A Short Introduction to XML Part 2. Definitions

Eugenia Sokolinski

In this section

XML vs HTML
XML Schema
XML rules

HTML

DISPLAY

Manuscripts
<p>The term M. (from Lat. <i>manu scriptum</i>, 'written
by hand') is applied to artefacts (codices,
unbound fascicles, scrolls, or even single
leaves) which carry handwritten texts.</p>

Manuscripts
The term M. (from Lat. *manu scriptum*, 'written
by hand') is applied to artefacts (codices,
unbound fascicles, scrolls, or even single
leaves) which carry handwritten texts.

https://www.w3schools.com/TAgs/default.asp
Only for display (presentational markup). Predefined tags.

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

XML

```
<title>Manuscripts</title>
<text>The term M. <comment>from Lat.
<foreign>manu scriptum <translation>written
by hand</translation></foreign>
</comment> is applied to artefacts
<comment>codices,
unbound fascicles, scrolls, or even single
leaves</comment> which carry handwritten
texts.</text>
```

DISPLAY

Manuscripts
The term M. (from Lat. *manu scriptum*, 'written by hand') is applied to artefacts (codices, unbound fascicles, scrolls, or even single leaves) which carry handwritten texts.

Manuscripts
The term M. -- from Lat. **manu scriptum**, = *written by hand* -- is applied to artefacts -- codices, unbound fascicles, scrolls, or even single leaves -- which carry handwritten texts.

eXtensible Markup Language

As rich and detailed as required

adding metadata (value/information to data)

nothing but text

XML

- Extensible set of tags
  - Add as many as you need!
- Content orientated
  - Tags are semantic
- Standard Data infrastructure
  - Set of rules is defined meaning the possibility to check if the grammar is correct (= validate)
- Allows multiple output forms
  - Different formatting, different bits of information (e.g. show only names, show only dates, show names with dates, show only text, etc etc etc)

HTML

- Fixed set of tags
  - You need to know the set and cannot add
- Presentation oriented
  - Tags are mainly for display
- No data validation capabilities
  - No way to check for errors (The possibilities are very limited, so not much space for error)
- Single presentation
  - What you see is what you get; the entire HTML file (page) is visualized as one single page

XML liberty and constrains

* **Extensible set of tags**
  * Add as many as you need!
* **Content orientated**
  * Tags are semantic
* **Standard Data infrastructure**
  * Set of rules is defined meaning the possibility to check if the grammar is correct (= validate)
* **Allows multiple output forms**
  * Different formatting, different bits of information (e.g. show only names, show only dates, show names with dates, show only text, etc etc etc)

* **Means many many ways to say the same thing**
  * Consistency is essential -> rules
* **Any bit of content may be encoded**
  * First human parsing! Then explain to computer
* **Schemas must be applied**
  * "Descriptive grammar" for your tag set helps and guides your encoding
* **Various transformations possible**
  * Queries and display transformations need to be additionally programmed, or one can use existing software if fits.
  * Data may be also printed directly to PDF as a camera-ready manuscript!

Validation with a schema

- check what should and should not be there
- check which names and values are allowed
- check content
- support data entry
- Separation of concerns: care about the meaning and structure, not about the visualization
- etc.

Greater control over data entry consistency

Validation with a schema

- check what should and should not be there
- check which names and values are allowed
- check content
- support data entry
- Separation of concerns: care about the meaning and structure, not about the visualization
- etc.

XML markup can

- **add** information about the text you are describing

- **elaborate** on the information structure and the contents

- **structure** and preserve structure

- add several **layers of meaning** to a single source text

- **control** the **quality** of the core information

- **check (validate) data** and **verify** results with highest possible precision

- keep **updating consistently freely changing output methods**

- **others can reuse your data**

XML rules and principles

An XML element is made up of a **start tag** (<...>), an **end tag** (</...>), and data in between. The start and end tag are the same string (word), the only difference is the slash (as in HTML).
Example:

    <editor>Alessandro Bausi</editor>

Example of another element with the same value:

    <author>Alessandro Bausi</author>

Tags do not necessarily need data between them, in which case they may be abbreviated

    <author></author> is the same as

Each tag can have an **attribute** to attach additional element to an element; attributes are a **name-value** pair separated by an **=** sign; the value must be in **quotation marks**
Example

    <editor type="chief">Alessandro Bausi</editor>
    <editor type="assistant">Eugenia Sokolinski</editor>

XML rules and principles

Tags may be nested to reflect the hierarchy of elements
Example
 <editors>
        <editor type="chief">Alessandro Bausi</editor>
        <editor type="assistant">Eugenia Sokolinski</editor>
</editors>

<book>
        <editors>
                <editor type="chief">Alessandro Bausi</editor>
                <editor type="assistant">Eugenia Sokolinski</editor>
        </editors>
</book>

Rules for a well-formed XML

- 1. Each tag which is opened must be closed

- 2. Nesting is Ok inside tags but not inside attributes. In case of nesting no overlapping is allowed

Example: wrong
```
<book>
    <editors>
        <editor type="<role>chief</role>">Alessandro Bausi</editor>
        <editor type="assistant">Eugenia Sokolinski</editor>
</book>
    </editors>
```

Example: correct
```
<book>
    <editors>
        <editor type="chief"><persName><forename>Alessandro</forename>
<surname>Bausi</surname></persName></editor>
        <editor type="assistant">Eugenia Sokolinski</editor>
    </editors>
</book>
```

12

Rules for a well-formed XML

- 3. The whole document must be contained within one root element

```
<TEI xmlns="http://www.tei-c.org/ns/1.0" type="place" xml:lang="en" xml:id="LOC3577Gondar">

  <teiHeader>

    <fileDesc>

      <titleStmt>

        <title>Gondar</title>

      </titleStmt>

        …

</TEI>
```

- 4. Reserved symbols have to be escaped, that is encoded otherwise

    < use &lt

    > use &gt

    & use &amp

Example 1

<book>
<page>
<paragraph>.....</paragraph>
<paragraph>These events suggest that there were influential groups of both pagans and Christians at Aphrodisias, and that there was some tension
</page>
<page>
between them; this may be the cause of the strife referred to in no. 64. One other tradition may be relevant.</paragraph>
</page>
</book>

Example 1

<book>
<page>
<paragraph>.....</paragraph>
<paragraph>These events suggest that there were influential groups of both pagans and Christians at Aphrodisias, and that there was some tension
</page>
<page>
between them; this may be the cause of the strife referred to in no. 64. One other tradition may be relevant.</paragraph>
</page>
</book>

Example 1

<book>
<page>
<paragraph>.....</paragraph>
<paragraph>These events suggest that there were influential groups of both pagans and Christians at Aphrodisias, and that there was some tension
</page>
<page>
between them; this may be the cause of the strife referred to in no. 64. One other tradition may be relevant.</paragraph>
</page>
</book>

Example 1

```xml
<book>
<page><paragraph>.....</paragraph>
<paragraph part="start">These events suggest that there were influential groups of both pagans and Christians at Aphrodisias, and that there was some tension</paragraph>
</page>
<page>
<paragraph part="end">between them; this may be the cause of the strife referred to in no. 64. One other tradition may be relevant.</paragraph>
</page>
</book>
```

Example 1

<book>
<paragraph>These events suggest that there were influential groups of both pagans and Christians at Aphrodisias, and that there was some tension
between them; this may be the cause of the strife referred to in no. 64. One other tradition may be relevant.</paragraph>
....
</book>

Example 2

<note>

<date type="foundation" cert="low">17th century</date>

<desc type="foundation">Although the name "Gondar" is first mentioned in the 14th century to designate a regiment of soldiers and it was subsequently used to designate a small market town inhabited by Muslim traders and Beta ʾƎsrāʾel craftsmen, it developed as a major centre only after ʾaṣe Fāsiladas chose it as his camp in 1636-1637.

<ab type="history">The town of Gondar is located in today's North Gondar Zone (historically Bagemdər province) in the ʾAmhārā Regional State.

…

</desc>

</ab>

</note>

Example 2

\<note\>

\<date type="foundation" cert="low"\>17th century\</date\>

\<desc type="foundation"\>Although the name "Gondar" is first mentioned in the 14th century to designate a regiment of soldiers and it was subsequently used to designate a small market town inhabited by Muslim traders and Beta ʾƎsrāʾel craftsmen, it developed as a major centre only after ʾaṣe Fāsiladas chose it as his camp in 1636-1637.

\<ab type="history"\>The town of Gondar is located in today's North Gondar Zone (historically Bagemdər province) in the ʾAmhārā Regional State.

…

\</desc\>

\</ab\>

\</note\>

Example 2

```
<note>

<date type="foundation" cert="low">17th century</date>

<desc type="foundation">Although the name "Gondar" is first mentioned in the 14th century to designate a regiment of soldiers and it was  subsequently used to designate a small market town inhabited by Muslim traders and Beta ʾƎsrāʾel craftsmen, it developed as a major centre only after ʾaṣe Fāsiladas chose it as his camp in 1636-1637. </desc>

<ab type="history">The town of Gondar is located in today's North Gondar Zone (historically Bagemdər province) in the ʾAmhārā Regional State.

…

</ab>

</note>
```

Example 3

<name key="galawdewos">ገላውዴዎስ፡</name>

<p>ማር፡ ገላውዴዎስ፡ ዘነግሠ፡ በ<ሀገር>ኢትዮጵያ፡ </ሀገር> </p>

<bibl><author>Getatchew Haile &amp Macomber</author></bibl>

Please continue with Part 3 on TEI and Guidelines