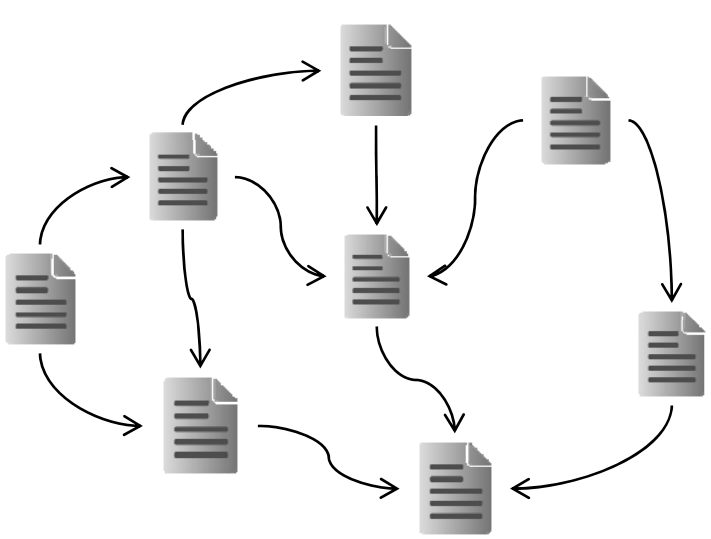
Best Practices for Multilingual Linked Open Data

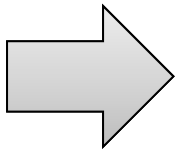
Jose Emilio Labra Gayo

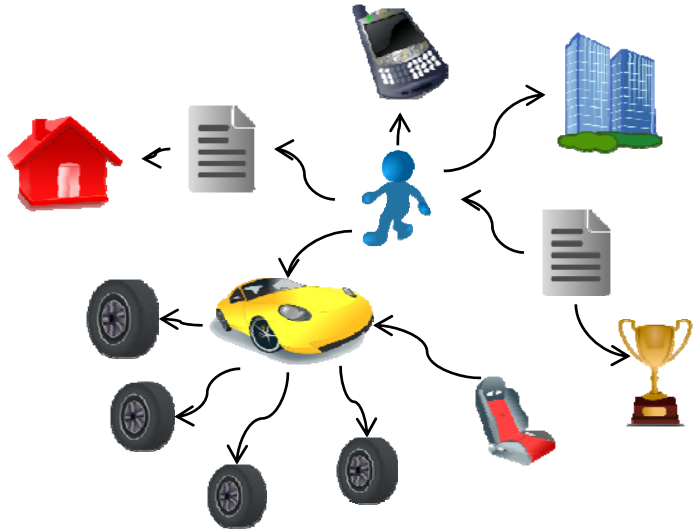
University of Oviedo, Spain

http://www.di.uniovi.es/~labra

Towards the web of data







Web of documents

Unit of information: Web page (HTML)
Human readable
Challenge: Multilingual pages

Web of Data

Unit of information: data (RDF)
Machine readable
Intrinsically Multilingual

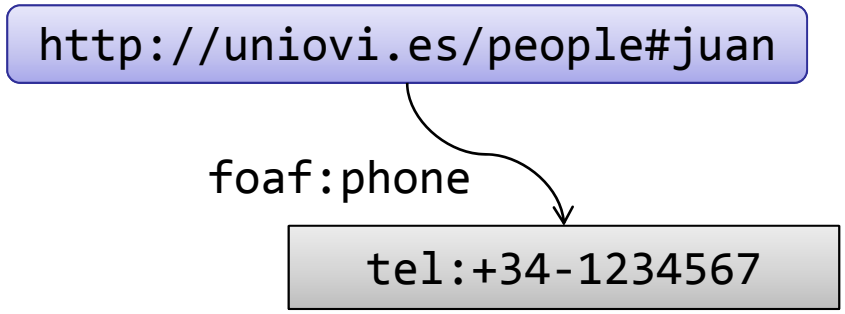
Example

English

```
<html lang="en">
<body>
<h1>Juan's Home page</h1>

<p>Juan is a Professor at the
University of Oviedo, Spain</p>

<p>Phone: +34-1234567</p>
</body>
</html>
```

Espanish

```
<html lang="es">
<body>
<h1>Página personal de Juan</h1>

<p>Juan es Catedrático en la
   Universidad de Oviedo, España</p>

<p>Tlfno: +34-1234567</p>
</body>
</html>
```



Intrinsically multilingual

Multilingual data

- Data that appears in a multilingual context
  - It contains labels/comments
  - Human-readable information
  - Using different languages/conventions

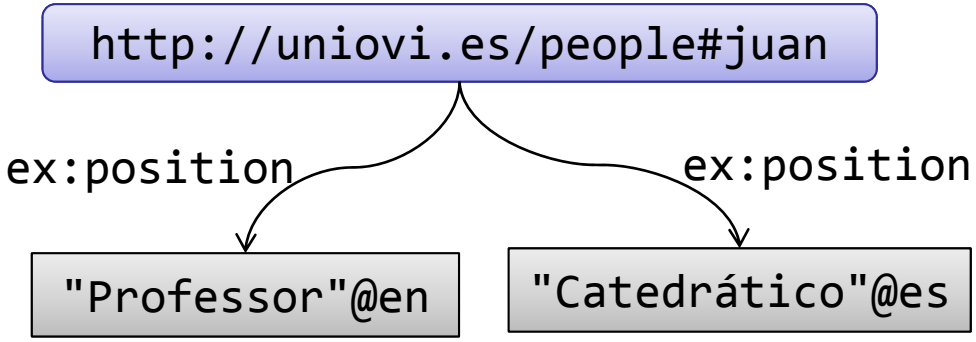
Example of Multilingual Data

English

```
<html lang="en">
<body>
<h1>Juan's Home page</h1>

<p>Juan is a Professor at the
University of Oviedo, Spain</p>

<p>Phone: +34-1234567</p>
</body>
</html>
```

Espanish

```
<html lang="es">
<body>
<h1>Página personal de Juan</h1>

<p>Juan es Catedrático en la
   Universidad de Oviedo, España</p>

<p>Tlfno: +34-1234567</p>
</body>
</html>
```

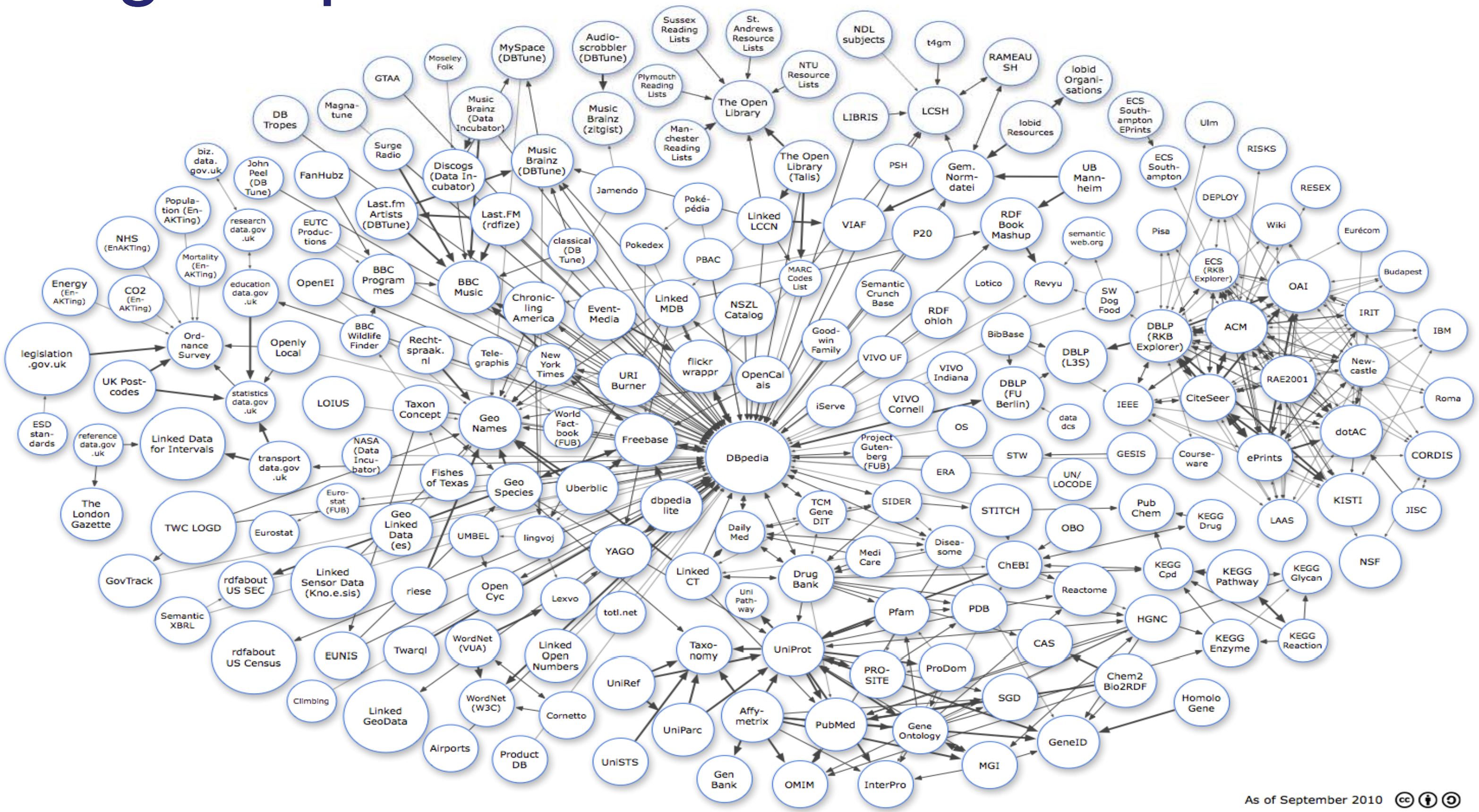


Web of Data

Unit of information: data (RDF)
Human + Machine readable
New Challenge: Multilingual

Linked Open Data

- Principles on how to publish data
- Increasing adoption



Best practices for LOD

Several proposals:

- Linked data book [Heath, Bizer, 2011]
- Linked data patterns [Dodds, Davis, 2012]
- Best Practices for Publishing Linked Data [Hyland et al]
- SemWeb Rules of thumb [R. Cyganiak]
- etc. . .

In this talk

- Best practices affected by multilinguality

Multilingual LOD patterns

1. Design a good IRI scheme
2. Separate domains by language
3. Model resources, not labels
4. Provide human-readable info
5. Labels for all
6. Use Multilingual literals
7. Language Content negotiation
8. Literals without language
9. Multilingual vocabularies
10. Interlanguage links

1. Design a good IRI scheme

Cool URIs → Cool IRIs

- Don't change
- Identify things
- If possible, use human-readable IRIs

`http://dbpedia.org/resource/Armenia`

`http://դբպեդիա.օրգ/ռեսուրսե/Հայաստան`

1. Design a good IRI scheme

- Most datasets use only URIs
- IRIs may be difficult to maintain
  - Domain names, phising, …
  - IRI support in current libraries
  - Some hybrid solutions

```
http://dbpedia.org/resource/Armenia

http://dbpedia.org/resource/Հայաստան

http://դբպեդիա.օրգ/ռեսուրսե/Հայաստան
```

2. Separate domains by language

Instead of

`http://dbpedia.org/resource/Հայաստան`

Language dependant URIs

`http://en.dbpedia.org/resource/Armenia`

`http://hy.dbpedia.org/resource/Հայաստան`

Language identifiers ≠ Country identifiers
Example: Use "hy" instead of "am"

2. Separate domains by language

Where should we put the language tag?

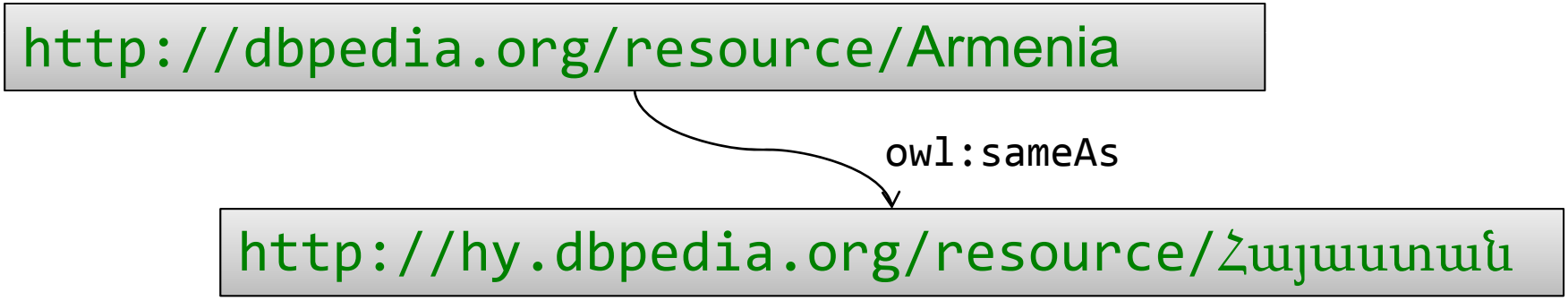
`http://hy.dbpedia.org/resource/Հայաստան`

`http://dbpedia.org/resource/hy/Հայաստան`

`http://dbpedia.org/resource/Հայաստան/hy`

3. Interlanguage links

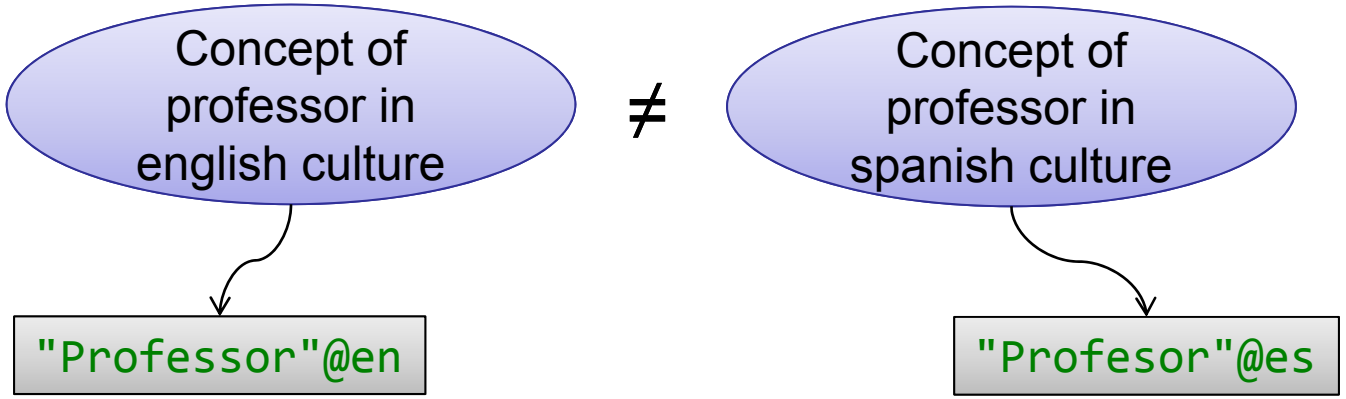
Provide links between concepts in different languages



3. Interlanguage links

Beware of cross-lingual mappings

Example:



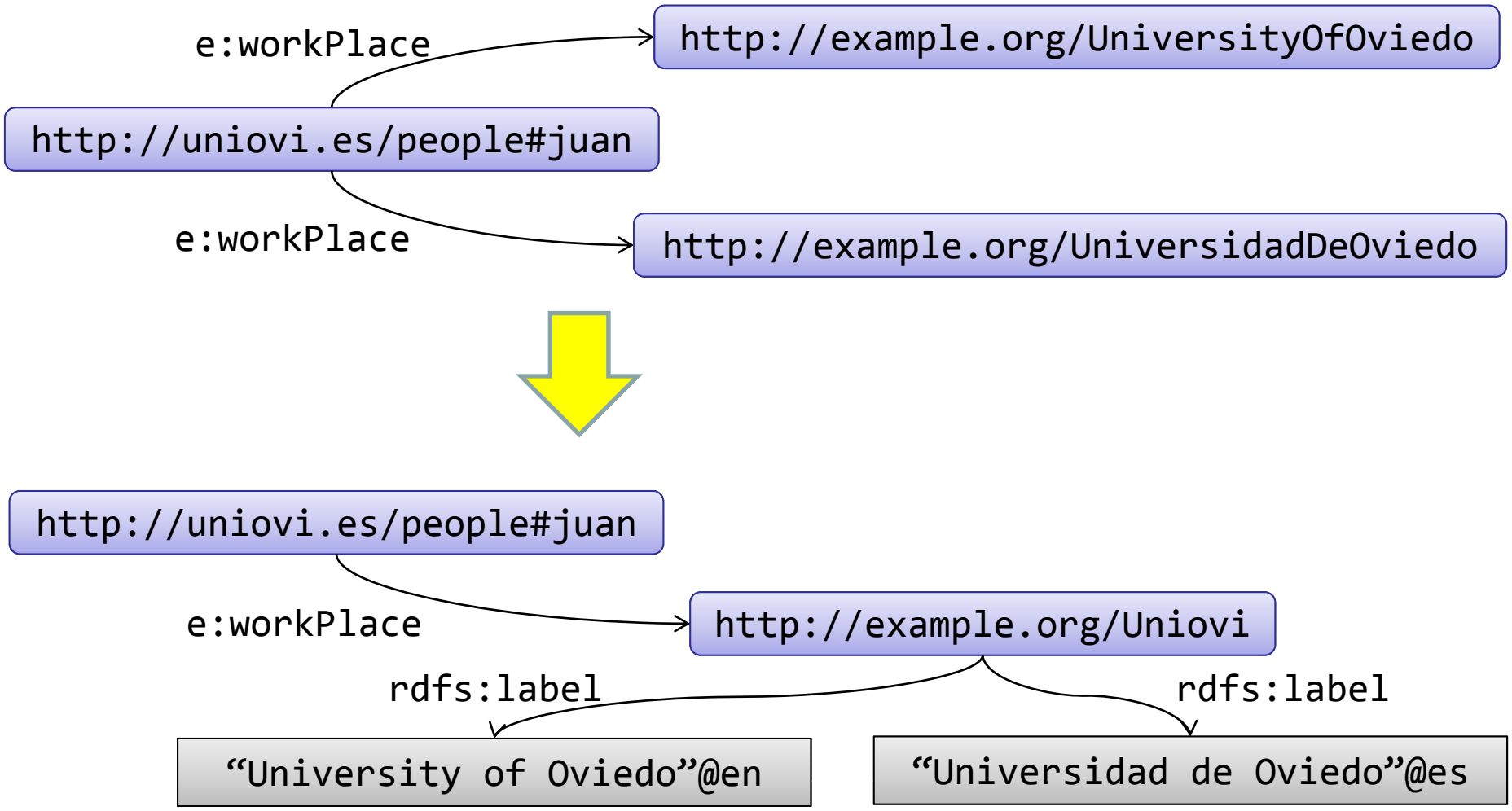
Use other properties to link:

```
dbo:interlanguageLink
rdfs:seeAlso
skos:related
...
```

4. Model resources, not labels

- Define URIs only for resources
  - Resources do not depend on a given language
  - Assign labels to those resources
- Do not mint separate URIs for labels

4. Model resources, not labels



4. Model resources, not labels

- Some domains may want to model labels
  - Linguistic resources
    - Thesaurus, corpora, etc.
    - Assertions and relations between labels
    - Examples:
      - SKOS-XL labels *
        - Resources of type sxosxl:Label
      - Strings URI-identifiable: NIF

*http://www.w3.org/TR/skos-reference/skos-xl.html

5. Provide human-readable info

Not only machine-readable information

Combine machine & human-readable info

Human-readable info must be multilingual

5. Provide human-readable info

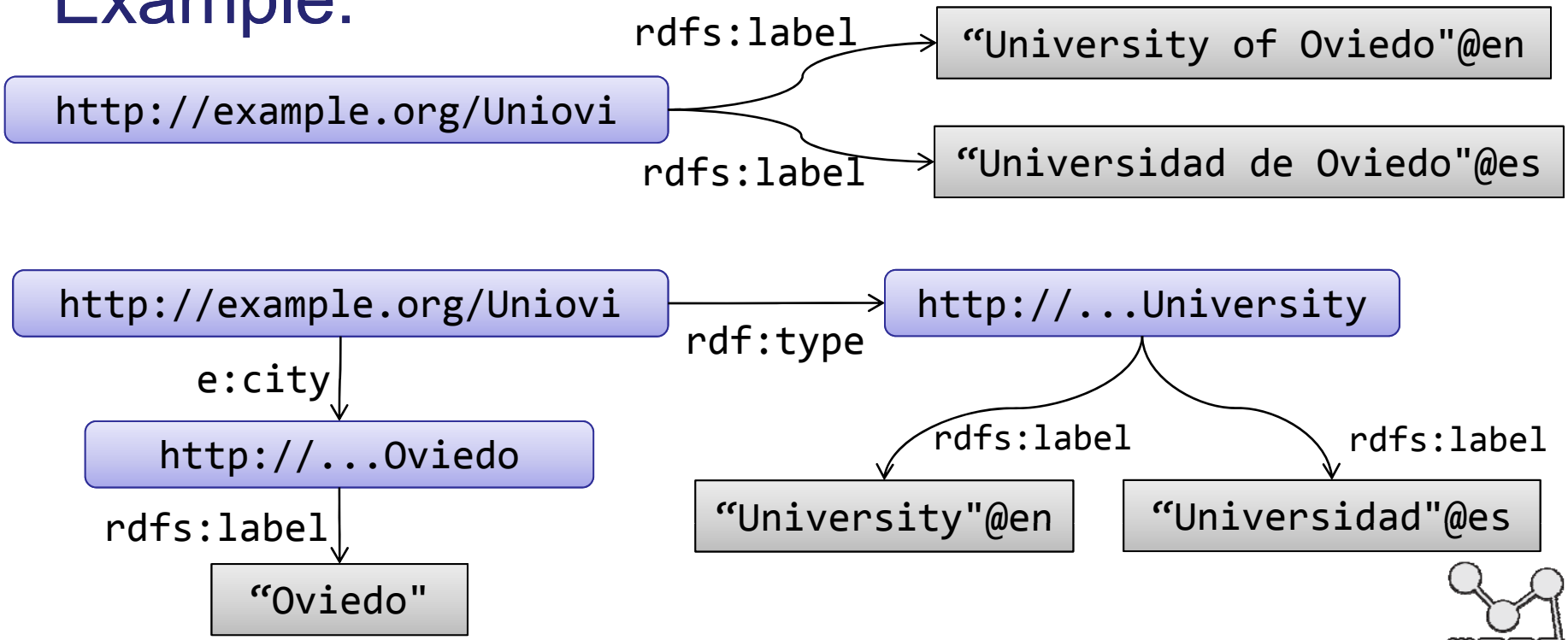
- Facilitates search over the web of data
- Linked data browsing
  - Applications can display labels instead of URIs
- Common properties:
  - `rdfs:label`
  - `skos:prefLabel`
  - `dcterms:title`
  - `dcterms:description`
  - `rdfs:comment`
  - `etc.`

5. Provide human-readable info

What is the right level of textual information?

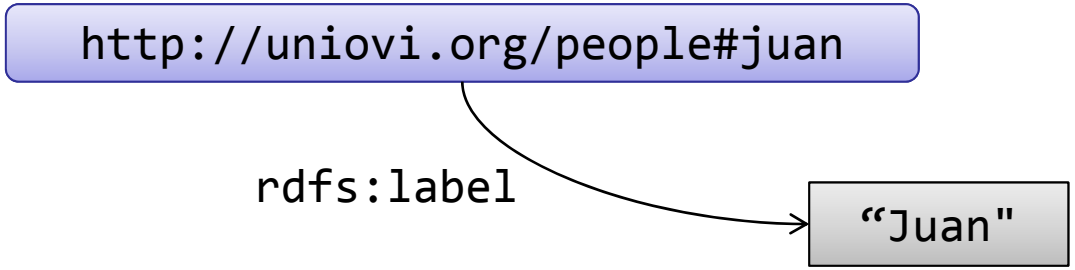
Balance between RDF/Textual world

Example:



6. Labels for all

- Provide labels for all URIs
  - Individuals / Concepts / Properties
  - Not just the main entities
- Displaying labels becomes easier and faster



6. Labels for all

- It may be difficult to select the right label
  - Don't provide more than one preferred label
  - Not feasible for some datasets
    - Only 38% non-information resources have labels

[B. Ell et al, 2011]

- Labels are for humans
  - Avoid camel case or similar notations
  - Guidelines for labelling
    - Upper case, space delimitters, etc

[Montiel et al, 2011]

7. Use Multilingual literals

Use language tags

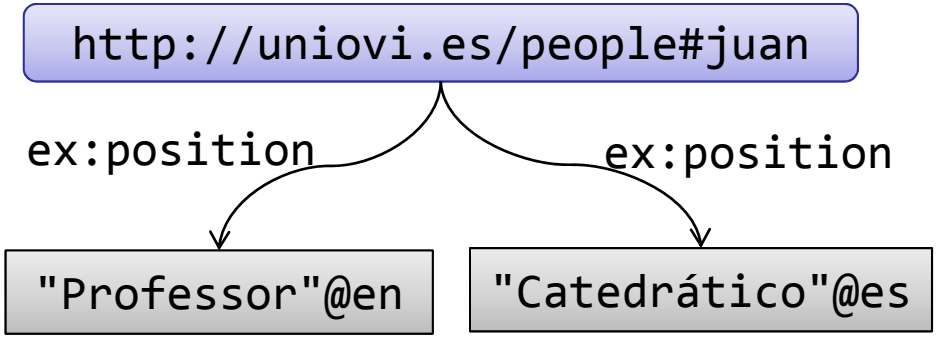
- Select the right IETF language tag (RFC 5646)

Example:

```
"University of Oviedo"@en
"Universidad de Oviedo"@es
"Universidá d'Uvieu"@ast
"Օվիեդոյի համալսարան"@hy
```

7. Use Multilingual literals

Multilingual literals & SPARQL




```
SELECT * WHERE {
  ?x ex:position "Professor" .
}
```

Returns Nothing

```
SELECT * WHERE {
  ?x ex:position "Professor"@en .
}
```

Returns <...#juan>

7. Use Multilingual literals

Underused feature

- 4.78% non info-resources have one language tag
- Only 0.7% datasets contain several language tags

Most commonly language used:

44.72% (en), 5.22% (de), 5.11% (fr), 3.96% (it),...

[B.Ell et al, 2011]

7. Use Multilingual literals

What about longer descriptions:

`dcterms:description, rdfs:comment`...

CDATA like or XML literals ?

Reuse existing practices in XML I18n

Problems:

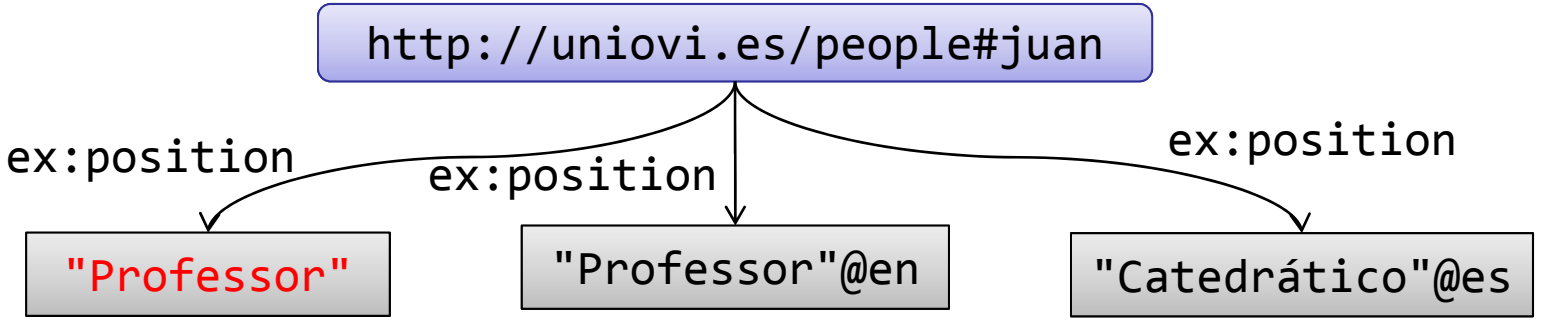
- Gap between descriptions and RDF model
- SPARQLing may be a challenge

8. Literals without language tag

Include literals without language-tag

SPARQL queries are easier

Example:




```
SELECT * WHERE {
  ?x ex:position "Professor" .
}
```

Returns <...#juan>

8. Literals without language tag

- Selecting a default language = controversial
  - Declare the primary language of a dataset
    - Some properties: `lexvo:language`
  - Consumers may not be aware of the default language

9. Language Content negotiation

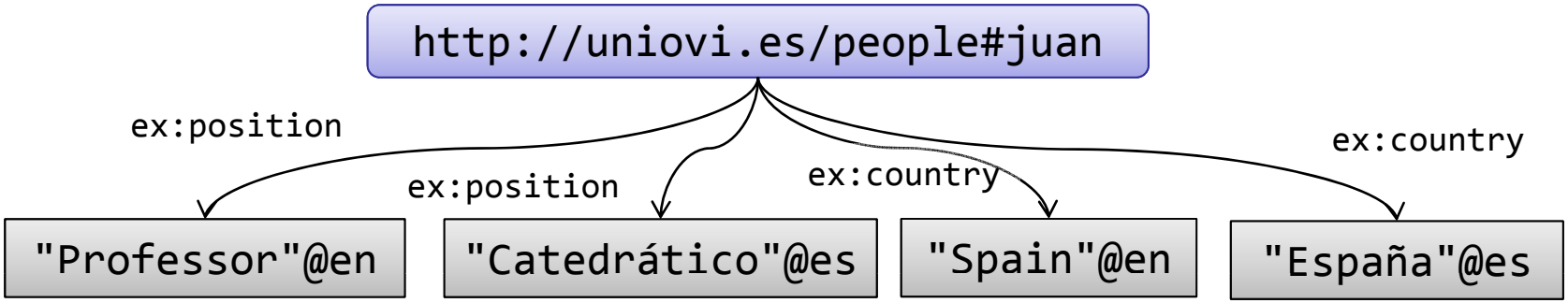
Use HTTP Accept-Language

Return different sets of labels

Reduce load in client applications

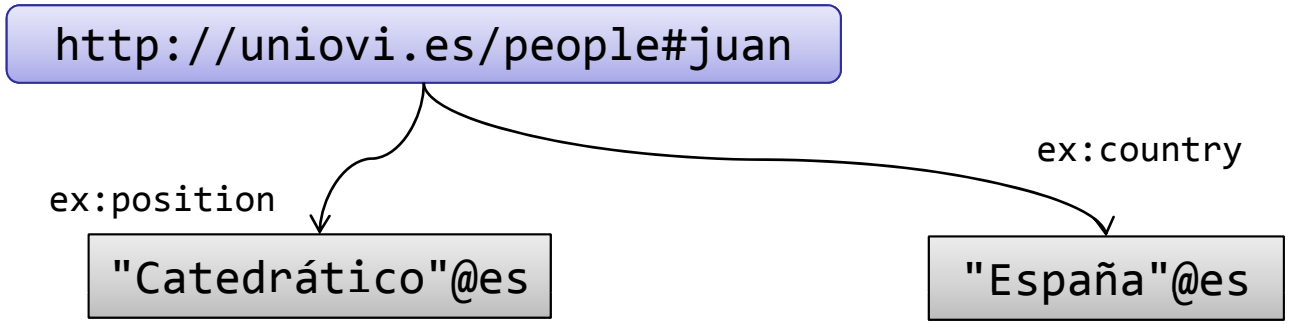
9. Language Content negotiation

No Accept-Language declaration (all)



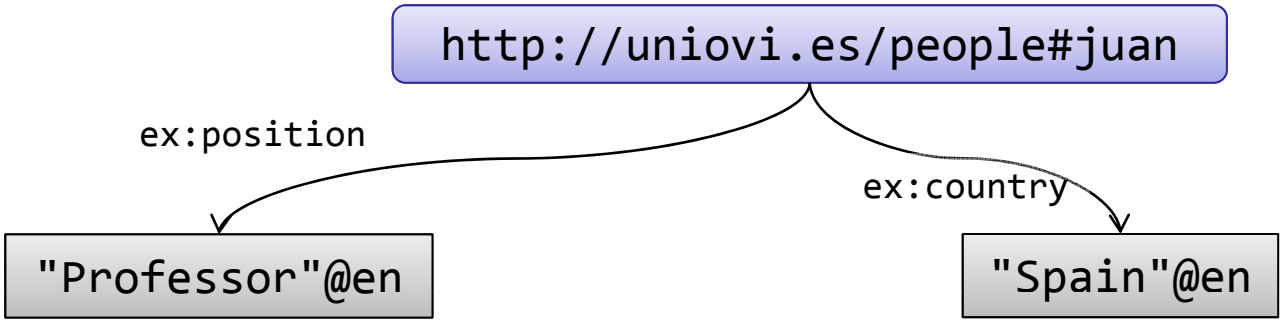
9. Language Content negotiation

```
Accept-language: es
```



9. Language Content negotiation

```
Accept-language: en
```



9. Languge Content negotiation

Not done in practice.

- Implementation issues?

Ensure equivalent representations for each language

10. Multilingual vocabularies

Link to existing vocabularies

Quality selection criteria for vocabularies

- Use vocabularies that contain descriptions in more than one language

[Hyland et al, 2012]

10. Multilingual vocabularies

Popular vocabularies are not localized

- Example: FOAF, DC, etc.

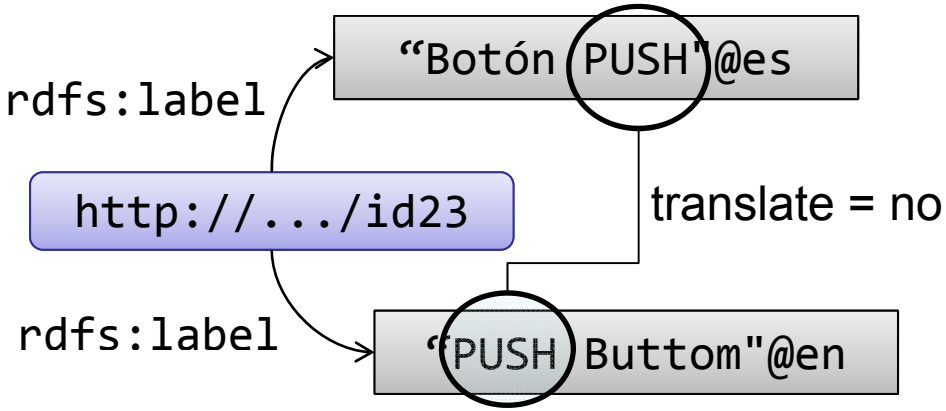
Should we extend it?

- Example:

```
dc:contributor rdfs:label "Colaborador"@es .
```

Other issues, not covered

- Unicode support in RDF
- Language declarations & Microdata
- Internationalization topics:
  - Text direction
  - Ruby annotations
  - Notes for localizers
  - Translation rules
- Linguistic topics
  - Ontology-lexicon, Lemon Model



[Gracia et al, 2011, Buitelaar et al, 2011, McCrae et al 2011]

Conclusions

Web of data is not just for machines

LOD applications will be used by humans

...and

Human users talk many different languages

Best? practices for Multilingual LOD

Acknowledgements

Jose María Álvarez Rodríguez
Sören Auer
Richard Cyganiak
Basil Ell
Sebastian Hellmann
Aidan Hogan
Dimitris Kontokostas
Pablo Mendes
Elena Montiel
Jeni Tennison
Boris Villazón-Terrazas

References

[Buitelaar et al, 2011] Ontology Lexicalisation: The lemon Perspective, 9th International Conference on Terminology and Artificial Intelligence, 2011

[Cyganiak] SemWeb Rules of thumb
http://www.w3.org/wiki/User:Rcygania2/RulesOfThumb

[Dodds, Davis, 2012] Linked data patterns
http://patterns.dataincubator.org/book/

[Ell et al, 2011] Labels in the Web of Data, ISWC 2011

[Gracia et al, 2011] Challenges for the Multilingual Web of Data, International Jounal on Semantic Web and Information Systems, 2011

[Hogan et al, 2012] An empirical study of Linked Data Conformance, Journal of Web Semantics, to appear.

[Heath, Bizer, 2011] Linked data: Evolving the Web into a Global Data Space
http://linkeddatabook.com/editions/1.0/

[Hyland et al] Best Practices for Publishing Linked Data
https://dvcs.w3.org/hg/gld/raw-file/default/bp/index.html#internationalized-resource-identifiers

[Hyland et al] Linked data cookbook. Open Government Linked Data
http://www.w3.org/2011/gld/wiki/Linked_Data_Cookbook

[McCrae et al, 2011] Linking Lexical Resources and Ontologies on the Semantic Web with lemon, ESWC, 2011

[Montiel et al, 11] Style Guidelines for Naming and Labeling Ontologies in the Multilingual Web, Elena Montiel-Ponsoda et al, DC-2011

End of presentation

http://purl.org/weso