Creating an ELTeC level zero document

Lou Burnard Consulting *2018-04-14*

This second tutorial walks you through the process of creating an ELTeC document. In the first part, you will create the first page by transcribing it from a page image, and then add the rest of the first chapter from some plain text OCR output. In the second part you will work on a version of the same text which has been saved in Word DOCX format. In the third part, you will combine the two.

1 Creating a document

- Open oXygen and click on the New icon (first left on the toolbar); or select File->New from the menu; or type CTRL-N
- The New File dialogue appears. Select XML Document from the New Document menu, and enter an output filename at the bottom (I suggest your name). Now click the Customize button at the bottom of the dialog in order to specify the kind of XML document you want to create.
- If this is the first time you've done this, oXygen needs to be told where to find the definition for the XML customization we want to use. This is provided by a file called a *schema*, and this dialogue would like you to specify where to find it in the box labelled URL of schema. The file you need is called eltec-0.rnc and you can get a local copy from your Work directory (it's in the folder Schema); alternatively you can find the same file online at http://distantreading.github.com/Schema/eltec-0.rnc. Use the arrow to the right of the window to navigate to the schema file and open it. Then press the Create button.
- The main oXygen editing window opens, with the beginnings of an ELTeC XML file in it. Note the following:
  - At the top of the file there are two purple lines beginning `<?xml`.
  - The file contains tags (in blue) and attribute names (in orange) but no text as yet.
  - To the right of the main window there is a status bar with an angry red square at the top and some red error flags. These correspond with parts of the text display underlined in red.
  - Underneath the editing window there is an error message corresponding with the first error flag: `element "sourceDesc" incomplete; expected element "bibl" or "p"`
- The red flags tell us that although oXygen has done its best some additional information is needed to create a document which is valid according to the rules specified in the schema we have named. Let's fix up the errors!
  - We need to add something inside the element sourceDesc. Put the cursor between the start and end tags for this element, i.e. after `<sourceDesc>` and before `</sourceDesc>`. Now type a <
  - Scan the menu of possibilities which opens: you can insert an element called bibl or p, a comment, or a few other oddities. Choose p and press RETURN.
  - The first red error flag disappears, and the error message at the bottom of the screen changes to `value of attribute xml:id is invalid`. This is because our schema requires an identifier for every text and we have not yet supplied one. Put

the cursor between the quote marks and type in the identifier for our sample novel which is (shall we say) `EN042`. The second error message disappears! To celebrate, add a value (`en`) for the attribute *@xml:lang*: this specifies the language code for the document.

– The third error is like the first: `element "body" incomplete...` so fix it in the same way. Put the cursor inside the `<body>` element and type a < to see what's legal here. We suggest you add a `<div type="chapter">`. Phew! no red marks anywhere! Your document is valid: click the disk icon (or type CTRL-S, or choose File-Save from the menu) to save it for further work.

Valid as it is, this document is not much use to anyone without some content. We need to start adding text, both in the header (where the metadata describing it will be held) and in the body (where the transcribed text itself will reside). For this exercise we will work on just the first chapter of a deservedly forgotten English novel: Mrs Grey's *Passages in the life of a fast young lady* (1862).

In your folder Work/Pages you will find page images of the chapter concerned. Take a look at the titlepage which is in file Work/Pages/006.png and browse a few of the other pages to get a taste for the kind of document you are dealing with.

oXygen is just like any other editor: you can type in text, correct, cut and paste, and so on. Let's begin by typing some minimal metadata into the TEI Header. (We'll return to this topic later)

- First enter the title of the digital text we are creating. For ELTeC, this should be the original main title (Passages in the life of a fast young lady), followed by the phrase `: ELTeC edition`.
- The `<author>` element should contain the author's name in a standardised format, with the surname first, and the author's dates in parentheses at the end: like this `Grey, Catherine Maria (1798-1870).`
- You should also take credit for your encoding by adding a `<respStmt>` element after the `<author>`. This element should contain a `<resp>` with the content `ELTeC encoding` and a `<name>` containing your name.
- We will look more closely at the other header elements later. For the moment, we suggest you just add `<p>` elements containing a few words for the publication statement and the source description (`<sourceDesc>`). If you feel ambitious, you could also add a `<revisionDesc>` containing a `<change>` element at the end.

All being well, your header should look like something like this when you've finished:

```
<teiHeader>
    <fileDesc>
        <titleStmt>
            <title>Passages in the life of a fast young lady : ELTeC edition</title>
            <author>Grey, Catherine Maria (1789-1870)</author>
            <respStmt><resp>ELTeC encoding</resp><name>Lou</name></respStmt>
        </titleStmt>
        <publicationStmt>
            <p>Unpublished exercise</p>
        </publicationStmt>
        <sourceDesc><p>OCR of Google scan of Bodleian copy</p>
        </sourceDesc>
    </fileDesc>
    <revisionDesc><change when="2018-03-28">Header added</change></revisionDesc>
</teiHeader>
```

Now we'll proceed to transcribe and tag the text itself.

First, open the page image for the beginning of the chapter in file Work/Pages/0008.png. You can do this directly within oXygen, or by clicking on it in the File Explorer. Look at the page,

and you will see that it contains a centered heading for the whole text (the title *Passages ... Lady*), followed by a heading for the first chapter (*Chapter I.*). There is then a paragraph of text *A yellow fog... on the heart*, and the first line of a second paragraph, *It was on a morning when our metropolis*. Finally there is a page footer containing the text *Vol I* and a signature *B.*. The words *yellow fog* are in small caps, but otherwise there is nothing particularly remarkable in the typography.

For ELTeC purposes, we do not care about line breaks or other purely typographic features, such as the centralised line between the two headings, or the page footer. But we need to show that there are two headings and place them correctly in the document structure. We don't need to capture line breaks explicitly, but we need to show the boundaries of the page, and of each paragraph.

- Insert a `<pb>` element as the first thing inside the `<body>` element. Give its *@n* attribute the value `1`.
- Insert a `<head>` element after the `<pb>`, but before the `<div>` you inserted to make the file valid. Then type the title of the whole work as content of this `<head>` element. No need to worry about the linebreaks or font changes.
- Now insert a second `<head>` element *inside* the `<div>` and type in its content 'Chapter I'.
- Finally insert a `<p>` element and start typing the first paragraph into it. As before, no need to worry about typographic variation or linebreaks, but you should respect the spelling and capitalisation of the original. If you make a mistake, use CTRL-Z to undo your last action.

When you've finished your first page, it should look something like this:

```
<text xml:id="EN042" xml:lang="en">
    <body>
        <pb n="3"/>
        <head>Passages in the life of a fast young lady</head>
        <div type="chapter">
            <head>Chapter I.</head>
            <p>A yellow fog in London! We all know from experience
                ... lie most heavily on the heart.</p>
            <p>It was on a morning when our metropolis</p>
        </div>
    </body>
</text>
```

Are you ready for the next page? Open the appropriate graphic file in Work/Pages/0008.png and take a look. Note the following:

- there is a running page header which repeats the title, as well as giving the page number; your transcription should include the latter (in a `<pb>`), but not the former.
- The text contains some long dashes — e.g. between 'garb' and 'that ' in the first line — which should be distinguished from hyphens: use the Insert Special Character command on the Edit menu to open a menu from which you can select the correct Unicode character.
- Some hyphens are simply typographic effects which can be ignored or discarded (for example the one at the end of the third line) ; others however should be retained. For example, starting on the 7th line but continuing to the 8th is a sequence which looks like 'orange-co-loured', but which should probably be transcribed as 'orange-coloured'.

If you're happy typing, please continue! Take care that you insert the second `<pb>` and the whole of the text of this second page *within* the paragraph you began at the foot of the first page... and it continues on to the third page.

If on the other hand you'd rather not have to do quite so much typing, there is a (more or less) tidied transcription of the next few pages of this chapter in the file EN042/pp2-8.txt. We'll use that in the next few steps.

- With the cursor inside the last on page one, i.e. immediately after the word 'metropolis', Select Document -> File -> Insert File from the menu. Navigate to the file pp2-8.txt in your Work folder and click Open. The content of this file is inserted into your document, ready for you to edit it.

The text you've just added needs at least two things done to it. Firstly we need to change the page numbers (which appear as single digits on a line of their own) into <pb> elements, like the one you typed in earlier. Secondly we need to divide the text up into paragraphs with proper tags rather than blank lines. Now, of course, we could do this by hand for a few pages, but that's not really a scalable solution. And a computer is not just a glorified typewriter! Let's use some of the special power of the digital to make our editing task easier.

oXygen includes a powerful Search and Replace facility, which you can use to tackle the first problem.

- With the cursor still immediately after the word 'metropolis', select Find -> Find/Replace, or type CTRL-F.
- In the dialog box that opens, first check that the option Regular expression is selected.
- Type the following incantation into the Find box: `\n(\d)[\s\n]+`. This is a regular expression: it means 'find a newline, followed by a single digit, followed by a sequence of one or more spaces and newlines'.
- Type the following incantation into the Replace with box: `<pb n="\1"/>\n`. This is what will be used to replace the part of the input text matched by the regular expression. The `\1` part refers to that part of the regexp which was parenthesized, i.e. the page number.
- Double check you typed the incantations correctly, take a deep breath, and press the Replace All button. Did it work? If not, press CTRL-Z, and try again!

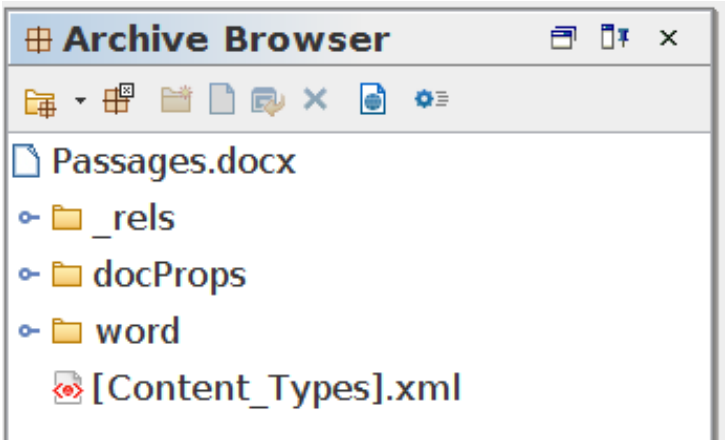
The same technique could be used to solve the other problem: you can probably work out how to define a regular expression for two consecutive newlines and replace it by the sequence . But instead I suggest you use a slightly slower but more accurate technique, relying on one of oXygen's built in XML manipulation facilities. As you can see if you look at the Outline view, or switch the display to Author mode, oXygen currently thinks that the whole of the text you just inserted is one big paragraph. Do you remember how to split that up?

- In Text Mode, scroll the text to the end of that interminable paragraph on page 2, i.e. between the words 'occupants of the cab.' and before 'Along dirty streets'.
- With the cursor at that point between the two paragraphs, press ALT-SHIFT-D, or select Document -> Markup -> Split Element from the menu.
- oXygen closes the current paragraph and opens a new one. Scroll down to the end of this new paragraph (after 'and that quickly."' at the top of page 4), and repeat the manipulation.
- Proceed in this way to the end of the file. Do you still have a jolly green square? You now have a valid ELTeC document, containing the first chapter of the novel. Well done! Save it in a safe place on your Desktop, and we will try to complete it in the next part of the exercise.

2 Working with Abbyy output

The text file you used in the preceding exercise was extracted from the file called Abbyy-output.txt in your Work directory. As the name suggests, this file was created by Abbyy Finereader in text format. As you saw yesterday, this OCR software can produce output in several other formats as well, including `docx`, the format used by Microsoft Word. Maybe you'd prefer to work on correcting OCR errors in this format, but will oXygen be able to add TEI tags to a Word file when you've finished? Of course it can!

- Open the file Abbyy-output.docx with oXygen.
- A docx file is actually an archive containing several files, so oXygen opens the *archive browser* window for you to select which one you want to edit.



- In the archive browser, click the blue circle to the left of the folder labelled word to open it up: it contains a file called document.xml which contains the text we want. Double-click to open it.
- This is an XML file we can edit it in oXygen, even though its schema is a bit far removed from TEI. Scroll down to reassure yourself that there is actually some useful content here.
- oXygen comes with some built in transformation tools, including one for converting this flavour of XML to TEI. With the document.xml file open in your main editing window, select Document -> Transformation -> Configure Transformation Scenario(s) from the Document menu. Or type `CTRL-SHIFT-C`. Or click the little spanner icon.
- The Configure Transformation Scenario appears, suggesting some relevant transformations for this file. Check the little box next to DOCX TEI P5 and press the Apply Associated button.
- The Word document in the main window is replaced by a valid TEI document. You may want to use the Format and Indent button (CTRL+SHIFT+P) button to make it a bit easier to see what's going on. You can also close the Archive Browser window now. Scroll through your new file to get a quick impression of the tags that have been introduced.

The text has been split up into paragraphs, each tagged `<p>`, and each carrying an attribute *@rend* with a coded value indicating something about the format of that paragraph in the original Word document, for example `Style 44`, `Style 50` etc. It looks as if `Style 50` paragraphs are text, `Style 44` paragraphs are chapter headings, and `Style 25` paragraphs are page footers, but there may be exceptions or inconsistencies.

```
<p rend="Style 50">Mrs. Lewis could not but feel gratified by this proof of the entire
    confidence this strange girl placed in her, and at the same time concurred in the
    prudence of the plan.</p>
<p rend="Style 44"><anchor xml:id="bookmark12"/>CHAPTER VII.</p>
<p rend="Style 50"><hi rend="Char_Style_52" xml:space="preserve">It </hi>was with sorrow
    and shrinking dread, that Linda saw one by one every excuse for protracting her stay
    hourly vanishing. She had really in her childish heart enjoyed the time she had
    spent in that humble abode, so kindly had she been treated by the widow and her
    little son.</p>
<p rend="Style 50">The girl's high spirit had been softened and subdued by the truly
    maternal and tender bearing of her lowly friend. She had never experienced the love
    of a mother; ties of blood relationship had in but a solitary instance exercised
    their uniting bond of affection upon her nature. Kindness</p>
<p rend="Style 25">G <hi rend="Char_Style_86"><seg rend="bold">2</seg></hi></p>
<p rend="Style 50">and consideration she had received, but not the peculiar motherly
    tenderness of a heart like Mrs. Lewis's, full as it was with the very milk of human
```

We could use this information to improve the tagging, if it is reasonably consistent. Let's start by investigating the paragraphs tagged as `Style 44`: are they in fact all chapter headings? And are all chapter headings tagged with this code?

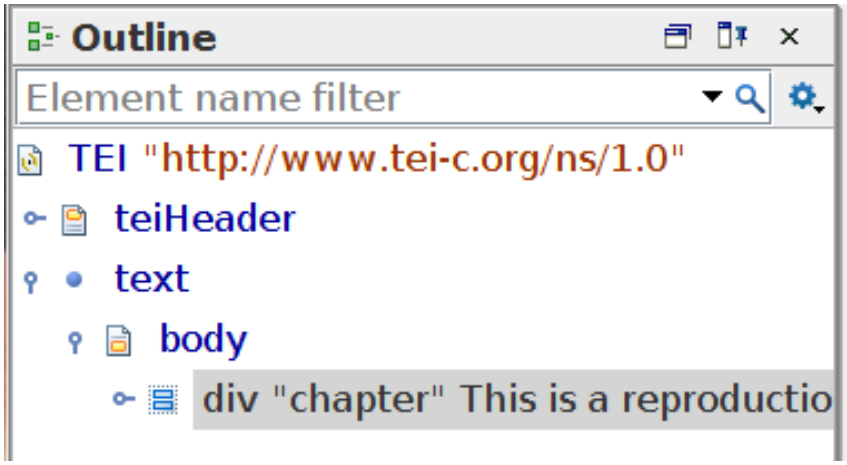
- At top left of the oXygen main screen, under the button bar, there is a box labelled XPath 2.0 which you can use to navigate the XML structure of your document. XPath is a standard language for describing and locating subtrees within an XML document.
- Type the following code into the box `//p[@rend="Style 44"]` and press RETURN. (The effect of this code is to locate any `<p>` element which has a *@rend* attribute with the value `Style 44`).
- A window opens at the foot of the screen, with the heading `Description - 21 items`. In the window there are 21 rows, each indicating where a matching `<p>` element is to be found in the document (the XPath location) and also displaying its content (e.g. `CHAPTER I.`).
- Look closely at the list and you will see that there are some chapters missing; nevertheless this is still a useful way of finding quickly where we should divide the text up into into chapters.

To divide the text up into chapters, we suggest you use the same method as before. (You will find it easier to see what you are doing if you open the Outline view first).

- First surround the whole text with one `<div>` as follows. Scroll up to the top of the document, and insert a `<div type="chapter">` tag after the `<body>` start-tag.
- As usual, oXygen will helpfully insert a `</div>` close tag: delete it.
- Scroll down to the end of the document, put the cursor in front of the end-tag for the `<body>`, and type `</` to close the `<div>` you just opened.

In the Outline box, you should now see something like this:
We now have a single `<div>` containing the whole text. We just need to locate the start of each chapter, and press ALT-SHIFT-D to divide it up.

- Click on the line in the window at the bottom of your screen which begins `CHAPTER I`.
- The cursor jumps to the paragraph containing this text. Press ALT-SHIFT-D and check to see what happens in the Outline window.
- Repeat with the next line in the lower window. You can move to it by just pressing the down-arrow key if your cursor is still in that window.
- Repeat to the end of the list of chapter headings. The Outline view should show that you have now divided the body of the text into 21 chapters.



Now that we have chapters, we need to transform the `style 44` paragraphs into headings. You can probably think of some ways of doing this for youself; here's a suggestion, which also introduces you to another feature of oXygen: the ability to *remove* tagging you don't want.

- Scroll to the top of the document, and repeat the XPath search for `//p[@rend='Style 44']`
- Click on the line starting `CHAPTER I.` in the bottom window, as before.
- The paragraph containing this text is highlighted. Carefully type ALT-SHIFT-X once to remove the tags around it. Type ALT-SHIFT-X again to remove the `<anchor>` element.
- Type CTRL-E and select or enter `<head>` to add that tag around the phrase `CHAPTER I.`.
- Repeat for the remaining chapters! If something goes wrong, press CTRL-Z to rewind and try again.

As we already noted, most paragraphs are coded `Style 50`. This is redundant information, especially since we are not interested in the formatting of the text. Use Find and Replace (CTRL-F) to remove the string `rend="Style 50"` (note the leading space). Then do another XPath search, this time for `//p[@rend]`. This will show all remaining styled paragraphs, the majority of which contain things you don't want such as page footers or signatures. Scroll the list to check, and see how many you can remove, or replace with a `<pb>` element. (Unfortunately, Word has not retained the pagination in a way that we can reuse it)

3 The real work starts here

You've now seen a few of the techniques oXygen offers to make the task of editing your XML document easier:

- You can use Validate (CTRL-SHIFT-V) to check that your document conforms to a schema, and thus detect any unexpected tags.
- You can use Find and Replace to search for strings or patterns in the text and replace them with other strings.
- You can type XPath expressions to locate parts of the XML structure of the document.
- You can use shortcuts like ALT-SHIFT-X to remove the tags of the adjacent element, CTRL-E to add tags to a selected string, or ALT-SHIFT-D to divide an XML element in two.

- You can use the Outline view to get an overview of the structure of your document and the Format and Indent button to reformat the editing window. You can also see your document in a styled format by clicking the Author tag.

In addition, of course, oXygen offers the usual range of facilities you'd expect in an editor such as cut and paste. It even has a spellchecker, which is also very useful for detecting OCR errors. (Try it out by selecting Edit->Check Spelling on the menu, or press F7). And oXygen allows you to operate on several files at the same time, which is particularly useful, as we shall see in the last part of this tutorial.

- Select File->Open, or CTRL-O and navigate to the Desktop or wherever you saved your ELTeC version of chapter 1. Click Open.
- A new editing window opens, containing the document you prepared earlier. Scroll down to the end of the file and put the cursor after the `</div>` tag at the end of the chapter.
- Click the tab above the edit window containing the name of the file you produced from word (it should be called Abbyy-output-TEI-P5.xml) to return to that file.
- Go to the Outline window and click on the div tag containing chapter 2.
- The corresponding text is highlighted in the edit window; type CTRL-C to copy it to the clipboard.
- Click on the tab corresponding with the edit window containing your ELTeC version of chapter 1.
- Click CTRL-V to paste the second chapter into your file.
- Click the Validate button (or type CTRL-SHIFT-V) to check that you have not introduced any validation errors : correct them, if you have!
- Continue in this way, pasting in chapters one at a time. When you get bored, try running the spell checker or reading the text for other errors...