Using XPath in oXygen

James Cummings *14 September 2014*

1 Introduction

In this exercise we will be using a subset sample of data from the Protestant Cemetery in Rome collected by Sebastian Rahtz. This is made into a browsable interface at: `http://tei.it.ox.ac.uk/pc/` and the full source is available from `http://github.com/sebastianrahtz/ProtestantCemetery`. You may also be interested in using the documentation for the XPath Tutorial at W3Schools as a reference:

- `http://www.w3schools.com/xpath`
- And the functions list at `http://www.w3schools.com/xpath/xpath_functions.asp`.

In our case we will be using `Work/PC/cem-sample.xml` as our datafile.

2 Starting up

- Start the oXygen XML Editor, and load up the file `Work/PC/cem-sample.xml` file.
- Have a look through the file. There may be sections that you don't understand but note the following items:
  - The file starts with a `<teiCorpus>` element.
  - The long `<teiHeader>` has several sections: `<fileDesc>` with metadata about this file, `<encodingDesc>` with some detailed hierarchical taxonomies, `<profileDesc>` with language identifications, and a `<revisionDesc>` element.
- After the `<teiHeader>` note that there is a `<facsimile>` element with information about general images of the stones.
- The rest of the file is filled with individual `<TEI>` elements each with their own `<teiHeader>` and `<facsimile>` before the `<text>` element which holds a transcription of the tombstone.
- Look through a couple `<TEI>` elements to see the kind of information that is given here rather than in the overal corpus header.

3 Using the XPath Search in the toolbar

- Assuming you haven't changed your toolbar setup, you should have and 'XPath 2.0' toolbar in the upper-left of the oXygen editor. (If you have got rid of this, ask and we'll show you how to restore it.)
- This toolbar should look like this:

It contains:

- a dropdown menu on the left for selecting the version of XPath you are using (for now use 'XPath 2.0')
- a box in which to type xpath queries for the current file

- a setting dropdown menu enabling you to change options (these should be unchecked by default)

- Let's say you want to find all of the entries we have in the sample file for stones. You know each one of these is in its own `<TEI>` element immediately as a child of the root `<teiCorpus>` element. So we should be able to find these with: `/teiCorpus/TEI`. (You may have to press enter twice, as it will be prompting you to select 'TEI' as one of the elements it knows exists there.)

- When you do a search in oXygen it lists the hits in a window at the bottom of the editor. This contains a description, the XPath location, what file it is in, and location. Click on one of the results and you will be taken to that result. This can be an easy way to navigate through a large file. Notice what it lists as the 'Description' and that next to the heading it also lists how many results there are.

- We could also get this using the `count(/teiCorpus/TEI)` function in the XPath box, but then wouldn't have access to each one. Try it and see!

- The `<teiHeaders>` for the individual stones have a `<profileDesc>` containing a `<particDesc>` containing a `<listPerson>` with one or more `<person>` elements in it such as:

```
<profileDesc>
 <particDesc>
  <listPerson>
   <person>
    <sex>2</sex>
    <persName>
     <forename>Anna</forename>
     <surname>Boynton</surname>
    </persName>
    <death when="1874"/>
    <nationality key="US"/>
   </person>
  </listPerson>
 </particDesc>
</profileDesc>
```

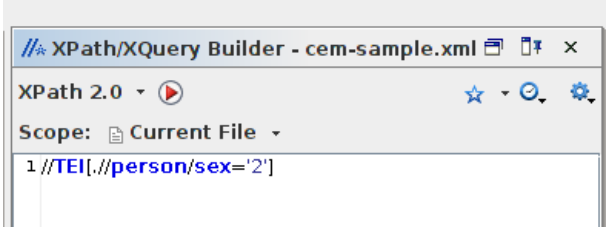
- **Find all person elements who are recorded as being women.** (Here women are denoted in Sebastian Rahtz's corpus of gravestones by having a `<sex>` element of '2'.) To do this we must look at that element in our XPath: `//person[sex='2']`. We can read this as 'give us and `<person>` element at any level in the file, which has a `<sex>` element inside it whose content is '2'. How many are there?

- **Now find all the stones which have women mentioned on them.** The difference is in whether you are finding `<person>` elements or `<TEI>` elements that have a female `<person>` listed on them. How many are there?

4 Using the XPath Builder

- If you type too long of an XPath in the XPath search box it will suggest you use the XPath Builder and offer to open it for you. Otherwise you can open it by going to `Window -> Show View -> XPath/XQuery Builder`. This will open up a window on the right-hand side of oXygen which enables you to type longer XPaths. (You can also run XQueries in them as well, but that is a slightly more complicated XPath-based query language for XML Databases that we won't teach you here!)

- The XPath/XQuery Builder looks like:



- On the left-hand side is a drop-down menu for choosing the version of XPath, and next to this a 'play' button to run your search. Beneath this you can control the scope of the XPath search to be one of:
    - Current File (Default and what you should be using
    - Project
    - Selected project resources
    - All opened files
    - Current DITA Map hierarchy
    - Working sets
    - And you can set some options
- To the right-hand side there is a star to allow you to 'favourite' queries, along with a drop-down list of those you have favourited. There is another drop-down list of recent queries. Finally, there is a drop-down settings menu.
- Try the queries you have made already in the XPath Builder and mark at least one of them as a 'favourite'.
- **Try to find all British Women.** To do this you need to use both the *@key* attribute of the `<nationality>` element (equal to 'GB') and the `<sex>` element. Sets of square bracket filters concatenated one after another create an 'and' condition. The XPath will be filtered by all of them.
- The 'contains()' function is a very useful one for finding nodes whose text contains some other bit of text. It is used by saying something like `//div[ab[contains(., 'DAUGHTER')]]`. This XPath looks complicated but let's disect it: We're asking for any `<div>`, but only where it has an `<ab>` element which itself contains there word 'DAUGHTER' (in upper-case as the stones are written and transcribed). The 'contains()' function takes two strings, it checks to see if the second string is present in the first. Because it works on individual strings of text, it can't take nodesets (and this is why we didn't write this as `//div[contains(ab, 'DAUGHTER')]` because that would provide multiple values to the contains() function.
- **Find and person elements where the `<forename>` element contains 'Elizabeth'.** How many are there in this sample? (You may notice two are repeated with slightly different forms of forename.)
- **Return a list of distinct-values of the `<condition>` element.** To do this you'll need to use the distinct-values() function which takes an XPath as its single parameter. Once you have done this, you can sort this by clicking on the 'Description' heading in the results box.

5 Using XPath with Find and Replace

- You can also use XPath to narrow the search context in the `Find -> Find/Replace` window.
- If you search for the word HEAD (type all in upper-case) and 'Find All', you will find many hits.
- You could narrow this in several ways. You should check the 'Enable XML search options' box to expose a sub-menu of places to look. In this case check 'Attribute values' to limit this to attribute values. (You could also check 'Case sensitive'.)

- You'll notice that one of the results is a `<category>` and the rest are `<objectDesc>` elements. If you limit the context to `//TEI` by using the XPath Box on this window and 'Find All' you'll see the `<category>` will disappear.
- Notice that if you close the 'Find/Replace' window and load it up again (control-f on windows, probably command-f on mac), that the options you had selected before are still selected. Always check this when doing a Find or Replace.
- Experiment finding more bits of the document through increasingly more complicated XPaths! Use the function list at: `http://www.w3schools.com/xpath/xpath_functions.asp`. For example, how would you use substring-before() to get the year from dates? How would you get the month number?