**Tagging document files**

Source documents often need to be structured into sections and subsections, and this information needs to be communicated to Greenstone so that it can preserve the hierarchical structure. Also, metadata - typically the title - might be associated with each section and subsection.

The source documents from an OCR process are typically a set of word processor files, including images. If these are represented as MicrosoftWord files, they can be input into Greenstone using the Word plugin. Alternatively, they can be converted to HTML and input using the HTML plugin.

In either case, the hierarchical structure of a document may be indicated by inserting tags in the text as follows:

<!--

<Section>

<Description>

<Metadata name="Title">Realizing human rights for poor people: Strategies for achieving the international development targets</Metadata>

</Description>

-->

*(text of section goes here)*

<!--

</Section>

-->

The <!-- ... --> markers are used because they indicate comments in HTML; thus these section tags will not affect document formatting. You must include these markers around your section tags, even if the document you are working with is not HTML (e.g. if it's a Microsoft Word file).

In the Description part (between the <Description> and </Description> tags) other kinds of metadata can be specified, but this is not done for the style of collections we are describing here.

It is important to remember that you are creating a hierarchical table of contents when you insert section tags into your document. This means that sections can be nested within other sections. In fact, all sections must be nested within a single enclosing section that encompasses the entire document.

The following example demonstrates a document with two chapters, the second of which contains two subsections. For real examples of sourcedocuments tagged in this way, look at the source documents for the Demo or DLS collections.

<!--

<Section>

<Description>

<Metadata name="Title">My Document</Metadata>

</Description>

<Section>

<Description>

<Metadata name="Title">Chapter 1</Metadata>

</Description>

-->

(text of chapter 1 goes here)

<!--

</Section>

<Section>

<Description>

<Metadata name="Title">Chapter 2</Metadata>

</Description>

<Section>

<Description>

<Metadata name="Title">Subsection 1</Metadata>

</Description>

-->

(text of sub-section 1 goes here)

<!--

</Section>

<Section>

<Description>

<Metadata name="Title">Subsection 2</Metadata>

</Description>

-->

(text of sub-section 2 goes here)

<!--

</Section>

</Section>

</Section>

-->

Note that metadata assigned from within a section tag in a source document takes precedence over that assigned to the document as a whole. This means that you should not explicitly specify Title metadata for the top-level section within a source document unless you want it to override the title you gave it when specifying metadata. In the above example, unless you want to override the document's existing title you should omit the line that reads:

<Metadata name="Title">My Document</Metadata>

# Greenstone tutorial exercise

[Back to wiki](http://greenstone.sourceforge.net/wiki/index.php/Tutorial_exercises)   
[Back to index](http://wiki.greenstone.org/gsdoc/tutorial/en/index.html)   
Devised for Greenstone version: 2.70w  
Modified for Greenstone version: 2.82

## Section tagging for HTML documents

1. In a browser, take a look at the Greenstone demo collection. Browse to one of the documents. This collection is based on HTML files, but they appear structured in the collection. This is because these HTML files were tagged by hand into sections.
2. Using a text editor (e.g. WordPad) open up one of the HTML files from the demo collection: *Greenstone → collect → demo → import → fb33fe →fb33fe.htm*. You will see some HTML comments which contain section information for Greenstone. They look like:

<!--  
<Section>  
  <Description>  
    <Metadata name="Title">Farming snails 1: Learning about snails;  
    Building a pen; Food and shelter plants</Metadata>  
  </Description>  
-->  
  
<!--  
</Section>  
<Section>  
  <Description>  
    <Metadata name="Title">Dew and rain</Metadata>  
  </Description>  
-->

When Greenstone encounters a <Section> tag in one of these comments, it will start a new subsection of the document. This will be closed when a </Section> tag is encountered. Metadata can also be added for each section—in this case, **Title** metadata has been added for each section. In the browser, find the **Farming snails 1** document in the demo collection (through the *Titles* browser). Look at its table of contents and compare it to the <Section> tags in the HTML document.

1. Add a new Section into this document. For example, lets add a new subsection into the **Introduction** chapter. In the text editor, add the following just after the Section tag for the **Introduction** section:

<!--  
<Section>  
  <Description>  
    <Metadata name="Title">Snails are good to eat.</Metadata>  
  </Description>  
-->

Then just before the next section tag (**What do you need to start?**), add the following:

<!--  
</Section>  
-->

The effect of these changes is to make a new subsection inside the **Introduction** chapter.

1. Open the Greenstone demo collection in the Librarian Interface. In the **Document Plugins** section of the **Design** panel, note that **HTMLPlugin** has the **description\_tags** option set. This option is needed when <Section> tags are used in the source documents.
2. **Build** and **preview** the collection. Look at the **Farming snails 1** document again and check that your new section has been added.

Copyright © 2005 2006 2007 2008 2009 2010 by the [New Zealand Digital Library Project](http://www.nzdl.org) at [the University of Waikato](http://www.waikato.ac.nz), New Zealand  
Permission is granted to copy, distribute and/or modify this document under the terms of the [GNU Free Documentation License](http://www.gnu.org/licenses/fdl.html), Version 1.2 or any later version published by the Free Software Foundation; with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts. A copy of the license is included in the section entitled [“GNU Free Documentation License.”](http://greenstonewiki.cs.waikato.ac.nz/wiki/gsdoc/GNUFDL.html)