Why use XML-TEI markup?

Alejandro Bia
Universidad Miguel Hernández
Spain

Some questions we asked ourselves…

- ***Why use markup at all?***
- ***Why choose XML, and TEI?***
- ***Why is XML better than SGML?***
- ***Why is TEI better than other markup vocabularies?***

Why use markup at all?

Markup:

- is any means of making explicit an interpretation of a text.
- both **for computers** and **for humans** (text encoders)
- useful for **data** and **metadata**

The alternatives are:

- Plain text (like the Gutenberg project)
- Facsimile formats: TIFF / JPEG / PDF / DejaVu
- Marked-up text + facsimiles → more functionality

Three types of markup:

- Punctuational
- Presentational or procedural
- **Descriptive or structural** ←
  (adds semantic value to the text)

When are facsimilies better than text?

Digital facsimiles are used to reproduce newspapers with historical value (e.g. "Izquierda Republicana"):

Or also to accurately reproduce old printings (e.g. "La Celestina"):

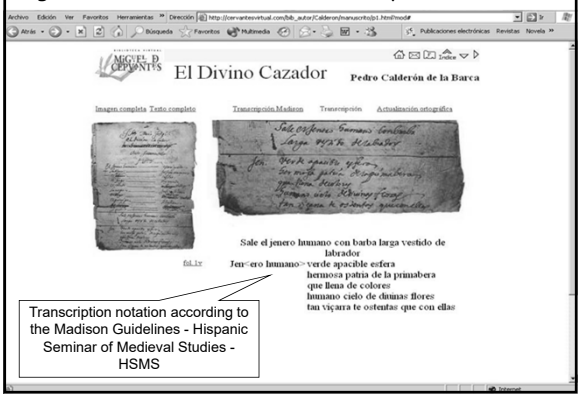
But the most interesting use is to reproduce manuscripts (e.g. "El Divino Cazador"):

And, for everything else, there's…

TEXT:
plain text, marked-up text, formated text…

Example images from the MCDL website.

Digital facsimiles with text transcriptions



Transcription notation according to the Madison Guidelines - Hispanic Seminar of Medieval Studies - HSMS

Markup is based on the OHCO Model

OHCO:

- ***Ordered Hierarchy of Content Objects***
- ***A book is a good example of what can be represented in this way.***
- ***Each Content Object is marked at its beginning and end with a tag.***

Limitations of the OHCO Model

- ***OHCO is good in most cases (Paretto law - 80/20)***
- ***Except, for instance…***

… when we need overlapping markup (forbidden by design)

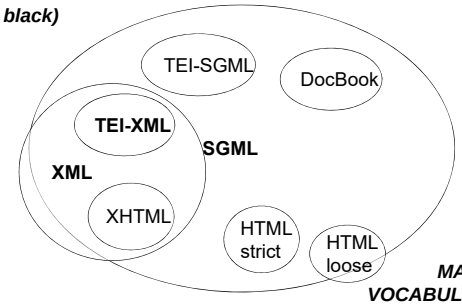
XML syntaxes for overlap, such as in TEI or in LMNL*, **adopt five different techniques**:

- **milestones**
- **fragmentation**
- **flattened**
- **multiple document**
- **standoff**

* LMNL: The Layered Markup and Annotation Language

Markup Languages and Vocabularies

MARKUP LANGUAGES (in black)



MARKUP VOCABULARIES (in red)

8

The rules of XML: compared to SGML

The XML markup language imposes strict rules:

- ***All opened tags must be closed***
- ***All empty tags should use the / to indicate that the tag is empty.***
- ***All attribute values within an XML document must be quoted.***
- ***The names of Elements and Attributes used in an XML document must be an exact match of how they are declared in the DTD (this includes letter-case).***

XML imposes **more constraints** than SGML.

Higher Constraint = Higher Predictability

This makes it **easier to use and to process**.

10

XML and digital preservation concerns…

***Digital preservation** is focused on:*

- Preventing data loss, mainly due to deterioration of the media
- Preventing obsolescence of the data formats

XML is plain text + tags:

- Tags are text as well: <tag also="text">
- … so XML is just text
- … the oldest, simplest, most widely used computer file format

However, we still have to be aware of changes in image formats:

- Tiff, Jpeg, Png…

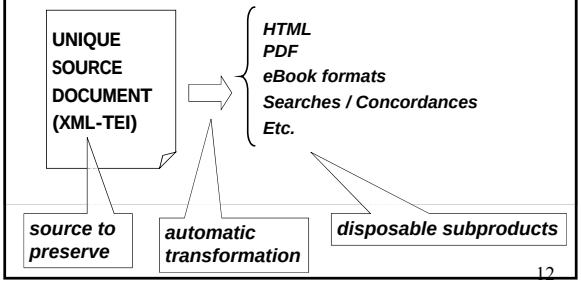
Many file formats rely on XML nowadays:

- MS .docx .pptx. .xlsx
- Epub, SVG
- Java JAR files
- And a long etc. …

Single source, multiple outputs, automatic transformation…

Avoid duplication of information

→ Principle of UNIQUE SOURCE

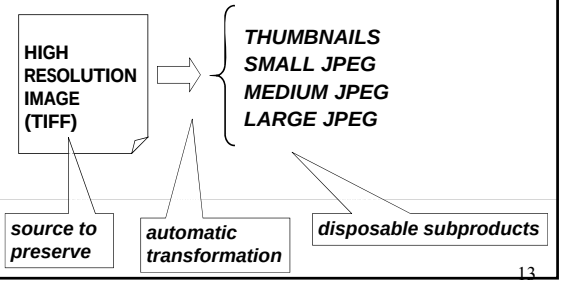


12

Likewise… with images:
single source, multiple outputs, and automatic transformation

Avoid duplication of information

→ PRINCIPLE OF UNIQUE SOURCE



13

XML as a family of technologies and tools

XML is not just a markup language but a family of standards and related tools that interoperate:

- ***XML***
- ***Namespaces***
- ***XPath***
- ***XSL or XSLT***
- ***DTDs and Schemas***
- ***XQuery***
- ***Xlink and XPointer***
- ***and many powerful software tools…***

14

So, why use XML?

DH Research potential

- Precise searches
- Automatic concordances
- Comparative analysis (parallel alignment)
- Linguistic research
- Statistical research

Multipurpose

- can be used for different research purposes
- can be rendered in several formats
- adequate for Web publishing

So, why use XML?

Ease of preservation

- It is text
- Easily converted to other formats
- Easily upgradeable

Drawbaks:

- encoding is more expensive
- requires skilled personnel
- overlapping markup is not allowed

And, why use TEI?

It's the most complete markup vocabulary

- More than 500 elements
- Specialized tagsets
- Customizable
- A de-facto standard for scholarship

Plus

- Community collaboration (TEI list, conference, SIGs)
- Documents interchange
- Tool sharing
- Support