Guidelines for Multilingual Linked Data

Asunción Gómez-Pérez
Universidad Politécnica de Madrid
Facultad de Informática
Departamento de Inteligencia Artificial
asun@fi.upm.es

Daniel Vila-Suero
Universidad Politécnica de Madrid
Facultad de Informática
Departamento de Inteligencia Artificial
dvila@fi.upm.es

Elena Montiel-Ponsoda
Universidad Politécnica de Madrid
Facultad de Informática
Departamento de Inteligencia Artificial
emontiel@fi.upm.es

Jorge Gracia
Universidad Politécnica de Madrid
Facultad de Informática
Departamento de Inteligencia Artificial
jgracia@fi.upm.es

Guadalupe Aguado-de-Cea
Universidad Politécnica de Madrid
Facultad de Informática
Departamento de Inteligencia Artificial
lupe@fi.upm.es

ABSTRACT

In this article, we argue that there is a growing number of linked datasets in different natural languages, and that there is a need for guidelines and mechanisms to ensure the quality and organic growth of this emerging multilingual data network. However, we have little knowledge regarding the actual state of this data network, its current practices, and the open challenges that it poses. Questions regarding the distribution of natural languages, the links that are established across data in different languages, or how linguistic features are represented, remain mostly unanswered. Addressing these and other language-related issues can help to identify existing problems, propose new mechanisms and guidelines or adapt the ones in use for publishing linked data including language-related features, and, ultimately, provide metrics to evaluate quality aspects. In this article we review, discuss, and extend current guidelines for publishing linked data by focusing on those methods, techniques and tools that can help RDF publishers to cope with language barriers. Whenever possible, we will illustrate and discuss each of these guidelines, methods, and tools on the basis of practical examples that we have encountered in the publication of the *datos.bne.es* dataset.

Categories and Subject Descriptors
Artificial Intelligence,

General Terms
Algorithms, Languages, Theory,

Keywords
Semantic Web, Linked Data, Multilingual

INTRODUCTION

The Linked Data paradigm [5, 6] has boosted the opportunities for applications to exploit the Web of Data at its full potential. This has been accompanied by a significant improvement in the methodological [17, 32] and technological support [4] involved in the publication and consumption of linked data. Currently, we count on a global data space that contains hundreds of Linked Data (LD) sets[1], as well as techniques, and technologies for the various activities involved in the publication of datasets as linked data on the Web. A myriad of technologies is available for (1) producing RDF (Resource Description Framework) datasets out of different types of data sources (e.g., relational databases [10, 29], spreadsheets [22], etc.), (2) discovering links between RDF datasets [14], or (3) publishing metadata describing linked datasets [1, 23], just to mention some examples. As both, process and technological support have gained maturity, new challenges have arisen that need to be faced. One of these major and exciting challenges is the idea of multilingualism, as a ubiquitous aspect of the Web, in the scope of the Web of Data.

With the increasing amount and heterogeneity of data being published on the so-called Web of Data, several empirical studies have been conducted in recent years on the data itself. However, most of these works do not take into account the multilingual dimension at all [11, 9] or if they do, it has been to a limited extent [13]. As identified in [16], the growing Web of Data offers an excellent opportunity to build a multilingual data network where users gain access to information independent of their native languages and the languages in which data have been published. As such, the Web of Data bears the potential "to create a level playing field for users with different cultural backgrounds, native languages and originating from different geo-political environments" [7], because of its language-independent nature. If we provide the mechanisms and the infrastructure necessary to manage multilingualism, we will be unlocking the potential of the Web of Data to the limit. Thus, some steps in the linked data generation process need to be reviewed and considered under the language perspective, for the process to be a seamless and efficient one.

Over the last years, several methodological guidelines for publishing linked data have been proposed [30, 17] and successfully applied to several domains of knowledge such as, mass media [20], geography [2, 28], or cultural heritage [18, 27]. These guidelines provide a principled way for publishing and consuming Linked Data through a series of clearly defined activities whose objective is to produce high quality best practice compliant linked data. However, existing guidelines have overlooked the language dimension and do not provide sufficient guidance and mechanisms for publishers that want to publish linked data in one or several natural languages. In this paper, we will provide some insights about language-related features to be taken into account during the linked data publication process. To

[1] See for example http://lod-cloud.net, http://datahub.io, or http://datacatalogs.org

illustrate the problem, we will use a real use case: *datos.bne.es*[2] [27], the publication of Linked Data out of the catalogue from the National Library of Spain (BNE, Biblioteca Nacional de España).

The purpose of *datos.bne.es* is two-fold: (1) facilitate the reuse of this valuable resource to other regional libraries and the general public, and (2) enrich the catalogue by linking it with data from other international institutions like VIAF[3] (Virtual International Authority File), or the German National Library[4] (DNB, Deutsche Nationalbibliothek). The BNE catalogue contains metadata in several languages describing persons, organizations, topics, and other library materials. For instance, the title of a work is generally registered in the original language, while titles and descriptions of the translations of that work are recorded in the translated form. Similarly, personal and organization names are registered in several languages to facilitate the task of cataloguing and retrieval of the different versions of a work (e.g., a Greek translation of "The old man and the sea" will likely use the translated form of Ernest Hemingway name, i.e., "Χέμινγουεϊ, Έρνεστ"). Additionally, standards for producing catalogue metadata (like ISBD[5], the International Standard for Bibliographic Description) are available in Spanish and BNE professionals use the Spanish version. Finally, linking to datasets like VIAF or the DNB means dealing with language heterogeneity as the data can be in German, French, etc.

In other words, *datos.bne.es* exemplifies three major issues related to language related features in the publication of linked data:

- Data sources may contain information in several natural languages: multilingual data like authors' names or monolingual data like works' titles.
- Vocabularies for describing the data may be also in several languages (multilingual vocabularies), or only in one language (monolingual vocabularies) that can be different from the language required by the publisher.
- Target datasets for linking and enriching the original data sources can be, in their turn, in several natural languages.

Under these conditions, the following two questions arise: (1) *Are available guidelines, best practices, and tools well suited for coping with these and other language related issues*? (2) *Do they provide appropriate guidance and mechanisms for producing high-quality data in such scenarios?*

In the following pages we aim to give principled answers to these questions. According to our experience in publishing the *datos.bne.es* dataset, guidance on these aspects is still very limited. Therefore, our purpose in this article is to review, discuss, and extend current guidelines for publishing linked data (1) by focusing on those methods, techniques and tools that can help RDF publishers to cope with language barriers, and (2) by identifying existing gaps, remaining research and technical challenges. Whenever possible, we will illustrate and discuss each of these guidelines, methods, and tools on the basis of practical examples that we have encountered in the publication of the *datos.bne.es* dataset.

The remaining of the article is organized as follows. In Section 2, we give a detailed overview of the current state of RDF datasets from a language perspective. In particular, we present the results from an in-depth analysis that has been performed for assessing and analyzing the multilingual dimension of the Web of Data. In Section 3, we introduce the methodological guidelines that will be used as starting point for this article. Then, from Section 4 to Section 8, we will deal with each of the activities included in the guidelines. Finally, in Section 9, we will conclude by providing some lessons learnt.

1. THE MULTILINGUAL WEB OF DATA: CURRENT STATE

In this article, we argue that there is a growing number of linked datasets in different natural languages, and that there is a need for guidelines and mechanisms to ensure the quality and organic growth of this emerging multilingual data network. However, we have little knowledge regarding the actual state of this data network, its current practices, and the open challenges that it poses. Questions regarding the distribution of natural languages, the links that are established across data in different languages, or how linguistic features are represented, remain mostly unanswered. Addressing these and other language-related issues can help to identify existing problems, propose new mechanisms and guidelines or adapt the ones in use for publishing linked data including language-related features, and, ultimately, provide metrics to evaluate quality aspects.

As mentioned in the introduction, only Ell et al. took the multilingual dimension into account, although their empirical study limit its analysis to figures about the usage of language tags to indicate the natural language of RDF literals. More specifically, they conclude that (1) most datasets contained at most one language (2.2%), which also indicates a very low usage of language tags, (2) only 0.7% contained several language tags, and that (3) the most used language tags are *en* (English) (44.72%), *de* (German) (5.22%), and *fr* (French) (5.11%). These figures, although interesting, lack: (1) a more in-depth analysis of the language distribution within the analyzed corpus, the BTC (Billion Triples Challenge) dataset[6], and (2) a classification of datasets according to the used natural language.

In this section, we present a study whose purpose is to focus on these two limitations. In particular we make use of the periodical snapshots gathered by DyLDO [19], a framework to monitor Linked Data over an extended period of time. The rationale for selecting this corpora over the BTC dataset is two-fold: a) DyLDO corpora are published monthly, allowing us to better capture the evolution over time, while BTC is published yearly, and b) DyLDO crawling strategy, and seed URIs are stable and do not change from one corpus to another, while BTC has been changing its

Table 1. Summary of the studied corpora

Corpus	Nº Datasets	Nº Monolingual	Nº Multilingual	Nº literals	Nº literals without language tag	Nº literals with language tag
January	2,255	1,906	349	12,818,260	10,250,936 (79.97%)	2,567,324 (20.03%)
June	2,836	2,201	635	13,749,117	10,594,338 (77.05%)	3,154,779 (22.95%)
December	2,660	1,984	676	15,638,736	12.272.806 (78.47%)	3,365,930 (21.53%)

[2] http://datos.bne.es

[3] http://viaf.org

[4] http://www.dnb.de/EN/Service/DigitaleDienste/LinkedData/linkeddata_node

[5] http://www.ifla.org/publications/international-standard-bibliographic-description

[6] The authors analyzed the Billion Triples Challenge dataset (BTC) released in 2011: see http://km.aifb.kit.edu/projects/btc-

crawling strategy and seed URIs each year, which has led to very different corpora from one year to another, in terms of the data

they contain.

For our study we used three snapshots or corpora gathered in January, June, and December of 2012[7]. From these corpora we analyzed a total of 42,206,113 RDF literals: 12,818,260 from January, 13,749,117 from June, and 15,638,736 from December. The study consisted of a number of data extraction and analytical jobs focusing on the following features: (1) distribution of natural language across RDF datasets, and (2) usage of language tags, for indicating the natural language of RDF literals. Table 1 summarizes the characteristics of the studied corpora.

1.1 Distribution of natural languages across RDF datasets

Classifying datasets by the natural languages that they use is useful to understand current practices and to identify datasets for the evaluation of specific techniques and tools. In particular, in this section we present the distribution of monolingual vs. multilingual datasets. For the purposes of this study, we consider a dataset to be monolingual when all its RDF literals are in the same language (e.g., Spanish), and multilingual, when its RDF literals are at least in two different languages (e.g., Spanish and English).

From the results shown in Table 1 we extract the following conclusions (Figure 1):

- The majority of the datasets in the studied corpora are monolingual (78.90% of all datasets on average).
- Between January and December of 2012 the number of multilingual datasets has doubled.

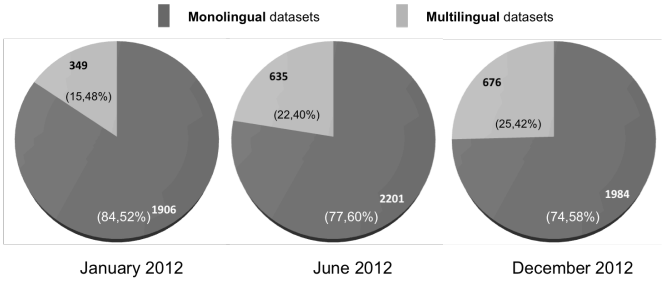

Fig. 1. Monolingual and multilingual datasets in the studied corpora

1.2 Usage of language tags

The RDF data model provides a means for indicating or tagging the language of literals (i.e., UNICODE strings). In particular, the RDF specification[8] allows plain literals (RDF Literals) to optionally include a language tag as defined by RFC-3066[9] normalized to lowercase. In this section we assess and contrast how such language tags are used in the corpora described above (i.e., January 2012, June 2012, and December 2012). More specifically, we provide: (1) the distribution of usage of language tags, (2) distribution of literals tagged as English vs. literals tagged as other languages, and (3) a more in-depth comparison of literals tagged in languages other than English.

1.2.1 Analysis of literals with language tag versus literals without language tag

From the results shown in Table 1 and Figure2 we extract the following conclusions:

- The use of language tags is low (21.50% of all literals on average). This seems to indicate the need of appropriate mechanisms and guidelines for LD publishers to better tag the language of literals.
- We have not observed substantial differences between the studied corpora (the increment/decrement is 1.94% on average). Thus, we do not observe a positive increment in the usage of language tags.
- Additionally, we have identified incorrect usage of language tags in two ways: (1) wrong ISO codes (e.g. spa for Spanish, or i18n), and (2) tags that do not follow the lowercase recommendation of the RDF specification (e.g., EN-US for US English, or ES for Spanish).

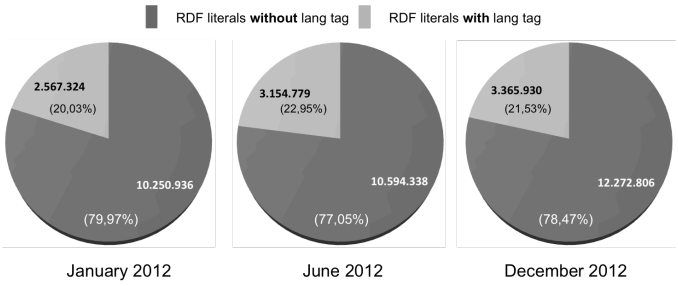

Fig. 2. General usage of language tags in the studied corpora

1.2.2 Analysis of literals in English and literals in other languages

As presented above, only 21.50% literals on average used language tags. We have observed that the presence of English (including US, Australian, and UK English tags) in the studied corpora is higher than that of other languages: January (2,135,664 literals, 83.19% of all tagged literals), June (2,751,065 literals, 87.20% of all tagged literals), and December (2,808,145 literals, 83.42% of all tagged literals).

From the results shown in Figure 3 we extract the following conclusions:

- English is the predominant language in the studied corpora (84.60% of all tagged literals on average).
- The number of literals from January to December rose in 798,606 literals, of which 672,481 literals (84.21%) are tagged as English, and 126,125 literals (15.79%) are tagged as other languages. These figures of literals

[7] http://swse.deri.org/DyLDO/data

[8] RDF Concepts 2004, see http://www.w3.org/TR/rdf-concepts/

[9] RFC-3066 2001 see http://www.isi.edu/in-notes/rfc3066.txt

increment are consistent with the average distribution shown in the three corpora (84.60% for English tags, and 13.40%).

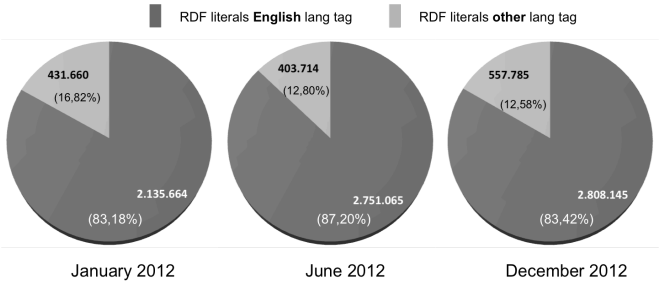

Fig. 3. Usage of English language tags vs. other language tags in the studied corpora

1.2.3 Analysis of literals in languages other than English

In order to analyze more in detail the presence of languages other than English, we looked at the evolution of the 10 most used languages across the three corpora as presented in Table 2 and Figure 1.

Figure 2 depicts the distribution of the 5 most used language tags, and the other languages in the top 10 languages analyzed above. From the results shown above we extract the following conclusion:

- Italian, French, Spanish, and German are the predominant languages in the studied corpora (approximately 50% of all tagged literals on average).

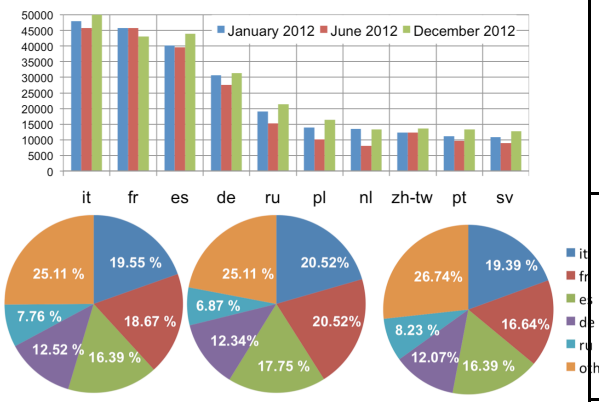

Fig. 4. Evolution of top 10 most used language tags in languages other than English (top graph), and distribution of top 5 languages bottom graph) in the studied corpora

2. METHODOLOGICAL GUIDELINES

In this chapter, we build on the method proposed in [30]. This method follows an iterative incremental model that covers the following activities: (1) *specification*, for analyzing and selecting data sources, (2) *modeling*, for developing the model that represents the information domain of the data sources, (3) *generation*, for transforming the data sources into RDF datasets, (4) *linking*, for creating links between different RDF datasets, (5) *publication*, for publishing the model, RDF datasets and links generated on the Web, and (6) *exploitation*, for developing applications that consume the dataset in question. In its turn, each activity is decomposed into one or more tasks.

So, we will review, discuss, and extend the proposed activities for publishing linked data (1) by focusing on those methods, techniques and tools that can help publishers to cope with language barriers, and (2) by identifying existing gaps, remaining research and technical challenges. Table 2 provides an overview of the activities, the main tasks they are decomposed into, and it also indicates those that will be reviewed in the next sections.

Phase (Goal)	**Tasks**	**Reviewed**
1.Specification Analyzing and describing data (data sources and RDF data) characteristics	1.Identify and analyze the data sources	Yes
	2. Design the URIs/IRIs	Yes
	3. Define license and provenance information	No
2.Modelling Creating/selecting vocabularies to describe the RDF resources	1.Analyze and select domain vocabularies	Yes
	2.Develop the vocabulary	Yes
	3. Vocabulary for provenance information	No
3.Generation Producing RDF datasets from the data sources	1. Technologies for producing RDF	Yes
	2. Create mappings between the vocabulary & sources	No
	3. Transform the data sources into RDF	Yes
4.Linking Connecting the RDF dataset	1.Select target datasets to link the entities in the dataset	Yes
	2. Discover the links with the target datasets	Yes
	3. Validate the links	No
5.Publication Making the dataset available and discoverable on the Web	1. Publish the dataset	No
	2. Publish metadata describing the dataset	Yes

Table 2. Analysing LD activities [30] from a language perspective

3. SPECIFICATION

In this section we explain how to deal with language issues during the specification phase. Basically, we need to analyze whether the original data sources are documented in different natural languages, or if they are intrinsically multilingual. A further aspect to take into account at this stage is the design of precise resource identifiers (URIs or IRIs).

3.1 Analysis of the data sources and their model

The first activity of the LD publication process is to analyze, and specify the data sources that will be used for publishing LD, as well as the data model(s) used within said sources. In this section, we analyze this activity taking into account the natural language dimension, and the scenarios introduced in Section 4.1. In other words, we review how language-related features affect the process of specification, and how can publishers approach this task in a sensible way with regards to natural language.

In Figure 3, we show the metadata descriptions about "Ernest Hemingway" and an edition of "The old man and the sea" in Spanish (i.e., the corresponding title in Spanish is "El viejo y el mar"). As shown in the figure, the data model corresponds to the types of entity[10] (i.e., person, and book), and the different attributes and relationships[11] (e.g., title, is author of, etc.), whereas the content corresponds to the value of each attribute (e.g., "El viejo y el mar", "XX844022", etc.)

In this task, we have to take into account two layers: (1) the *data model*, in this case defined by the MARC 21 format, and (2) the *content of the sources* (i.e., the data itself), in this case the book titles, authors' names, standard identifiers, dates, etc.

Data model. The data model (including standards, terminology, etc.) used for the description of entities, attributes and relationships can be found in the language of the dataset publisher, or in other languages. In this task we recommend compiling all available information about the data model used in the sources, and identifying the natural languages that will be used for designing the domain vocabulary.

Content. Content can be language independent or language-dependent. Some properties such as identifiers, numbers, and some date formats are usually language-independent, whereas names, titles, textual descriptions, and some date formats are normally language-dependent, as they are bound to a specific language. Language-dependent properties do not always make explicit the language of the content they carry. Given this situation, we recommend specifying and classifying attributes in the following way: (1) language independent, or language-dependent, based on the content they carry, (2) for language-dependent attributes, the language can be explicit (e.g., using a metadata annotation, a pointer to the language description or code, etc.) or unspecified (e.g., "Title" shown in Figure 3 is a language-dependent attribute, with unspecified language). For the former case (i.e., explicit language), the mechanisms that are used to indicate the language should be documented. For the latter case, (i.e., unspecified language), the dataset publisher should apply language identification techniques in the generation activity, as we will discuss in Section 6.

[10] Classes in RDF terminology

[11] Properties in RDF terminology

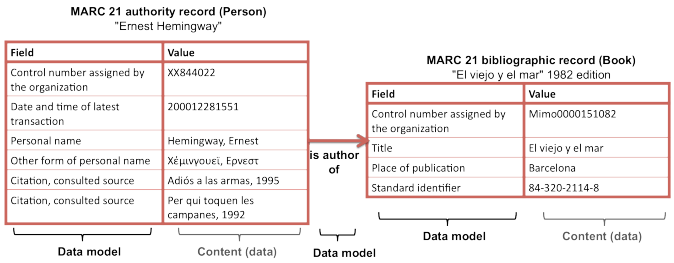

Fig. 5. "Ernest Hemingway" and "El viejo y el mar" MARC 21 records

3.2 URIs and IRIs design

The goal of this task is to design the structure of the resource identifiers that will be used to name RDF resources, either the TBox (classes and properties) or the Abox (instances). In both cases, we basically have two options: to use meaningful or descriptive resource identifiers, i.e., the use of natural language descriptions in the local name of URIs and IRIs (e.g., http://vocab.org/frbr/core#Expression, corresponds to the URI of the class Expression in the FRBR Core vocabulary[12]), or rather, the employment of opaque resource identifiers, i.e., non-human readable local names (e.g., http://iflastandards.info/ns/fr/frbr/frbrer/C1002 is the URI for the class Expression of FRBR defined by IFLA). Both approaches have well-known advantages and disadvantages that we will try to summarize on the light of the multilingual dimension.

From a technical point of view, in a multilingual scenario we have several options:

- Use of **meaningful URIs** or *descriptive URIs*, in which the local name is normally in English or any other Latin-based language which makes use of the ASCII characters only, e.g.,http://vocab.org/frbr/core#Expression; http://example.org/frbr/core#Expresion (fictitious meaningful URI in Spanish without the diacritical mark, compulsory in the Spanish word *expresión*).
- Use of **full IRIs** (Internationalized Resource Identifier) [21], created with the aim of allowing the use of Unicode characters for languages that do not follow the Latin alphabet. This enables the use of Unicode characters not only for local names but also in the domain part. E.g., http://frbr-español.org/Expresión (in correct Spanish).
- Use of **Internationalized Local Names**, which are IRIs in which the domain part is restricted to ASCII characters while the local name can use Unicode characters, [21]. E.g., http://example.org/frbr/core#Expresión (fictitious URI with the diacritical mark).

Additionally, if our starting point is a multilingual resource in which TBox and ABox contain information in several languages, more fundamental questions should be brought up: which language should we use for the local names in meaningful URIs or IRIs? Should English be the default language? In which language was the dataset originally created? Does it contain preferred labels in that language (by means of the skos:prefLabel annotation property)? Or should we opt for opaque URIs to avoid

[12] http://vocab.org/frbr/core

any language bias? Would a hybrid approach work (meaningful URIs for the TBox and opaque URIs for the ABox)? Moreover, if we decide to use meaningful URIs or unrestricted IRIs, which format should we follow in the local name (CamelCase strategy, use of space or underscores as word separators)? E.g., (http://vocab.org/frbr/core#creatorOf vs. "http://vocab.org/frbr/core#creator Of", or http://vocab.org/frbr/core#is_creator_of). These are some questions that should be addressed beforehand in order to choose the naming format.

However, there are also some other arguments that support the use of opaque URIs or IRIs, as summarized in [25]. For example, in a Semantic Web context, resource identifiers are intended for machine consumption, so that there is no need for them to be human-readable. It is also well accepted that opaque URIs make ontologies more stable, so once the ontology has been published and adopted by a community of users, local names should not change even if the natural language descriptions associated to them are modified (unless the actual meaning of concepts has changed). Furthermore, opaque URIs may also be a good solution if we have a multilingual data source and we want to avoid any language bias.

4. MODELING

Once the specification activity has been completed, it is time to build the model to be followed for the selected data sources. The most important recommendation is to reuse available vocabularies as much as possible. In this direction, current methodological guidelines divide this activity into two core tasks: (1) analysis and selection of domain vocabularies to maximize reuse of widely-deployed vocabularies, and (2) development of the domain vocabulary reusing as many terms as possible and creating those concepts that are not covered by the vocabularies analyzed in the previous task. From a multilingual perspective, however, we argue that such guidelines are insufficient and do not account for linguistic and cultural varieties. It is frequently the case that the linked data publishers want to provide descriptions to vocabulary classes and properties in their own language, or even in several languages to improve vocabulary usability, data visualization, and so on. For this reason, in this chapter we propose an optional task, (4) "Vocabulary localization" that deals with this issue. Additionally, we review tasks (1) and (2) to account for the multilingual dimension. In the following we present and discuss the aforementioned tasks.

4.1 Analysis and selection of domain vocabularies

The goal of this task is to analyze and select already available domain vocabularies that will be used to model the RDF data. Currently, there are several catalogues and services suitable for finding available vocabularies on the Web such as the Semantic Web Search Engine[13] (SWSE), Sindice[14], the Datahub[15], Falcons[16], or LOV[17] (Linked Open Vocabularies). These catalogues and services allow users to: (1) *search for similar data* within similar domains (SWSE, Sindice, and Datahub), and (2) *search for vocabularies or specific terms* (Falcons, and LOV). Thinking about the multilingual dimension of vocabularies we question ourselves: *do existing catalogues and services take this dimension into account, facilitating discovery of terms no matter the language they are described in?*

In order to assess their support, we issued a search query with the keyword "شخص" ("person " in Arabic), and with the keyword "プロジェクト" ("project" in Japanese) given that this term is included in DOAP (Description Of A Project)[18]

Table 4 shows the results from the analysus discussed above, specifically: (1) *Indexing capabilities*: the service or catalogue is capable of indexing data in several languages (✓:Yes, **X**: No), (2) *User Interface (UI) support*: the degree of support for natural languages in the interface showing the search results, (- : Low, + : medium, ++ : high) (3) *Search facet*: it facilitates filtering the results by language, and (4) *Additional information*: some additional remarks about the service or catalogue.

Summarizing, we observe that current multilingual support is still limited although services like Falcons and LOV can be used for finding terms in different languages. In particular, according to our analysis, LOV represents the best option due to the following reasons: (1) it is able to index multilingual labels , (2) it provides the best UI support for languages, and (3) it is a well-established repository with long term support from the Open Knowledge Foundation (OKF) and a clear curation strategy[19].

Table 4. Comparative study of catalogues and services for vocabulary reuse from a multilingual perspective

Catalogue	Indexing	UI	Search	Additional information
SWSE	**X**	_	**X**	Issues parsing UNICODE characters
Sindice	✓	_	**X**	Difficult to grasp results for vocabulary reuse
Datahub	N/A	_	**X**	Focuses on datasets, Difficult to grasp results for vocabulary reuse
Falcons	✓	_	**X**	UI does not properly account for languages
LOV	✓	++	**X**	Stable and long-term support, does not cover highly specialized domains

Finally, in order to assess the availability of multilingual vocabularies in the recommend service, we have performed a classification of vocabularies in LOV according to natural languages. The results are presented in Table 5,

Table 5. Classification of vocabularies in LOV according to the natural language dimension (retrieved 12.04.2013)

Type	Number of vocabularies
Monolingual (only en)	223

[13] http://swse.deri.org/

[14] http://sindice.com

[15] http://datahub.io

[16] http://ws.nju.edu.cn/falcons

[17] http://swse.deri.org/

[18] http://usefulinc.com/ns/doap#Project

[19] http://lov.okfn.org/dataset/lov/suggest/

Multilingual	53
Non specified language	42
Monolingual (no en)	8
Total	326

4.2 Ontology localization

The term "ontology localization" was defined in [15, 9] as the process of adapting an ontology to the needs of a particular (linguistic and cultural) community. A localized ontology can be understood as an ontology adapted to the target community and language, and used independently of the original ontology, or, most commonly, as an ontology in which the vocabulary or TBox has been translated to one or several natural languages, so that it contains terms in several languages for describing classes and properties [16]. When extrapolating this to the linked data context, if the vocabulary publisher reuses an available vocabulary and decides to translate or localize the vocabulary terms into other languages, this could be understood as vocabulary localization and the result would be a multilingual vocabulary.

For this purpose, publishers could make use of some ontology localization tools such as LabelTranslator [, 2009] or the ontology translation component developed in the Monnet project[20] (specially tuned for the financial domain), to (semi)-automatically translate the vocabulary. However, either making use of these tools or following a manual approach, publishers should decide which representational model to follow according to their multilingual and linguistic needs. In this sense, three main alternatives have been identified in [30]: (1) *Multilingual labelling approach*, (2) *Association of the vocabulary to an external lexicon model*, and (2) *Cross-lingual linking or matching approach*.

To illustrate the three approaches, we will use as example the localization of the ISBD standard introduced in section 5.2. As already mentioned, English labels for classes and properties were translated into Spanish, resulting in a multilingual vocabulary in English and Spanish. For this specific case, the publishers decided to rely on the SKOS annotation property for preferred labels (skos:prefLabel), and agreed on the use of only one preferred label per language (see section 5.3.1). However, the Spanish translation of this vocabulary revealed a problem which was not apparent in the English version, namely, that some labels were adjectives (*cartographic* in English), which in Spanish require a form change depending on whether the word they modify is masculine (*cartográfico*) or feminine (*cartográfica*). Because of the agreed restriction, compounds such as "cartográfico/a" were suggested (skos:prefLabel "cartográfico/a"@es), which have some problems, such as the fact that these compounds would not naturally appear in free texts.

4.2.1 Multilingual labeling approach

The first alternative relies on a single conceptual or data structure to which alternative labeling information is provided in the form of plain literals represented as properties of concepts. This is supported by RDFS or SKOS. Some additional support for describing lexical entities is provided by SKOS-XL. In this extension of the SKOS syntax, labels are considered SKOS classes and, therefore, assertions can be made on these classes. Below you can see examples of this approach in RDFS, SKOS and SKOS-XL.

```
isbd:T1001    rdfs:label "cartográfico"@es;
              rdfs:label "cartográfica"@es.
```

Listing 1. Example in RDFS

```
isbd:T1001    skos:prefLabel    "cartográfico/a"@es.
```

Listing 2. Example in SKOS

```
isbd:T1001 skosxl:prefLabel :cartografico.
:cartografico a skosxl:Label;
    skosxl:literalForm "cartográfico"@es.
isbd:T1001 skosxl:prefLabel :cartografica.
:cartografica a skosxl:Label;
    skosxl:literalForm "cartográfica"@es.
              rdfs:label "cartográfica"@es.
```

Listing 3. Example in SKOS-XL

The main disadvantage of the labeling facility of RDFS and SKOS is that the set of labels that can be related with one vocabulary term result in a bunch of unrelated labels whose motivation cannot be asserted and for which further properties cannot be specified (for instance, specify the gender -masculine and feminine- in the cartographic example). This is, in a sense, solved by the SKOS-XL description, which allows for labels to be treated as RDF classes, so, in principle, additional assertions could be made. However, SKOS-XL does not provide a principled way for specifying linguistic properties of labels, nor is it conceived to linguistically enrich vocabulary terms (for instance, specifying that the plural forms of *cartográfico* and *cartográfica* are obtained by adding an –s, etc.). It is for these reasons that linguistic models have been proposed to enrich ontologies and vocabularies, as explained in the next section.

4.2.2 Association of the vocabulary to an external lexicon model

The second alternative consists in associating the vocabulary to a lexicon model that contains the lexical and linguistic information relative to that vocabulary (in one or several languages). Examples of these ontology-lexicon models are LexInfo [10], LIR [29] or lemon [24]. In fact, the lemon model has an RDF implementation, which allows publishing linguistic and lexical information in the linked data format. *lemon*[21] considers the possibility of associating lexical and terminological descriptions to vocabularies and ontologies of the following type: linguistic properties (part-of-speech, gender, number, etc.), lexical and terminological variation, decompositions of phrase structures (representation of multi-word expressions), syntactic frames and their mappings to the logical predicates in the ontology, and morphological decomposition of lexical forms.

[20] http://www.monnet-project.eu

[21] Here it is also worth mentioning the OntoLex W3C community effort

In order to illustrate the potentiality of such models, we present how *lemon* allows for the inclusion of the two adjectival forms of the cartographic adjective in Spanish, the masculine and the feminine, by linking them to that property in the ontology by means of a LexicalEntry with two LexicalForm (masculine and feminine). The model is also able to represent that these are form variants of the same lexical entry. Isocat categories are used in the example to represent the grammatical gender.

isbd:T1001 lemon:isReferenceOf [lemon:isSenseOf :cartographic].

```
:cartographic a lemon:LexicalEntry;
    lemon:form [lemon:writtenRep "cartográfico"@es;
        isocat:grammaticalGender isocat:masculine];
    lemon:form [lemon:writtenRep "cartográfica"@es;
        isocat:grammaticalGender isocat:feminine].
isocat:grammaticalGender rdfs:subPropertyOf lemon:property.
```

Listing 4. Example in *lemon*

4.2.3 Cross-lingual linking or matching approach

Regarding the third possibility, the cross-lingual or matching approach, it can be followed whenever we count on two or several vocabularies defined in different natural languages, but covering the same or similar subject domains. In this approach links are established between the terms or lexical entries that describe the two vocabularies. In this sense, the scenario also involves the automatic discovery of links, another crucial issue in the Multilingual Semantic Web.

A number of recently developed cross-lingual ontology alignment tools can be used to that end[22]. Currently, equivalent links can be represented by means of properties of current Semantic Web languages such as OWL (owl:sameAs to link individuals in ontologies, or owl:equivalentClass and owl:equivalentProperty to link classes and properties in ontologies that have the same extension, as well as with other commonly used vocabularies such as SKOS (skos:closeMatch to link two concepts that are sufficiently similar and skos:exactMatch, when the similarity degree is even higher). It could be argued that such links can be reused for the purpose of establishing links between classes, properties and individuals expressed in different natural languages in the Linked Data cloud. However, we claim that some of these cross-lingual equivalences need to be analyzed carefully within the multilingual dimension, since we may want to establish cross-lingual and cross-cultural equivalences that may not admit the strong ontological commitments that current links make. Additionally, we also consider the possibility of establishing links between lexical entries of ontology-lexicon models are associated with ontologies and linked data vocabularies. For more on this, see [25b]

Continuing with our example, we show a very simple example of cross-lingual link between the entity "Bibliographical searching" as it is represented in the BNE dataset ("Busqueda bibliográfica") and the BNF dataset ("Recherche documentaire"):

```
http://data.bnf.fr/ark:/12148/cb11941422b a skos:Concept
skos:prefLabel "Recherche documentaire"@fr.

http://datos.bne.es/resource/XX528311 a skos:Concept;
  rdfs:label "Búsqueda bibliográfica"@es;
  owl:sameAs http://data.bnf.fr/ark:/12148/cb11941422b.
```

Listing 5. Example of cross-lingual mapping

4.2.4 Discussion

The main difference between the first two approaches (section 5.3.1 and section 5.3.2) is that the first option considerably restricts the amount and type of linguistic information that can be related to vocabulary elements, whereas the second one allows for the inclusion of as much linguistic information as needed by the final application. The choice between one and the other model will depend on the linguistic requirements of each use. As for the third approach, (section 5.3.3), it depends on the availability of similar vocabularies in different natural languages on the same domain.

5. GENERATION

This activity deals with the transformation of the data sources selected in the *specification* activity (presented in section 4) using the model developed in the *modeling* activity (treated in section 5). This is a crucial activity in the process of publication, and is, of course, influenced by language-related features. In this chapter we focus on two core aspects of the RDF generation, and point the reader to other relevant works.

Language identification. As reviewed in section 2, the current usage of language tags in RDF datasets is still limited (only 21,50% of all analyzed literals on average), and there is a need of adequate guidelines and techniques for tagging the language. Also, as discussed in section 4.2, *language-dependent* properties can (1) explicitly specify the language of the content that they carry (via language codes, external information, etc.), or (2) leave the language *unspecified*. For the former case, the generation activity should include mechanisms to leverage the specified language to properly tag the language of the generated RDF literals. For the latter scenario, it might be necessary to automatically "guess" or "identify" the language, using so-called "language identification" techniques [12]. For this, we find extensive literature that can be useful for the case of RDF properties, where literals are usually short [31, 15], as well as some available tools23.

Encoding issues. An important aspect when working with languages whose scripts make use of characters not included in ASCII is the appropriate handling of the encoding of such characters. The *generation* activity is probably the most important activity in order to assure proper encoding, thus producing quality RDF data. When generating LD encoding issues affect several levels: (1) URIs and IRIs handling, (2) Different RDF serialization formats (e.g., RDF/XML, NTriples, etc.), and (3)

[22] See for instance CIDER-CL (http://www.oeg-upm.net/files/cider-cl) or the set of systems that participated in OAEI2012 (http://oaei.ontologymatching.org/2012/multifarm/index.html)

[23] See for example: http://tika.apache.org/, http://code.google.com/p/language-detection and http://nutch.apache.org

libraries and tools for RDF (e.g., triple-stores[24], APIs[25], RDF generation tools[26], etc.). Taking informed decision in the selection of technologies, serialization formats, and unique identifiers (IRIs or URIs) will lead to better quality RDF data and avoid problems for RDF consumers. In this sense, we point the reader to [Auer et al., 2008], who provides an in-depth survey on these issues that might help publishers to make suitable choices.

6. INTERLINKING

In a multilingual Web of Data, semantic data with lexical representations in one natural language are mapped to equivalent or related information in other languages, thus allowing navigation across multilingual information by software agents [16]. Several activities have to be carried out for cross-lingual interlinking: (1) the selection of relevant and authoritative mono/multilingual datasets to link, (2) the automatic discovery of equivalent and/or related entities between the dataset and the selected external resources, and finally (3) the representation and storage of the discovered links.

6.1 Selection of target datasets or vocabularies to link the entities

The goal of this task is to identify those RDF datasets about similar topics that can provide extra information to the entities in the dataset. To that end we refer to the systems described in Section 5.2. However, not all of them are suited for the discovery of RDF datasets, such as LOV, which is focused on vocabularies only. Others, such as Datahub, Sindice, etc. are suitable for that purpose, in the sense that they store metadata about RDF datasets, but the selection task is hampered by the fact that language is not an explicit search parameter.

In our running example, we were able to identify other relevant datasets that could enrich the information contained in *datos.bne.es,* in order to allow the consumers to navigate related resources. That is the case of *VIAF* (Virtual International Authority File), the German National Library (DNB, Deutsche Nationalbibliothek), or the French National Library (BNF, Bibliothèque nationale de France). All of them are reachable by means of Datahub.

6.2 Cross-lingual link discovery and representation

This activity involves the automatic discovery of relationships between data items to increase the external connectivity of the RDF dataset. Automatic discovery of relationships among data items in a multilingual scenario poses an added challenge because of data sources being available in different natural languages. There are many tools and techniques for discovering links between data items of different RDF datasets (see [17] for a survey). Nevertheless, none of these techniques consider multilingualism as an explicit feature and do not include specific techniques to deal with language diversity during the process of link discovery. Therefore, more research is also needed on *automatic* methods for cross-lingual instance matching.

As for the representation of links, current solutions may fall short of representing cross-lingual and cross-cultural equivalences, as briefly introduced in section 5.3. Richer options for representing such mappings at the linguistic level need further exploration and remain as an interesting challenge.

7. PUBLICATION

The publication of multilingual resources would involve the same tasks as in a monolingual process: (1) dataset publication, (2) metadata publication, and (3) enabling effective discovery. In the context of this chapter, we limit the scope to the second task. In recent years we have found two major vocabularies for publishing metadata describing datasets and catalogs: VoID[27] (Vocabulary of Interlinked Datasets) [1], and DCAT[28] (Data Catalog Vocabulary) [27] both published in the context of the W3C. In this section we show through examples how to account for the language dimension of datasets using these two vocabularies.

Although there might be other areas where language could be involved (for instance, when the dataset contains cross-lingual links), the most basic aspect to describe is *the language or languages used in the dataset*. Surprisingly, the language dimension in VoID is not included in its specification. DCAT, on the other hand, includes a property to indicate language by means of the dcterms:language property, and defines the range of the property in the following way: (1) use resources defined by the Library of Congress[29], [30], (2) if an ISO 639-1 (two-letter) code is defined for language, then its corresponding IRI *should* be used;, otherwise (3) if no ISO 639-1 code is defined, then the IRI corresponding to the ISO 639-2 (three-letter) code *should* be used. As both VoID and DCAT reuse the *Dublin Core Metadata Terms*[31] vocabulary for providing basic metadata (e.g., dcterms:publisher, dcterms:title, etc.), it seems natural to recommend publishers to follow the recommendation found in DCAT also for building VoID descriptions. Therefore, in Listing we provide an example of the recommended mechanism to indicate the dataset language for VoID and DCAT.

```
# VoiD description
:bne a void:Dataset;
    dcterms:language <http://id.loc.gov/vocabulary/iso639-1/es> .
# DCAT description
:bne a dcat:Dataset;
    dcterms:language <http://id.loc.gov/vocabulary/iso639-1/es>;
```

Listing 6. VoID and DCAT descriptions indicating the language used in the dataset

[24] Some examples are Virtuoso (http://openlinksw.com), 4Store (http://4store.org), and Allegrograph (http://www.franz.com/agraph/allegrograph/)

[25] Some examples are Apache Jena (http://jena.apache.org/), Sesame (http://www.openrdf.org/) and ARC2 (https://github.com/semsol/arc2)

[26] Some examples are RDF refine (http://refine.deri.ie/), and Apache Any23 (http://any23.apache.org/)

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

[28] http://www.w3.org/TR/vocab-dcat/

[29] http://id.loc.gov/vocabulary/iso639-1.html

[30] http://id.loc.gov/vocabulary/iso639-2.html

[31] http://dublincore.org/documents/2010/10/11/dcmi-terms/

8. CONCLUSIONS

This contribution aims at throwing some light on some issues raised by multilingualism in the Linked Data world/field. Especially, we have focused on methodological guidelines to support users in the transformation and publication of data sources according to the Linked Data paradigm. As we have shown in this chapter, some methods, technologies and tools currently used for publishing and consuming linked data can be directly applied to multilingual resources, whereas others need to be enhanced to cope with linguistic diversity. As a summary and to conclude this contribution, we provide a list of the main lessons learnt:

1. In the specification activity, *perform a careful analysis of the data sources regarding the natural languages in which your data are described*.
2. To properly use language tags for identifying the language of RDF literals, in the specification activity, *identify language-dependent properties and document the mechanisms used in the data sources to indicate the language*.
3. As part of the specification activity, *consider the advantages and disadvantages of meaningful vs. opaque URIs*.
4. In the generation activity, when mapping the data sources and the domain model, *be sensible to language-dependent properties, and tag the language in the RDF produced* (e.g., using language identification techniques, etc.)
5. If you use languages (e.g., in the model, the data, or IRIs) with characters not included in ASCII, *take informed decisions when selecting the technologies and serializations for producing RDF*.
6. *Reuse existing vocabularies*, *if possible those that are described in several languages*.
7. *When creating new vocabulary classes and properties, localize them*. In the modeling activity, *identify possible existing NORs to leverage the localization process*,
8. *Identify which localization strategy suits your requirements better* (multilingual labeling, external lexicon, or cross-lingual linking).
9. *Link your vocabulary and dataset to others linked datasets in the same or other languages*, and *think about the possibilities and ontology commitments of the different types of links from a multilingual perspective*.
10. *Specify the natural languages used within your dataset when publishing your VoID and/or DCAT* dataset descriptions.

ACKNOWLEDGMENTS

This work has been supported by the BabelData (TIN2010-17550) and myBigData (TIN2010-17060) Spanish projects and by the FP7 European project Monnet (FP7-ICT-4-248458)

REFERENCES

[1] K. Alexander, R. Cyganiak, M. Hausenblas, and J. Zhao. Describing Linked Datasets with the VoID Vocabulary. W3C in- terest group note, W3C, 2011. http://www.w3.org/TR/void/.

[2] [Auer, S., Lehmann, J., & Hellmann, S. (2009). Linkedgeodata: Adding a spatial dimension to the web of data. In The Semantic Web-ISWC 2009 (pp. 731-746). Springer Berlin Heidelberg.

[3] Auer, S., Weidl, M., Lehmann, J., Zaveri, A. J., and Choi, K. S. (2010). I18n of semantic web applications. In The Semantic Web–ISWC 2010 (pp. 1-16). Springer Berlin Heidelberg.

[4] Auer, S., Bühmann, L., Dirschl, C., Erling, O., Hausenblas, M., Isele, R., and Williams, H. (2012). Managing the life-cycle of Linked Data with the LOD2 Stack. In The Semantic Web–ISWC 2012 (pp. 1-16). Springer Berlin Heidelberg.

[5] Design issues: Linked Data. Online resource available at http://www.w3.org/DesignIssues/LinkedData (last viewed April 2013)

[6] Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked data-the story so far.International Journal on Semantic Web and Information Systems (IJSWIS),5(3), 1-22.

[7] P. Buitelaar, K.S. Choi, P. Cimiano, E. Hovy. Report on the Dagstuhl Seminar: “The Multilingual Semantic Web", November 22, 2012 (to be published).

[8] Cimiano, Philip, Paul Buitelaar, John McCrae, and Michael Sintek. (2010). LexInfo: A Declarative Model for the Lexicon-Ontology Interface. Journal of Web

[9] Cimiano, P., Montiel-Ponsoda, E., Buitelaar, P., Espinoza, M., & Gómez-Pérez, A. (2010). A note on ontology localization. *Applied Ontology*, *5*(2), 127-137.

[10] Cimiano, P., Buitelaar, P., McCrae, J., & Sintek, M. (2011). LexInfo: A declarative model for the lexicon-ontology interface. *Web Semantics: Science, Services and Agents on the World Wide Web*, *9*(1), 29-51.

[11] M. d'Aquin, C. Baldassarre, L. Gridinoc, S. Angeletou, M. Sabou, and E. Motta. Characterizing knowledge on the semantic web with watson. In R. Garcia-Castro, D. Vrandecic, A. Gmez-Prez, Y. Sure, and Z. Huang, editors, EON, volume 329 of CEUR Workshop Proceedings, pages 1{10. CEUR-WS.org, 2007.

[12] Das, S., Sundara, S., and Cyganiak, R. (2012). R2RML: RDB to RDF Mapping Language. W3C Recommendation

[13] L. Ding and T. Finin. Characterizing the semantic web on the web. In Proceedingsof the 5th International Semantic Web Conference, 2006.

[14] Dunning, T., Laboratory, N.M.S.U.C.R.: Statistical Identification of Language. Memoranda in computer and cognitive science. Computing Research Laboratory, New Mexico State University (1994)

[15] Ell, B., Vrandecic, D., Simperl, E.P.B.: Labels in the web of data. In Aroyo, L., Welty, C., Alani, H., Taylor, J., Bernstein, A., Kagal, L., Noy, N.F., Blomqvist, E., eds.: International Semantic Web Conference (1). Volume 7031 of Lecture Notes in Computer Science., Springer (2011) 162–176

[16] Espinoza, M., Gómez-Pérez, A., & Mena, E. (2008). Enriching an ontology with multilingual information. In *The Semantic Web: Research and Applications*(pp. 333-347). Springer Berlin Heidelberg.

[17] Espinoza, M., Montiel-Ponsoda, E., & Gómez-Pérez, A. (2009, September). Ontology localization. In *Proceedings of the fifth international conference on Knowledge capture* (pp. 33-40). ACM.

[18] Ferrara, A., Nikolov, A., & Scharffe, F. (2011). Data linking for the semantic web. International Journal on Semantic Web and Information Systems (IJSWIS), 7(3), 46-76.

[19] Gottron, T., Lipka, N.: A comparison of language identification approaches on short, query-style texts. In Gurrin, C., He, Y., Kazai, G., Kruschwitz, U., Little, S., Roelleke, T., Ru ger, S.M., van Rijsbergen, K., eds.: ECIR. Volume 5993 of Lecture Notes in Computer Science., Springer (2010) 611–614

[20] J. Gracia, E. Montiel-Ponsoda, P. Cimiano, A. Gómez-Pérez, P. Buitelaar, J. McCrae. Challenges for the multilingual Web of Data. In Web Semantics: Science, Servicesand Agents on the World Wide Web 11, p 63-71, 2011.

[21] Tom Heath and Christian Bizer (2011) *Linked Data: Evolving the Web into a Global Data Space* (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1-136. Morgan & Claypool.

[22] B. Hyland, B. Villazón-Terrazas, G. Atemezig. Best Practices for Publishing Linked Data. W3C Note 18 April 2013. Available at http://www.w3.org/TR/gld-bp/.

[23] A. Isaac and B. Haslhofer. Europeana Linked Open Data – data.europeana.eu. Semantic Web Journal, to appear. Available from http://www.semantic-web-journal.net/

[24] Tobias Käfer, Jürgen Umbrich, Aidan Hogan and Axel Polleres, Towards a Dynamic Linked Data Observatory, in the Proceedings of the Linked Data on the Web WWW2012 Workshop (LDOW 2012), Lyon, France, 16 April, 2012.

[25] Kobilarov, G., Scott, T., Raimond, Y., Oliver, S., Sizemore, C., Smethurst, M., ... & Lee, R. (2009). Media meets semantic web–how the bbc uses dbpedia and linked data to make connections. In The Semantic Web: Research and Applications (pp. 723-737). Springer Berlin Heidelberg.

[26] Jose Emilio Labra Gayo, Dimitris Kontokostas, Soeren Auer, Multilingual Linked Open Data Patterns. Semantic Web journal [under review], 2013. Available from http://www.semantic-web-journal.net/

[27] Maali, F., Cyganiak, R., & Peristeras, V. (2012). A publishing pipeline for linked government data. In The Semantic Web: Research and Applications (pp. 778-792). Springer Berlin Heidelberg.

[28] F. Maali, J. Erickson, P. Archer. Data Catalog Vocabulary (DCAT) W3C Working Draft 12 March 2013. Available at http://www.w3.org/TR/vocab-dcat/.

[29] McCrae, John, Guadalupe Aguado de Cea, Paul Buitelaar, Philipp Cimiano, Thierry Declerck, Asunción Gómez- Pérez, Jorge Gracia, Laura Hollink, Elena Montiel-Ponsoda, Dennis Spohr, and Tobias Wunner. (2012). Interchanging Lexical Resources in the Semantic Web. Language Resources and Evaluation 46, (4),p. 701-719.

[30] E. Montiel-Ponsoda, G. Aguado de Cea, A. Gómez-Pérez, and W. Peters, Enriching ontologies with multilingual information. In Journal of Natural Language Engineering 17 (3): 283–309. 2009

[31] E. Montiel-Ponsoda, D. Vila-Suero, B. Villazón-Terrazas, G. Dunsire, E. Escolano Rodríguez, A. Gómez-Pérez. 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.

[32] Montiel-Ponsoda, E., Gracia, J., Aguado de Cea, G., Gómez-Pérez, A. (2011). Representing Translations on the Semantic Web. En actas del workshop *MSW 2011 – Workshop on the Multilingual Semantic Web*, CEUR-Proceedings Vol-775, pp. 25-37.

[33] Vila-Suero, D., Villazón-Terrazas, B. and Gómez-Pérez, A. (2013), "datos.bne.es: A library linked dataset". Semantic Web Journal, to appear. Available from http://www.semantic-web-journal.net/.

[34] Vilches-Blázquez, L. M., Villazón-Terrazas, B., Corcho, O., & Gómez-Pérez, A. (2013). Integrating geographical information in the Linked Digital Earth. International Journal of Digital Earth, (just-accepted).

[35] Villazón-Terrazas, Boris, Mari Carmen Suárez-Figueroa, and Asunción Gómez-Pérez. "A pattern-based method for re-engineering non-ontological resources into ontologies." International Jounal on Semantic Web and Information Systems 6.4 (2010): 27-63.

[36] B. Villazón-Terrazas, L. Vilches-Blázquez, O. Corcho, A. Gómez-Pérez. Methodological Guidelines for Publishing Government Linked Data. In Wood, D. (ed.): Linking Government Data. Springer New York, p. 27–49, 2011.

[37] Villazón-Terrazas, B., Vila-Suero, D., Garijo, D., Vilches-Blazquez, L. M., Poveda-Villalon, M., Mora, J., & Gomez-Perez, A. Publishing Linked Data-There is no One-Size-Fits-All Formula. European Data Forum (2012)

[38] Vojtek, P., Bieliková, M.: M.: Comparing natural language identification methods based on markov processes. In: In: Slovko, International Seminar on Computer Treatment of Slavic and East European Languages. (2007)