BOOK OF ABSTRACTS

DATA AND INFORMATION FOR THE COMING KNOWLEDGE MILLENNIUM:
Science and Technology in the Quest for a Better World

15-19 October 2000
Baveno, Italy

CODATA Secretariat, 51 Bld de Montmorency, 75016, Paris.
Ph: + 33 1 45 25 04 96, Fax: + 33 1 42 88 14 66, E-mail:codata@ diat.oleane.com, http://www.codata.org

17th International CODATA Conference
Session: Theme I-6 " Innovative Web Design and Applications"

XML in the Documentation Field: Designing Hints for 'Semantic Web' Applications

Paola Carrara, Giuseppe Fresta

ITIM-CNR - via Ampère 56 - 20131 Milano - Italy
Email addresses of co-authors: paola.carrara@itim.mi.cnr.it, giuseppe.fresta@cnuce.cnr.it

The rapid growth of the web is more and more dependent on technologies developed as from common rules and standards: this is much more true in the documentation field where one of the main problems is as much content structuring as their form and appearance. There is a need of "semantic" structures, corresponding to powerful search functionalities, able to give meaningful answers, and to operate in limited contexts (items and/or structures) following the users' request.

XML applications seem to be suitable to these problems and moreover, are effective as far as document rendering, as they guarantee the reuse of information (freed from formatting aspects) in various environments, assuring its coherence, completeness, and reliability.

The paper suggests methodologies derived from Database and Information Retrieval (IR) design experience to design XML application in the documentation field: in particular, the ER (Entity-Relationship) approach is adopted to describe the 'universe' of the applications, while methods derived from the IR practice are combined to significant items aimed to create a "semantic web" application.

The introduction of the ER description has a double role: from the designer point of view it is extremely useful in defining which are the entities of the application, their attributes and their relationships. This avoid many usual drawbacks of naive design such as redundant (sometimes inconsistent) data, bad structure definition in term of

granularity, etc. From the user point of view it can be adopted as a schema for suggesting useful hints on the content structures, thus guiding the formulation of non-generic queries.

The paper will describe both aspects, and furthermore discusses another important problem which can be addressed by the use of a schema, that is how XML documents can be stored in a database (native or not): in the paper three possible scenarios are discussed. In the first whole documents are stored in the database; in the second documents are fully fragmented following their tag stucturation; in the third, a hybrid one, documents are decomposed in three sections corresponding respectively to structured data (i.e. records), unstructured data (i.e. text), and context data aimed to rebuild hierarchical information lost in document decomposition.

In IR Systems, documents are represented by the so called surrogates, which are meaningful descriptions of the document contents (and sometimes of their structures), suitable to be matched with users' queries. The paper suggests to derive from the ER schema richer definitions of tags, i.e. tags with name, type, timestamp, etc., to give them the role of metainformation able to represent in a less ambiguous way "knowledge" and "meaning" of the document content.

For example, in the following choice:
... the basilica is mentioned for the first time already in the <item type="data" role="citation"> year 375</item> and its importance ... The existing building almost entirely dates back to the <item type="data" role="foundation">Xth century</item> ...

tags represent concepts and their meaning more thoroughly and explicitly with respect to the choice:
... the basilica is mentioned for the first time already in the <dating> year 375</dating> and its importance ... The existing building almost entirely dates back to the <construction>Xth century</construction> ...

thus allowing a real usability in terms of searching and filtering mechanisms (thesauri, dictionaries, etc.). In fact, tags of the first form, with their attributes, allow to best fit web answers to users' requests and to filter needs depending on both objective criteria (i.e. levels of privacy, security, different types of platform, purpose and target of the application - education, tourisms, on-line help, e-commerce, etc.-), and subjective requirements (scientific background or interests, handicaps, etc.)