Topic Maps and MLA Information Resources

Author:
Richard Light

What is the Semantic Web?

The Semantic Web ("SW") is a broad concept, but Tim Berners-Lee sees its essence as "a universal medium for the exchange of data where data can be shared and processed by automated tools as well as by people"[1] or, more simply, a "web of data"[2]. In other words, it will do for data what the original Web did for documents: enables a scale of operation which transcends anything we can imagine today.

Berners-Lee foresees the SW being built up in a number of layers:

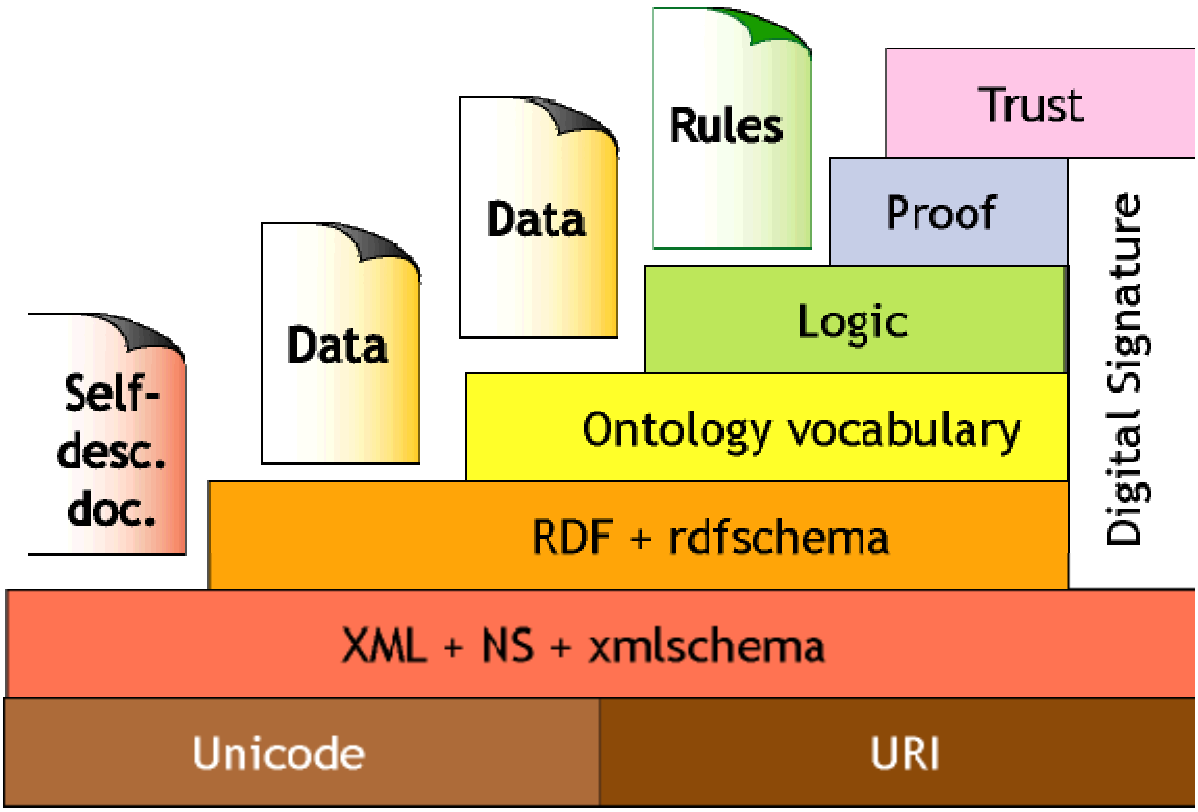

Figure 1: Semantic Web Architecture (http://www.w3.org/2000/Talks/1206-xml2k-tbl/slide10-0.html)

1 http://www.consortiuminfo.org/bulletins/semanticweb.php
2 http://www.w3.org/2001/sw/

The layers are built upon the foundations of today's technology: Unicode, XML and RDF. A standard vocabulary for describing ontologies – the Web Ontology Language ("OWL") - has been formalised as a W3C Recommendation. The intention is to build on this by adding a framework for querying data, and support for machine-based logic (essentially first-order predicate logic).

Topic Maps ("TMs") are another initiative, which has similar goals but a different history. TMs were developed as an ISO Standard (ISO/IEC 13250:2000), based on attempts to generalise the process of indexing online resources. The TM standard is expressed as an abstract Data Model[3], and an XML syntax for interchange is also under development[4].

Initially there was concern that RDF/OWL and Topic Maps were competing to solve the same problem. However, representatives of the two communities have worked to establish a large degree of interoperability between the two frameworks.

In this paper I will concentrate on the potential use of Topic Maps as a means of expressing concepts of interest to museums, libraries and archives.

Topic Maps

Topic Maps provide a mechanism for encoding knowledge and connecting this encoded knowledge to relevant information resources. They consist of:

- Topics: the concepts themselves
- Associations: assertions about a relationship between two or more Topics
- Occurrences: resources which relate to a specified Topic

Thus, a "museum" view of the world and everything in it could be expressed as a Topic Map. The concepts which underpin any particular framework (such as the CRM, or SPECTRUM, for

3 http://www.isotopicmaps.org/sam/sam-model/
4 http://www.isotopicmaps.org/sam/sam-xtm/

museum information) can also be expressed as Topics. Individual Topics can then be assigned as instances of a particular generic concept. It is also possible to indicate superclass and subclass relationships between Topics. These relationships are transitive, which means that if A is a subclass of B, and B is a subclass of C, then one can infer that A is a subclass of C. Apart from these built-in relationships, Topic Maps can define and use any type of relationship which is meaningful in the context of the knowledge framework being described.

Topics have *names*, which are used to communicate their significance to human readers. Examples of names would be "Light, Richard Brian" or "Cambridge". However, these names are not guaranteed to be unique, and so in addition each Topic has one or more *item identifiers* or *subject identifiers*. The Data Model states that if two or more Topics share the same item or subject identifier, they should be considered to be the "same" Topic, and all of their properties should be merged to create a single (new) Topic. Item and subject identifiers are just strings (strictly speaking, IRIs), but subject identifiers refer to *subject indicators*: information resources which attempt to help human beings by unambiguously identifying the subject represented by a Topic.

The concept of *scope* is central to Topic Maps. Scope represents the context within which an assertion is valid. Thus one might scope the name "Suomi" with the language "Finnish" in the context of the Topic representing the country "Finland".

Playing with Topic Maps

Given that Topic Maps are meant to contain data which is queriable, it makes sense to create them by converting existing data resources, rather than create them by hand. One straightforward method of achieving this is to extract existing data in XML format, then use an XSLT[5] transformation to convert it to the Topic Map interchange format.

5 http://www.w3.org/TR/xslt

Alternatively, tools exist which allow the manual creation and editing of Topic Maps. For example, TM-Tab[6] is a tab-widget plug-in for Protégé[7] which enables the use of Protégé as a Topic Map editor.

Once a Topic Map has been created, it is useful to have a tool to check its validity and allow users to browse and query it in various ways. The Omnigator[8] from Ontopia fulfills this need, providing a web-based interface which checks that a Topic Map is correctly structured, then allows the user to browse the Topic Map, view topics in a graphical format, or query the Topic Map using the tolog language described above.

Work has begun on a Query Language for Topic Maps (TMQL), but this is still at an early stage. For practical experimentation I intend to use a Prolog-like language, tolog, which has been developed by Ontopia A/S, a Norwegian company specialising in Topic Map products and support. Tolog "is a logic-based query language, which means that the basic operation is for the user to ask tolog in which cases a certain assertion holds true, and tolog will then respond with all the sets of values that make the assertion true"[9]

Museum catalogues as Topic Maps

Museum Collection Management Systems and catalogue databases are full of "assertions". For example, a catalogue entry might state that object BCRTM:1930.1 was made by the clockmaker Andrew Dickie of Edinburgh, about 1866, and that the same object was purchased from Francis Ferry on 16 April 1930.

In Topic Map terms, each entity of interest within these assertions is a Topic. Here we have an object, two people, a place, an occupation, and two dates. All of these would normally be Topics (although the dates could alternatively be treated as Occurrences), e.g.:

6 http://www.techquila.com/tmtab/index.html
7 http://protege.stanford.edu/
8 http://www.ontopia.net/download/index.html
9 http://www.ontopia.net/omnigator/docs/query/tutorial.html

Topic

Instance of: person

Item identifier: clockmaker-Andrew-Dickie-of-Edinburgh

Topic name: Andrew Dickie

Topic

Instance of: place

Item identifier: Edinburgh-Lothian-Scotland-U.K.

Topic name: Edinburgh

Each assertion which makes a link between two or more entities is an Association. Here we have one Association making statements about the production of the object, and one describing its acquisition by the museum. Associations consist of a *type* and a number of *roles*:

Association

Type: production

Role (type: maker) Andrew Dickie

Role (type: place) Edinburgh

Role (type: date) 1866

Association

Type: acquisition

Role (type: method) purchased

Role (type: source) Francis Ferry

Role (type: date) 16 April 1930

If the source database record is an addressable resource, it can be cited as the source of these Associations. In order to do this, it is necessary to *reify* each Association by creating a Topic whose subject is the specific event or activity which the Association represents. The source database record can then be cited as an Occurrence of each of these topics.

A museum frame of reference

As mentioned above, every concept in a Topic Map has to be declared as a Topic. This includes the low-level terms we have casually used in the examples above, such as "production", "maker" and "place". These concepts could be declared locally each time an institution creates a Topic Map. However, the scope for interoperability would be greatly enhanced if agreement could be reached on the use of museum standards such as the CIDOC Conceptual Reference Model (CRM) or SPECTRUM for the naming of low-level "museum" topics.

In general terms, this type of consensus can be achieved by the declaration and use of *Published Subject Identifiers* ("PSIs"). This involves a "publisher" issuing a set of subject identifiers, i.e. URIs, covering the Topics in question. These URIs should:

- resolve to *Published Subject Indicators*, which:
    - provide both human-readable and machine-processable metadata
    - indicate that they are intended to be a PSI
    - identify the publisher of the PSI
- be stable
- be permanently available

In my past projects I have created a Topic Map representing the CRM, but the act of publishing a set of PSIs for the CRM is something which only the body responsible for the CRM (CRM-SIG, or perhaps CIDOC) has the authority to do. Also, it should be noted that while it would be convenient and helpful to write and publish a Topic Map containing the Published Subject Identifiers (and especially so if this Topic Map declared the superclass-subclass relationships between CRM concepts), the Published Subject Indicators themselves do not form part of this Topic Map, and are stored elsewhere.

In a similar vein, the MDA might choose to publish the SPECTRUM Units of Information as a set of PSIs.

Once these resources are in place, it becomes possible for different Topic Maps to share information about museum-related subjects without ambiguity, since they can use the same PSIs to represent these concepts. The standard Topic Map processing model will then automatically treat them as instances of the "same" concept.

Other "global" concepts/subjects

While museums have a particular view of the world, they share many aspects of understanding with other branches of human endeavour. It therefore makes sense to look for existing Published Subjects covering areas which are common, rather than museum-specific.

One example I came across while preparing this paper is that there is a set of Published Subjects for dates according to the Gregorian calendar[10]. This includes support for dates down to the year, month or day, e.g.:

http://psi.semagia.com/iso8601/2006 for the year 2006

http://psi.semagia.com/iso8601/2006-06 for June 2006

http://psi.semagia.com/iso8601/2006-06-20 for the 20th June 2006

In addition, Semagia publish a set of date/time subjects, such as:

http://psi.semagia.com/datetime/month

http://psi.semagia.com/datetime/March

The existence of these PSIs means that it makes sense to record dates as Topics, rather than storing them as a string value within an Occurrence. This makes it possible to pose Topic Map queries with a temporal aspect.

Turning to other "global" subjects, two obvious possibilities are people and places.

10 http://psi.semagia.com/iso8601/

There are already published resources detailing people, for example Getty's ULAN (Union List of Artist Names)[11]. However, these simply serve to highlight the scale of the task if one is to attempt to produce an authoritative source for *all* people: ULAN (the result of a major collaborative effort over a number of years) "only" contains about 120,000 entries. This suggests that it will only be feasible to produce Published Subjects for "well-known" people, and that there is scope for many separate initiatives to identify and catalogue the people of interest to a specific discipline, geographical area, etc.

Place names are a more promising area. Apart from the Getty's TGN (Thesaurus of Geographic Names)[12], which contains entries for 912,000 places, there is the Alexandria Digital Library's Gazetteer[13], which contains no less than 5.9 million geographic names. Both contain relationships between places (e.g. "Burgess Hill is in West Sussex") which could be converted into Associations and so used when searching Topic Maps.

There may be scope for subject-specific published resources. For example, in taxonomy both Species 2000[14] and ITIS[15] have developed significant databases.

All of these large-scale initiatives are database-driven, and so could be adapted to act as a Published Subjects server if their publishers could be persuaded of the benefits in so doing.

"local" subjects; scope

What happens once we have exhausted the possibilities of finding "global" PSIs for the subjects described in our museum database? Does interchange become impossible? On the contrary, this is where one of the design features of Topic Maps – scope – comes into play.

Taking the names given to objects as an example, each museum will probably have its own

11 http://www.getty.edu/research/conducting_research/vocabularies/ulan/
12 http://www.getty.edu/research/conducting_research/vocabularies/tgn/
13 http://www.alexandria.ucsb.edu/gazetteer/
14 http://www.sp2000.org/
15 http://www.itis.usda.gov/

vocabulary of object names. While this terminology may be controlled within the museum, it will not be shared at the semantic level with other institutions – even if they use many of the same names for their own objects. Therefore the museum needs to declare its own set of Subject Identifiers (assuming that it has resources, such as an authority file, which can act as human-readable Subject Indicators), or else use Item Identifiers which are consistent and each uniquely assigned to a single subject (i.e. object name).

While the Topic Map is being used on its own, within the museum, that's all that needs to be done: the "local" Identifiers will do just as good a job as the "global" ones. This is because the control previously exercised while building the source database ensures that the same term is always used to describe the same subject.

However, if you want to merge this Topic Map with others and put queries to the combined resources, you need to do more. Each time a "local" Subject or Item Identifier is used, it must be qualified by a *scope* entry, which indicates that this Identifier is only valid within its source Topic Map.

Once several Topic Maps have been merged, one useful exercise would be to compare the sets of "local" Identifiers, and add a set of statements (Associations) which assert that certain Identifiers describe the same subject. This should allow queries to treat the equivalent "local" Identifiers as interchangeable. Alternatively, by adding another system's Identifier to a Topic, you can cause the two Topics to be automatically merged. This means that it is possible to carry out "bottom-up" harmonization of authority Topics.

In addition to denoting the source of Identifiers, scope can be used, for example:

- to express uncertainty about an attribution (see "about 1866" above)
- to indicate the specific authority for an attribution

As noted above, names for a Topic can be provided in multiple languages, each with a scope specifying which language the name belongs to.

Other types of source

Apart from museum catalogues and other databases, it is possible to extract Topic Map information from structured prose sources, such as:

- descriptive museum publications, e.g. exhibition catalogues
- archival descriptions (ISAD(G); possibly encoded using EAD)
- general humanities resources (possibly encoded using TEI)

In each case, Topics of interest within the resource need to have been marked up in some manner. For word processed documents (exhibition catalogues; textual archival descriptions) this might be done by use of the indexing facility; EAD and TEI have their own methods for indicating names, etc., of interest. It would also be helpful to convert the source to XML, to facilitate the extraction of the Topic Map information. This can be done, for example, by saving word processor documents in Open Office format.

One weakness of this type of resource – you might argue that it is a weakness in most metadata schemes – is that the Topics mentioned in the resource are all independent of each other: there is no means of constructing the sort of interesting multi-faceted Associations that can be derived from structured catalogue data.

Conclusions

The Topic Maps standard (ISO 13250) offers a stable data model, which can be exploited to develop new and more powerful ways of searching and browsing "museum" information resources. It can be used in a number of ways, for example:

- extracting Topic Map information from an existing database and building a separate, free-standing Topic Map
- developing a new interface to an existing database, thus turning it into a "Virtual Topic Map"
-

- extracting Topic Map information from a number of source databases using a protocol such as OAI, and merging the results to produce a combined Topic Map

Agreement on the semantics of information is central to the effectiveness of any combined Topic Map. This should ideally be achieved through the use of Published Subjects. This goal can be achieved at a basic level by developing PSIs for museum standards such as the CRM and SPECTRUM. Where possible, publishers of existing authorities should be encouraged to publish PSIs as well, but "bottom-up" merging of local authorities is also quite possible.

The use of the concept of "scope" allows effective searching of a Topic Map containing material from a number of sources, even where its contents conform to different naming schemes.

Topic Maps can be used to develop searchable cross-disciplinary resources, taking material from archives, museums and even libraries and using a variety of sources: databases, word processor documents and XML documents.