Practical Exercise 1: Creating an XML Document with Basic Markup

James Cummings *19 September 2013*

1 Exercise 1: Creating an XML File

1.1 Learning Outcomes

When you successfully complete this exercise you should be able to:

- start a new XML file
- insert a text file into an XML editor
- mark up basic structural features of document using multiple methods
- add elements and attributes to an XML document
- create a well-formed XML document
- format and indent a well-formed XML document

1.2 Summary

This exercise will walk you through creating an XML document in the oXygen editor and introduce a variety of ways to mark this document up. You will first start a new document, then insert some unmarked up text into the editor, and then mark up the structural sections of the document. You will learn how to check that your document is well-formed and then format-and-indent it.

1.3 Starting A New XML File

Let's start a new XML file by following the following steps:

1. Load up the *oXygen XML Editor* if it isn't already loaded (depending on your operating system this may be by using a Menu, or double-clicking the icon on the desktop).
2. If oXygen prompts you for a license key, then please ask and one will be provided. oXygen is a piece of commercial software and though inexpensive they are very kind to give us licenses for training purposes.
3. oXygen may also present you with a 'Welcome' screen with tips and shortcuts. Close this.
4. Once the editor has fully loaded from the `File` menu select `New` and expand the `New Document` section so you can select `XML Document`. Doing this should open up a blank document with an XML Declaration added.
5. An XML Declaration looks like:

```
<?xml version="1.0" encoding="UTF-8"?>
```

The XML declaration in the element tells anything processing your XML file, including the oXygen editor, that this is an XML file. It also conveys which characters the program may expect in attribute @encoding. UTF-8 (Universal Character Set Transformation Format - 8 bit) contains most characters that people tend to need, though UTF-16 is also possible. The XML declaration needs no closing tag as it takes the form of a special processing-instruction that starts and ends with an angle-bracket and a question mark.

1.4 Creating a <text> Element

Let's create some structure for our file using the `<text>` element. This is a generic division or section element.

1. On the line below the XML declaration type: `<text>`
2. Notice what happens when you type the final '>'. oXygen is trying to help you and inserts the closing `</text>` tag. This is because it knows the rules of XML, and knows that if you type an opening `<text>` you are required to have a closing `</text>` sooner or later.
3. We haven't said what type of text this is, so lets categorise it as 'letter' by adding a @type attribute. Move the cursor back until your just after the last letter in the opening tag. Press space, and then type: `type="` and notice what happens when you type the quotation mark. oXygen is again trying to help you by putting the closing quotation mark, because it knows that attribute values must always be quoted.
4. In between the quotation marks type `letter` to categorise our text as being a letter.
5. Move the cursor back until you are directly in between the opening `<text>` tag and the closing `</text>` tag. Press 'enter' a couple times to give yourself some space inside the element.
6. Add a `<body>` element, inside the `<text>` element and a few blank lines inside that. Spaces and new lines are generally considered disposal in XML files.
7. You should now have something that looks like:

```
 <?xml version="1.0" encoding="UTF-8"?>
<text>
 <body>
<!-- some line spaces here -->
 </body>
</text>
```

1.5 Inserting Some Text

As a sample we are going to use a letter Wilfred Owen wrote to his cousin Leslie Gunston for this exercise. But it would waste a lot of time if we asked you to type the whole letter so we've done it for you. The letter talks about a forthcoming address to the Field Club, and contains a partial draft of a poem 'The Wrestlers'. It was written in July 1917 from Craiglockhart War Hospital, Edinburgh, Scotland.

1. Make sure your cursor is in between the opening `<body>` and the closing `</body>` and go to the `Document` menu and select `File` and from there then `Insert File`. **Note: This is from the `Document` menu on the menu bar, *not* the `File` one!**
2. Select the file 'letter.txt' as the file to insert.
3. The start of your document should look like:

```
 <?xml version="1.0" encoding="UTF-8"?>
<text>
 <body> [1]
   Craiglockhart.
   July 1917.
```

```
      Wednesday

      Dear L.
      Thanks for yours of this morning. I hope
      you [have] had my card -of- posted last Monday.
      On Mond. next I lecture the "Field Club" -
      a Nat. Hist. Association, in the lines of our
      old Society - Geological, (you + me) + Botanical
      (New) Do you remember: -my- you old
      Black Molt? Well, the days have
      come when I am [one of the] founders of a real
      [... a lot more text...]

    </body>
  </text>
```

4. Read through the letter and notice the character-based markup that was provided by the transcriber. (Page numbers are provided in square brackets, hyphens immediately surrounding a word mean it is deleted, underscores mean something was underlined, etc.) We won't do much with these in this exercise.

5. Briefly compare this transcription to the images letter1.jpg and letter2.jpg which are at the end of this exercise.

1.6 Encoding the Basic Structure of the Letter

1. At the very top, let's change the '[1]' into a page-break element `<pb/>`. First highlight the '[1]' and then just type in `<pb/>` watching how oXygen tries to help you. Note that `<pb/>` is one of those milestone-like 'empty' elements. We could put an *@n* attribute with a value of '1' here, but it is probably unnecessary (computers can count). There is a '[2]' further down, replace this with the same markup.

2. The letter is divided into 5 sections:

   (a) An opener (including a place and dateline)

   (b) A section of prose

   (c) A section of verse

   (d) Another section of prose

   (e) A closer (including a salutation and signature)

   We're going to start by marking up each of these sections.

3. Highlight from 'Craiglockhart.' to the end of 'Dear L.'

4. While that is highlighted press control-e as a shortcut key (on Windows, on Mac this may be command-e) by holding down the 'Ctrl' key and pressing 'e'. Or you can right-click and under 'Refactoring' select 'Surround with Tags'. A dialog box should pop up and type `opener` into it. Notice how oXygen helps you again by putting the opening tag before what you had highlighted and the closing tag afterwards.

5. Your document should now look like:

```
 <?xml version="1.0" encoding="UTF-8"?>
<text>
 <body>
  <pb/>
  <opener>Craiglockhart.
     July 1917.
     Wednesday

     Dear L.</opener>
   Thanks for yours of this morning. I hope
   you [have] had my card -of- posted last Monday.
   On Mond. next I lecture the "Field Club" -
   a Nat. Hist. Association, in the lines of our
   old Society - Geological, (you + me) + Botanical
   (New) Do you remember: -my- you old
   [... a lot more text...]
 </body>
</text>
```

6. Highlight from 'Thanks for yours of...' all the way down to the end of 'nearly licked old Herk.)'
7. Use the same method as above to surround this with a `<div>` element.
8. Then highlight from '... How Earth herself empowered' all the way to the end of this bit of poem at 'and flickered from his eyes.'
9. Although you can use the same method as above to surround this with a `<div>` element, you could also use 'control-/' (control held down then the '/' key on Windows) to surround this with whatever you provided to surround-with-element last time. Try it!
10. Highlight from 'I had seen your Song' all the way to 'olde Ballad! Heigh ho!' and surround it with a `<div>` element as well.
11. This leaves a final closing section which says '_Ever Yours_WEO', surround this with a `<closer>` element instead.
12. Your document should now look like:

```
 <?xml version="1.0" encoding="UTF-8"?>
<text>
 <body>
  <pb/>
  <opener>Craiglockhart.
     July 1917.
     Wednesday

     Dear L.</opener>
  <div>Thanks for yours of this morning. I hope
     you [have] had my card -of- posted last Monday.
     [...]
     nearly licked old Herk.)
  </div>
   --
 <div>... How Earth herself empowered him with her touch,
     [...]
     Stirred on his face, and flickered from his eyes.
  </div>--
```

```
 <div>I had seen your Song. May the music
     [...]
     replicas of the olde Ballad! Heigh ho!
  </div>
  <closer>_Ever Yours_WEO</closer>
 </body>
</text>
```

That is, it should contain an `<opener>`, 3 `<div>` elements, and a `<closer>`.

1.7 What type of division are these?

We have three divisions, two of which have prose and one of which has poetry. Let's pretend that we might be interested in easily finding all these different sections and add a *@type* attribute to each of the divisions.

1. Move the cursor back to just after the 'v' in the first `<div>` tag.
2. Press space, and then type `type="prose"` noting that oXygen provides the second double quotation mark for you.
3. Add a `type="verse"` to the second division
4. Add a `type="prose"` to the third division
5. Notice that the transcriber had put in two hyphens '- -' before and after the verse division to separate it off. You can delete these now since we have separated it off as a division.
6. Your divisions should now look like:

```
<div type="prose">Thanks for yours of this morning. I hope
[...]
nearly licked old Herk.)
</div>
<div type="verse">... How Earth herself empowered him with her touch,
[...]
Stirred on his face, and flickered from his eyes.
</div>
<div type="prose">I had seen your Song. May the music
[...]
replicas of the olde Ballad! Heigh ho!
</div>
```

1.8 Marking Paragraphs and Verse Lines

One thing we haven't done is mark paragraphs inside the divisions!

1. We can use a combination of 'surround-with-element' and 'split-element' to mark up lots of text quickly. In this case our prose divisions each have two paragraphs.
2. Highlight from 'Thanks for yours' all the way to the end of 'nearly licked old Herk.)' without including the `<div>` tags.
3. Use control-e or 'surround-with-element' to surround this with a single `<p>` element.
4. However, it really is two paragraphs, so move the cursor back between the end of 'Lessons at the Berlitz, Edin.' and the start of 'Last week I wrote (to order) a strong' and press 'alt-shift-d' (on Windows) or select `Refactoring -> Split Element` from the right-click menu. This should result in the first division having two paragraphs.

5. In the last prose division do the same, splitting the paragraph into two after 'Its now at home.' and before 'I see Swinburne'.

6. The verse division does not have paragraphs inside it, but lines of verse. Use whichever of the methods you learned above to mark up each line of verse with an <l> element.

7. In the end your divisions should look like this:

```
<div type="prose">
 <p>Thanks for yours of this morning. I hope
   [...]
   Lessons at the Berlitz, Edin.
 </p>
 <p> Last week I wrote (to order) a strong
   [...]
   nearly licked old Herk.)</p>
</div>
<div type="verse">
 <l>... How Earth herself empowered him with her touch,</l>
 <l>Gave him the grip and stringency of Winter,</l>
 <l>And all the ardour of th' invincible Spring;</l>
 <l>How all the blood of June glutted his heart,</l>
 <l>And all the glow of huge autumnal storms</l>
 <l>Stirred on his face, and flickered from his eyes.</l>
</div>
<div type="prose">
 <p>I had seen your Song. May the music
   [...]
   will. Its now at home.
 </p>
 <p> I see Swinburne also wrote a number of
   replicas of the olde Ballad! Heigh ho!</p>
</div>
<closer>_Ever Yours_WEO</closer>
```

Don't worry if you used the split-element method on the verse lines and it placed them as </l><l>, the whitespace does not matter here as long as the structure is good.

1.9 Linebreaks in Prose

The transcriber was very careful to press 'return' before each new line in the letter. It seems a shame to waste this intellectual effort so let's mark the prose linebreaks where they are significant.

1. In the <opener> put a 'linebreak' element <lb/> just after 'Craiglockhart.', '1917.' and 'Wednesday'. You can put this in once and then highlight it, copy it (control-c on windows) and paste it (control-v on windows) to make this quicker. If you do so oXygen might indent the lines for you slightly for clarity.

2. Do the same after each prose line of the two divisions we marked as 'prose'.

3. Whether you add a <lb/> at the beginning of each line, or at the end of it is an editorial decision that is up to you. As long as you are consistent in your decision it shouldn't matter too much later.

1.10 Formatting and Indenting a Well-formed XML Document

1. Make sure that your file is 'well-formed'. You'll be able to tell it is well-formed because oXygen will have a happy green square in the upper right-hand corner. If it is an angry

red square, you'd better find the problem (where in the file is indicated by a red bar on the right-hand side and is underlined in red) and correct the mistake! (Ask for help if you need it!)

2. Now let's format and indent our file. This tidies up some of the whitespace and indents elements based on their place in the hierarchy. Either select the 'Format and Indent' icon from the toolbar (it looks like some indented lines), or go to the menus: 'Document' -> 'Source' -> 'Format and Indent'.

3. Formatting and indenting your markup is not necessary, it could all be on one big long line, but it makes it much easier for other people to read.

4. In the end your file should look something like this:

```
 <?xml version="1.0" encoding="UTF-8"?>
<text>
 <body>
  <pb/>
  <opener>Craiglockhart.<lb/> July 1917. <lb/> Wednesday<lb/> Dear L.</opener>
  <div type="prose">
   <p>Thanks for yours of this morning. I hope<lb/> you [have] had my card -of-
posted last Monday. <lb/> On Mond. next I lecture the "Field Club" - <lb/> a
Nat. Hist. Association, in the lines of our <lb/> old Society - Geological,
(you + me) + Botanical <lb/> (New) Do you remember: -my- you old <lb/> Black
Molt? Well, the days have <lb/> come when I am [one of the] founders of a real
<lb/> learned society. My subject -is- has <lb/> the rather journalese Title
of "Do Plants <lb/> Think? - a study of the Response to <lb/> Stimuli -shown-
+ Devices for Fertilisation, <lb/> etc. I have no books yet, but I remember
a number of useful points<lb/> from your big Cassels' (I think it was <lb/>
Cassels') studied a 1911. Meanwhile <lb/>
    <pb/> I'm beastly bothered with our Mag.<lb/> (herewith) _and_I'm take German
<lb/> Lessons at the Berlitz, Edin.<lb/>
   </p>
   <p>Last week I wrote (to order) a strong <lb/> bit of Blank: on _Antaeus
v. Heracles_. <lb/> These are the best lines, methinks: <lb/> (N.B. Antaeus
deriving strength from his Mother Earth<lb/> nearly licked old Herk.)<lb/>
   </p>
  </div>
  <div type="verse">
   <l>... How Earth herself empowered him with her touch,</l>
   <l>Gave him the grip and stringency of Winter,</l>
   <l>And all the ardour of th' invincible Spring;</l>
   <l>How all the blood of June glutted his heart,</l>
   <l>And all the glow of huge autumnal storms</l>
   <l>Stirred on his face, and flickered from his eyes.</l>
  </div>
  <div type="prose">
   <p>I had seen your Song. May the music<lb/> be equally happy. You
_are_lucky!<lb/> You shall have my Locke's "Usurper" if you<lb/> will. Its
now at home. <lb/>
   </p>
   <p> I see Swinburne also wrote a number of<lb/> replicas of the olde Ballad!
Heigh ho!<lb/>
   </p>
  </div>
  <closer>_Ever Yours_


     WEO</closer>
 </body>
</text>
```

1.11 Saving Your Work

Let's save our work:

- Is your work well-formed? Do you have a happy green square or an angry red one?
- From the 'File' menu select 'Save' or click on the Save icon (looks like an old-style 3.5" disk)
- Save the file using the name 'exercise01.xml' or another name of your choice.

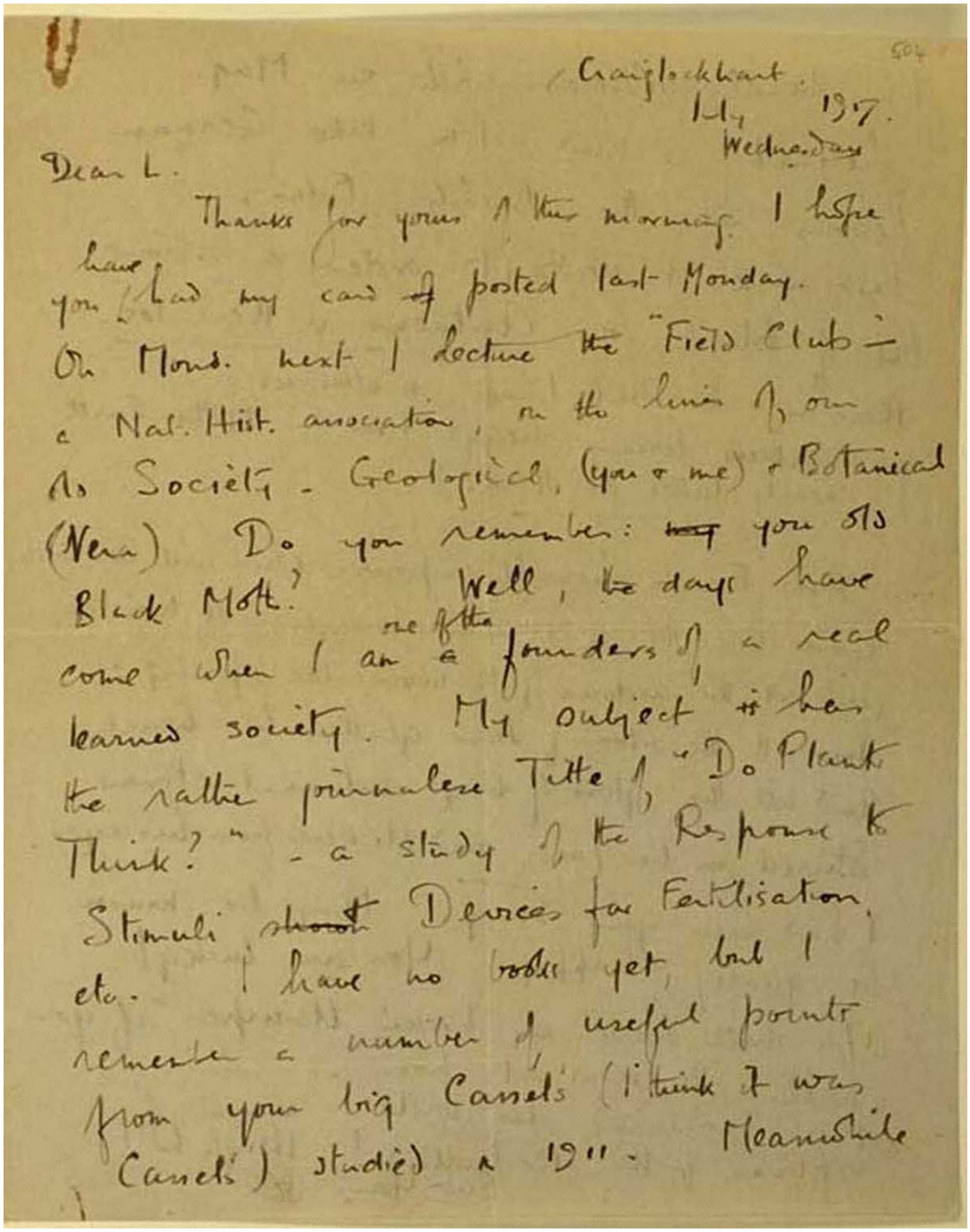
1.12 Self-Assessment

Check that you understand some of the core principles of this exercise by answering the following questions to yourself:

- How do you start a new XML document in oXygen?
- What is an XML declaration?
- What is a well-formed document?
- How do I 'Surround with tag' and repeat that action quickly?
- Why might using the 'Split element' approach be useful?
- What is the function of each element and attribute in your current file?
- What is the advantage of formatting and indenting your markup?

1.13 Next?

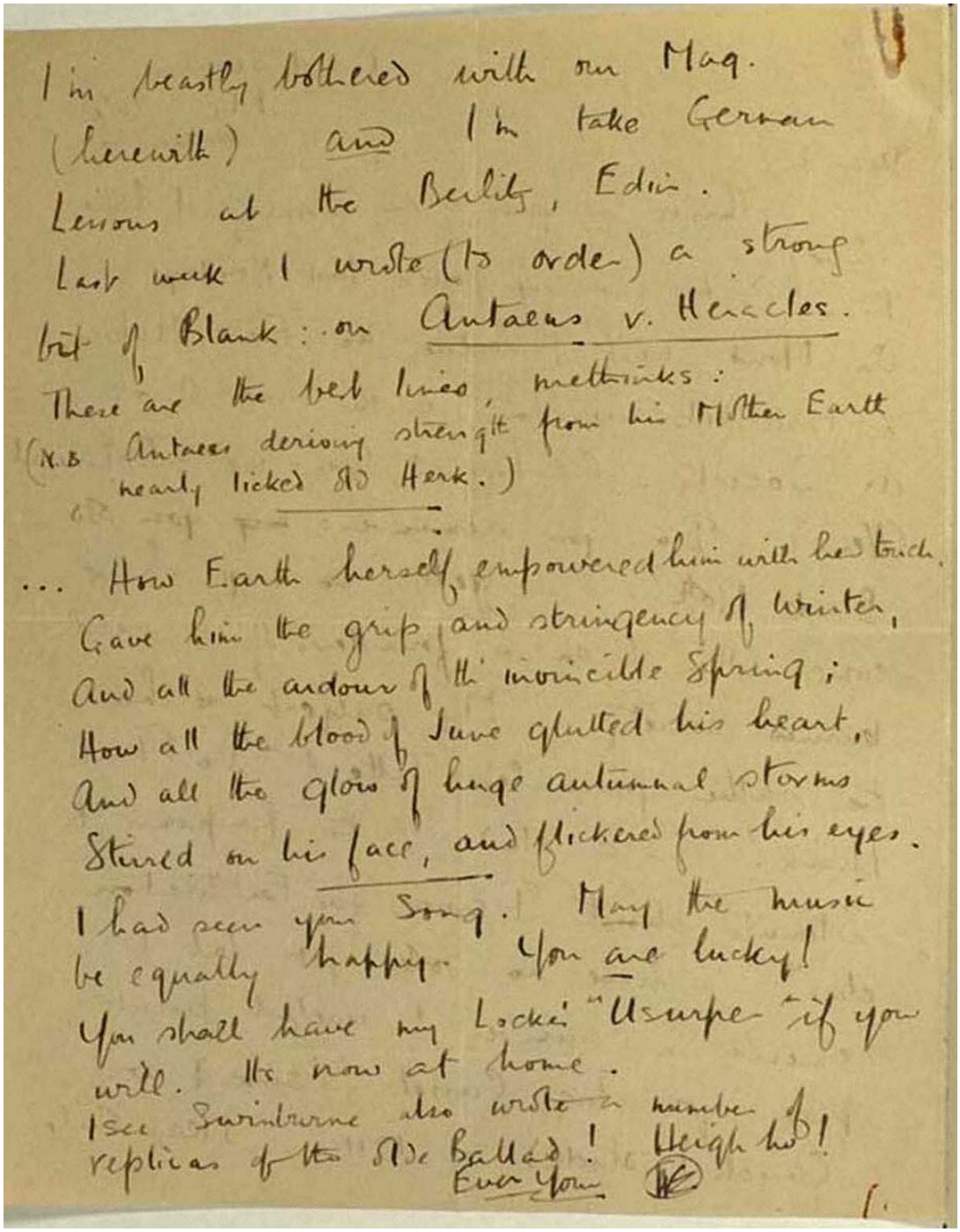
Your XML file may be well-formed but it is not yet **valid** because it doesn't validate against a particular schema (such as those which are customisations of the TEI). Next we will have a short introduction to the structure of TEI documents and some of the most frequently used elements. If you are finished early you may wish to browse through the TEI Guidelines online at http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index-toc.html. In particular you might want to look at the 'Elements' appendix for reference pages of individual elements. Consider looking up all the elements you've used in this file to see how they are defined!



Craiglockhart.
14 May 1917.
Wednesday

Dear L.

Thanks for yours of the morning. I hope you have had my card posted last Monday. On Mond. next I lecture the "Field Club" — a Nat. Hist. association, on the lines of our As Society - Geological, (you & me) & Botanical (Nevin) Do you remember: you sts Black Moth? Well, the days have come when I am one of the founders of a real learned society. My subject is here the rather pronunciese Title of "Do Plants Think?" - a study of the Response to Stimuli Devices for Fertilisation, etc. I have no books yet, but I remember a number of useful points from your big Camels (I think it was Camels) studies of 1911. Meanwhile



I'm beastly bothered with our Mag.
(herewith) and I'm take German
Lessons at the Berlitz, Edin.
Last week I wrote (to order) a strong
bit of Blank: on Antaeus v. Heracles.
There are the best lines, methinks:
(N.B Antaeus deriving strength from his Mother Earth
nearly licked Old Herk.)

... How Earth herself, empowered him with her touch.
Gave him the grip and stringency of Winter,
And all the ardour of th' invincible Spring;
How all the blood of June glutted his heart,
And all the glow of huge autumnal storms
Stirred on his face, and flickered from his eyes.

I had seen your Song. May the music
be equally happy. You are lucky!
You shall have my Locke "Usurper" if you
will. It's now at home.
I see Swinburne also wrote a number of
replicas of the Old Ballad! Heigh ho!
Ever Yours