

Facilitando el acceso computacional a colecciones digitales

Universidad de Alicante

Gustavo Candela, Pilar Escobar, María Dolores Sáez



Facilitando el acceso computacional a colecciones digitales



Universitat d'Alacant
Universidad de Alicante



BIBLIOTECA VIRTUAL
MIGUEL DE CERVANTES

With many thanks to:

Mahendra Mahey, Sally Chambers, Abbey Potter, Kristy Kokegei, Sarah Ames, Sophie Wagner, Thomas Padilla, Milena Dobрева-McPherson, Caleb Derven, Ditte Laursen, Armin Straube, Caleb Derven, Katrine Gasser, Aisha Al-Abdulla, Lotte Wilms, Paula Bray, Tim Sherratt, Juan Carlos García, Patricia Murrieta-Flores, Javier Pereda, David Abián, Wikimedia Spain and Deutschland

Facilitando el acceso computacional a colecciones digitales

- Parte 1
 - Introducción
 - Proyecto Jupyter
 - Conclusiones
- Parte 2
 - Ejecución de Jupyter Notebooks



Introducción

¿Quiénes somos?

- Dr. Gustavo Candela
- Dr. María Pilar Escobar
- MSC. María Dolores Sáez (Estudiante doctorado)
- Prof. Manuel Marco Such



Universitat d'Alacant
Universidad de Alicante

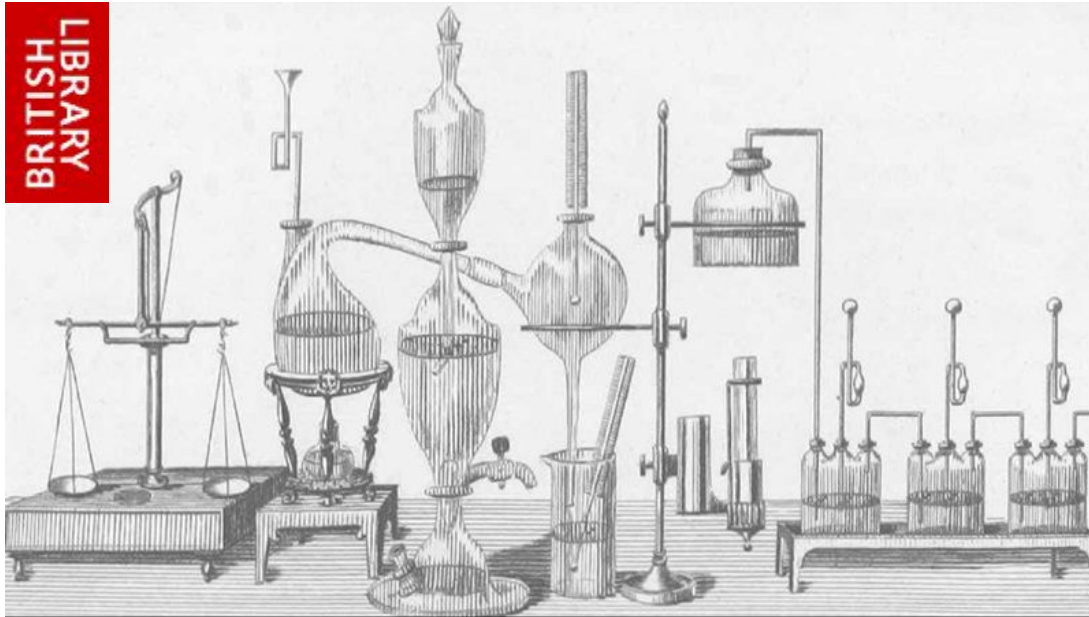


BIBLIOTECA VIRTUAL
MIGUEL DE CERVANTES



Introducción

GLAM Lab: Reutilizando colecciones digitales de forma innovadora y creativa



<https://www.bl.uk/projects/british-library-labs>



<https://glamlabs.io/books/>
<http://www.cervantesvirtual.com/obra/open-a-glam-lab-1066249/>

Introducción

International GLAM Labs Community



¿Te interesan los GLAM labs?

<https://glamlabs.io/>

Introducción

GLAM Workbench



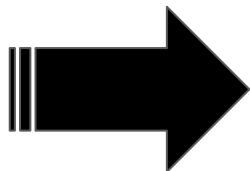
<https://glam-workbench.net/>

Sherratt, Tim, Wilms, Lotte, & Lingstadt, Kirsty. (2020, April). LIBER Webinar: Setting Up A GLAM Workbench In Your Library. Zenodo. <http://doi.org/10.5281/zenodo.3743193>

Introducción

Collections as Data

- Documentos de texto
- Imágenes
- Metadatos
- Application Programming Interface (APIs)
- Datos enlazados (Linked Open Data)

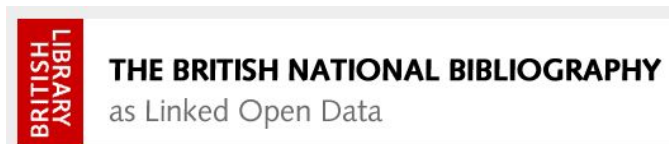
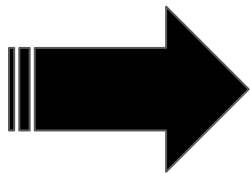


<https://doi.org/10.5281/zenodo.3152935>

<https://collectionsasdata.github.io/fiftythings/>

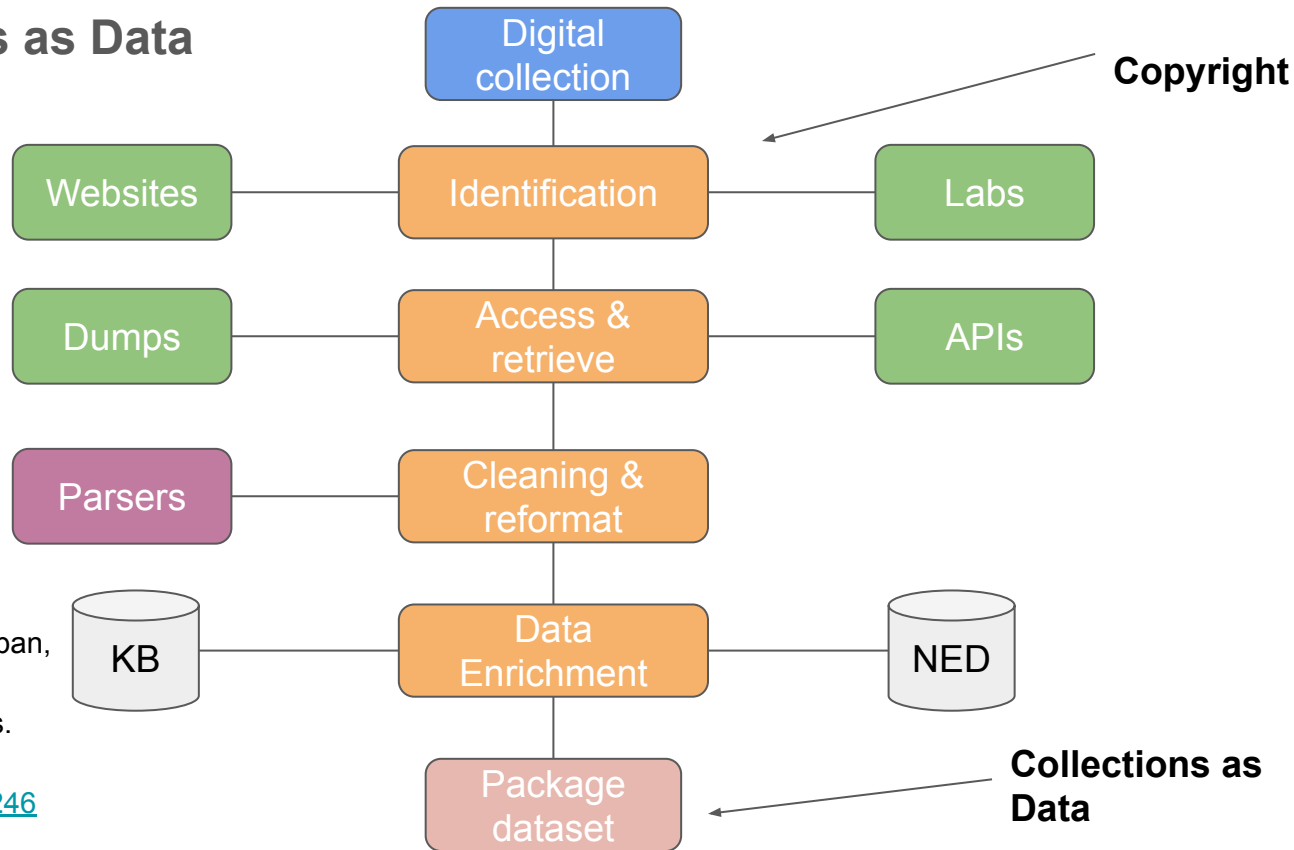
Introducción

Candela, G., Sáez, M. D., Escobar Esteban, Mp., & Marco-Such, M. (2020). Reusing digital collections from GLAM institutions. *Journal of Information Science*. <https://doi.org/10.1177/0165551520950246>



Introducción

Creando Collections as Data



Candela, G., Sáez, M. D., Escobar Esteban, Mp., & Marco-Such, M. (2020). Reusing digital collections from GLAM institutions.

Journal of Information Science.

<https://doi.org/10.1177/0165551520950246>

Jupyter Notebooks

Proyecto Jupyter

- Software de **código abierto**
- Facilita un **interfaz** sencillo de edición y ejecución
- Combina **documentación textual con código ejecutable**
- Herramienta **colaborativa, reproducible** y reutilizable
- Permite el uso de **lenguajes de programación** como Python y Java



Jupyter Notebooks

The image shows a Jupyter Notebook interface with several annotations. The interface includes a top bar with the Jupyter logo, 'Index (autosaved)', and buttons for 'Join this repo's Video Chat' and 'Visit repo'. Below this is a menu bar with 'File', 'Edit', 'View', 'Insert', 'Cell', 'Kernel', 'Widgets', and 'Help'. A toolbar contains icons for adding, deleting, and running cells, along with a dropdown menu set to 'Markdown' and buttons for 'Download', 'GitHub', and 'Binder'. The main content area displays a 'Welcome to Jupyter!' message and a list of links for further learning.

Añadir /eliminar celdas (An arrow points to the '+' icon in the toolbar)

celdas (An arrow points to the main content area)

ejecutar (An arrow points to the 'Run' button in the toolbar)

texto o código? (An arrow points to the 'Markdown' dropdown menu in the toolbar)

Index (autosaved) (Text in the top bar)

Join this repo's Video Chat (Button in the top bar)

Visit repo (Button in the top bar)

File Edit View Insert Cell Kernel Widgets Help (Menu bar)

Trusted (Text in the top bar)

Run (Button in the toolbar)

Download (Button in the toolbar)

GitHub (Button in the toolbar)

Binder (Button in the toolbar)

Memory (Text in the top bar)

Welcome to Jupyter!

This repo contains an introduction to [Jupyter](#) and [IPython](#).

Outline of some basics:

- [Notebook Basics](#)
- [IPython - beyond plain python](#)
- [Markdown Cells](#)
- [Rich Display System](#)
- [Custom Display logic](#)
- [Running a Secure Public Notebook Server](#)
- [How Jupyter works](#) to run code in different languages.

You can also get this tutorial and run it on your laptop:

```
git clone https://github.com/ipython/ipython-in-depth
```

Install IPython and Jupyter:

with [conda](#):

Jupyter Notebooks

[GLAM WorkBench](#)

[LC Maps for Robots](#)

[BL Labs](#)

[Archives Unleashed
Notebooks](#)

[LC - Newspapers
navigator](#)

[National Library of
Scotland](#)

[ONB Labs](#)

[BVMC](#)



Reproducible

Popular en la
comunidad

Código

Ejecutable en la nube



Fácil de
utilizar

No es necesario
escribir código

Documentación

Herramienta
pedagógica

Jupyter Notebooks

Estructura general

- Introducción (colección, licencia, etc.)
- Importación de librerías de software
- Descarga de la colección digital
- Preprocesamiento de datos
- Análisis

Topic Modeling based on Digitised Volumes of theatrical English, Scottish, and Irish playbills between 1600 - 1902 from data.bl.uk

Topic Models are a type of statistical language models used for discovering hidden structure in a collection of texts.

This example is based on a dataset that comprises 264 volumes of digitised theatrical playbills published between 1660 – 1902 (mostly 19th century) from England, Scotland, Wales and Ireland. Digitised from the British Library's physical collection of over 500 volumes of playbills, the dataset contains text files in Optical Character Recognition (OCR) format. More information about the dataset at <https://data.bl.uk/playbills/>

Setting up things

```
] : import sys
import requests
import pandas as pd
import re
import gensim
from gensim.utils import simple_preprocess
from nltk.corpus import wordnet
from nltk.tokenize import word_tokenize
from nltk.stem.porter import PorterStemmer
```

Jupyter Notebooks

¿Qué necesitamos para su ejecución?

- Un navegador

Usuarios avanzados

- Repositorio de software: [GitHub](#)
- Plataforma de ejecución en la nube: [MyBinder](#)
- [Proyecto Jupyter](#)
- [Google Colab](#) vs MyBinder
- [Zenodo](#), [figshare](#)



GitHub



colab

Jupyter Notebooks

The notebooks are available in GitHub classified by type of project: [images](#), [Linked Open Data](#), and [metadata and text](#). Additional notebooks provided by [ONB Labs](#) and developed by Stefan Karner, have been integrated into the collection.

They are also citable (in Zenodo) and have been assigned a DOI.

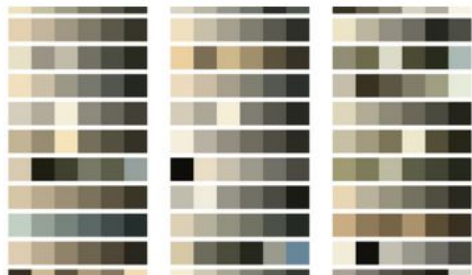
DOI [10.5281/zenodo.3765188](https://doi.org/10.5281/zenodo.3765188)

DOI [10.5281/zenodo.3765186](https://doi.org/10.5281/zenodo.3765186)

DOI [10.5281/zenodo.3765184](https://doi.org/10.5281/zenodo.3765184)

In order to launch the notebooks in the cloud, click on the Binder button. Each notebook is based on a dataset provided by a GLAM institution.

Have fun!



Binder



Binder

| | subject | count |
|----|---|-------|
| 0 | amateur | 143 |
| 1 | Glasgow | 690 |
| 2 | Transport | 548 |
| 3 | Leisure and Recreation | 518 |
| 4 | television news | 471 |
| 5 | Edinburgh | 466 |
| 6 | amateur | 417 |
| 7 | local topical | 392 |
| 8 | Celebrations, Traditions and Customs | 358 |
| 9 | Employment, Industry and Industrial Relations | 344 |
| 10 | documentary | 340 |
| 11 | Sporting Activities | 333 |

Binder

<http://data.cervantesvirtual.com/blog/notebooks/>

Jupyter Notebooks

Thanks to Google Cloud, OVH, GESIS Notebooks and the Turing Institute for supporting us! 🐛



Starting repository: hibernator11/notebook-texts-example
/master

Your launch may take longer the first few times a repository is used. This is because our machine needs to create your environment.

Build logs

show

<http://data.cervantesvirtual.com/blog/notebooks/>

Jupyter Notebooks



[Visit repo](#) [Copy Binder link](#) [Quit](#)

Files [Running](#) [Clusters](#)

Select items to perform actions on them.

[Upload](#) [New](#) [Refresh](#)

| <input type="checkbox"/> 0 | / | Name | Last Modified | File size |
|----------------------------|---------------------------------|------|---------------|-----------|
| <input type="checkbox"/> | images | | hace 8 meses | |
| <input type="checkbox"/> | Moving-Image-Archive | | hace 8 meses | |
| <input type="checkbox"/> | playbills-ocr-text | | hace 8 meses | |
| <input type="checkbox"/> | dataset-extraction-images.ipynb | | hace 8 meses | 8.42 kB |
| <input type="checkbox"/> | topic-modeling-billing.ipynb | | hace 8 meses | 12 kB |
| <input type="checkbox"/> | datapackage.zip | | hace 8 meses | 2.52 MB |
| <input type="checkbox"/> | LICENSE | | hace 8 meses | 35.1 kB |
| <input type="checkbox"/> | marc_records.csv | | hace 8 meses | 2.52 MB |
| <input type="checkbox"/> | README.md | | hace 8 meses | 2.1 kB |
| <input type="checkbox"/> | requirements.txt | | hace 8 meses | 118 B |
| <input type="checkbox"/> | runtime.txt | | hace 8 meses | 11 B |

<http://data.cervantesvirtual.com/blog/notebooks/>

Jupyter Notebooks

Topic modeling

BRITISH LIBRARY
data.bl.uk



Map visualization



BRITISH LIBRARY
THE BRITISH NATIONAL BIBLIOGRAPHY
as Linked Open Data

MARC extraction



| | subject | count |
|----|---|-------|
| 0 | amateur | 1438 |
| 1 | Glasgow | 690 |
| 2 | Transport | 548 |
| 3 | Leisure and Recreation | 518 |
| 4 | television news | 471 |
| 5 | Edinburgh | 466 |
| 6 | amateur | 417 |
| 7 | local topical | 392 |
| 8 | Celebrations, Traditions and Customs | 356 |
| 9 | Employment, Industry and Industrial Relations | 344 |
| 10 | documentary | 340 |
| 11 | Sporting Activities | 332 |

Jupyter Notebooks

Ventajas

- Facilita el tratamiento de las **colecciones publicadas** por instituciones GLAM
- Elimina **barreras de entrada** para los usuarios con menos formación en el uso de la tecnología



- Permite la aplicación de **métodos de investigación**
- **Guía a los usuarios** y permite la edición y ejecución del código

Jupyter Notebooks

¿Dónde encontrar colecciones de datos?

| Institución | Colección | URL |
|----------------------------------|----------------------------|---|
| Bibliothèque nationale de France | BnF API et jeux de données | http://api.bnf.fr/ |
| British Library | data.bl.uk | https://data.bl.uk/ |
| Det Kgl. Bibliotek | KB Labs | https://labs.kb.dk/ |
| National Library of Netherlands | KB Lab | https://lab.kb.nl/ |
| National Library of Scotland | Data Foundry | https://data.nls.uk/ |
| Library of Congress | LC for Robots | https://labs.loc.gov/lc-for-robots/ |

Conclusiones

Impacto

- [British Library Labs Public Awards 2020](#)
- Presentación en [Research Libraries UK](#)
- Artículo Journal of Information Science <http://rua.ua.es/dspace/handle/10045/109460>
- Tutorial Programming Historian [sobre GLAM Labs](#)
- Colaboración con [Lancaster University Digital Humanities Hub](#)



Conclusiones

- Notebooks pueden servir como herramienta pedagógica
- Las instituciones GLAM están comenzando a adoptar **Collections as Data**
- Existen multitud de formas de hacer públicas las colecciones digitales
- **GLAM Labs** son el espacio ideal para recolectar colecciones digitales
- **Copyright** debe ser claro ([Creative Commons Public Domain](#))
- ¿Estamos reutilizando las colecciones digitales de forma eficiente?
¿Investigadores externos las usan?

Facilitando el acceso computacional a
colecciones digitales

Muchas gracias

gcandela@ua.es, mpilar.escobar@ua.es,
md.saez@ua.es

