Thursday' talks

Who cares about DBpedia properties? Use infoboxes' properties to create better SPARQL queries

Mariano Rico, May 2016

@marianorico

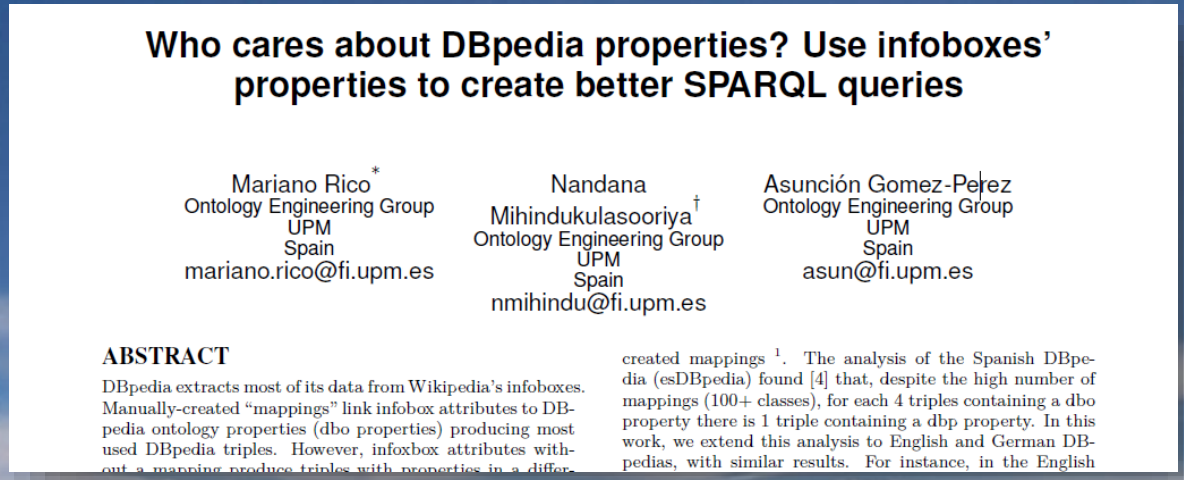
mariano.rico@upm.es

General motivation

- You learn something
- I get feedback
- Implicit agreement
  - I do my best trying to teach you something **valuable**
  - You give me **valuable** feedback
    - Please, be as polite as possible ☺

My motivation

- Long tail work on DBpedia properties
- Submitted as short paper at iSEMANTICS



Who cares about DBpedia properties? Use infoboxes' properties to create better SPARQL queries

Mariano Rico*
Ontology Engineering Group
UPM
Spain
mariano.rico@fi.upm.es

Nandana Mihindukulasooriya†
Ontology Engineering Group
UPM
Spain
nmihindu@fi.upm.es

Asunción Gomez-Perez
Ontology Engineering Group
UPM
Spain
asun@fi.upm.es

ABSTRACT

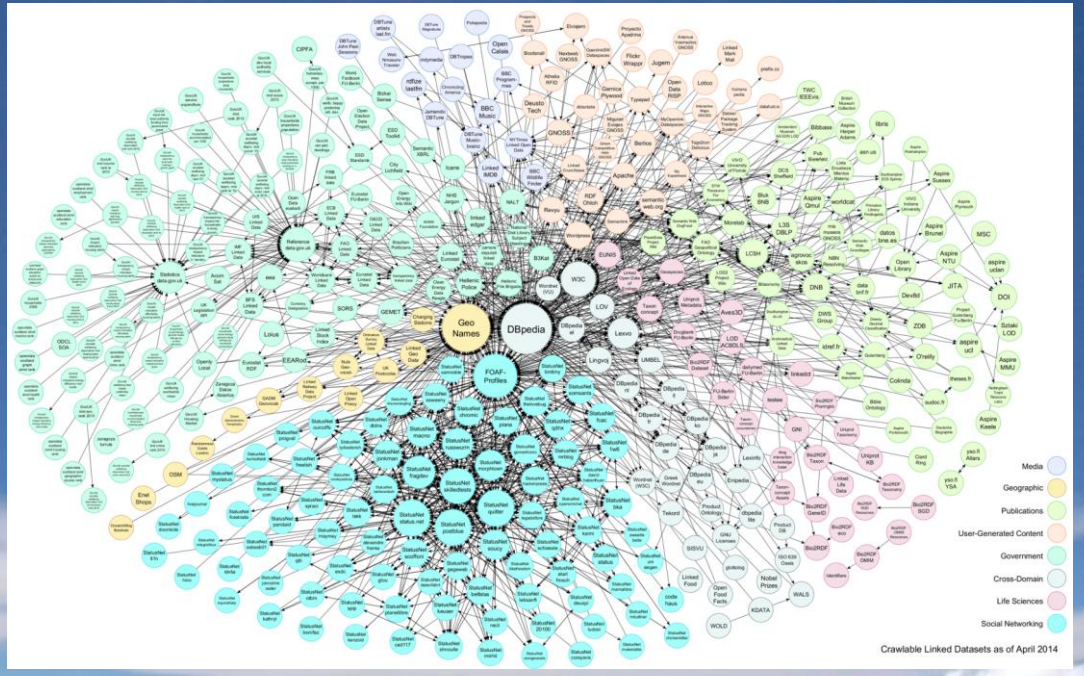
DBpedia extracts most of its data from Wikipedia's infoboxes. Manually-created "mappings" link infobox attributes to DBpedia ontology properties (dbo properties) producing most used DBpedia triples. However, infoxbox attributes without a mapping produce triples with properties in a differ-

created mappings [1]. The analysis of the Spanish DBpedia (esDBpedia) found [4] that, despite the high number of mappings (100+ classes), for each 4 triples containing a dbo property there is 1 triple containing a dbp property. In this work, we extend this analysis to English and German DBpedias, with similar results. For instance, in the English

In the LOD galaxy…

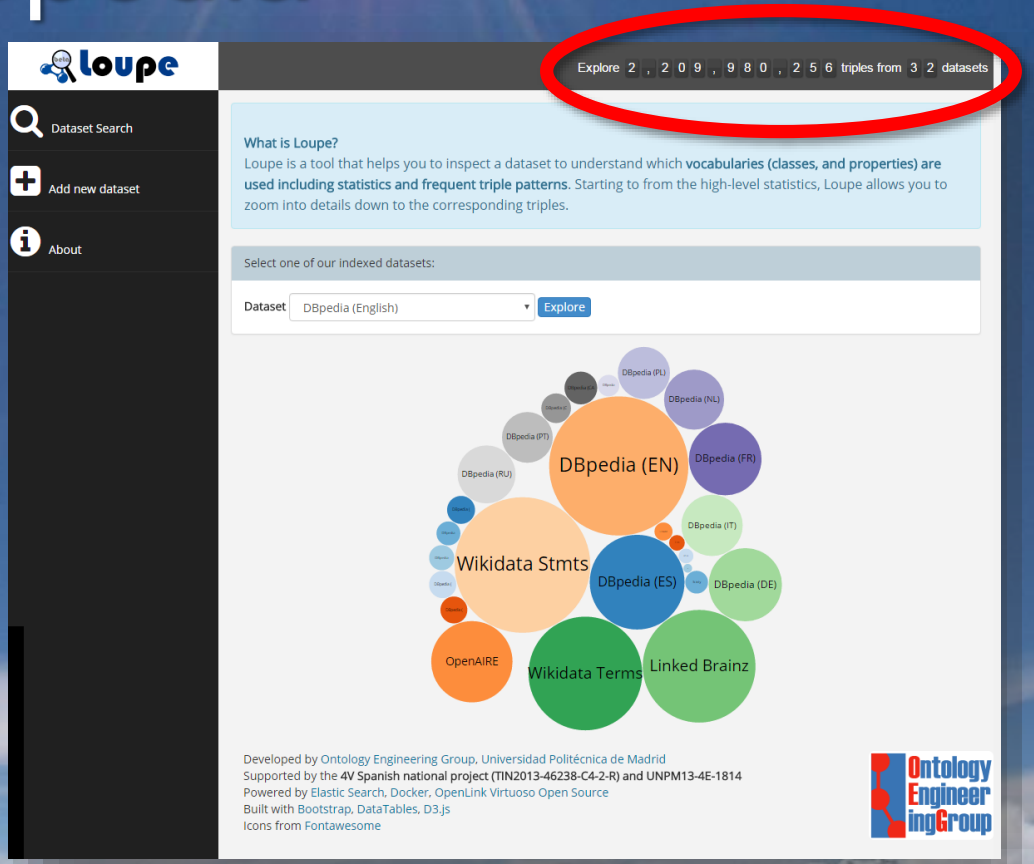
DBpedia

- Kernel of the LOD
- Several DBpedias
  - English
  - Spanish
  - … up to 16



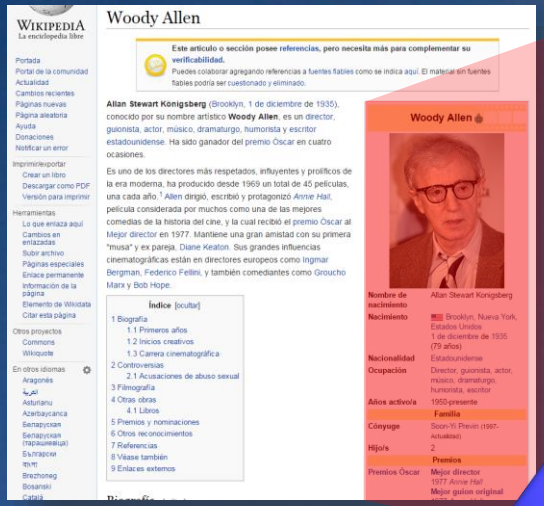
DBpedia

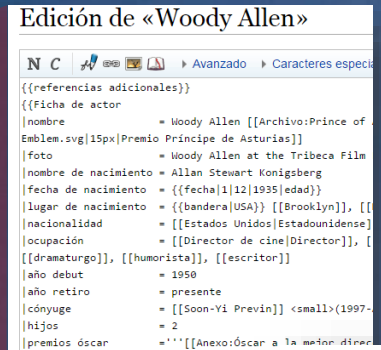
- How to explore these **huge** datasets?
  - By means of SPARQL queries ☹
  - Explore them using LOUPE ☺ loupe.linkeddata.es

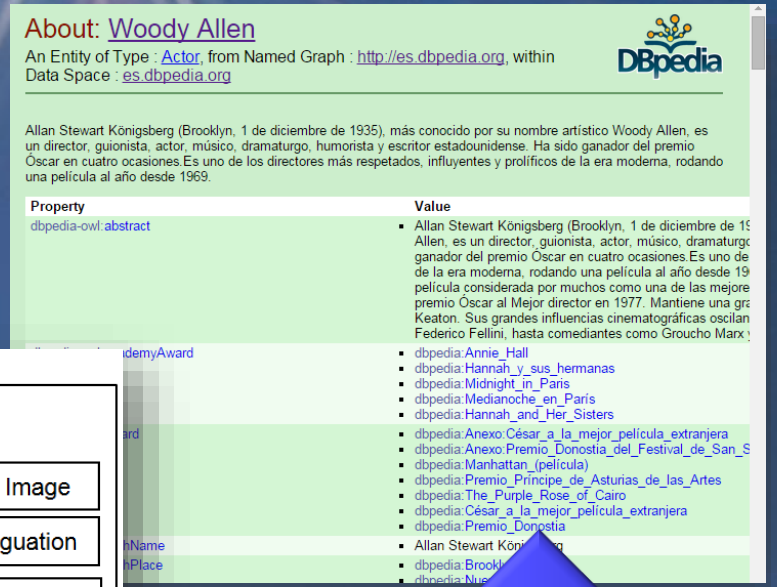


Claims

- C1: DBpedia has more dbp properties than expected
  - you know dbo properties
- C2: SPARQL queries barely use dbp properties
  - We need logs
- C3: The method proposed can enhance DBpedia SPARQL queries
  - Enhance? (more results)







infobox

mappings

Extraction Job

Extractors

Label	Category	Image
Redirect	Disambiguation	
Abstract	Geo	Pagelink
Generic Infobox		
Mapping-based Infobox		

Parsers

DateTime	Units	Geo
String-List	Numbers	

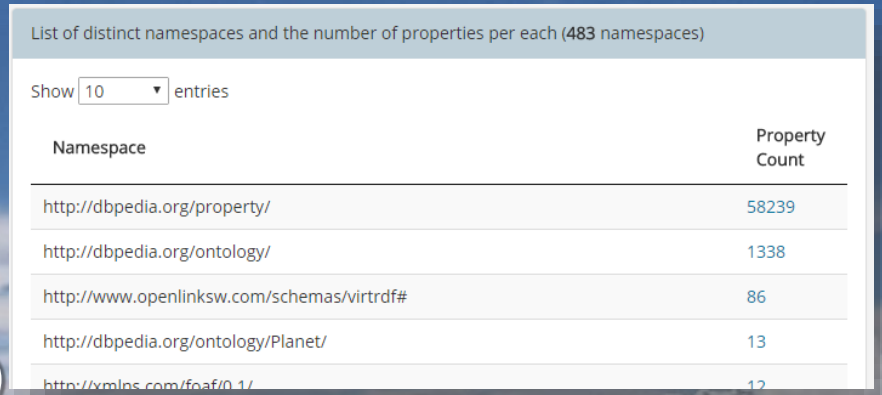
RDF

Rendering

Triple store

Claim C1

- DBpedia has more dbp properties than expected
  - At least for the 3 DBpedias Analyzed (Eng, Spa, Ger)
- We get this from LOUPE (Eng)
  - Number of properties for the namespaces
  - But, what about triples using that properties?
    - LOUPE does not know (yet)



List of distinct namespaces and the number of properties per each (**483** namespaces)

Show 10 entries

Namespace	Property Count
http://dbpedia.org/property/	58239
http://dbpedia.org/ontology/	1338
http://www.openlinksw.com/schemas/virtrdf#	86
http://dbpedia.org/ontology/Planet/	13
http://xmlns.com/foaf/0.1/	12

Claim C1

- We need a “strong” SPARQL query
  - Any EP would produce timeout
  - We have superpowers ☺

Table 1: Top-10 dbp properties for the English, Spanish and German DBpedias (2015-04 version).

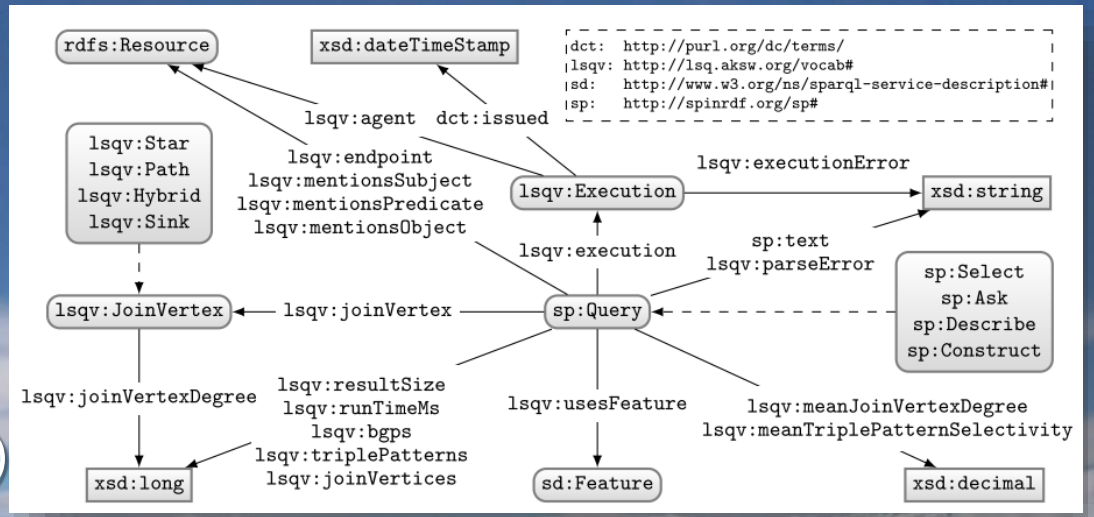
English DBpedia		Spanish DBpedia		German DBpedia	
dbp: http://dbpedia.org/property/					
URI	Triples	URI	Triples	URI	Triples
dbp:hasPhotoCollection	4,041,585	dbp:wikiPageUsesTemplate	3,402,499	dbp:name	494,852
dbp:name	4,021,368	dbp:nombre	558,837	dbp:geburtsort	305,063
dbp:title	1,452,504	dbp:título	327,498	dbp:kurzbeschreibung	283,695
dbp:subdivisionType	1,257,766	dbp:name	230,763	dbp:geburtsdatum	283,405
dbp:shortDescription	1,194,274	dbp:tipoSuperior	225,868	dbp:typ	232,702
dbp:dateOfBirth	1,023,951	dbp:horario	203,890	dbp:viaf	169,145
dbp:subdivisionName	1,004,294	dbp:imagen	183,887	dbp:gnd	165,362
dbp:goals	969,216	dbp:familia	152,430	dbp:jahre	156,498
dbp:placeOfBirth	908,819	dbp:title	144,724	dbp:sterbedatum	144,209
dbp:birthPlace	903,529	dbp:ordo	142,196	dbp:alternativnamen	143,893
#props dbp	58,239	#props dbp	17,111	#props dbp	12,167
#props dbo	1,338	#props dbo	559	#props dbo	534
#triples dbp	78,125,087	#triples dbp	28,234,292	#triples dbp	10,483,987
#triples dbo	82,369,408	#triples dbo	90,389,560	#triples dbo	50,750,486

Claim C2

- SPARQL queries barely use dbp properties
  - We need to analyze SPARQL repositories (logs)
  - Fortunately
    - We have logs for the Spanish DBpedia
    - There is a public repository at aksw.github.io/LSQ
      - DBpedia (logs from 30/04/2010–20/07/2010). 1.7M queries
      - Linked Geo Data (24/11/2010–06/07/2011). 1.6 M queries
      - Semantic Web Dog Food (16/05/2014–12/11/2014). 1.4M queries
      - British Museum (08/11/2014–01/12/2014) 0.8M queries

Claim C2

- SPARQL queries barely use dbp properties
  - In the English DBpedia
    - We looked for SPARQL queries cointaining **both** dbp and dbo properties
  - Results
    - Out of 1,208,762 distinct queries
      - Only 2,328 queries
    - Out of 3,041 distinct agents (IPs)
      - Only 473 IPs



Claim C3

- The method proposed can enhance DBpedia SPARQL queries
  - More results
- How? See an example

```
PREFIX dbo: <http://dbpedia.org/ontology/>
select ?s ?bp {
   ?s dbo:birthPlace ?bp .
}
```

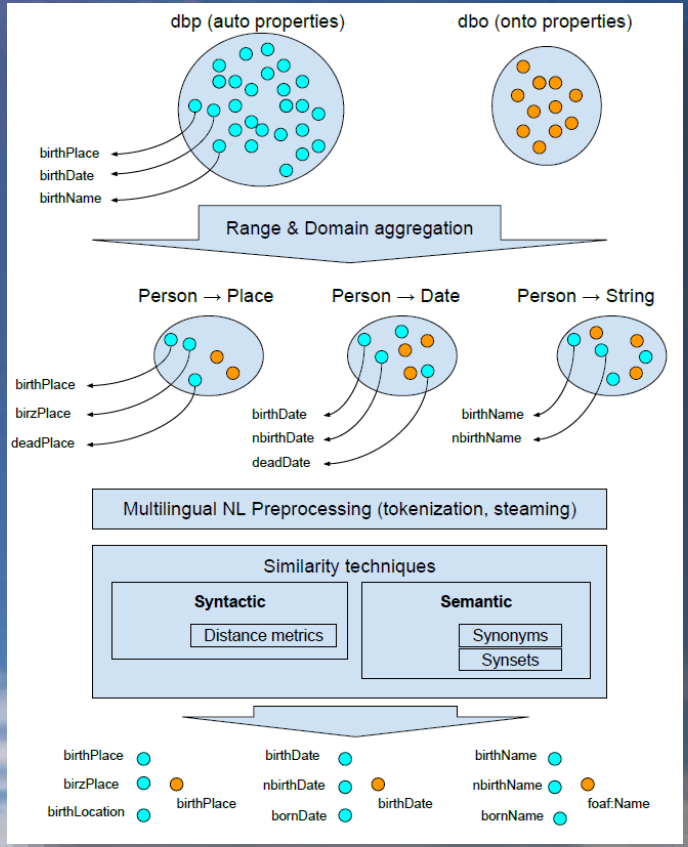
Listing 1: Original SPARQL query

```
PREFIX dbo: <http://dbpedia.org/ontology/>
PREFIX dbp: <http://dbpedia.org/property/>

select ?s ?bp  where {
 ?s ?p ?bp .
 VALUES ?p {
  dbo:birthPlace #typical dbo property
  #Alternative dbp properties
  dbp:birthPlcace dbp:birthplace
  dbp:birhPlace   dbp:bithPlace
  dbp:birtPlace   dbp:biRthPlace
 }
}
```

Claim C3

- The method
- Step 1: aggregate properties into groups according to their domain and range
- Step 2: Multilingual NL preprocessing
- Step 3: aggregate properties by similarity (syntactic and semantic)



Claim C3

- Results. dbp→dbo mapping. Example for dbo:birthPlace for Eng, Spa, Ger.

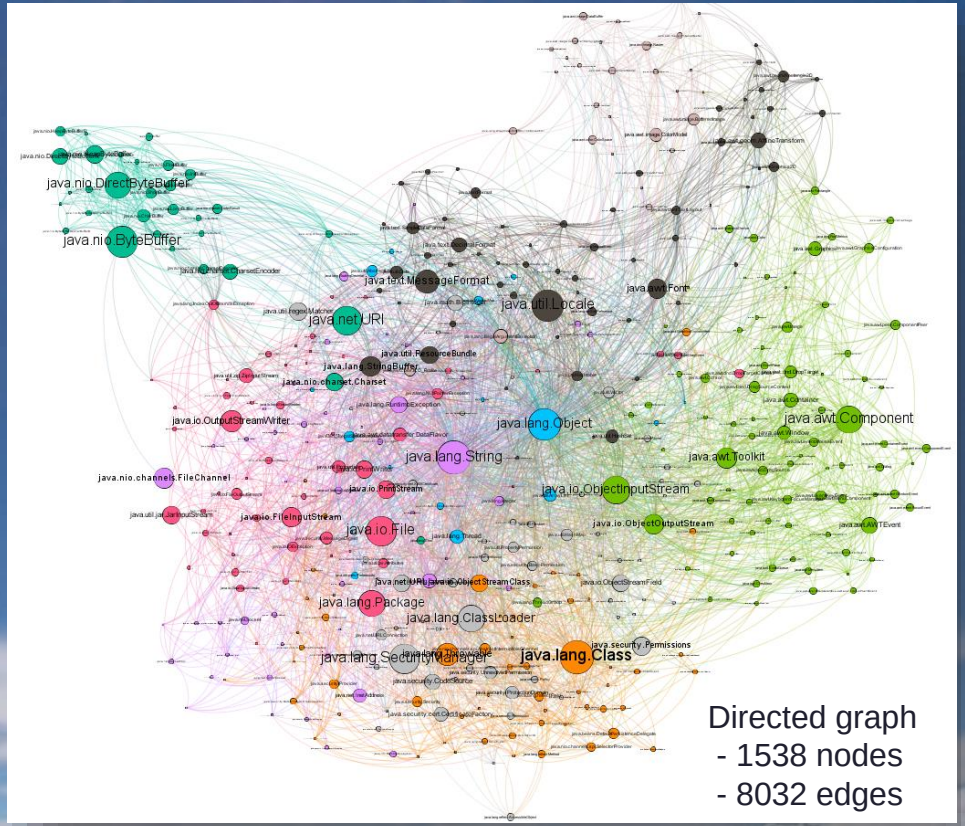
DBpedia	dbo prop	dbp prop		Δ_1
		Syntactic	Semantic	
English	birthPlace	birthPlace birthplace placeofbirth birthPlac birthdplace birthPalce birthPlace PlaceOfBirth laceOfBirth oplaceOfBirth birthPlace. birthPlacE birthPalce birthPlae birthPace birthPlaxe birtPlace birthPlcace bithPlace brithPlace nbirthPlace birthplace birghPlace birthdplace biRthPlace birth placebirth placeOfBirth placOfBirth birthPlaceOf birthPlae	cityofbirth cityofbirthPlace cityOfBirth birthLocation	350%
Spanish	birthPlace	lugarDeNacimiento lugarNacimiento lugarNacimiento lugarnacimiento lugardenacimiento lugarNacimento lugarNaciento	ciudaddenacimiento ciudadDenacimiento paisdenacimiento paisNacimiento birthPlace birthplace placeOfBirth	221%
German	birthPlace	geburtsort birthplace birthPlace placeOfBirth placeofbirth	geburtsland countryofbirth	134%

Future work

- Ideal objective: a stable mapping for a given language
  - Drawbacks
    - Fine tuning of algorithms and parameters
      - This requires users validation
    - Detailed study of precision and recall
      - For many common classes
      - For many common SPARQL queries

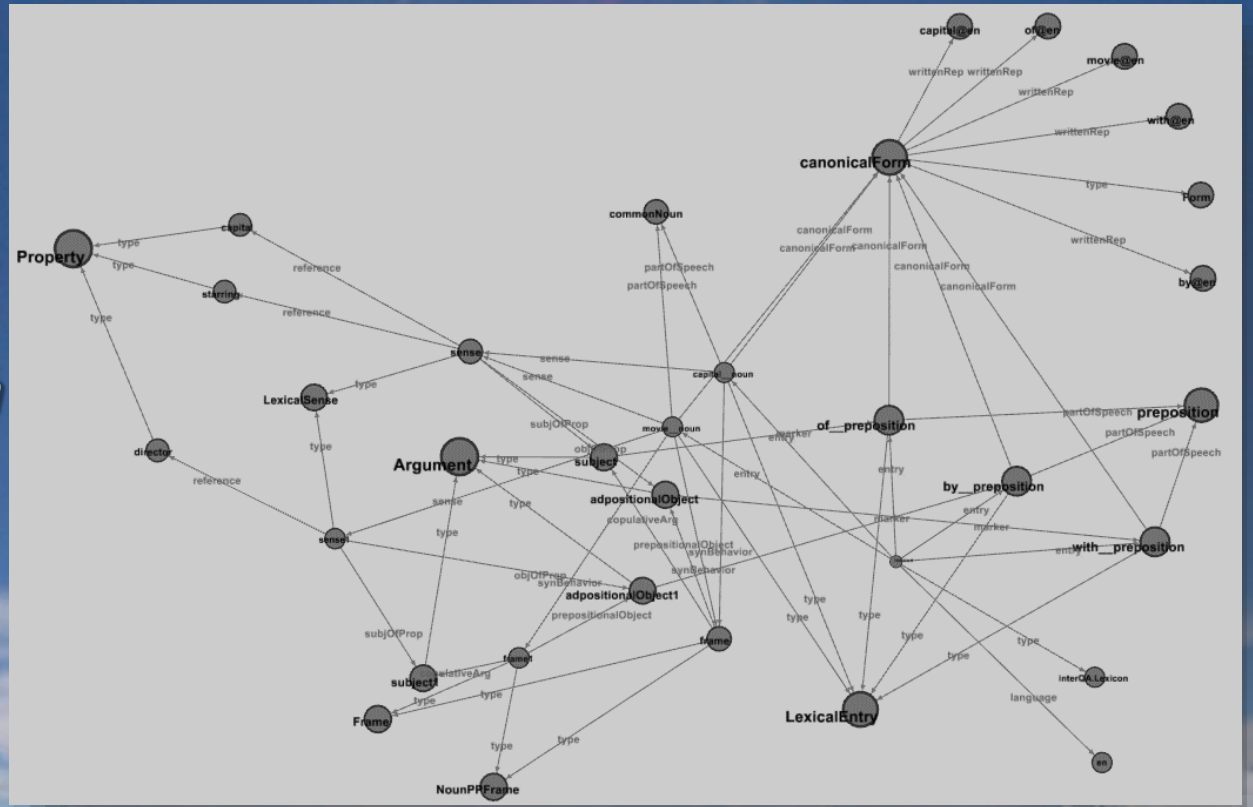
And another thing…

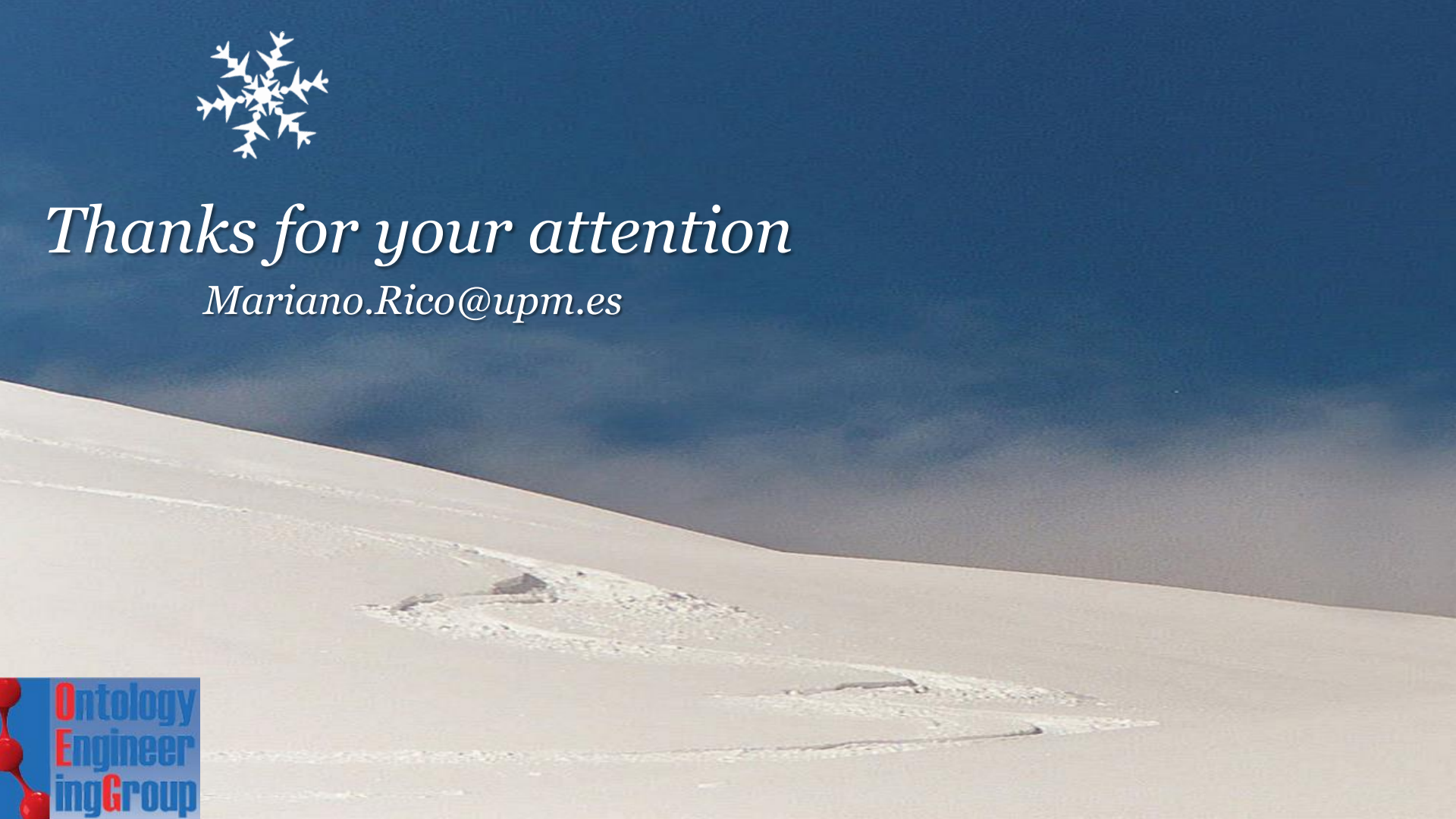
- Visualizing large datasets
  - E.g. The Java Compile-Time Dependency graph
    - Nodes are Java classes and directed edges are compile-time dependencies between two classes.
    - The data provided contains the dependencies for all classes under the java.* packages for JDK 1.4.2



And another thing…

- Visualizing (very) large datasets
  - E.g. a *Lemon* dataset




Thanks for your attention
Mariano.Rico@upm.es
Ontology
Engineer
ingGroup