Updated 2 Jan. 2020

GENEALOGIES OF KNOWLEDGE

How to clean up documents and prepare them for the internet corpus

Purpose: To produce a machine-readable document that corresponds to the contents of the online article or other item to be incorporated into the internet corpus.

This involves copy-and-pasting the material from the online source into a .xml file and then using a text editor (jEdit or Oxygen, for example) to annotate the text with XML metadata. jEdit is used here for illustration.

N.B. If you do not already have jEdit installed on your computer, it can be downloaded here: http://www.jedit.org/

Setting Up jEdit (you only have to do this once):

On the menu bar: Select UTILITIES → GLOBAL OPTIONS …

Select ENCODINGS (on the left-hand side)

Choose DEFAULT CHARACTER ENCODING: UTF-8

Deselect AUTO-DETECT FILE ENCODING WHEN POSSIBLE

Select EDITING (from the same JEdit tree on the left-hand side)

Set WORD WRAP to 'soft'. Click on Apply and Ok.

On menu bar: Choose PLUGINS → PLUGIN MANAGER

Select INSTALL tab (top left)

Tick XML (you will need to scroll down to find it on the list)

Click on INSTALL button and then CLOSE

Back to menu bar: Select PLUGINS → SIDEKICK

Tick HIGHLIGHT MARKERS IN SIDEKICK TREE

It is useful to have a printed copy of the 'goktext.dtd' and 'gokheader.dtd' files to hand when adding the metadata. These documents include all the possible metadata tags that you can use to annotate the text and it is against these .dtds that jEdit will check for errors.

Creating the text file:

1. In jEdit, click FILE > NEW and copy-and-paste the full text of the online article into this new file. You do not need to include any of the text included in the website's menus or advertising banners etc., only the written content relevant to that particular article: e.g. the title, the author's name, editor's introduction, the main body of the text, etc..

2. Click FILE > SAVE AS and save this text as an .xml file with the filename you have been given for this article (e.g. int000083.xml - please note that all filenames in the internet corpus should begin with **int** and then a **6-digit number**). Make sure you are saving the file in the same folder on your computer's hard drive as the 'goktext.dtd' and 'gokheader.dtd' files.

3. Make sure there are no tabbed spaces or empty lines in this document (e.g. between paragraphs). Double check that this text corresponds accurately to the full written content of the original article.

Inserting the xml metadata:

4. Insert the following lines at the very top of this text:

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<!DOCTYPE goktext SYSTEM "goktext.dtd">
<goktext>
<section id='s1'>
```

   At the very end of the text, place these lines:

```
</section>
</goktext>
```

5. Place the title between <title>…</title> tags and any subtitle(s) between <shead>…</shead> tags.

   NB All tags need to have two parts, one that opens them and one that closes them (<title>…..</title>).

6. Throughout the main body of the text, a <p/> tag should be placed at the end of each paragraph.

   Note that <p/> is a special case in that, unlike all other tags, it does not require an opening tag at the beginning of the paragraph. Simply place <p/> at the end of each paragraph.

7. Any web addresses (URLs) should be placed between <url>…</url> tags.

8. Please note that **paratextual elements should NOT be tagged** as such in any internet corpus files. The following tags **are only intended to be used in the Modern English corpus**: <frontmatter>, <backmatter>, <endorsements>, <acknowledgements>, <preface>, <introduction>, <appendix>, <foreword>, <afterword>, <epigraph>, <bibliography>, <footnote> and <endnote>.

9. To check if you have entered the metadata correctly, SAVE and CLOSE the file and then re-open it again. An 'ErrorList' window should pop up with a list of any issues. If this does not work, click on PLUGINS > ERROR LIST > ERROR LIST.

   The most common errors are when the closing tag (e.g. </footnote>) has been omitted or if you have used a tag element that is not included in the .dtd.

Creating the header file:

10. We also need to create an xml header file for the text. This provides all the metadata relevant to the article: the date of publication, the names of the author(s), translator(s), source language etc.

11. To begin, click FILE > NEW and then save this new file as a .hed file with the same filename as the text file (e.g. int000083.hed - please note that all filenames in the internet corpus should begin with **int** and then a **6-digit number**).

12. Paste the following template into this file:

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<!DOCTYPE gokheader SYSTEM "gokheader.dtd">
<gokheader>
 <document format="" subcorpusid="internet" filename="">
 </document>
  <section id="s1"  publication_date="" authorship_date="" outlet="" internet_outlet=""
       language="">
       <title></title>
       <author>
       <name></name>
       </author>
          <summary></summary>
  <translation status="translation" language="English">
      <translator certainty="certain">
      <name></name>
      </translator>
   <source date="" language=""><original_title></original_title></source>
   </translation>
  <copyright status="">
   <terms></terms>
   <holder></holder>
  </copyright>
  <comments></comments>
 </section>
</gokheader>
```

13. Insert relevant information into each field (e.g. title, author, filename, summary etc.).

    In cases where you want to add more than one name (e.g. translators) just copy and paste that line/section. For example, a text with one translator would look like this:

```
<translator>
<name>Helen Zimmern</name>
 </translator>
```

Whereas a text with two translators will look like this:

```
<translator>
<name>Helen Zimmern</name>
<name>Peter Bush</name>
</translator>
```

14. For the 'publication_date' attribute, simply insert the year in which the text was published online. The day and month of publication (if available) can be placed between the <comments>…</comments> tags.

15. The 'authorship_date' attribute should be filled with the year in which the text was written (or translated in the case of translations). In the Internet corpus, the authorship_date will in almost all cases be the same as the publication_date.

16. The source_date is only used in the case of translations: to refer to the year in which the original source text was published.

17. 'internet_outlet' refers to the name of the platform on which the text has been published e.g. ROAR Magazine. These outlets are defined in the GOKHEADER.DTD so must be spelled in the header in exactly the same manner as in this DTD (otherwise the header will not validate).
    Here's an example of what a header file should look like when it is finished:

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<!DOCTYPE gokheader SYSTEM "gokheader.dtd">
<gokheader>
<document filename="int000074" format="report_journal_article" subcorpusid="internet">
</document>
<section id="s1" language="English" outlet="website" internet_outlet="ROAR_Magazine"
publication_date="2015" authorship_date="2015">
   <title>Reopening the Revolutionary Question</title>
    <author><name>Amador Fernández-Savater</name></author>
          <summary>Starting out form The Invisible Committee's publications, The Coming Insurrection and
          To our Friends, this article addresses the relationship between revolution and common life. The
          true revolutionary spirit is founded not in a delineated ideology but in a continuous if inconclusive
          ethics. Revolution is a process which is not as much about proper government as it is about self-
          organization: the current struggle dictates conduct more directly than any abstract structure
          could.</summary>
   <translation status="translation" language="English">
      <translator>
     <name>Liz Mason-Deese</name>
    </translator>
    <source date="2015" language="Spanish"><original_title>Reabrir la cuestión
revolucionaria</original_title></source>
  </translation>
    <copyright status='copyright_granted'>
    <terms></terms>
    <holder>ROAR Magazine</holder>
  </copyright>
  <comments>Text published 09 December 2015. Original source text available at:
http://www.eldiario.es/interferencias/comite_invisible-revolucion_6_348975119.html</comments>
  </section>
```

18. If the text you are working with is not a translation (i.e. a document originally written in English), you will need to delete everything between the <translation>...</translation> tags (including the tags themselves) to ensure that the text validates:

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<!DOCTYPE gokheader SYSTEM "gokheader.dtd">
<gokheader>
<document filename="int000091" format="report_journal_article" subcorpusid="internet">
</document>
<section id="s1" language="English" outlet="website" internet_outlet="Discover_Society"
publication_date="2015" authorship_date="2015">
    <title>RADICAL EDUCATION NOW</title>
     <author><name>Nick Stevenson</name></author>
    <summary>This article draws on the work of Ivan Illich to argue that the debate about education needs
to be radically democratised. It contends that rather than designing schools and universities to meet the
needs of the corporate sector, education policy should take into consideration the voices of those who have
previously been dismissed as the waste products and failures of the system.</summary>
    <copyright status='copyright_granted'>
     <terms></terms>
     <holder>Discover Society</holder>
   </copyright>
   <comments>Text published 02 November 2015</comments>
   </section>
```