Command Line Converters for Office Formats

Chameleon tools

No-one can avoid the ubiquitous office formats. Command line converters ensure that the contents of these document types remains accessible even if the original application is not. We guide you through the tools available, so you can convert from proprietary formats to more useful open standards.

BY ANDREA MÜLLER

visipix.com

Do your business contacts insist on sending you MS Word documents, although you have repeatedly asked them not to? Not prepared to wait five minutes for an application to load, just to take a quick glance at a StarWriter file?

A fundamental aversion to huge office suites is not the only reason for wanting to convert their proprietary formats into quicker or more generally readable ones.

A number of command line conversion tools are available to tackle this problem, table 1 provides an overview. The table also includes some tools which we will not be discussing in this article.

OpenOffice and StarOffice

Despite the fact that OpenOffice and StarOffice are available on Linux, not everybody installs these large footprint packages. *o3read* [1] provides a quick option for viewing the content of documents created by these packages. The tool can handle documents produced both by the word processing and spreadsheet components.

The *o3read* package includes three output modules with *o3tohtml* producing the best results. The other modules are *o3totxt* for document to text format conversion and *o3read* itself that outputs **XML** tags and their values as a table.

Operating these three programs takes some getting used to at first, they cannot handle *sxw* and *sxc* files. These file formats, which are used by OpenOffice and StarOffice version 6.0 or later for word processing and spread documents respectively are compressed archives. *xml* files are revealed when you extract them. The document content can be found in *content.xml*.

The *o3read* tools leave it to the user to extract the archive files. You will need *unzip* to do so. The following syntax runs *o3tohtml* for the *content.xml* file:

```
unzip -p document.sxw↩
content.xml | o3tohtml >↩
document.html
```

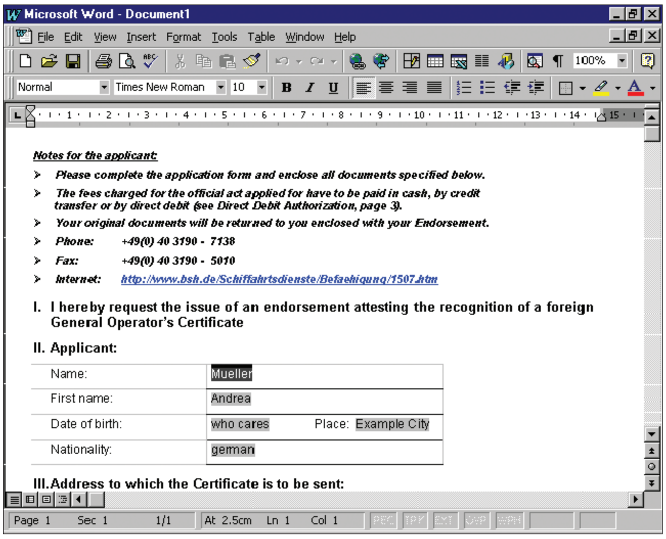
This command extracts the *content.xml* file from a StarWriter or OpenWriter document called *document.sxw* and outputs it to standard output. This output is piped (|) to *o3tohtml*. The arrow (>) then redirects this output to a file called *document.html*.

This is also the best way to launch the other output modules belonging to the *o3read* package. If you discover that non-standard characters are not displaying correctly after conversion, this may be due to the fact that the original document was UTF8 encoded. However, there is a solution for this: the *utf8tolatin1* tool supplied with *o3read*. The complete command is thus as follows:

```
unzip -p table.sxc content.xml↩
| o3tohtml | utf8tolatin1 >↩
table.html
```

GLOSSARY

XML: "Extensible Markup Language" is a meta-language for document type definitions. In contrast to HTML, XML does not only possess tags for formatting and outline definitions, but also semantic tags with purely descriptive content.


Microsoft Word - Document1
File Edit View Insert Format Tools Table Window Help
100%
Normal
Times New Roman
10
Page 1 Sec 1 1/1 At 2.5cm Ln 1 Col 1

Notes for the applicant:

- ***Please complete the application form and enclose all documents specified below.***
- ***The fees charged for the official act applied for have to be paid in cash, by credit transfer or by direct debit (see Direct Debit Authorization, page 3).***
- ***Your original documents will be returned to you enclosed with your Endorsement.***
- ***Phone:*** ***+49(0) 40 3190 - 7138***
- ***Fax:*** ***+49(0) 40 3190 - 5010***
- ***Internet:*** ***http://www.bsh.de/Schiffahrtsdienste/Befaehigung/1507.htm***

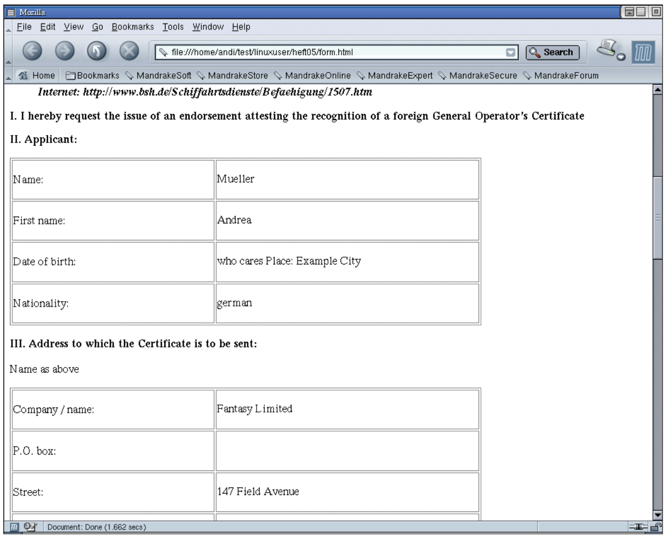
I. I hereby request the issue of an endorsement attesting the recognition of a foreign General Operator's Certificate

II. Applicant:

Name:	Mueller
First name:	Andrea
Date of birth:	who cares Place: Example City
Nationality:	german

III.Address to which the Certificate is to be sent:

Figure 1: The original Word form before conversion


Mozilla
File Edit View Go Bookmarks Tools Window Help
file:///home/andi/test/linuxuser/heft05/form.html
Search
Home Bookmarks MandrakeSoft MandrakeStore MandrakeOnline MandrakeExpert MandrakeSecure MandrakeForum
Document: Done (1.662 secs)

Internet: http://www.bsh.de/Schiffahrtsdienste/Befaehigung/1507.htm

I. I hereby request the issue of an endorsement attesting the recognition of a foreign General Operator's Certificate

II. Applicant:

Name:	Mueller
First name:	Andrea
Date of birth:	who cares Place: Example City
Nationality:	german

III. Address to which the Certificate is to be sent:

Name as above

Company / name:	Fantasy Limited
P.O. box:	
Street:	147 Field Avenue

Figure 2: After wvHtml conversion

The output from *o3tohtml* is piped to *utf8tolatin1*, where character set conversion is performed, before the output is stored in a file.

If you consider the way the program works, its weakness soon becomes apparent. *sxw* and *sxc* archives also contain other *xml* files with meta-information on the document; they may also contain subfolders with embedded images. All of this formatting information is lost during conversion.

However, some formatting, such as italics or underlined text are retained along with the basic structure of the document.

o3totxt is no use at all for spreadsheets. As the document is output line by line, you completely lose track. However, *o3tohtml* provides fairly useful results in this discipline.

Word Conversions

All it takes to make your friends and colleagues stand back in amazement is a simple sentence like "I don't use Word". *Microsoft Word* is included with nearly every new PC you can purchase and is the de facto standard in the world of word processing. Due to its pervasiveness, it is hardly surprising that the bulk of converters are written with this Office application in mind. Although Microsoft has not published details of this proprietary format, *Word* converters have nothing to be ashamed of. The most popular application in this field is probably *wv* [2].

The package includes over 15 individual applications. Some of them provide purely informational output, such as *wvVersion*, which finds out what Word version was used to create a document. *wvSummary* outputs details such as the title and author for MS Office documents. Most of the applications are actually output modules. Their names are self-explanatory, for example *wvHtml* converts Word files to HTML format, whereas *wvLatex* converts them to *tex* files. Additional output modules handle conversion to PostScript, RTF, or DVI, with some modules requiring external utilities. Calling these tools is extremely simple:

```
wvHtml letter.doc letter.html
```

Don't forget to change the file suffix for the output file to reflect the output module used.

wvWare provides access to password protected Word files. The following command:

```
wvWare -p secret letter.doc
> letter.html
```

Table 1: Command Line Office Converters

Name	Scope	Quality of Results
o3read	Converts *sxw* and *sxc* files to text or HTML.	See text.
sxw2html [8] and *sxw2txt* [9]	Converts *sxw* files to text or HTML using shell scripts. Launches *lynx* for HTML display.	Structure retained, formatting lost. Non-standard characters not displayed correctly due to lack of character table function.
sdw2txt [10]	Converts StarWriter 5.x files to text.	Document structure partially retained. Some sections of document lost, e.g. tables or sender's address in letters. Only for simple documents.
wv	Converts Word files to various formats.	See text.
antiword	Converts Word files to text or Postscript.	See text.
catdoc [11]	Converts Word files to text. The program provides a viewer based on the Tk library for output functionality and a tool for converting Excel files to comma-separated lists.	Document structure partially retained. However, not particularly good for forms or tables, as the out put is too cluttered.
xlhtml	Converts Excel documents to HTML files, conversion of specific areas to clear text also possible. *ppthtml* extracts text from PowerPoint files.	See text.
chmlib	Converts typical Windows help file files to *chm* format using multiple HTML files.	See text.
wp2x [12]	Converts WordPerfect 5.1 files to various formats.	Document structure and formatting mainly retained. Restricted usefulness due to rarity of original format.

converts the a file called *document.doc*, which is password protected with the *secret* password, to HTML. As *wvWare* uses standard output, in contrast to the individual modules, you will probably want to redirect this output into a file.

The *wv* programs all returned useful results (Figures 1 and 2), retaining both the document structure, graphics and formatting, provided the specified output format supports this functionality.

Depending on the document you are converting, it may be worthwhile experimenting with various output formats. *wvPDF* provides the best results for embedded graphics, although it does remove any Euro characters the document may contain. Thus, *wvHtml* may be preferable for documents containing price lists, for example.

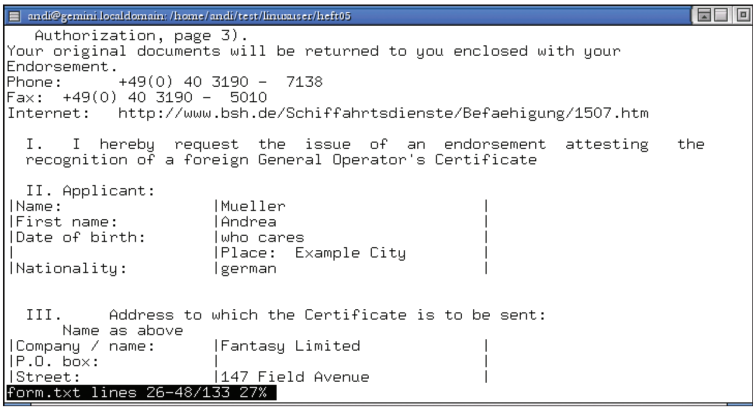
Worthy Opponent

antiword [3,4] is another popular contender that converts *doc* files either to text or PostScript format. The text conversion feature is particularly useful as it extracts as much information as possible from a format. *antiword* retains the original structure wherever possible (Figure 3) and even uses a placeholder (*[pic]*) to indicate where pictures occurred in the original document.

To additionally view the images, you can convert the original document to PostScript format as follows:

```
antiword -p a4 ➥
-i 3 document.doc ➥
> document.ps
```

The *-p* flag specifies the page format. No matter what image level you use (parameter *-i*), *antiword* tends to distort some graphics. If the document contains Euro characters, you can use the *-m 8859-15* flag to specify the character set.



```
   Authorization, page 3).
Your original documents will be returned to you enclosed with your
Endorsement.
Phone:      +49(0) 40 3190 -  7138
Fax:  +49(0) 40 3190 -  5010
Internet:   http://www.bsh.de/Schiffahrtsdienste/Befaehigung/1507.htm

  I.   I  hereby  request  the  issue  of  an  endorsement  attesting   the
  recognition of a foreign General Operator's Certificate

  II. Applicant:
|Name:                 |Mueller                      |
|First name:           |Andrea                       |
|Date of birth:        |who cares                    |
|                      |Place:  Example City         |
|Nationality:          |german                       |


  III.     Address to which the Certificate is to be sent:
      Name as above
|Company / name:       |Fantasy Limited              |
|P.O. box:             |                             |
|Street:               |147 Field Avenue             |
form.txt lines 26-48/133 27%
```

Figure 3: A text version courtesy of antiword

If you do not want to store the document, but simply need to create a hard copy, you can enter the following

```
antiword -p a4 -i 3➥
document.doc | lp
```

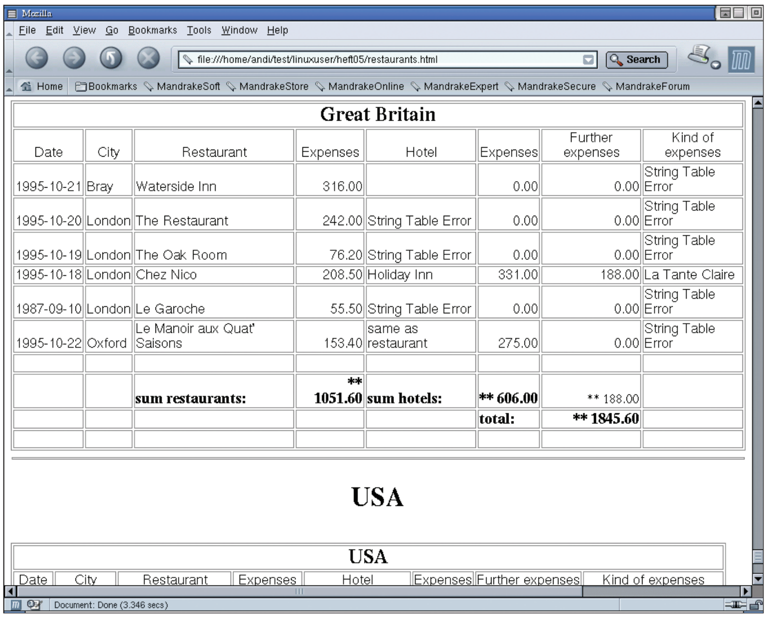
to send the document to your printer.

Spreadsheet Converters

If you are a regular reader of Linux Magazine, you may remember reading about *xlhtml* [5,6] a while back; this is a tool for converting Excel files to HTML. The following command:

```
xlhtml table.xls > table.html
```

creates HTML that any browser can display, and numerous parameters additionally allow you to parse individual rows (*-xr*), columns (*-xc*) or pages (*-xp*).


Mozilla
File Edit View Go Bookmarks Tools Window Help
file:///home/andi/test/linuxuser/heft05/restaurants.html
Search
Home
Bookmarks
MandrakeSoft
MandrakeStore
MandrakeOnline
MandrakeExpert
MandrakeSecure
MandrakeForum
Document: Done (3.346 secs)

Great Britain

Date	City	Restaurant	Expenses	Hotel	Expenses	Further expenses	Kind of expenses
1995-10-21	Bray	Waterside Inn	316.00		0.00	0.00	String Table Error
1995-10-20	London	The Restaurant	242.00	String Table Error	0.00	0.00	String Table Error
1995-10-19	London	The Oak Room	76.20	String Table Error	0.00	0.00	String Table Error
1995-10-18	London	Chez Nico	208.50	Holiday Inn	331.00	188.00	La Tante Claire
1987-09-10	London	Le Garoche	55.50	String Table Error	0.00	0.00	String Table Error
1995-10-22	Oxford	Le Manoir aux Quat' Saisons	153.40	same as restaurant	275.00	0.00	String Table Error
		sum restaurants:	**** 1051.60**	**sum hotels:**	**** 606.00**	** 188.00	
					total:	**** 1845.60**	

USA

USA

Date	City	Restaurant	Expenses	Hotel	Expenses	Further expenses	Kind of expenses

Figure 4: An Excel file converted by xlhtml

If you use one of these scope parameters, you can either specify text output (*-asc*), or export the results to a comma-separated (*-csv*) list. The latter can then be imported into a MySQL database, for example. The quality of the results is mediocre and largely depends on the original document.

xlhtml converts simple documents perfectly, but the results are far less so, if the lines contain the results of calculated operations.

Each of these cells is surrounded by two *, and the footer informs the user that the results may not be correct, although this warning proved unnecessary judging from our test results. Incidentally the converter uses three asterisks to indicate formats it does not support.

If you cannot read the information in the HTML output provided by *xlhtml*, as the program has produced black print on a black background, for example, you might like to try the *-nc* ("*no color*") flag as a last resort. This flag stipulates monochrome output. The reason for this strange behavior would seem to be the creativity of the original author of the document, as the problem was reproducible when documents contained cells where different background colors had been used.

Strangely enough, the blackout did not only effect the areas which previously had colored backgrounds, but areas whose backgrounds had originally been white.

xlhtml is not thrown by larger Excel spreadsheets with one or two minor exceptions. Our test document was an Excel spreadsheet that contained nine tables, each one of which had over 2000 records. The following command

```
xlhtml -nc rst.xls ➥
> shops.html
```

created a HTML file that displayed the individual tables one after the other in an orderly fashion (Figure 4).

The *String Table Error* that occurred in some lines is nothing to worry about, as the cells were empty in the original document. The error is probably caused by the fact that the original document contained type formatting for empty cells.

If you simply need a quick glance at the content of an Excel file, *xlhtml* is definitely useful, as it does not drop cell content or display it incorrectly. However, more complex documents may cause some concern due to the messages xhtml produces, particularly if you cannot check the conversion results by referring to the original.

xlhtml‘s author also provides a tool called *ppthtml* for PowerPoint files, however, this can be regarded as a statement of intent rather than an actual conversion tool at present. The tool simply extracts the text from *ppt* files, without retaining even a modicum of formatting information. Thus, *ppthtml* is unacceptable, particularly when one considers that presentations typically contain graphics and diagrams.

Box 1: Avoiding A Confusion of Formats

PDF is a format that more or less any system can read. Thus, the "*Portable Document Format*" is perfect for propagating documents for read-only access. One simple way of creating PDF documents is to print the output of an Office application to a file and convert the results to *pdf* format using *ps2pdf*. Unsurprisingly, there are also command line tools to perform this task, for example, using the following command:

```
ps2pdf letter.ps outputletter.pdf
```

ps2pdf is part of the *ghostscript* package and thus typically installed on most distributions.

Users of the latest KDE versions have an even easier option, as they can print to *pdf* file format directly. If you are one of those lucky users, simply choose the *Print to File (PDF)* entry in the printer dialog instead of your normal printer.

GLOSSARY

chm: *Abbreviation for "Compiled HTML Manual". The advantage of this help file format for newer Windows in comparison to pure HTML is the fact that it is a compressed format and thus has a smaller footprint. Additionally, Windows provides extended functionality in the form of navigational aids like trees.*

Illegible Help ...

The profusion of ***chm*** files is constant source of irritation. Whether for brochures or manuals for new hardware – more and more manufacturers seem to think that this format is a good thing for their customers. Whereas Windows systems provide native tools, Linux and Windows 95 users, are initially left up the creek.

chmlib [7] is not a true converter, but an interface that provides access to *chm* files. Besides the library, which will be particularly interesting for programmers, the author provides a collection of, admittedly unpolished, sample applications.

Installing the library can be quite tricky: the first pitfall is the *Makefile* which assumes the *gcc-3.2* compiler. If you do not use this version, you should change the following lines

```
CC=gcc-3.2
LD=gcc-3.2
```

to

```
CC=gcc
LD=gcc
```

If you have installed *gcc*, ensure that the compiler is in the shell's search path, if only in the form of a link to the executable.

You can then launch *make* and *make install* to compile and install the file, ensuring that you are *root* before performing the latter step. Having completed these steps, you will still not be able to access the sample applications; to do so you need to change to the source directory for *chmlib* and enter the following

```
make examples
```

The programs created by this step *enum_chmLib*, *test_chmLib*, *extract_chmLib*, and *enumdir_chmLib* should then be copied by the superuser *root* to the */usr/local/bin* directory.

...made legible

To extract the content of a *chm* file, enter

```
extract_chmLib help.chm↩
 outputdirectory
```

This creates a HTML file with the content of the *chm* file in the output directory, and although a navigational index is normally not available, at least the links in the HTML files, and images are displayed correctly.

The *enum_chmLib* program lists the content of an *chm* file. This shows that a file of this type has a kind of tree structure internally. If you note one of the paths, you can enter a command such as

```
test_chmLib help.chm /↩
requiredfile.htm output.html
```

to extract individual files. This procedure does mean that any links in the exported file will point to empty space as the target files do not exist.

All this effort sometimes makes you forget that none of these contortions would be necessary, if authors were more considerate. If you want to set a good example, and help reduce the Babylonian confusion of formats, you can refer to Box 1 for tips on creating PDF documents, which are displayable on more or less any system. ■

INFO

[1] o3read: *http://siag.nu/o3read/*

[2] wv: *http://wvware.sourceforge.net/*

[3] antiword: *http://www.winfield.demon.nl/*

[4] Christian Perle: "Against It!", Linux Magazine, Issue 15, p82

[5] Christian Perle: "eXcellent", Linux Magazine, Issue 23, p84

[6] xlhtml: *http://chicago.sourceforge.net/xlhtml*

[7] chmlib: *http://66.93.236.84/~jedwin/projects/chmlib*

[8] sxw2html: *http://massaint.com/sxw2html.tar.gz*

[9] sxw2txt: *http://massaint.com/sxw2txt.tar.gz*

[10] sdw2txt: *http://sdw2txt.sourceforge.net/*

[11] catdoc: *http://www.45.free.net/~vitus/ice/catdoc/ver-0.9.html*

[12] wp2x: *ftp://ftp.penguin.cz/pub/users/mhi/wp2x/*