XML as and for metadata

Airi Salminen
University of Jyväskylä

http://www.cs.jyu.fi/~airi/

Outline

1. What is XML?
2. Why XML evolved
3. XML as metadata
4. XML for metadata
5. Summary

XML = Extensible Markup Language

A set of rules for defining and representing information as structured documents for applications on the Internet; a restricted form of SGML (Standard Generalized Markup Language)

T. Bray, J. Paoli, C. M. Sperberg-McQueen, and E. Maler (Eds.), Extensible Markup Language (XML) 1.0 (Second Edition), W3C Recommendation 6 October 2000, http://www.w3.org/TR/2000/REC-xml-20001006

1. What is XML?

- Rule 1: Information is represented in units called *XML documents.*
- Rule 2: An XML document contains one or more *elements.*
- Rule 3: An element has a name, it is denoted in the document by explicit markup, it can contain other elements, and it can be associated with *attributes.*

and lots of other rules ...

Example of an XML document

```
<?xml version = "1.0"?>
<poem author = ”Murasaki Shikibu” author_born = ”974”>
<info_link  xmlns:xlink=”http://www.w3.org/1999/xlink”
   xlink:type="simple”
   xlink:href=
   ”http://digital.library.upenn.edu/women/omori/court/murasaki.html”>
      About the author
</info_link>
<stanza>
<line>This life of ours would not cause you sorrow</line>
<line>if you thought of it as like </line>
<line>the mountain cherry blossoms</line>
<line>which bloom and fade in a day. </line>
</stanza>
</poem>
```

Note: The text of the line elements is taken from http://www.slip.net/~knabb/rexroth/translations/japanese.htm, containing Kenneth Rexroth’s translations of Japanese poetry

XML is a metalanguage, not a specific language

- Defines the rules how to mark up a document — does not define the names used in markup.
- Includes capability to prescribe a document type by a collection of declarations to constrain the markup permitted in a class of documents.
- Intended for *all* natural languages, regardless of character set, orientation of script, etc.

Document type declaration for a poem

```
<!DOCTYPE poem [
<!ELEMENT poem   (info_link? title?, stanza+)>
<!ATTLIST poem
      author CDATA  #REQUIRED
      author_born  CDATA    #IMPLIED>
<!ELEMENT title          (#PCDATA) >
<!ELEMENT info_link    (#PCDATA) >
<!ATTLIST info_link
     xmlns:xlink  CDATA  #FIXED "http://www.w3.org/1999/xlink"
     xlink:type  CDATA    #FIXED "simple"
     xlink:href  CDATA    #REQUIRED >
<!ELEMENT stanza       (line+)   >
<!ELEMENT line          (#PCDATA) >]
```

2. Why XML evolved

1960-1980 Infrastructure for the Internet

1986 SGML for defining and representing structured documents

1991 WWW and HTML introduced for the Internet

1995 Business adopts the WWW technology; huge expansion in the use of the Internet; new kinds of businesses evolve, based on the connectivity of applications built by various software providers (B2C, B2B)

Urgent need for a new, common data format for the Internet

2. Why XML evolved

- Needs:
  - Simple, common rules that are easy to understand by people with different backgrounds (like HTML)
  - Capability to describe Internet resources and their relationships (like HTML)
  - Capability to define information structures for different kinds of business sectors (*unlike* HTML, like SGML)

2. Why XML evolved

- Needs (cont'd):
  - Format formal enough for computers and clear enough to be human-legible (like SGML)
  - Rules simple enough to allow easy building of software (*unlike* SGML)
  - Strong support for diverse natural languages (*unlike* SGML)

metadata = data about data

- The markup used in a document serves as metadata in relationship to the character data
- The declarations associated with a class of documents serve as metadata in relationship to the documents.

3. XML as metadata

This life of ours would not cause you sorrow if you thought of it as like the mountain cherry blossoms which bloom and fade in a day.

About the author

This life of ours would not cause you sorrow
if you thought of it as like
the mountain cherry blossoms
which bloom and fade in a day.

About the author

Metadata expressed in the markup (slide 5):

- The document is called a poem and it consists of elements called info_link and stanza, and the stanza consists of elements called line.
- The author of the poem is Murasaki Shikibu, born in 974.
- The element info_link with the text content "About the author" is a simple link referring to the Web resource at **http://digital.library.upenn.edu/women/omori/court/murasaki.html**
- **...**

Metadata Expressen in the DTD (slide 7) and associated with a document collection:

- The documents are poems.
- A poem may contain a title and it always contains one or more stanzas.
- A poem may be linked to a resource by a simple link.
- For each poem there is information about the author and possibly about the year of birth of the author.
- **...**

The metadata can be used, for example, to access information:

- Find poems authored by "Murasaki Shikibu"
- Find poems whose author was born at least 1000 years ago
- Find poems with two lines

4. XML for metadata

There is a wide variety of applications where XML has been used especially for bibliographic metadata, for example,

- BiblioML - XML for UNIMARC Bibliographic Records
- bibteXML - XML for BibTeX
- OAI - Open Archives Initiative
- PRISM - Publishing Requirements for Industry Standard Metadata

BiblioML

- XML-based format for the interchange of UNIMARC bibliographic records between applications
- Sponsored by the Ministère de la culture et de la communication, France
- DTD under development, latest version 0.3 from May 2000, defines 224 elements, the top level element is BiblioRecord

BibteXML

- Expresses an XML markup similar to the BibTeX language earlier specified for LaTeX
- For researchers to maintain a bibliography in XML format

BibteXML example

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE bibtex:file PUBLIC "bibteXML”...  >
<bibtex:file xmlns:bibtex="http://www.science.uva.nl/~zegerh/bibteXML/">
<bibtex:entry bibtex:id=”Salminen1999a">
  <bibtex:article>
    <bibtex:author>A. Salminen, F.W. Tompa</bibtex:author>
    <bibtex:title>Grammars++ for modelling information in text</bibtex:title>
    <bibtex:journal>Information Systems</bibtex:journal>
    <bibtex:year>1999</bibtex:year>
    <bibtex:volume>24</bibtex:volume>
    <bibtex:number>1</bibtex:number>
    <bibtex:pages>1-24</bibtex:pages>
  </bibtex:article>
</bibtex:entry>
...
```

OAI

- The Open Archives Initiative has its roots in an effort to enhance access to e-print archives as a means of increasing the availability of scholarly communication.
- The interoperability framework that is defined in the Open Archives Metadata Harvesting Protocol.
- The Open Archives Metadata Harvesting Protocol defines a mechanism for harvesting records containing metadata from repositories.
- The metadata is expressed in the Dublin Core format, using XML.

PRISM

- PRISM = Publishing Requirements for Industry Standard Metadata
- Developing a standard XML metadata vocabulary for publishing industry
- For syndicating, aggregating, post-processing and multi-purposing content from magazines, news, catalogs, books and mainstream journals.

5. Summary

- XML is a metalanguage defining rules to mark up documents and to define specific markup languages for specific purposes.
- XML was developed to the needs of data interchange and distribution on the Internet.
- The markup always carries metadata that can be used, for example, for information retrieval purposes.
- Several XML-based languages for bibliographic metadata are under development.

Tack!