



INTERNATIONAL
OPEN DATA
HACKATHON



Scraping con programación y sin ella, extracción de datos de PDFs, fuentes de datos comunes...

Renato Luis Ramírez Rivero

Contenido

- Quien soy.
- Presentación.
- Ejercicio.
- Preguntas

odinetnoC

- Yo pregunto
- Ejercicio
- Presentación
- Quien soy.

Todo está

<https://github.com/renatolrr/OpenDataDay2015>

Contenido

- Scraping (definición y conocimientos previos).
- Scraping para no programadores.
- Extracción de datos en PDFs.
- Fuentes de datos comunes.
- Conclusión.

Scraping (definición y conocimientos previos)

Definición

Según Wikipedia:

“Web Scraping es una técnica utilizada mediante programas de software para extraer información de sitios web”

Buenas costumbres en scraping

- Definir previamente lo que se busca. Planificar.
- Copiar web.
- Conocimientos previos de programación.
- Conservar fuentes.
- Guardar los datos utilizando estándares.

Aspectos legales

“No estarán autorizadas la extracción y/o reutilización repetidas o sistemáticas de partes no sustanciales del contenido de una base de datos que supongan actos contrarios a una explotación normal de dicha base o que causen un perjuicio injustificado a los intereses legítimos del fabricante”.

- Artículo 133 del Real Decreto Legislativo 1/1996, de 12 de abril, por el que se aprueba el Texto Refundido de la Ley de Propiedad Intelectual.

Planificar

- Open project
- Redmine
- Github

Copia web: HTTrack

Actividades Iceweasel mar 19:49

Download HTTrack Website Copier 3.48-19 - HTTrack Website Copier - Free Software Offline Browser (GNU GPL) - Iceweasel

File Edit View History Bookmarks Tools Help

Download HTTrack ...

www.httrack.com/page/2/en/index.html

slideshare

HTTrack WEBSITE COPIER

Free software offline browser

About Download Manual Forum Blog Information Français

Advertisement:

Business File Sharing

Business-class Secure File Sharing. 4 STARS by PC Magazine. Free Trial.

Download HTTrack Website Copier 3.48-19

- **READ THIS BEFORE DOWNLOADING:** This [free software](#) program is not guaranteed, and is provided "as is".

Platform	Choose file to download	Version
Windows 2000/XP/Vista/Seven/8 installer version WinHTTrack (also included: command line version)	httrack-3.48.19.exe [alternate site]	3.48-19 3.96 MiB (4151768 B) (28/Jul/2014)

Copia web: HTTrack

- Descarga:<http://www.httrack.com/page/2/en/index.html>
- Manual:<http://www.httrack.com/html/fcguide.html>
- `httrack "http://lujoyglamour.net/" -O
"/tmp/www.all.net" "+.all.net/" -v`

Html5, W3c

- Firebug
- HTML Regex Data Extractor
- Clearly

Perl, Python, Java, Php, R...



Instant Web Scraping with Java 26 agosto 2013

de Ryan Mitchell

Versión Kindle

EUR 13,99

Disponible para descargar ya

Tapa blanda

EUR 22,71 ✓Premium

Recíbelo el **martes, 24 febrero**

Más opciones de compra

EUR 22,71 usado y nuevo (4 ofertas)



Web Scraping with Python: A Comprehensive Guide to Data Collection Solutions

25 mayo 2015

de Ryan Mitchell

Tapa blanda

EUR 24,70 ✓Premium

Disponible en pre-venta. Este producto saldrá a la venta el 25 mayo 2015

Envío GRATIS disponible (ver página del producto).

Perl, Python, Java, Php, R...



Phparchitect's Guide to Web Scraping

septiembre 2010

de Matthew Turland

Tapa blanda

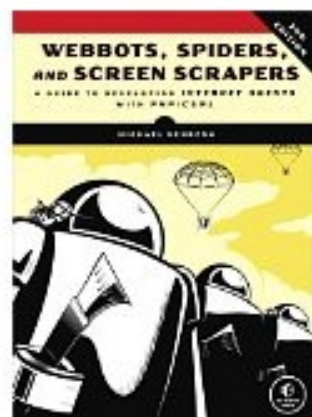
EUR 37,59 ✓Premium

Sólo hay 3 en stock. Cómpralo cuanto antes.

Más opciones de compra

EUR 34,28 usado y nuevo (10 ofertas)

Envío GRATIS disponible (ver página del producto).



Webbots, Spiders, and Screen Scrapers: A Guide to Developing Internet Agents with PHP/CURL

4 marzo 2012

de Michael Schrenk

Versión Kindle

EUR 17,99

Disponible para descargar ya

Tapa blanda

EUR 31,62 ✓Premium

Reservado el precio 10,45 euros

Perl, Python, Java, Php, R...



Instant PHP Web Scraping 26 julio 2013
de Jacob Vahrdt

Tapa blanda
EUR 17,77  Premium
Recíbelo el **martes, 24 febrero**

Más opciones de compra
EUR 17,77 **usado y nuevo** (3 ofertas)

Versión Kindle
EUR 10,99

Disponible para descargar ya

Envío GRATIS disponible (ver página del producto).



Spidering Hacks: 100 Industrial-Strength Tips & Tools 7 noviembre 2003
de Morbus Iffy Tara Calishain

Tapa blanda
EUR 22,80  Premium
Sólo hay 3 en stock. Cómpralo cuanto antes.

Más opciones de compra
EUR 1,35 **usado y nuevo** (17 ofertas)

Envío GRATIS disponible (ver página del producto).



Perl & LWP (Classique Us) 30 junio 2002
de Sean M. Burke

Tapa blanda
EUR 37,05  Premium
Sólo hay 3 en stock. Cómpralo cuanto antes.

Más opciones de compra
EUR 12,02 **usado y nuevo** (13 ofertas)

Versión Kindle
EUR 20,32

Disponible para descargar ya

Envío GRATIS disponible (ver página del producto).

Python

Scraping express por Serafín Velez Barrera

Scraping Web Pages with Scrappy - YouTube

Scraping express
El arte de recuperar datos

Serafín Vélez Barrera
serafa12000@gmail.com – @seravb



oficina de
software
libre



Perl

<https://github.com/oslugr/datos-ugr/tree/master/scripts>

- By Óscar Zafra

```
if($file_data =~ /\(-?\d{1,3}\.\d{3}(\,\d{2})?\)/u){  
#Quitamos los puntos a los miles  
$file_data =~ s/(\d{1,3})\.\d{3}\.\d{3}\.\d{3}(\,\d{2})?/$1$2$3$4$5/g;  
$file_data =~ s/(\d{1,3})\.\d{3}\.\d{3}\.\d{3}(\,\d{2})?/$1$2$3$4$5/g;  
$file_data =~ s/(\d{1,3})\.\d{3}\.\d{3}(\,\d{2})?/$1$2$3$4/g;  
$file_data =~ s/(\d{1,3})\.\d{3}\.\d{3}(\,\d{2})?/$1$2$3$4/g;  
$file_data =~ s/(\d{1,3})\.\d{3}(\,\d{2})?/$1$2$3/g;  
$file_data =~ s/(\d{1,3})\.\d{3}(\,\d{2})?/$1$2$3/g;  
}
```

Modern Perl

```
use Modern::Perl; use autodie;

use LWP::Simple; use Mojo::DOM; use JSON;

my $url = "http://www.europapress.es/trafico/";

my $dom = Mojo::DOM->new( get $url );

my $estados_granada = $dom->find("table#tblTrafico tr")-
>grep(qr/Granada/i);

my %estados; for my $estado (@$estados_granada ) { push
@{$estados{$estado->at("td.lugar")->text}} , [$estado->at("td.fecha_tr")-
>text , $estado->find("td img")->map(attr =>'alt')->join(" | " )->to_string]; } say
encode_json \%estados; {% endhighlight %}
```

- <https://github.com/JJ/perl-moderno>

-

R

<http://www.r-bloggers.com/?s=Web+Scraping>

<http://cran.r-project.org/web/packages/htrr/htrr.pdf>

The screenshot shows the R-bloggers website interface. At the top, the logo "R-bloggers" is displayed with the tagline "R news and tutorials contributed by (573) R bloggers". Below the logo is a navigation bar with links: Home, About, RSS, add your blog!, R jobs, and Contact us. The main content area features a search result for "Web Scraping" with 152 results. The top result is titled "rvest: easy web scraping with R" by Hadley Wickham, dated November 24, 2014. The article description states: "rvest is new package that makes it easy to scrape (or harvest) data from html web pages, by libraries like beautiful soup. It is designed to work with magrittr so that you can express complex operations as elegant pipelines composed of simple, easily understood pieces. Install it with: install.packages('rvest') rvest in action To see rvest". The article has 17 comments and a "Read more" link. To the left of the article is a sidebar with a "WELCOME!" message, a "Follow @rbloggers" button (15.2K followers), and a "Subscribe" button (15765 readers). Below this is a "On Facebook" section showing 21,561 people who like R bloggers. To the right of the article is another sidebar titled "TOP 3 POSTS FROM THE PAST 2 DAYS" listing: "An R tutorial for Microsoft Excel users", "Installing R packages", and "Scatterplots". Below this is a search bar with the text "Search & Hit Enter". At the bottom of the right sidebar is a section titled "TOP 9 ARTICLES OF THE WEEK" listing: 1. Installing R packages, 2. Eight New Ideas From Data Visualization Experts, 3. Scatterplots, 4. Using apply, sapply, lapply in R, 5. In-depth introduction to machine learning in 15 hours of expert videos, 6. Visualization series: Insight from Cleveland and Tufte on plotting numeric data by groups, 7. 12 nifty tips for scientists who use computers.

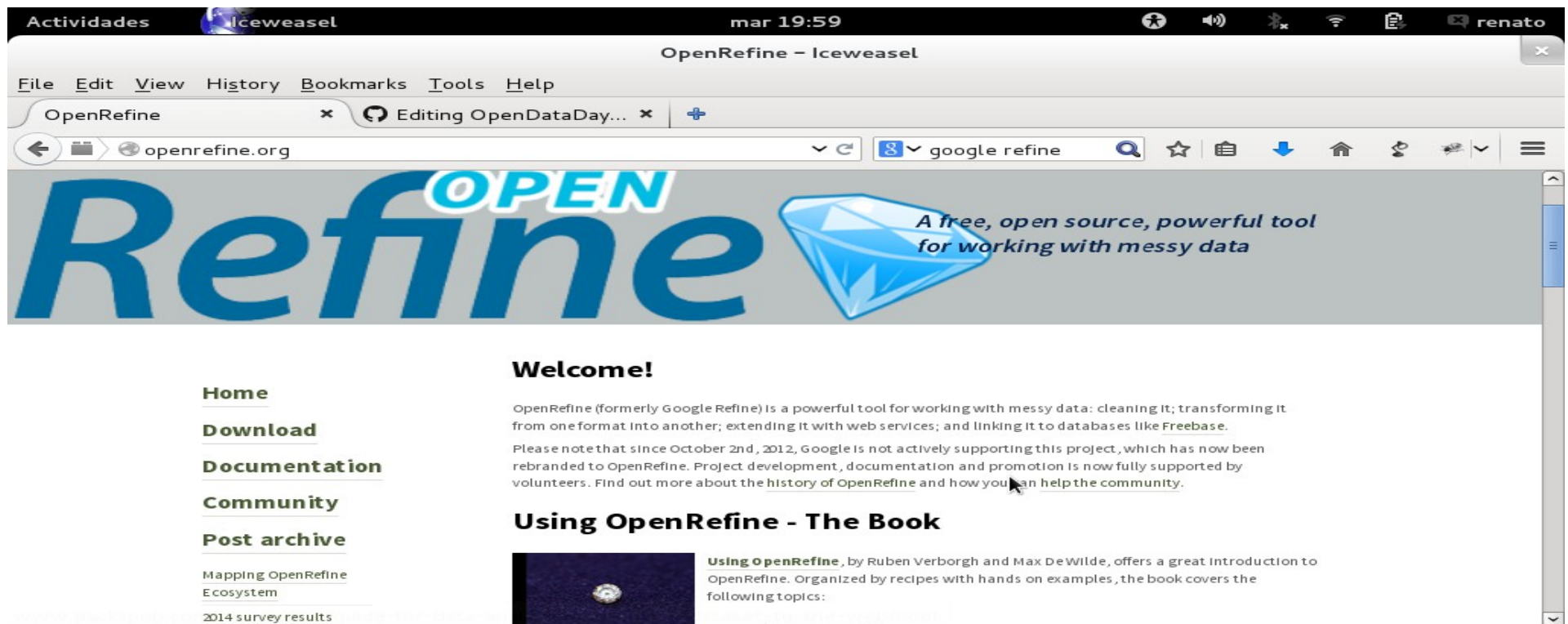
Scraping para no programadores.

Google Spreadsheet

- ImportHTML()
- importxml()

Google Refine

- Open Refine
- <http://openrefine.org/>



Extracción de datos en PDFs

Scraper Wiki

<http://scraperwiki.com>

Perl

Cpan

Tabula

- Java
- Tabula

<http://tabula.technology/>

Java

Actividades Iceweasel mar 18:58 renato

Verificar la versión de Java - Iceweasel

File Edit View History Bookmarks Tools Help

renatoIrr/OpenData... x Verificar la versión d... x

Oracle Corporation (US) https://www.java.com/es/download/installed.js java

Java™ Descargar Ayuda

Recursos de ayuda

- » ¿Qué es Java?
- » Eliminar versiones anteriores de Java
- » Desactivar Java
- » Mensajes de error
- » Solucionar problemas de Java
- » Otra ayuda

Mac OS X Chrome

Verificar la versión de Java

No hemos podido verificar si Java está instalado y activado en el explorador.

Si ha instalado Java y se ha producido un error con la verificación, se podría generar un problema de configuración (por ejemplo, en la configuración del explorador, del panel de control de Java o de seguridad) o el explorador podría bloquear el plugin de Java. Intente reiniciar el explorador antes de intentar verificar la instalación de nuevo y compruebe que el explorador permite la ejecución de Java.

- » Consulte las preguntas frecuentes sobre solución de problemas
- » Intente verificar Java de nuevo

Actividades Iceweasel mar 20:11 renato

Descargar software de Java para Linux - Iceweasel

File Edit View History Bookmarks Tools Help

Descargar software ... x OpenDataDay2015/... x

https://www.java.com/es/download/linux_manual.jsp?locale=es java

Java™ Descargar Ayuda

Recursos de ayuda

- » Solucionar problemas de Java

Java 7

- » ¿Dónde puedo obtener Java 7?

JDK

- » ¿Busca JDK?






Descargas Java para Linux

Recomendado Version 8 Update 31

En función del sistema operativo de su computadora, seleccione un archivo de la siguiente lista para obtener la versión más reciente de Java.

- > Todas las descargas de Java
- > Eliminar versiones anteriores de Java
- > ¿Qué es Java?

Al descargar Java, confirma que ha leído y aceptado los términos del acuerdo de licencia de usuario final

	Linux	
	Linux RPM	Tamaño de archivo: 40.4 MB
	Linux	Tamaño de archivo: 61.5 MB
	Linux x64	Tamaño de archivo: 59.8 MB
	Linux x64 RPM	Tamaño de archivo: 40.6 MB

[Instrucciones](#)

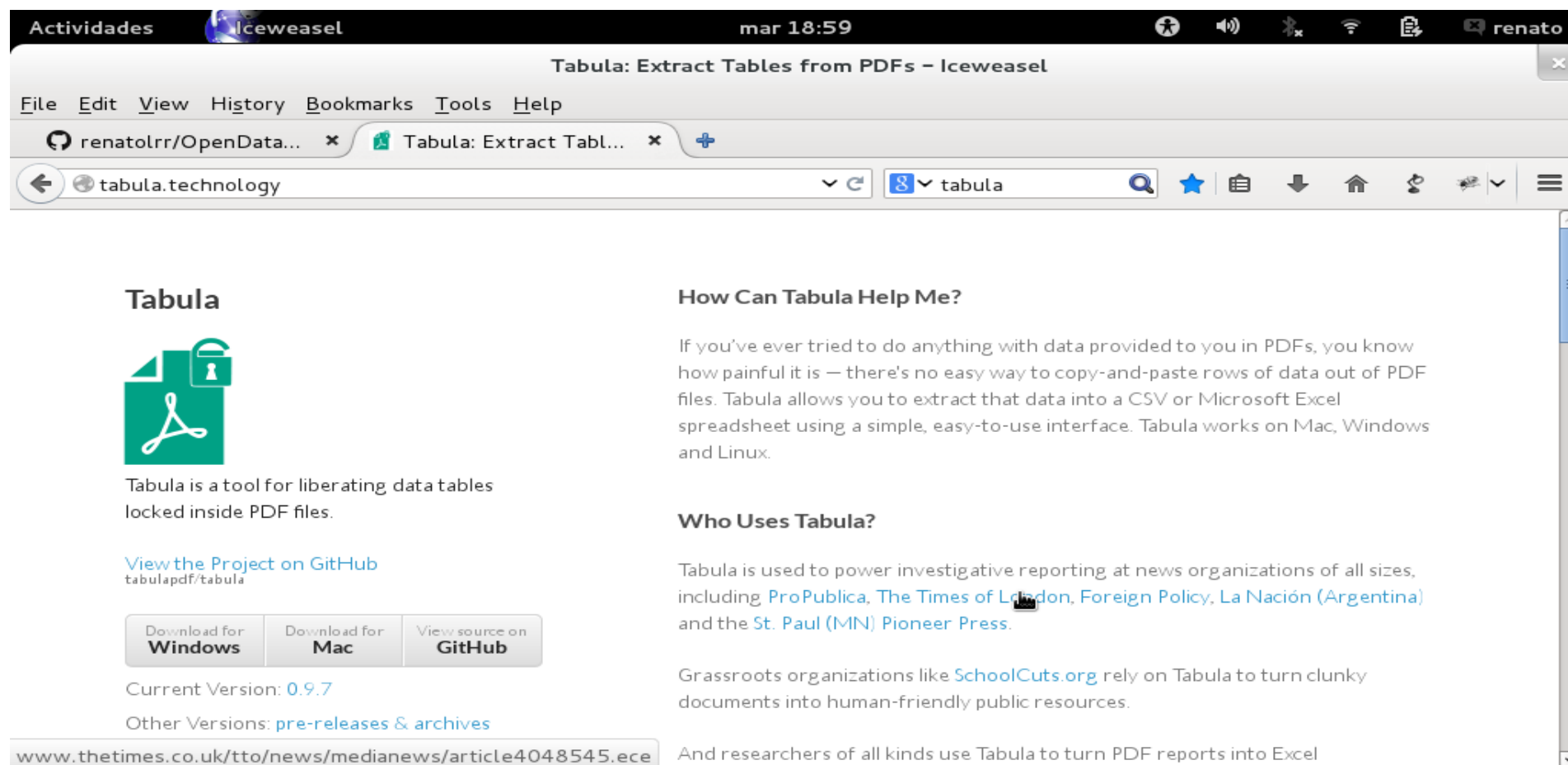
[Instrucciones](#)

[Instrucciones](#)

[Instrucciones](#)

Tras la instalación de Java, deberá activar Java en el explorador.

Tabula



The screenshot shows a web browser window titled "Tabula: Extract Tables from PDFs - Iceweasel". The address bar shows "tabula.technology". The page content includes the Tabula logo (a green square with a white PDF icon and a padlock), a description of the tool, and links to download it for Windows, Mac, or view the source on GitHub. The current version is 0.9.7. The page also features sections titled "How Can Tabula Help Me?" and "Who Uses Tabula?", which describe the tool's utility for extracting data from PDFs and list various organizations that use it, such as ProPublica, The Times of London, and SchoolCuts.org. A URL is visible in the browser's address bar: "www.thetimes.co.uk/tto/news/medianews/article4048545.ece".

Tabula

Tabula is a tool for liberating data tables locked inside PDF files.

[View the Project on GitHub](#)
tabulapdf/tabula

Download for **Windows** | Download for **Mac** | View source on **GitHub**

Current Version: **0.9.7**

Other Versions: [pre-releases & archives](#)

www.thetimes.co.uk/tto/news/medianews/article4048545.ece

How Can Tabula Help Me?

If you've ever tried to do anything with data provided to you in PDFs, you know how painful it is — there's no easy way to copy-and-paste rows of data out of PDF files. Tabula allows you to extract that data into a CSV or Microsoft Excel spreadsheet using a simple, easy-to-use interface. Tabula works on Mac, Windows and Linux.

Who Uses Tabula?

Tabula is used to power investigative reporting at news organizations of all sizes, including [ProPublica](#), [The Times of London](#), [Foreign Policy](#), [La Nación \(Argentina\)](#) and the [St. Paul \(MN\) Pioneer Press](#).

Grassroots organizations like [SchoolCuts.org](#) rely on Tabula to turn clunky documents into human-friendly public resources.

And researchers of all kinds use Tabula to turn PDF reports into Excel

Tabula

Actividades Iceweasel mar 19:00 renato

Tabula: Extract Tables from PDFs - Iceweasel


File Edit View History Bookmarks Tools Help

renatoIrr/OpenData... x Tabula: Extract Tabl... x

tabula.technology

tabula

Tabula



Tabula is a tool for liberating data tables locked inside PDF files.

[View the Project on GitHub](#)
tabulapdf/tabula

[Download for Windows](#) [Download for Mac](#) [View source on GitHub](#)

Current Version: [0.9.7](#)

Other Versions: [pre-releases & archives](#)

<https://github.com/tabulapdf/tabula/releases/download/v0.9.7/tabula-jar-0.9.7.zip>

And researchers of all kinds use Tabula to turn PDF reports into Excel spreadsheets, CSVs, and JSON files for use in analysis and database applications.

How to Install Tabula

Note: You'll need a copy of [Java](#) installed. You can [download Java here](#).

1. Download the version of Tabula for your operating system:
 - o Windows: [tabula-win.zip](#)
 - o Mac OS X: [tabula-mac.zip](#)
 - OS X 10.8+ users: if you have issues opening the app, [see the two notes at bottom of this page](#)
 - o Linux/Other: [tabula-jar.zip](#) (view README.txt inside for instructions)
2. Extract the zip file. (Instructions: [Windows](#), [Mac](#))
3. Go into the folder you just extracted. Run the "Tabula" program inside.
4. A web browser will open. If it doesn't, open your web browser, and go to <http://localhost:8080>. There's Tabula!

Tabula

Actividades Iceweasel mar 19:03 renato

Tabula: Extract Tables from PDFs – Iceweasel


File Edit View History Bookmarks Tools Help

renatolrr/OpenData... x Tabula: Extract Tabl... x

tabula.technology

tabula

Tabula



Tabula is a tool for liberating data tables locked inside PDF files.

[View the Project on GitHub](#)
tabulapdf/tabula

Download for Windows Download for Mac

View source on GitHub Current Version: 0.9.7 Other Versions: pre-releases & archives

Need help? Open an [issue on Github](#).

How to Use Tabula

1. Upload a PDF file containing a data table.
2. Select the table by clicking the top left corner of a table and dragging the mouse to the bottom right corner, until all of the data is included in the shaded selection area.
3. A window will then appear containing your data. Inspect the data to make sure it looks correct. If data is missing, you may have to slightly expand your selection.
4. Click the Download button.
5. Now you can work with your data as text file or a spreadsheet rather than a PDF!
(You can open the downloaded file in Microsoft Excel or the free [LibreOffice Calc](#))

Note: Tabula only works on text-based PDFs, not scanned documents.

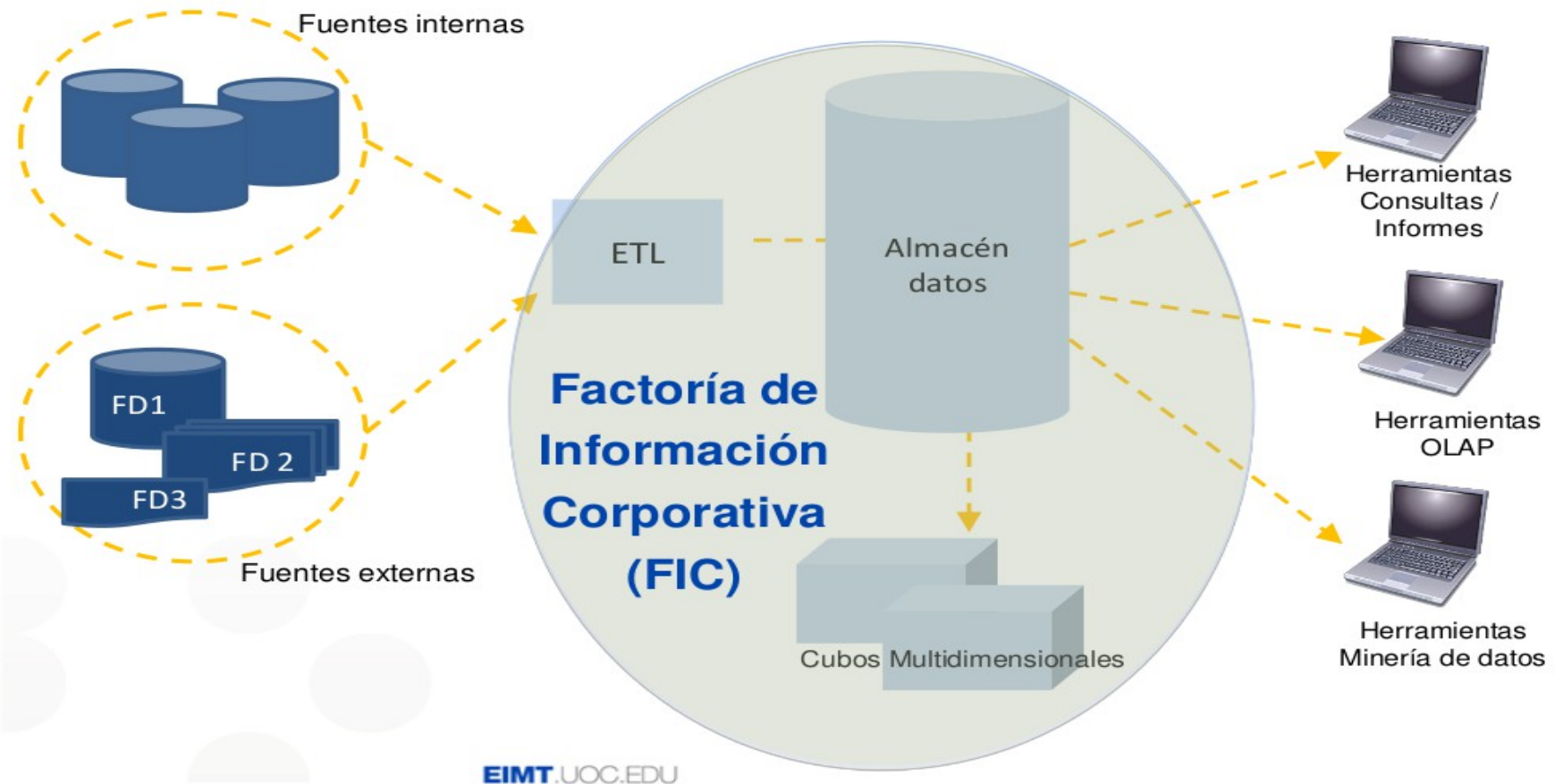
Your web browser may lack some functionality expected by Popcorn Maker to function properly. Please upgrade your browser or [file a bug](#) to find out why your browser isn't fully supported. Click [here](#) to remove this warning.

Fuentes de datos comunes

Data Warehouse

“Es una colección de datos orientados al tema, integrados, no-volátiles e historizados, organizados para dar soporte a los procesos de ayuda a la decisión.”

Data Warehouse



Data Warehouse

TRANSFORMACIÓN

- Cambiar formato o tipo de datos (ejemplo formato fecha).
- Reestructurar campos (fusionar o dividir campos).
- Cambiar las unidades o códigos de transformación (cambios de moneda).
- Cambiar el grado de agregación (calcular las ventas mensuales a partir de las diarias).
- Añadir información temporal (período validez de los datos).

DEPURACIÓN

- Detectar y corregir valores inconsistentes.
- Añadir valores por defecto a los campos con valores no definidos
- Detectar y corregir información duplicada.

INTEGRACION

- El proceso de integración dependerá si realizamos la carga inicial del almacén de datos o una actualización.
- Principal problema: Detectar datos que representan el mismo concepto.
- Se transforman los datos para homogeneizar la representación y eliminar la información duplicada.

Granjero

- Accede a información de forma predecible y repetitiva.
- Sólo accede a su parcela de información: extrae datos para mejorar el funcionamiento de la empresa.
- Utiliza herramientas OLAP (On-Line Analytical Processing).

Explorador

- Explora gran cantidad de datos.
- Accede a información de forma impredecible e irregular.
- Perfil informático o estadístico.
- Objetivo: Obtener información que proporcione ventaja competitiva.

Turista

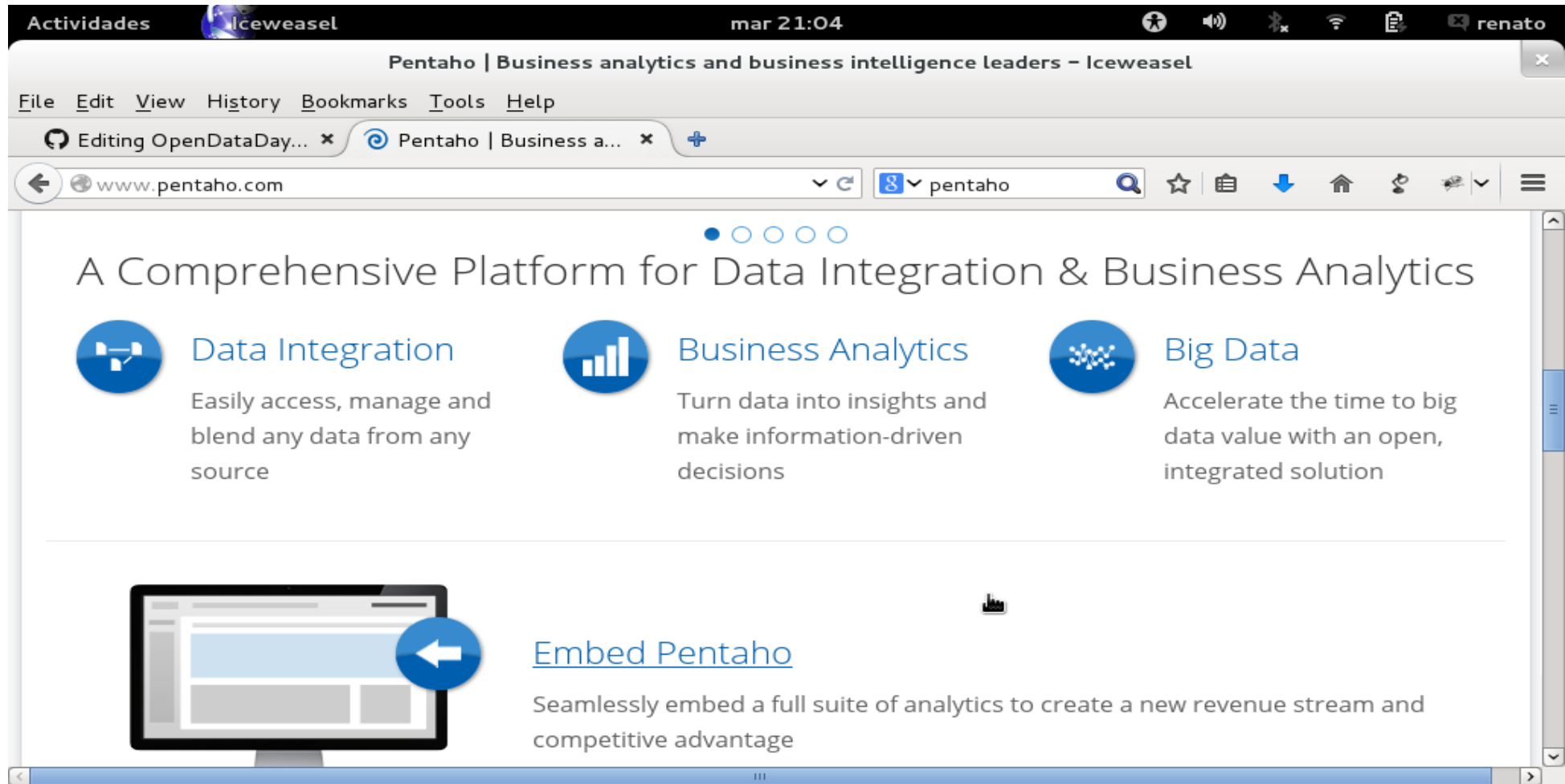
- Grupo de dos o más personas.
- Un perfil con conocimientos del negocio y visión global empresa.
- Segundo perfil con conocimientos informáticos.
- Consulta datos y metadatos.
- Acceden sin ningún patrón de acceso.
- Las herramientas que utiliza suelen ser navegadores o buscadores.
- Su resultado serán proyectos para los usuarios granjero y explorador.
-

Pentaho

- <http://es.wikipedia.org/wiki/Pentaho>

Pentaho BI Suite es un conjunto de programas libres para generar inteligencia empresarial (Business Intelligence). Incluye herramientas integradas para generar informes, minería de datos, ETL, etc.

Pentaho



Pentaho

Kettle – Pentaho Data Integration

Formado por cuatro componentes:

- Spoon: entorno gráfico para el desarrollo de transformaciones y trabajos.
- Pan: permite ejecutar transformaciones.
- Kitchen: permite ejecutar trabajos.
- Carte: es un servidor remoto que permite la ejecución de transformaciones y trabajos.

Problema codificación

http://es.wikipedia.org/wiki/Codificaci%C3%B3n_de_caracteres

<http://ubuntudriver.blogspot.com.es/2011/06/cambiar-codificacion-de>

Problema codificación

Minúsculas				
carácter	ISO-8859-1	UTF-8	UTF-16	
a	0x61	0x61	0x00	0x61
b	0x62	0x62	0x00	0x62
c	0x63	0x63	0x00	0x63
d	0x64	0x64	0x00	0x64
e	0x65	0x65	0x00	0x65
f	0x66	0x66	0x00	0x66
g	0x67	0x67	0x00	0x67
h	0x68	0x68	0x00	0x68
i	0x69	0x69	0x00	0x69
j	0x6a	0x6a	0x00	0x6a
k	0x6b	0x6b	0x00	0x6b
l	0x6c	0x6c	0x00	0x6c
m	0x6d	0x6d	0x00	0x6d
n	0x6e	0x6e	0x00	0x6e
o	0x6f	0x6f	0x00	0x6f
p	0x70	0x70	0x00	0x70
q	0x71	0x71	0x00	0x71
r	0x72	0x72	0x00	0x72
s	0x73	0x73	0x00	0x73
t	0x74	0x74	0x00	0x74
u	0x75	0x75	0x00	0x75
v	0x76	0x76	0x00	0x76
w	0x77	0x77	0x00	0x77
x	0x78	0x78	0x00	0x78
y	0x79	0x79	0x00	0x79
z	0x7a	0x7a	0x00	0x7a

Mayúsculas				
carácter	ISO-8859-1	UTF-8	UTF-16	
A	0x41	0x41	0x00	0x41
B	0x42	0x42	0x00	0x42
C	0x43	0x43	0x00	0x43
D	0x44	0x44	0x00	0x44
E	0x45	0x45	0x00	0x45
F	0x46	0x46	0x00	0x46
G	0x47	0x47	0x00	0x47
H	0x48	0x48	0x00	0x48
I	0x49	0x49	0x00	0x49
J	0x4a	0x4a	0x00	0x4a
K	0x4b	0x4b	0x00	0x4b
L	0x4c	0x4c	0x00	0x4c
M	0x4d	0x4d	0x00	0x4d
N	0x4e	0x4e	0x00	0x4e
O	0x4f	0x4f	0x00	0x4f
P	0x50	0x50	0x00	0x50
Q	0x51	0x51	0x00	0x51
R	0x52	0x52	0x00	0x52
S	0x53	0x53	0x00	0x53
T	0x54	0x54	0x00	0x54
U	0x55	0x55	0x00	0x55
V	0x56	0x56	0x00	0x56
W	0x57	0x57	0x00	0x57
X	0x58	0x58	0x00	0x58
Y	0x59	0x59	0x00	0x59
Z	0x5a	0x5a	0x00	0x5a

Acentos y tildes				
carácter	ISO-8859-1	UTF-8	UTF-16	
á	0xe1	0xc3 0xa1	0x00	0xe1
Á	0xc1	0xc3 0x81	0x00	0xc1
é	0xe9	0xc3 0xa9	0x00	0xe9
É	0xc9	0xc3 0x89	0x00	0xc9
í	0xed	0xc3 0xad	0x00	0xed
Í	0xcd	0xc3 0x8d	0x00	0xcd
ó	0xf3	0xc3 0xb3	0x00	0xf3
Ó	0xd3	0xc3 0x93	0x00	0xd3
ú	0xfa	0xc3 0xba	0x00	0xfa
Ú	0xda	0xc3 0x9a	0x00	0xda
ü	0xfc	0xc3 0xbc	0x00	0xfc
Ü	0xdc	0xc3 0x9c	0x00	0xdc
ñ	0xf1	0xc3 0xb1	0x00	0xf1
Ñ	0xd1	0xc3 0x91	0x00	0xd1

Símbolos				
carácter	ISO-8859-1	UTF-8	UTF-16	
¿	0xbf	0xc2 0xbf	0x00	0xbf
?	0x3f	0x3f	0x00	0x3f
¡	0xa1	0xc2 0xa1	0x00	0xa1
!	0x21	0x21	0x00	0x21

Conclusión

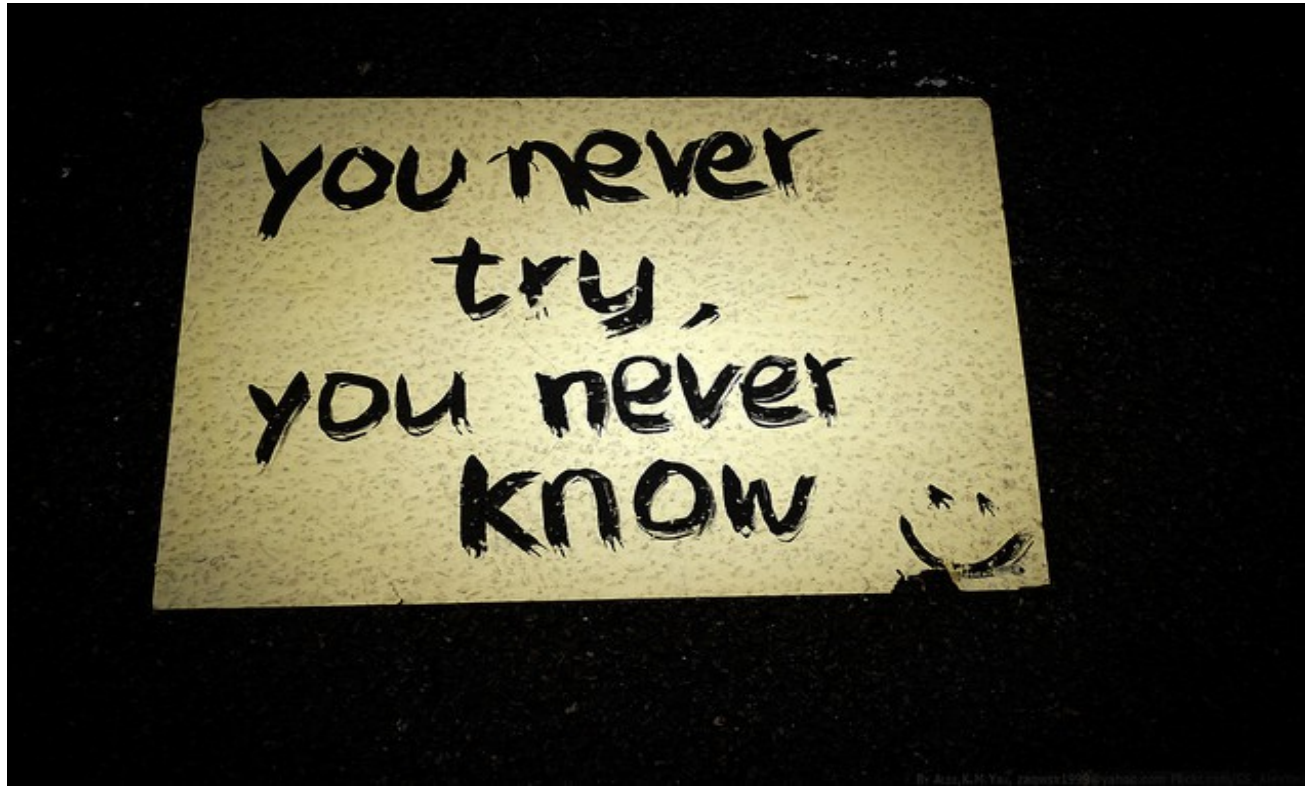


Image credit: Umbrella movement, Alex, K.M. Yau, Flickr, CC BY

Cursos

- Desarrollo de software colaborativo con Git
- Introducción al lenguaje de programación Python
- Programación Avanzada en Python
- Programación en Perl
- Programación Avanzada en Perl

Cevug

Buscar un hacker

