Recipe document analysis

Simon Mahony

From an original document by Susan Hockey

This document is part of a collection of presentations and exercises on XML. For full details of this and the rest of the collection see the cover sheet at: http://ucloer.eprints-hosting.org/id/eprint/19

Version 1.0

Session 4

Matters arising

More on Oxygen

Tables (cf letter 3: Zoe Zettern)

Text Encoding Initiative: TEI Lite

National Trust: more structure

Deconstruct and model the letter

<addressee>
Ms Jane Bloggs — name (title / firstname / lastname)
23 Nice Drive — addressline
West Ham — addressline
London — town / city
EC9 2LZ — postcode
</addressee>
<date>
10 January 2004
</date>
<body>
Dear Ms Bloggs greeting

MA in Library and Information Studies — heading

Thank you for coming for an interview etc etc — para

Yours sincerely — closingtext

Vanda Broughton — name (title / firstname / lastname)
Programme Director — jobtitle
</body>

All Elements need to be declared:

<!ELEMENT letters (letter+)>

<!ELEMENT letter (addressee, date, body)>

<!ELEMENT addressee (name, addrline+,town,state?,postcode,country?)>

<!ELEMENT date (#PCDATA)>

<!ELEMENT body (greeting, heading, p+, closer)>

<!ELEMENT name (title?, firstname, lastname)>
<!ELEMENT title (#PCDATA)>
<!ELEMENT firstname (#PCDATA)>
<!ELEMENT lastname (#PCDATA)>

<!ELEMENT addrline (#PCDATA)>
<!ELEMENT town (#PCDATA)>
<!ELEMENT state (#PCDATA)>
<!ELEMENT postcode (#PCDATA)>
<!ELEMENT country (#PCDATA)>

```
<!ELEMENT letters (letter+)>
<!ELEMENT letter (addressee, date, body)>
<!ATTLIST letter type CDATA #REQUIRED>
<!ELEMENT addressee (name, addrline+,town,state?, postcode,country?)>
<!ELEMENT name (title?, firstname, lastname)>
<!ELEMENT title (#PCDATA)>
<!ELEMENT firstname (#PCDATA)>
<!ELEMENT lastname (#PCDATA)>
<!ELEMENT addrline (#PCDATA)>
<!ELEMENT town (#PCDATA)>
<!ELEMENT state (#PCDATA)>
<!ELEMENT postcode (#PCDATA)>
<!ELEMENT country (#PCDATA)>
<!ELEMENT date (#PCDATA)>
<!ELEMENT body (greeting, heading, p+, closer)>
<!ELEMENT greeting (greetingtext, title, lastname)>
<!ELEMENT greetingtext (#PCDATA)>
<!ELEMENT heading (#PCDATA)>
<!ELEMENT p (#PCDATA|table)*>
<!ELEMENT table (row+)>
<!ELEMENT row (column+)>
<!ELEMENT column (#PCDATA)>
<!ELEMENT closer (closingtext,title?, firstname, lastname,jobtitle)>
<!ELEMENT closingtext (#PCDATA)>
<!ELEMENT  jobtitle (#PCDATA)>
```

Oxygen offsite

- XML Editor (not Author)
- 30 free trial licence
  - If you're away from internet
- Buy your own copy
  - Academic and education price: $64 (at time of writing)
  - Useful if you don't have regular internet access
- Remember: these will be newer versions than on College network

Using Oxygen to check and tidy files (EGS)

- DTD
  - red tick to check validation
- XML
  - red tick for validation
  - blue tick for well formed
  - blue horizontal lines to tidy layout
  - Document > Edit > Toggle Line Wrap
- Both
  - red marker in right-hand margin indicates an error
  - asterisk on file name tag = not saved
  - practice tags

Oxygen shortcuts (EGS)

- XML file with a DTD (internal or linked)
- To add elements
  - type < and you will be offered valid elements
  - Ctrl+e will also offer valid elements
  - Both with opening and closing tags
- To add attributes
  - put cursor at end of start tag (before >) type a space
  - A list of valid attributes appears
- To add attribute values
  - valid attribute values will be offered if #REQUIRED not if #IMPLIED or CDATA

DTD Language Syntax Summary

- Declaration
  - Internal / System / Public
- Elements
  - Declaration / Data element / Empty element / One or more
  - Alternatives / Combinations / Occurrence / Mixed content
- Attributes
  - Declaration / Types / Values
- Entities
  - Declaration / Default / Character entities / General

Document analysis: Letters (worksheet 3)

- Formal structure
  - Letter type?
  - Who is it addressed to? Addressee / Details
    - Name / address / postcode
  - Date
  - Body
    - Greeting
    - Heading
    - Stuff at the end / closer

Letter 3: Zoe Zettern (table XML)

```
<p> Please be aware that the programme consists of
    <table>
        <row>
            <column>Modules</column>
            <column>8</column>
            <column>10 credits each</column>
        </row>
        <row>
            <column>Dissertation</column>
            <column>1</column>
            <column>40 credits</column>
        </row>
    </table>
</p>
```

Letter 3: Zoe Zettern (table DTD)

- <!ELEMENT p (#PCDATA | table)*>
  - p contains text or table zero or many times (some letters have no tables)
- <!ELEMENT table (row+)>
  - table must have at least one row
- <!ELEMENT row (column+)>
  - row must have at least one column
- <!ELEMENT column (#PCDATA)>
  - column (here cell) contains ONLY text (leaves of the tree)

Ambiguous content models

- Content models must not be ambiguous
- Bibliographic item: title followed by author or author followed by title allowing for no author.

<!ELEMENT bibitem ((title, author?) | (author?, title))>

- Here title could match first or last – not permissible

- Consider: <bibitem><title>Beowulf</title></bibitem>

<!ELEMENT bibitem ((title, author?) | (author, title))>

Making author element required when before title removes ambiguity.

Documentation

- In the 'real world' documentation is essential
- Often not done due to pressure of time
- Allows others to follow
- Makes research repeatable
- Reproducible evidence and method
- Allows scholarly citation in notes (refs to markup, DTDs etc)
- Annotate files with comments <!-- … -->
- Add readme.txt files to folders
- Project wikis are common although generally hidden
- Has to be done as you progress through not retrospectively
- Larger the project the more important
- Good working practice

Text Encoding Initiative

- TEI: "a consortium which collectively develops and maintains a standard for the representation of texts in digital form"
- Guidelines for machine-readable texts
- TEI council / Workgroups / SIGs
- Projects listing
- Annual conference (sponsored by Oxygen)
- TEL Lite cut down version suitable for all but highly specialised use

National Trust website

- National Trust: http://www.nationaltrust.org.uk/
- Places to visit:
  - Bembridge Windmill
  - Bateman's

- Note: again highly structured format of data
- Structure consistent across all the pages
- Differences in print and online

- Create a data model and DTD to generate XML file