



XML

Stephen Miller

AAC / ÖAW

Texts & Files WS 2004



XML Base Specification



XML: Base Specifications

- Extensible Markup Language 1.0 (XML)
 - W3C Recommendation 10 February 1998
 - www.w3.org/TR/2000/REC-xml-20001006
- Namespaces in XML (XML Names)
 - W3C Recommendation 14 January 1999
 - www.w3.org/TR/1999/REC-xml-names-19990114



Associated Standards

- Unicode and ISO/IEC 10646
 - glyphs
- Internet RFC 1766
 - language identification tags
- ISO 639
 - language name codes
- ISO 3166
 - country name codes



Design Goals of XML



XML

- Extensible **Markup Language**
- Based on SGML
 - an *application profile* of SGML
 - (3) XML shall be compatible with SGML. Most of the people involved in the XML effort come from organizations that have a large, in some cases staggering, amount of material in SGML. XML was designed pragmatically, to be compatible with existing standards while solving the relatively new problem of sending richly structured documents over the web.



XML and the Internet

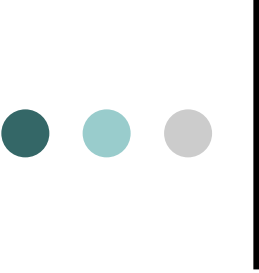


Goals of XML (1)

(1) It shall be straightforward to use XML over the Internet. Users must be able to view XML documents as quickly and easily as HTML documents. In practice, this will only be possible when XML browsers are as robust and widely available as HTML browsers, but the principle remains.

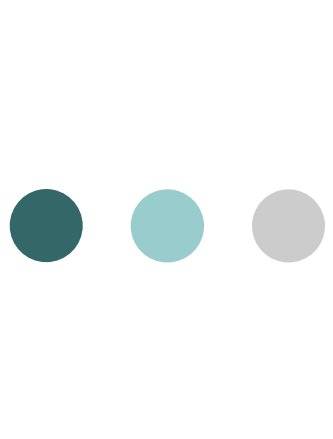


The Prime Aim of XML



Goals of XML (2)

(2) XML shall support a wide variety of applications. XML should be beneficial to a wide variety of diverse applications: authoring, browsing, content analysis, etc. Although the initial focus is on serving structured documents over the web, it is not meant to narrowly define XML.



The Goals of XML



Goals of XML (1 to 5)

- (1) XML shall be straightforwardly usable over the Internet.
- (2) XML shall support a wide variety of applications.
- (3) XML shall be compatible with SGML.
- (4) It shall be easy to write programs which process XML documents.
- (5) The number of optional features in XML is to be kept to the absolute minimum, ideally zero.



Goals of XML (6 to 10)

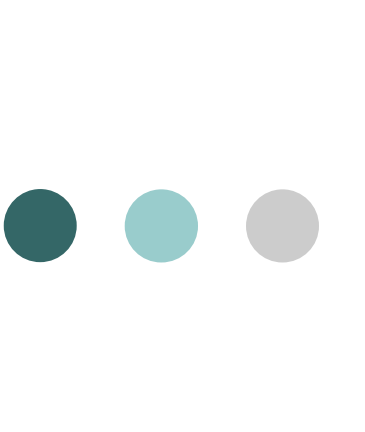
- (6) XML documents should be human-legible and reasonably clear.
- (7) The XML design should be prepared quickly.
- (8) The design of XML shall be formal and concise.
- (9) XML documents shall be easy to create.
- (10) Terseness in XML markup is of minimal importance.



XML in 10 Points...

- *XML in 10 points*

- www.w3.org/XML/1999/XML-in-10-points



XML as “Well-Formed”



XML & Conformance

1. “Well-Formed”
 - standalone file
 - DTDless
2. Valid
 - conform to a DTD
 - all valid XML instances are “well-formed”



XML as “Self-Describing Data”



“Self-Describing Data”

○ Elements

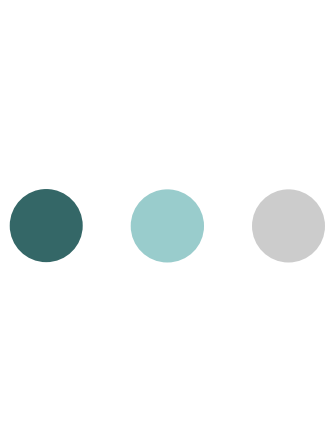
- names are user-defined
- names are case-sensitive
- elements must nest
- empty elements must be closed

○ Attributes

- values must be quoted

○ Character Entities

- declared using Unicode



XML as Application

Stephen Miller
AAC / ÖAW



Index Card

"Ec yn Fiddleryn"

Coll. 1896 by John Strachan & Father
Richard Henebry from Thomas
Kermode, Bradda, Rushen.

Text (Manx), no tune, pub. ZCP i
(1897): 54-58.



XML format

```
<?xml version="1.0" standalone="yes" ?>
<catalog>
  <entry>
    <canonical_title>Ec yn Fidderlyn</canonical_title>
    <collected_title>Ec yn Fidderlyn</collected_title>
    <text lang="glv">Yes</text><tune>No</tune>
    <singer gender="male">Thomas Kermode</singer>
    <collector>Strachan, John </collector>
    <collector>Henebry, Father Richard</collector>
    <collected date>1896</date>
    <source date>1897</date>
    <place>Man : Rushen : Rushen : Bradda</place>
    <source type="printed"/>
    <reference>ZCP i (1897): 54-58</reference>
  </entry>
</catalog>
```



XML as Data



Data / Document

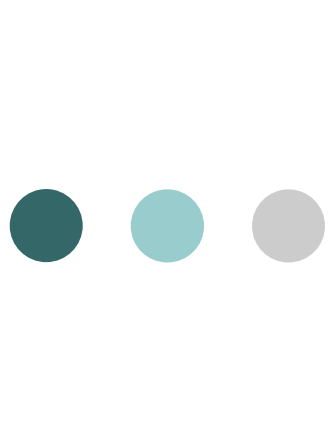
- Data
 - application
- Documents
 - users



Elements v Attributes

- Datacentric v Docucentric issue

- <http://www.oasis-open.org/cover/elementsAndAttrs.html>
- <http://www.xml everywhere.com/newsletters/20000525.htm>



XML & HTML



Whither HTML?

- HTML 2.0
- HTML 3.0
- HTML 4.0
- HTML 4.01



Cometh XHTML...

- XHTML 1.0 is a reformulation of HTML 4.01 in XML
 - 3 flavours
 - XHTML 1.0 Strict
 - XHTML 1.0 Transitional
 - XHTML 1.0 Frameset
 - <http://www.w3.org/MarkUp>



XML & TEI



TEI P4

- TEI Guidelines now expressed in XML
 - <http://www.tei-c.org/P4X/>
- Guidelines remain in development, P5 is expected end of 2004
 - <http://www.tei-c.org/P5/p5.html>