WP2 Metadata in ECHO
WP2-TR08-2003 Version 1

Metadata in ECHO

Peter Wittenburg, Daan Broeder
March 10, 2003

Definition

The term “metadata” in general stands for “data about data”, i.e. it refers to data that is expressed as a relationship to another type of data.

According to this general definition metadata can be

- annotations of multimedia signals
- linguistic type of annotations (morphosyntax) associated with texts
- polygon coordinates identifying an object in a 2D image
- keyword type of descriptions of resources as a whole
- many other things

Metadata as a Digital Library Term

Traditionally the term metadata is used in the world of libraries for a set of descriptors describing documents such as books and dissertations that are stored in them. Each such document is associated with a card containing such descriptions and it is used for discovery and management purposes. In the world of digital libraries, the term “metadata” had its revival. In the era of the Internet it will become even more difficult to discover a resource that may be useful for the work in mind. In ever growing repositories of digital resources that are highly related also management is an increasingly difficult task.

Therefore, the term “metadata” now usually refers to machine readable structured data of the keyword/value type describing Internet resources as a whole. These can be used to discover and manage these resources that can be distributed all over the Internet. The set of descriptors and their structural arrangement are specified by DTDs or XML schema and their semantics should be defined carefully according to ISO standards. Of course, this type of metadata has to be openly accessible.

Dublin Core Metadata Set

The most well-known metadata proposal of this sort is the Dublin Core Metadata Set. It was designed to allow describing all types of authored resources with an emphasis on the description library information issues. The DC initiative finally came up with 15 descriptors that are necessarily defined vaguely to accommodate with the various types of resources. The Dublin Core set was a result of a process of necessary “pidginization” as Lagoze describes it.

It was obvious very early during the process of developing the DC set that communities and sub-communities will come up with their own sets that are more suitable to their discipline. This process of modularization can be called “creolization”. Indeed many metadata sets - some of them being reformulations of descriptions systems that were used for local discovery and management purposes for a while already - occurred, since the underlying goals became increasingly important.

Necessity of Metadata

Metadata in the restricted sense as keyword type description of whole resources are increasingly important in a world of extremely increasing number of resources. Content-based search will not help professional users to quickly find the resources they want to work on. There are several reasons for this such as

- Many resources such as images, sound or video files need explicit descriptions since their content is not accessible from the stored pattern.
- Many textual resources have contents that don't describe attributes that are relevant to investigators. A resource with some text does not say who created the text. A resource containing an interview does not say who the interviewee was and what kind of educational background this person has.

In the ECHO project we have sub-projects such as “Archimedes”. Here it is often argued that the most important part for researchers is to draw relations between textual entries and in doing so to allow new insights into existing material. The point is that within such a project the researchers select the documents that belong to the scope of such a project. Given this background it is understandable that History of Science is not interested in descriptive metadata. Here comes the role of metadata. It allows the user to discover relevant material with a high precision and recall. In a scenario where is has to be possible for individual researchers to quickly define the scope of resources to be included in a more complex operation, metadata is omissible.

Another aspect making keyword type of metadata very important has to do with reasoning and multilingualism. It seems to be relatively easy to convert metadata descriptions into different languages making metadata a primary candidate to offer multilingual information about resources. Further, metadata are an excellent platform to allow interdisciplinary operations and after all definitions and relations have been made explicit (see below) metadata are a primary candidate for applying reasoning, i.e. to let software agents operate on them and combine them with other type of knowledge.

Exchange and Interoperability

The scenario of distributed repositories with different holdings (archives) became obvious very early. Therefore, the Open Archives Initiative (OAI) worked out a simple lightweight protocol that allows harvesting metadata from different registered repositories to come to a central index that can be used for efficient searching. The protocol allows harvesting all types of structured metadata records, however, OAI also requires supporting a mapping to DC.

So the DC set can be seen as a way to achieve interoperability. If every community provides a mapping to DC we can use DC as a platform of issuing queries including all registered holdings. However, as for example the MPEG7 and IMDI initiatives have shown much of the useful information for the more specialist users.

The Semantic Web (SW) will shift interoperability towards new shores. The bases of the SW scenario are open and machine-readable definitions of the data categories used and their structural and semantic relations within and across metadata sets. This open scenario will allow users to use existing mapping schemes between two metadata sets or to redefine the mappings according to their particular needs. These mappings will be typically stored in practical ontologies as RDF assertions. This open scenario of schema definitions, data category repositories based on XML and relations in ontologies create completely new perspectives to tackle the problem of interoperability in a modular, service based fashion.