Briefing paper:

Metadata Schema Registries

Emma Tonkin

What is a metadata schema registry?

To explain this, we will have to begin with a brief overview of some specialist language used in the metadata world.

Metadata itself is often called 'data about data', information that typically describes what a document is, what it is about, and where one may locate it (Zeng & Jin, 2008). There are various metadata standards, such as Dublin Core and IEEE LOM. Each standard contains an element set or scheme that defines the structure and semantics of elements, such as the DC Metadata Element Set, now referred to as Simple DC, that defines 15 core elements that can be used to describe distributed information resources.

Metadata elements are the actual data fields; title, author, etc. The scheme in which these elements are placed may be a flat structure, a hierarchical structure, etc. Elements may contain encoding conventions – for example, dates may be encoded in ISO 8601 – and content limitations, such as the need to use controlled vocabularies such as LCSH. These encoding conventions are referred to by Zeng & Jin as value encoding schemes.

The defined element sets may be represented in XML; Zeng & Jin tell us that the resultant structures are metadata schemas. Generally, schemas are a set of statements, expressed in any of several data definition languages (XML Schema Definition, for example) that describe the organisation or structure of some form of database or information record, so this is a specialised use of a general term.

Another related piece of vocabulary worth defining here is terminologies; terminologies are sets of vocabularies available for use in information systems and services (quoting http://www.ukoln.ac.uk/projects/trss/)

Let's look at a brief example:

title	Metadata schema registries
creator	Emma Tonkin
publisher	UKOLN
url	http://some-site.com/content.pdf
dc.subject.lcsh	Electronic data processing

The metadata schema used here is the Dublin Core Metadata Element Set. The elements used are title, creator, publisher and url. The URL is encoded according to RFC 2396, Uniform Resource Identifiers (URI): Generic Syntax. The subject information provided (admittedly completely the wrong classification for this piece!) has been taken from the Library of Congress Subject Headings (LCSH), a controlled vocabulary and one of many possible terminologies.

A metadata schema registry has initially to describe the structure and semantics of elements, describing the number of data fields, the structures they are in and the encoding conventions (value encoding schemes) that they may contain. One question to ask is how much of this should be held internally to a given registry.

One might do a partial job of schema description and refuse to describe some of this information internally – for example, merely stating that a date element uses ISO 8601, rather than containing an actual description of what this means.

One might do an extremely partial job and simply create a list that describes metadata schemas in a very vague way, providing no details at all except perhaps for a brief description of the intended use of the schema:

"DCMES Dublin Core Metadata Element Set http://the_appropriate_web_site"

The decision has to do with the use cases that one has in mind – supporting humans who are looking for information about the various schemas on offer? Supporting machines in the task of deciding whether a given data record is compliant to a given schema?

Zeng & Jin (p.274) provide the following general definition of metadata registries: "a metadata registry collects data regarding metadata schemas for reuse of existing metadata terms to achieve interoperability among metadata element sets. The basic components of a metadata registry may include identification of data models, elements, element sets, encoding schemas, application profiles, element usage information, and element cross-walks. The primary functions of metadata registries include registering, publishing and managing schemas and application profiles, as well as making the registry easily searchable within the registry. A registry also provides services for crosslinking and crosswalking among schemas and application profiles."

The Application Profile Complication

In the Dublin Core world, people also speak about application profiles. Originally, the application profile was a description of a specific form of usage of a schema, usually a narrower set of guidelines than the original schema allows. The point of this was to ensure that people with a particular use case or set of needs could customise schemas to their needs, by explaining to their community how this should occur.

Nowadays, application profiles have a number of slightly more technical definitions; they may describe the extent to which one retains the original schema; they may describe 'an assemblage of metadata elements selected from one or more metadata schemas, combined in a compound schema' (Zeng & Jin, p.112), optimised for a given local application. It may be that APs even contain elements from multiple schemas as such – that is, elements from IEEE LOM within an AP that itself is in the form of a DC schema.

In other words, for those of us who come from the world of software engineering, application profiles are essentially component-based software engineering applied to metadata; that is, the basis of this design is that elements are decomposable into unique components (there is no interdependence between them), and it is therefore possible to build 'mix and match' novel AP from these component elements with no loss of coherence or interoperability. Note: this equivalent-semantic reusability assumption is extremely major, fairly novel in this arena, and fairly untested.

APs may also be stored within metadata schema registries. They are assemblies or customisations of existing elements from various schemas, so in order to be able to describe an AP effectively, the metadata schema registry must contain all of the elements and schemas to which an AP refers. This means that to assemble an AP successfully, somebody somewhere first has to input every single element/schema/value encoding scheme (vocabularies and standards) to which that AP refers.

Metadata Schema Registries vs Terminology Registries

One can draw out the relationship between metadata schema registries and terminology registries using a Venn diagram such as the example shown to the right.

To some extent, the terminology registry is either a subset of the schema registry, or could be seen as a service required to complete the task of the schema registry, especially in circumstances in which APs are registered as well. There are scenarios in which one could build two APs, with the only differences between them being in the allowed range of value encoding schemes. If these were not properly represented, they would look identical!

Architecture of a Metadata Schema Registry

There are many possible ways of producing something that does this job. The IEMSR registry approach is to represent the information in an RDF store, importing it from RDF files. Then it may be searched by means of SPARQL, a query language for RDF. This provides a 'ready-made' machine-to-machine interface which in theory can be replicated using any of a large set of open source or commercially available tools. Practically, the hard part of this problem is to decide what to represent and how it should be represented.

Conclusion

Metadata schema registries are designed to describe, to some level or another, metadata schemas. In practice, this means describing the elements of that metadata schema and their value encoding schemes, and the structure(s) permissible according to that schema.

This is a difficult problem for a number of reasons, primarily because the goalposts are reorganised every few years as the various standards organisations change the way in which they approach their metadata schemes, alter their schemas, and enable novel and more complex encodings and interactions between them. There are also several organisational issues, almost all of which are social rather than technical. Application profiles in particular often come to represent a set of aspirations rather than a common-sense set of practical guidelines. Hence, we return to the question of 'what, conceptually speaking, is an AP? What is a schema? How can it be effectively modelled?' This is the Achilles' heel of metadata standards – introspection leading ever deeper into modelling and remodelling, with the aim of developing a Grand Unified Theory that encompasses every possibility – and it is a remarkably disruptive process.

One therefore suggests that metadata schema registries should concentrate on a relatively modest description of metadata schemas as they are, taking as their model the following comment by Wilks (2008): “the heart of the issue is the creation of meaning by some interaction of (unstructured language) usage and the interpretations to be given to higher level concepts.“ At any one time, the most useful definitions of metadata schemas are those that most accurately describe the ways in which it is actually used; provided our model is sufficient to describe this, the ideal formal description to be applied at any one time should, arguably, be of very little interest to us.

References

Zeng, Marcia Lei, and Qin, Jian (2008). Metadata. Facet publishing. London, UK.

Wilks, Yorick (2008). The Semantic Web as the apotheosis of annotation, but what are its semantics? IEEE Intelligent Systems May/June 2008.