Angel Pablo Hinojosa

Obtención de Datos

Open Data

Esto va de publicar datos

Pero (obviamente) también de obtenerlos

Tipos de datos (por su origen)

Datos primarios:

Datos de nuestras propias fuentes: Contabilidad, informes, investigaciones…

Pero también usuarios, tráfico, logs, mails, redes, aplicaciones biométricas...

Tipos de datos (por su origen)

Datos de terceros:

Fuentes abiertas, proveedores de datasets…

(Por ejemplo `http://opendata.ugr.es/`)

...y scraping

Tipos de datos (por su licencia)

Licencias libres

- Que permiten la reutilización

Licencias privativas o cerradas

- Que no la permiten

Licencias

Public Domain Dedication and License (PDDL)

- Como “Dominio público”
- Permite reutilización sin condiciones.

Licencias

Attribution License (ODC-By)

- Como la anterior, pero exige atribución

Licencias

Open Database License (ODC-OdbL)

- También exige atribución
- Clausula “Copyleft”

Licencias

Más recursos legales en:

http://opendatacommons.org

Tipos de datos (por su formato)

No todos los formatos son iguales:

- Formatos manipulables y no manipulables
- Formatos abiertos y cerrados

Formatos abiertos y cerrados

- Formatos abiertos son estándares y permiten sus uso.

- Formatos cerrados no permiten el libre acceso al contenido.

Manipulables y no manipulables

- ¿Puede accederse a ellos de forma automatizada?
- ¿Hace falta revisión manual?
- ¿Hay que copiarlos a mano?

El mejor de los casos

Datos accesibles, en un formato abierto y manipulable, y con licencia libre que permita reutilización.

El peor de los casos

Datos ilocalizables, en un formato cerrado o no manipulable, sin licencia o con una que no permita reutilización.

El infierno de los formatos

Son tus amigos:

- JSON
- CSV

El infierno de los formatos

…y el PDF como ejemplo de todo lo malo

`http://sl.ugr.es/bigdataPDF`

El infierno de los formatos

Ejemplos:

- PRIMER_TRIMESTRE_2015.pdf
- presupuesto_ayuntamiento_2015.pdf
- presupuesto_parque_ciencias_2015.pdf
- tabla.pdf

Herramientas online:

Multiconversor:

`http://www.cometdocs.com/`

(con Limitaciones si no pagas)

Herramientas online:

Extraer tablas de PDF:

`https://pdftables.com/`

(Sólo lee las tablas)

Herramientas online:

PDF a Excell (y otros):

`https://www.pdftoexcelonline.com/`

Lo envía a tu correo

Herramientas online:

OCR de PDF

`http://www.onlineocr.net/`

(una sola pagina)

Herramientas online:

OCR de PDF

`http://free-online-ocr.com/`

(hasta 10 páginas)

NOTA: Ningún OCR es bueno

Lento, complejo, propenso a errores

Requiere supervisión y revisión posterior

Huye del OCR como de la peste

En local

Tabula

`http://tabula.technology/`

(fácil y simple, pero son OCR)

Conversor de formatos

Pandoc

`pandoc.org/`

(no convierte DE PDF)

Para programadores (Python)

pdfminer

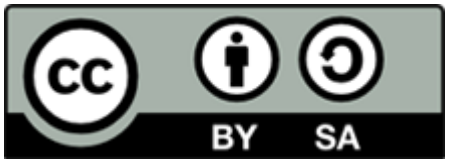
`https://euske.github.io/pdfminer/`

pypdfocr

`https://pypi.python.org/pypi/pypdfocr`

Gracias

(Ruegos y preguntas)



http://www.psicobyte.com/descargas/ODPAS2.pdf