GENEALOGIES OF KNOWLEDGE

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

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

N.B. This is a two-stage process: first, you will need to 'clean' the document in Microsoft Word (i.e. ensure that the spelling and paragraphing corresponds to that of the original document), and then second, you will need to use an xml editor (jEdit or Oxygen, for example) to annotate the text with metadata. jEdit is used here for illustration.

CLEANING THE WORD FILE

1. You will have been sent an electronic scan/copy of the original document. If this is not already contained in a word document (.doc), you will need to convert it using software such as Adobe Acrobat.

   Save this .doc file under the file name that you have been given for this document (e.g. mod000005). Please note that all filenames in the modern corpus should begin with **mod** and then a **6-digit number**.

2. You can delete any of the copyright and publication information that is often included on the inside front cover of books. This information will be included in the header file when we come to the jEdit stage. Any Epigraphs, Tables of Contents and/or Acknowledgements should be left in this document however. You should not include the index.)

3. Make sure the spelling in the .doc file corresponds with the spelling in the original book. Make sure that spellcheck is activated and that you check all the underlined words/phrases. In some cases, the errors may be just a word that MS Word does not know or a syntax that is unusual. Always check with the original document, if in doubt.

   If the doc file has been created by scanning the original it will likely contain many OCR errors. In the example pasted below, you will see for instance that the title of this work (highlighted in yellow) has been rendered 'Retreatin9 the Political' and so this will need to be corrected.

This collection of essays presents, for the first time in English, some of the key issues at the heart of Philippe Lacoue-Labarthe and Jean-Luc Nancy's work. Including several unpublished essays, *Retreatin9 the Political* offers some highly original perspectives on the relationship between philosophy and the political.

Some of these errors will recur throughout the text and can be quickly corrected using the Find & Replace function within Word.

Note that any special characters in the original text are unlikely to be recognised by the OCR text scanner and so it is again useful to have a copy of the original document here for reference.

4. Make sure that the paragraphs in the .doc file correspond to those in the document. Clicking on the 'Show ¶' button within Word's main toolbar will help you to identify more easily the end of each paragraph in the .doc file.

5. There should be no empty lines in the .doc file. For example, below is how it should look like in the .doc file, even though in the book, chapter three starts on a new page and the new paragraph begins a couple of lines later.

Using the term "*a priori*" in its (genetic) meaning of "anticipatory", one might characterise the psychological (and also the epistemological) "theory of preformation" advocated here, the deductivist-empiricist view, by means of the formulation: *There are, indeed, synthetic a priori judgements, but a posteriori they are often false.*
Chapter III
THE PROBLEM OF INDUCTION
5. The infinite regression (Hume's argument). Hume1 was the first who presented, with exemplary clarity, the difficulties surrounding the problem of universal empirical statements, or the problem of induction ("Can we know more than we know?"). He demonstrated that every attempt at *inductive generalisation* will be defeated by a *circular inference*.

You should also remove any tabbed spaces and other non-essential 'white space' from the document. Please see the instructions on using Regular Expressions for a quick and easy means of doing this: http://genealogiesofknowledge.net/genealogies-knowledge-corpus/corpus-text-preparation/

6. If at any point there is a table or diagram in the text, you will need to re-arrange the formatting so that each line of the table or diagram exists as a separate paragraph. A table in the document that looks like this, for example:

Table 2

	Inductivism	Deductivism
Rationalism	3	1
Empiricism	2	4

must be rendered like this:

Table 2

Inductivism Deductivism

Rationalism 3 1

Empiricism 2 4

7. If the work contains any footnotes, these should be left where they are on the page (see example below). These will be placed in <footnote> tags when we come to the xml annotation stage. Original text:

Kant's theory of knowledge (leaving aside the preliminary work of his friend J.H. Lambert) is the first attempt at a critical synthesis of the classical opposition between rationalism and empiricism. Kant set himself the task of determining, through this synthesis, the "formal" and the "material" side of knowledge: the formal side by taking over elements of rationalism, the material side by taking over empiricist elements. (Perhaps this tendency is expressed in its purest form in the first two "Postulates of Empirical Thought in General",[4] here quoted in Section 11, as well as in his famous formulation: "Thoughts without content are empty, intuitions without concepts are blind."[5])

In this way Kant's critique of pure reason attempts to solve essentially the same problems that I called here (Section 1) the fundamental problems of the theory of knowledge.

[4] Immanuel Kant, *Kritik der reinen Vernunft* (2nd ed., 1787), pp. 265 f. [English translation by N. Kemp Smith (1929): *Critique of Pure Reason*, p. 196, A. Pickel (2nd version c.1993), pp. 239 f. Tr.].

[5] Immanuel Kant, *op. cit.*; p. 75 [English translation, *op. cit.*, p. 93. Tr.].

'Cleaned' word document:

> Kant's theory of knowledge (leaving aside the preliminary work of his friend J.H. Lambert) is the first attempt at a critical synthesis of the classical opposition between rationalism and empiricism. Kant set himself the task of determining, through this synthesis, the "formal" and the "material" side of knowledge: the formal side by taking over elements of rationalism, the material side by taking over empiricist elements. (Perhaps this tendency is expressed in its purest form in the first two "Postulates of Empirical Thought in General",4 here quoted in Section 11, as well as in his famous formulation: "Thoughts without content are empty, intuitions without concepts are blind."5)
> In this way Kant's critique of pure reason attempts to solve essentially the same problems that I called here (Section 1) the fundamental problems of the theory of knowledge.
> 4 Immanuel Kant, *Kritik der reinen Vernunft* (2nd ed., 1787), pp. 265 *f.* [English translation by N. Kemp Smith (1929): *Critique of Pure Reason*, p. 196, A. Pickel (2nd version c.1993), pp. 239 *f.* Tr.].
> 5 Immanuel Kant, *op. cit.*; p. 75 [English translation, *op. cit.*, p. 93. Tr.].

Endnotes should be left at the end of the document. They will be contained within the <backmatter> tags when we come to the xml annotation stage.

8. Return any text placed in columns to a standard, single-column format and remove any headers/footers from the document (e.g. page numbers, place markers etc.).

9. When you are finished, select all of the text (Ctrl + A) and go to Styles -> Clear All.

10. Finally, save the document as a plain text file (.txt) with the same filename (e.g. mod000005). After you click Save, a dialogue box will open asking how you want to convert the document. Select 'Other Encoding' and choose 'Unicode (UTF-8)' from the menu on the right-hand side.

11. You should now have two files, a .doc and a .txt, with the same name.

    IMPORTANT: Make sure these files are both in the same folder in your computer as the files 'gokheader.dtd' and 'goktext.dtd'.

USING JEDIT TO ADD METADATA:

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 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 xml text file:

1. Open the text file in jEdit. Go to FILE > SAVE AS and save it as an .xml file (e.g. mod000005.xml). Please note that all filenames in the modern corpus should begin with **mod** and then a **6-digit number**.

2. Insert the following lines at the top of this file:

```
<?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>

(NB If the text is an edited collection – i.e. each chapter of the book was originally written by a different author and/or published at a different time – then you will have to create a separate section for each chapter. For more details on how to deal with edited collections, please see the instructions that follow at the end of this document.)

3. Place the title between <title>…</title> tags and any subtitle between <shead>…</shead> tags.

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

4. The main body of the text now needs to be divided up into three parts: frontmatter, chapters and backmatter.

   In between the <frontmatter></frontmatter> tags you should place:

   Epigraphs - between <epigraph></epigraph> tags

   Acknowledgements - between <acknowledgements></acknowledgements> tags

   Tables of contents – between <table_contents></table_contents> tags

   Prefaces - between <preface></preface> tags

   Introductions – between <introduction></introduction> tags

   Each chapter should be placed between <chapter></chapter>tags and numbered in the correct order, as follows:

   <chapter n=”1”>….</chapter>
   <chapter n=”2”>….</chapter>
   <chapter n=”3”>….</chapter>
   Etc.

   The <backmatter></backmatter> tags should include any endnotes, bibliographic references, etc.

   e.g.

   <endnote>'Opening Address', p. 114, below.</endnote>
   <endnote>Nancy, The Inoperative Community, p. xxxvi.</endnote>
   <endnote>'Annexe', p. 140, below.</endnote>

<endnote>'Opening Address', p. 113, below.</endnote>
<endnote>Les fins de l'homme, p. 15.</endnote>

5. Go through the text and place the tag <p/> at the end of every paragraph in the book.

   NB <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.

6. Place footnotes between <footnote></footnote> tags:

Kant's theory of knowledge (leaving aside the preliminary work of his friend J.H. Lambert) is the first attempt at a critical synthesis of the classical opposition between rationalism and empiricism. Kant set himself the task of determining, through this synthesis, the "formal" and the "material" side of knowledge: the formal side by taking over elements of rationalism, the material side by taking over empiricist elements. (Perhaps this tendency is expressed in its purest form in the first two "Postulates of Empirical Thought in General",4 here quoted in Section 11, as well as in his famous formulation: "Thoughts without content are empty, intuitions without concepts are blind."5) In this way Kant's critique of pure reason attempts to solve essentially the same problems that I called here (Section 1) the fundamental problems of the theory of knowledge. <footnote>4 Immanuel Kant, *Kritik der reinen Vernunft* (2nd ed., 1787), pp. 265 *f.* [English translation by N. Kemp Smith (1929): *Critique of Pure Reason*, p. 196, A. Pickel (2nd version c.1993), pp. 239 *f.* Tr.].</footnote> <footnote>5 Immanuel Kant, *op. cit.*; p. 75 [English translation, *op. cit.*, p. 93. Tr.].</footnote>

7. Any tables must be placed between <list type="tables"></list> tags.

   Diagrams can be placed between <diagram></diagram> tags.

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

8. To check for errors in the metadata, click on PLUGINS → ERROR LIST → ERROR LIST to see them. 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:

9. Finally, we need to create an xml header file for the text. This provides all the metadata relevant to the text: the date of publication, the names of the author(s), translator(s), etc.

10. To begin, click FILE > NEW and then save this as a .hed file with the same filename as the text file (e.g. mod000005.hed). Please note that all filenames in the modern corpus should begin with **mod** and then a **6-digit number**.

11. Paste the following lines into this file:

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<!DOCTYPE gokheader SYSTEM "gokheader.dtd">
<gokheader>
 <document format="" subcorpusid="modern" filename="">
      </document>
<section id="s1"  publication_date="" authorship_date="" 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>
```

12. Insert relevant information in each field (e.g. author, title, 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 book with one translator would look like this:

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

    Whereas a book with two translators will look like this:

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

13. Please note that for most texts included in the Modern English corpus, the authorship_date and publication_date will be the same.

14. For translations from classical Greek, the <original_title> element should be filled with the Greek title transliterated into the roman alphabet e.g. <original_title>Thoukudidou Historiai</original_title> or <original_title>Apologia Sōkratous</original_title>.

15. For classical Greek and Roman texts, the source date should be formatted as follows: "c.400 BC" or "c.485-424 BC" or "125 AD".

16. N.B. Unless you are working with an edited collection, the section id will always equal “s1”:

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 format="report_journal_article" subcorpusid="modern" filename="mod000057">
    </document>
        <section id="s1"  publication_date='2011' authorship_date='2011' outlet='publisher'
        language='English'>
          <title>Trends Come and Go, the Community Remains</title>
                <author>
                <name>Tarik Dahou</name>
                </author>
          <summary>This paper examines the concept of ‘community’ in the context of development theory.
          Through a focus on Africa, it criticises the ways in which the notion has been instrumentalised in
          the social sciences as a means of overcoming the complexity of specific socio-political situations
          and of defining strategic channels through which citizens must participate in
          democracy.</summary>
                    <translation status="translation" language="English">
                            <translator certainty="certain">
                          <name>JPD Systems</name>
                           </translator>
                                    <source date="2011" language="French">
                           <original_title>Les modes passent, la communauté reste</original_title>
                                     </source>
                  </translation>
  <copyright status='copyright_granted'>
           <terms>Not available for commercial purposes</terms>
           <holder>Cairn International on behalf of P.U.F.</holder>
  </copyright>
          <comments>Available online at: http://www.cairn-int.info/article-E_CEA_202_0395--trends-come-
          and-go-the-community-remains.htm</comments>
  </section>
</gokheader>
```

17. If the text you are working with is not a translation (i.e. a document originally written in English), you will need to delete the whole section of the header relating to translation from the header (<translation>...</translation>), to ensure that the text validates.

18. Finally, check for errors in the same way as above for the xml text file: click on PLUGINS → ERROR LIST → ERROR LIST.

Dealing with edited collections:

xml text file:

Because edited collections generally contain many different texts written at different times by different authors, we cannot place all of the volume's content in a single section. Instead, each individual text contained in the collection must be placed in its own section within the xml text file, as follows:

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<!DOCTYPE goktext SYSTEM "goktext.dtd">
<goktext>
        <section id="s1">
                [Insert first text/chapter here]
        </section>
        <section id="s2">
                [Insert second text/chapter here]
        </section>
        <section id="s3">
                [Insert third text/chapter here]
        </section>
        …
</goktext>
```

NB The first section of the text (section id="s1") will in most cases contain only the frontmatter (editor's preface and/or introduction, acknowledgements etc.). You must still place this content within <frontmatter></frontmatter> tags, as you would with a single section text.

```
<section id="s1">
        <frontmatter>
            [Insert frontmatter here]
        </frontmatter>
</section>
```

Likewise, the last section of the text (e.g. section id="s14") will in most cases contain only the backmatter (bibliography, endnotes, etc.). You must still place this content within <backmatter></backmatter> tags, as you would with a single section text.

```
<section id="s14">
        <backmatter>
            [Insert backmatter here]
        </backmatter>
 </section>
```

Header file:

When you create the header file for the edited collection, you will then need to create a separate set of metadata for each section. Some of the metadata will be the same for every section (e.g. the summary and copyright information) but others will be different (e.g. title, author, source date etc.).

Please note that when dealing with an edited collection, instead of placing the book summary within the <section> tags, you should instead place this within the <document> tags (highlighted in red in the example below).

An example header file for an edited collection is pasted below. The beginning and end of each section are highlighted in yellow and the main differences between each section are highlighted in green. Note that there is a new set of elements within the <document>…</document> tags which provides details about the volume as a whole (highlighted in blue):

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<!DOCTYPE gokheader SYSTEM "gokheader.dtd">
<gokheader>
<document filename="mod000007" subcorpusid="modern" format="edited_collection" >
        <collection_title>Retreating the Political</collection_title>
        <editor>Simon Sparks</editor>
                <place_publication>London and New York</place_publication>
                <summary>This collection of essays by Philippe Lacoue-Labarthe
                and Jean-Luc Nancy explores the nature of the link between
                philosophy and politics. Specifically, it highlights the current
                importance of questioning the political from a philosophical point of
                view in order to understand the cultural conditions that underlie
                totalitarianism and democracy.</summary>
</document>
        <section  id="s1" language="English" outlet="publisher"  publication_date="1997"
        authorship_date="1997">
           <title>Frontmatter</title>
             <author>
              <name>Simon Sparks</name>
              </author>
           <copyright status='copyright_granted'>
                 <terms>Not available for commercial purposes</terms>
                  <holder>Routledge</holder>
           </copyright>
                  <comments></comments>
        </section>
        <section  id="s2" language="English" outlet="publisher"  publication_date="1997">
           <title>La Panique Politique</title>
             <author>
                   <name>Philippe Lacoue-Labarthe</name>
                   <name>Jean-Luc Nancy</name>
              </author>
                <translation status="translation" language="English">
                        <translator>
                                <name>Celine Surprenant</name>
                        </translator>
                        <source date="1979" language="French">
                                <original_title>La Panique Politique</original_title>
                        </source>
                </translation>
                <copyright status='copyright_granted'>
                        <terms>Not available for commercial purposes</terms>
```

```
                    <holder>Routledge</holder>
                </copyright>
            <comments></comments>
        </section>
        <section  id="s3" language="English" outlet="publisher"  publication_date="1997">
            <title>The Free Voice of Man</title>
                <author>
                    <name>Jean-Luc Nancy</name>
                </author>
                <translation status="translation" language="English">
                    <translator>
                        <name>Richard Stamp</name>
                    </translator>
                    <source date="1981" language="French">
                        <original_title>La voix libre de l'homme</original_title>
                    </source>
                </translation>
                <copyright status='copyright_granted'>
                    <terms>Not available for commercial purposes</terms>
                    <holder>Routledge</holder>
                </copyright>
                <comments></comments>
            </section>

        …

</gokheader>
```