

# 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*

# THE WORLD OF DATA

TWEETS  
PER  
DAY

50  
MILLION

TOTAL MINUTES  
SPENT ON  
FACEBOOK  
EACH MONTH

700  
BILLION

DATA SENT  
AND RECEIVED  
BY MOBILE  
INTERNET USERS

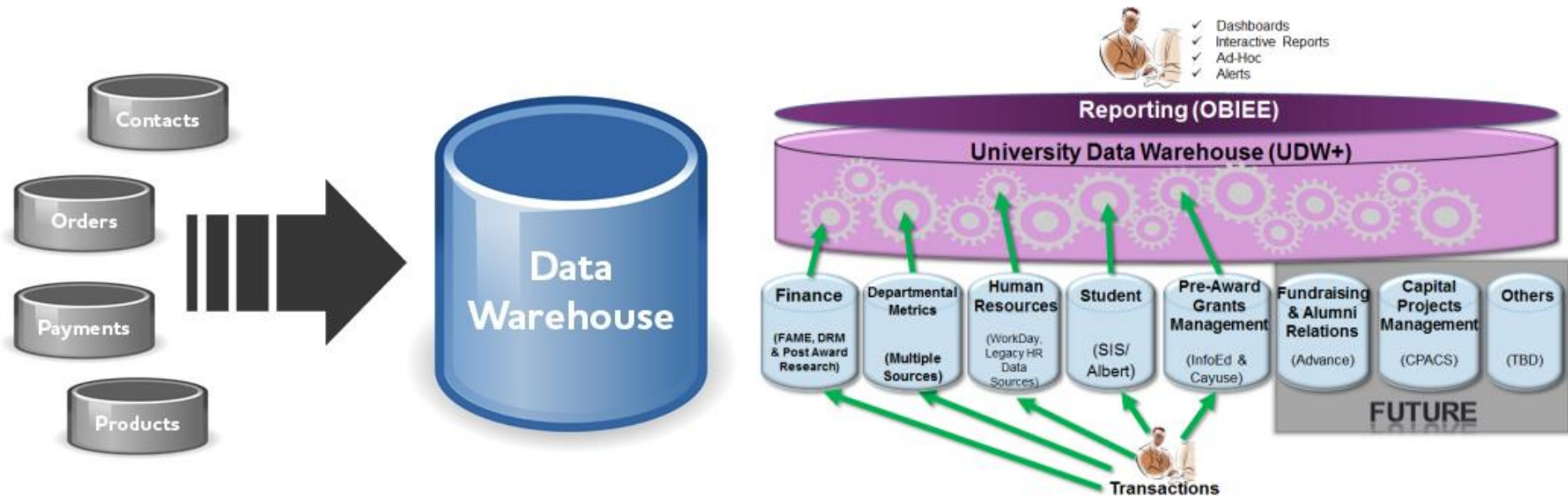
1.3  
EXABYTES

PRODUCTS  
ORDERED ON  
AMAZON PER  
SECOND

72.9  
ITEMS



# Datos en organizaciones y universidades

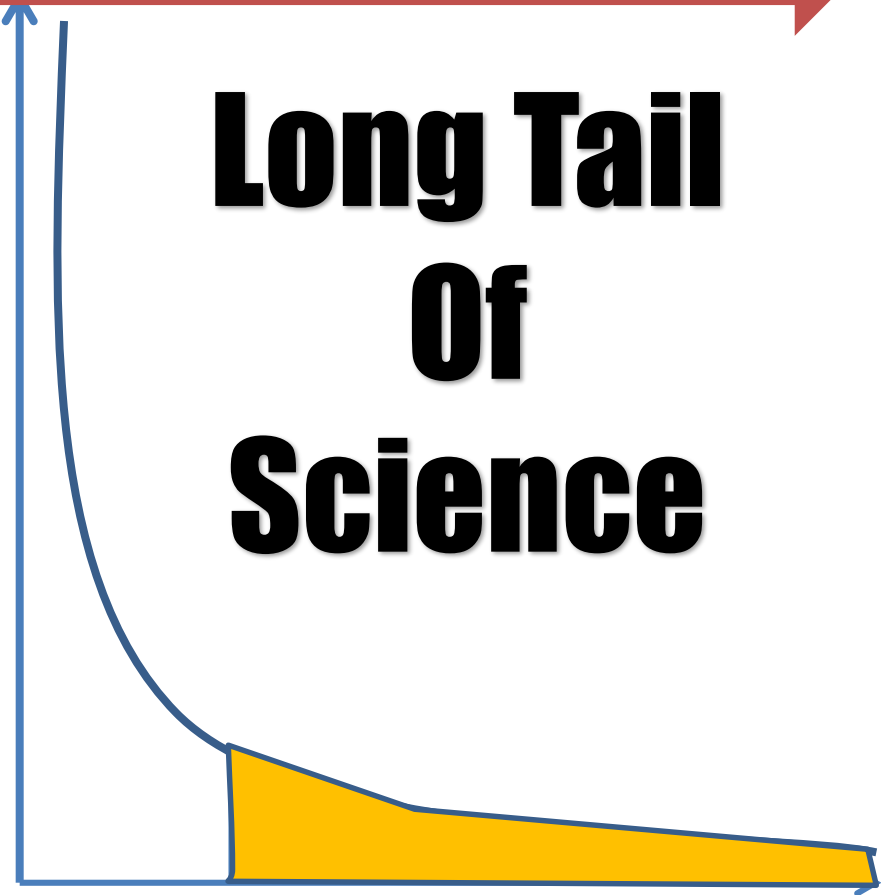


New York University Warehouse Plus

<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)

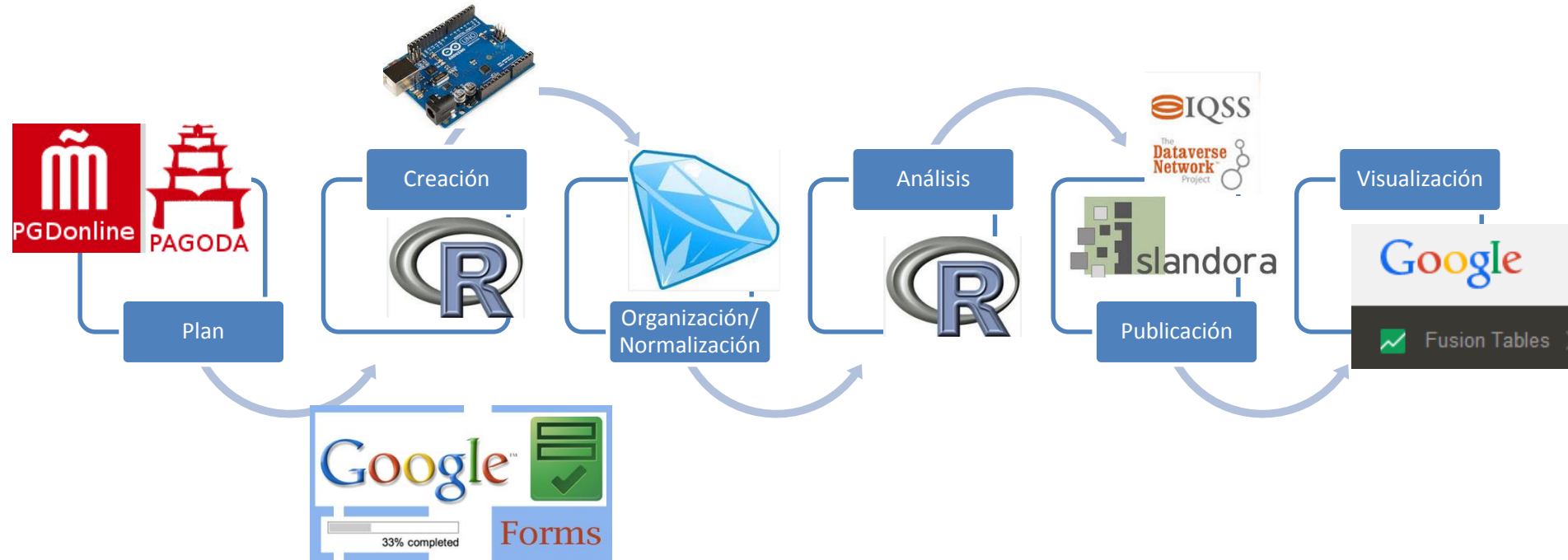
*“**requires, in all proposals** a supplementary document of no more than two pages describing **a Data Management Plan** for the proposed research. “*

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

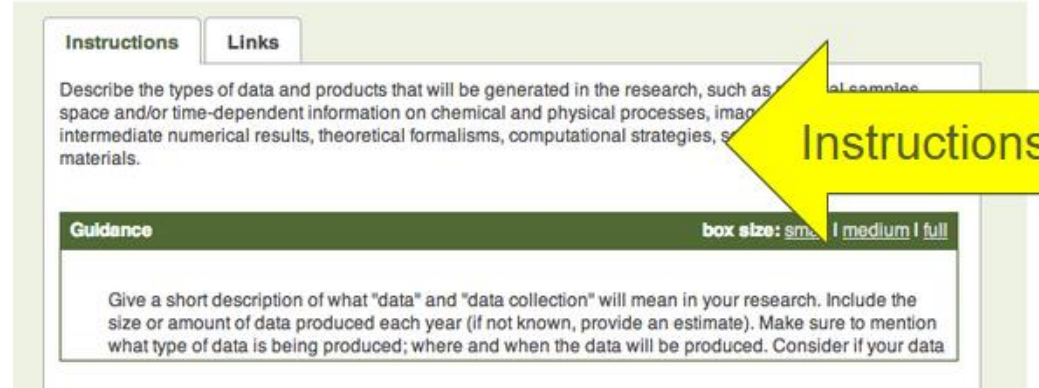
# Herramientas a lo largo del ciclo de vida de los datos





# Planes de gestión de datos

*“Documentos que describen que harás con tus datos durante tu investigación y una vez que termines con tu proyecto”*



The image shows a screenshot of a data management plan form. At the top, there are two tabs: 'Instructions' and 'Links'. The 'Instructions' tab is active, showing a text area with the following text: "Describe the types of data and products that will be generated in the research, such as ... al samples ... space and/or time-dependent information on chemical and physical processes, image ... intermediate numerical results, theoretical formalisms, computational strategies, ... materials." Below this is a 'Guidance' section with a green header. The guidance text reads: "Give a short description of what 'data' and 'data collection' will mean in your research. Include the size or amount of data produced each year (if not known, provide an estimate). Make sure to mention what type of data is being produced; where and when the data will be produced. Consider if your data". To the right of the 'Instructions' tab, a large yellow arrow points left towards the 'Instructions' text.

## 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

Registrado como Luis Martinez

Ver planes Crear un plan Acercas de Ayuda

## Mi proyecto (Horizon 2020)

No se ha respondido a las preguntas

(Rellenar todos los campos en inglés)

Detalles **PGD inicial** Revisión intermedia Revisión final Compartir Exportar

Para cada dataset, especifique lo siguiente (5 questions, 0 answered)

El PGD debe centrarse en los puntos siguientes dataset por dataset y debe reflejar el estado actual de la reflexión sobre los datos que se producirán

Nombre y referencia del dataset

Orientación

Se creará un identificador para el dataset

Guardar

Aún no respondido/a

Descripción del dataset

Orientación

Descripción de los datos que serán generados o recolectados, su origen

# CREAR Y ALMACENAR DATOS



# Herramientas de encuestas





- Instrucciones Google Forms  
<http://bit.ly/1Luobgj>
- Creamos una encuesta usando la herramienta <http://bit.ly/17Jdmlx>

## Encuesta proyecto investigación

Form Description

**Nombre y apellidos\***

Tu nombre y apellidos

**Universidad\***

Institución

- Universidad Complutense de Madrid
- Universidad de Barcelona
- Universidad de Córdoba
- Otra

**Departamento\***

Departamento de la Universidad

- Investigación Operativa
- Matemáticas
- Estadística

# 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](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.scrapewiki.com/2011/06/knight-foundation-finance-scrapewiki-for-journalism/>



# Organizar los datos

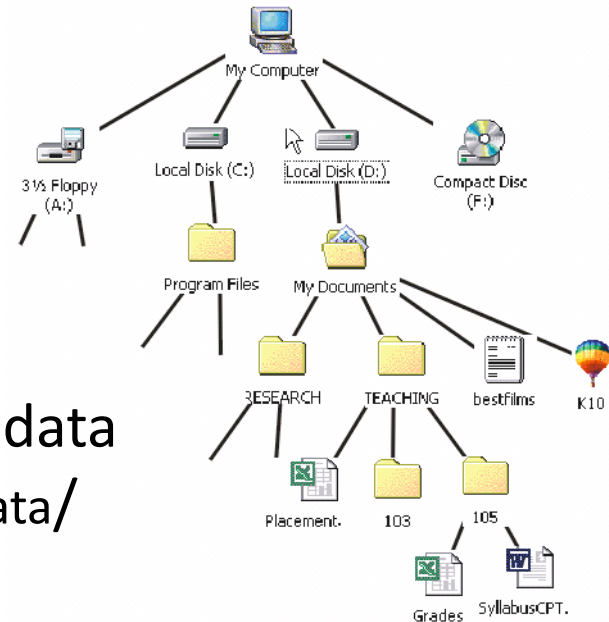
- Ficheros de datos y carpetas tiene que denotarse y organizarse de forma sistemática.

- 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



- Curso de metadatos MANTRA

[http://datalib.edina.ac.uk/mantra/documentation\\_metadata\\_citation/](http://datalib.edina.ac.uk/mantra/documentation_metadata_citation/)

## WHAT IS METADATA?

Metadata is **data about data**.

Metadata can describe a single piece of data, a dataset or collection.

Metadata can be used to describe *anything* - both physical or digital.



- Estándares

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

### Search by Discipline



Biology



Earth Science



General Research Data



Physical Science



Social Science & Humanities

# TRANSFORMAR E INTEGRAR





# Open Refine

Google refine

A power tool for working with messy data.

Create Project

Open Project

Import Project

« Start Over

Configure Parsing Options

## Assignee/Applicant

1.	Galen (Chemicals) Limited,Dublin,IE
2.	Aplus Flash Technology Inc.,Saratoga,CA,US
3.	MACEACHERN; WILLIAM A
4.	Eastman Kodak Company,Rochester,NY,US
5.	SIEGER; ARLETTE
6.	Daniels Pull Plow Inc.,East Dundee,IL,US
7.	Telecom Medical Inc.,San Francisco,CA,US
8.	Sabre Oxidation Technologies Inc.,Odessa,TX,US
9.	Rumber Materials Inc.,Austin,TX,US
10.	Priority Call Management Inc.,Wilmington,MA,US

## Parse data as

### CSV / TSV / separator-based files

Line-based text files

Fixed-width field text files

PC-Axis text files

JSON files

RDF/N3 files

XML files

Open Document Format spreadsheets  
(.ods)

RDF/XML files

Character encoding

Columns are separated

commas (CSV)

tabs (TSV)

custom \t

Escape special charact

- Tutoriales básicos

<http://bit.ly/1zYWGn8>

<http://bit.ly/1B7pqRA>

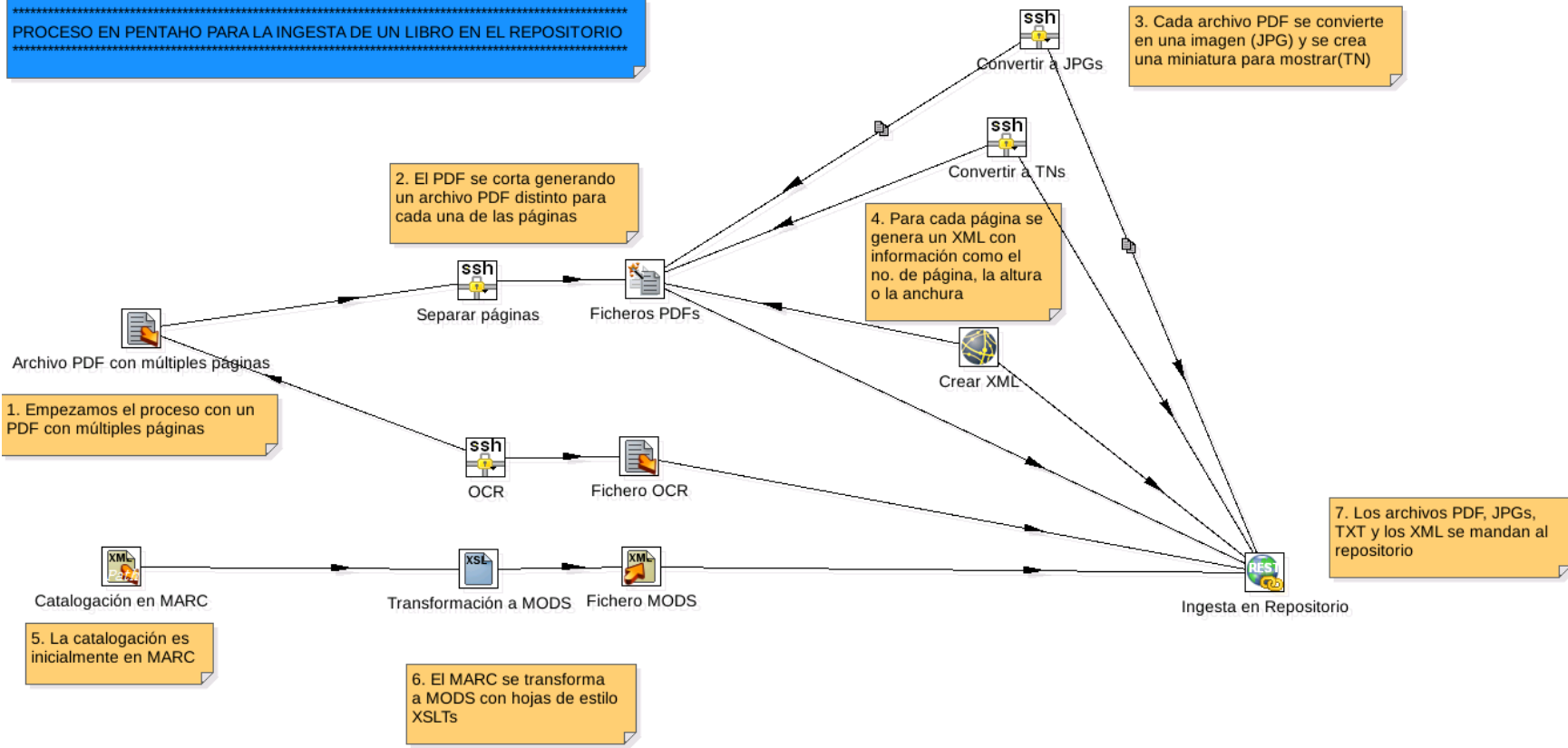
- Demostración de desambiguación de nombres de persona



Version 2.5 [r2407]

Help  
About

## PROCESO EN PENTAHO PARA LA INGESTA DE UN LIBRO EN EL REPOSITORIO



# ANALIZAR Y VISUALIZAR

**UCM Matemáticas**

- tesis Durán ✓
- 2014/2015 Plan

**DLib**

- texto ✓
- imágenes ✓

**Prototipo Sit** todos los años

- diseño ✓

**IASISIST**

**FJM Isterdix.NET**

**Datos Música**

- mapas
- timeline

**WORLD OF DATA** → **SCIENTIF DATA** → **INTRO DATA**

**DATA LIBRARIAN**

- EDS 70s
- Access + Support
- Data Support
- Davis, STS, STAB, ARES

**RESEARCH DATA MAN**

- OA + Early Access
- Papers, Plus, Legal deposit
- Data Library
- Papers, Web Enquiry

**DATA CURATOR**

- Digitalization
- Metadata, preservation, harmonization
- Digital Metadata/Tools
- ETL

**DATA SCIENTIST**

- "Big data"
- ?
- OED Data Science
- Analytics, Visualization

**EVOLUCIÓN**

Evolutiva  
Profunda  
FJM app

**VISION AMPLIA**

SECCIONES PERDIDAS + METADATOS  
LIBRARIAS FALTA

**OPORTUNIDAD**  
BIBLIOTECA  
PERFORMANCE

**CONTEXTO**

**JOB DESCRIPTION**

**FJM**

**SKILLS**

**CONCLUSIONES**

- se pueden hacer cosas

**EVENTOS**

- + BIREGIAL - PORTO
- 13-18 Octubre
- + Open Access Week
- 23 Octubre
- + Open Repositories - G
- Mayo

**FORMACION**

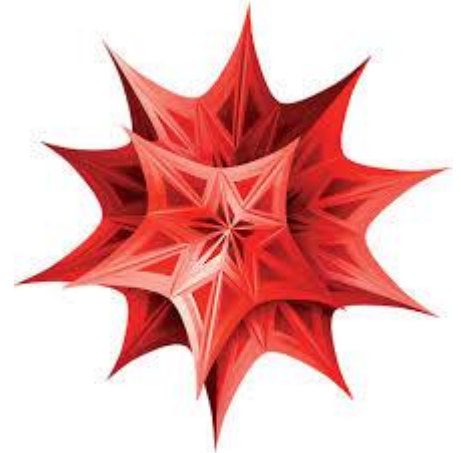
- + Data Science
- cursos, Jobs H
- + Repositorios
- Solo
- + Big data
- Hadoop

**CAJAS DE TEXTO**

1. INTRODUCCIÓN  
2. DESCRIPCIÓN  
3. DESCRIPCIÓN DE DATOS  
4. DESCRIPCIÓN DE METADATOS  
5. DESCRIPCIÓN DE METADATOS

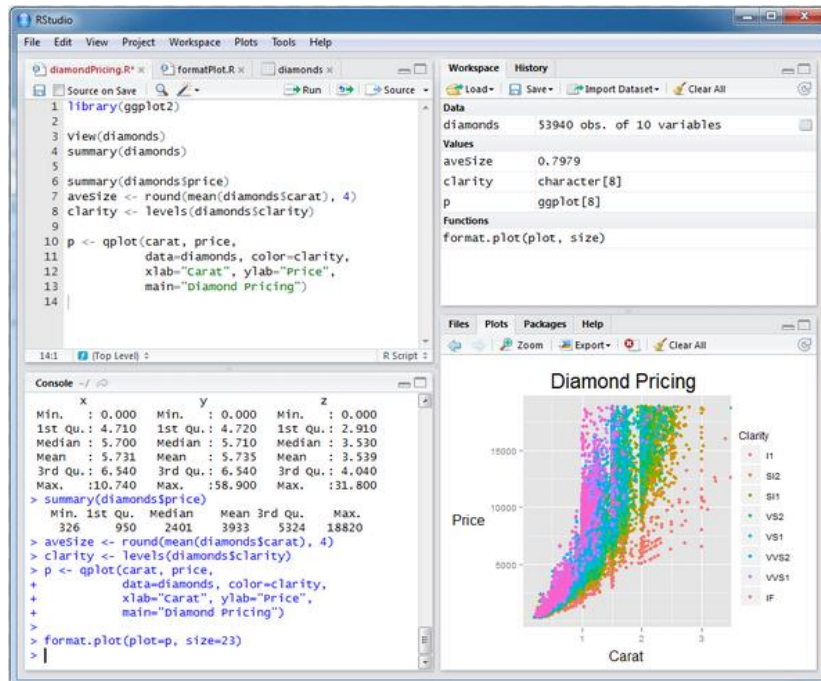
**CONCLUSIONES**

• se pueden hacer cosas





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







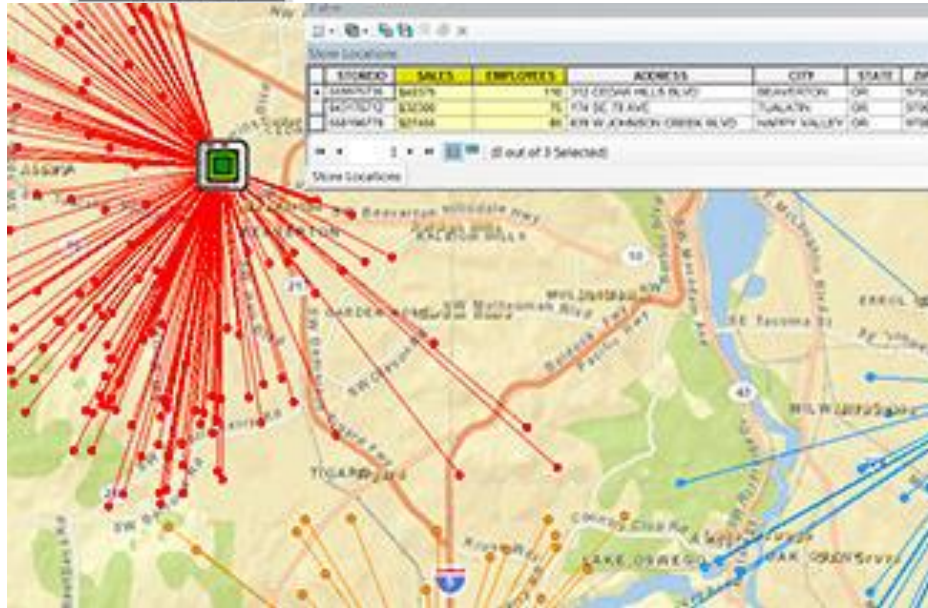


# ArcGIS®

ESRI



## atlas.ti



atlas.ti P.6: Happiness quotes

Think, Act, Be  
Think happy  
Act happy  
Be happy  
David Barry

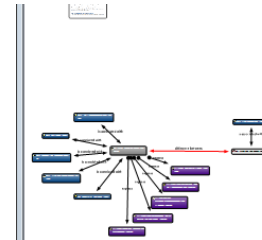
audio wave form

preview images

media controls

Preview of full video

Quotes: 616



Explaining the research findings (13)

Creator: Susanne

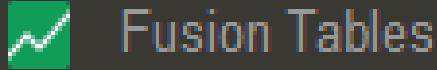
Created: 07.03.2012 08:55:18

Size: 13

Count: 18

right-click to open context menu

- Video Preview
- Show Audio Waveform
- Show Time Indications
- Combine Audio Channels
- Left Audio Channel
- Right Audio Channel
- Waveform Zoom Factor
- Auto-Scroll Margin
- Show Quotation Overlay



- Datos de paro por municipio  
<http://bit.ly/1HnUtl2>
- Datos de población por municipio  
<http://bit.ly/1xahmbX>
- Demostración de cómo representarlos  
usando Google Fusion Tables  
<http://bit.ly/1Adqyxa>





# Timelines

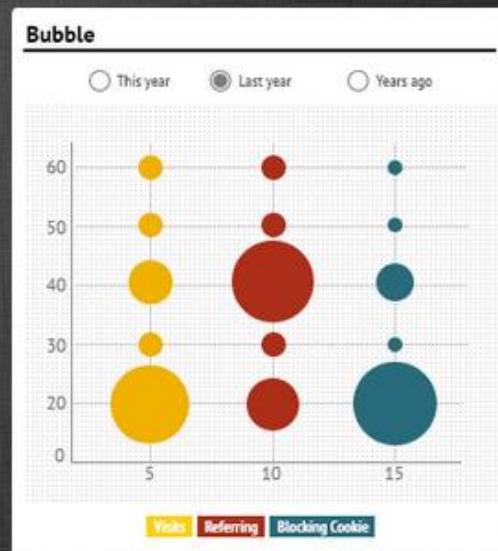
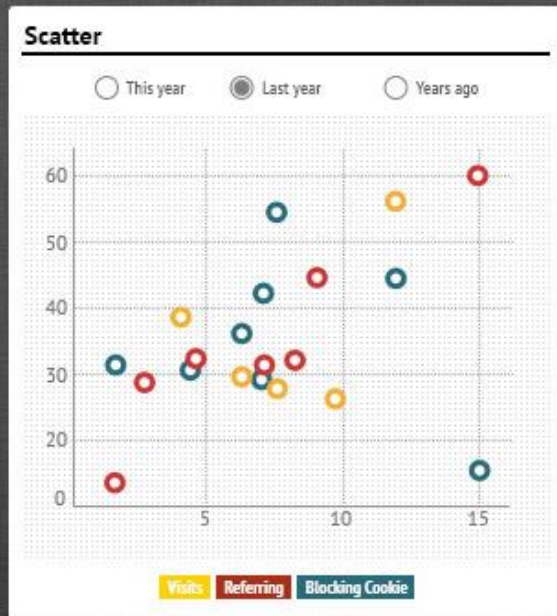
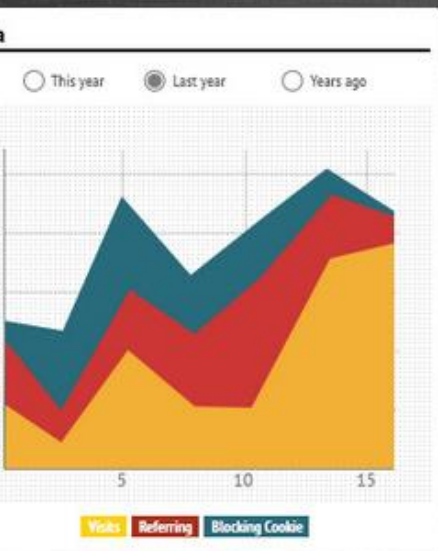
**Timeline** <sup>JS</sup>

Beautifully crafted timelines that are easy and intuitive to use.

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

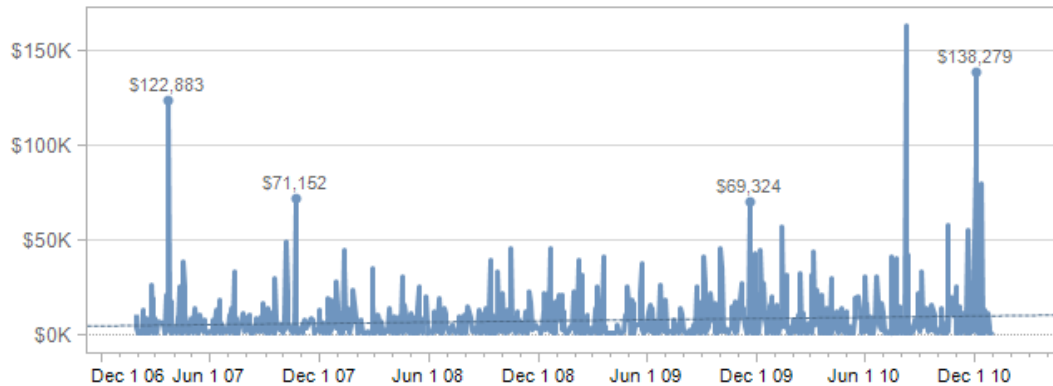


# Chart type



Use it

# Daily Sales Dashboard



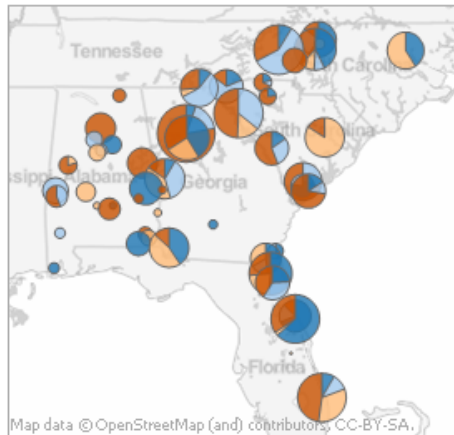
**Date**  
 1/27/2007 12/31/2010

**Region**  
 (All)  
 Central  
 East  
 South  
 West

**Customer Search Box**

**Customer Segment**  
 Bakeries  
 Catering  
 Distributors  
 Restaurants

## Sales by Segment



## Customer Detail

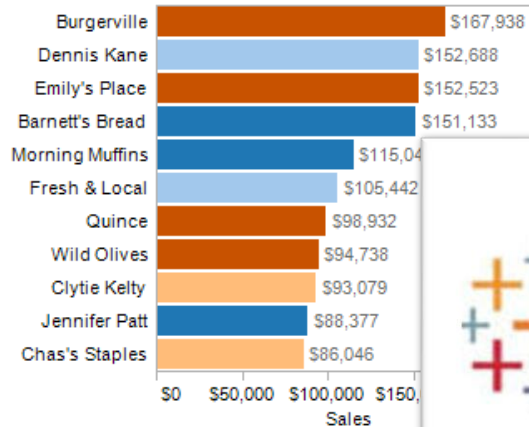


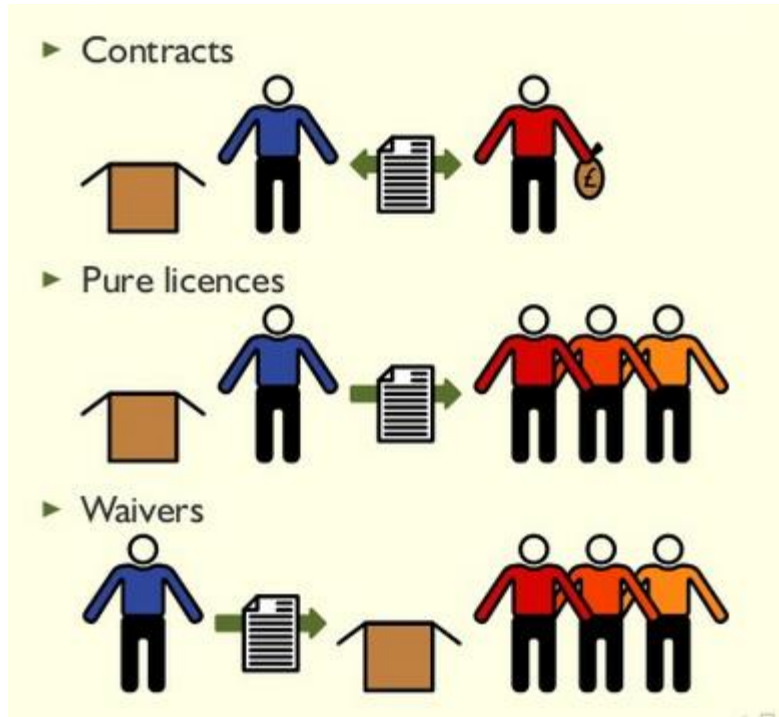
Imagen de <https://www.flickr.com/photos/cogdog/>

Life is sharing

**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>

# Publicar código

The screenshot shows a GitHub repository page for 'luismart / CursoDatosOSRepositorios'. The repository has 10 commits, 1 branch, 0 releases, and 1 contributor. The current branch is 'master'. The repository was signed-off by luismart <l.martinezuribe@gmail.com> 9 days ago. The latest commit is 56e75b4cf3. The repository contains files: Ejemplo Web Scraping.R (9 days ago), NombresOpenRefine.txt (15 days ago), and README.md (14 days ago). The README.md file is open, showing the title 'Taller de Datos - Data Toolbox [WORKING PROGRESS]' and the subtitle 'XIV Workshop Rebiun -VI Jornadas OS-Repositorios'. The text in the README describes the workshop's focus on data management and sharing in research libraries.

This repository

Search

Explore Gist Blog Help

luismart / CursoDatosOSRepositorios

Unwatch 1

Description Website

Short description of this repository

Website for this repository (optional)

Save or Cancel

10 commits 1 branch 0 releases 1 contributor

branch: master CursoDatosOSRepositorios / +

Signed-off-by: luismart <l.martinezuribe@gmail.com>

luismart authored 9 days ago latest commit 56e75b4cf3

Ejemplo Web Scraping.R	Signed-off-by: luismart <l.martinezuribe@gmail.com>	9 days ago
NombresOpenRefine.txt	Signed-off-by: luismart <l.martinezuribe@gmail.com>	15 days ago
README.md	Signed-off-by: luismart <l.martinezuribe@gmail.com>	14 days ago

README.md

## Taller de Datos - Data Toolbox [WORKING PROGRESS]

### XIV Workshop Rebiun -VI Jornadas OS-Repositorios

Tras varios años tratando sobre las posibles labores de las bibliotecas con los datos de investigación, las políticas de la Unión Europea empiezan a tener su efecto. Estas políticas se han cristalizado en la necesidad de presentar planes de gestión de datos con las propuestas de proyectos de investigación. Los investigadores van a necesitar planificar cómo gestionar sus datos, dejando documentando qué datos y cómo van a crearse, la manera en que se van a almacenar y además de la forma en que se van a compartir o publicar. Las bibliotecas en muchos países ya apoyan en la elaboración de estos planes además de ofrecer una amplia variedad de servicios con datos

- Publica código
  - Controla versiones
  - Ejemplo de este curso
- <http://bit.ly/1Hujl0l>

## Harvard Dataverse Network



Create Account

Log In

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

**i** 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 Inference 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

**i** 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
<b>i</b> PRODES Brazil Weg Gis by Clarissa Gandour	Jun 30, 2014
<b>i</b> 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
<b>i</b> 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 >](#)



## Basic Image Collection

View

Manage



Grid view List view



Among the Big Guns,  
Camp Brighton



Cardigan Rural  
Setting



Farmers Bank of  
Rustico / Le Banque  
de Rustico



Seafarers Haven,  
Prince Edward Island,  
Canada

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

# Motores de búsqueda y sistemas de recomendaciones



## Advanced Full-Text Search Capabilities

Powered by Lucene™, Solr enables powerful matching capabilities including phrases, wildcards, joins, grouping and much more across any data type



## Optimized for High Volume Traffic

Solr is proven at extremely large scales the world over



## Standards Based Open Interfaces - XML, JSON and HTTP

Solr uses the tools you use to make application building a snap



## Comprehensive Administration Interfaces

Solr ships with a built-in, responsive administrative user interface to make it easy to control your Solr instances



## Easy Monitoring

Need more insight into your instances? Solr publishes loads of metric data via JMX



## Highly Scalable and Fault Tolerant

Built on the battle-tested Apache Zookeeper, Solr makes it easy to scale up and down. Solr takes in replication, distribution, rebalancing and fault tolerance out of the box.

## ROY LICHTENSTEIN BEGINNING TO END



En cualquier campo

Título

Editorial

Año de publicación  > índice de autores

Artista  > índice de artistas

Fecha desde  Hasta:

Idioma

Página:

Idioma

Modo de búsqueda

Fundación Juan March, Madrid]. Madrid: Fundación Juan March, 2007

**Fecha y lugar de exposición:**  
2007, 2 de febrero - 20 de mayo. Fundación Juan March, Madrid.

[LEER CATÁLOGO](#)

[DESCARGAR PDF \(39.36 MB\)](#)

> Consultar en la Biblioteca

## Otras ediciones

> *Roy Lichtenstein. Evolution* [cat. expo., Fundación Juan March, Madrid]. Madrid: Fundación Juan March, 2007

> *Roy Lichtenstein. De principio a fin* [cat. expo., Fundación Juan March, Madrid]. Madrid: Fundación Juan March, 2007

## Otros catálogos relacionados



Roy Lichtenstein [2007]



Roy Lichtenstein [1970-1980] [1981]



Lichtenstein [2005]



Colección Leo Castelli [1988]

Estr  
repi

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

