



SESSION ON

MULTILINGUAL VOCABULARY

Development and extension

Daniel Vila-Suero (dvila@fi.upm.es)
Ontology Engineering Group, UPM (Madrid)

ACKNOWLEDGEMENTS:

BabelData Project (TIN2010-17550),
Elena Montiel-Ponsoda, Asunción Gómez Pérez

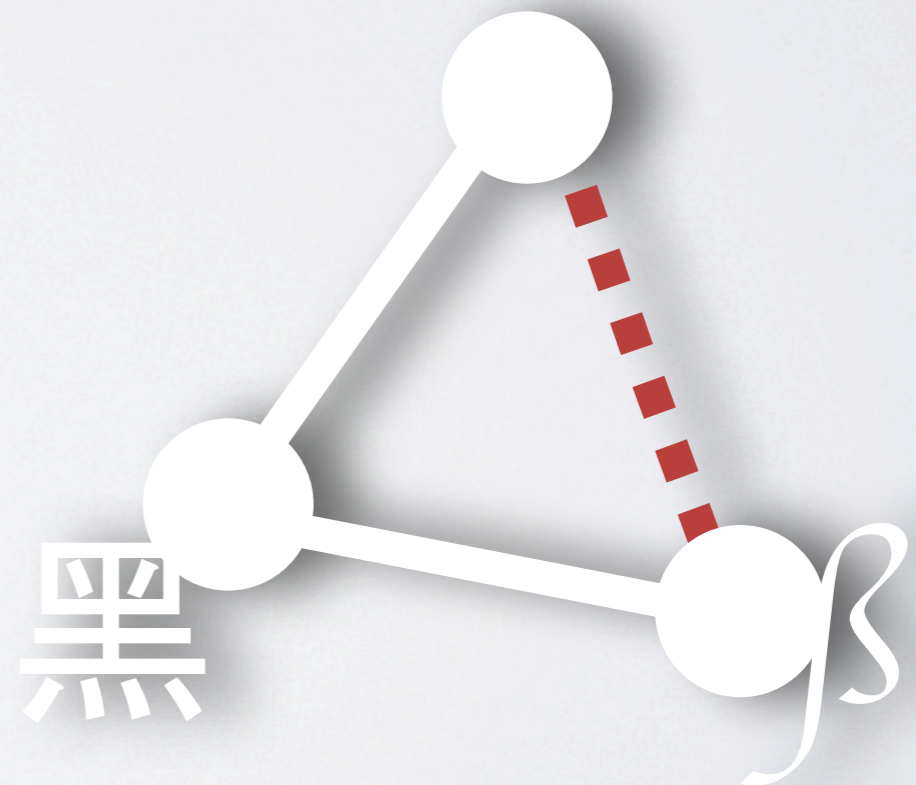
GOAL

Open discussion around **vocabularies** enabled **for multilingual environments (WWW)**

Introduce some **examples**: current situation and efforts.

More open **questions** than answers.

Promote **collaboration**



SESSION OUTLINE

OPEN DISCUSSION

1. Introduction to the session and the topic
2. *“Representing multilingual lexical and terminological information in RDF vocabularies”*
Elena Montiel-Ponsoda, OEG-UPM
3. *“Metadata registry of the Publications office of the EU”*
Michael Düro. PO-EU

OPEN DISCUSSION

THIS TALK

1. **Why** should we care about multilingual vocabularies?
2. **What** is a multilingual vocabulary?
3. Current situation: **when** and **who**

WHY

The primary design principle underlying the Web's usefulness and growth is **universality**. When you make a **link**, you can **link to anything**. That means people must be able to put **anything on the Web**, no matter what computer they have, software they use or **human language they speak...**

Tim Berners-Lee



WHY

The primary design principle underlying the usefulness and growth of the Web is **uniformity**. If you have a **link**, you can **link** to it. People must be able to find what they want, no matter what **language** they use or **human** they speak...

Vocabularies are becoming a central part of the WWW

Tim Berners-Lee

黒



LANGUAGES ARE USEFUL

- For **Humans**

- ★ Finding vocabularies, terms, etc.
- ★ Understanding their semantics, how to use them
- ★ ...

- and **Machines...**

- ★ Search, ranking, resource discovery
- ★ Natural Language Processing applications: multilingual question answering, localized presentation of data
- ★

WHY

Linked Open Vocabularies (LOV)

atures gives you the possibility to search for an existing element (property, class or vocabulary) in the
aries Catalogue.

oint and metrics about the use of vocabularies in the Semantic Web are used to bring you some relevant



Search for
プロジェクト

24 results in 1 vocabulary

doap:Project (rdfs:Class)

rdfs:label プロジェクト @ja

rdfs:comment プログラミングのプロジェクト @ja

doap (voaf:Vocabulary)

dce:description プロジェクトの説明の語彙(DOAP)。W3C RDF... @ja

doap:helper (rdf:Property)

rdfs:comment このプロジェクトの貢献者 @ja

WHY

24 ranked results
including the term
project in Japanese

24 results in 1 vocabulary

doap:Project (rdfs:Class)

rdfs:label プロジェクト @ja

rdfs:comment プログラミングのプロジェクト @ja

score:0.704



doap (voaf:Vocabulary)

score:0.363



dce:description プロジェクトの説明の語彙(DOAP)。W3C RDF... @ja

doap:helper (rdf:Property)

score:0.298



rdfs:comment このプロジェクトの貢献者 @ja

doap:category (rdf:Property)

score:0.227



rdfs:comment このプロジェクトの分類。 @ja

doap:release (rdf:Property)

score:0.216



rdfs:comment このプロジェクトのリリース @ja

doap:developer (rdf:Property)

score:0.197



rdfs:comment プロジェクトのソフトウェアの開発者 @ja

doap:wiki (rdf:Property)

score:0.183



rdfs:comment このプロジェクトの討論用ウィキ @ja

doap:documenter (rdf:Property)

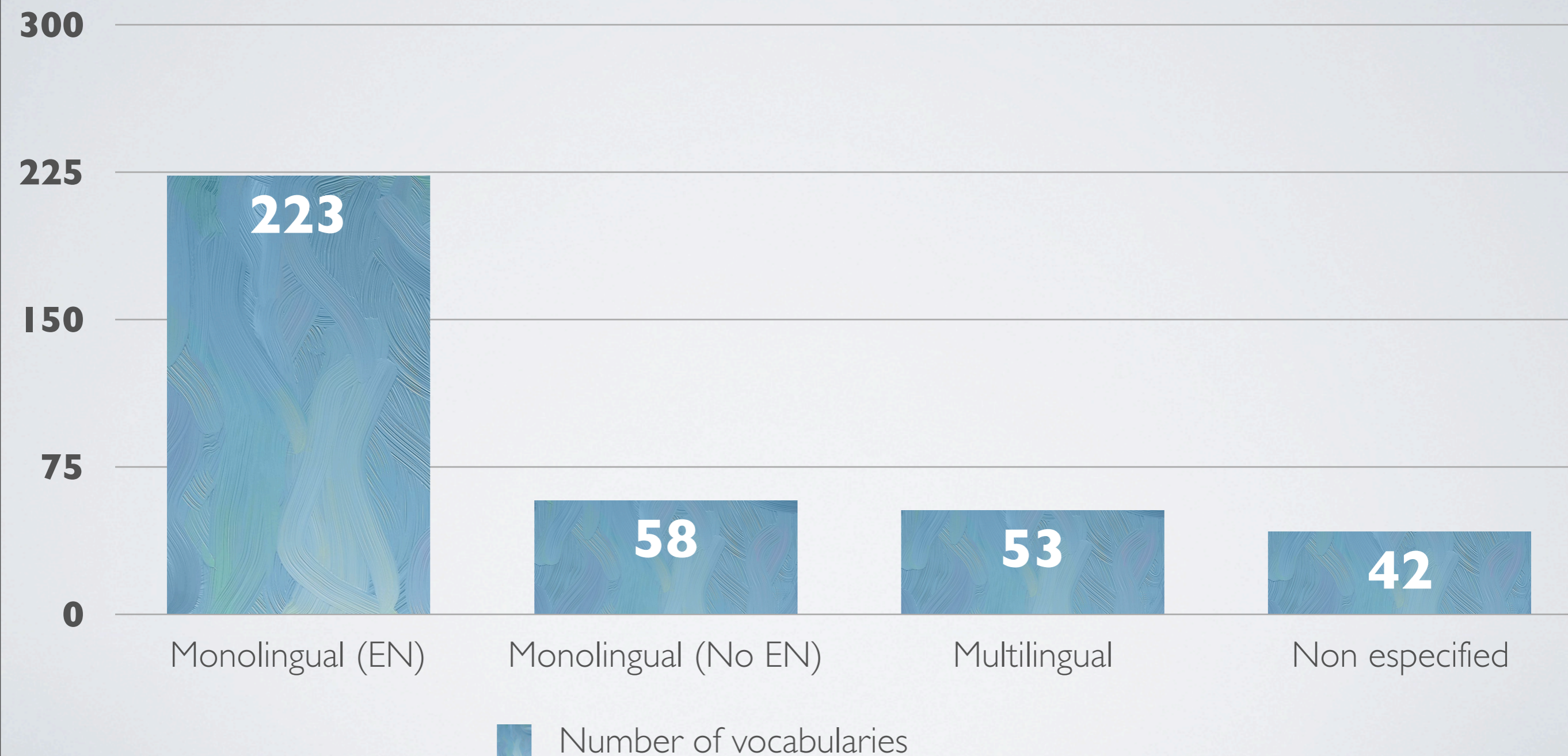
score:0.172



rdfs:comment このプロジェクトのドキュメントの貢献者 @ja

SOME FACTS ABOUT LOV

- Data retrieved 12.04.2013* out of 326 vocabs



* "Guidelines for Multilingual Linked Data" Gómez-Pérez et al., 2013

SOME FACTS ABOUT LOV

- LOV loves multilingual descriptions: indexing, ranked search results.
- But, still very **low usage of language tags** for vocabulary elements **< 60%**
- Other semantic search engines (Sindice, Falcons, SWSE..) lack support for multiple languages

WHAT IS AN ML VOCAB?

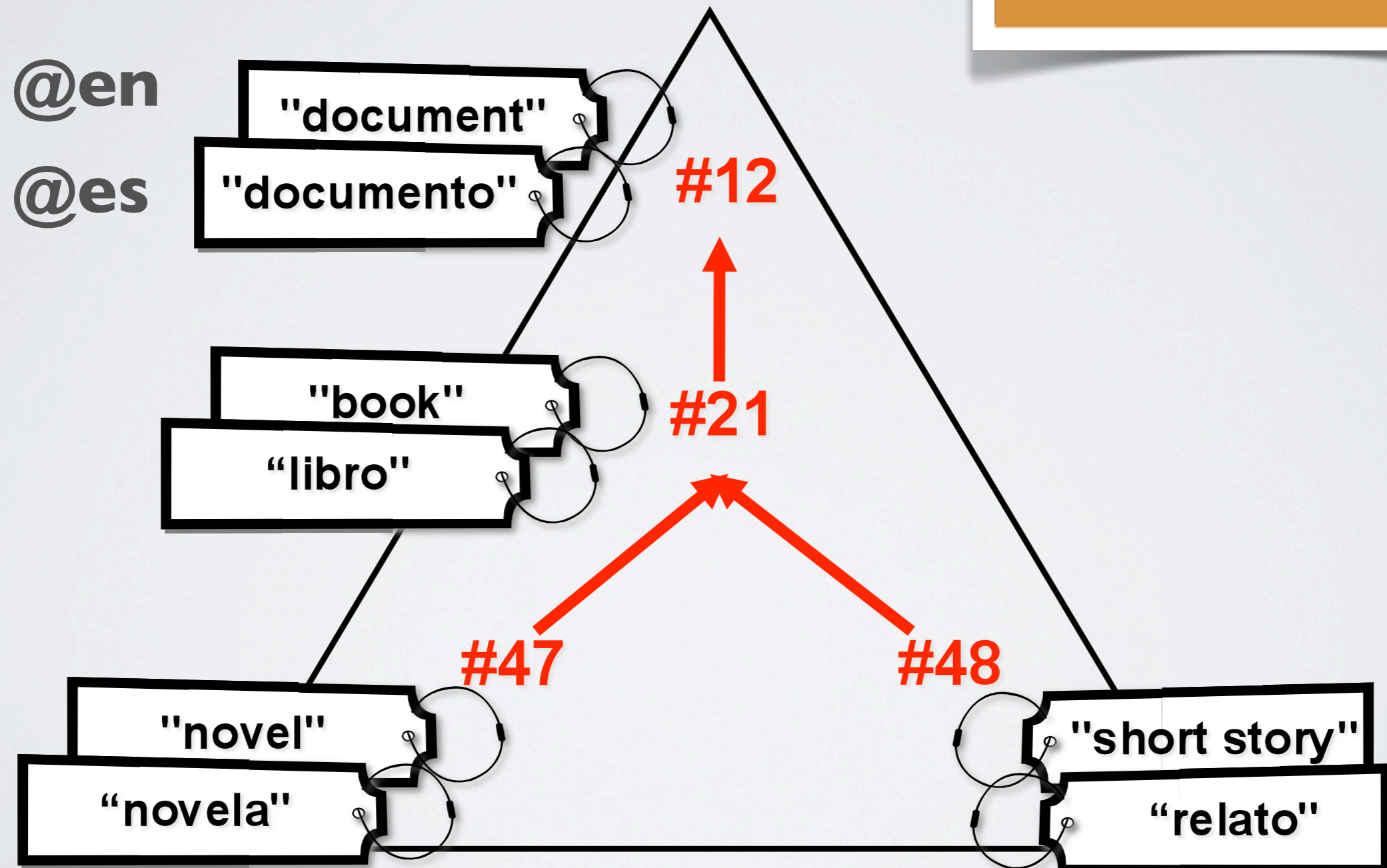
- Simple (general) answer:

“
A vocabulary which includes **labels and**
”
documentation in multiple languages”

- Are there **other flavors** of multilingual vocabularies?

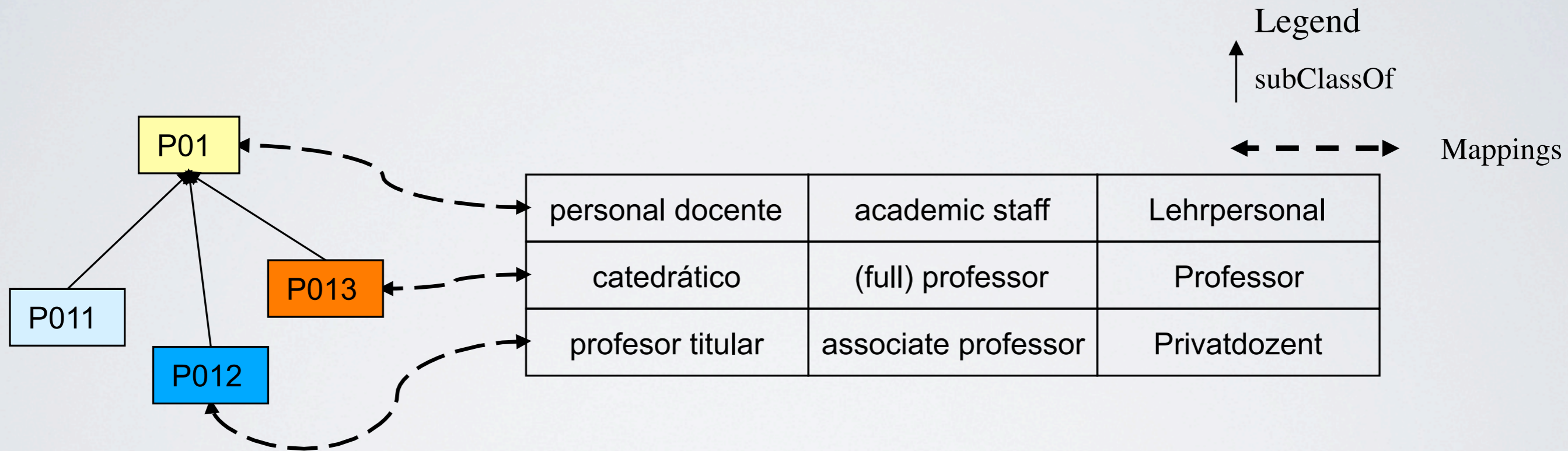
FLAVORS

I. LABELLING



FLAVORS

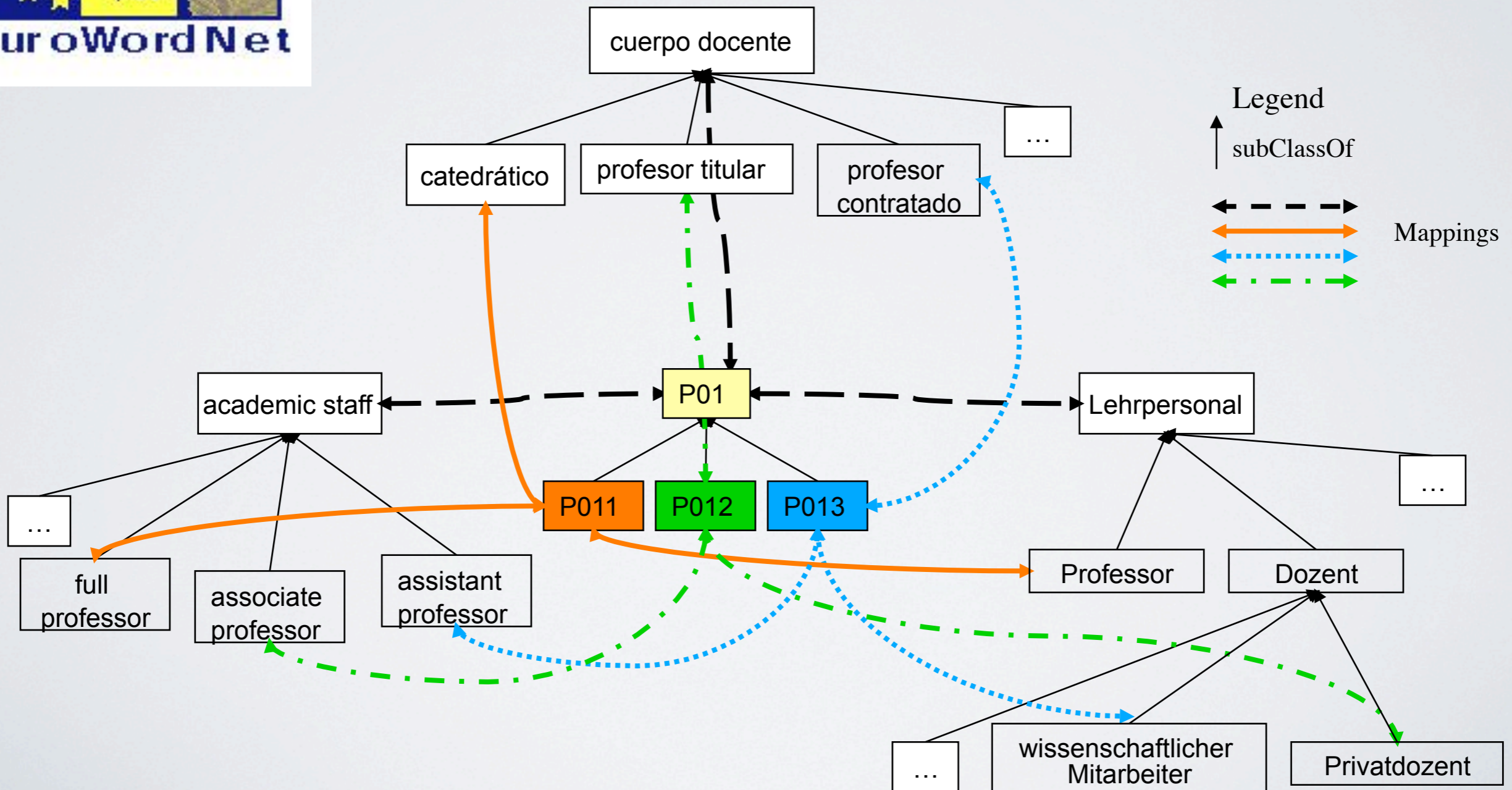
2. EXTERNAL MODEL



ISSUES:

Directionality of links, different namespaces, resolution of URIs (at http level with header, htaccess, external service..-)

3. MAPPING MODEL



WHAT FLAVOR IS MINE ?

WHAT FLAVOR IS MINE?

- Depends on a number of factors:
 - ★ Your starting point (starting from scratch? can you modify the terms within your original namespace? are there similar vocabularies in other langs?)
 - ★ Your needs (linguistically complex model, simplicity, efficiency, et)
 - ★ Your available resources (time, people, money...)
 - ★
- Selection should be **USE CASE DRIVEN**

LAYERED FRAMEWORK

TECHNICAL



INFRASTRUCTURE

REPRESENTATION

PROCESS

POLICY

ORGANIZATIONAL

POLICY

POLICY

★ Vocabulary publishers should commit to a **translation policy**:

e.g., What are the protocols for including/developing/validating a new translation?

★ Establish the necessary mechanisms to **manage** and **assess the quality, synchronization** and appropriate **coverage** between different languages.

★ Again, should be based on **requirements**, goals, etc. and be UC driven

PROCESS

PROCESS

★ Translation **workflows**: versioning, notification, edition, validation mechanisms, etc.

★ Develop **methodologies, guidelines and best practices** for translating and including new languages.

★ Establish **communication protocols** between the responsables of the different translations (languages)

★ Coordination among the **people** involved

REPRESENTATION

REPRESENTATION

- ★ Choose your modelling approach:
 - ★ rdfs and skos labels and descriptions
 - ★ Specialized models (lemon, ontolex etc.)
 - ★ Mappings
- ★ Guidelines for:
 - ★ **Naming**: coining new URIs for terms
 - ★ **Labeling**: Defining the structure of the labels (should we use verbs, full sentences, etc.)

INFRASTRUCTURE

INFRASTRUCTURE

★ Manage different aspects:

★ **Management** of translation/edition workflows: notifications, review process, versioning, etc.

★ **Access** to vocabulary elements: localize access? different namespace for the linguistic descriptions?

★ Generation of human-readable **documentation**

★ Look at **MLOD patterns** and guidelines

WHEN AND WHO

- Learn from (successful) initiatives:
 - ★ FAO's AGROVOC
 - ★ EUROVOC
 - ★ WORDNET
 - ★ IFLA Vocabularies and Guidelines for translations
 - ★
- Get involved in initiatives around the topic:
 - ★ W3C Internationalization Activity
 - ★ W3C Best practices for Multilingual LOD CG
 - ★ W3C Ontology-Lexica CG
 - ★ EU Lider project

W3C BPMLOD

INFRASTRUCTURE

REPRESENTATION



W3C Community and Business Groups

CURRENT GROUPS

REPORTS

ABOUT

 Mailing List

 Wiki

 Chat

 RSS

 Contact Group

Home / Best Practices for...

Best Practices for Multilingual Linked Open Data Community Group

The target for this group is to crowd-source ideas from the community regarding best practises for producing multilingual linked open data. The topics for discussion are mainly focused on naming, labelling, interlinking, and quality of multilingual linked data, among others. Use cases will be identified to motivate discussions. Participation both from academia and industry is expected. The main outcome of the group will be the documentation of patterns and best practices for the creation, linking, and use of multilingual linked data.

This group will not create specifications.

Use cases wanted!

REPRESENTATION




W3C Community and Business Groups

CURRENT GROUPS

REPORTS

ABOUT

 Mailing List

 Wiki

 RSS

 Contact Group

Home / Ontology-Lexica Community...

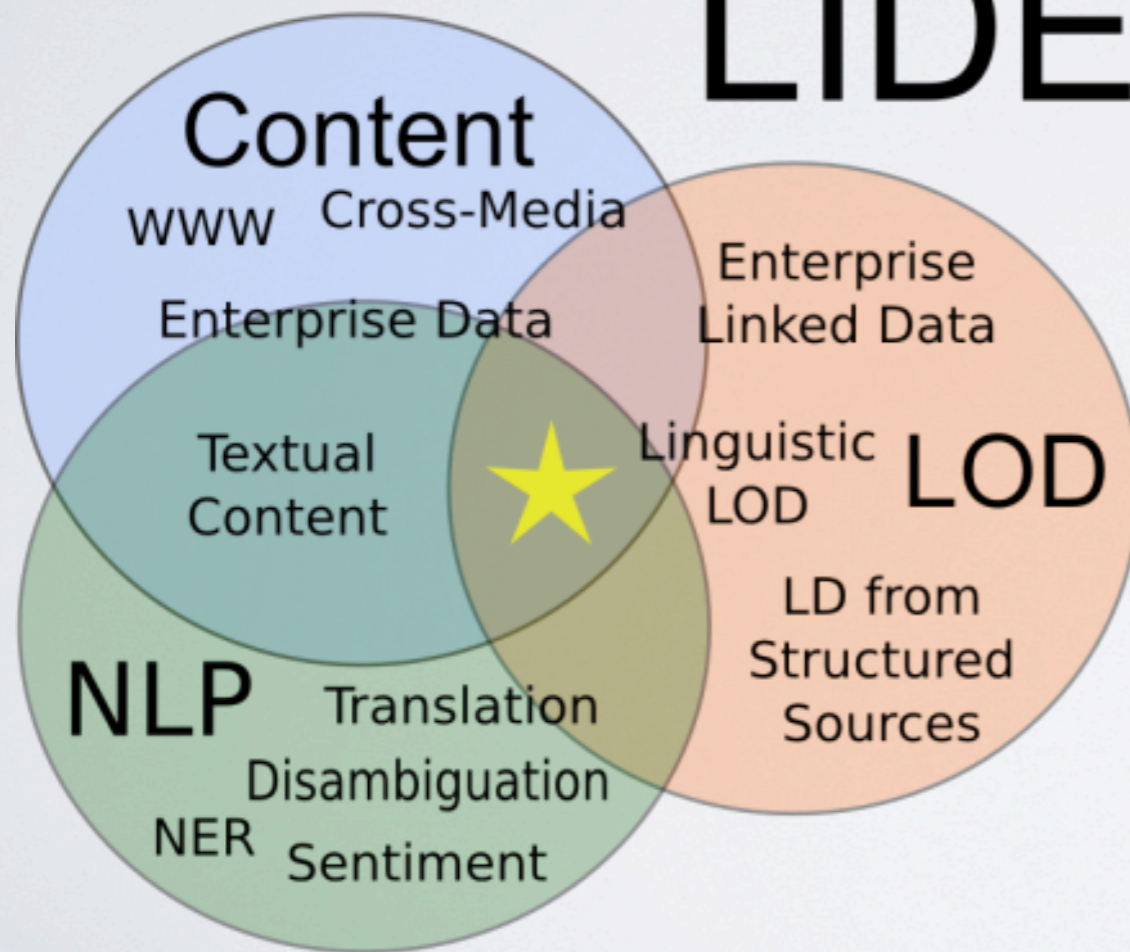
Ontology-Lexica Community Group

The mission of the Ontology-Lexicon community group is to: (1) Develop models for the representation of lexica (and machine readable dictionaries) relative to ontologies. These lexicon models are intended to represent lexical entries containing information about how ontology elements (classes, properties, individuals etc.) are realized in multiple languages. In addition, the lexical entries contain appropriate linguistic (syntactic, morphological, semantic and pragmatic) information that constrains the usage of the entry. (2) Demonstrate the added value of representing lexica on the Semantic Web, in particularly focusing on how the use of linked data principles can allow for the re-use of existing linguistic information from resource such as WordNet. (3) Provide best practices for the use of linguistic data categories in combination with lexica. (4) Demonstrate that the creation of such lexica in combination with the semantics contained in ontologies can improve the performance of NLP tools. (5) Bring together people working on standards for representing linguistic information (syntactic, morphological, semantic and pragmatic) building on existing initiatives, and identifying collaboration tracks for the future. (6) Cater for interoperability among existing models to represent and structure linguistic information. (7) Demonstrate the added value of applications relying on the use of the combination of lexica and ontologies.

LIDER-PROJECT.EU

Linguistic Linked Data (including **vocabularies**) can serve as an enabler technology for content analytics on the Multilingual Web.

LIDER



Universidad Politécnica de Madrid (Spain)



Trinity College (Ireland)



DFKI (Germany)



National University of Ireland , Galway (Ireland)



Institut für Angewandte Informatik (Germany)



Universität Bielefeld (Germany)



Università Roma la Sapienza (Italy)



W3C/ERCIM (France)

LIDER-PROJECT.EU

- Development of best practices and guidelines for publishing multilingual linked data resources (including vocabularies).
- Events: W3C Multilingual Web workshop, hackathons, industrial events, etc.
- **Help organizations** with publishing Multilingual Linked Data resources

Get involved!

THANK YOU VERY MUCH

dvila@fi.upm, [@dvilasuero](https://twitter.com/dvilasuero)