Kernerman DICTIONARY News

Number 23 • June 2015
kdictionaries.com/kdn

# Multilingual dictionaries and the Web of Data

Jorge Gracia

## 1. Introduction

Nowadays, we are witnessing a growing trend in publishing language resources (lexicons, corpora, dictionaries, etc) as Linked Data (LD) on the Web. LD refers to a set of best practices for exposing, sharing and connecting data on the Web (Bizer et al 2009). In short, the LD paradigm requires that (i) resources are represented on the Web via HTTP URIs (Unique Resource Identifiers), (ii) once a resource is accessed via its URI, information about it is obtained, and (iii) such information contains links to other resources. The basic mechanism to support the representation of resources and their related information is the Resource Description Framework (RDF[1]), which follows the *subject-object-predicate* pattern. Resources can be anything, including documents, people, physical objects and abstract concepts. Following LD principles, a 'Web of Data' emerges in which links are at the level of data, as a counterpart to the "traditional" Web in which links are established at the level of documents (e.g. hyperlinks between webpages).

Publishing language resources as LD offers clear advantages to both the data owners and data users, such as higher independence from domain-specific data formats or vendor-specific APIs, as well as easier access and re-use of linguistic data by semantic-aware software agents. Further, we argue that reaching a critical mass of linguistic data as LD on the Web will set the basis for a new generation of LD-aware Natural Language Processing (NLP) services, with improved scalability and better interoperability level. The latter is, in fact, one of the motivations of LIDER[2], a European project that is driving a remarkable community effort in that direction.

In this context, the Ontology Engineering Group (OEG[3]) at Universidad Politécnica de Madrid has started converting a series of bilingual dictionaries and multilingual terminologies and publishing them as LD on the Web. In the following paragraphs we briefly present the RDF conversion process that we have followed, and report on our experience with two of these datasets: Apertium and Terminesp.

## 2. RDF generation of bilingual and multilingual dictionaries

Recently the W3C Best Practises for Multilingual Linked Open Data (BPMLOD) community group[4] has proposed a set of guidelines for the LD generation of language resources. In particular, the guidelines for bilingual dictionaries[5] identify five steps, namely: (i) vocabulary selection, (ii) modelling, (iii) URI

[1] http://w3.org/TR/rdf11-primer/

[2] http://lider-project.eu/

[3] http://oeg-upm.net/

[4] http://w3.org/community/bpmlod/

[5] http://bpmlod.github.io/report/bilingual-dictionaries/index.html/

Editor | **Ilan Kernerman**

This issue is dedicated to the memory of **Adam Kilgarriff**

**K DICTIONARIES LTD**
8 Nahum Hanavi St.
Tel Aviv 63503 Israel
Tel: +972-3-5468102
kdl@kdictionaries.com
http://kdictionaries.com

ISSN 1565-4745

**Jorge Gracia** is post-doctoral researcher at the Ontology Engineering Group, Universidad Politécnica de Madrid. He got his PhD in Computer Science at University of Zaragoza in 2009 with a thesis about heterogeneity issues on the Semantic Web. His current research interests include multilingualism and Linked Data, linguistic linked data, and cross-lingual matching and information access on the Semantic Web. Currently he is exploring how to move language resources (lexica, dictionaries, corpora, etc) from their data silos into the multilingual Web of Data and make them interoperable, in support of future generation Linked Data-aware NLP tools.

http://jogracia.url.ph/web/

design, (iv) generation, and (v) publication. A similar approach can be followed for multilingual dictionaries as well. We are not going into the details here, but will just highlight key aspects.

In order to represent the lexical information contained in the original dictionaries, we relied on the LExicon Model for ONtologies (*lemon*[6]), a de-facto standard for representing ontology lexica. We used the *lemon translation module*[7] to represent explicit translations between languages (Gracia et al 2014). As a result of the conversion into RDF of a bilingual dictionary, a *lemon* lexicon is defined per language, where all the *translations* corresponding to a pair of languages are grouped under the same *translation set*. A *translation set* groups a set of translations sharing certain properties, for instance stemming from the same language resource, or belonging to the same organisation, etc.

To design the URIs of our RDF datasets, we adopted the patterns and recommendations proposed in the context of the ISA program (Archer et al 2012). In order to construct the URIs of the lexical entries, their senses and other elements, we preserved the identifiers of the original data whenever possible, propagating them into the RDF representation. This is, for example, the URI that points to the Apertium English-Spanish translation set: http://linguistic.linkeddata.es/id/apertium/tranSetEN-ES/.

One of the most interesting aspects of our conversion methodology is that, by being consistent with the generation rules of every URI assigned to every element in the model (lexicons, lexical entries, lexical senses, translations, etc), each time a bilingual dictionary is converted into LD, its monolingual lexicon is not created again if it already exists but its lexical entries are shared by two or more translation sets (see Figure 1). This allows the dynamic growth of monolingual lexicons and, more significantly, shared lexical entries serve as pivot nodes in the graph to allow getting indirect translations from two languages initially disconnected in the original dictionaries.

The generation step deals with the transformation into RDF of the selected data sources using the chosen representation scheme and modelling patterns. Depending on the format of the data source, there are a number of tools that can be used to support this task. In our work we have used Open Refine[8] (and its RDF plug-in) in a preferred way.

Finally, the generated RDF data has been loaded in a triple store and made accessible through a single SPARQL endpoint. In that way, all the data from the original dictionaries were made accessible as LD on the Web in a unified graph with lexical entries, senses, translations, etc, as nodes. All the nodes were identified with dereferenceable URIs. Such data can be accessed by means of SPARQL clients and RDF and HTML browsers.

## 3. Apertium

Apertium[9] is a free open-source machine translation platform. Its translation engine consists of a series of assembled modules that communicate with each other using text streams. One of the modules, the lexical transfer module, uses a bilingual dictionary to deliver the corresponding target lexical forms from a given lexical form in the source language. Many of the Apertium bilingual dictionaries are available in the Lexical Markup Framework (LMF) XML-based format[10]. We took such LMF dictionaries as a starting point in our process to publish the Apertium data as LD.

We used *lemon* and its *translation module* as representation schemes. Once the conversion into RDF was completed we published the Apertium data on the Web in accordance with the LD principles. The result is a set of 22 Apertium RDF bilingual dictionaries, which can be found in the LLOD cloud[11]. The following languages are currently represented in Apertium RDF: Spanish, Catalan, English, French, Italian, Romanian, Asturian, Aragonese, Basque, Galician, Portuguese, Occitan, Esperanto. The whole dataset contains 400,808 translations and, after its conversion into RDF, 8,842,510 RDF triples were created. The LD version of Apertium groups the data of the (originally disparate) Apertium bilingual dictionaries in the same graph, interconnected through the common lexical entries of the monolingual lexicons that they share. Figure 2 shows the network of interconnected languages in Apertium RDF.

We made all the generated information accessible on the Web both for humans (via a Web interface[12]) and software agents (with a SPARQL endpoint[13]).

[6] http://lemon-model.net/

[7] http://purl.org/net/translation/

[8] http://openrefine.org/

[9] http://apertium.org/

[10] A complete list can be found at http://lod.iula.upf.edu/types/Lexica/by/standards/.

[11] http://linguistic-lod.org/, and see also back cover.

[12] http://linguistic.linkeddata.es/apertium/

[13] http://linguistic.linkeddata.es/apertium/sparql-editor/

## 4. Terminesp

Terminesp is a multilingual terminological database created by AETER (Asociación Española de Terminología)[14] that contains the terms and definitions from Spanish technological norms (standards) including more than thirty thousand terms, many of them with translations available into another language (English, French, German, Italian, Swedish). The terminology contained in Terminesp is highly technical and specific of domains such as electrical engineering, aeronautics, marine technology, etc.

We converted the original data (an MS Access database) into RDF by using the *lemon-ontolex* model as representation scheme. The *lemon-ontolex* model is the next version of *lemon*, developed under the umbrella of the W3C Ontology Lexica (Ontolex) community group[15]. At the time of writing, the *lemon-ontolex* model is nearly finished and waiting for final corrections by the community to be officially released.

We established an automatic mechanism to extract the lexical entries from Terminesp database and instantiate the *lemon* lexicons. Translation sets were created between Spanish to French (13,996 translations), German (12,593), English (14,936), Italian (802) and Swedish (67). Differently from the Apertium RDF graph, the Terminesp RDF graph follows a star topology (see Figure 3), with Spanish as hub and the other languages as peripheral nodes.

In addition to accounting for explicit translations, we extended the original Terminesp dataset with part-of-speech and syntactic information that was not explicitly declared in the original data (e.g. nominal, prepositional and adjectival phrases), as well as some terminological variations (Bosque-Gil et al 2015). The whole conversion into RDF resulted in 1,095,051 triples. Since *lemon-ontolex* is still under development, the resultant RDF files are not published as LD yet. However, a preliminary version of Terminesp RDF, restricted to Spanish, English and German, and with less rich syntactical information, was already published in October 2013 as LD using *lemon*[16].

## 5. The emergence of a unified single graph of translations

The publication of the Apertium dictionaries as LD resulted in the creation of a large unified graph of linked lexical entries, senses and translations on the Web. The URIs of all these elements can be seen as the nodes of this graph. Every URI is dereferenceable, meaning that when it is accessed, a response is obtained with its attributes and links to other elements in RDF. There are several ways to access and explore the graph (both for software agents and humans), such as by querying through the SPARQL endpoint, by using dedicated search interfaces[17], or by following the links as they are represented in an LD interface such as Pubby[18].

Some of the advantages of having all the lexical information and translations in the same RDF graph are:

- As we have seen above (Figure 1), monolingual lexicons grow with time

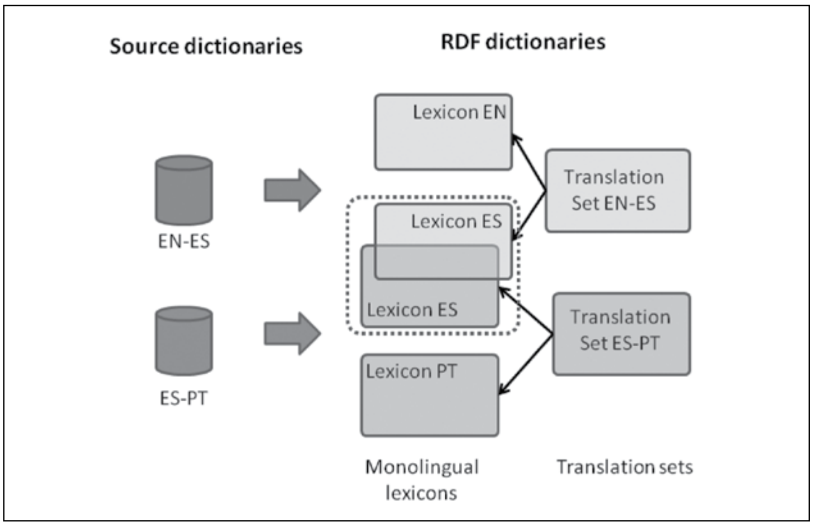

**Figure 1: Example of the conversion of two bilingual dictionaries (EN-ES and ES-PT) into RDF**

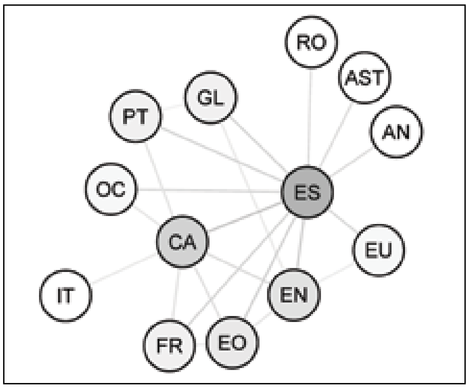

**Figure 2: Network of languages in the Apertium RDF Graph (nodes are languages and edges are translation sets)**

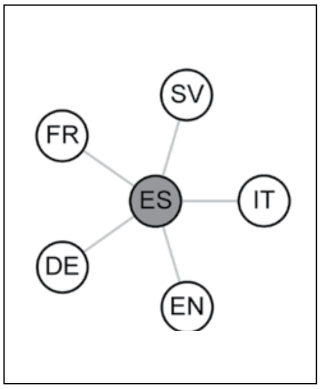

**Figure 3: Network of languages in the Terminesp RDF graph**

[14] http://aeter.org/

[15] https://w3.org/community/ontolex/

[16] http://linguistic.linkeddata.es/terminesp/

[17] http://linguistic.linkeddata.es/search/

[18] For example, one can introduce the URI of a translation set in a Web browser http://linguistic.linkeddata.es/id/apertium/tranSetEN-ES/, and its properties and links will be shown in a human readable way (as the links are clickable, "manual" navigation through the graph is possible).

### 1st Summer Datathon on Linguistic Linked Data (SD-LLOD-15)

The 1st Summer Datathon on Linguistic Linked Data (SD-LLOD-15) will be held in Cercedilla (Madrid, Spain) from 15 to 19 June 2015. It is co-organized by Jorge Gracia from Madrid Polytechnic University and John McCrae from Bielefeld University (Germany) as part of the LIDER project. Its main goal is to offer persons from the industry and academia practical knowledge in the field of Linked Data applied to linguistics, and the final aim is to allow participants to migrate their own (or other's) linguistic data and publish it as Linked Data on the Web. This datathon is the first organized on this topic worldwide and is supported by the LIDER FP7 Support Action. It will join around seventy participants (including attendees, speakers and tutors) from all around the world and will constitute an invaluable forum not only for learning but also for the exchange of experiences and ideas related to linguistic Linked Data.

http://datathon.lider-project.eu/

as more dictionaries of the same family are published as LD. There is no need to create a new and separate monolingual lexicon every time.

- Useful information can be obtained through a single SPARQL query in a manner that otherwise would be more difficult to get if the isolated original data sources must be queried. For instance, one can get all the possible direct translations of "network"@en into any available target language in the graph with just a single query (not needing to specify which dictionaries to look up), getting as a result the list of translated forms: {"xarxa"@ca, "red"@es, "rede"@gl, "reto"@eo, "sarera_konektatu"@eu, ...}.
- Indirect translations can be obtained between language pairs that were initially unconnected. For instance, a translation of the English term "network"@en to Italian can be obtained, with a single SPARQL query, by using Catalan as pivot language. The result is "rete"@it[19]. Notice however that some strategies have to be introduced in order to detect and exclude wrongly inferred translations. To that end we propose the use of the one time inverse consultation algorithm (Tanaka and Umemura 1994).
- Further, direct connections to other datasets in the Web of Data are possible so the original information can be enriched with additional relevant data. For instance, a high number of lexical senses in Apertium RDF have been linked to BabelNet (Navigli and Ponzetto 2012). In that way, additional descriptions, lexical relations, or even pictures, can be obtained by querying BabelNet to enrich the information that can be obtained in Apertium. For instance, one of the ontological references of "network"@en, when translated as "red"@es, is the babelsynset http://babelnet.org/rdf/s00030258n/i/, from which an English definition, not initially present in Apertium, could be obtained: "(electronics) a system of interconnected electronic components or circuits".
- Several "families" of bilingual dictionaries (Apertium and Terminesp in our case) can be published as LD under the same domain or default graph (http://linguistic.linkeddata.es/, in our case). In that way, we can have a common access point to all of them and unified SPARQL queries can be built to access these sub-graphs at the same time. For instance, a search for translations of "red"@es in Apertium could be extended with the languages covered by Terminesp, obtaining for example the German translation "Netz"@de, which is not available in Apertium originally.

In conclusion, generating linguistic Linked Data is a growing trend in the community of language resources, with clear advantages such as standardised ways of representing and accessing the data, the possibility of linking to other resources on the Web of Data, and enabling enhanced ways of discovering and aggregating the data. In this article we have briefly reported our recent experiences with the LD generation of the Apertium bilingual dictionaries and the Terminesp multilingual terminological database, and commented on the benefits of publishing their information as unified RDF graphs.

**Acknowledgement**

This work is supported by the FP7 European project LIDER (610782) and by the Spanish Ministry of Economy and Competitiveness (project TIN2013-46238-C4-2-R).

**References**

**Archer, P., Goedertier, S. and Loutas, N. 2012.** *Study on persistent URIs. Technical Report, Interoperability Solutions for European Public Administrations*. ISA

**Bizer, C., Heath, T. and Berners-Lee, T. 2009.** Linked data – the story so far. *International Journal on Semantic Web and Information Systems* (IJSWIS), 5.3: 1-22.

**Bosque-Gil, J., Gracia, J., Aguado-de Cea, G. and Montiel-Ponsoda, E. 2015.** Applying the OntoLex model to a multilingual terminological resource. In *Proceedings of 4th Workshop of the Multilingual Semantic Web) at 12th ESWC, Portoroz, Slovenia* (MSW'15, to appear). CEUR-WS.

**Gracia, J., Montiel-Ponsoda, E., Vila-Suero, D. and Aguado-de Cea, G. 2014.** Enabling language resources to expose translations as linked data on the web. In *Proceedings of the 9th Language Resources and Evaluation Conference, Reykjavik, (LREC'14)*. Paris: European Language Resources Association (ELRA): 409-413.

**Navigli, R. and Ponzetto, S.P. 2012.** BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193: 217-250.

**Tanaka, K. and Umemura, K. 1994.** Construction of a bilingual dictionary intermediated by a third language. In *COLING*: 297-303.

[19] Examples of queries in Apertium RDF can be found at http://dx.doi.org/10.6084/m9.figshare.1352066/.