Guidelines for Multilingual Linked Data generation and publication

Jorge Gracia, Daniel Vila-Suero
jgracia, dvila@fi.upm.es
ISWC Tutorial “Building the Multilingual Semantic Web”, Trentino (Italy)
20th October 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!
• Report: Guidelines for Linguistic Linked Data Generation: Bilingual Dictionaries

• Available at:

• We will use this use case (bilingual dictionary) as **running example** but guidelines are general.
Guidelines for ML Linked Data

• Set of main activities:
  1. Analysis of data sources
  2. Modelling
  3. URI/IRI design
  4. Generation
  5. Publication

• Each activity composed of several tasks
Guidelines for ML Linked Data

• Set of 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 following activities.

– Important 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
  – ....
<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>
  ...

• **Set of main activities:**

1. Analysis of data sources
2. Modelling
3. URI/IRI design
4. Generation
5. Publication
1. Analysis and selection of domain vocabularies
2. Mapping of data sources and vocabularies
3. Vocabulary for representing licensing and provenance information
Modelling

Analysis and selection of domain vocabularies

Use http://lov.okfn.org/

NIF
NLP Interchange Format

LexInfo

Wordnet
Translation Module

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

Translation Set

Translation

Lexical Sense

Translation Category

directEquivalent

culturalEquivalent

lexicalEquivalent

Resource

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

The translation model

The translation model
EXAMPLE

Mapping of data sources

EN-ES dictionary

EN-ES RDF dictionary

Lexicon ES

Lexicon EN

Translation Set EN-ES

trans. module

lemon
Example: Modelling
Mapping of data sources in detail

```
"bench"@en

lemon:Lexicon
  lexiconEN

lemon:Form
  lemon:writtenRep
  lemon:lexicalForm
  lemon:entry
  lemon:LexicalEntry

lemon:LexicalSense
  lemon:isSenseOf
  tr:translationSource
  tr:Translation
  tr:TranslationSet
  translationSetEN-ES
  lemon:LexicalSense
  tr:trans
  tr:translationTarget

lemon:Entry
  lemon:Lexicon
  lexiconES

lemon:Form
  lemon:writtenRep
  "banco"@es
```
• Set of 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
Guidelines for ML Linked Data

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

```rdfs
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 skeleton below specifies how the RDF data that will get generated from your grid-shaped data. The cells in each record of your data will get placed into nodes within the skeleton. Configure the skeleton by specifying which column to substitute into which node.

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

### RDF Schema Alignment

<table>
<thead>
<tr>
<th>Available Prefixes:</th>
<th>do rdfs lexinfo foaf owl rdf lemon</th>
<th>add prefix</th>
<th>manage prefixes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>LexicalEntry-id URI</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lemon:LexicalEntry</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>add rdf:type</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>X &gt; lexinfo : partOfSpeech</td>
<td></td>
<td></td>
</tr>
<tr>
<td>X &gt; lemon : lexicalForm</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X &gt; df : source</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lexiconES</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lemon:Lexicon</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>add rdf:type</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>X &gt; lemon:entry</td>
<td></td>
<td></td>
</tr>
<tr>
<td>X &gt; lemon : language</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>add property</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><a href="http://hdl.handle.net/10230/17110">http://hdl.handle.net/10230/17110</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>add rdf:type</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>LexicalEntry-POS URI</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lemon:Form</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>add rdf:type</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>X &gt; lemon : writtenRep</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>LexicalEntry-writtenFo</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>add property</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Add another root node*
• Set of 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 LD best practices
- Facilitate dataset discovery and consumption
Vocabulary for licensing and provenance

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

OUTPUT:
- Selection of standard vocabularies

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**
  - Use URI of standard license, e.g., CC0

- **No**
  - 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

LD FRONTEND

SPARQL STORE

SPARQL QUERY LANGUAGE

SPARQL ENDPOINT
How:

1) Register dataset in datahub.io
2) (Extend generated DCAT file and link to datahub.io one)
http://datahub.io/dataset/apertium-en-es

  <owl:sameAs rdf:resource="http://datahub.io/dataset/apertium-en-es"></owl:sameAs>

  <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 an 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