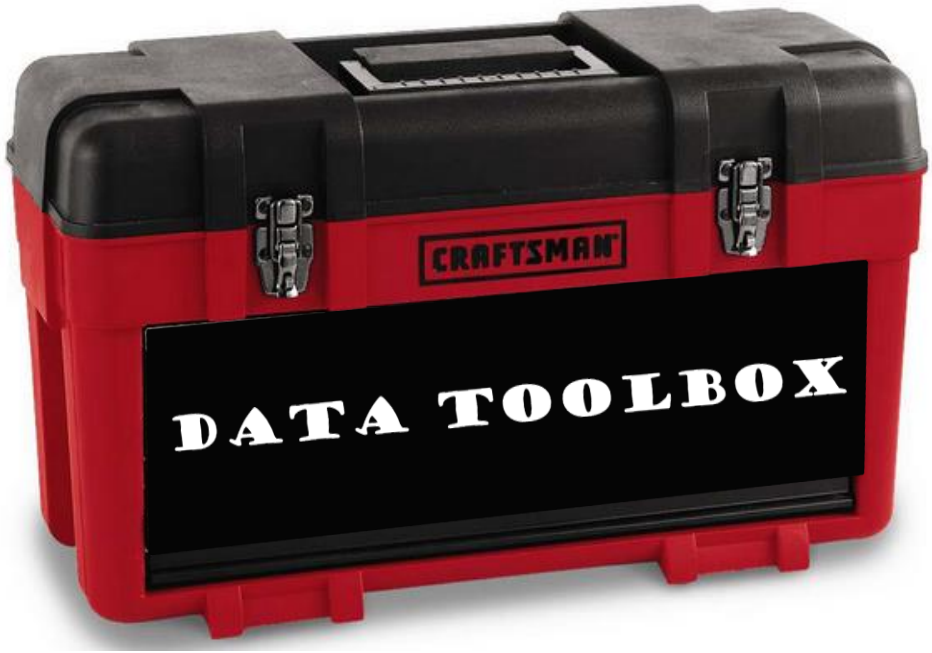
Taller de datos

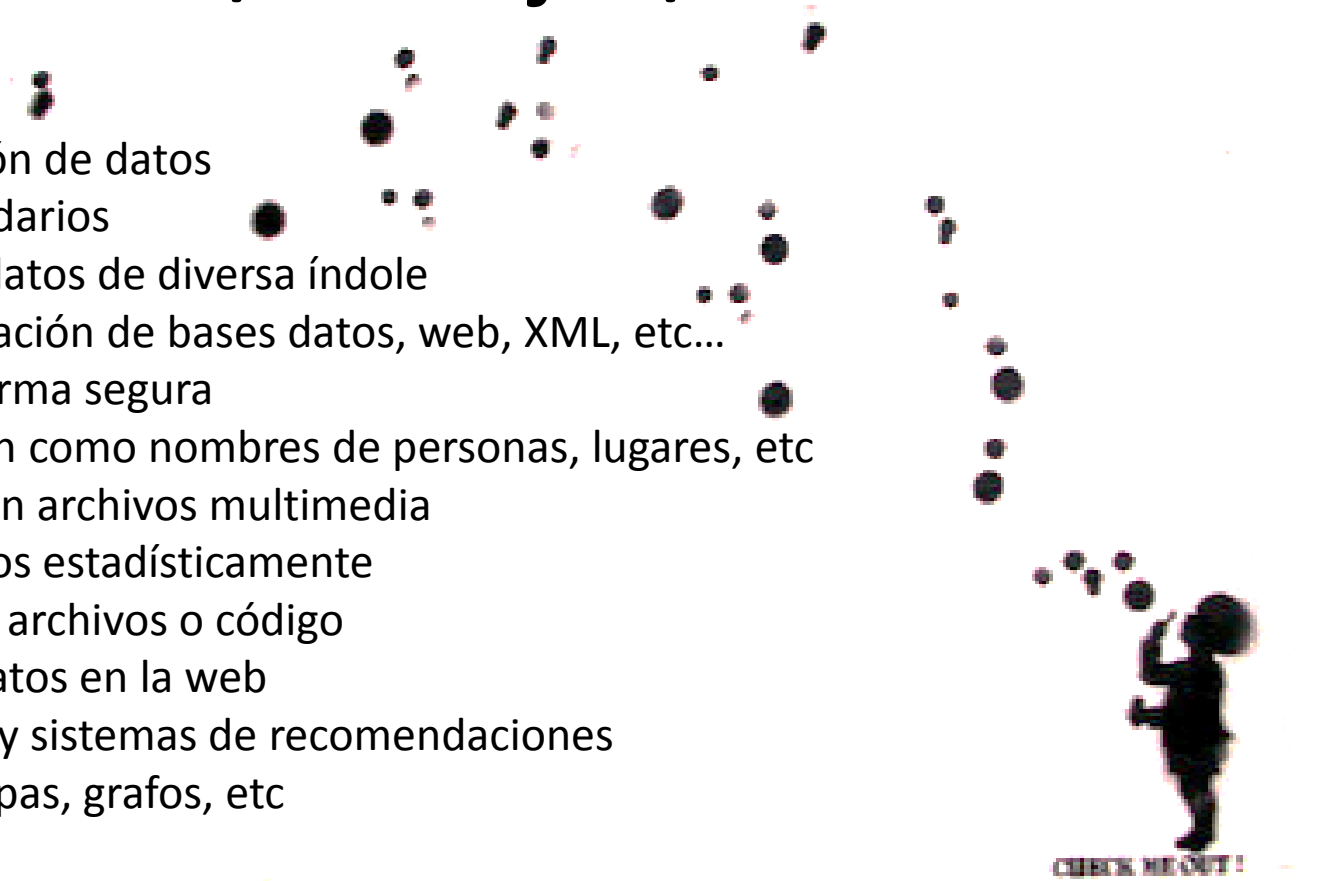


Luis Martínez-Uribe, Biblioteca Fundación Juan March

XIV Workshop REBIUN -VI Jornadas OS-Repositorios
12 Marzo 2015

Problemas/trabajos/retos

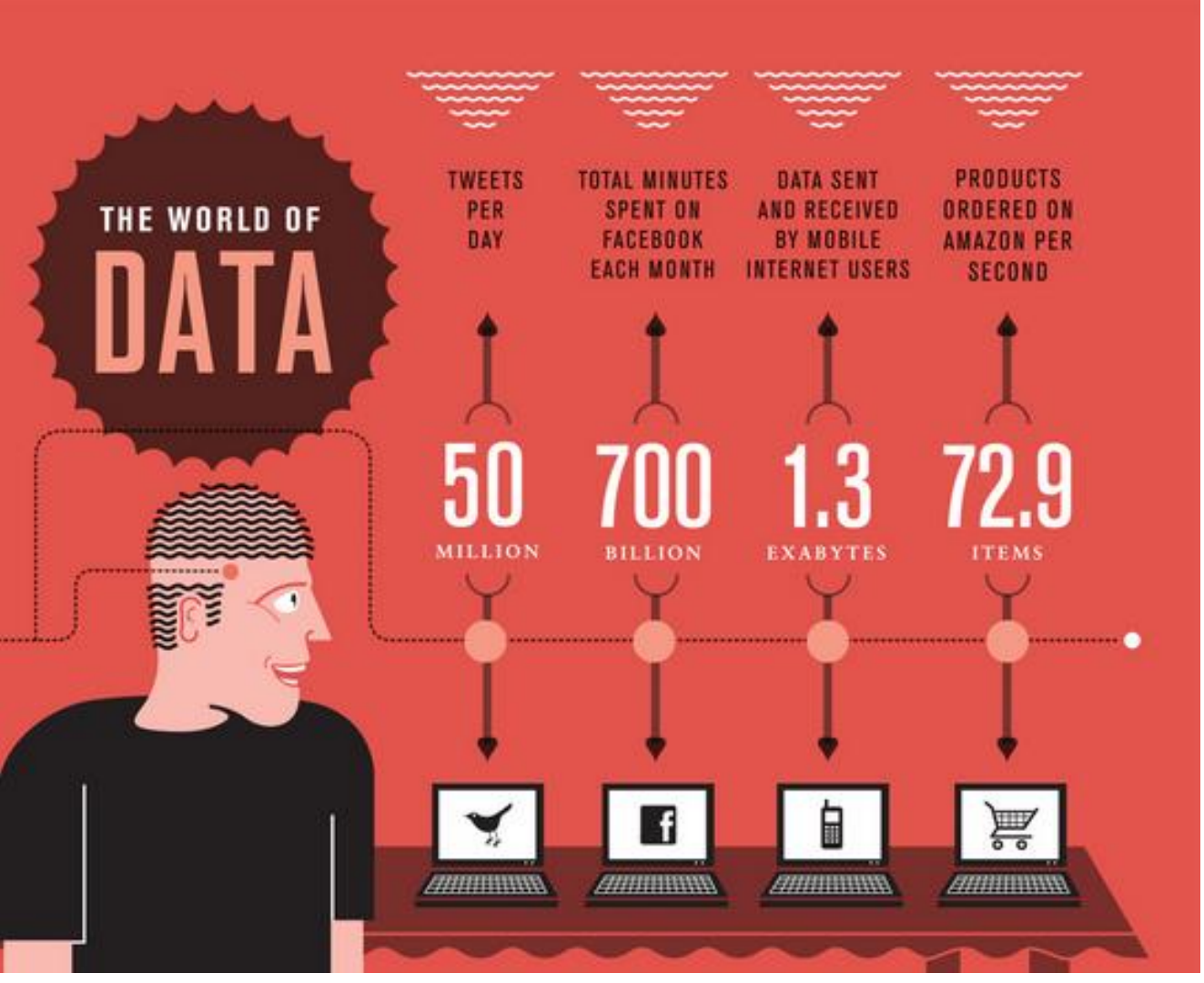
- Crear un plan de gestión de datos
- Encontrar datos secundarios
- Generar datos y metadatos de diversa índole
- Extraer y tratar información de bases datos, web, XML, etc…
- Almacenar datos de forma segura
- Normalizar información como nombres de personas, lugares, etc
- Gestionar y trabajar con archivos multimedia
- Explorar y analizar datos estadísticamente
- Controlar versiones de archivos o código
- Publicar y compartir datos en la web
- Motores de búsqueda y sistemas de recomendaciones
- Visualizar datos en mapas, grafos, etc



Objetivos del taller

- ***visión amplia y práctica*** *de los trabajos con datos que pueden formar parte de los* ***servicios de datos*** *desde bibliotecas*

- *ejemplos de herramientas, metodologías y estrategias:*
  - *planificar*
  - *generar y documentar*
  - *almacenar*
  - *tratar y normalizar*
  - *analizar*
  - *visualizar*
  - *compartir*
  - *publicar*



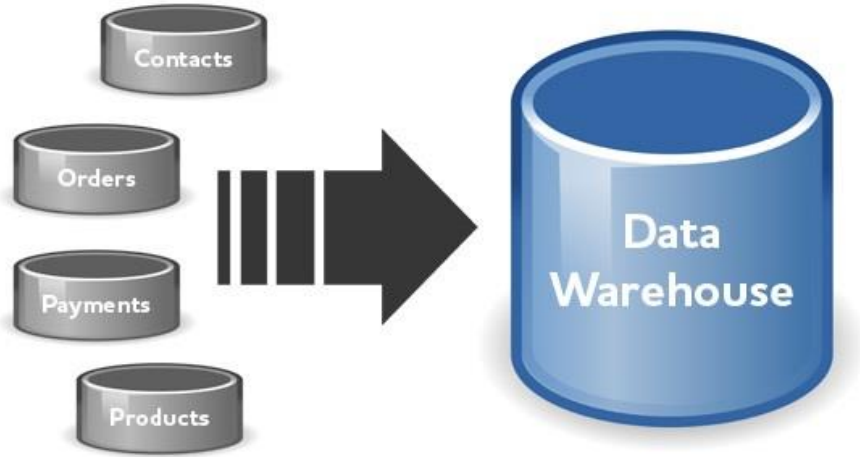
IBM

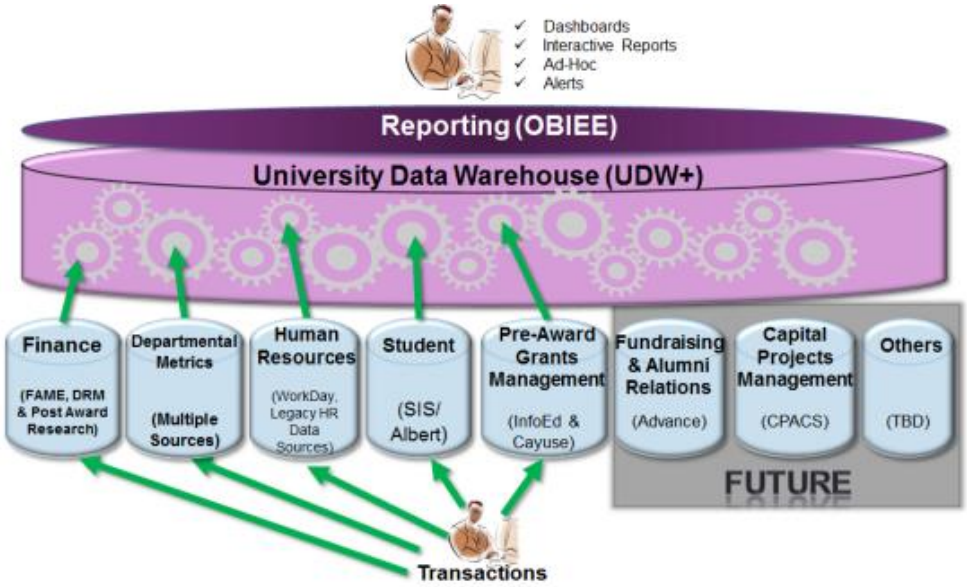
ebay

amazon

hp

Datos en organizaciones y universidades





New York University Warehouse Plus
http://www.nyu.edu/employees/resources-and-services/administrative-services/university-data-warehouse-plus.html

Long Tail Of Science

Requisitos de las agencias de financiación

*"… **open access to scientific data** should be adopted as the international norm for the exchange of scientific data derived **from publicly funded research**."*

OECD Principles and Guidelines for Access to Research Data from Public Funding (2004-2007)

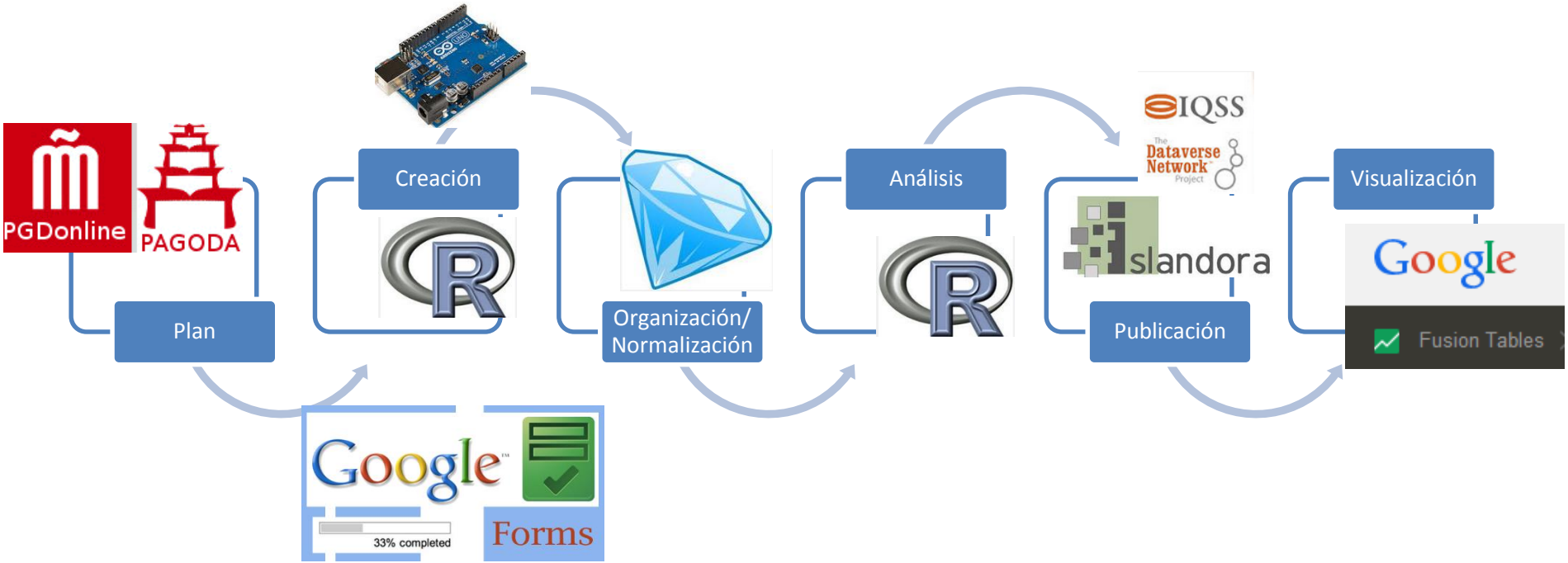
*"**requires, in all proposals** a supplementary document of no more than two pages describing **a Data Management Plan** for the proposed research. "*

The National Science Foundation, January 2011

*"…**primary data**, as well as data-related products such as computer codes, is deposited in the relevant databases **as soon as possible**, preferably immediately after publication and in any case **not later than six months after the date of publication**."*

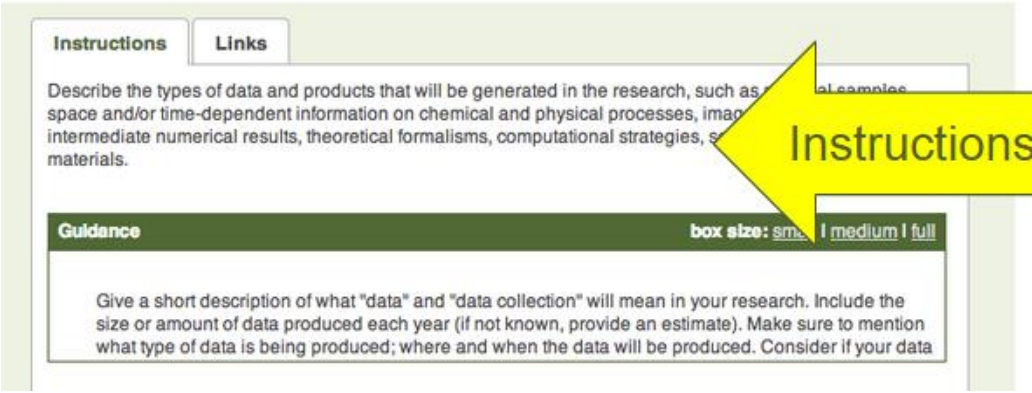
European Research Council , Open Access Guidelines for Researchers, June 2012

Herramientas a lo largo del ciclo de vida de los datos



Planes de gestión de datos

"Documentos que describen que harás con tus datos durante tu investigación y una vez que termines con tu proyecto"



Herramientas

- DMPTool (EEUU)
- DMPOnline (Reino Unido)
- PGDOnline (España)

- EU Guidelines on DMP Horizonte 2020 http://bit.ly/1LukfMV
- DCC DMP Checklist http://bit.ly/1zxssZt
- Creamos un DMP usando PGDOnline

CREAR Y ALMACENAR DATOS

Herramientas de encuestas

- Instrucciones Google Forms http://bit.ly/1LuobgJ
- Creamos una encuesta usando la herramienta http://bit.ly/17JdmIx

Instrumentos con Open Source Hardware

Arduino - a class of open source microcontrollers useful for automating equipment

Raspberry Pi - credit-card sized computer running Linux

Red Pitaya - open source measurement and control tool

Sensorica - an Open Value Network providing sensing and automation solutions.

3D printable science equipment - 3D print your lab

- Subir datos a internet con arduino
  - http://bit.ly/1NEHMgc
  - https://thingspeak.com/channels/9

Imágenes e información de:
http://www.appropedia.org/Building_research_equipment_with_free,_open-source_hardware

Web Scraping

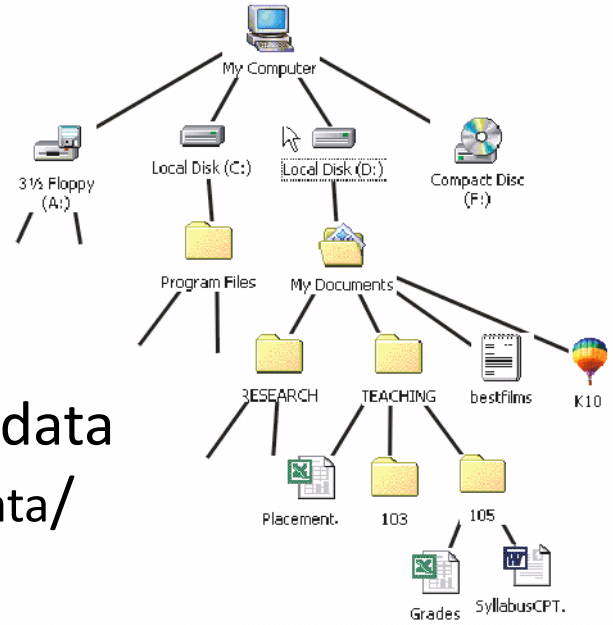
- Tutoriales básicos
  http://bit.ly/1zxzope
  http://bit.ly/1EFvbn6
- Demostración de web scraping con R
  http://bit.ly/1AjnSSa

Imagen de:
https://blog.scraperwiki.com/2011/06/knight-foundation-finance-scraperwiki-for-journalism/

Organizar los datos

- Ficheros de datos y carpetas tiene que denotarse y organizarse de forma sistemática.
- Criterios
  - Organización
  - Contexto
  - Consistencia
- Research Data Mantra – Organising data

http://datalib.edina.ac.uk/mantra/organisingdata/



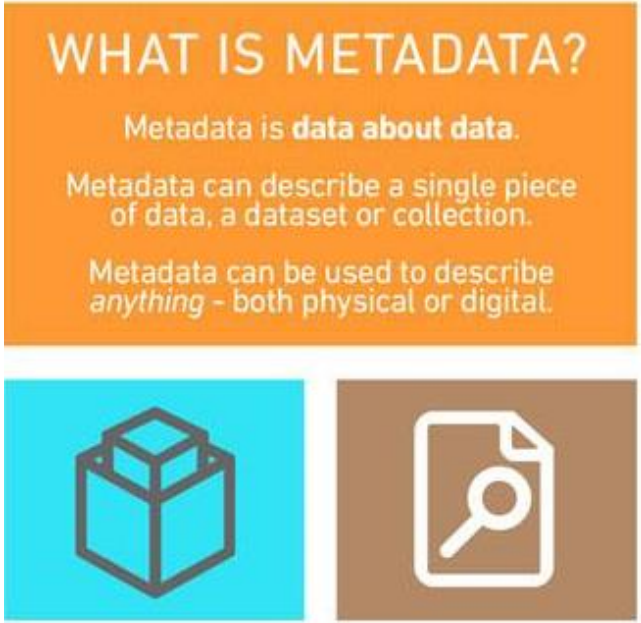
Donde almacenarlos

- Servidores en red
  - Gestionados por informáticos
  - Con back-ups
- Ordenadores personales y portátiles
  - Discos duros pueden fallar
  - Portátiles se pueden perder
- Unidades de almacenamiento externo
  - Longevidad no garantizada
  - Fácilmente se estropean o pierden

Fotos tomadas de *www.**data-archive**.ac.**uk**/ukda40/*

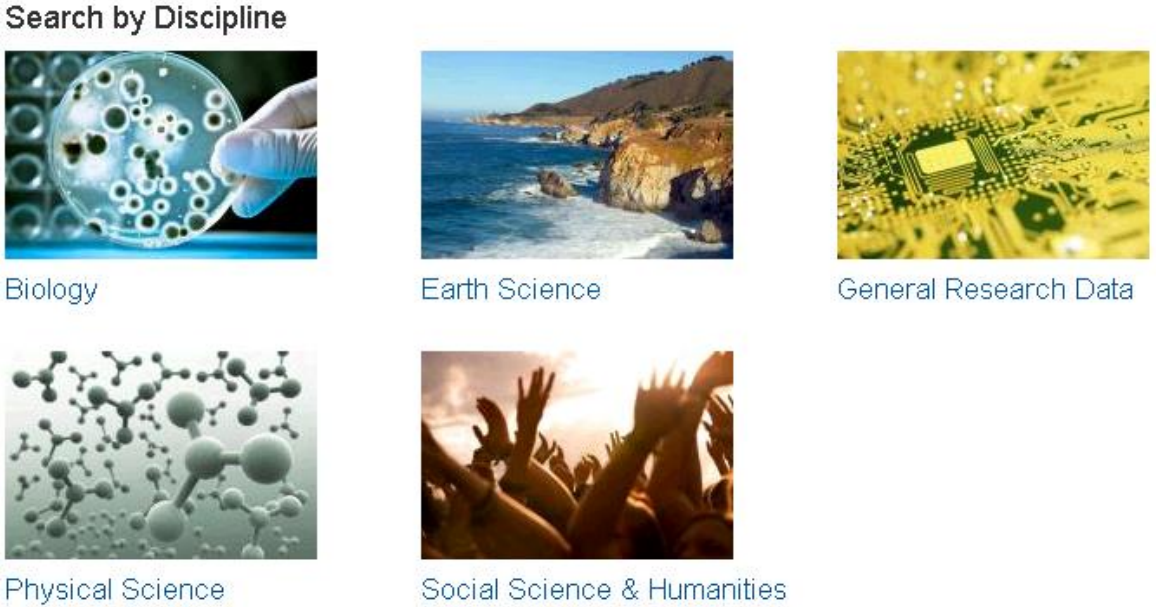
- Curso de metadatos MANTRA

http://datalib.edina.ac.uk/mantra/documentation_metadata_citation/



- Estándares

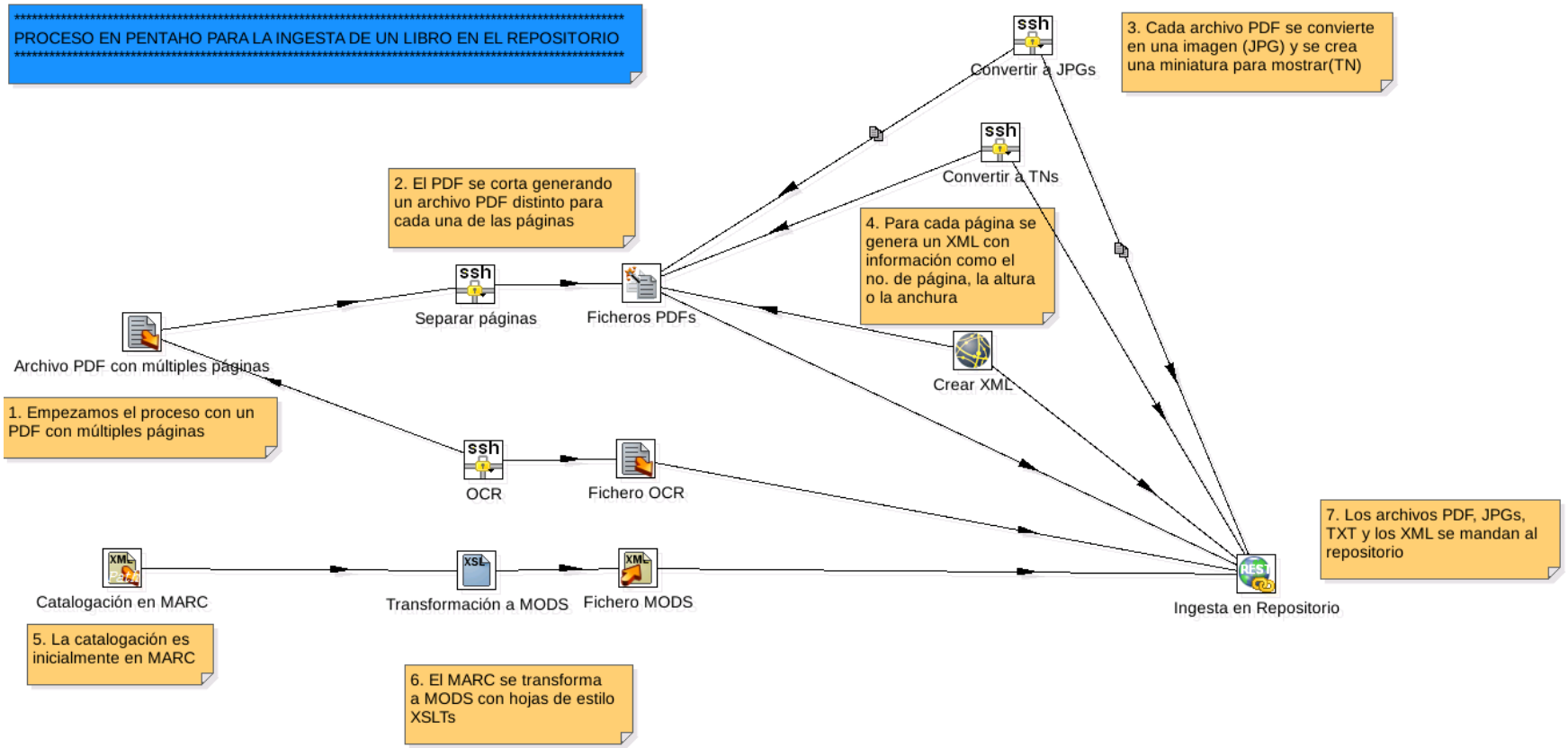
http://www.dcc.ac.uk/resources/metadata-standards




TRANSFORMAR
E
INTEGRAR

Open Refine

- Tutoriales básicos
  http://bit.ly/1zYWGn8
  http://bit.ly/1B7pqRA
- Demostración de desambiguación de nombres de persona

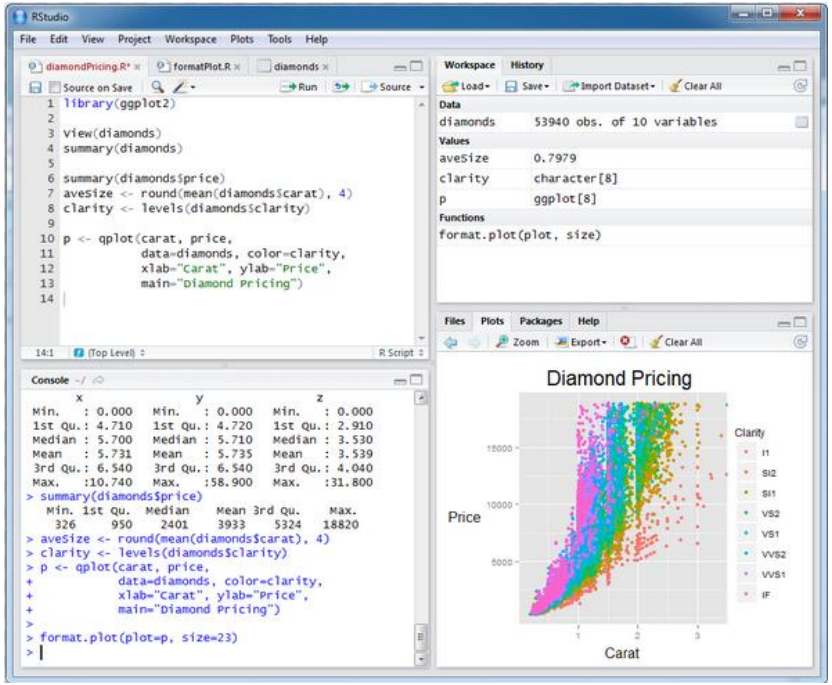

****
PROCESO EN PENTAHO PARA LA INGESTA DE UN LIBRO EN EL REPOSITORIO
****
Convertir a JPGs
3. Cada archivo PDF se convierte en una imagen (JPG) y se crea una miniatura para mostrar(TN)
Convertir a TNs
2. El PDF se corta generando un archivo PDF distinto para cada una de las páginas
4. Para cada página se genera un XML con información como el no. de página, la altura o la anchura
Separar páginas
Ficheros PDFs
Archivo PDF con múltiples páginas
Crear XML
1. Empezamos el proceso con un PDF con múltiples páginas
OCR
Fichero OCR
7. Los archivos PDF, JPGs, TXT y los XML se mandan al repositorio
Catalogación en MARC
Transformación a MODS
Fichero MODS
Ingesta en Repositorio
5. La catalogación es inicialmente en MARC
6. El MARC se transforma a MODS con hojas de estilo XSLTs

ANALIZAR Y VISUALIZAR


SPSS®


STATA®


sas®



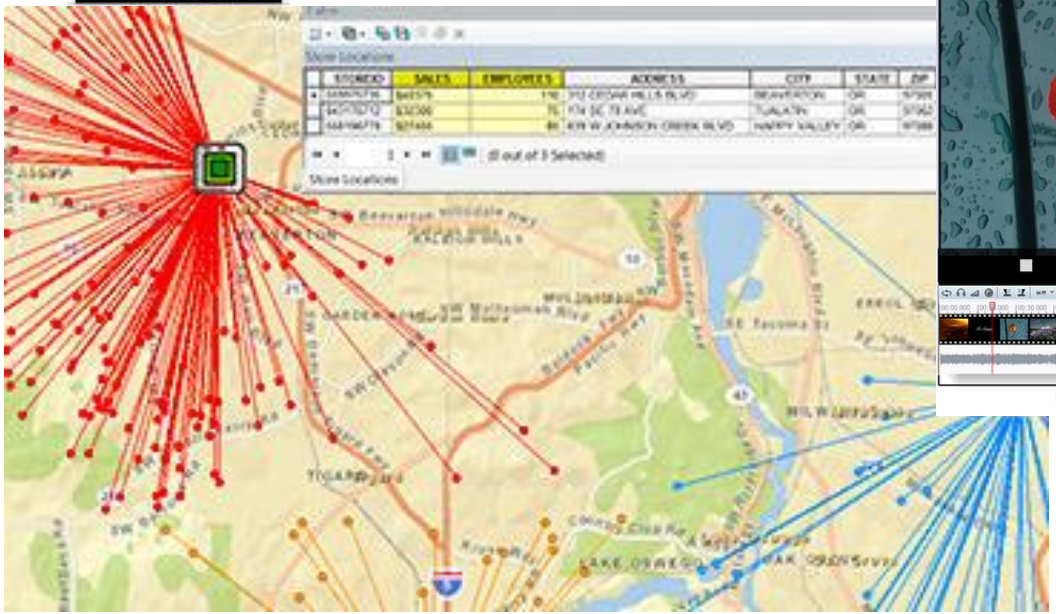
- Demostración de análisis básico de datos con R http://bit.ly/1MnqPFv
- Demostración de captura de datos de twitter http://bit.ly/1GjV5gL

Wordle™

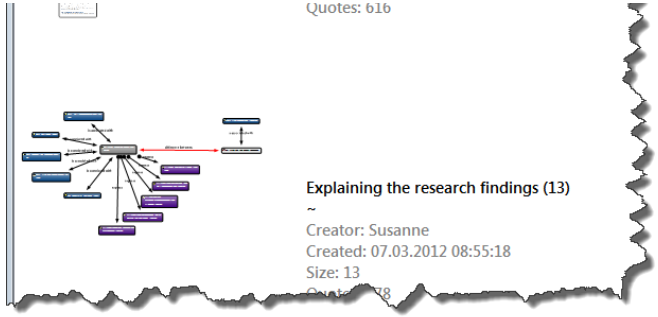

@TeresaRodr
L6Neldebate
28FUnidosPorAndalucía
ProgramaAndalucía
Andalucía
gente
cambio
Londres
https://t.co/M2tEJ84Nk0
https://t.co/wT3oN4WiPC
https://t.co/tHCuaVWWnY
@PFeminismos
igualdad
Gobierno
preocupa
vota
acto
encuestas
exilio
debate
Mujer
país
comienza
organizado
foto
tierra
Encuentro
Podemos
estatal
Vota
Rajoy
día
financie
Continuamos
Eduardo
interviene
decidir
Entra
embargo
medicamentos
primavera
emocionada
firmas
vive
después
expresa
virtudes
pierdas
políticos
servicio
volver
cosas
digna
sequías
contigo
hepatitis
hace
streaming
partir
va
Cádiz
ser
vida
Economía
test
lugar
Gran
griegos
duelen
programa
junto
experiencia
conmigo
llegado
eligen
decidan
democratizar
venga
Parece
andaluzas
Siguela
nunca
pagar
apoyo
Comenzamos
urgente
mismo
semillas
emigrante
Artículo
Mariano
posible
duras
parte
hacer
movieron
muestran
Directo
entregan
ciudadana
feliz
emotivo
recuperar
organización
participación
pobreza
confiar
entraron
mañana
Programa
decisión
millones
@SextaNocheTV
28FUnidosPorAndalucía
vendrá
importante
fraude
demás
paseo
trabajar
celebrado
tener
Contamos
conozca
hombres
andaluces
encuentro
hoy
dos
gran
Hoy
economía

ArcGIS®

ESRI



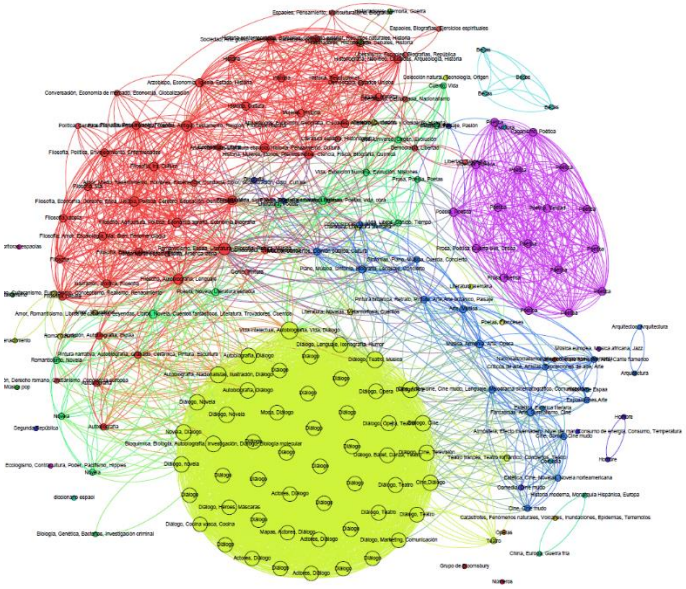
atlas.ti



Google

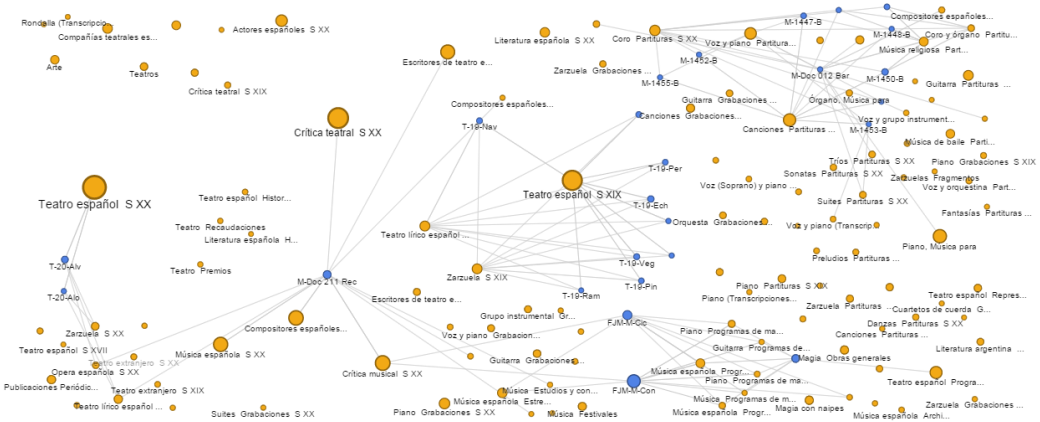
Fusion Tables

- Datos de paro por municipio
  http://bit.ly/1HnUtI2
- Datos de población por municipio
  http://bit.ly/1xahmbX
- Demostración de cómo representarlos usando Google Fusion Tables
  http://bit.ly/1Adqyxa



Gephi

- Ejemplo de grafico de redes
  http://bit.ly/18834Ce



Google Fusion Tables

Timelines

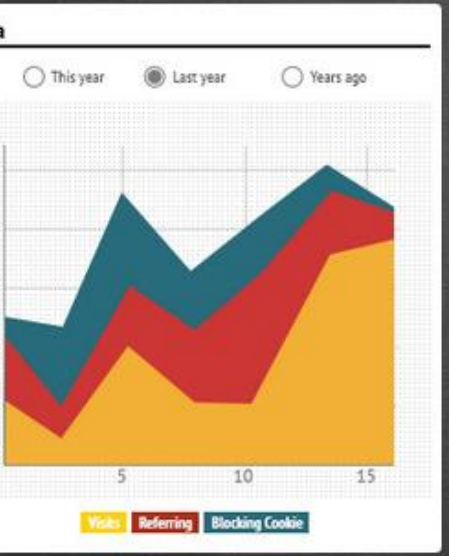
Timeline JS

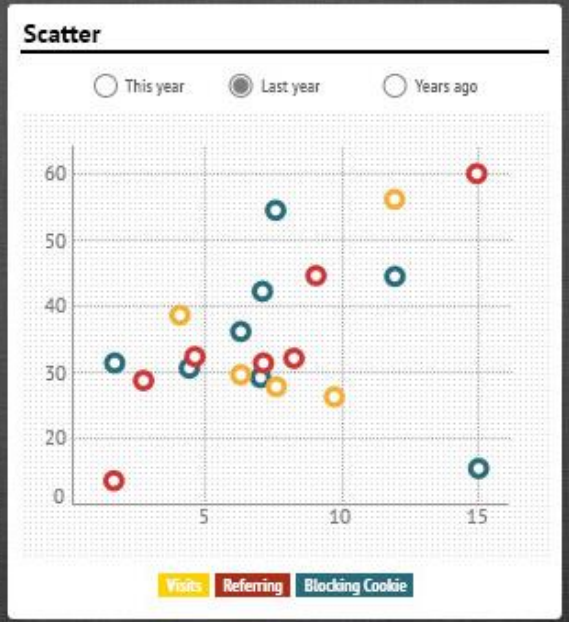
Beautifully crafted timelines that are easy and intuitive to use.

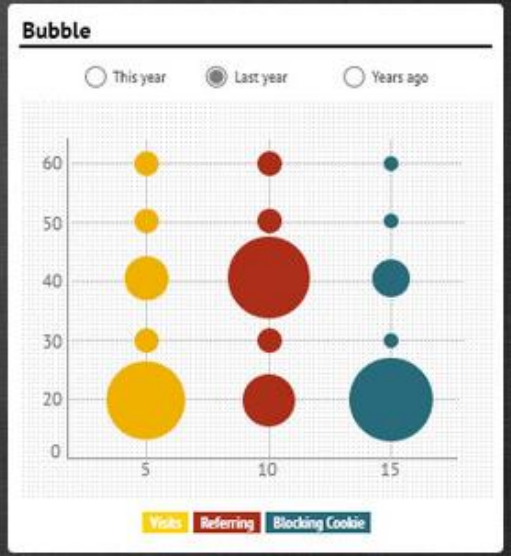
- Ejemplo http://bit.ly/1NEYTyy

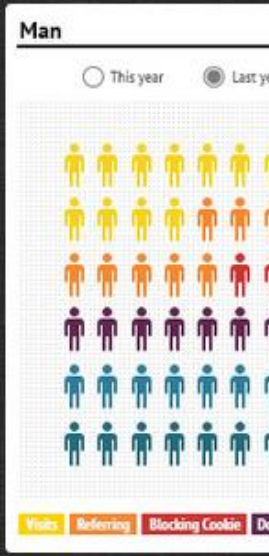
infogr.am

Chart type



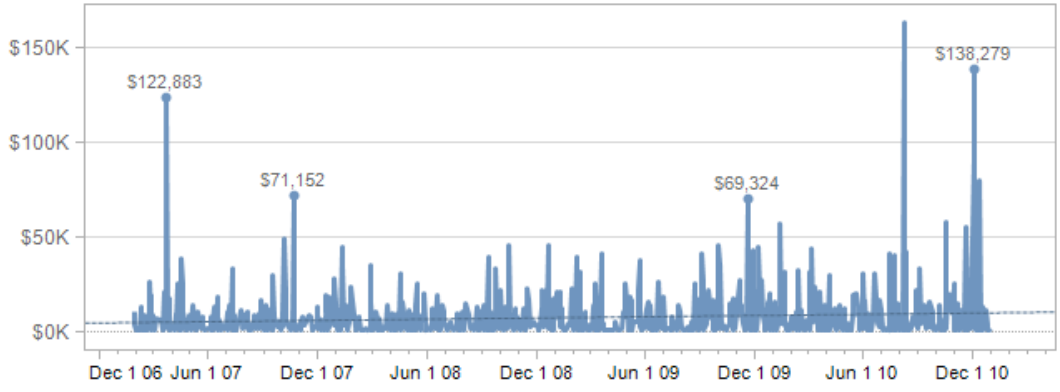






Use it

Daily Sales Dashboard



Date

1/27/2007 12/31/2010

Region

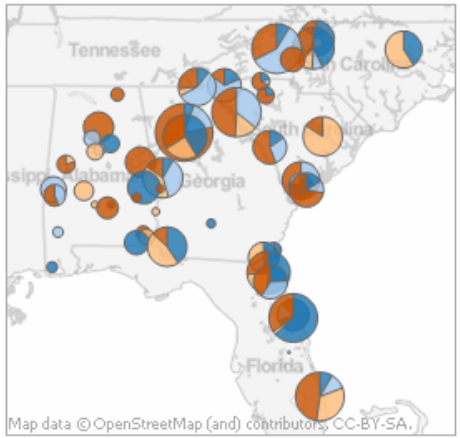
- (All)
- Central
- East
- South
- West

Customer Search Box

Customer Segment

- Bakeries
- Catering
- Distributors
- Restaurants

Sales by Segment



Customer Detail

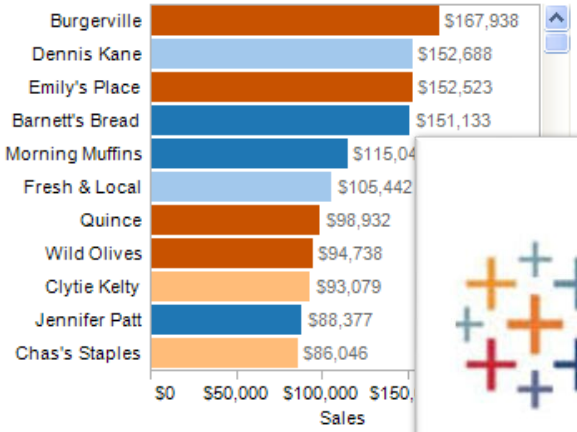
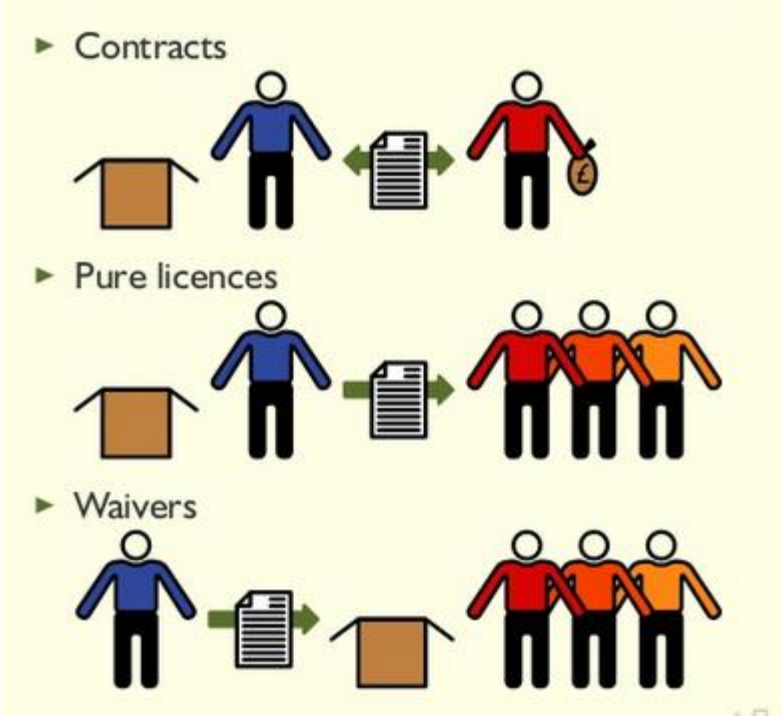


Imagen de https://www.flickr.com/photos/cogdog/

PUBLICAR Y COMPARTIR

Licencias



- Tipos
  - Creative Commons
  - Open Data Commons
  - Open Government Licence
  - GILF/AusGOAL Licences
  - Design Science Licence
  - Public Domain
- http://www.dcc.ac.uk/resources/how-guides/license-research-data
- http://www.ausgoal.gov.au/research-data-faqs

http://www.slideshare.net/AlexBall/ball2011ddh-slides

Publicar código

- Publica código
- Controla versiones
- Ejemplo de este curso
  http://bit.ly/1Hujl0I

IQSS
The Institute for Quantitative Social Science
HARVARD UNIVERSITY

HARVARD LIBRARY

Share, Cite, Reuse, Archive Research Data
Scientific data for reproducible research

POWERED BY THE Dataverse Network PROJECT v. 3.6.2

Harvard Dataverse Network

Create Account | Log In

Search this Dataverse Network | Search

Advanced Search | Tips

We're redesigning Dataverse and want your feedback! Please check out our Beta Site

The Harvard Dataverse Network is open to all scientific data from all disciplines worldwide. It includes the world's largest collection of social science research data. Learn more about the Dataverse Network.

Dataverses

Create Dataverse

732 Dataverses

A ***Dataverse*** is a container for research data studies, customized and managed by its owner.

RECENTLY RELEASED DATAVERSES

COS	Jun 26, 2014
Salomon & Cimpian: The Inherence Heuristic as a Source of Essentialist Thought	Jun 25, 2014
PSI/Laos	Jun 24, 2014
PSI/Nigeria	Jun 23, 2014
PSI/DR Congo	Jun 20, 2014

View More >

CfA Dataverses
Dataverses: 27
Harvard-Smithsonian Center for Astrophysics (CfA) and affiliated Dataverses

Studies

54,332 Studies, 744,247 Files, 1,031,873 Downloads

A ***study*** is a container for a research data set. It includes cataloging information, data files and complementary files.

RECENTLY RELEASED STUDIES

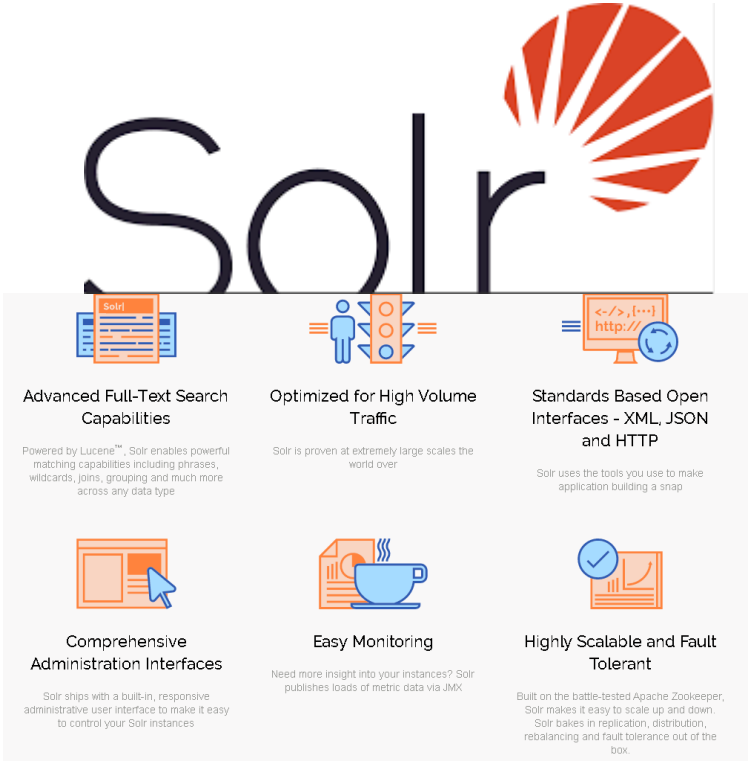
Brain Genomics Superstruct Project (GSP) by Buckner, Randy L.; Roffman, Joshua L. ; Smoller, Jordan W.	Jun 30, 2014
PRODES Brazil Weg Gis by Clarissa Gandour	Jun 30, 2014
Nicaragua (2013): Barreras que las PCV enfrentan para la adherencia al tratamiento y para el uso de condon. Costa Rica, El Salvador, Nicaragua y Panama. Estudio con Personas con VIH. by PSI; PASMO	Jun 30, 2014
El Salvador (2013): Barreras que las PCV enfrentan para la adherencia al tratamiento y para el uso de condon. Costa Rica, El Salvador, Nicaragua y Panama. Estudio con personas con VIH. by Nieto-Andrade Benjamin; Fortin Isolda; Alvarenga, Fredy Orlando; Palma, Carlos	Jun 30, 2014
Replication data for: Effects of transformation processes on plant species richness and diversity in homegardens of the Nuba Mountains, Sudan by Martin Wiehle; Sven Goenster; Jens Gebauer; Seifeldin Ali Mohamed; Andreas Buerkert; Katja Kehlenbeck	Jun 30, 2014

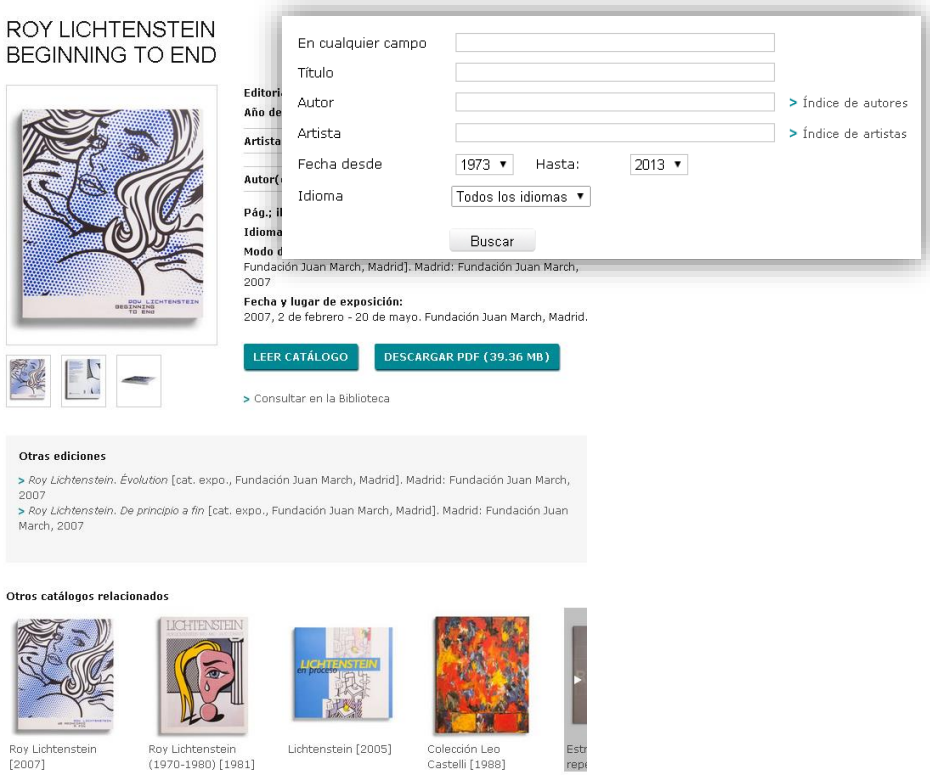
View More >

islandora

- Ejemplo
  - Creación de colección de fotos en http://sandbox.islandora.ca/

Motores de búsqueda y sistemas de recomendaciones






I DON'T KNOW HOW TO DO STATISTICS BUT IT DOESN'T MATTER BECAUSE I DIDN'T HAVE DATA.