Workshop on
Multimedia and Internet Technologies
26th -28th February, 2001
DRTC, Bangalore

Paper: AI

Bibliographic Data Elements for Internet Resources using XML: A Proposal

Prasenjit Kar
Tata Energy Research Institute, Bangalore 560 052

Abstract

The paper focuses on the possibilities of better organization and retrieval of Internet resources with the advent of eXtensible Markup Language (XML). The paper advocates the need of identifying necessary and sufficient bibliographic data fields for Internet resources, and generating numerical XML elements for the data fields using a standard syntax which brings together existing tag codes. The paper also brings into the issue of standard out put of search results.

1 Introduction

Internet is a distributed information system, consisting thousands of disjoint sources and millions of documents. As Internet is an ungoverned information system the major activity on the Net, that is electronic publishing, is being done in quiet an unorganized way. The lack of knowledge in organizing and retrieving of huge quantities of information can be visualized in every affair connected to Internet.

In any information system, we have to have certain tools and techniques to tell us which document contains what information. These tools and techniques can be classification, cataloguing, indexing etc. in the case of traditional information systems. As Internet is a distributed information system with disjoint information sources spread across the globe, we cannot bring the related documents together in the retrieval process. Hence the tools and techniques meant for pinpointing information on the Net need to be more accurate.

2 The present day's search engines

Search engines claim to be the information pinpointer on the Net. But the approaches of the existing search engines are ill planned and inefficient to say the least. This is obvious from the two following important observations:

1. The number of results of many a search is so high in recall (often in thousands) and low in precision that it leaves us wondering whether we made any progress with search at all.

2. The output of the search result has no standard of any kind.

The search engines on the Net use robots to collect information about the documents to generate keyword indexes. Generally different algorithms are used to assign keywords to the document. Some of the robots collect keywords from the title whereas some collect from the first few lines of the document (say 50 k). But the very pre-assumption of the search robots, that the titles are expressive and the beginning few lines of a document focuses the content, do not hold good for maximum documents.

To give an example, if we search the Net for ‘bond’, we should be sure to get results at least regarding ‘chemical bonds’, ‘James Bond’, ‘financial bond’, etc. Certainly the recall will be very high. The retrieval of irrelevant result can be reduced by subject gateway approach, but cannot be eliminated.

The root cause of the problem lies in HTML (*Hyper Text Markup Language*), the defacto standard for web publication. The major problem with HTML is its ‘*fixed tagset*’. This *tagset* is mainly for display of the content and HTML provides no tag to address the content precisely. XML (*eXtensible Markup Language*) designed by W3C (World Wide Web Consortium) promises a possible solution to this problem. The major advantage of XML over HTML is its extensibility ie, provision of user defined tags and attributes to identify the structural elements of a document. XML also provides structural complexity to define document structure that can be nested at any level of complexity.

3 What is XML?

Extensible Markup Language is a text-based format that lets developers describe, deliver and exchange structured data between a range of applications to clients for local display and manipulation. XML also facilitates the transfer of structured data between servers. Vast stores of legacy information exist today, distributed across disparate, incompatible databases. XML allows the identification, exchange and processing of this data in a manner that is mutually understood, using custom formats for particular applications if needed.

XML resembles and complements HTML. XML describes data, such as city name, temperature and barometric pressure, and HTML defines tags that describe how the data should be displayed, such as with a bulleted list or a table. XML, however, allows developers to define an unlimited set of tags, bringing great flexibility to authors, who can decide which data to use and determines its appropriate standard or custom tags.

This data could be displayed in many different ways, or it could be handed off to other applications for further processing. Style sheets could help by transforming structured data into different HTML views for display in a browser, or even to other display formats on other platforms running other applications.

Today with XML, *Document Type Definitions* (DTDs) may accompany a document, essentially defining the rules of the document, such as which elements are present and the structural relationship between the elements. DTDs help to validate the data when the receiving application does not have a built-in description of the incoming data. With XML, however, DTDs are optional.

Data sent along with a DTD is known as '*valid*' XML. In this case, an XML parser could check incoming data against the rules defined in the DTD to make sure data was structured correctly. Data sent without a DTD is known as '*well-formed*' XML.

With both valid and well-formed XML, XML encoded data is self-describing. The open and flexible format used by XML allows it to be employed anywhere a need exists for the exchange and transfer of information. This makes it powerful.

For instance, XML can be used to describe information about HTML pages, or it can be used to describe data contained in business rules or objects in an electronic-commerce transaction, such as invoices, purchase orders and order forms. XML is separate from HTML, but XML could also be added inside HTML documents. By embedding XML data inside an HTML page, multiple views could be generated from the delivered data, using the semantic information contained in the XML. Moreover, XML can be used for such compelling applications as distributed printing, database searches, and others.

4 XML and more precise search

Imagine a search engine that understands and uses contextual information when performing a full-text search. Searching for information about the 'Java' programming language will no longer yield to coffee sites or the 'Island of Java'. This is because searching for the term 'Java' is narrowed down to those fields tagged as a *'programming language*'. As a result, the speed and accuracy of the search is dramatically improved. Widespread use of XML repository technology on Web servers will play a vital role in easing the "information overload" currently suffered by Internet users.

In bibliographic data elements, once we have the standard tagset for Internet resources, the search engines can easily build the keyword index from the terms appearing in the tag denoting subject descriptors of the document. This approach will certainly be more effective and efficient than depending on the expressiveness of the title and searching for the needle in the hay stack of first few lines.

Data can be uniquely tagged with XML, allowing a customer to specify books *by*, rather than *about*, S R Ranganathan, for example. In contrast, searches using present methods would probably yield both types of books mixed together.

Anyhow, there are things, which must happen before XML can solve the search problem.

- Wider browser acceptability
- Each industry will have to set up its standard structure
- Web site content providers will have to tag the pages according to the standard structures
- Search engine indexing applications will have to hold the tag information as Meta data
- Search engines will have to learn each standard structure in the collection of indexed documents

5 A proposal

The direct impact of XML will be in locating information on the Internet, pin-pointedly, exhaustively and expediously. XML will also enhance intra-disciplinary and inter-disciplinary information exchange. On the other hand, XML will surely enhance resource sharing and data reusability of bibliographic data. But one of the prerequisite for the above said advantages is common understanding among the concerned communities.

More specifically we require a standard mark up tagset for describing bibliographic data elements for resources on the Net. Our experiences with data exchange across the globe prompt us to have numeric tagset for bibliographic data element so that the tagset can cut across the language barrier. One unfortunate thing with traditional/existing information system is that it has a plethora of standards to describe bibliographic data elements defeating the purpose of standardization.

Proposing one more standard for the Internet documents will add up to the confusion. So a standard syntax has to be worked out to bring together all the existing standards.

And again, creating standard tagset specific to each industry will be an impossible task. So instead of creating industry specific standard tagsets the much more feasible solution might lie in identifying the necessary and sufficient data fields to describe an Internet resource. And XML elements can be generated using a standard syntax. *Anglo-American Cataloguing Rules, 2d ed., 1988 revision* (AACR2), has put down the rules for cataloguing computer files. The same rules can be reviewed and followed, after necessary changes, for Internet resources. Necessary tags from each tag-code can be identified for this purpose and the XML element can be generated following a standard syntax. The repeatability of each of the XML element for bibliographic data elements is governed by the rules of the code under use.

6 Towards standard output format

Apart from the fact that the number of results of many a search is so high in recall and low in precision, another characteristic of the search engines is the conspicuous absence of authors' names in the search result. The search results normally present the title and a part of first few sentences. This kind of presentation does not generally convey about the relevance of the result and definitely this is not the way a document surrogate should look like. It is true that the web document cannot be described as a general document. The web may not have publisher name, place of publication but similarities can be brought in easily, like the organization publishing the document, date of the document, URL etc.

An adhoc proposal for bibliographic descriptive elements of a document can be as follow:

a. Title of the document
b. Author(s)
c. Publishing organization
d. Site address of the host server (URL)
e. Date of the document
f. Subject descriptors
g. Abstract

The provision of this information will be much more congenial for search engines and the users of the web. As this information gives much better context to the end user in selecting the document for further information.

7 Conclusion

The basic focus of the paper is to advocate the possibilities of creating standard tagset for defining bibliographic data of Internet resources using XML. This will inturn lead to more effective and efficient functioning of Internet. For the same purpose a standard syntax, which can bring together the existing tag codes, has to be worked out. The search engines have to accept one such standard syntax and train themselves about the semantics of each element. This will help in getting far more precise search results. If the new syntax does not disturb the existing standards and their rules, two-way conversion i.e. database to the Web and Web to database, can be done easily.

8 References

1. BRADELY (Cara). The XML files: the truth will be out there. *http://www.slis.ualberta.ca/538-99/cbradley/xml.htm*
2. CLARK (James) and DEACH (Stephen) [Ed.]. Extensible style language(XSL):version 1.0(working draft). World Wide Web Consortium, 1998. *http://www.w3.org/TR/WD-xsl*
3. CLARK (James) and DeROSE (Steve) [Ed.]. XML path language (XPath): version 1.0(W3C recommendation). World Wide Web Consortium, 1999. *http://www.w3.org/TR/xpath.htm*
4. HAROLD (Elliotte Rusty). XML bible. IDG : New Delhi, 2000.
5. HERWIJINEN (Eric van). The impact of XML on library procedures and services. CERN : Geneva, 2000.
6. LANDER (Rich'd). XML: the new markup wave. 1997. *http://www.csclub.uwaterloo.ca/u/relander/XML/Wave/summary.html*
7. MALER (Eve) and DeROSE (Steve) [Ed.]. XML linking language (XLink)(W3C working draft). World Wide Web Consortium, 1998. *http://www.w3.org/TR/1998/WD-xlink.htm*
8. NAVARRO (Ann), WHITE (Chuck) and BURMAN (Linda). Mastering XML. BPB : New Delhi, 1999.
9. NORTH (Simon) and HERMANS (Paul). SAMS teach yourself in 21 days. Techmedia : New Delhi, 1999.