Best Practises for Multilingual Linked Open Data: a Community Effort

Jorge Gracia  
Ontology Engineering Group (OEG)  
Universidad Politécnica de Madrid (UPM)  
jgracia@fi.upm.es

Jose Labra  
Web Semantics Oviedo (WESO)  
University of Oviedo  
labra@uniovi.es

Multilingual Web Workshop  
Madrid (Spain)  
7-8 May 2014
Outline

- Motivation
- The group
- Main goals
- Activities
- Where are we now?
Motivation
The Web of Data is increasingly multilingual

Standard Semantic Web techniques (e.g. language tagging) are however underused.

Lots of **design decisions** have to be made when publishing/consuming/linking multilingual linked data.

For instance...
“How do I publish my data on the Multilingual Web of Data?”

Vocabulary selection → RDF generation → Data Interlinking → Web Publishing
“How do I identify ‘things’ on the Multilingual Web of Data?”

http://example.org/Spain

http://example.org/I23AX45

http://example.org/España
“How do I create a Linked Data version of my bilingual dictionary?”
“How do I localise an existing **ontology** in my own language”
The group
The group

W3C community group on **Best Practises for Multilingual Linked (Open) Data**

https://www.w3.org/community/bpmlod

Started on June 2013
bi-weekly telcos
3 chairs. Currently:

José Labra  Jorge Gracia  John McCrae

67 members from academia and industry
and many others…
Main goals
Main goals

Crowdsourcing ideas from the community regarding **best practices** to produce **multilingual** linked (open) data.

**Documenting** patterns and best practices for the creation, linking, and use of multilingual linked data.
Relation to other W3C groups

Linked Data for Language Technologies (LD4LT)
Use Cases
BP for LD in LT
BP for Data on the Web
BP for Multilingual Data on the Web

BPMLOD

Ontology lexica (Ontolex)
lemon specification
BP for using lemon

Data on the Web Best Practices
Activities

TOPIC classification

USE CASES

PATTERNS

BEST PRACTISES & GUIDELINES
TOPIC classification
Topics

- **Naming**
  Opaque URIs, Descriptive URIs, IRIs, ...

- **Textual information**
  Language tags, linguistic information, ...

- **Linking**
  Interlanguage links, owl:sameAs, ...

- **Ontologies and vocabularies**
  Mono/multilingual vocabularies, ontology localisation...

- **Quality of MLOD**

- **Tools and examples of MLOD**

- **Other related aspects**
  Licensing, legal aspects, ...

https://www.w3.org/community/bpmlod/wiki/Topic_classification
TOPIC classification

USE CASES
Use cases

USE CASES
1. Localization workflow [D. Lewis]
2. Lexicalisation of RDF Datasets [E. Montiel, G. Dunshire]
3. Ontology localisation [E. Montiel, L. Aguado, G. Dunshire]
4. Crosslingual linked data matching [J. Gracia]
6. Application localization [J. McCrae]

CASE STUDIES
1. Translations of multilingual terminologies for libraries [G. Dunshire]

https://www.w3.org/community/bpmlod/wiki/Use_cases_definition
TOPIC classification

USE CASES

PATTERNS
Difficult to establish a boundary between Patterns vs Best Practices vs Bad smells

By now: we identify the main practices
Bad/Good may depend on the context/use case

Examples:
- Patterns for naming and dereferencing
Patterns for Naming

Example: URI for Armenia?

Descriptive URIs

- http://example.org/Armenia

Human-readable
Good tool support
Independence between concept and language
Maintenance: changes in text don't affect URI
Readability (for one language)

May be unreadable for non-Latin alphabet users
Difficult to be descriptive enough in some contexts
Non-human-readable
Difficult to handle by developers
Security issues (spoofing)
Unreadable for speakers of other languages
Unreadable for speakers of other languages

Language in Path

- http://en.example.org/Armenia
- http://example.org/Armenia.en
- http://example.org/en/Armenia
- http://example.org/Armenia?lang=en

Practical reasons
Independent development of datasets by language
Compatible with content negotiation

Dialects

- hy-Latin-IT-arevela

Where should we put the language tag?
Dialects can become unwieldy
Example: languages & sublanguages

- http://example.org/Armenia?lang=en

Multilingual Web Workshop Madrid, May 2014
Patterns for dereferencing

Which data should I return when accessing a URI?

No language content negotiation

- Ignore Accept-language...all the data
- Clients have to filter triples in other languages

- Easy to develop
- Consistency of data
- Bandwidth overhead

Language content redirection

- Difficult to implement
- Loses data
- More difficult to implement
- Not always feasible

Keep track overhead

- Language content redirection
- Language content negotiation

Keep track overhead

http://example.org/Armenia

http://example.org/Armenia

http://example.org/Armenia

http://example.org/Armenia

http://example.org/Armenia

http://example.org/Armenia

http://example.org/Armenia
TOPICS classification

USE CASES

PATTERNS

BEST PRACTISES & GUIDELINES
Best practices and guidelines
Some (future) EXAMPLES. Guidelines for:

- Linguistic Linked Data generation
- RDF and Ontology translation
- Multilingual Linked Data generation, publication and exploitation

...
Where are we now?
TOPICS classification

USE CASES

PATTERNS

BEST PRACTISES & GUIDELINES

We are here (Patterns for textual information)
Thanks… and get involved!

Next telco: Thursday 22nd May 10:00 CEST

https://www.w3.org/community/bpmlod