TEI XML : Practical Exercises

Lou Burnard

1 Before you start

These exercises assume you will be using the TEI-Knoppix CD, release 3.6 or later. Unless you are using this TEI-Knoppix CD, you will need access to the net and a web browser such as Firefox or Opera. You will also need an XML-aware editor (in our case, oXygen or TEI Emacs).

If you are using the Knoppix CD, you will need to do the following before you start:

- Put the CD into the CD drive of your computer and reboot the computer. All being well, your computer will Do the Right Thing (you may need to adjust the BIOS to make it reboot from the CD rather than the hard disk or network).
- The TEI-Knoppix CD has to check out all the components of your system and load the appropriate drivers before it can start loading the application software bundled with it. This can take five minutes or so, depending on your system specification. Just hit the RETURN key if it pauses during this process. When you see the TEI Knoppix web page, the startup is complete; you will also a random 'helpful' tip from the Knoppix operating system, which you can dismiss without further attention.
- This version of Knoppix starts up with a UK keyboard. If you want a different one, click on the national flag in the bottom right corner. If the flag you want doesn't appear for selection, right click with your mouse to open the "Add mapping" dialog, which will allow you to add it.

The example files used in these exercises are all available in a folder called samples on your ramdisk, if you are using the Knoppix CD. If not, you can download them from the website associated with this tutorial on the TEI website.

2 Editing an XML file

There are literally dozens of different editors you can choose from to work with XML. For this course we will use a commercially-produced piece of software called oXygen, which has the following merits:

- it works in exactly the same way on most flavours of Linux, Windows or Macintosh,
- it has a very full range of features — we will show only a handful of them in this workshop
- it has an enlightened (i.e. cheap) academic licensing policy
- it comes pre-customised to work with the TEI schema (amongst others)
- it looks familiar to Windows users

If you find oXygen totally unfriendly, don't panic. There are other editors: you could try Emacs for example, which is also included on the TEI-Knoppix CD.

- Start oXygen. If you can't find its icon on your desktop, you'll find it by clicking the Big K button at bottom left, and selecting Editors from the All Applications menu that opens up.

- The oXygen application window opens. Wait for the program to initialize. It will also display a 'helpful' tip on startup, which you can safely dismiss without further attention.

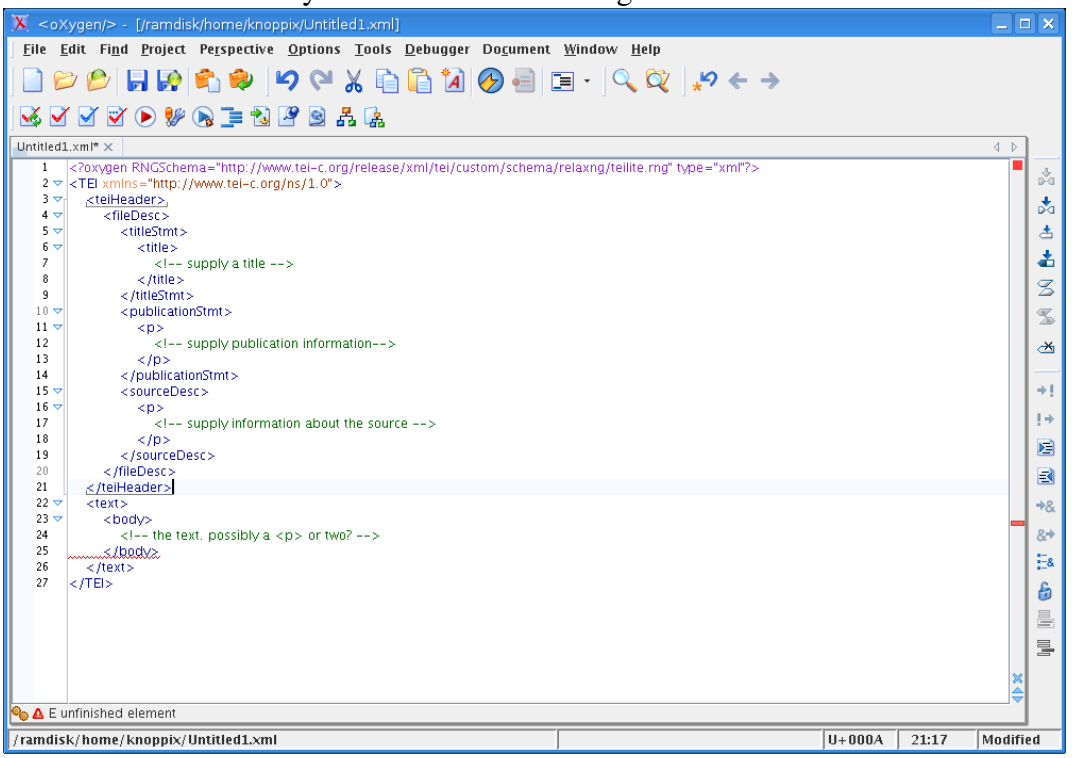
3 Building an XML document

In this exercise, we'll encode various parts of the Punch page using as schema a fairly small subset of the TEI called TEI Lite (the full documentation for TEI Lite is included in your pack).

oXygen comes with predefined template files which can be used to help you build files conforming to this schema.

- Select New from Template from the File menu, or press the corresponding button (it has an orange briefcase on it). Choose TEI P5 (experimental) Lite from the sub menu that opens.

- oXygen replaces the current window with one containing the bare minimum for a TEI conformant document: you should see something like this:



- This shows the structure of a typical TEI document, with a header at the top, and the body of the document below. XML comments such as 'supply a title' are in a different colour from the tags, attributes, and text. Note that the tags are just as easily edited as the rest of the text, as we will see in a moment.

- The blinking cursor marks the insertion point, and indicates where keyboard input will be inserted if you start typing. At present, it is at the end of the file. Move it around (using the mouse or the arrow keys), and see how oXygen highlights the element around the insertion point.

- The red wavy line at the bottom of the document shows you that it is currently invalid. At bottom left, you see an error message explaining why it is invalid: "E unfinished element". Before fixing that, we suggest you familiarize yourself a bit more with the oXygen editing environment.
- Use the mouse or arrow keys to move the insertion point (the cursor position) anywhere in the document, and type or delete a couple of characters. What happens if you do this inside a tag?
- If you make the document invalid, by changing one of the tags, another red wavy line will appear to mark the offending changes. oXygen is configured to constantly check that your document is valid, and warn you when it is not. The red wavy line moves to the first invalid point in the document.
- If you introduced a new error, you can undo the last change you made by selecting Undo from the Edit menu, by typing ctrl-Z, or by pressing the Undo button on the toolbar, in the usual way. You can also correct the file by retyping or deleting characters in the usual way: most — but not all — of the keys behave in the way you'd expect. Or you can simply close the file (choose Close from the File menu) and re-open it.
- Let's supply some missing parts of this TEI Header, following the suggestions given as XML comments <!– like this –>. We suggest you give your text a title such as 'A page from Punch' and an author; supply a publication statement such as 'Unpublished exercise'; and as source description you could specify something like 'oXygen exercise for OUCS courses: February 2006'

We will now start editing the Punch page in earnest. To make life a bit easier, we've done part of the job for you — but not very well.

- Make sure that the insertion point (cursor) is inside the `<body>` element. As before, you should see only one red line on the screen.
- Select the Insert File command from the submenu which opens when you select File on the Document menu. Navigate to the file samples/verse.xml and press OPEN.

If you did this right, the error message on the status bar disappears to show that your document is now valid. Congratulations! However, validity is not the same as truth... As you will see, if you scroll down a little,we haven't actually done a very good job of tagging this poem: the last 'line' actually contains several lines and stanzas run together. In this part of the exercise we'll try to improve on the tagging.

- Using the mouse or the arrow keys, put the cursor after 'upper lid;"' in the third stanza. Type the two characters </. Observe what happens.
- Move the cursor to the start of the next line (before 'So I blew') and quickly type `<l>`. Observe what happens.
- You may notice that oXygen is trying to guess what you want to do, and is suggesting what the element you need to insert is. Now you need to repeat these two steps to add a `</l>`-tag at the end of this line, and a `<l>` tag at the start of the next one.
- When you have put the `</l>` tag at the end of the refrain 'And I still had a fly in my eye', type </ again. Can you explain what happens?

- The sequence `</l>` closes the current element (the line). If you type `</l>` again, you close the next element up the hierarchy (the line-group). If you type it again, what will happen? Check your understanding by trying! Use the Undo command, or type ctrl-z, to repair the document before proceeding. What tags must you insert at the start of the line 'And then Sir...'?

As you have probably noticed, oXygen knows about all the elements in the schema you are using: both where they can appear and what they are to be used for. We will use this facility in the next part of the exercise.

4 Making life easier

Computers are supposed to simplify boring repetitive jobs like typing in tags. Fortunately oXygen has a few nice features to reduce the amount of typing needed for repetitive tasks like this.

- Make sure your text is still valid. If it isn't, make it so!
- Put the cursor at the end of the next verse line. ('And then Sir...').
- Click the Split Element button on the toolbar (fifth one down on the toolbar to the right of the window), type ctrl-shift-/ or select Split Element from the XML Refactoring command on the Document menu
- Repeat for each line. Don't forget you will need to split the stanza too!
- To tidy up the appearance of your markup, you can click on the Format and Indent button on the lower horizontal toolbar (or type ctrl-shift-p or select Format from the XML Document command on the Document menu) if you like

5 Adding more tags

Some of the verse lines in your example actually take up more than one typographic line. The tag `<lb/>` should be inserted at the point where a linebreak occurs, if you want to mark this fact. The poem also contains several emphasised words (they are in italic, in the original printed version), which you might want to tag using the `<emph>` tag.

First we will try oXygen's tag completion feature.

- put the cursor in front of the emphatic 'I've' in the refrain of the second stanza and type `<`
- a menu of possible elements you can insert at this point pops up, together with a brief gloss explaining what they mean
- scroll down the list to find the one you want (`<emph>` in this case) and press Return

This inserts both start and end tag for the new element in the document. This would be appropriate if we were typing in new text, but less useful in the present situation, where the text is already there and we want to add tags *around* it. For this, we will try oXygen's Surround feature

- Press `ctrl-z` or select Undo from the Edit menu to remove the `<emph>` element you just added

- highlight the word I've with the mouse
- Click the Surround with Tag button (second one down on the right hand toolbar) or press `ctrl-e`, or select Surround with Tag from the XML Refactoring menu
- A popup shows you the elements which are valid for the string you have highlighted. Select the one you want (`<emph>`) and press Return.
- Now move the cursor around and experiment to see what elements can be added at different points in your document.
- If you want to add an attribute to an element, type a space inside its start tag: a menu will appear showing the available attributes, as before. Select the one you want and press Return

6 Adding character entities

The poem contains a number of dashes, which have been represented as double hyphens. The dash is a normal Unicode character, but like accented letters and other special punctuation marks is difficult to enter from the keyboard. The most portable way of inserting it is by means of a numeric character entity reference.

- Put the cursor in front of the two hyphens in the first stanza, used to represent a dash. Type `&#x2014;` — this is the Unicode number for the mdash. Delete the two hyphens.
- oXygen has a find-and-replace facility like any other editor. Select Find/Replace from the Find menu (or type ctrl-f) to open it. Experiment with its options!
- Talking of semicolons, you may have noticed that this document has redundant whitespace in front of them. Why not use the find and replace tool to tidy them up?

Alternatively, a suitable keyboard utility such as KCharSelect can be used to enter any Unicode character directly into the file.

- Open KCharSelect by selecting it from the Knoppix menu (Utilities -> More Applications -> KCharSelect)
- Select an appropriate font (Lucida Bright is a good choice)
- The display shows you how each of the possible Unicode characters is rendered in your selected font, 256 characters at a time. To select the next 256 characters, increment the number in the Table box. If you know the approximate character number, enter its hex value in the Unicode code point box. Press Return to refresh the display.
- Hover the mouse over any character to see what it looks like, and its character number. Click on the character and it is added to the row at the bottom of the display.
- Click on the To Clipboard button and the content of the bottom row is added to the Clipboard. Return to oXygen and select Paste from the Edit menu (or press ctrl-V) to copy it from the Clipboard into your document.

7 More things to try

Using the tricks you've learned so far to find out what tags are available, you should be able to build up a complete TEI document representing the whole of the Punch page. You will find versions of the other parts of the Punch page called cartoon.xml, play.xml and paras.xml in the samples directory.

You can add some metadata to your TEI header if you like.

Good luck!

8 And finally...

Want to see what your document looks like without all those annoying tags?

- First make sure your document is valid! Press the red tick (Validate Document) button... If the message `Document is valid` appears at the bottom of the screen, you're OK. If not, you need to fix the errors before going on. Type ctrl-dot (or select Next error from the Validate As You Type command on the Document menu) to move to the next invalid spot, if there is one.

- If your document is valid, you may want to prevent any further changes to its tagging. You can do this by clicking the Lock/Unlock XML tags button on the right hand toolbar (it looks like a padlock, and changes when you click on it) or select this command from the Source command on the Document menu. Click the button again if you change your mind.

- To visualise your document, you can simply transform it into HTML and let your web browser show it to you. Click the Apply Transformation button (a red triangle), press ctrl-shift-t, or select Apply transformation scenario from the XML Document command on the Document menu.

- You should see the message `Transformation successful`. Click on the Firefox button in the main Knoppix toolbar, if necessary, to see the result.

You'll find out more about how transformations of XML are done tomorrow. For now, don't forget to save your work when you have finished! Use the Save command from the File menu, or the Save As command if you want to save it under a different name. If you're working with the Knoppix CD, remember also to copy the file from the Ramdisk to some external medium such as a floppy disk or USB key, or to email it to yourself, since everything on the Ramdisk is lost when the machine closes down.