Best Practices for Multilingual Linked Open Data

Jose Emilio Labra Gayo

University of Oviedo, Spain

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

About me

WESO Research Group (Web Semantics Oviedo, since 2004)

Several projects involving Multilingual LOD

Example: EU Public procurement notices (MOLDEAS)

Catalog of product schema clasifications (1842053 triples)

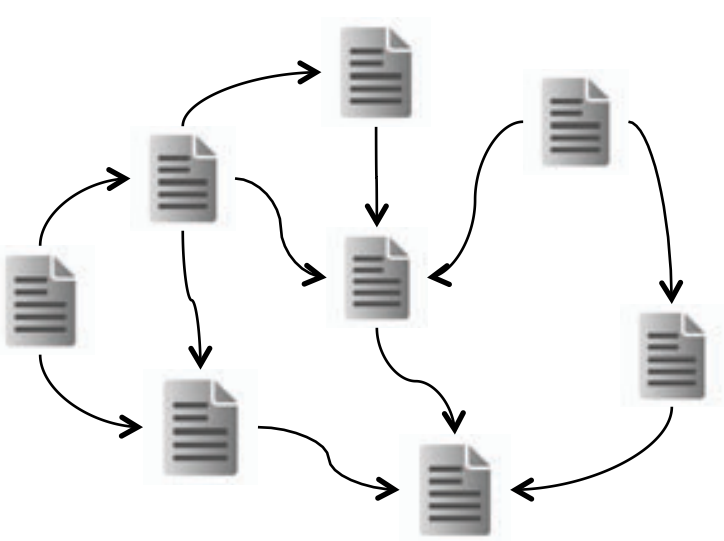
□tt□r//t□□□□t□□p□g□□□/□□t□h□t/□h□hs□□t□□□□p□□

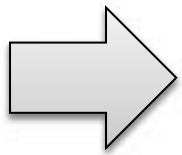
Common Procurement vocabulary (803311 triples)

□tt□r//t□□□□t□□p□g□□□/□□t□h□t/□□:s3jjf□

23 EU languages

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

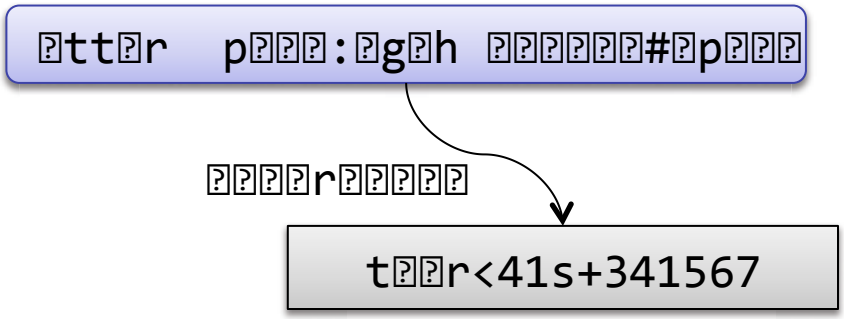
Example

English

```
=□t□□□□□□□□mn□□n"
=□□□d"
=□+"□p□□8h□□□□□□□□□□□= □+"
□
=□"□p□□□□h□□□□□□□□□hh□□□□t□t□□□
□□□:□□h□td□□□□□:□□□□o□□□□□□= □"
□
=□
= □□□d"□
= □t□□"□
```

Espanish

```
=□t□□□□□□□□mn□hn"
=□□□d"
=□+"□a□□□□□□□□□h□□□□□□□□□□p□□= □+"
□
=□"□p□□□□h□□□t□□□at□□□□□□□□□□□□□□□
□□□□□□□:□□h□□□□□□□□□:□□□□o□□h□□u□= □
□
=□
= □□□d"□
= □t□□"□
```



Intrinsically multilingual

Multilingual data

Data that appears in a multilingual context

- It contains labels/comments
- Human-readable information
- Using different languages/conventions

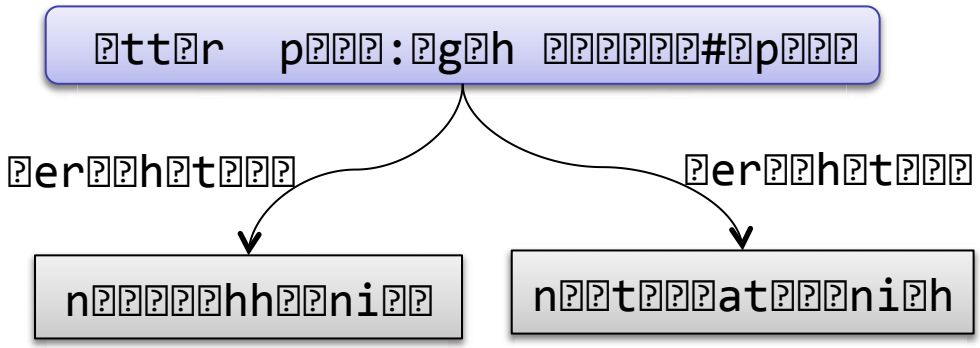
Example of Multilingual Data

English

```
=□t□□□□□□□□mn□□n"
=□□□d"
=□+"□p□□8h□□□□□□□□□□□= □+"
□
                                    □t□t□□□
□□□:□□h□td□□□□□:□□□□o□□□□□□= □"
□
=□"□□□□□r□<41s+341567= □"
= □□□d"□
= □t□□"□
```

Espanish

```
=□t□□□□□□□□mn□hn"
=□□□d"
=□+"□a□□□□□□□□□h□□□□□□□□□□p□□= □+"
□
                                    □□□□□□□□□□□
□□□□□□□:□□h□□□□□□□□□□:□□□□o□□h□□u□= □
□
=□"□□□□□r□<41s+341567= □"
= □□□d"□
= □t□□"□
```



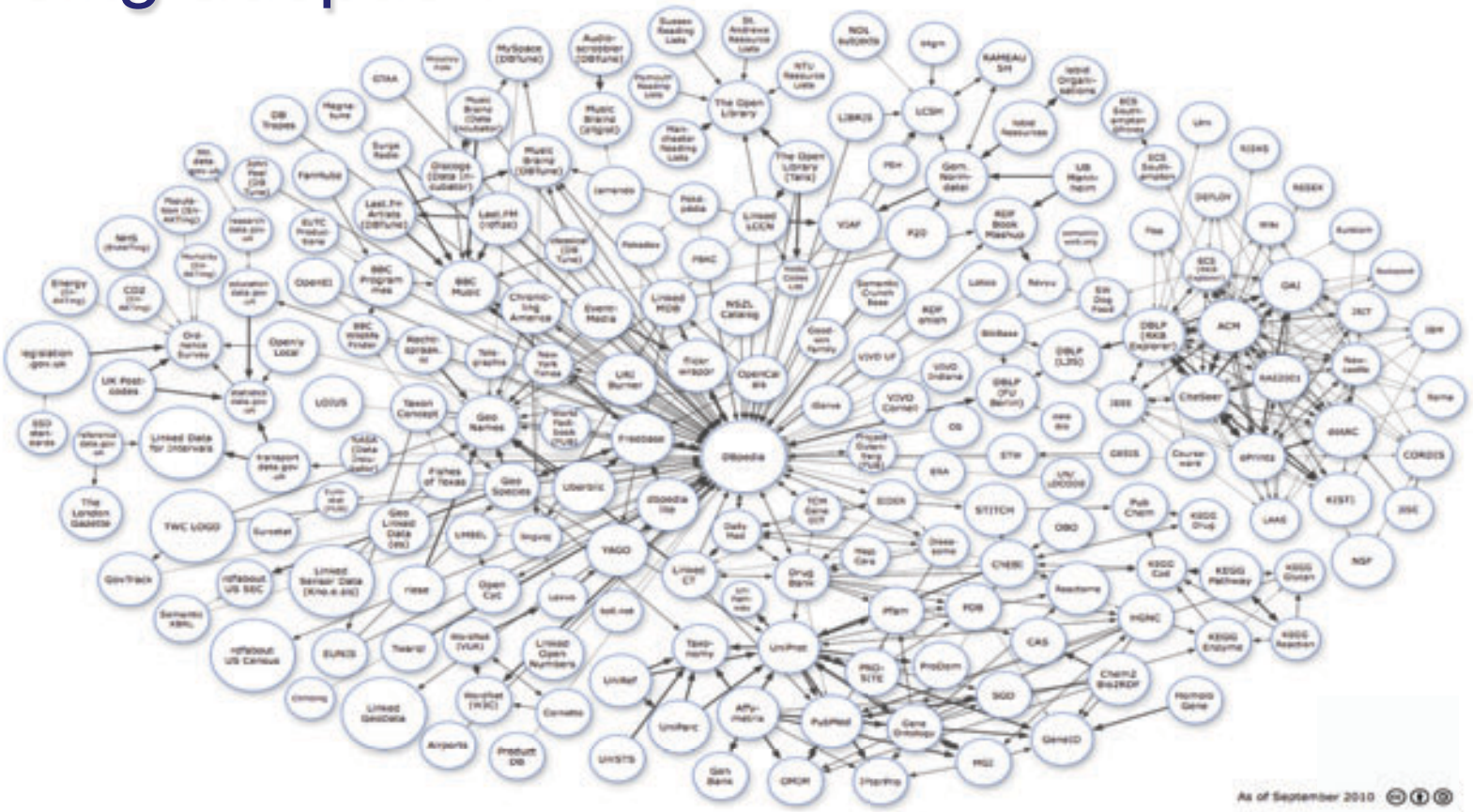
Web of Data

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 practices

1. Design a good URI scheme
2. Model resources, not labels
3. Use human-readable info
4. Labels for all
5. Use Multilingual literals
6. Content negotiation
7. Literals without language
8. Multilingual vocabularies

1. Design a good URI scheme

Cool URIs

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

1. Design a good URI scheme

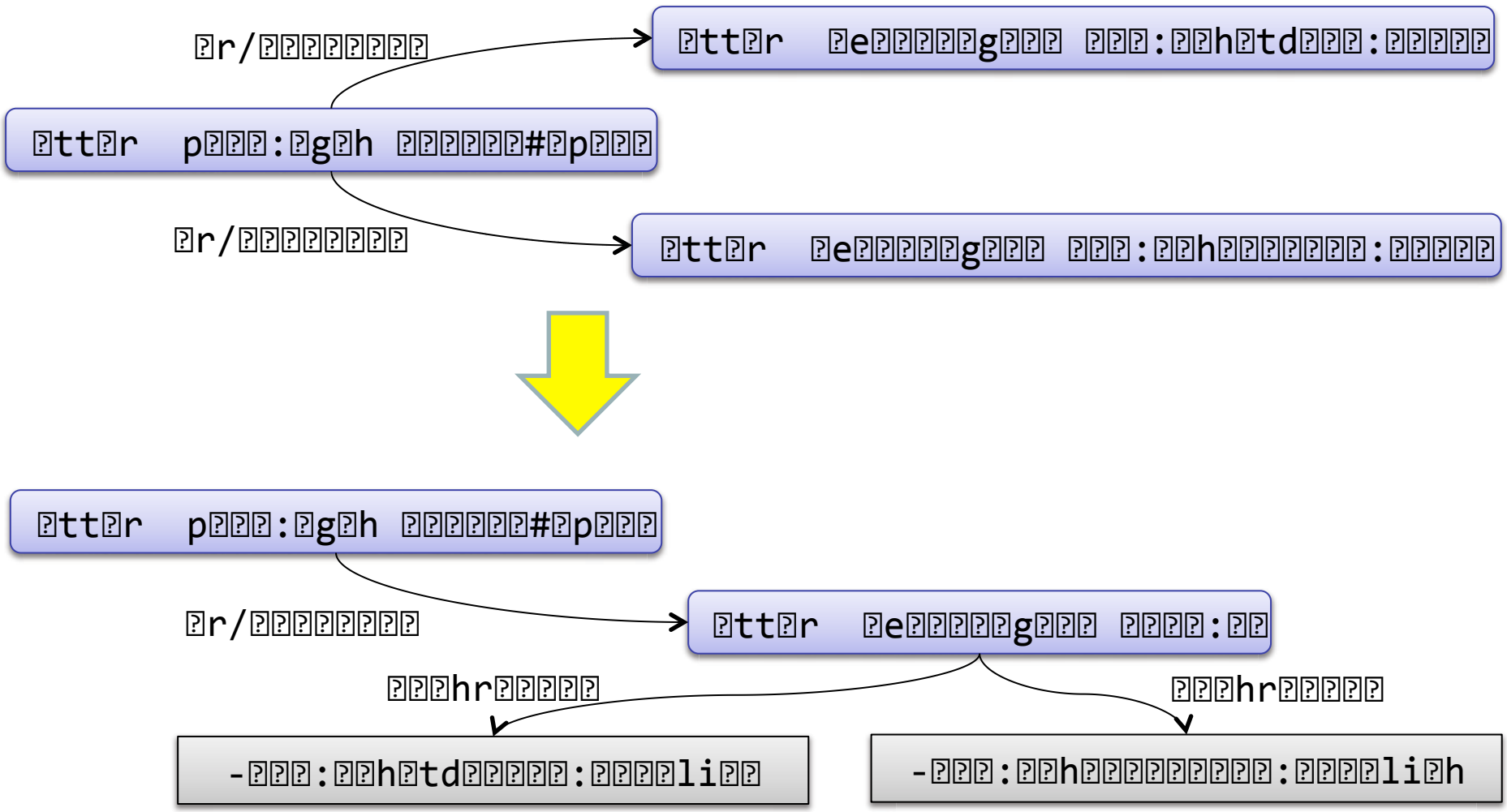
- Use IRIs?
- Most datasets use only URIs
- IRIs may be difficult to maintain
  - Domain names, phising, …
  - IRI support in current libraries
  - Human-readability?

```
□tt□r  □□□□□□□□g□□□  □□h□p□□□  Armenia
□tt□r  □□□□□□□□g□□□  □□h□p□□□  Հայաստան
հտտպ://դբպեդիա.օրգ/րեսօուրսե/Հայաստան□□
```

2. 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

2. Model resources, not labels



2. Model resources, not labels

- Some domains may require to model labels
  - Thesaurus
  - Assertions and relations between labels
  - Example: SKOS-XL labels
    - Resources of type sxosxl:Label
    - Labels are URI-identifiable

2. Model resources, not labels

Mint different URIs for each language?

Localized URIs

□tt□r □□□□□□□□g□□□ □□h□p□□□ Armenia□□

□tt□r □□□□□□□□g□□□ □□h□p□□□ Հայաստան□□

Language dependant URIs

□tt□r □□□□□□□□g□□□ □□h□p□□□ Armenia/en□□

□tt□r □□□□□□□□g□□□ □□h□p□□□ Armenia/hy□□

3. Use human-readable info

Not only machine-readable information

- Combine machine & human-readable info
- Human-readable info must be multilingual

3. Use human-readable info

Facilitates search over the web of data

Linked data browsing

- Applications can display labels instead of URIs

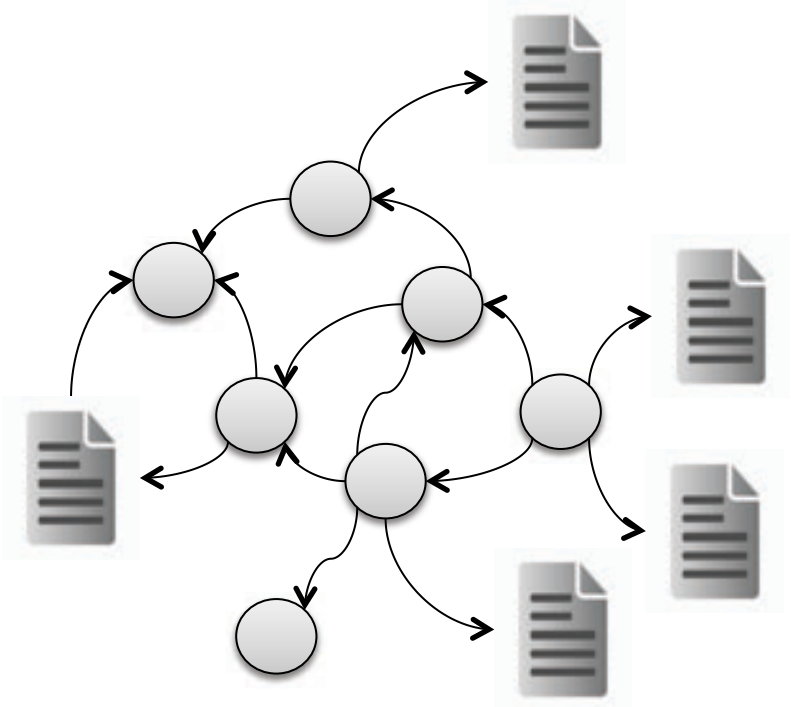
Some common properties:

- ???hr??????
- h??hr??????????
- ??t???hrt?t???
- ??t???hr??h????t????
- ???hr???????t?
- ?t?g?

3. Use Human-readable info

What is the right level of textual information?

Balance between HTML/RDF world



4. Labels for all

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

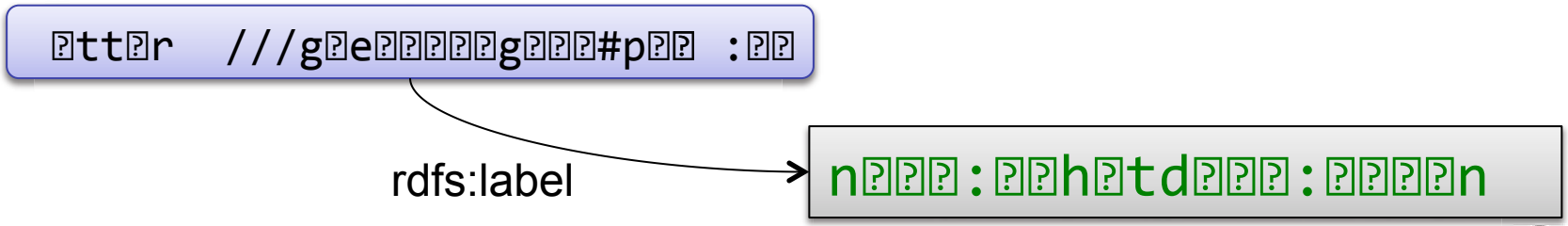
4. Labels for all

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]

Avoid camel case or similar notations



5. Use Multilingual literals

Use language tags

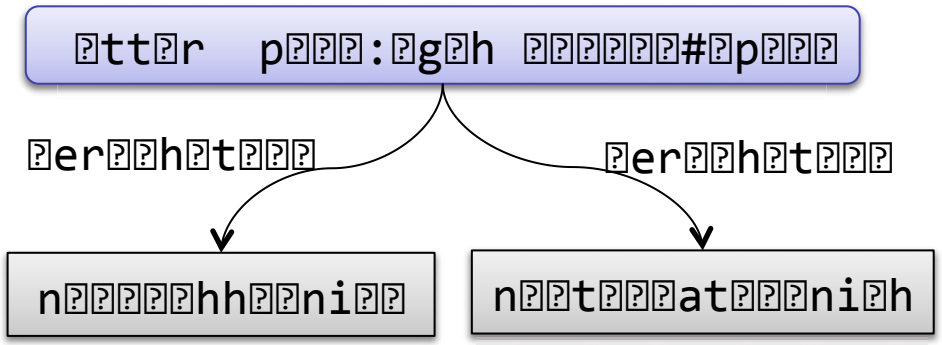
Select the right IETF language tag (RFC 5646)

Example:

```
□n□□□:□□h□td□□□□□□:□□□□ni□□□
□n□□□:□□h□□□□□□□□□□□:□□□□ni□h□
□n□□□:□□h□□a□□8□:□□pni□ht□
□n0վիեդոյի համալսարանում"i□d□
□
```

5. Use Multilingual literals

Multilingual literals & SPARQL




```
????????0????????v?
??ce??er??h?t????n?????hh??n?g?
2
```

Returns Nothing

```
????????0????????v?
??ce??er??h?t????n?????hh??ni???g?
2
```

Returns =ggg#?p??"?

5. 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]

5. Use Multilingual literals

What about longer descriptions:

□□t□□□hr□□h□□□□t□□□o□□□□hr□□□□□t…

CDATA like or XML literals ?

Reuse existing practices in XML I18n

Problems:

- Gap between descriptions and RDF model
- SPARQL maybe a challenge

6. Content negotiation

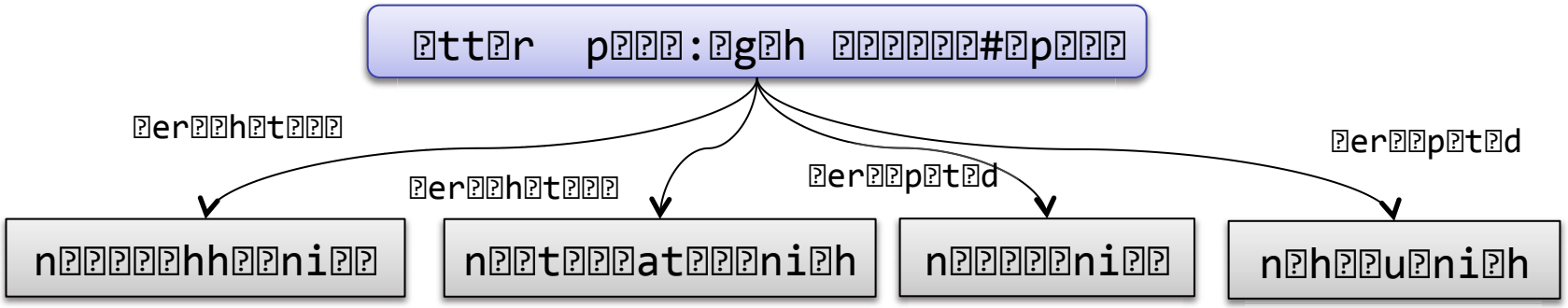
Use HTTP Accept-Language

Return different sets of labels

Reduce load in client applications

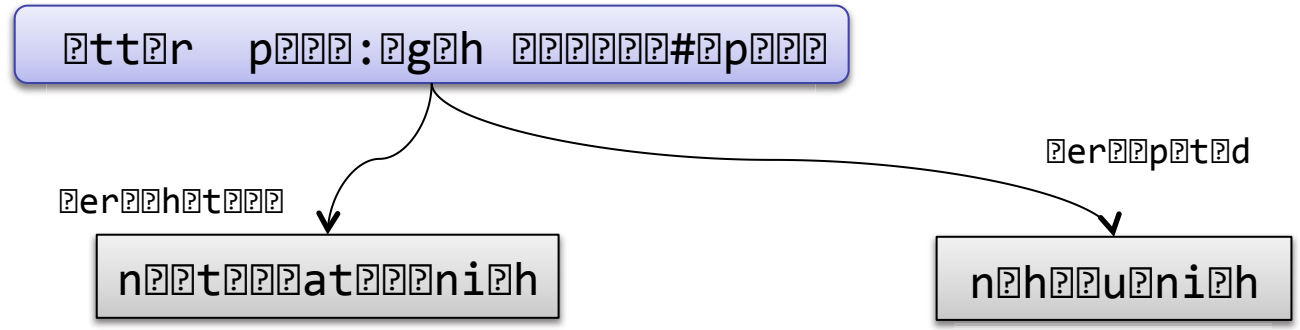
6. Content negotiation

No Accept-Language declaration (all)



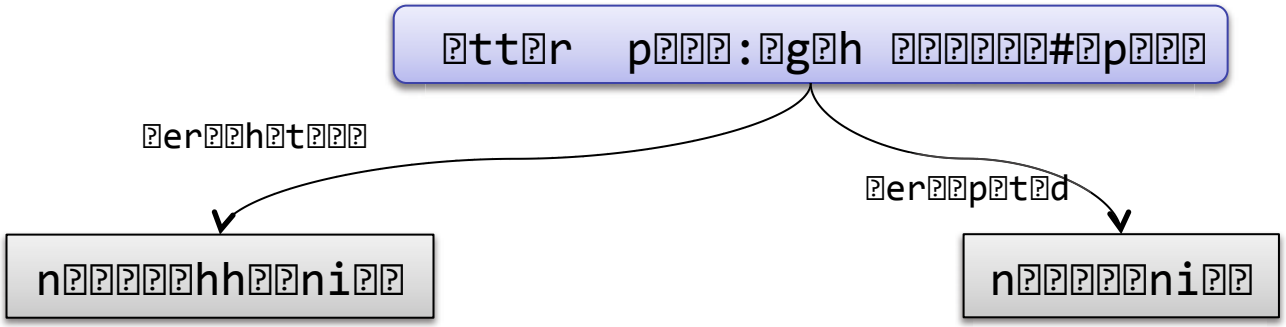
6. Content negotiation

?????ts????p???r??h?



6. Content negotiation

??????ts????p????r????



6. Content negotiation

Implementation issues

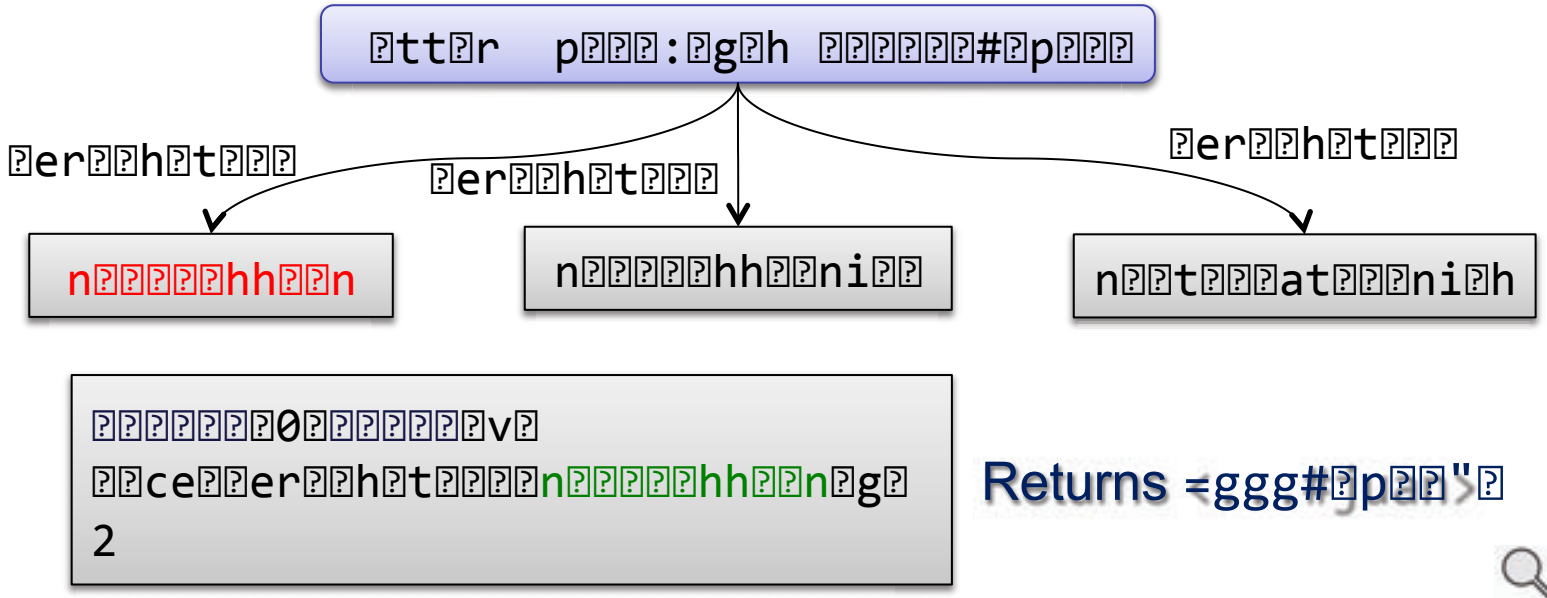
Return equivalent representations for each language

7. Literals without language tag

Include literals without language-tag

SPARQL queries are easier

Example:



7. Literals without language tag

- Selecting a default language maybe controversial
- How to declare the primary language of a dataset?

8. Multilingual vocabularies

- Link to existing vocabularies
- Quality selection criteria for vocabularies
  - Vocabularies should contain descriptions in more than one language

[Hyland et al, 2012]

8. Multilingual vocabularies

What to do if they are not localized?

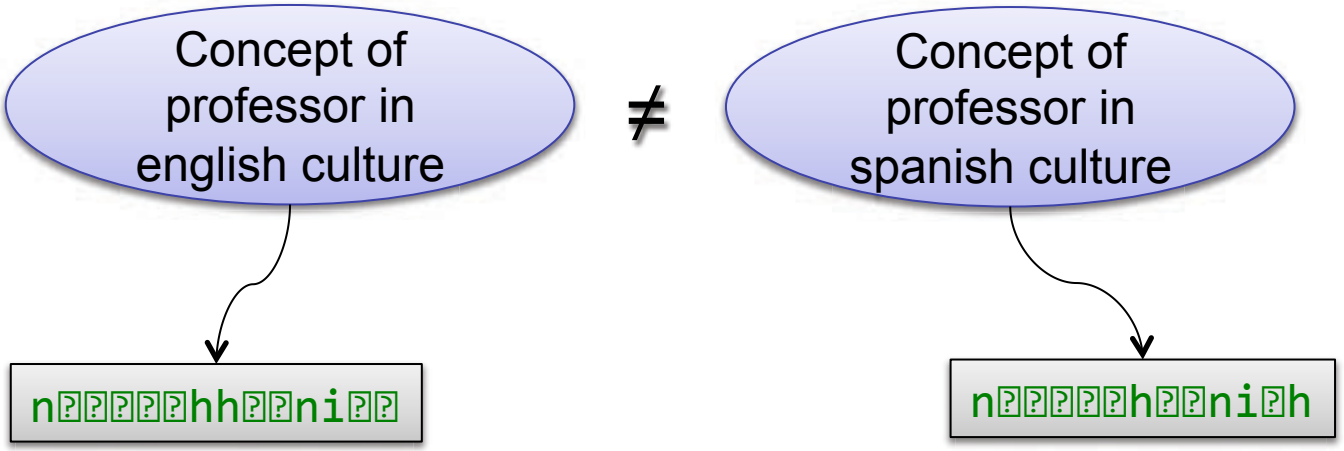
- Enrich vocabularies with translated extensions?
- Example:

  □□r□□□t□□□pt□□□□□□hr□□□□□□n□□□□□ □ □□□ni□h□g□

8. Multilingual vocabularies

Beware of cross-lingual mappings

Example:



Possible solutions:

Ontology-lexicon, Lemon Model

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

Other issues not covered

- Unicode support in N-Triples
- Language declarations in Microdata
- Internationalization topics:
  - Text direction
  - Ruby annotations
  - Notes for localizers
  - Translation rules

Conclusions

LOD adoption offers new challenges

Web of data is not just for machines

At the end, human users will employ LOD applications.

- Human users speak different languages

Challenge:

- **Best?** practices for multilingual LOD

Acknowledgements

Aidan Hogan

Richard Cyganiak

Basil Ell

Jose María Álvarez Rodríguez

Elena Montiel

Jeni Tennison

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.

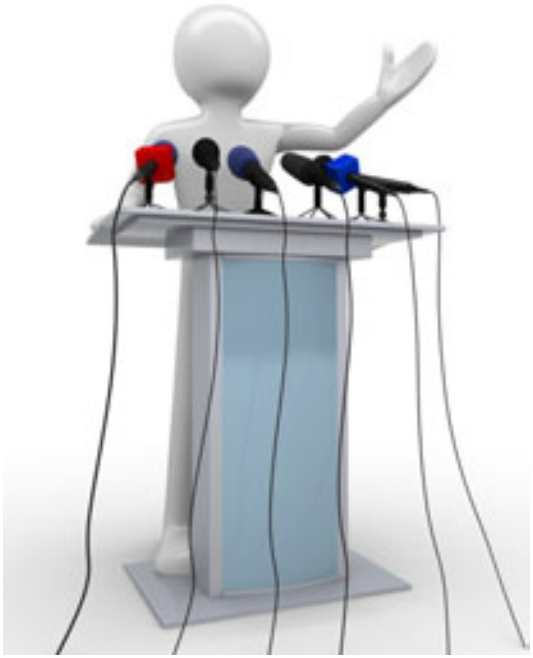
[Heath, Bizer, 2011] Linked data: Evolving the Web into a Global Data Space
http://linkeddatabook.com/editions/1.0/

[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

End of presentation



http://purl.org/weso