An RDF Model for Multi-Level Hypertext in Digital Libraries

Gudrun Fischer, Norbert Fuhr

University of Dortmund, Germany

1 Multi-level hypertext in digital libraries

A core concept of the Semantic Web [1] is to enrich Web documents with machine-readable metadata (i. e. data describing the resources). Such metadata and the corresponding documents are already provided by a growing number of digital libraries on the Web. The Open Archives Initiative Protocol for Metadata Harvesting ([VdSL01]) provides a standard for harvesting metadata from any digital library that implements the protocol. The services developed on top of the Open Archives protocol, e. g. in the EU project Cyclades[2], can thus use metadata in a variety of formats and from many different libraries.

For searching and browsing with various schemas available, [Fuh99] identifies the following levels according to the concept of *Multi-Level Hypertext* ([ACG91]):

- The *schema level* contains information about the metadata formats available at a digital library, i. e. which features of documents are distinguished in which format. These features are called *attributes*.
- At the *attribute value level*, the user looks at existing values of document attributes, e. g. the names of the authors whose publications are available at the digital library, without having to look at individual documents or at data about individual documents.
- At the *metadata level*, the user explores data about individual documents, without reading the documents as well.
- The *document level* is the level where the user finally looks at the documents themselves. This level will not be considered in the following.

The Open Archives protocol provides the means to access metadata from many different digital libraries in different schemas, but for enabling users to access all levels of the multi-level hypertext thus induced (i. e. also the information about the schemas themselves), we need a representation that allows for searching and browsing in and between these levels in a uniform way.

[1] http://www.w3.org/2001/sw/Activity, http://www.semanticweb.org/

[2] http://www.ercim.org/cyclades

2 Existing approaches

Many digital libraries have one or more built-in metadata schemas: Which attributes of documents can be searched can usually be determined by the search interface of a digital library. In some cases, also possible values for the attributes are offered to choose from when formulating a query. Thus, the higher levels of multi-level hypertext are often present implicitly, but they are not exported in a well-defined way.

XML is a well-known standard for sharing structured and semi-structured data across applications, and there are XML DTDs and schemas for several metadata schemas. With XML schemas, both the metadata and its schema are represented in XML, and thus XML query languages can be used for querying both. However, there is no defined way to formulate a search that combines both schema level and metadata level conditions (e. g. *Search in all attributes of the data type "Name" for the value "Salton"*).

RDF ([LS99] is a resource description language developed for the Semantic Web. RDF statements consist of a subject (the resource being described), a predicate (also called a property), and an object (which is the value of the property). They are commonly represented as triples (subject, predicate, object) or as graphs, with the subject and object resources as ellipses, and the predicate an arc from subject to object. RDF Schema ([BG02]) adds to this the concept of classes (using the property *rdf:type*), inheritance (via the properties *rdfs:subClassOf* and *rdfs:subPropertyOf*), and the possibility to restrict the domain and range of a property (using the properties *rdfs:domain* and *rdfs:range*). In the following, we will not distinguish between RDF and RDF Schema, but use the name RDF for both. One key feature of RDF is the possibility of adding statements to a resource description in a different document without having to change the original description. Thus, to model the complete multi-level hypertext, existing RDF representations of individual metadata schemas can be combined unchanged to represent the metadata level, leaving only the schema level to be represented in addition, as we will show in the next section. Thus, an RDF model seems to be the representation approach of choice.

3 An RDF representation for multi-level hypertext

Metadata can be seen as a set of attributes (properties) of a resource together with their values. Thus, it is straightforward to express metadata about a resource as a set of RDF statements of the form (resource,attribute,value). Most existing RDF representations of metadata schemas follow this approach of representing attributes as properties, e. g. [BBM01] and [KS01] (both Dublin Core) and the schema for the semantic units in the Basic Semantic Registry (BSR[3]).

Building on this, the fact of one metadata attribute refining another can be expressed by defining a *subPropertyOf* relationship between them. Examples are again [KS01] (for qualified Dublin Core) and the BSR schema. The lower part of figure 1 shows an example

[3] http://www.gils.net/bsr-gils.rdfs

of a Dublin Core record. The properties from the *dc* namespace are representations of unqualified DC elements, the property *dcq:replaces* is an example for a refined DC element from qualified Dublin Core.

If a metadata attribute is expressed as an RDF property, then the type of its values is expressed by its range. Analogously, the type of resources that this attribute can describe is expressed by the domain. For example, in figure 1, the attribute *InformationResource.Name* has the domain *InformationResource* and the range *Name*, i. e. it can only be applied to resources of the type *InformationResource*, and its values must be of the type *Name*. Note that this representation makes use of data types in the form of RDF classes.

So far, we have used only concepts from existing metadata representations, but we have no way of describing the metadata schemas themselves yet. As a metadata schema can be seen as a set of attributes, we define an RDF class *MetadataSchema* and an RDF property *Attribute* with *MetadataSchema* as its domain, and with the *rdf:Property* class as range (see the upper part of figure 1, above the first dashed line). This means that an instance of *MetadataSchema* can have an arbitrary number of *Attribute* arcs which point to those *Properties* that represent the metadata attributes belonging to the schema.

In the example in figure 1, we use this technique to describe the Dublin Core and BSR metadata schemas mentioned earlier. For Dublin Core, we use the representation recommended in [KS01] and additionally define a resource *DublinCore* of the type *MetadataSchema*. For each DC element (i. e. each attribute of the schema), we write an *Attribute* statement as described above. Analogously, we define a resource *BSRGILS* of the type *MetadataSchema* and specify the BSR units (e. g. *InformationResource.Name*) as its *Attributes*. Figure 1 shows a part of the resulting RDF graph.

Note that in figure 1, we have three different layers of instantiation (distinguished by the dashed horizontal lines), which correspond to the *metadata level* (the example DC record), the *schema level* and a conceptual level. The *attribute value level* is not represented explicitly, but can be seen as a partial view of the metadata level with respect to the attributes at the schema level. For the attribute *dc:creator*, this is indicated by the dotted area.

4 Searching and browsing in multi-level hypertext

Browsing a multi-level hypertext means navigating its structure, i. e. following the property arcs of the model, while searching involves the formulation of conditions. If following an arc is viewed as a query on the structure of the data, then browsing can be seen as a special case of query formulation. For querying RDF, various languages are available, and most of them allow queries on the proposed model both in and between the different levels of the represented multi-level hypertext.

Examples for queries (with respect to the data in figure 1) within the levels are:

Schema level: *Which units of BSR GILS are names?*

Get those properties that are *Attribute* of *BSRGILS* and whose range is the class *bsrgils:Name*.

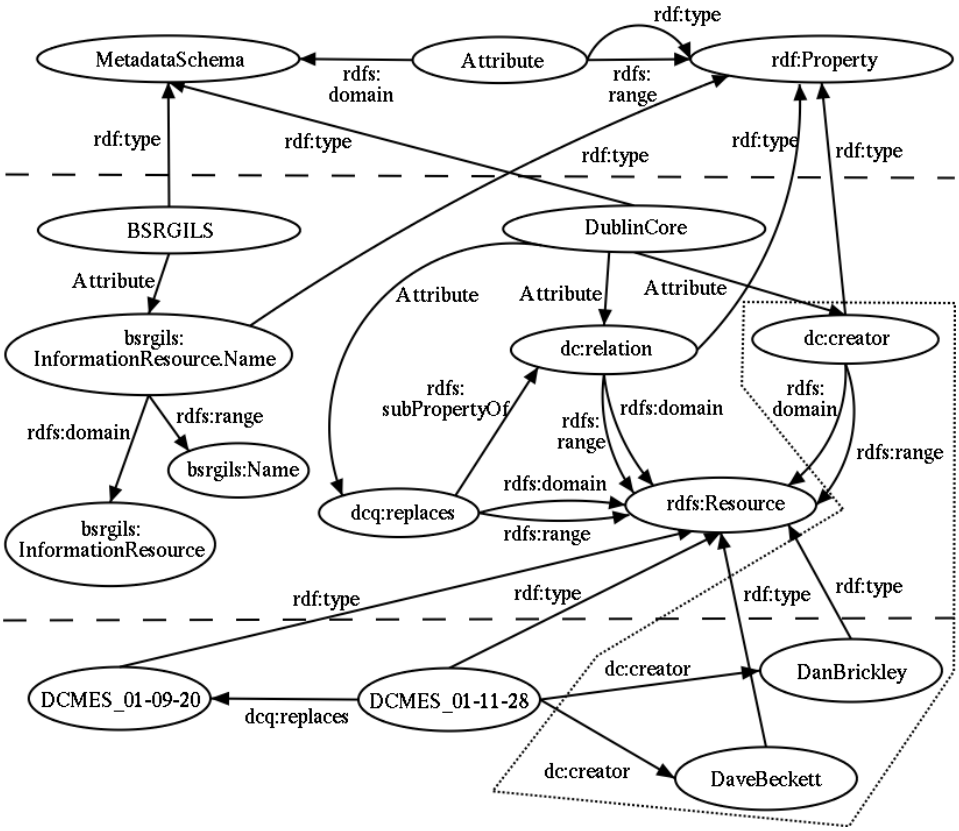

Figure 1: Example – levels of two multi-level hypertexts, represented in RDF

Schema level: *Which attributes are specializations of the attribute dc:relation?*
Get those properties that are a *subPropertyOf* the property *dc:relation*.

Attribute value level: *Browse the values of the Dublin Core attribute dc:creator*
List the resources that are pointed to by instances of the property *dc:creator* (in the example, this would return the resources *DaveBeckett* and *DanBrickley*).

Metadata level: *Which resources were created by Dave Beckett?*
Get those resources that have a property *dc:creator* pointing to the resource *DaveBeckett*.

For a query involving different levels, consider e. g. :

Schema and metadata levels: *In which BSR GILS units of the type "Name" does the value "Salton" appear?*
Get those properties that are *Attribute* of *BSRGILS*, whose range is the class *bsrgils:Name*, and which appear as the predicate in statements where the object is "Salton".

To sum up, the proposed RDF representation provides a uniform model for all levels of multi-level hypertext for any metadata schema, integrating also existing proposals of repre-

senting specific schemas, and allowing searching and browsing in and between the levels. We have implemented this approach using the RDF Schema Specific DataBase (RSSDB) and the RDF Query Language (RQL) ([KCP01]).

5 Conclusion and future work

In this paper, we propose an RDF model for multi-level hypertext in digital libraries. In contrast to other approaches, this model covers all levels of multi-level hypertext in one representation, thus enabling searching and browsing of all levels using an appropriate RDF tool. Existing RDF representations of metadata schemas can be integrated without modification, if they model metadata attributes as RDF properties. Data types can be specified for attributes, both as a restriction on the kind of resource the attribute can apply to, as well as a restriction on the values the attribute can have.

RDF currently does not support imprecision and vagueness, which are essential concepts for performing information retrieval. Thus, appropriate extensions of RDF should be developed.

Literature

[ACG91] M. Agosti, R. Colotti, and G. Gradenigo. A two-level hypertext retrieval model for legal data. In *Proceedings of the Fourteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 316–325, New York, 1991. ACM.

[BBM01] D. Beckett, D. Brickley, and E. Miller. Expressing Simple Dublin Core in RDF/XML, November 2001. `http://dublincore.org/documents/2001/11/28/dcmes-xml/`.

[BG02] D. Brickley and R.V. Guha. RDF Vocabulary Description Language 1.0: RDF Schema, W3C Working Draft, April 2002. `http://www.w3.org/TR/2002/WD-rdf-schema-20020430/`.

[Fuh99] N. Fuhr. Information Retrieval in Digitalen Bibliotheken. In *21. DGI-Online-Tagung – Aufbruch ins Wissensmanagement.*, Frankfurt, 1999. DGI.

[KCP01] G. Karvounarakis, V. Christophides, and D. Plexousakis. RQL: A Declarative Query Language for RDF. *D-Lib Magazine*, December 2001.

[KS01] Stefan Kokkelink and Roland Schwänzl. Expressing Qualified Dublin Core in RDF / XML. Technical report, Universität Osnabrück, Fb. Mathematik / Informatik, November 2001. `http://www.dublincore.org/documents/2001/11/30/dcq-rdf-xml/`.

[LS99] O. Lassila and R. Swick. Resource Description Framework (RDF) Model and Syntax Specification, W3C Recommendation, February 1999. `http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/`.

[VdSL01] Herbert Van de Sompel and Carl Lagoze. The Open Archives Initiative Protocol for Metadata Harvesting, Version 1.1. Technical Report 1.1, Open Archives Initiative, June 2001. `http://www.openarchives.org/OAI/openarchivesprotocol.html`.