http://virtuoso.openlinksw.com/

[57] http://jena.sourceforge.net/

[58] http://www.openrdf.org/

[59] http://4store.org/

[60] http://sw.deri.org/2004/06/yars/

[61] http://www.ontotext.com/owlim

[62] http://www4.wiwiss.fu-berlin.de/pubby/

[63] http://www.joseki.org/

[64] http://www.talis.com/platform/

[65] http://www.w3.org/TR/void/

[66] http://openprovenance.org/

Challenge Series[67]. The provenance of the government datasets plays an important role when browsing and exploring government resources.

2.3.4.3 Enable effective discovery

The last task of the publication activity is the one related to enable the effective discovery and synchronization of the government dataset. This task consists in the following steps

- In this step we deal with Sitemaps[68] that are the standard way to let crawlers know about the pages on a website. When sitemaps provide time indications using *lastmod*, *changefreq* and *priority* fields, they can be used to have (semantic) web search engines download only new data and changed pages. This step aims at allowing (semantic) web search engines to discover what is new or recently changed in the government dataset in an efficient and timely manner. In this step is necessary (1) to generate a set of `sitemap.xml` files from the government SPARQL endpoint, and (2) to submit the `sitemap.xml` files into (semantic) web search engines, such as Google[69] and Sindice[70]. In this step we can rely on automatic tools like sitemap4rdf[71].
- The second step aims to include the government dataset in the LOD cloud diagram[72]. To this end, we have to add an entry of dataset in the CKAN repository[73]. The Linking Open Data Task Force provides some guidelines for collecting metadata on linked datasets in CKAN at their site[74].
- The goal of the final step is to include the dataset in the available open data government catalogues, such as datacatalogs.org[75], and Open Government Data Catalog[76].

2.3.5 Exploitation

The final goal of opening government data (legacy data, streaming data, and services), is to enable transparency, deliver more public applications, and encourage

[67] http://twiki.ipaw.info/bin/view/Challenge/OPM

[68] http://www.sitemaps.org/

[69] https://www.google.com/webmasters/tools/

[70] http://sindice.com/main/submit

[71] http://lab.linkeddata.deri.ie/2010/sitemap4rdf/

[72] http://richard.cyganiak.de/2007/10/lod/

[73] http://ckan.net/group/lodcloud

[74] http://www.w3.org/wiki/TaskForces/CommunityProjects/LinkingOpenData/DataSets/CKANmetainformation

[75] http://datacatalogs.org/

[76] http://datos.fundacionctic.org/sandbox/catalog/

public and commercial use and re-use of the governmental information. Therefore, we have to develop applications on top of the Linked Open Government Data that exploit these data and provide rich graphical user interfaces to the citizens.

According to [7] we can categorize the Linked Data applications in generic applications and domain-specific applications. Regarding generic applications, we can have (i) Linked Data Browsers, e.g., Disco[77], Tabulator browser[78], LinkSailor[79], and LOD Browser Switch[80]; (ii) Linked Data Search Engines, e.g., Sig.ma[81], and VisiNav[82]. As for domain-specific applications, we have US Global Foreing Aid[83] that combines and visualizes data from different branches of the US Government, Talis Aspire[84] that helps educators to create and manage lists of learning resouces, and DBPedia Mobile[85] that helps tourists to explore a city.

It is worth mentioning that Linked Data applications have to integrate data from different provides (governmental and non-governmental) in a more comprehensive view. In section 2.4 we provide examples of specific applications that exploit the government linked data by providing rich graphical user interface to the final users.

2.4 Use Cases

In order to validate the understandability, applicability and usability of the guidelines proposed in this chapter, we conducted two experiments in real case scenarios within GeoLinkedData and AEMETLinkedData.

2.4.1 GeoLinkedData

GeoLinkedData[86] is an open initiative whose aim is to enrich the Web of Data with Spanish geospatial data into the context of INSPIRE themes[87]. This initiative has started off by publishing diverse information sources belonging to the National Geographic Institute of Spain, onwards IGN, and the National Statistic Institute in Spain, onwards INE. Such sources are made available as RDF knowledge bases ac-

[77] http://www4.wiwiss.fu-berlin.de/bizer/ng4j/disco/

[78] http://www.w3.org/2005/ajar/tab

[79] http://linksailor.com/

[80] http://browse.semanticweb.org/

[81] http://sig.ma/

[82] http://visinav.deri.org/

[83] http://data-gov.tw.rpi.edu/demo/USForeignAid/demo-1554.html

[84] http://www.talisaspire.com/

[85] http://wiki.dbpedia.org/DBpediaMobile

[86] http://geo.linkeddata.es/

[87] The INSPIRE Directive addresses 34 spatial data themes needed for environmental applications. http://inspire.jrc.ec.europa.eu/index.cfm/pageid/2/list/7

cording to the Linked Data principles, and following the guidelines proposed in this chapter.

2.4.1.1 Specification

This section presents the specification of GeoLinkedData according to our guidelines. This specification is not intended to be exahustive, but it just describes the most important points.

Identification and analysis of the government data sources

Regarding the government data sources we followed two paths

- We reused and leveraged on data already opened up and published by INE, at its open catalog[88].
- We opened and published data that IGN had not yet opened up and published.

Table 2.1 depicts the datasets that we have chosen for their publication, together with the format in which they are available. All the datasets correspond to Spain, so their content is available in Spanish or in any of the other official languages in Spain (Basque, Catalan and Galician).

Table 2.1 Government Datasets

Data	Provenance	Format
Population	INE	Spreadsheet
Dwelling	INE	Spreadsheet
Industry	INE	Spreadsheet
Building Trade	INE	Spreadsheet
Hydrography	IGN	Relational database (Oracle)
Beaches	IGN	Relational database (MySQL)
Administrative boundaries	IGN	Relational database (MySQL)

URI design

Following the guidelines introduced in section 2.3.1.2, within GeoLinkedData we are using meaningful URIs, and slash (303) URIs. Moreover, we manage the following URI elements

- Base URI structure. For the Spanish Linked Data initiatives we have bought the domain `http://linkeddata.es/`, and specifically, for the Spanish geospatial information we have created the subdomain `http://geo.linkeddata.es/`.

[88] `http://www.ine.es/inebmenu/indice.htm`

- TBox URIs. We appended the word *ontology* to the base URI structure for including concepts and properties available in our ontologies
  http://geo.linkeddata.es/ontology/{conceptorproperty}.
- ABox URIs. We appended the word *resource* to the base URI structure for including the available instances. In addition we include the type of resource in the URI, e.g.,
  http://geo.linkeddata.es/resource/{resourcetype}/{resourcename}

Definition of the license

In the case of GeoLinkedData, we are reusing the original license of the government data sources. IGN and INE data sources have their own license, similar to Attribution-ShareAlike 2.5 Generic License[89].

2.4.1.2 Modelling

In the case of GeoLinkedData our chosen datasets contain information such as time, administrative boundaries, unemployment, etc. For modelling of the information contained in the datasets we have created an ontology network [5]. The vocabulary that models the information contained in the datasets has been developed by reusing available vocabularies/ontologies. Next, we describe briefly each one the subvocabularies that compose the resultant vocabulary/ontology.

For describing statistics, we chose the **Statistical Core Vocabulary (SCOVO)** [6], which provides a modelling framework for statistical information. This vocabulary[90] is currently defined in RDF(S) and terms and labels are provided in English. However, we are going to change it for **RDF Data Cube Vocabulary**[91] that is an extension and improved vocabulary for modelling statistical information.

Regarding geospatial vocabulary we chose diverse ontologies.

- The **FAO Geopolitical Ontology**[92]. This OWL ontology includes information about continents, countries, and so on, in English. We have extended it to cover the main characteristics of the Spanish administrative division.
- Regarding the hydrographical phenomena (rivers, lakes, etc.) we chose **hydrOntology** [21], an OWL ontology that attempts to cover most of the concepts of the hydrographical domain. Its main goal is to harmonize heterogeneous information sources coming from several cartographic agencies and other international resources.

[89] http://creativecommons.org/licenses/by-sa/2.5/

[90] http://purl.org/NET/scovo

[91] http://publishing-statistical-data.googlecode.com/svn/trunk/specs/src/main/html/cube.html

[92] http://www.fao.org/countryprofiles/geoinfo.asp?lang=en

- With respect to geometrical representation and positioning we reuse the **GML ontology**[93] (an OWL ontology for the representation of information structured according to the OGC Geography Markup Language - GML3.0-) and the **WSG84 Vocabulary**[94] (a basic RDF vocabulary, published by the W3C Semantic Web Interest Group, that provides a namespace for representing lat(itude), long(itude) and other information about spatially-located things, using WGS84 as a reference datum).
- Regarding the time information we chose the **Time Ontology**[95], an ontology for temporal concepts developed into the context of World Wide Web Consortium (W3C). This ontology provides a vocabulary for expressing facts about topological relations among instants and intervals, together with information about durations, and about date-time information.

2.4.1.3 Generation

As described in section 2.3.3, RDF is the standard data model in which the government information has to be made available. Therefore, the following tasks describe the transformation of the INE and IGN data sources to RDF following the model we have developed.

Transformation

Given the different formats in which the selected datasets were available, we used three different RDF-izers for the conversion of data into RDF. Next we describe some details of them.

The generation of RDF from spreadsheets was performed using the NOR_2O [22] software library. This library performs an (ETL) process of the legacy data sources, transforming these non-ontological resources (NORs) [22] into ontology instances.

The transformation of the relational database content into RDF was done using the integrated framework R_2O+ and ODEMapster+ [1], which is available as a NeOn Toolkit plugin[96]. This framework allows the formal specification, evaluation, verification and exploitation of semantic mappings between ontologies and relational databases.

For transforming the geospatial information we have used geometry2rdf[97] that converts geometrical data, which could be availabe in GML[98] or WKT[99], into RDF.

[93] http://loki.cae.drexel.edu/~wbs/ontology/2004/09/ogc-gml.owl

[94] http://www.w3.org/2003/01/geo/wgs84_pos

[95] http://www.w3.org/TR/owl-time/

[96] http://www.neon-toolkit.org

[97] http://www.oeg-upm.net/index.php/en/downloads/151-geometry2rdf

[98] http://www.opengeospatial.org/standards/gml

[99] http://iwkt.com/

Data cleansing

In the context of GeoLinkedData we have identified and fixed errors like unescaped characters and encoding problems. Those errors produced by the fact we were generating and publishing linked data in Spanish, therefore we had URIs that contain special characters such as ***á***,***é***, ***ñ***, and it was necessary to encoded those URIs, for example

For the province of *Málaga*

http://geo.linkeddata.es/resource/Provincia/M%C3%A1laga

For the *Miñor* river

http://geo.linkeddata.es/resource/R%C3%ADo/
Mi%C3%B1or%2C%20R%C3%ADo

Linking

In the context of GeoLinkedData, we have identified as initial data sets to link with DBpedia[100] and Geonames[101], because these data sets include similar topics of GeoLinkedData.

The task of discovering relationships between data items was based on the SILK framework. First, we have used SILK to discover relationships between RDF published of Spanish provinces, DBpedia[102] and GeoNames[103] data sources. This process allows setting (*owl:sameAs*) relationships between data of these sources. Next, we present an example of these relationships:

```
<http://geo.linkeddata.es/resource/Provincia/Granada>
<owl:sameAs>
<http://dbpedia.org/resource/Province_of_Granada>
```

The task of result validation was performed by domain experts. This task shows a value of accuracy equal to 86%.

2.4.1.4 Publication

In GeoLinkedData, for the publication of the RDF data we relied on Virtuoso Universal Server[104]. On top of it, Pubby[105] was used for the visualization and navigation of the raw RDF data.

[100] http://dbpedia.org/

[101] http://www.geonames.org/

[102] http://dbpedia.org/About

[103] http://www.geonames.org/

[104] http://virtuoso.openlinksw.com/

[105] http://www4.wiwiss.fu-berlin.de/pubby/

We used VoID for describing the government dataset, and we already created an entry in CKAN for this dataset[106]. Finally, we submitted the sitemap files, generated by sitemap4rdf, to Sindice.

2.4.1.5 Exploitation

As described in section 2.3.5 we need to develop applications that unlock the value of data to provide benefits to citizens. To this end, we have developed an application, map4rdf[107], to enhance the visualization of the aggregated information. This interface combines the faceted browsing paradigm [14] with map-based visualization using the Google Maps API[108]. Thus for instance, the application is able to render on the map distinct geometrical representations such as *LineStrings* that depict to hydrographical features (reservoirs, beaches, rivers, etc.), or *Points* that show province capitals.

2.4.2 AEMETLinkedData

AEMETLinkedData[109] is an open initiative whose aim is to enrich the Web of Data with Spanish metereological data. Within this initiative we are publishing information resources from the *Agencia Estatal de Meteorlogía* (Spanish Metereological Office), ownwards AEMET, as Linked Data.

2.4.2.1 Specification

Here we present the specification of AEMETLinkedData according to our guidelines. This specification just describes the most important points.

Identification and analysis of the government data sources

Regarding the government data sources we reused and leveraged on data already opened up and published by AEMET. Recently, AEMET made publicly available meteorological and climatic data registered by its weather stations, radars, lightning detectors and ozone soundings. AEMET has around 250 automatic weather stations registering pressure, temperature, humidity, precipitation and wind data every 10 minutes. These data from the different stations are provided in CSV files, updated every hour and kept for seven days in the AEMET FTP server, linked from its website.

[106] http://ckan.net/package/geolinkeddata

[107] http://oegdev.dia.fi.upm.es/projects/map4rdf/

[108] http://code.google.com/apis/maps/index.html

[109] http://aemet.linkeddata.es/

URI design

Following the guidelines introduced in section 2.3.1.2, within AEMETLinkedData we are using meaningful URIs, and slash (303) URIs, and managing the following URI elements

- Base URI structure. For the Spanish metereological information we have created the subdomain http://aemet.linkeddata.es/.
- TBox URIs. We appended the word *ontology* to the base URI structure for including concepts and properties available in our ontologies http://aemet.linkeddata.es/ontology/{conceptorproperty}.
- ABox URIs. We appended the word *resource* to the base URI structure for including the available instances we have. In addition we include the type of resource in the URI, e.g., http://aemet.linkeddata.es/resource/{resourcetype}/{resourcename}

Definition of the license

In the case of AEMETLinkedData, we are reusing the original license of the government data sources. AEMET data sources have their own license, an Spanish copyright license. However, this government agency is changing the publication policy and therefore their data sources will adopt a new license in the near future.

2.4.2.2 Modelling

In the case of AEMETLinkedData our chosen datasets contain information related to the metereology domain. For modelling that domain we have developed a network of ontologies [5], by reusing available ontologies and non-ontological resources. Next, we present a high level overview each one of the vocabularies that compose the resultant ontology.

- **Observations ontology**. This vocabulary models the knowledge related to meteorological observations. For its development the NOR$_2$O[110] tool was used to transform non-ontological resources provided by AEMET to ontological resources, i.e., ontology of measurements.
- **Location ontology**. The vocabulary models the knowledge about locations, such as administrative limits and coordinates. The WGS84 vocabulary has been reused with the aim of supporting the representation of geospatial positioning by means of the *Point* concept.

[110] http://www.oeg-upm.net/index.php/en/downloads/57-nor2o

- **Time ontology**. The ontology is for representing knowledge about time such as temporal entities, units, instants, intervals, etc. This ontology was mainly developed by reusing the OWL Time ontology[111].
- **Sensors ontology**. The vocabulary models sensors networks and weather stations. For this ontology we have been reused the Semantic Sensor Network Ontology (SSN)[112].

2.4.2.3 Generation

As described in section 2.3.3, RDF is the standard data model in which the government information has to be made available. Therefore, the following sections describe the transformation of the AEMET data sources to RDF following the model we have developed.

Transformation

The RDF was generated with ad-hoc Python scripts that were executed in two steps, integrating with ease the generation of RDF and the crawling of the FTP server where the CSV files are located. Next, we describe briefly the two steps.

- The first step generates the RDF data about the automatic stations. Since this information is static, only needs to be executed once.
- The second step generates the RDF data about the observations. The observations are obtained by crawling the AEMET FTP server. Whenever new files are added or old files are modified, the script downloads and processes the files.

Data cleansing

In AEMETLinkedData we are finishing the first iteration of the process, and so far we have not yet deeply analyzed the RDF generated. We are planning to do it in the next iteration of the process.

Linking

Within AEMETLinkedData we have identified as initial dataset to link with GeoLinkedData, since we are working with Spanish metereological data.

The task of discovering relationships between data items was based on the SILK framework. First, we have used SILK to discover relationships between AEMET weather stations and their locations in GeoLinkedData resources. This process allows setting (*geo:isLocatedIn*) relationships between data of these sources. Next, we present an example of these relationships:

[111] http://www.w3.org/TR/owl-time/

[112] http://www.w3.org/2005/Incubator/ssn/ssnx/ssn

```
<http://aemet.linkeddata.es/resource/Estacion/Estacion_08430>
<geo:isLocatedIn>
<http://geo.linkeddata.es/resource/Provincia/Murcia>
```

The task of result validation was performed by one person from AEMET. This task shows a value of accuracy equal to 80%.

2.4.2.4 Publication

In AEMETLinkedData, for the publication of the RDF data we relied on Virtuoso Universal Server[113]. On top of it, Pubby[114] was used for the visualization and navigation of the raw RDF data.

We used VoID for describing the government dataset, and we already created an entry in CKAN for this dataset[115]. Finally, we submitted the sitemap files, generated by sitemap4rdf, to Sindice.

2.4.2.5 Exploitation

As described in section 2.3.5 applications have to be developed to unlock the value of data to provide benefits to citizens. Within AEMETLinkedData we have enhanced the visualization capabilities of map4rdf, by including a chart that displays the evolution of a given variable, e.g., temperature. Figure 2.2 shows an example of this visualization.

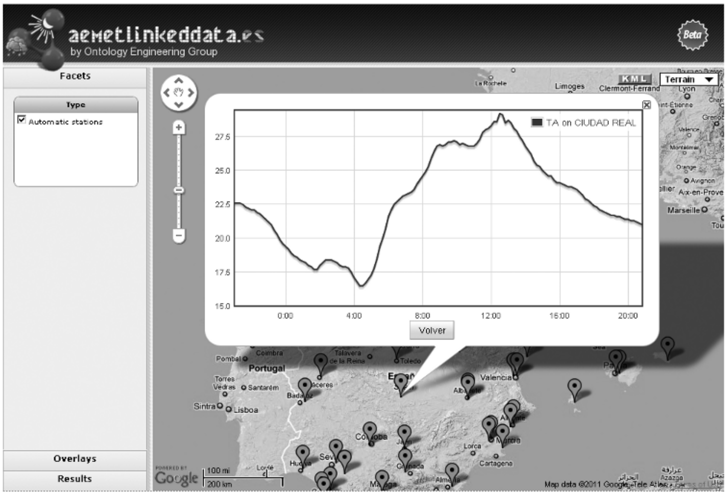

Fig. 2.2 Overview of the Metereological Linked Data Application

[113] http://virtuoso.openlinksw.com/

[114] http://www4.wiwiss.fu-berlin.de/pubby/

[115] http://ckan.net/package/aemet

2.5 Conclusions

In this chapter we have introduced a set of preliminary guidelines for generating, publishing and exploiting Linked Government Data. These guidelines are based on our experience in the production of Linked Data in several Governmental contexts.

According to our experience this process has an iterative incremental life cycle model, which is based on the continuous improvement and extension of the Government Linked Data resulted from performing several iterations. It is worth mentioning that the order of the activities and tasks, involved in this process, might be changed base on particular needs of the government bodies. Moreover, in order to validate the understandability, applicability and usability of the guidelines proposed in this chapter, we have presented two experiments in real case scenarios within GeoLinkedData and AEMET.

As future work, we will (1) continue formalizing the experiences we are gained in the different government contexts we are working; (2) develop more applications for the exploitation of the Government Linked Data; (3) include a validation activity, in which government agencies will validate the results according to the requirements identified in the specification activity; (4) perform more experiments to validate and refine our guidelines.

2.6 Acknowledgments

This work has been supported by the BabelData (TIN2010-17550) and myBigData (TIN2010-17060) Spanish projects. We would like to kindly thank the chapter reviewers for the very good comments. Finally, we would like to thank all OEG members involved in the Linked Data initiatives.

References

1. Barrasa, J., Corcho, O., and Gómez-Pérez, A. R2O, an Extensible and Semantically Based Database-to-Ontology Mapping Language. In *Second Workshop on Semantic Web and Databases (SWDB2004)*, 2004.
2. Bizer, C., Cyganiak, R. and Heath, T. How to publish Linked Data on the Web. Web page, 2007. Revised 2008. Accessed 01/01/2011.
3. Volz, J., Bizer, C., Gaedke, M., Kobilarov, G.: Discovering and Maintaining Links on the Web of Data. In: Bernstein, A., Karger, D. (eds.) The Semantic Web - ISWC 2009, pp. 731-746. Springer, Heidelberg (2009)
4. Cyganiak, R., Maali, F. and Peristeras, V. Self-service linked government data with dcat and gridworks. In *Proceedings of the 6th International Conference on Semantic Systems*, I-SEMANTICS '10, pages 37:1–37:3, New York, NY, USA, 2010. ACM.
5. Haase, P., Rudolph, S., Wang, Y., Brockmans, S., Palma, R., Euzenat, J. and d'Aquin, M. Networked Ontology Model. Technical report, NeOn project deliverable D1.1.1, 2006.
6. Hausenblas, M., Halb, W., Raimond, Y., Feigenbaum, L. and Ayers, D. SCOVO: Using Statistics on the Web of Data. In *ESWC*, volume 5554 of *LNCS*, pages 708–722. Springer, 2009.

7. Heath, T. and Bizer, C. *Linked Data: Evolving the Web into a Global Data Space*, volume 1. Morgan & Claypool, 2011.
8. Hogan, A., Harth, A., Passant, A., Decker, S. and Polleres, A. Weaving the Pedantic Web. In *Linked Data on the Web Workshop (LDOW2010) at WWW'2010*, 2010.
9. Kimball, R. and Caserta, J. *The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleanin*. John Wiley & Sons, 2004.
10. Maali, F. and Cyganiak, R. Re-using Cool URIs : Entity Reconciliation Against LOD Hubs. *Library*, 2011.
11. Montiel-Ponsoda, E., Vila-Suero, D., Villazón-Terrazas, B., Dunsire, G., Rodríguez, E.E. and Gómez-Pérezi, A. Style Guidelines for Naming and Labeling Ontologies in the Multilingual Web. In *Proceedings of the 2011 International Conference on Dublin Core and Metadata Applications*, DCMI '11. Dublin Core Metadata Initiative, 2011.
12. Ngonga Ngomo, A.-C. and Auer, S. Limes - a time-efficient approach for large-scale link discovery on the web of data, 2011.
13. Nikolov, A. and dAquin, M. Identifying Relevant Sources for Data Linking using a Semantic Web Index. *Search*, 2011.
14. Oren, E., Delbru, R. and Decker, S. Extending faceted navigation for RDF data. In *ISWC*, pages 559–572, 2006.
15. Sauermann, L., Cyganiak, R., Ayers, D. and Volkel, M. Cool URIs for the semantic web. Interest Group Note 20080331, W3C. Web page, 2008.
16. I. E. E. E. Computer Society, Sponsored B. The, and Software Engineering Standards Committee. IEEE Recommended Practice for Software Requirements Specifications IEEE Std 830-1998. Technical report, 1998.
17. Sommer, G.G. *The World of E-Government*. Haworth Press, January 2005.
18. Suárez-Figueroa, M.C. *NeOn Methodology for Building Ontology Networks: Specification, Scheduling and Reuse*. PhD thesis, Facultad de Informática, Universidad Politécnica de Madrid, Madrid, Spain, 2010.
19. Suarez-Figueroa, M.C, and Gómez-Pérez, A. NeOn Methodology for Building Ontology Networks: a Scenario-based Methodology. In *(S3T 2009)*, 2009.
20. Taylor, J. *Project Scheduling and Cost Control: Planning, Monitoring and Controlling the Baseline*, volume 1. J. Ross Publishing, 2008.
21. Vilches-Blázquez, L.M., Gargantilla, J.A.R., López-Pellicer, F.J., Corcho, O., and Nogueras-Iso, J. An Approach to Comparing Different Ontologies in the Context of Hydrographical Information. In *IF&GIS*, pages 193–207, 2009.
22. Villazón-Terrazas, B., Suárez-Figueroa, M.C. and Gómez-Pérez, A. A Pattern-Based Method for Re-Engineering Non-Ontological Resources into Ontologies. *International Journal on Semantic Web and Information Systems*, 6(4):27–63, 2010.
23. W3C. Publishing Open Government Data. W3C Working Draft. Web page, 2009.