Cahiers GUTenberg

INTEGRATION OF RESOURCES ON THE WORLD WIDE WEB USING XML

Roberta FAGGIAN

Cahiers GUTenberg, nº 35-36 (2000), p. 157-167.

<http://cahiers.gutenberg.eu.org/fitem?id=CG_2000___35-36_157_0>

Integration of resources on the World Wide Web using XML

Roberta FAGGIAN

CERN, Genève, Suisse

Abstract. *An initiative to explain High Energy Physics to the general public has been started at CERN. The use of the Web has been identified as crucial to the success of this initiative. An integral part of this project is the construction of a Web-based information system, which collects many different resources on the Web (information published by many European, and US, Particle Physics institutes). This paper proposes a solution to the problem of integration and reuse of heterogeneous information by enriching existing content semantic with metadata in order to improve understanding and discovery. The main part of the work is the study of RDF standard for representing metadata, and its implementation using the XML syntax.*

1. Introduction

The spread of the Internet and the enormous growth of the World Wide Web have allowed everybody to publish and distribute electronic documents throughout the world. At the same time, however, the enormous amount of information on the Internet makes it very difficult to find specific items. This is the main reason why search engines were invented: to search the Web for particular subjects. Search engines, by definition, are very good at searching the Web, but the answers we are looking for are often buried inside an enoumous amount of irrelevant, at least to us, information. The reason for this is that the authors of Web pages did not consider structuring their information and using **metadata** to help the search engines; furthermore, the language commonly used for publishing on the Web, HTML, is, on the one hand, easy to use, but, on the other hand, has several limitations:

— no extensibility, it is not possible to define new **tags** to better classify the information;
— poor structuring of documents, making it insufficient for publishing complex information;
— there are no specifications for the validation of documents;

— a standard formatting is associated to each tag.

The new releases of the HTML language (DHTML, XHTML) try to overcome these limitations and produce documents which can be better understood and processed. New approaches are trying to offer better alternatives and, at the moment, XML seems to be a good solution: it enriches document structure, meaning, comprehensibility, validation, representation and more. But what should we do with the millions of HTML documents already on the Web?

This article proposes how to organise many heterogeneus Web resources in a way that they can be classified and searched using appropriate metadata to describe their contents, the RDF model to structure the metadata and the XML syntax to implement the model.

The case of study described here concerns the field of High Energy Physics (HEP), but it could be used in any other field of knowledge.

2. The problem of information retrieval

The Web contains information about a huge number of subjects intended for a wide variety of purposes, and are structured and formatted in many different ways. In order to allow Web resources to be properly searched and processed metadata needs to be used.

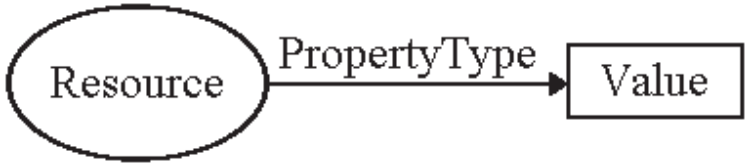
2.1. Metadata

Metadata is information about information, i.e. descriptive/semantic information about documents which is structured in such a way that it can be used for classification and retrieval; allowing documents to be better used and shared over the Net. It is unlikely that everyone will agree to use the same metadata implementation, nevertheless, metadata functionality have a lot in common, even when the metadata are different. RDF (Resource Description Framework) is an effort to identify these common threads and defines a way for web architects to use them to provide useful Web metadata.

2.2. RDF

RDF (Resource Description Framework) defines a standard mechanism for representing and interchanging metadata, which is suitable for describing information about any subject/domain. Each Web resource (a Web page, a graphic,

an audio file, a movie clip, etc.) can be referred to by an address and described using metadata in the form of Statements. A **Statement** is a triplet which associates a specific resource, a named property and the value the property takes for that resource. A Statement can be graphically represented as shown below:



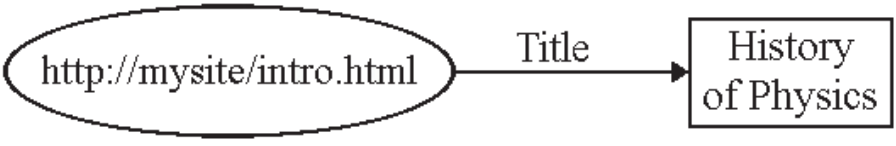
A simple example of a statement is:

```
The title of the page identified by http://mysite/intro.html is
'History of Physics'.
```

The corresponding triplet is

```
(http://mysite/intro.html, Title, 'Hystory of Physics')
```

which can be represented as follows:



RDF properties may be thought of as attributes of resources and also as relationships between resources.

Since XML seems to be a good solution as an exchange format on the Web it has been proposed as the syntax for the implementation of the RDF metadata model. For example, the syntax of the statement in the previous example is:

```
<rdf:RDF>
  <rdf:Description about="http://mysite/intro.html">
    <s:Title>Hystory of Physics</s:Title>
  </rdf:Description>
</rdf:RDF>
```

that becomes, with the basic abbreviated syntax:

```
<rdf:RDF>
  <rdf:Description about="http://mysite/intro.html"
     s:Title="Hystory of Physics">
</rdf:RDF>
```

RDF can be used in a variety of application areas:

- in resource discovery, to provide better search engine capabilities;
- in cataloging, to describe the content and content relationships available at a particular Web site, page, or digital library;
- by intelligent software agents, to facilitate knowledge sharing and exchange;
- in content rating, for content selection;
- in describing collections of pages that represent a single logical "document";
- for describing intellectual property rights of Web pages;
- for expressing the privacy preferences of a user as well as the privacy policies of a Web site.

2.3. RDF Schemas

When defining a meta-representation of information it is fundamental to avoid ambiguous descriptions. The author and the reader must agree on the meaning of the terms used for representing properties and property values. In a global system, like the WWW, one cannot assume that all terms will necessarly mean the same thing for everybody. It is necessary, therefore, to be precise and unambiguous. The RDF Schema specification provides a mechanism that can be used to define special vocabularies for a variety of application domains. This should help people to communicate effectively.

Independent communities can develop vocabularies that suit their specific needs and share vocabularies with other communities. A schema defines the meaning, characteristics, and relationships of a set of properties (the vocabulary's terms), and this may include constraints on potential values and the inheritance of

properties from other schemas. The RDF language allows each document containing metadata to clarify which vocabulary is being used by assigning each vocabulary a Web address. If two applications use the same tag names XML Namespaces become important in resolving conflicts.

One of the best-known schemas is the Dublin Core, invented by the library community. The Dublin Core Metadata Element Set was intended to describe any electronic document or resource. It is made up of the following items: Title, Creator, Subject, Description, Publisher, Contributor, Date, Type, Format, Identifier, Source, Language, Relation, Coverage, Rights.

3. The project

The European Particle Physics Outreach Group was born at CERN in 1997 as an initiative of ECFA (European Commitee for Future Accelerators). It was felt that as a publicly funded activity, HEP needed to better explain to the Europeen tax-payers what exacly they were doing. The group consists of a network of people, representing the CERN member states, who are involved in the popularisation of Particle Physics.

It was identified at a very early stage that the WWW would play an important role in the work of the group and the project to construct a Web-based information system begun in 1998. The project's aims were to collect data and publications related to HEP, merge and organise information already published on other Web sites by European, and US, HEP research institites, universities, museums, laboratories, etc. Today the project includes more than 140 Web sites.

In implementing this system, the main problem we had to face was that information is heterogeneous (in structure and type), is continuously changing and is scattered across the Web.
The solution we are developing is based on the idea that all the information we want to organise can be considered as Web resources which can be indentified by an unique identifier. Each resource can be described with the appropriate set of metadata (**Properties**), which we will define. A database of metadata can be created and queried by users to retrieve the links to the resources in which they are interested.

In our system we identify each Web resource by its location, the URL (Uniform Resource Locator), and we associate to it a description of its contents. In this way we can work (search and reorganise the Web resources) at a level of abstraction independent from the different implementations and we can access

Property	*Value*
Institute	type of the institute publishing the information (resear
Category	category of information (news, events, seminars, people,...)
Kind	type of information (text, image, slide, video,...)
Where	where the institute is situated
When	date releted to the information, if events or conferences...
Contact	the persons responsable for this resource
Expiry	expiry date, if it exists
Related	secondary subjects

Table 1: Properties chosen for describing the resources.

information objects distributed across multiple locations and systems in the same simple way.

4. Arriving at a solution

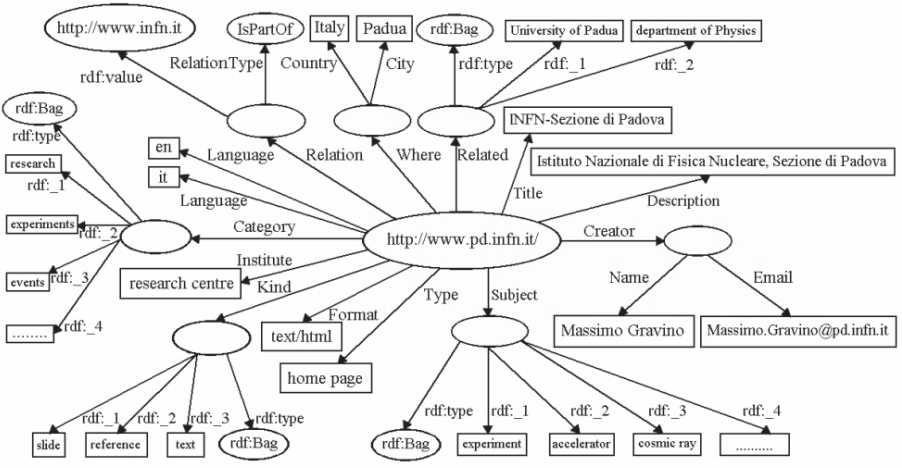
The first step is the choice of Properties which best describe our resources. We have decided to adopt the following:

— the Dublin Core Elements Set,
— other properties corresponding to the needs of the domain of information we are going to treat; these are listed and briefly described in table 1.

This vocabulary of terms will be used for describing the resources which will take part in our information system. We have to define their meaning and the possible values they can assume in an RDF Schema.

Our model is ready: we have defined a special vocabulary for our application domain and a precise meaning for its terms. This means we will have a common-undestanding of terms and subjects for everybody sharing these information.

For instance, we can consider the RDF model of the metadata related to the resource referred to by the URL `http://www.pd.infn.it/`, it can be graphically represented in this way:



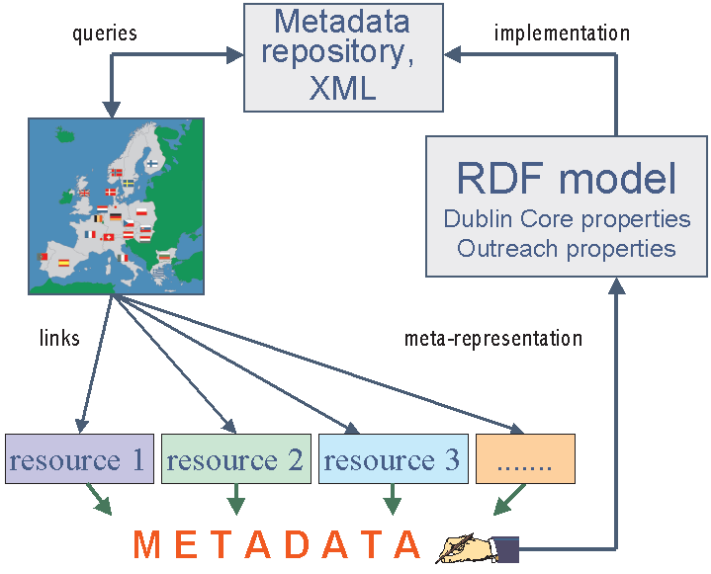
XML can be used for the implementation:

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:dc="http://purl.org/metadata/dublin_core#"
xmlns:dcq="http://purl.org/metadata/dublin_core_qualifiers#"
xmlns:otrc="http://outreach.cern.ch/rdf/elements.htm#"
xmlns:ent="http://outreach.cern.ch/rdf/entity.htm#">
xmlns:pla="http://outreach.cern.ch/rdf/place.htm#">
 <rdf:Description about="http://www.infn.pd.it/"
   <dc:Title>INFN - Sezione di Padova</dc:Title>
   <dc:Creator>
       <ent:Name>Massimo Gravino</p:name>
       <ent:Email>Massimo.Gravino@pd.infn.it</p:Email>
   </dc:Creator>
   <dc:Subject>
       <rdf:Bag>
    <rdf:li>experiment</rdf:li>
    <rdf:li>accelerator</rdf:li>
    <rdf:li>cosmic ray</rdf:li>
    <rdf:li>neutrino beam</rdf:li>
    <rdf:li>nuclear physics</rdf:li>
    <rdf:li>theoretical physics</rdf:li>
    <rdf:li>high-energy physics</rdf:li>
```

```
    .......
       </rdf:Bag>
  </dc:Subject>
  <dc:Description>Istituto Nazionale di Fisica Nucleare,
      Sezione di Padova</dc:Description>
  <dc:Type>home page</dc:Type>
  <dc:Format>text/html</dc:Format>
  <dc:Language>en</dc:Language>
  <dc:Language>it</dc:Language>
  <dc:Relation rdf:parseType="Resource">
    <dcq:RelationType rdf:resource=
    "http://purl.org/metadata/dublin_core_qualifiers#IsPartOf"/>
    <rdf:value resource="http://www.infn.it"/>
  </dc:Relation>
  <otrc:Institute>research centre</otrc:Institute>
  <otrc:Category>
<rdf:Bag>
     <rdf:li>research</rdf:li>
     <rdf:li>experiment</rdf:li>
     <rdf:li>event</rdf:li>
     <rdf:li>people</rdf:li>
     .......
</rdf:Bag>
  </otrc:Category>
  <otrc:Kind>
<rdf:Bag>
     <rdf:li>text</rdf:li>
     <rdf:li>reference</rdf:li>
     <rdf:li>slide</rdf:li>
</rdf:Bag>
  </otrc:Kind>
  <otrc:Where>
<pla:Country>Italy</pla:Country>
<pla:City>Padua</pla:City>
  </otrc:Where>
  <dc:Related>
      <rdf:Bag>
   <rdf:li>University of Padua</rdf:li>
   <rdf:li>department of Physics</rdf:li>
      </rdf:Bag>
  </dc:Related>
</rdf:Description>
```

Data describing the resources can be stored in a central repository. This repository can be organised as a database holding information about properties and property values for each resource. This database stores a meta-representation of the information we want to organise, and it can be used to execute a metadata-level query on the resources it stores. In this way we can obtain high-precision results and the integration of heterogeneus information. XML, together with the related standards, will be used as the syntax for the representation of the results.

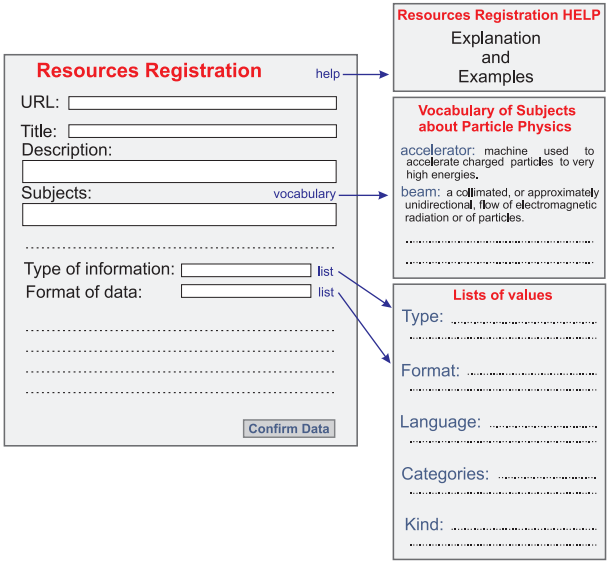
The following schema gives an idea of the system operating cycle:



5. The Web interface

How can we define the metadata describing all the resources of our system? Fortunately we can count on the collaboration of the persons involved in the Outreach project. People in charge of filling up the metadata repository don't need to have a technical profile: a simple Web interface (password protected) will allow them to update the model in a very easy way. This interface is

represented here:



Below, we see an example of how the meta-information can be used to help the user finding resources. An alphabetical list of the subjects can be consulted, and for each subject the corresponding resources will be listed.

6. Conclusions

The key points of our solution are:

— the definition of a personalised, descriptive RDF model of the resources;
— the flexibility of the XML implementation.

Using our solution, the problem of integration of resources becomes a problem of definition of their meta-representations. These meta-representations can be used to query the whole information system at a metadata-level of abstraction instead of working on the actual documents.

The advantages of our solution are:

— we have defined a common basis for understanding for everyone sharing these resources and using our vocabulary;
— metadata is updated throughout the collaboration, i.e. the work of adding to the knowledge base is distributed;
— metadata stored in the repository has been entered manually and so will have significant added value over computer-generated input;
— metadata can be processed independently from the resources to which they refer. In this way the resource is not bound to only one description;
— a single metadata-entry in the repository can represent an entry point into an arbitrarily complex resource. The entry point to a complete Web site would be its Home Page;
— different kinds of resource can be included, each identified by a URL;
— the responsability of the repository maintenance is centralised;
— the RDF model can be extended, if necessary, simply by adding new terms and properties to the vocabulary;
— the solution is independent of any proprietary technology;
— the solution can be applied to any other domain by redefining the properties particular to this model;
— the use of XML (metadata can be shared, reformatted, etc.) which is becoming the de-facto standard for sharing information on the Web.