Creating WESS corpus

Objective is to have only main body text. Start from Chapter 1 or Introduction. Need to remove:

* Cover, front matter, overview, etc.
* Bibliography
* Footnotes
* Sidebars, figures, charts, tables

It’s ok to have the text of the boxes

If have the original .doc file, this is the best place to start.

If not, get a PDF and open it in Acrobat (not the Reader).

From there, you can select the object editor tool and start deleting elements.

* Footnotes, charts, tables, etc.

Header and footer, as well as sidebars can be removed by cropping the pages. Use odd/even to crop batches, but check the output (!)

Also need to delete non-showing elements [GET THE RIGHT COMMAND]

Finally, delete the pages that are not needed.

When you have the text as basic as possible, export to Word document.

In Word, the text will appear as text boxes. To fix this do: ctrl-A, ctrl-Q

Now need to go through word and clean up the text. Use find/replace, etc. Pay attention to ligatures (fi, ffl, etc) that may have been created.

You can try to manually join split paragraphs, but this is also easy to do in TextWrangler with the following grep search:

(?<=[a-z])\r(?=^[a-z]) (case sensitive, grep)

Spend as much time as you are willing to clean up the text.

End file should start with: “World Economic….” [year]: [Title]

Save as a text file (.txt)

Open in TextWrangler and run the commands: Zap Gremlins and Normalize line endings. Save.

Last, in the Terminal, run the following commands to make sure the file is encoded properly:

cd [path to the .txt file(s)]

textutil -convert txt -encoding UTF-8 \*.txt

Now join the text files into a single file if needed (Torch) or just keep them in a directory (Mallet)

One option to join .txt files: from Terminal:

cd [path to directory]

cat \*.txt > [filename]

Note: Don’t add extension to [filename]. Will join all the .txt files. The names define the order. If you want to check the order, start the command with ‘echo’ and it will output the file list order