the Web of Data

Daniel Vila-Suero

Ontology Engineering Group, Universidad Politécnica de Madrid

Campus de Montegacedo s/n, Boadilla del Monte 28660, Madrid, Spain

dvila@fi.upm.es

The Web has witnessed an enormous growth in the amount of semantic information published in recent years, stimulated to a large extent by the emergence of Linked Data.

The so-called **Linked Open Data** (LOD) cloud presents a great opportunity for building a truly **universal network of data** as opposed to the current set of disconnected data "silos" where useful data is kept within the organizations' boundaries.

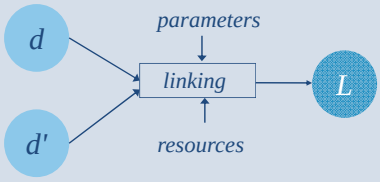
However, **methods** and **techniques** need to be developed in order to avoid that this emerging network of data builds upon "linked data silos" where datasets in **different natural languages remain disconnected**.

*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…***

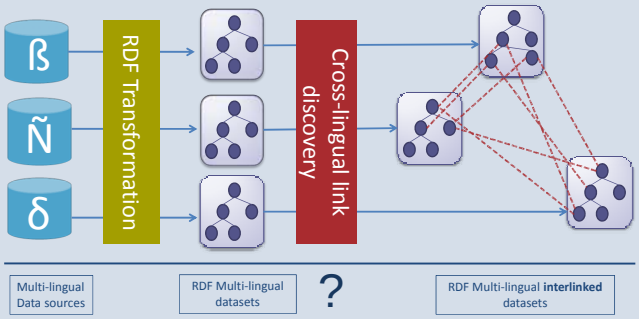
Tim Berners-Lee

Research problem

Finding **relationships** between RDF resources within **different datasets** and across **different** natural **languages**



Where D and D' are RDF datasets in different natural languages and L is a set of links between resources from D and D'



Current situation

Techniques*

Granularity	Type of evidence	Source of evidence	Techniques
Value matching	Data-level	Internal	String similarity metrics
		External	Keyword search
	Knowledge-level	External	Linguistic resources
			Formal domain models
Instance matching	Data-level	Internal	Aggregation methods
		External	Transitivity of owl:sameAs
	Knowledge-level	Internal	Ontological restrictions
		External	Ontology mappings
Dataset matching	Data-level	Internal	Similarity propagation
		External	Graph analysis
	Knowledge-level	Internal	Ontology matching
		External	Ext. repositories: mappings and LD

* categorization based on the results presented in Datalift D4.1: "Methods for automated dataset interlinking"

Systems

- Existing systems use a combination of these techniques.
- Very few systems take into account that resources can be described in different languages. All use **translation services**.
- **Automation** of existing systems **can be** further **improved**.
- Few systems take advantage of ontology matching techniques.

Other considerations

- Clear **definition** of **the meaning of mappings** is needed (it ranges from weak correspondences to strong relationships of identity: skos vs. owl:sameAs)
- Most of the solutions evaluate the mappings based on **similarity techniques** but they are used to generate strong identity **owl:sameAs mappings**

Next steps

We have already generated cross-lingual links using existing systems in several domains: datos.bne.es (libraries), geolinkeddata.es (geo), etc.

Taking these results as a starting point, next steps are:

1. BENCHMARKING: Existing systems and techniques in both domain-specific and general settings.
2. PROPOSAL: Based on the analysis of existing techniques and systems, propose a solution involving **ontology matching** and **localization techniques**.
3. IMPLEMENT, EVALUATE, REFINE

Acknowledgements

This work is supported in part by the Spanish Ministry of Science and Innovation under Grant No. TIN2010-17550 for the **BabelData** project, and by the joint- project with **National Library of Spain** "Generación de datos enlazados a partir del logo de autoridades y bibliográfico"