1 Indexing MS Office documents with swish-e

This only partly works.

```
# Example configuration file

# Tell Swish-e what to directories to index
IndexDir /Users/jkitchin/Dropbox


# where to save the index
IndexFile /Users/jkitchin/.swish-e/office-index.swish-e

# What to index
IndexOnly .docx .pptx .xlsx

# Tell Swish-e that .txt files are to use the text parser.
IndexContents XML* .docx .pptx .xlsx

FileFilter  .docx  /sw/bin/unzip "-p \"%p\" word/document.xml
FileFilter  .pptx  /sw/bin/unzip "-p \"%p\" ppt/slides/slide*.xml
FileFilter  .xlsx  /sw/bin/unzip "-p \"%p\" xl/worksheets/sheet*.xml

MetaNames swishtitle

# Ask libxml2 to report any parsing errors and warnings or
# any UTF-8 to 8859-1 conversion errors
ParserWarnLevel 9
```

```
swish-e -c ~/.swish-e/swish-office.conf
```

```
Indexing Data Source: "File-System"
Indexing "/Users/jkitchin/Dropbox"

Warning: filter '/sw/bin/unzip' exited with non-zero status: [9]

Warning: filter '/sw/bin/unzip' exited with non-zero status: [9]

Warning: filter '/sw/bin/unzip' exited with non-zero status: [11]

Warning: filter '/sw/bin/unzip' exited with non-zero status: [9]

Warning: filter '/sw/bin/unzip' exited with non-zero status: [9]

Warning: filter '/sw/bin/unzip' exited with non-zero status: [9]

Warning: filter '/sw/bin/unzip' exited with non-zero status: [9]
```

```
Warning: filter '/sw/bin/unzip' exited with non-zero status: [9]

Warning: filter '/sw/bin/unzip' exited with non-zero status: [11]

Warning: filter '/sw/bin/unzip' exited with non-zero status: [11]

Warning: filter '/sw/bin/unzip' exited with non-zero status: [11]

Warning: filter '/sw/bin/unzip' exited with non-zero status: [11]

Warning: filter '/sw/bin/unzip' exited with non-zero status: [9]

Warning: filter '/sw/bin/unzip' exited with non-zero status: [9]

Warning: filter '/sw/bin/unzip' exited with non-zero status: [11]

Warning: filter '/sw/bin/unzip' exited with non-zero status: [9]

Warning: filter '/sw/bin/unzip' exited with non-zero status: [9]

Warning: filter '/sw/bin/unzip' exited with non-zero status: [9]

Warning: filter '/sw/bin/unzip' exited with non-zero status: [9]

Warning: filter '/sw/bin/unzip' exited with non-zero status: [9]

Warning: filter '/sw/bin/unzip' exited with non-zero status: [9]

Warning: filter '/sw/bin/unzip' exited with non-zero status: [11]

Warning: filter '/sw/bin/unzip' exited with non-zero status: [9]

Warning: filter '/sw/bin/unzip' exited with non-zero status: [9]

Warning: filter '/sw/bin/unzip' exited with non-zero status: [9]

Warning: filter '/sw/bin/unzip' exited with non-zero status: [9]

Warning: filter '/sw/bin/unzip' exited with non-zero status: [9]
```

```
Warning: filter '/sw/bin/unzip' exited with non-zero status: [9]

Warning: filter '/sw/bin/unzip' exited with non-zero status: [11]

Warning: filter '/sw/bin/unzip' exited with non-zero status: [9]

Warning: filter '/sw/bin/unzip' exited with non-zero status: [9]

Warning: filter '/sw/bin/unzip' exited with non-zero status: [9]

Warning: filter '/sw/bin/unzip' exited with non-zero status: [9]

Warning: filter '/sw/bin/unzip' exited with non-zero status: [9]

Warning: filter '/sw/bin/unzip' exited with non-zero status: [9]

Warning: filter '/sw/bin/unzip' exited with non-zero status: [11]

Warning: filter '/sw/bin/unzip' exited with non-zero status: [9]

Warning: filter '/sw/bin/unzip' exited with non-zero status: [11]

Warning: filter '/sw/bin/unzip' exited with non-zero status: [11]

Warning: filter '/sw/bin/unzip' exited with non-zero status: [9]

Warning: filter '/sw/bin/unzip' exited with non-zero status: [9]

Warning: filter '/sw/bin/unzip' exited with non-zero status: [9]

Warning: filter '/sw/bin/unzip' exited with non-zero status: [9]

Warning: filter '/sw/bin/unzip' exited with non-zero status: [9]
Removing very common words...
no words removed.
Writing main index...
Sorting words ...
Sorting 1,382,832 words alphabetically
Writing header ...
```

```
Writing index entries ...
  Writing word text: ...
  Writing word text:  10%
  Writing word text:  20%
  Writing word text:  30%
  Writing word text:  40%
  Writing word text:  50%
  Writing word text:  60%
  Writing word text:  70%
  Writing word text:  80%
  Writing word text:  90%
  Writing word text: 100%
  Writing word text: Complete
  Writing word hash: ...
  Writing word hash:  10%
  Writing word hash:  20%
  Writing word hash:  30%
  Writing word hash:  40%
  Writing word hash:  50%
  Writing word hash:  60%
  Writing word hash:  70%
  Writing word hash:  80%
  Writing word hash:  90%
  Writing word hash: 100%
  Writing word hash: Complete
  Writing word data: ...
  Writing word data:   9%
  Writing word data:  19%
  Writing word data:  29%
  Writing word data:  39%
  Writing word data:  49%
  Writing word data:  59%
  Writing word data:  69%
  Writing word data:  79%
  Writing word data:  89%
  Writing word data:  99%
  Writing word data: Complete
1,382,832 unique words indexed.
Sorting property: swishdocpath
Sorting property: swishtitle
```

```
Sorting property: swishdocsize
Sorting property: swishlastmodified
4 properties sorted.
3,598 files indexed.  5,154,618,509 total bytes.  19,696,066 total words.
Elapsed time: 00:01:57 CPU time: 00:01:36
Indexing done!
```

```
1 swish-e -f ~/.swish-e/office-index.swish-e -x '%r\t%p\n' -w alesi hydrogen | grep pptx
```

```
430     /Users/jkitchin/Dropbox/CMU/projects/archive/IAES/amine-sorbents/presentation
388     /Users/jkitchin/Dropbox/CMU/projects/archive/IAES/amine-sorbents/presentation
388     /Users/jkitchin/Dropbox/CMU/archive/group/students/Rich Alesi/walesi-presenta
388     /Users/jkitchin/Dropbox/CMU/projects/archive/IAES/2007-2009/CO2 management th
335     /Users/jkitchin/Dropbox/CMU/projects/archive/IAES/amine-sorbents/presentation
335     /Users/jkitchin/Dropbox/CMU/projects/archive/IAES/amine-sorbents/presentation
335     /Users/jkitchin/Dropbox/CMU/projects/archive/IAES/2007-2009/CO2 management th
260     /Users/jkitchin/Dropbox/CMU/meetings/@archive/2009/kitchin-Dow/@kitchin-dow-o
```