Before we begin

- Launch <oXygen/>
- Open a web browser and go to
  - http://www2.flwi.ugent.be/download/l.txt
- Copy the entire license key (including the "start" and "end" lines)
- Paste into the dialog box in <oXygen/>
- Workshop web site
  - http://ghent.obdurodon.org

Session 1: Introduction to XML

Historical documents, digital approaches
University of Ghent, 2013-09-05

David J. Birnbaum
djbpitt@gmail.com
http://www.obdurodon.org

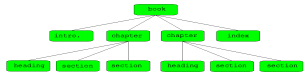
Outline

- Overview: XML and text
- Pseudo-markup and markup
- Elements
- Attributes
- Well-formedness
- The creation of a digital text
- Editing XML in <oXygen/>
- [Hands-on practice editing XML]

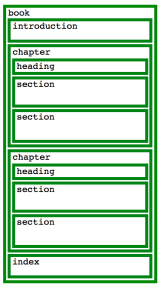
Overview

- OHCO: ordered hierarchy of content objects
- Three views of XML
  - Boxes
  - Tree
  - Serialization

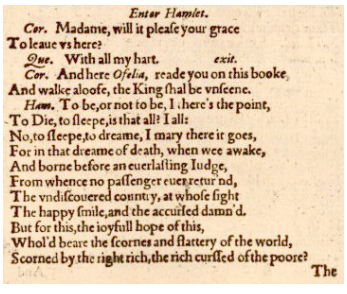
Three views of a document




```
<book>
  <introduction> … </introduction>
  <chapter>
    <heading> … </heading>
    <section> … </section>
    <section> … </section>
  </chapter>
  <chapter>
    <heading> … </heading>
    <section> … </section>
    <section> … </section>
  </chapter>
  <index> … </index>
</book>
```



Hamlet, First quarto, 1603



Enter Hamlet.
Cor. Madame, will it pleaſe your grace
To leaue vs here?
Que. With all my hart. *exit.*
Cor. And here *Ofelia*, reade you on this booke,
And walke aloofe, the King ſhal be vnſeene.
Ham. To be, or not to be, I there's the point,
To Die, to ſleepe, is that all? I all:
No, to ſleepe, to dreame, I mary there it goes,
For in that dreame of death, when wee awake,
And borne before an euerlaſting Iudge,
From whence no paſſenger euer retur'nd,
The vndiſcouered country, at whoſe ſight
The happy ſmile, and the accurſed damn'd.
But for this, the ioyfull hope of this,
Whol'd beare the ſcornes and flattery of the world,
Scorned by the right rich, the rich curſſed of the poore?
The

Pseudo-markup

- Hamlet
  - Stage directions
  - Speeches
  - Speakers
  - Other characters
  - Metrical lines
- General
  - Paragraph spacing and indentation
  - Centering and bolding of titles
  - Hanging indentation for bibliographic lists
  - Italics for emphasis, foreign words, book titles, etc.

The XML view of content and markup

- Content is the textual data
  - Transcribed from source (e.g., a manuscript)
  - Created by the editor (e.g., manuscript catalogue)
- Markup describes the role of different data components
- No pseudo markup in your content
  - Editorial parentheses, square brackets, angle brackets, slashes and backslashes, etc.

XML building blocks

- Textual (character data) content
- Elements
- Attributes

Elements

- Elements have matching start and end tags
  <title> … </title>
- (Some elements are empty and self-closing)
  <bookmark/>
- Element names must begin with a letter and may contain letters, digits, and underscores (no spaces; no other punctuation)
  Underscore: <personal_name>
  Camel case: <personalName>

The “X” in XML

- eXtensible Markup Language
  - The user determines the tag set
  - Pro: you determine how to characterize your data
  - Con: you are responsible for determining how to characterize your data
- You decide
  - What to tag
  - How to tag it (what to call it)

Three types of markup

- Descriptive: what the object is (emphasized)
  - yes!
- Presentational: what the object looks like (italicized)
  - yes!
- Procedural
  - [instructions to the machine]

Why DH projects use descriptive markup

- Italics: emphasis, foreign, book title, etc.
- Emphasis: italic, bold, loud (audio device), etc.
- Separation of levels: content and presentation
  - Encode descriptively
  - Transform to presentational final form for rendering (HTML, PDF, etc.)
- Multipurposing: format the same content objects different way for different purposes

Texts and trees

- Why XML looks at texts as trees
  - Computers can traverse trees quickly
  - Documents *are* hierarchical, right?
- Hierarchical challenges
  - Multiple, overlapping hierarchies
    - Physical hierarchy: folios, lines
    - Intellectual hierarchy: texts (with subelements: chapters, sections, paragraphs, etc.)
  - Relationships at a distance
    - Cross-references and other pointers
    - References and pointers to other documents

Attributes

- Qualifying information about elements
- Encoded inside the start tag, after the element name
  - Attribute name="value" pair
- <place xml:lang="fr">Paris</place>
- <title type="journal">Journal of Digital Humanities</title>
- Attribute names are subject to the same rules as element names
- Attribute values must be quoted (matching single or double straight quotation marks)

An XML document must be well-formed

- Single root element
- Proper nesting (no overlapping tags)
  Good: <foreign>oui!</foreign>
  Bad: <foreign>oui!</foreign>
- Name and name start characters for element and attribute names
- Attribute values must be quoted (single or double)
- Reserved characters must be encoded as *entities*
  - & &
  - < <
  - > >

What's wrong?

```
<author>Michael Kay</author>
<title edition = 4>XSLT 2.0 and XPath 2.0 Programmer's
    Reference</title>
<published date = 2008>
  <publisher>John Wiley & Sons, Inc.<publisher>
  <pubPlace>10475 Crosspoint Boulevard,
    Indianapolis, IN 46256</pubPlace>
</published>
<ISBN num="978-0-470-19274-0">
<dedication>
  <i><b>To Anyone Who Uses This Book
  To Make the World a Better Place</i></b>
</dedication>
```

Creating a digital text

- In theory
  1. Document analysis, then …
  2. Schema development, then …
  3. Markup
- In practice
  - The preceding is a cycle, and not a sequence
  - Markup is part of the process of document analysis
- Nonetheless
  - Start with document analysis, not with angle brackets

Why use an XML editor?

- <oXygen/> (http://www.oxygenxml.com)
- XML-aware
  - Real-time and on-demand validation
  - Completion hinting
  - Multiple views
  - (Schema-aware … stay tuned)
- IDE (integrated development environment)
  - XSLT (eXtensible stylesheet language transformations)
  - Debugger
  - Other development tools

Editing XML in <oXygen/>

- Create a new file
  - File → New → New document → XML document
  - Short cuts: Ctrl+n; leftmost icon at top of screen
- Create an element
  - Type a start tag (in angle brackets)
    - <oXygen/> automatically creates the matching end tag
- Change an element
  - Change the start tag; the end tag changes automatically to match
- Wrap text in an element
  - Select the text, type Ctrl+e (for 'element'), type the element name
  - To use the same wrapper as last time, select and type Ctrl+/
- Split an element
  - Put the cursor at the split point and type Shift+Alt+d
- Pretty-print (wrap) the text
  - Shift +Ctrl+p; pretty-print (indentation) icon

Hands on

- Choose a document with a regular structure
  - Google a recipe for your favorite food
  - Find a menu from your favorite restaurant
  - Encode a letter by Oscar Wilde (http://law2.umkc.edu/faculty/projects/ftrials/wilde/lettersfromwilde.html)
  - Encode a sonnet by William Shakespeare (http://www.shakespeares-sonnets.com/all.php)
  - … or choose your own text
- Copy into a new XML document in <oXygen/>
- Mark it up in XML
  - Imagine a research or other context where you're marking up your text for a reason
  - Identify and tag major structural components
  - Identify and tag small, in-line items that might be useful
  - explore tagging options at http://www.tei-c.org

Credits

- Three views of a document
  - http://www.wwp.brown.edu/outreach/seminars/uvic_xslt_2012/presentations/xslt/xml_and_xpath_01.xhtm