The ISE, TEI and XML Encoding

A workshop and consultation

In this workshop, we're going to give a brief introduction to XML and TEI (primer for some, refresher for others). We're then going to take you through the process of encoding a fragment of Henry VIII, using the Oxygen XML editor and some schemas and other tools we have been developing for use by ISE editors.

For those new to XML and TEI, we hope to demonstrate that it's not difficult; for those who have done encoding before, we hope to show that with a well-prepared set of project tools and schemas, encoding can be faster, more accurate and more rewarding than you might be used to. For the ISE project, we would like to get feedback from all of you on how this nascent toolset works, and how it can be improved.

Encoding in TEI with Oxygen: downloading and opening the Oxygen project

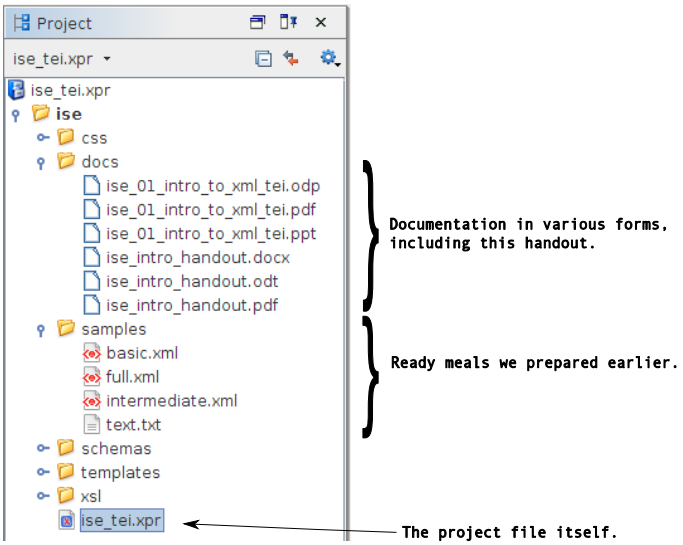
In your browser, go to this URL:

http://hcmc.uvic.ca/workshops/ise/

You will see a small set of materials. Download the file **ise_tei.zip**, and unzip it somewhere convenient (such as your desktop). You should see a folder called **ise**.

Start the Oxygen editor, and click on Project / Open Project.
Browse to the **ise** folder, open it, and choose the **ise_xpr** file.

You should see something like this:

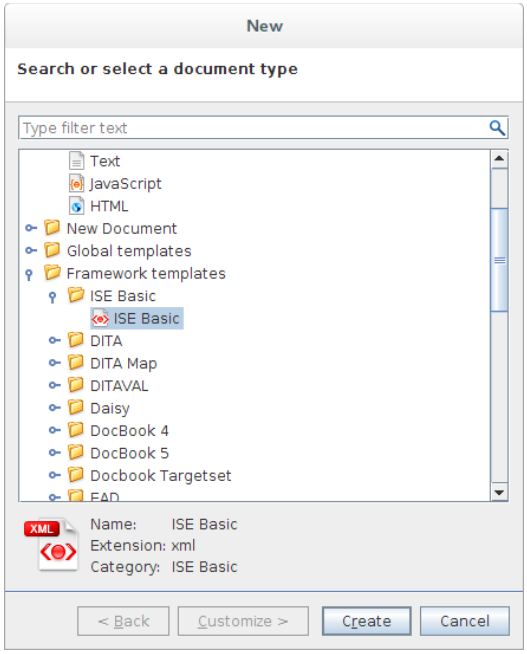


Starting a new ISE XML file

First, look at the project tree on the left of Oxygen and find the **samples/text.txt** file. Double-click it to open it. You should see a plain-text transcription of the first part of Henry VIII (the prologue and part of Act 1, Scene 1). This is going to be the base text we use for encoding.

In Oxygen, click on **File** / **New**, and scroll down to where you see **Framework templates**. Choose **ISE Basic**.

This will create a very rudimentary XML file. Most of this file consists of the *TEI Header* (<teiHeader>...</teiHeader>), which we're not going to worry about right now; that's metadata. You might add the title of the play in the <title> element at the top, and the date in the <change> element, but that's not important for now. We'll be focused on the <text> element in this workshop.



Encoding Phase 1: What things are

At the top of your TEI file, you should see this:

<TEI xmlns="http://www.tei-c.org/ns/1.0" rend="ise:basic">

In our Oxygen project, rend="ise:basic" causes our file to be validated using a very basic TEI schema, which mainly consists of elements and attributes concerned with *what things are* in the text (as opposed to what they look like, or what they relate to). In this part of the workshop, we're going to work on copy/pasting bits of text from the text file into the appropriate places in the TEI file, and tagging them appropriately to identify what they are (titles, headings, lines, speeches, speakers and stage directions). We'll be using these tags:

<titlePart> <head> <l> <sp> <speaker> <stage>

We'd like to make sure everyone completes a few lines of the prologue, and also a couple of speeches from Act 1 Scene 1. Don't forget to **validate your file** frequently as you work.

However, when we get to the end of this phase, you can serve yourself a ready meal we prepared earlier, by opening the **samples/basic.xml** file.

Encoding Phase 2: What things look like

Open the **samples/basic.xml** file and look at the top (the root <TEI> element). Change the @rend attribute to this:

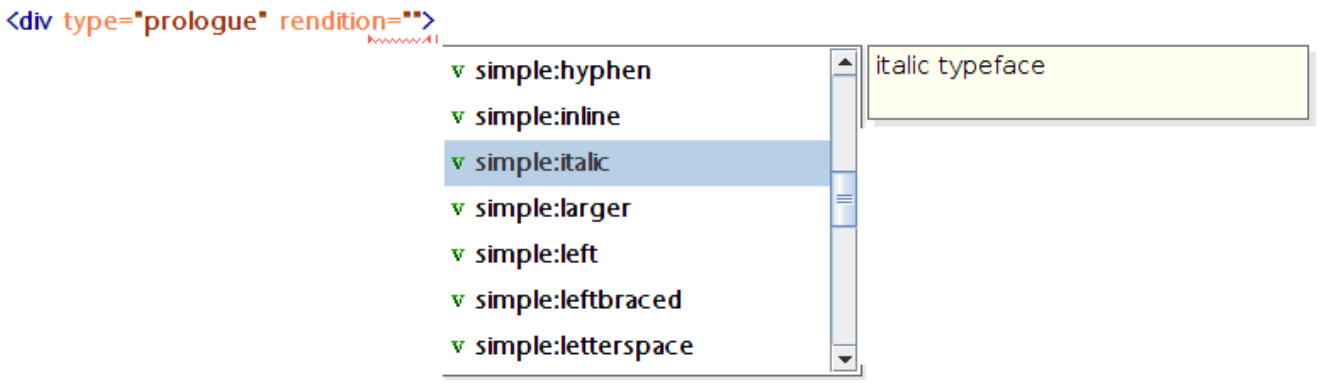
rend="ise:intermediate"

You won't see any visible change, but the effect of this is to switch the file from validating with the basic schema to validating with a larger schema that includes more stuff. That new stuff enables us to start encoding information about how the text appears.

The first thing we're going to do is to encode a few line breaks. We do this with the <lb/> element. (The slash means this is an empty element; it's equivalent to <lb></lb>.)

There are line breaks in the document title, as well as in many of the stage directions. Add them at the appropriate places.

Next, we're going to handle some italics. The prologue is entirely in italics, so we're going to encode that, using the @rendition attribute. Find the appropriate div, and add the new attribute. When you type the attribute in, you should see a drop-down list of all the values you can use:



The schema includes a set of values which are supported by the project. In this case, the values come from the TEI Simple initiative. Each value is accompanied by a brief description to help you choose appropriately. Select **simple:italic**.

Now we're going to use another component of the Oxygen interface which can be very helpful, especially for people new to encoding. At the bottom of the window with your XML file in it, you'll see three buttons, **Text**, **Grid**, and **Author**. Click on the **Author** button.

You should see a formatted view of your text which is much more convenient for reading and proofing, especially as more tags get added. You can switch between these two views any time you like, and you can make edits in the Author view as well as in the Text view. The styling is provided by a combination of a default stylesheet we have provided, along with what you have encoded in your text. So, for instance, the stage directions are centred because you encoded them that way, but the speaker names are rendered in italics by default.

One other important thing we're going to handle at this stage is the long s and the ligatures. Our transcription is a normalized diplomatic transcription which does not include these glyphs, but for an "old

spelling" edition it would be important to record them. The document title has one: the "st" in "History" is actually a long-s-t ligature: Hiſtory. Select the "st", then press **Control + e**, and type "g". This will insert the <g> element ("glyph"); this signifies that the original characters were not quite as they are transcribed. Add the attribute ref="ise:longStLigature"

```
Hi<g ref="ise:longStLigature">st</g>ory
```

You'll see there's a long list of available special characters and ligatures built into the schema.

Now if you go to Author view, you'll see that the characters you've tagged as a <g> have a grey background, so you can easily find them.

To finish off this phase, open the **samples/intermediate.xml** file to see all the extra encoding we would do at this level.

You can also now try something new. Click on the Transform (red triangle) button on the toolbar. Oxygen should automatically build you a web page from the file. This has some useful statistics, and also allows you to switch special characters on and off, so you can proof them more effectively.

Encoding Phase 3: What things link to

For the next phase, change the @rend attribute in the TEI element to this:

```
rend="ise:full"
```

That switches us to a third schema, with even more useful stuff in it. In this phase, you would add encoding which links to external resources; for instance, you might tag placenames and link them to a gazetteer or authority record. For this workshop, we're going to add two things: through line numbers (McKerrow-Hinman), and speaker identifiers.

A TLN can be encoded with a special line break element, like this:

```
<lb type="tln" n="1"/>
```

This says "TLN #1 starts here.". These are not the same a regular line breaks; in your source text or edition, actual line breaks may occur in different places from the TLNs. The first TLN appears right before the <head> element in the prologue, and subsequent ones appear right before each of the lines (<l> elements).

Finally, we're going to link each speech to its speaker in a formal way which is more processable than the abbreviated speaker names that occur in the text. On the <sp> element, add a new attribute, with its value selected from the available list:

```
<sp who="ise:h8_buckingham">
```

These values are taken from a central list of roles created by the ISE project.

You can see the ready-meal version of this phase in **samples/full.xml**. Try transforming it to a web page, to see some new statistical features.