Opening the Data in Documents on the Web

John Snelson, Semantics Lead Engineer
MarkLogic Corporation
<john.snelson@marklogic.com>

MarkLogic Overview

MarkLogic is a Schema-less Enterprise NoSQL database, coupled with powerful integrated search and flexible application services. MarkLogic has over 300 enterprise and government customers who trust us for mission-critical Big Data applications.

The Web of Documents Contains Valuable Data

RDF excels as a way to publish open datasets and to easily combine valuable data from different sources. The value of the "Web of Data" is plain, but the route to that ideal has proved problematic.

Our customers find great value in being able use MarkLogic to handle poly-structured data - a mixture of highly structured data alongside human language document content, often marked up to varying degrees. Estimates have been made that 80% or more of the world's information is poly-structured, human oriented content - this is the substance of the so-called "Web of Documents".

Although containing large amounts of information, only a small amount of the most structured data that is contained in these documents has been made available to the typical tools of the open data movement. RDFa and microformats have enabled some of this, but the use of these formats has sometimes suffered due to lack of immediately apparent motivation.

Ultimately one of the biggest problems is that RDF is suitable for modelling facts, but not human language content. This leads to an ETL mentaility to pulling out facts from a document, leaving much valuable information out of the queryable domain.

Opening the Data in Documents

The techniques for modelling, enriching, and querying human language content are well known - markup technologies including HTML and XML are designed specifically for that problem. No one familiar with the "Web of Documents" should be surprised that full-text search is also needed to make any sense of this kind of data, in combination with structured query capabilities.

MarkLogic provides sophisticated full-text search capabilities to its customers, a feature which typically forms the back bone of any data application that is built on top of the database. This is provided in the context of structured query capabilities over the stored XML content using XQuery, giving access to all the information in polystructured data.

Many MarkLogic customers also choose to use semantic technologies with or alongside MarkLogic, to take advantage of graph merging, querying, and inference. It's important to realise that data modelling is enriched by using both RDF and XML, and querying is most effective using SPARQL, XQuery, and full-text queries. The open data ecosystem could be enriched by investing in a greater integration between these technologies.

Data Pragmatism

MarkLogic customers often describe a process of "pragmatic" data modelling that mixes RDF and XML according to application need. Structured information can be moved from the XML into the RDF to enable graph navigation and greater flexibility in combining data sources. Human langauge content remains solely in the XML, or is put into the RDF only in summary.

However other reasons also exist for leaving strutured data in XML formats - the most common reason being efficiency of querying and retrieval. RDF is a highly granular data format, and querying it with SPARQL necessarily involves a large number of joins. Using joins to reconstruct information can often be avoided by using intelligent denormalization to XML based around common access patterns.

Since different open data based applications will require different access patterns over data and documents, it seems valuable to invest in integration between RDF and XML. There is great potential for methods of conversion between the two formats, for ways to effectively embed them in each other, and for ways to query across both using their native query languages SPARQL and XQuery.