

CONVERSIÓN DE DOCUMENTOS Y MARKDOWN

versión 1 23-6-2015

El formato en los documentos se indica siempre con marcas. Lo más habitual y deseable es que estos documentos sean archivos de texto plano puesto que así son accesibles sin restricciones, como por ejemplo html, markdown, xml, LaTeX, etc. Incluso odt y docx son archivos de texto, solo que están comprimidos con el formato zip. Formatos más antiguos como doc, sxw, wp, etc. codifican la información en binario en vez de en ASCII por lo que si no se conoce esa codificación resulta muy difícil crear programas que los lean (realmente un fichero en ASCII también está en binario, como no podía ser de otra manera, lo que ocurre es que el código ASCII es el estándar para codificar letras, números y caracteres).

- 1 En la carpeta `convertir_documentos_y_markdown` descomprime los ficheros con extensión `odt` y `docx` y busca en qué fichero se encuentra el texto del documento.
- 2 Extrae el texto de los ficheros anteriores en los ficheros `pdf.txt`, `odt.txt` y `doc.txt`.

Para ello usa los siguientes comandos:

```
pdftotext fichero.pdf
odt2txt fichero.odt > fichero.txt
antiword fichero.doc > fichero.txt
```

Si es necesario instala estos paquetes: `odt2txt`, `poppler-utils` (para `pdftotext`) y `antiword`.

- 3 Con `Pandoc` podemos transformar un fichero, con las consiguientes limitaciones, entre diferentes lenguajes de marcas, por ejemplo: html, LaTeX, markdown, odt, epub, etc.
 - markdown a html: `pandoc -s fichero.txt -o fichero.html`
 - markdown a odt: `pandoc fichero.txt -o fichero.odt`
 - LaTeX a html:


```
latex2html -html_version 4.0,unicode -split 0 -nonavigation
-show_section_numbers fichero.tex
```
 - etc.
- 4 Transforma, con `pandoc`, el fichero `manual_markdown.txt` a `odt` y `html` y comprueba el resultado.
- 5 Con la ayuda del manual anterior escribe en markdown un documento de tema libre que contenga título, subtítulo, párrafos, enlaces, negritas, cursivas, viñetas, listas numeradas y texto de ancho fijo (para escribir código). Posteriormente expórtalo a `odt` y `html`.
- 6 Transforma, con `latex2html`, el fichero `apuntes_TIC_1_bach.tex` y comprueba el resultado.