Haslhofer, Bernhard

Uniform SPARQL Access to Interlinked (Digital Library) Sources

Looking at the Web today, we notice that it has become the largest source of information for practically any topic. People publish their photos on Flickr, share their videos on YouTube, write Blogs, and contribute their knowledge to online encyclopaedias such as Wikipedia. Institutions, ranging from broadcast companies[1] to libraries and archives[2], are opening their repositories and make their contents and – in many cases – also structured metadata available on the Web. For end users it is desirable to search and browse related contents of interest in the available sources via a single access point, no matter if the source is an online image database, a Digital Library system, or some other kind of content repository.

In general, these various kinds of sources are decoupled from each other. Uniform access to their contents and metadata is impeded by the heterogeneous nature of the systems and incompatible information models. An important *first step* to eliminate part of the heterogeneities and to connect the related (meta-)data stored in these sources is currently performed by the Linked Data Initiative: recently, sources such as Wikipedia[3], GeoNames[4], or DBLP[5] have exposed their data in RDF, provided a SPARQL query interface, and have connected data that wasn't previously linked. More sources, especially structured sources from the Digital Library domain, are expected to follow in the near future. But even if all data sources were interlinked and have their metadata exposed in RDF, the central problem which impedes uniform query access to these sources is yet unresolved: the semantic and structural incompatibility of the metadata in the data sources. Taking the digital libraries domain as an example, standards for metadata schemes (e.g. MARC-21, MODS, DC, VRA, PRISM, ONIX, etc.) exist but do not solve the problem simply because of the fact that many sources do not adhere to any standard or implement it only partially.

In this presentation, we therefore focus on a solution for providing uniform access to Digital Libraries and other online services. In order to enable uniform query access to heterogeneous sources, we must provide metadata interoperability in a way that a query language – in this case SPARQL – can cope with the incompatibility of the metadata in various sources without changing their already existing information models. Therefore, we propose a metadata mapping approach consisting of the following building blocks:

- A *mapping formalism* which allows for the declaration of complex mapping relations between information model elements, both on a semantic and structural level. The formalism also

[1] e.g. BBC (http://creativearchive.bbc.co.uk/index.html)
[2] Open Archives Initiative (http://www.openarchives.org/)
[3] Wikipedia data set exposed via DBpedia.org (http://dbpedia.org/docs/)
[4] Geonames Ontology and RDF representation (http://www.geonames.org/ontology/)
[5] DBLP Computer Science Bibliography (http://dblp.uni-trier.de)

considers the instances of an information model, an aspect that is often neglected by mapping approaches.

- A *query transformation* mechanism for rewriting incoming SPARQL queries according to a mapping specification.
- A machine-readable and *exchangeable format* for integration specifications so that mappings can be registered, discovered, and reused at schema registries.

We believe that SPARQL could be the future for enabling integrated access to online metadata sources. Besides being a query language it also defines a Web-based protocol for transmitting queries and receiving results. Different from other query languages (e.g. XQuery, SQL) it is concept based, thus formulated over real-world concepts, and not expressed over document or table structures. With our mapping approach we want to provide the mechanism to adapt SPARQL queries to the semantic and structural heterogeneities of the data sources so that at least in a certain integration context, uniform access to distributed, perhaps interlinked, data sources can be established.