Tools for Editing and Publishing TEI Documents

James Cummings

April 2009
This section provides a brief look at some of the technology for editing and publishing in XML, especially for TEI XML users, and issues related to that in the area of data capture and editing.
Summary

How does a TEI user do the following?

- Data capture
- Editing
What tools do we need?

- Appropriately expressive vocabularies (eg TEI XML)
- Syntax-checking document creation tools (ie editors)
- Document transformation tools
- Document delivery tools
- Document storage and management tools
- Programming interfaces
- Specialized applications
Two stages to get a TEI text

- capture the text
- create the markup

Often they occur simultaneously; but often not. Note that the markup does not necessarily all have to be in the same file.
Categories of creation tools

- scanning/OCR
- data-entry vendors
- software to add tagging automatically
- editors

followed by

- validators, well-formedness checkers
- proofing aids, data integrity checkers
Scanning and OCR software generally produce only minimal HTML or Word (e.g., recognizing paragraph breaks, font changes etc).

Data-entry vendors in theory would insert whatever markup you wanted, but at a price. They generally prefer HTML or TEI Lite or some such well-known DTD.

TEI is creating a standard slimmed-down vocabulary for initial encoding that may be useful in mass-digitisation projects called 'TEI Tite'.
Editor types

Editing tools cover a wide spectrum:

- Basic text editors
- General programmers' editors
- XML-aware programmers' editors
- XML-specific editors
- Word-processors which can export XML
- Data-entry forms
- Image-specific editors

It is likely that people in different roles need different tools.
Things to look for in specialist XML editors

- schema-aware
- constraining element entry
- IDE features
- customizable
- validation, preferably continual
- Multiple display views (as tree, with tags, formatted etc)
- folding structures
- context-sensitive help

For XML editing, oXygen, Emacs, jEdit, XMetaL, XMLSpy, Stylus Studio, Arbortext Adept are all worth a look.

For image editing, try University of Victoria Image Markup Tool or Edition Production and Presentation Technology (EPPT).
1. **Multiple modes for editing XML documents**
   - Visual editing mode - CSS based - for XML documents
   - Grid mode - spreadsheet like editing for regular XML data structures
   - Text mode - full access to the XML source

2. **Ready-to-use support for TEI P4 and P5**
   - New TEI document templates
   - Visual presentation and editing of TEI documents in the Author mode
   - TEI specific actions for inlines, structure, images, lists and tables
   - Pre-configured transformation scenario to get PDF and HTML
   - Pre-configured XML catalog to resolve remote references locally
oXygen Features (2)

• The possibility to add, extend and deploy the support for a framework
  • Through a configuration GUI - no need for coding - you can configure similar support as oXygen has for TEI, DocBook, DITA and XHTML for another framework or you can extend an existing framework with new actions.
  • A Java developer API and all the code for the common actions that are used for TEI, DocBook, DITA and XHTML is open source.

• Support for all the schema languages: Relax NG, Schematron, XML Schema, DTD, NVDL.
  • The support for Relax NG is specially important in relation with TEI as the main TEI P5 schemas are in Relax NG.
  • oXygen offers content completion proposals based on the Relax NG schema, presents the schema annotations as tooltips.
  • Schematron allows extra constraint-checking.
  • The NVDL support allows easily handling of TEI documents that contain foreign markup.
• **XQuery support against XML databases** TEI documents are now often stored in XML databases like eXist. oXygen offers support for all XML databases to explore, edit and query the data.

• **XSLT and FOP support**
Converting TEI to different output formats is done using XSLT and in case of PDF output also using a Formatting Objects Processor (FOP).

• **WebDAV and FTP support**
View and edit files from any WebDAV enabled repository (including WebDAV enabled XML database or Content Management Systems) or from remote FTP servers.
Facilitates collaborative work - content change tracking and SVN support

In version 10.2 oXygen (beta stage right now and planned to be available in a couple of weeks) adds visual change tracking management for content changes supporting multiple authors. oXygen also includes a full Subversion client allowing collaborative work on large projects.

Spell-checking support

Spell checking is xml:lang aware and acts on the whole document or as you type.
But maybe most important...

- Multi-platform and flexible licenses
  - oXygen is available on Windows, Mac OS X, Linux, Solaris, etc.
  - They have an enlightened academic pricing policy
  - The named-user based license allows the same user to use any oXygen distribution on any platform or machine: the same license covers you are work, laptop, and home.
<head>THE LEGEND OF SLEEPY HOLLOW.</head>

<head>FOUND AMONG THE PAPERS OF THE LATE DIEDRICH KNICKERBOCKER</head>

<lg>
A pleasing land of drowsy head it was,
Of dreams that wave before the half-shut eye;
And of gay castles in the clouds that pass,
For ever flushing round a summer sky.
</lg>

<trailer>CASTLE OF INDIFFERENCE.</trailer>

<lg>
IN the bosom of one of those spacious coves which indent the eastern shore of the Hudson, at that broad expansion of the river denominated by the ancient Dutch navigators the Tappan Zee, and where they always prudently shortened sail, and implored the protection of St. Nicholas when they crossed, there lies a small market-town or rural port, which by some is called Greensburgh, but which is more generally and properly known by the name of Tarry Town. This name was given, we are told, in former days, by the good housewives of the adjacent
the dominant dignity and absolute sway with which he lorded it in his little empire, the school, and became wonderfully gentle and ingratiating. He found favor in the eyes of the mothers, by petting the children, particularly the youngest; and like the lion bold, which whilst so magnanimously the lamb did hold, he would sit with a child on one knee, and rock a cradle with his foot for whole hours together. </p>

<p>In addition to his other vocations, he was the singing-master of the neighborhood, and picked up many bright shillings by instructing the young folks in psalmody. It was a matter of no little vanity to him, on Sundays, to take his station in front of the church gallery, with a band of chosen singers; where, in his own mind, he completely carried away the palm from the parson. Certain it is, his voice resounded for above all the rest of the congregation; and there are peculiar quavers still heard half a mile off, on Sunday morning, which are Ichabod Crane. Thus, by commonly denominated "by hook and by crook," the worthy pedagogue got on tolerably enough, and was thought, by all who understood nothing of the labor of headwork, to have a wonderfully easy life of it. </p>

<p>The schoolmaster is generally a man of some importance in the female circle of a rural neighborhood; being considered a kind of idle gentlemanlike personage, of vastly superior taste and accomplishments to the rough country swains, and, indeed, inferior in learning only to the parson. His appearance, therefore, is apt to occasion some little stir at the tea-table of a farmhouse, and the addition of a supernumerary dish of cakes or sweetmeats, or, peradventure, the parade of a silver tea-pot. Our men of letters, therefore, was peculiarly happy in the smiles of all the country damsels. How he would figure among them in the churchyard, between services on Sundays! gathering grapes for them from the wild vines that overran the surrounding trees; reciting for their amusement all the epitaphs on the tombstones; or sauntering, with a whole bevy of them, along the banks of the adjacent mill-pond; while the more bashful country bumpkins hung
the dominant dignity and absolute sway with which he lorded it in his little empire, the school, and became wonderfully gentle and ingratiating. He found favor in the eyes of the mothers, by petting the children, particularly the youngest; and like the lion bold, which whilst so magnanimously the lamb did hold, he would sit with a child on one knee, and rock a cradle with his foot for whole hours together. </p>

<p>In addition to his other vocations, he was the singing-master of the neighborhood, and picked up many bright shillings by instructing the young folks in psalmody. It was matter of no little vanity to him, on Sundays, to take his station in front of the church gallery, with a band of chosen singers; where, in his own mind, he completely carried away the palm from the parson. Certain it is, his voice resounded far above all the rest of the congregation; and there are peculiar quavers still to be heard in that church, which may even be heard half a mile off, quite to the opposite side of the mill-pond, on a still Sunday morning, which are said to be legitimately descended from the nose of <name>Ichabod Crane</name>. Thus, by divers little make-shifts in that ingenious way which is commonly denominated “by hook and by crook,” the worthy pedagogue got on tolerably enough, and was thought, by all who understood nothing of the labor of headwork, to have a wonderfully easy life of it. </p>

<p>The schoolmaster is generally a man of some importance in the female circle of a rural neighborhood; being considered a kind of idle gentlemen-like personage, of vastly superior tastes and accomplishments to the rough country swains, and, indeed, inferior in learning only to the parson. His appearance, therefore, is apt to occasion some little stir at the tea-table of a farmhouse, and the addition of a supernumerary dish of cakes or sweetmeats, or, peradventure, the parade of a silver tea-pot. Our man of letters, therefore, was peculiarly happy in the smiles of all the country damsels. How he would figure among them in the churchyard, between services on Sundays! gathering grapes for them from the wild vines that overrun the surrounding trees; reciting for their amusement all the epitaphs on the tombstones; or countering, with a whole bevy of them, along the
Tagless editing in oXygen

The Legend of Sleepy Hollow
Washington Irving

THE LEGEND OF SLEEPY HOLLOW, FOUND AMONG THE PAPERS OF THE LATE DIEDRICH KNICKERBOCKER.
A pleasing land of drowsy head it was,
Of dreams that wave before the half-shut eye,
And of gay castles in the mist that pass,
For ever rushing on the summer sky.
CASTLE OF INDOLENCCE...
Editing Options

EPPT
Editing Options

UVic IMT screenshot 2
What is missing, or hard, in the TEI editing world

• Only a few editors like oXygen or XMetaL which combine visual feedback with code editing
• Visual, or WYSIWYG, editors embedded in web applications (eg in a CMS); most web editors are for XHTML (cf Google Docs)
• Reliable conversion to and from Word and OpenOffice styles.

Note:
• the general inability of word-processors to nest inline inside inline, or block inside block
• the difficulty of extrapolating a hierarchical structure from a sequence of free-standing headings at assorted levels
• the tedious programming required to trace the ancestry of styles in Word and OpenOffice
• the lack of a facility in word processors to stop the user formatting by hand
• but the TEI does provide partial transformations to/from both Word and OpenOffice
Using TEI XSL

This is a family of XSL stylesheets which are designed to render simple TEI documents. For the purpose of the TEI Consortium, they

• Implement the processing of ODD files behind Roma to make schemas and documentation
• … and thus generate the TEI Guidelines in HTML
• … and transform the TEI Guidelines to LaTeX for typesetting
• Render TEI Lite documents to
  1. HTML
  2. XHTML
  3. XSL FO (formatting objects, for page makeup)
  4. LaTeX (for typesetting)
Limitations

These stylesheets only do what were designed to do!

• They do not provide a rendering of all TEI elements
• They do not implement all possible values of every `@rend` attribute
• The different output formats are not always in sync, or give the same result

but they do deal with quite a few common problems.
Related stylesheets

We also maintain in XSLT:

- A simple Docbook to TEI conversion
- Conversions to and from OpenOffice XML
- Conversions to and from Word 2007 XML
- Conversion from TEI P4 to TEI P5
Output assumptions

The stylesheets attempt to work in the same way with each of the three supported output formats, but note:

- The HTML output is designed to work with an associated CSS stylesheet, which takes care of much of the detailed spacing and font work; however, the HTML is in charge of features such as the numbering of sections.

- The LaTeX output is designed for people who understand how to use existing LaTeX packages and classes; it therefore tries to produce reasonably readable TeX markup, with high-level commands whose effects will be determined by LaTeX (including numbering and spacing).

- The XSL FO output produces a very detailed specification of the output layout, with all the details of fonts, numbering, vertical and horizontal spacing specified *in situ*. The FO processor is only responsible for line and page breaking, and hyphenation.
• Transform to HTML: no dynamic content
• Off-the-shelf metadata retrieval: Philologic, XTF, or now-dated TEI-Publisher
• Using Apache's Cocoon: dynamic transformations
• Build on top of an XML Database: eXist and XQuery
• Roll-your-own: Build a bespoke solution