Data Source Interconnection/Integration

Gilles Falquet

2019

Accessing different data sources

Main interest of SW/Linked data:

applications may access/discover data from several sources

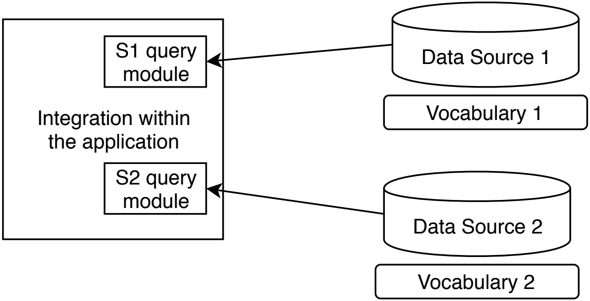
Heterogeneity

- Normally an RDF graph is uniform in terms of vocabulary
- Heterogeneity occurs when an application needs data from different sources
- Different sources often use/refer to different vocabularies

Using Multiple Sources in Applications

Direct access to the sources ⇒

- several data access modules in the application
- the application must deal with different vocabularies/representations
- internal data integration

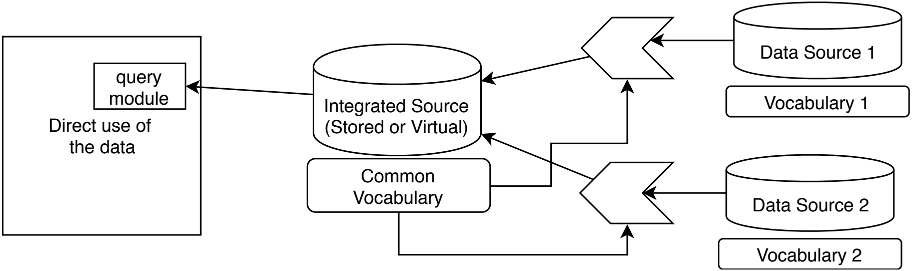


Using Multiple Sources in Applications

External integration ⇒

the application sees

- one vocabulary
- one or more sources that use the same vocabulary
- does not care about integration



Data Integration or Interconnection

Goal: provide data users (applications) with a unified view of the data

The view may be

- fully computed and stored in a repository (data warehouse)
- virtual: computed on-demand by
  - a central "mediator" system
  - a set of "wrappers", one for each source

The Data Integration Problem

How to build an integrated view from heterogenous sources?

1. What are the soures of heterogeneity?
2. How to resolve each type of heterogeneity?

Syntactic heterogeneity [1]

When two vocabularies are not expressed in the same modelling language.

- RDFS,
- XML Schema,
- OWL,
- First order logic,
- Relational schema,
- ...

A one to one translation is possible only when the target languages has an equivalent or higher expressive power

- RDFS to OWL is OK
- not OWL to RDFS

[1] Adapted from , S. Ontology Alignment in the Urban Domain in Ontologies in urban development projects (2011) G. Falquet, C. Métral, J. Teller, C. Tweed (Eds), Advanced Information and Knowledge Processing, Springer, 2011.

Terminological heterogeneity [2]

Variations in names when referring to the same entities in different vocabularies.

Can be caused by the

- use of different natural languages, e.g. Paper vs. Articulo, different technical sublanguages, e.g. Paper vs. Memo,
- use of synonyms, e.g., Paper vs. Article.

[2] ibid.

Conceptual heterogeneity [3]

Differences in modelling the same domain of interest.

Also called *semantic heterogeneity* or *logical mismatch*

Causes

- use of different axioms for defining concepts
- use of totally different concepts
    - geometry axiomatised with points as primitive objects or with spheres as primitive objects.

(Benerecetti et al. 2001) identifies three reasons

- difference in coverage,
- difference in granularity
- difference in perspective.

[3] ibid.

Semiotic heterogeneity [4]

Also called **pragmatic heterogeneity**

This heterogeneity is concerned with how entities are interpreted by people.

How different people, in different contexts, interpret what is *not explicitly stated/defined* (implicature) in the vocabulary.

How apparent ambiguities are solved.

[4] ibid.

Semantic/semiotic conflicts

Goh et al. [5] identify three main causes for semantic heterogeneity:

- Confounding conflicts occur when information items seem to have the same meaning, but differ in reality e.g., due to different temporal contexts.
- Scaling conflicts occur when different reference systems are used to measure a value. Examples are different currencies.
- Naming conflicts occur when naming schemes of information differ significantly. A frequent phenomenon is the presence of homonyms and synonyms".

[5] Cheng Hian Goh, Stephane Bressan, Stuart Madnick, and Michael Siegel, Context Interchange: New Features and Formalisms for the Intelligent Integration of Information, ACM Transactions on Information Systems, Vol 17, No. 3, pp. 270-293, 1999.