Luchando contra datos no amigables

Por David Cabo

La cantidad de datos disponibles en formatos abiertos y con licencias que permiten la reutilización es cada mayor, gracias a la creciente sensibilización de las administraciones públicas. Sin embargo, es todavía muy frecuente encontrar información en formatos que dificultan su uso, como PDFs, páginas HTML o bases de datos Access. Existen algunas herramientas que facilitan el trabajo, incluso sin necesidad de grandes conocimientos técnicos.

PDFs

El formato PDF es uno de los más populares a la hora de publicar información, a pesar de que no está pensado para la reutilización de dicha información: al convertir a PDF mucha de la estructura original de la información se pierde, lo cual complica su uso posterior. Dentro de los ficheros PDF es fundamental distinguir, a la hora de trabajar con ellos, dos tipos, en función de su origen:

- PDFs generados a partir del escaneo de un documento físico: en estos casos el PDF es una foto (compuesta por puntos blancos y negros), y no contiene el texto original. Puedes saber que es una imagen haciendo click en el documento. Verás que se te ilumina todo en azul.

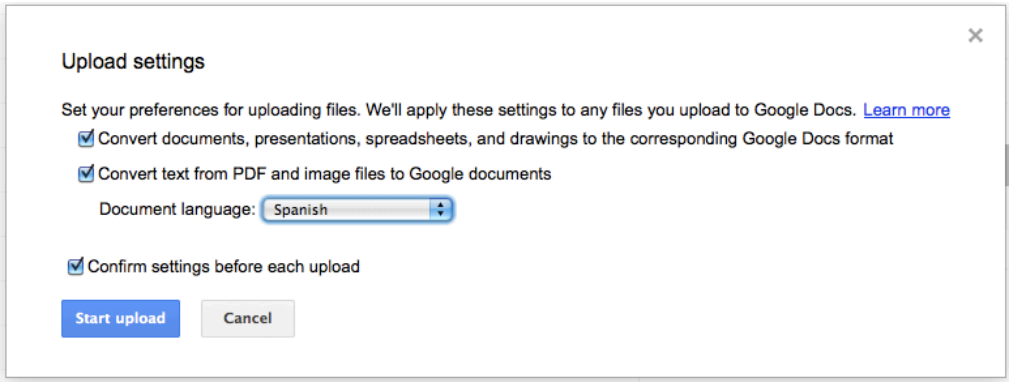
- PDFs generados "digitalmente", a partir de programas como Word, Excel u Open Office. Se distinguen porque al abrirlos podemos seleccionar un fragmento de texto, y también podemos hacer búsquedas por palabras.

PDFs Escaneados

Para extraer la información de este tipo de ficheros es necesario emplear programas de "Reconocimiento Óptico de Carácteres" (OCR en inglés), que intentan reconocer el texto original a partir de "las manchas negras sobre fondo blanco" que aparecen en la imagen. La precisión del resultado depende de la nitidez de la imagen original, y de la calidad del programa empleado, pero nunca es del 100%, y requiere una revisión "manual" posterior.

Existen programas comerciales como Abbyy FineReader, Acrobat Pro o Solid, así como alternativas de código abierto OCRopus y Tesseract, aunque estas últimas no son demasiado fáciles de instalar y utilizar.

Una alternativa gratuita, sencilla, y de calidad generalmente aceptable, es usar la conversión integrada en Google Docs. Al subir un fichero PDF a este servicio podemos indicarle que intente convertir imágenes a texto:



La opción "Convert text from PDF and image files to Google documents" tiene que estar activada al subir un fichero PDF a Google Docs. Elegir el idioma del documento ayuda a mejorar la calidad del resultado.

En el caso de que ninguna de estas operaciones funcione y no cuentes con la ayuda de un programador que te pueda ayudar a extraerlos, la única opción de trabajar con los datos será picarlos manualmente en una hoja de cálculo. Es recomendable que cuando hagas entrada manual de datos lo hagas dos veces, por ejemplo, poniendo a dos personas a meter los datos y luego comparando el resultado. Así evitarás errores.

Otra opción es que pidas ayuda a tus lectores con la entrada de datos. La campaña de Twitter #adoptaunsenador pidió a la gente que metiese los datos de las declaraciones de bienes de los senadores en una hoja de cálculo. En menos de 48 horas, el trabajo estaba hecho. Por supuesto, siempre que pidas ayuda a tus lectores/audiencia tienes que revisar el resultado, pero eso lleva menos tiempo que hacerlo tú mismo desde cero.

PDFs Digitales

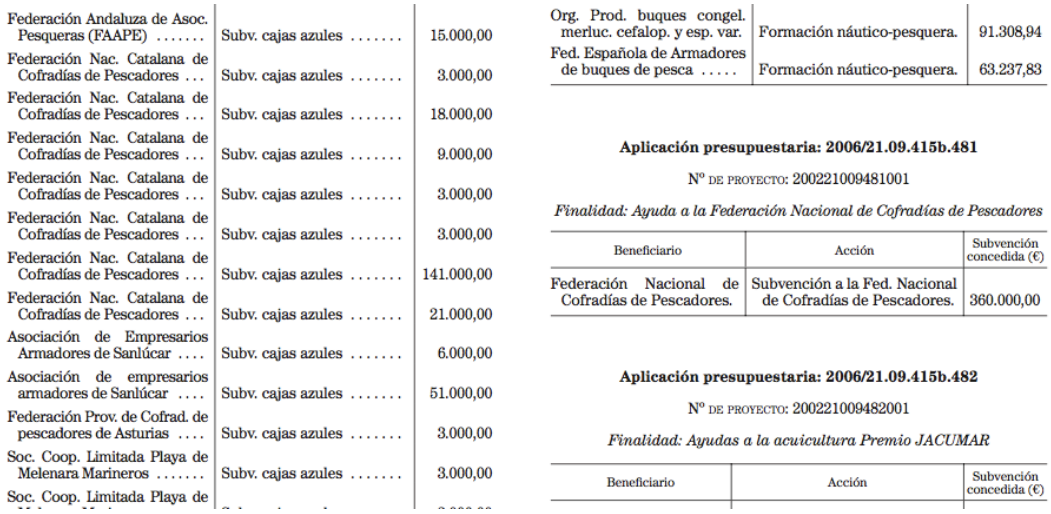
Para convertir un PDF "digital" a un documento con el que podamos trabajar (es decir, en una hoja de cálculo, como Excel, por ejemplo), podemos usar programas comerciales (PDF Converter, Nitro PDF, Acrobat Pro...), pero existen también servicios webs gratuitos, entre los que destacan:

- cometdocs.com
- pdftoexcelonline.com
- zamzar.com

El funcionamiento en los tres casos es similar: subimos el fichero PDF a convertir a su web, y al cabo de unos minutos nos llegará a la dirección de correo que suministremos el fichero convertido. Es muy importante que revises los datos y los limpies, porque no siempre la conversión funciona perfectamente.

PDFs Digitales Complejos

Los servicios de conversión automáticos dan buenos resultados con ficheros PDF "sencillos" que contienen tablas bien delimitadas, pero generalmente no son capaces de convertir correctamente casos más complejos como una página de un boletín oficial, en la que el texto aparece en varias columnas, con líneas partidas y múltiples encabezamientos, como en el siguiente ejemplo:



Federación Andaluza de Asoc. Pesqueras (FAAPE)	Subv. cajas azules	15.000,00
Federación Nac. Catalana de Cofradías de Pescadores ...	Subv. cajas azules	3.000,00
Federación Nac. Catalana de Cofradías de Pescadores ...	Subv. cajas azules	18.000,00
Federación Nac. Catalana de Cofradías de Pescadores ...	Subv. cajas azules	9.000,00
Federación Nac. Catalana de Cofradías de Pescadores ...	Subv. cajas azules	3.000,00
Federación Nac. Catalana de Cofradías de Pescadores ...	Subv. cajas azules	3.000,00
Federación Nac. Catalana de Cofradías de Pescadores ...	Subv. cajas azules	141.000,00
Federación Nac. Catalana de Cofradías de Pescadores ...	Subv. cajas azules	21.000,00
Asociación de Empresarios Armadores de Sanlúcar	Subv. cajas azules	6.000,00
Asociación de empresarios armadores de Sanlúcar	Subv. cajas azules	51.000,00
Federación Prov. de Cofrad. de pescadores de Asturias	Subv. cajas azules	3.000,00
Soc. Coop. Limitada Playa de Melenara Marineros	Subv. cajas azules	3.000,00
Soc. Coop. Limitada Playa de		

Org. Prod. buques congel. merluc. cefalop. y esp. var.	Formación náutico-pesquera.	91.308,94
Fed. Española de Armadores de buques de pesca	Formación náutico-pesquera.	63.237,83

Aplicación presupuestaria: 2006/21.09.415b.481

Nº DE PROYECTO: 200221009481001

Finalidad: Ayuda a la Federación Nacional de Cofradías de Pescadores

Beneficiario	Acción	Subvención concedida (€)
Federación Nacional de Cofradías de Pescadores.	Subvención a la Fed. Nacional de Cofradías de Pescadores.	360.000,00

Aplicación presupuestaria: 2006/21.09.415b.482

Nº DE PROYECTO: 200221009482001

Finalidad: Ayudas a la acuicultura Premio JACUMAR

Beneficiario	Acción	Subvención concedida (€)

En estos casos es necesario:

1. Convertir el fichero PDF a formato texto con una herramienta gratuita, pdftotext.

2. Escribir un programa a medida (en Perl, Python, Ruby...) para extraer los datos. Para eso probablemente necesitarás la ayuda de un programador.

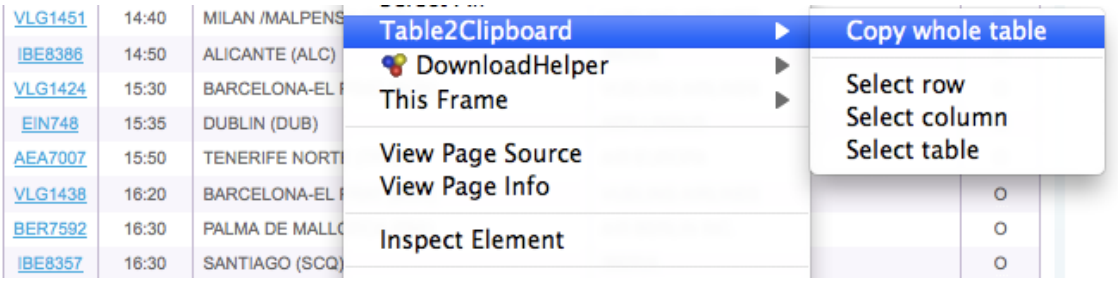
Páginas HTML

Si queremos extraer una tabla de una página web la opción de copiar en el navegador y pegar en la hoja de cálculo no suele funcionar correctamente, ya que generalmente se pierde la estructura de la tabla (es decir, el hecho de que los datos están distribuidos en filas y columnas). Pero podemos ayudarnos de extensiones en nuestros navegadores:

- Table2Clipboard en Firefox. *(Una vez bajado el add-on hay que reiniciar el navegador para que se instale)*

- TableCapture para Chrome

Table2Clipboard, por ejemplo, añade una opción al menú del botón derecho de nuestro navegador, de forma que al hacer click con el botón derecho sobre una tabla podemos copiar todo su contenido manteniendo la estructura original:



Luego, simplemente tienes que ir a tu hoja de cálculo y darle al botón de pegar. Asegúrate bien de que se muestran todos los resultados de la tabla, y no hay varias páginas con datos. En este caso, tendrías que ir repitiendo el proceso página por página.

Otra alternativa es utilizar la [función para importar datos de páginas web en Google Docs], que permite extraer información de listas o tablas. Por ejemplo, si queremos extraer la [tabla con los ganadores del Premio Nobel de Literatura según la Wikipedia], escribimos en cualquier celda de la hoja de cálculo que tengamos abierta en Google Docs:

```
=importhtml("http://es.wikipedia.org/wiki/
Anexo:Ganadores_del_Premio_Nobel_de_Literatura","table",2)
```

dónde el primer parámetro es la dirección de la página a utilizar, el segundo indica que queremos extraer una tabla, y el tercero especifica que queremos la segunda tabla de la página (ten en cuenta que en este caso el índice de contenidos también es una tabla). A continuación Google Docs irá a la página indicada, buscará la segunda tabla y rellenará la hoja de cálculo con la información encontrada. Nos quedaría algo así:

	A	B	C	D	E	F
1	Año	Imagen	Laureado	País	Idioma	Motivación
2	1901		Sully Prudhomme	Francia	Francés	«en reconocimiento especial a su composición poética, lo cual da pruebas de un elevado idealismo, una perfección artística y una rara combinación de las cualidades tanto del corazón como del intelecto».[11]
3						«el más grandioso maestro con vida del arte de la escritura histórica, con una especial

Bases de datos propietarias (Access)

Microsoft Access es una base de datos popular en las administraciones públicas, que guarda la información en ficheros con extensión .MDB, que requieren disponer del programa original para ser leídos. Aún peor, Microsoft Access sólo está disponible en Windows.

Es posible extraer toda la información de ficheros .MDB usando la herramienta gratuita y de código abierto, MDB Tools. Una vez instalado el programa, y suponiendo que el fichero se llama "datos.mdb", podemos ejecutar desde la línea de comandos:

```
mdb-export datos.mdb nombre_de_tabla
```

para convertir el contenido de la tabla "nombre_de_tabla" en un fichero de texto CSV (Comma Separated Values; valores separados por comas). (Más información.)