“Methodology and tools for Multilingual Linguistic Linked Data generation”

Jorge Gracia, Daniel Vila-Suero
Ontology Engineering Group (OEG)
Universidad Politécnica de Madrid (UPM)
jgracia, dvila@fi.upm.es

EKAW Tutorial on Language Resources and Linked Data
Linköping (Sweden), 25th November 2014
Different methods and guidelines available:

• LOD2
• Datalift
• W3C Linked Data cookbook
• W3C Best Practices for Linked Data
• Guidelines for Multilingual Linked Data
• W3C Best Practices for Multilingual Linked Open Data (BPMLOD) community group Get Involved!
• We will use a particular use case: “bilingual dictionaries” as running example (although the proposed methodology is general)
• W3C BPMLOD community group draft report: “Guidelines for Linguistic Linked Data Generation: Bilingual Dictionaries”
• Available at: http://bpmlod.github.io/report/bilingual-dictionaries/index.html
Main activities:

1. Analysis of data sources
2. Modelling
3. URI/IRI design
4. Generation
5. Publication

Each activity composed of several tasks
Main activities:

1. Analysis of data sources
2. Modelling
3. URI/IRI design
4. Generation
5. Publication
The goal is to:

• Specify and analyse the data sources in order to plan and manage the subsequent activities

• Main aspects to specify are:
  – Format
  – Identifiers structure
  – Access methods: file, webservice, etc.
  – Data models: Standards, terminologies, etc.
  – Language representation: how languages are tagged, represented, etc.
  – License and provenance: existing license of data sources
Documentation of data sources:

- Type of data: *Bilingual dictionary (English and Spanish)*
- Data model: *LMF (Lexical Markup Framework)*
- Format: XML files
- License: GPL 3.0
- Provenance: Apertium EN-ES
- ....
Analysis of data sources EXAMPLE

```
<Lexicon>
    <feat att="language" val="en"/>
    ...
    <LexicalEntry id="bench-n-en">
        <feat att="partOfSpeech" val="n"/>
        <Lemma>
            <feat att="writtenForm" val="bench"/>
        </Lemma>
        <Sense id="bench_banco-n-l"/>
    </LexicalEntry>
    ...
</Lexicon>
```
• Main activities:
  1. Analysis of data sources
  2. Modelling
  3. URI/IRI design
  4. Generation
  5. Publication
Modelling tasks

1. Analysis and selection of domain vocabularies
2. Mapping of data sources and vocabularies
3. Vocabulary for representing licensing and provenance information
Modelling

Analysis of vocabularies

Use http://lov.okfn.org/

NLP Interchange Format

DCAT

PROV-O

Dublin Core
Modelling

Translation Module

http://purl.org/net/translation.owl

TranslationSet

trans

Translation

translationSource

LexicalSense

translationTarget

translationCategory
double

translationConfidence

Resource

culturelEquivalent

directEquivalent

context

culturalEquivalent

context

lexicalEquivalent

Translation Categories

http://purl.org/net/translation-categories

25/11/2014

Jorge Gracia, Daniel Vila-Suero
Modelling EXAMPLE

Mapping of data sources

EN-ES dictionary

EN-ES RDF dictionary

Lexicon ES

Lexicon EN

Translation Set EN-ES

trans. module

EN-ES
Modelling EXAMPLE

"bench"@en

lemon:writtenRep

lemon:Lexicon
lexiconEN

lemon:Form

lemon:LexicalSense

lemon:isSenseOf
tr:translationSource

tr:Translation

tr:trans
tr:transTarget

translationSetEN-ES

lemon:LexicalSense

lemon:isSenseOf

lemon:LexicalEntry

lemon:entry

lemon:Lexicon
lexiconES

lemon:Form

lemon:writtenRep

"banco"@es
Main activities:

1. Analysis of data sources
2. Modelling
3. URI/IRI design
4. Generation
5. Publication
The goal is to:

- Define **URI/IRI patterns and namespaces to be used**
- Ensure that LD best practices are followed
Some good practices...

1. Define **namespace(s)** (that you own or have control over).

2. Define how to create the **ID of resources** (reuse original data source keys if possible)

3. Define the structure of the **URI space** to organize the resources in different addresses and **avoid collision**.

**Useful guidance at:**

- ISA - *Study on persistent URIs Archer et al.,*
- *Linked Data patterns book online → URI patterns*
Following ISA* recommendations:

```
http://{domain}/{type}/{concept}/{reference}
```

where:

- **{type}**: a value from the set of type of resources, examples are 'id' or 'item' for real world objects; 'doc' for documents that describe those objects; 'def' for concepts; 'set' for datasets

*ISA - Study on persistent URIs, Archer et al.*
http://{domain}/{type}/{concept}/{reference}

{domain}: http://linguistic.linkeddata.es/
{type}: id (real-world object)
{concept}: apertium
{reference}: resource ID

# Apertium English lexicon:
http://linguistic.linkeddata.es/id/apertium/lexiconEN

# Apertium Spanish lexicon:
http://linguistic.linkeddata.es/id/apertium/lexiconES

# Apertium English-Spanish translation set:
http://linguistic.linkeddata.es/id/apertium/tranSetEN-ES
• **Set of main activities:**

1. Analysis of data sources
2. Modelling
3. URI/IRI design
4. Generation
5. Publication
1. Selection, extension or development of technologies for RDF generation
2. Mapping of data sources to RDF
3. Transformation of data sources to RDF
Goal:

```turtle
apertium:lexiconEN a lemon:Lexicon ;
    dc:source <http://hdl.handle.net/10230/17110> .
...
apertium:lexiconEN lemon:entry
apertium:lexiconEN/bench-n-en .

apertium:lexiconEN/bench-n-en a lemon:LexicalEntry ;
    lemon:lexicalForm apertium:lexiconEN/bench-n-en-form ;
    lexinfo:partOfSpeech lexinfo:noun .

apertium:lexiconEN/bench-n-en-form a lemon:Form ;
    lemon:writtenRep "bench"@en .
```
The RDF schema alignment tool allows you to specify how RDF data will be generated from grid-shaped data. Each cell in each record of your data will get placed into nodes within the schema. Configure the schema by specifying which column to substitute into which node.

Base URI: http://linguistic.linkeddata.es/id/apertium/edit

Add another root node

Available Prefixes:
- dc: dctext
- lexinfo: partOfSpeech
- lemon: Lexicon
- rdf: type
- http://hdl.handle.net/10230/17110
- es

Add prefix
Manage prefixes
Main activities:

1. Analysis of data sources
2. Modelling
3. URI/IRI design
4. Generation
5. Publication
• The goal is to:
  – Make available the RDF dataset following Linked Data best practices
  – Facilitate dataset discovery and consumption
Vocabulary for licensing and provenance

→ INPUT:
  - Documentation of data sources (licensing and provenance)

OUTPUT →
  - Selection of standard vocabs

- cc creative commons
- ODRL Open Digital Rights Language
- PROV W3C Provenance Ontology
Add "rights" metadata in the dataset description (e.g., VoID, DCAT)

Use standard predicates to declare "rights" statements (e.g., Dublin Core terms: dc:rights, dct:license)

Standard license available

Yes

3a Use URI of standard license e.g., CC0

No

3b Use rights declaration language, e.g., ODRL
Dataset and vocabulary publication on the Web

**CONFIGURATION FILE**
- Location of the RDF data
- Define access methods
- and even the presentation of the data

**HTTP**
**SPARQL QUERY LANGUAGE**

**LD FRONTEND**  
**SPARQL ENDPOINT**

**SPARQL STORE**
Metadata definition and publication using DCAT

1) Register dataset in datahub.io
2) Extend generated DCAT file and link to datahub.io one
Extending DCAT description

```xml
  <dct:source rdf:resource="http://hdl.handle.net/10230/17110"></dct:source>
  <rdfs:seeAlso rdf:resource="http://dbpedia.org/resource/Apertium"></rdfs:seeAlso>
  <rdfs:seeAlso rdf:resource="http://purl.org/ms-lod/UPF-MetadataRecords.ttl#Apertium-en-es_resource-5v2"></rdfs:seeAlso>
</dcat:Dataset>
```
• Loading the RDF data into a **SPARQL** endpoint **not enough for publishing LD:**
  – Why? We provide a queryable repository, but **URIs are not de-referenceable**

• **We need a mechanism to make our URIs de-referenceable:**
  – Through a common web server (as files)
  – **Linked Data front-ends:**
    • Pubby
    • More sophisticated: LD APIs (Puelia, Elda)
Conclusions

• Documentation of data sources and issues
• Language issues have to be taken into account during the whole process
• Metadata description is key for enabling reusing and discovery
• Vocabulary have to be documented and published following LD BPs