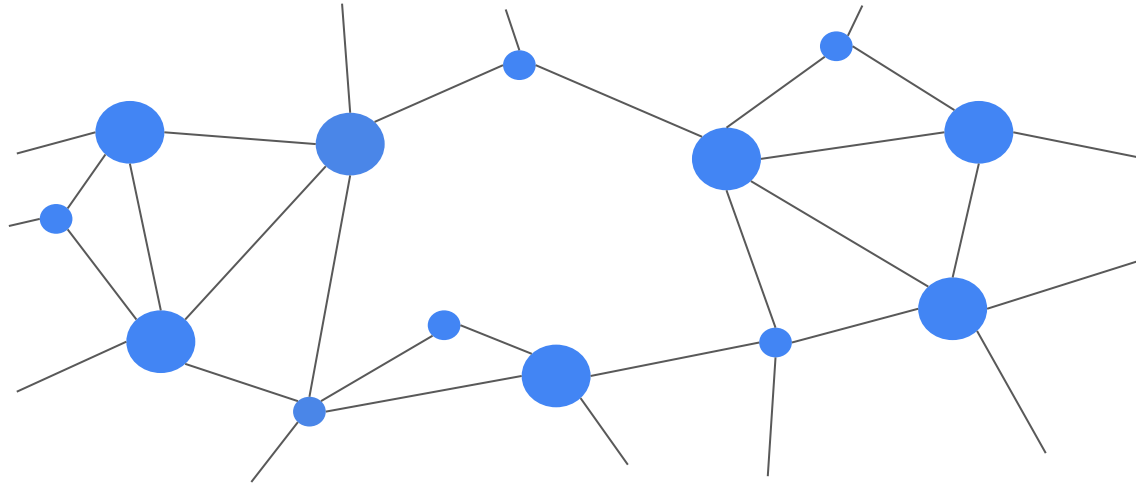


CLARIAH-CM: Workflows para la publicación de colecciones como datos

Gustavo Candela
31 de mayo de 2024



Índice

Gustavo Candela

- Desarrollador en la **BVMC** 2010-2023
- Profesor Ayudante Doctor
- [@gus_candela](https://twitter.com/gus_candela)
- [Publicaciones](#)
- gcandela@ua.es



International
**GLAM Labs
Community**

code{4}lib
JOURNAL



CLARIAH-ES



**BIBLIOTECA VIRTUAL
MIGUEL DE CERVANTES**

<https://www.cervantesvirtual.com/centro/>



Universitat d'Alacant
Universidad de Alicante

<https://www.ua.es/>

Introducción

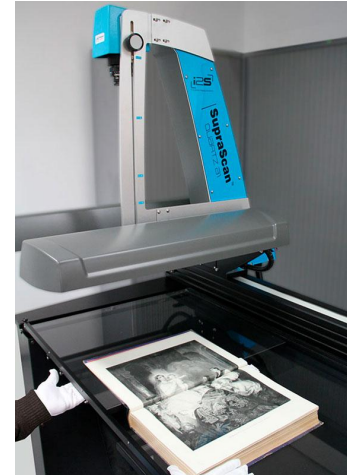
La Biblioteca Virtual Miguel de Cervantes ha sido transformada recientemente en un centro de Humanidades Digitales.



¿Sabías que...?



<https://www.cervantesvirtual.com/>



Introducción

La BVMC forma parte de **CLARIAH-ES** y ha participado activamente en la red **INTELE**.



Introducción

El proyecto **BVMC Labs** (data.cervantesvirtual.com) tiene como objetivo la publicación y reutilización de colecciones digitales de forma creativa

- Herramientas y prototipos
- Publicaciones
- Tutoriales
- Congresos
- Ejemplos de uso



<http://hdl.handle.net/10045/110281>



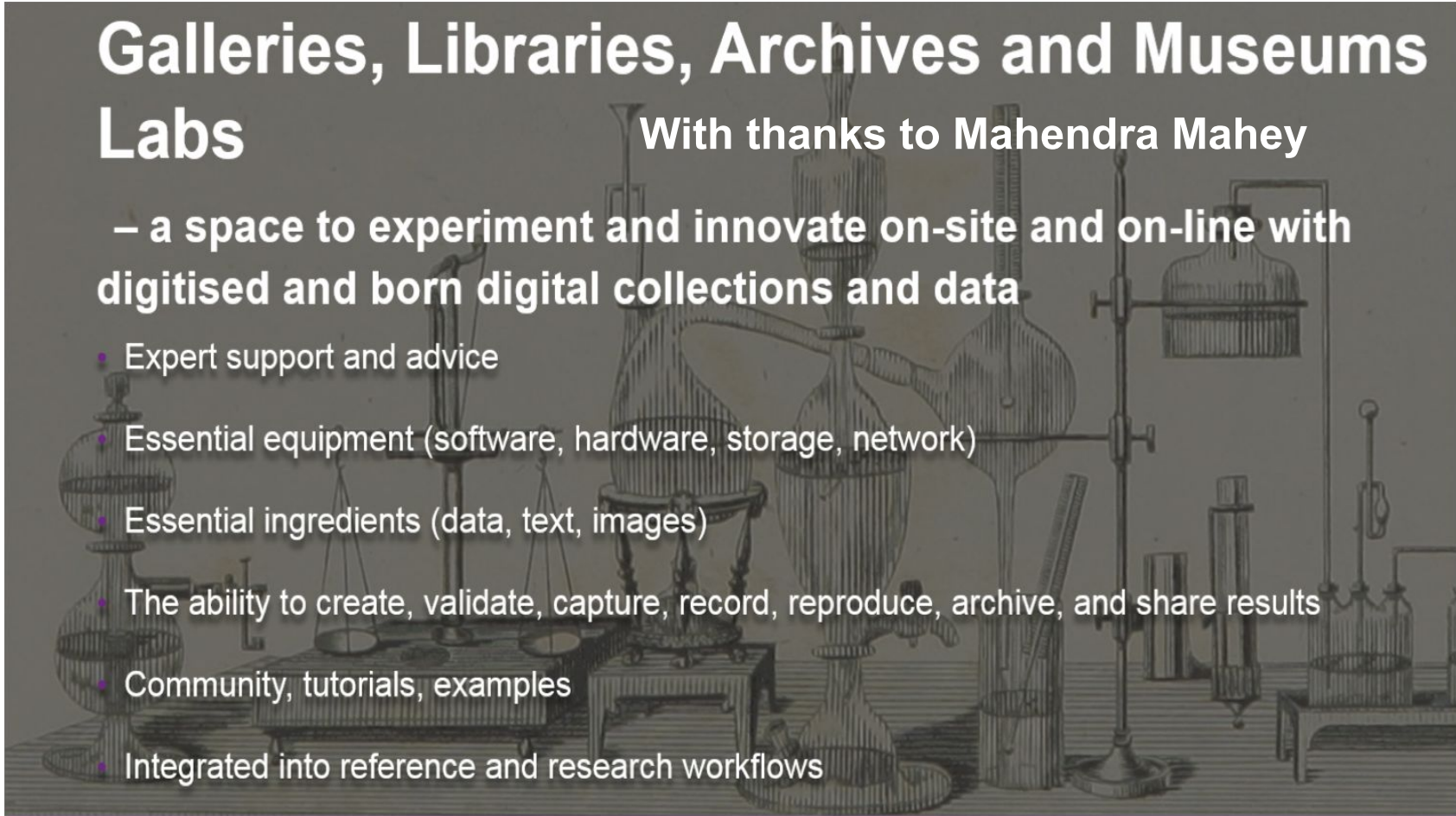
<https://collectionsasdata.github.io/>

Galleries, Libraries, Archives and Museums Labs

With thanks to Mahendra Mahey

– a space to experiment and innovate on-site and on-line with digitised and born digital collections and data

- Expert support and advice
- Essential equipment (software, hardware, storage, network)
- Essential ingredients (data, text, images)
- The ability to create, validate, capture, record, reproduce, archive, and share results
- Community, tutorials, examples
- Integrated into reference and research workflows



Introducción

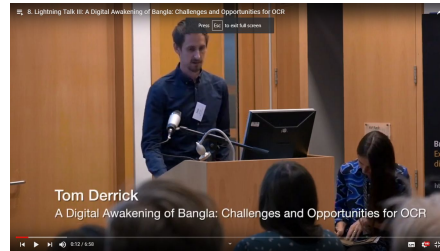
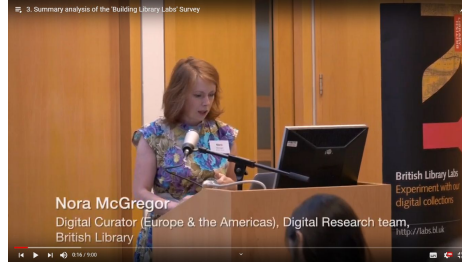
International GLAM Labs Community (<https://glamlabs.io/>)

- Library of Congress
- British Library
- Royal Danish Library
- National Library of the Netherlands
- National Library of Scotland
- Royal Library of Belgium
- Bibliotheca Alexandrina
- ...



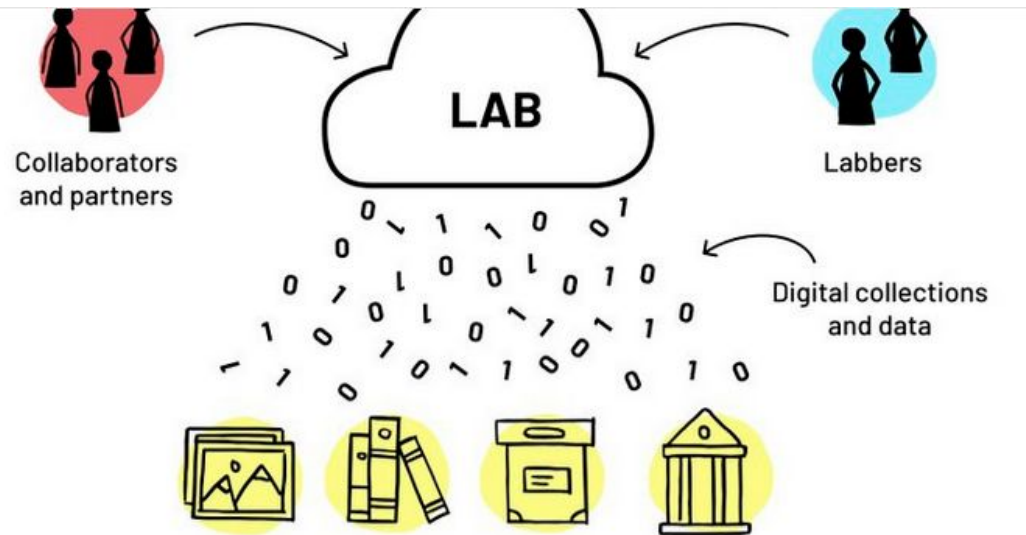
Jisc email subscription

It's all about people!



YouTube: <https://bit.ly/2YMxXYj>

International GLAM Labs Community



<https://glamlabs.io/books/>

<https://blogs.bl.uk/digital-scholarship/2019/07/invitation-to-join-digital-cultural-heritage-innovation-labs-book-sprint-doha-qatar-23-27-september.html>

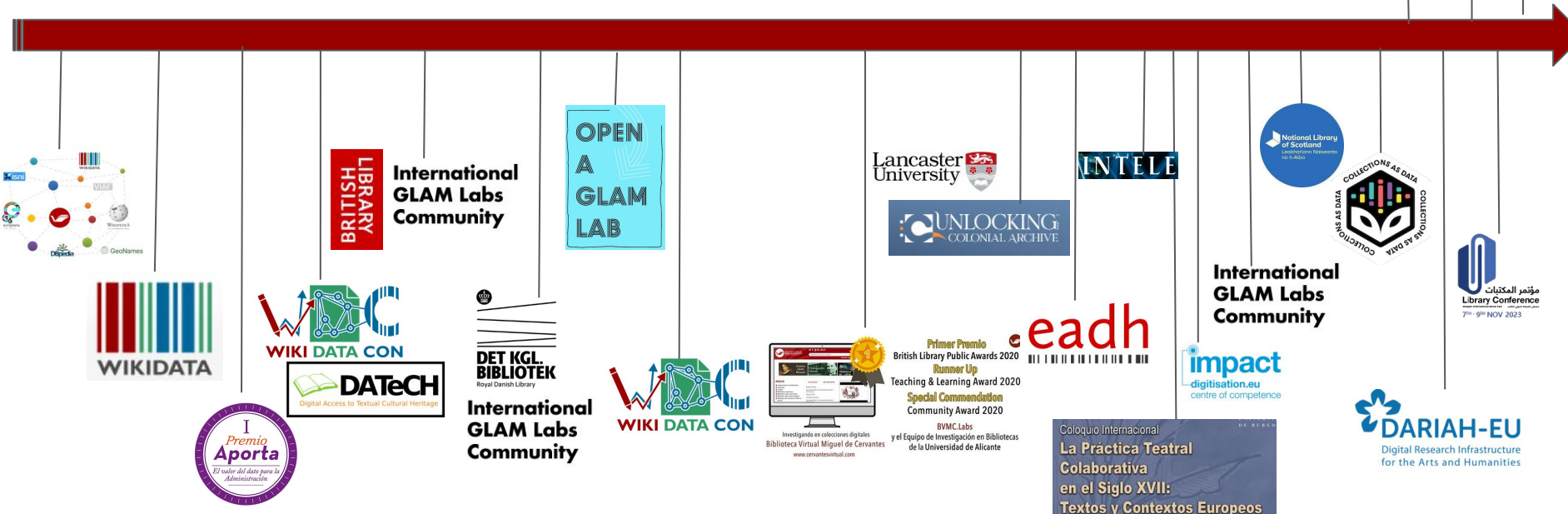
Introducción

Pasos BVMC



2015 2016 2017 2018 2019 2020 2021 2022 2023

Nueva
plataforma



International
GLAM Labs
Community



International
GLAM Labs
Community



Investigando en colecciones digitales
Biblioteca Virtual Miguel de Cervantes
www.cervantesvirtual.com



Primer Premio
British Library Public Awards 2020
Runner Up
Teaching & Learning Award 2020
Special Commendation
Community Award 2020

BVMC Labs
y el Equipo de Investigación en Bibliotecas
de la Universidad de Alicante



Coloquio Internacional
La Práctica Teatral
Colaborativa
en el Siglo XVII:
Textos y Contextos Europeos



International
GLAM Labs
Community



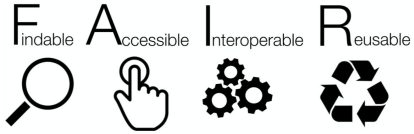
Collections as Data



“Developing cultural heritage collections that support computationally-driven research and teaching.”

(Always Already Computational: Collections as Data: <https://collectionsasdata.github.io>)

Collections as Data



Focused on characteristics of data



Complement FAIR considering both people and purpose

Findable

Rich metadata and persistent identifier

Accessible

Data and metadata understandable by humans and computers

Interoperable

Metadata described using controlled vocabularies

Reusable

License & provenance information

Collective benefit

Inclusive and equitable outcomes

Authority to control

Recognizing rights and interests

Responsibility

For indigenous languages and worldviews

Ethics

Minimizing harm, and for justice and future use

Berners-Lee:

5 estrellas

[“5-star Open Data”](#)

Relacionado con
FAIR, y abogando
por el paradigma
de los Datos
Enlazados Abiertos
(Linked Open Data)

5 ★ DATOS ABIERTOS

Tim Berners-Lee, el inventor de la Web e iniciador de los Datos Enlazados (Linked Data), sugirió un **esquema de desarrollo de 5 estrellas** para Datos Abiertos. A continuación te mostramos ejemplos para cada escalón o nivel de estrellas y te explicamos los costos y beneficios involucrados en cada caso.

El diagrama ilustra el camino hacia los Datos Abiertos de 5 estrellas. Comienza con un archivo PDF (nivel 1) y avanza a través de XLS (nivel 2), CSV (nivel 3), RDF (nivel 4) hasta LOD (nivel 5). Cada nivel está etiquetado con un número de estrellas y los requisitos OL, RE, OF, URI y LD. En la parte inferior del diagrama se muestran los logos de Creative Commons (CC), W3C, RDF, Open Data y MIT.



COLLECTIONS AS DATA: CHARTING INTERNATIONAL FUTURES

Thomas Padilla, Hannah Scates Kettler,
Yasmeen Shorish, Stewart Varner

WHAT IS "COLLECTIONS AS DATA?"

A Mellon Foundation funded grant team that seeks to foster responsible implementation and use of Collections as Data. The project funded 12 organizations to develop and document efforts to sustainably and ethically engage in the field. In April 2023, the grant team convened an international summit to expand the conversation to an intercontinental stakeholder group and share key findings.

WHY INTERNATIONAL?

In April 2023, the team convened a summit that had representation from 6 continents, 18 countries, and 62 organizations for two days in Vancouver, Canada. Summit outcomes included an update to the Santa Barbara Statement on Collections as Data, evaluation of transferable models of implementing projects at different institutional types, focused perspective from various parts of the world, and identification of challenges and opportunities for the field.

NEXT STEPS

The development of collections, staff, services, and partnerships that support multidisciplinary, multiprofessional, creative, and ethical computational engagement with library collections--going beyond the "datafication" of collections.


WHO ARE WE?

Given the variety of activities taking place through national initiatives, conferences, national library implementation, and collaborative 'collections as data' professional development offerings, it is clear that working collaboratively across contexts will result in more efficient processes and greater shared understanding.

COLLECTIONS AS DATA: STATE OF THE FIELD AND FUTURE DIRECTIONS SUMMIT

Focus on areas, such as artificial intelligence, that require further attention as we consider how to approach principles, infrastructure, and operations in a collaborative and ethically grounded way. Desire to engage more intentionally with communities that are under-represented geographically, as we work to codify as an internationalized field.



 Acknowledgments: the Mellon Foundation, Part to Whole cohort participants, and summit participants

WHAT IS "COLLECTIONS AS DATA?"

A Mellon Foundation funded grant team that seeks to foster responsible implementation and use of Collections as Data. The project funded 12 organizations to develop and document efforts to sustainably and ethically engage in the field. In April 2023, the grant team convened an international summit to expand the conversation to an intercontinental stakeholder group and share key findings.

Deliverables

1. Final Report - <https://doi.org/10.5281/zenodo.3152935>
2. Santa Barbara Statement on Collections as Data - <https://doi.org/10.5281/zenodo.3066209>
3. Facets - <https://doi.org/10.5281/zenodo.3066240>
4. Personas - <https://doi.org/10.5281/zenodo.3066515>
5. 50 Things - <https://doi.org/10.5281/zenodo.3066237>
6. Position Statements - <https://doi.org/10.5281/zenodo.3066161>
7. Methods - <https://doi.org/10.5281/zenodo.3146756>

<https://repository.ifla.org/handle/123456789/3085>

<https://osf.io/mx6uk/wiki/home/>

Collections as Data

Position Statements -> Collections as Data: State of the field and future directions

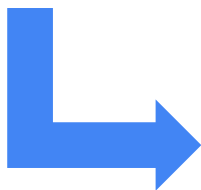
<https://zenodo.org/doi/10.5281/zenodo.7888575>

*We ask that you write a brief position statement (1-2 pages)
derived from direct or related experience salient to the scope of
work described in Collections as Data...*

<https://collectionsasdata.github.io/part2whole/recap/>



Summit Participants, Internet Archive Canada, 4/26/2023

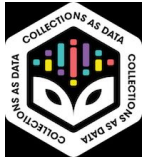


[Vancouver Statement on Collections-as-Data](#)
(translations to Spanish, Arabic, French, etc.)

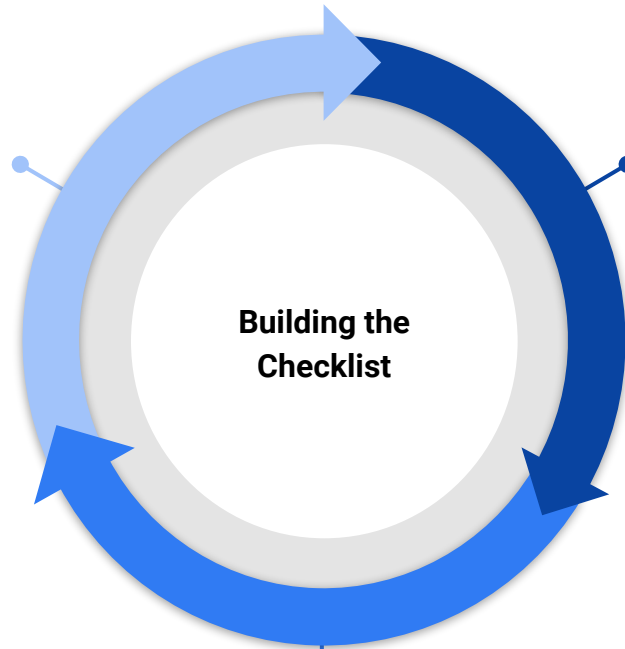
International GLAM Labs Community

Literature review

Best practices, research
articles, reports, etc.



...



Building the
Checklist

Community feedback

Questionnaire & webinar



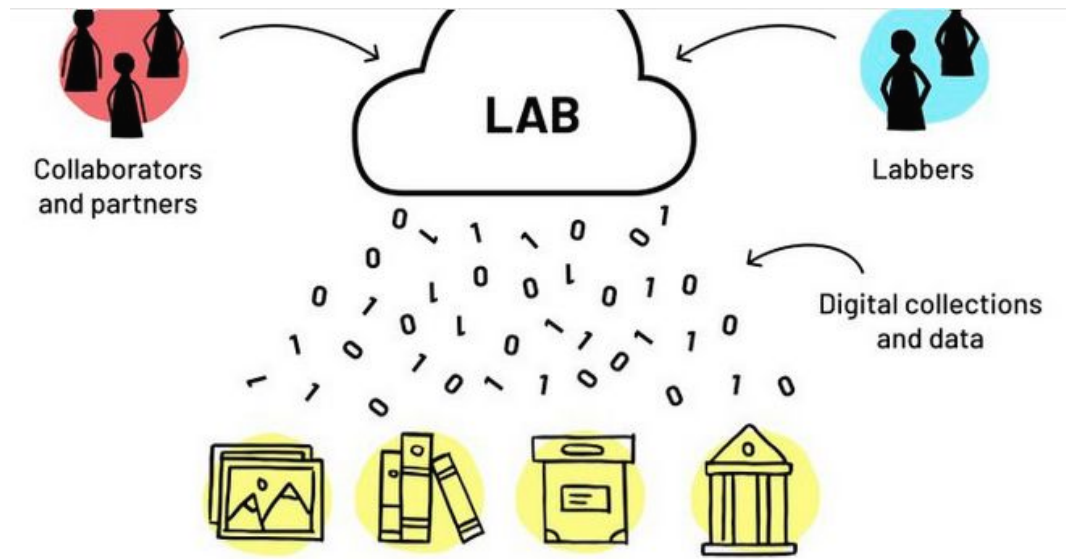
With many thanks to
Nele Gabriels

Checklist

Items to assess



International GLAM Labs Community



**Towards implementing Collections as Data in GLAM
institutions, Tuesday, October 25th 2022, 15:30-16:45 CET.**

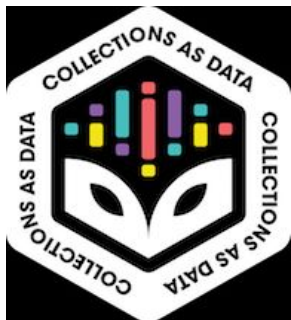
 @GLAM_labs

International GLAM Labs
Community Meetups

Collections as Data

Publicación

A checklist to make available digital collections suitable for computational use



**International
GLAM Labs
Community**

<https://doi.org/10.48550/arXiv.2304.02603>

[Global Knowledge, Memory and Communication](#)

✓ A Checklist to Publish Collections as Data in GLAM Institutions		
<input type="checkbox"/>	01	Provide a clear license allowing reuse of the dataset without restrictions <ul style="list-style-type: none">• CC0, CC BY, Public Domain Mark• National initiatives• No known copyright
<input type="checkbox"/>	02	Provide a suggestion of how to cite the dataset <ul style="list-style-type: none">• BibTeX, APA• DOI• Versions
<input type="checkbox"/>	03	Include documentation about the dataset <ul style="list-style-type: none">• README file• Tutorials and websites• Programming Historian & research articles
<input type="checkbox"/>	04	Use a public platform to publish the dataset <ul style="list-style-type: none">• GitHub, Zenodo, DataCite• Hosting
<input type="checkbox"/>	05	Share examples of use as additional documentation <ul style="list-style-type: none">• Prototypes & tools• Jupyter Notebooks (reproducible)• GLAM Labs
<input type="checkbox"/>	06	Give structure to the dataset <ul style="list-style-type: none">• Folder structure• Using self-describing folder and file names• BagIt File Packaging Format & Data Package
<input type="checkbox"/>	07	Provide machine-readable metadata (about the dataset itself) <ul style="list-style-type: none">• Dublin Core• Vocabulary of Interlinked Datasets (VoID)• Data Catalog Vocabulary (DCAT)
<input type="checkbox"/>	08	Include your dataset in collaborative edition platforms <ul style="list-style-type: none">• Increase visibility• Title, author, location, license, main subject, etc.
<input type="checkbox"/>	09	Offer an API access to your repository <ul style="list-style-type: none">• OAI-PMH, JSON, XML• IIIF• SPARQL
<input type="checkbox"/>	10	Develop a portal page <ul style="list-style-type: none">• GitHub Pages• New section in the Lab• e.g. <i>Chronicle America & Data Foundry</i>
<input type="checkbox"/>	11	Add a terms of use <ul style="list-style-type: none">• e.g. section detailing copyright, liability and access statements

[Inicio](#) / [Archivos](#) / [Vol. 18 \(2024\)](#) / [Bibliotecas y servicios de información y documentación](#)

Collections as data: Acceso computacional a colecciones digitales

Gustavo Candela

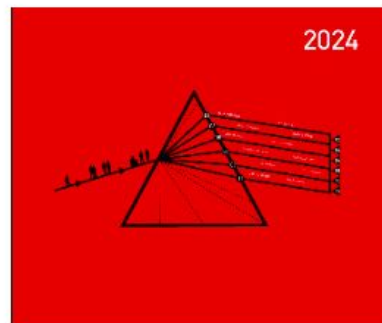
Universidad de Alicante

 <https://orcid.org/0000-0001-6122-0777>

DOI: <https://doi.org/10.3145/thinkepi.2024.e18a06>

Palabras clave: Collections as Data, Acceso computacional, Bibliotecas, GLAM

Resumen



Anuario
ThinkEPI 2024

Análisis de tendencias en información
y comunicación

Idioma

[English](#)

[Español \(España\)](#)

Información

[Para lectores/as](#)

[Para bibliotecarios/as](#)

¿Por qué un workflow?

- Más práctico y sistemático
- Reduce barreras
- Adoptar buenas prácticas
- Desarrollo iterativo
- Más fácil de implementar



Automatización de la
publicación de
Collections as Data

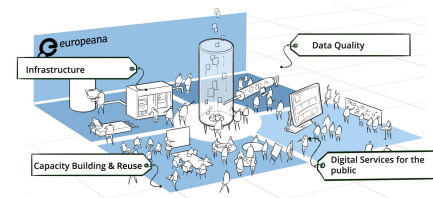
Workflow - Proceso de creación

A Checklist to Publish Collections as Data in GLAM Institutions

<input type="checkbox"/> 01	Provide a clear format already based on the chosen cultural institutions	<ul style="list-style-type: none"> CC0, CC BY, Public Domain Mark, Attribution-NonCommercial, No Rights Reserved
<input type="checkbox"/> 02	Provide a description of items on the dataset	<ul style="list-style-type: none"> Basic info DOI Access
<input type="checkbox"/> 03	Include documentation about the dataset	<ul style="list-style-type: none"> README file Metadata and vocabularies Provenance/History & related activity
<input type="checkbox"/> 04	Use a public platform to publish the dataset	<ul style="list-style-type: none"> GitHub, Zenodo, DataCite, etc.
<input type="checkbox"/> 05	Share examples of use as additional documentation	<ul style="list-style-type: none"> Workshop & tools Public webinars (reproducibility, etc.)
<input type="checkbox"/> 06	Give structure to the dataset	<ul style="list-style-type: none"> Folder structure Long and short description files and file names Input file (Package format & Data Package)
<input type="checkbox"/> 07	Provide machine-readable metadata about the dataset (such)	<ul style="list-style-type: none"> Quality Code Availability of machine-readable datasets (DSDS) Data Catalogue (DataCite)
<input type="checkbox"/> 08	Include user consent in metadata and related guidelines	<ul style="list-style-type: none"> Increase visibility File names, structure, format, user interface, etc.
<input type="checkbox"/> 09	Offer an API access to your repository	<ul style="list-style-type: none"> API (Open API, XML, JSON) API
<input type="checkbox"/> 10	Develop a portal page	<ul style="list-style-type: none"> GitHub pages Web version on the lab EU Strategy on Research & Data Funding
<input type="checkbox"/> 11	Add a version of use	<ul style="list-style-type: none"> It's better readability, copyright, safety and access conditions

<https://beta.europeana.eu/collect>

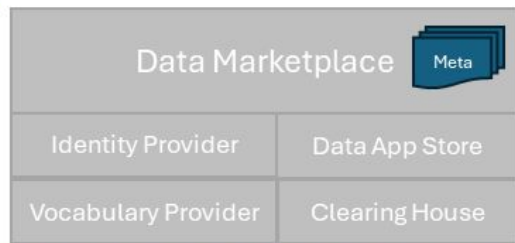
International
GLAM Labs
Community



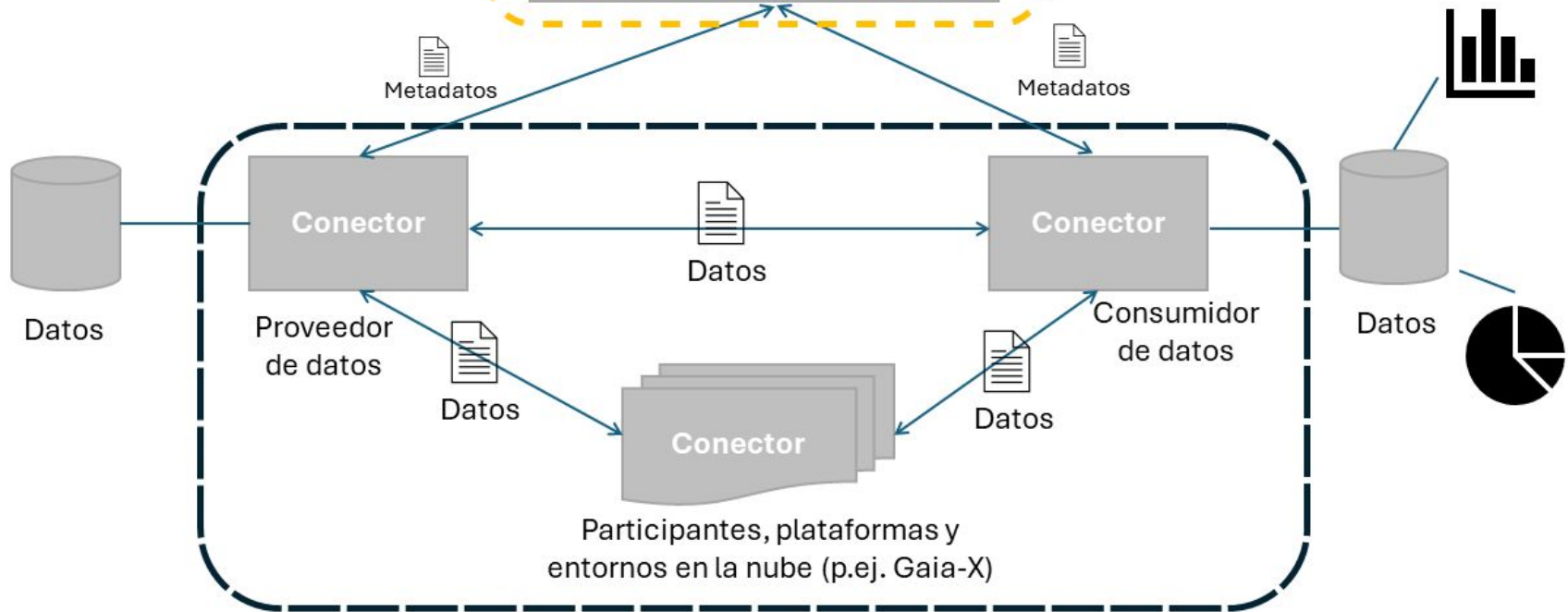
European data space
for cultural heritage

Espacios de datos

Servicios de CH Data Space



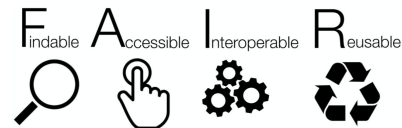
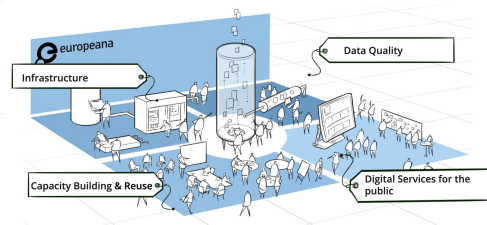
<https://pro.europeana.eu/page/com-mon-european-data-space-for-cultural-heritage>



A Collections as Data Workflow on the SSH Open Marketplace



The screenshot shows the SSH Open Marketplace interface. At the top, there is a navigation bar with the logo 'SSH Open Marketplace' (Social Sciences and Humanities Open Marketplace) and a search bar. The main content area features a breadcrumb trail: Home / Workflows / Publishing Collections as data in a Cultural Heritage data space. The title of the workflow is 'Publishing Collections as data in a Cultural Heritage data space', with a 'Copy to clipboard' link. Below the title, a note states: 'This is a draft version that is currently under review'. The main text describes how Cultural Heritage institutions have been making digital collections available for public use, and how advances in technology like AI and ML have provided a new context for computational use. It mentions 'Collections as data' and 'GLAM Labs' (Galleries, Libraries, Archives and Museums). A section on 'Data spaces' explains how they provide a set of steps for publishing collections as data in a Cultural Heritage data space. The 'Details' section lists: License: Creative Commons Attribution 4.0 International; CATEGORISATION: Activity (Description, Sharing, Disseminating); Keyword: digital collections, Data, computational methods, Collections as data; Discipline: Digital accessibility, Library.



<https://marketplace.sshopencloud.eu/workflow/I3JvP6>

With many thanks to Sally Chambers & Alba Irollo

A Collections as Data Workflow - 10 pasos

- 1** Provide a clear license allowing reuse of the dataset without restrictions Expand ▾
- 2** Provide a suggested citation for the dataset so reusers are aware of how to cite it Expand ▾
- 3** Include documentation about the dataset Expand ▾
- 4** Use a public platform to make available the dataset for the public
- 5** Share examples of use to demonstrate how the dataset can be reused
- 6** Think about a structure for the dataset for a better understanding of how to reuse the content Expand ▾
- 7** Include machine-readable metadata about the content provided in the dataset Expand ▾
- 8** Use an existing collaborative-edition platform to include the information about the dataset Expand ▾
- 9** Provide the dataset by means of an existing API Expand ▾
- 10** Create a website to present and describe the dataset to encourage its reuse Expand ▾

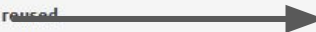


<https://marketplace.sshopencloud.eu/workflow/13JvP6>

Workflow steps (10)

- 1 Provide a clear license and terms of use allowing reuse of the dataset without restrictions
- 2 Provide a suggested citation for the dataset so reusers are aware of how to cite it
- 3 Include documentation about the dataset
- 4 Use a public platform to make available the dataset for the public
- 5 Share examples of use to demonstrate how the dataset can be reused
- 6 Think about a structure for the dataset for a better understanding of how to reuse the content
- 7 Include machine-readable metadata about the content provided in the dataset
- 8 Use an existing collaborative-edition platform to include the information about the dataset
- 9 Provide the dataset by means of an existing API
- 10 Create a website to present and describe the dataset to encourage its reuse

Detalles



¿Cómo usarlo?

Expand ▾

Expand ▾

Expand ▾

Expand ▾

5 Share examples of use to demonstrate how the dataset can be reused Collapse ▾

Why Examples of use of the contents provided by a digital collection are useful to inspire researchers. For example, reproducible code provided by Jupyter Notebooks and prototypes help to better understand how to use the content provided by a dataset. Additional approaches include the definition of research scenarios to address a particular topic.

How For example, the [Jupyter Notebooks to process the Europeana newspaper text resources](#) have been designed to help getting started with the processing of historical texts (from [Europeana Newspapers](#)) using Natural Language Processing (NLP) tools. Additional examples of reuse based on Jupyter Notebooks are provided by the [International GLAM Labs Community](#). Other approaches proposed by KBR, the Royal Library of Belgium's [DATA-KBR-BE project](#) are based on the definition of research scenarios focused on a particular topic such as mapping the publication of literature in Belgian newspapers across the first century of the Belgian nation state.

Related items (5)



Reutilizar colecciones digitales: GLAM Labs

Esta lección muestra cómo reutilizar colecciones digitales publicadas por instituciones de patrimonio cultural.

[Read more](#)



Reusing digital collections from GLAM institutions

For some decades now, Galleries, Libraries, Archives and Museums (GLAM) institutions have published and provided access to information resources in digital format. Recently, innovative approaches have appeared such as th...

[Read more](#)



Jupyter notebooks for Europeana newspaper text resource processing with CLARIN NLP tools

Binder These notebooks have been designed to help getting started with the processing historical text resources (from Europeana Newspapers) with



Introduction to Jupyter Notebooks

Jupyter notebooks provide an environment where you can freely combine human-readable narrative with computer-readable code. This lesson

<https://marketplace.sshopencloud.eu/workflow/l3JvP6>



TRAINING WORKSHOP

A workflow to publish Collections as Data

13 February 2024 | Online



Image: Exposing Online the European Cultural Heritage: The impact of Cultural Heritage on the Digital Transformation of The Society. Photo by Sebastiaan ter Burg, CC BY 4.0



¿Cuál de estos elementos es más importante para tu trabajo?

1. Incluir licencia y términos de uso
2. Incluir una sugerencia de cita
3. Documentación sobre el conjunto de datos
4. Utilizar una plataforma para publicar el conjunto de datos
5. Proveer ejemplos de reutilización

¿Cuál de estos pasos es más complejo de aplicar en tu organización?

1. Dar estructura a la colección de datos
2. Proporcionar metadatos legibles por máquina
3. Utilizar una plataforma de edición colaborativa
4. Utilizar un API para publicar la colección
5. Crear una web para la colección de datos

¿Dónde podemos buscar colecciones digitales?



**International
GLAM Labs
Community**



CARIBBEAN NEWSPAPERS



Collections as Data



<https://doi.org/10.5281/zenodo.8051036>

 Tímarit.is

<https://timarit.is/>

LIBRARY | LABS
LIBRARY OF CONGRESS

<https://id.loc.gov>

 europeana pro

<https://pro.europeana.eu/index.php/page/harvesting-and-downloads>



National Library
of Sweden

<https://libris.kb.se/sparql>

SAAM

<https://americanart.si.edu/about/lod>



<https://data.nationallibrary.fi>

**LIBRARY
HSILIRB**

<https://bnb.data.bl.uk>

DATOS·BNE·ES

<https://datos.bne.es/>

{ BnF } Data

<https://data.bnf.fr/>

KB LAB

<https://data.bibliotheken.nl>



<https://data.cervantesvirtual.com/>

B Bibliothèque nationale du Luxembourg
Open Data

<https://data.bnl.lu/>

**DEUTSCHE
NATIONAL
BIBLIOTHEK**

<https://www.dnb.de/EN/lds>



<https://labs.onb.ac.at/en/dataset/lod>

Collections as Data (nacional)

Diferente contenido y formato



DOI [10.5281/zenodo.3634442](https://doi.org/10.5281/zenodo.3634442)

Corpus of Spanish Golden-Age Sonnets



Biblioteca Virtual Miguel de Cervantes & RDA

En 2015 la BVMC transformó el catálogo a RDA para facilitar su **reutilización**

- Basado en la Web **Semántica** y **Linked Open Data**
- **Modelado de datos** siguiendo el vocabulario [RDA Registry](#)
- **Enriquecimiento** con repositorios externos como **Wikidata**
- Disponible en [SPARQL API](#)

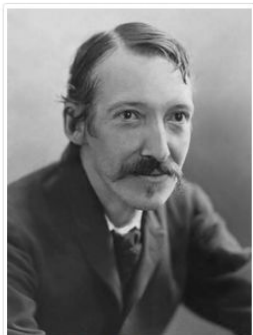


Biblioteca Virtual Miguel de Cervantes

La BVMC y Wikidata están conectadas = metadatos enriquecidos

[Inicio](#) / [Datos enlazados](#) / Stevenson, Robert Louis, 1850-1894

Stevenson, Robert Louis, 1850-1894



[Wikimedia Commons](#)

novelista, poeta y ensayista escocés

Información extraída de [Wikipedia](#) / [\(CC BY-SA 3.0\)](#)

Identificador: 79071

Nombre: Stevenson, Robert Louis

Fecha de nacimiento: [1850](#)

Fecha de fallecimiento: [1894](#)

VIAF: <https://viaf.org/viaf/95207986>

Lugar de nacimiento: [Edimburgo](#)

Lugar de fallecimiento: [0548806](#)

Nacionalidad: Reino Unido

Nombre de pila: Robert

Género: masculino

Enlaces externos:

- <https://viaf.org/viaf/95207986>
- [BVMC](#)
- [Wikidata](#)

EXPORTAR:

RDF

JSON

<https://data.cervantesvirtual.com/person/79071>

Biblioteca Virtual Miguel de Cervantes

La BVMC y Wikidata están conectadas = metadatos enriquecidos

Nuevas opciones
para explorar el
catálogo

The screenshot displays search results for 'Miguel de Cervantes' on the Cervantes Virtual website. The results are enriched with Wikidata data, which is highlighted in red boxes. Two blue arrows point from the text 'Nuevas opciones para explorar el catálogo' to these enriched sections.

Nacionalidad del autor
Datos extraídos de Wikidata

- España 2568
- México 227
- Estados Unidos 153
- Argentina 125
- Francia 101
- Chile 85
- Venezuela 80
- Italia 49
- Uruguay 48
- Perú 44

[Ver más]

Movimiento del autor
Datos extraídos de Wikidata

- Siglo de Oro 612
- Romanticismo 187
- Realismo literario 153
- Generación del 98 84
- Barroco 34
- Naturalismo 23
- Costumbrismo literario 16
- Literatura del barroco 16
- Neoclasicismo 15
- Generación del 27 12

[Ver más]

Publicación Periódica

Título: Cervantes : Bulletin of the Cervantes Society of America - [Registro bibliográfico]

Portales: Literatura | Miguel de Cervantes | Cervantes : Bulletin of the Cervantes Society of America

Mat. aut.: Cervantes Saavedra, Miguel de (1547-1616) -- Crítica e interpretación | Cervantes Saavedra, Miguel de, 1547-1616 -- Publicaciones periódicas

Fondo: 61 tomos

<https://www.cervantesvirtual.com/buscador/?q=miguel+de+cervantes>



- Main page
- Community portal
- Project chat
- Create a new item
- Recent changes
- Random item
- Query Service
- Nearby
- Help
- Donate

- Lexicographical data
- Create a new Lexeme
- Recent changes
- Random Lexeme

Tools

- What links here
- Related changes
- Special pages
- Permanent link
- Page information
- Concept URI
- Cite this page

Item **Discussion**

Read **View history**

Search Wikidata

LOD data.cervantesvirtual.com (Q111396572)

Linked Open Data from Biblioteca Virtual Miguel de Cervantes

edit

▼ In more languages

Configure

Language	Label	Description	Also known as
English	LOD data.cervantesvirtual.com	Linked Open Data from Biblioteca Virtual Miguel de Cervantes	
Spanish	LOD data.cervantesvirtual.com	Proyecto de datos abiertos de la Biblioteca Virtual Miguel de Cervantes	
Catalan	No label defined	No description defined	
Galician	No label defined	No description defined	

Statements

instance of



open data portal

edit

▼ 0 references

+ add reference

+ add value

Biblioteca Virtual Miguel de Cervantes: aplicando el workflow



<https://data.cervantesvirtual.com>



1

Provide a clear license and terms of use allowing reuse of the dataset without restrictions

¿Bajo qué licencia se han publicado los datos? ¿Y términos de uso?

CONDICIONES DE USO

El sitio web data.cervantesvirtual.com está dirigido por la Biblioteca Virtual Miguel de Cervantes y da acceso a un amplio elenco de patrimonio cultural digitalizado de toda Europa y del resto del mundo. Este material lo proporcionan un gran número de instituciones y organizaciones. La Biblioteca Virtual Miguel de Cervantes trabaja para que todos los recursos de este sitio web estén disponibles para su reutilización en otros espacios. Es por ello que, todos los metadatos (información textual sobre el patrimonio cultural digitalizado) en esta web, se publican sin ningún tipo de restricción. La mayoría de los demás materiales, como las previsualizaciones del patrimonio cultural digitalizado, están claramente etiquetados indicando si son reutilizables y en qué condiciones. Todos los metadatos disponibles en data.cervantesvirtual.com se publican sin restricciones, bajo las condiciones de Creative Commons CC0 1.0 Universal Public Domain Dedication. Sin embargo, si reutiliza los datos publicados por la Biblioteca Virtual Miguel de Cervantes, le recomendamos que siga las pautas de uso de esta para metadatos y cite la fuente de procedencia siempre que sea posible.

<https://data.cervantesvirtual.com/condiciones-de-uso/>



¿Cómo puedo citarlo?

REFERENCIAS

- Candela, G., Escobar, P., Carrasco, R. y Marco-Such, M. (2020). Evaluating the quality of linked open data in digital libraries. *Journal of Information Science*. <https://doi.org/10.1177/0165551520930951>
- Candela, G., Escobar, P., Carrasco, R. y Marco-Such, M. (28-29 de octubre de 2019). *Evaluating the quality of linked open data in digital libraries*. WikidataCon. Berlín, Alemania. https://www.wikidata.org/wiki/Wikidata:WikidataCon_2019/Program/Sessions/Libraries_panel
- Candela, G., Escobar, P., Carrasco, R. y Marco-Such, M. (2019). A linked open data framework to enhance the discoverability and impact of culture heritage. *Journal of Information Science*, 45(6), 756-766. <https://doi.org/10.1177/0165551518812658>
- Candela, G., Escobar, P., Carrasco, R. y Marco-Such, M. (2018). Migration of a Library Catalogue into RDA Linked Open Data. *Semantic Web Journal*, 9(4), 481-491. <http://dx.doi.org/10.3233/SW-170274> [Versión preprint]

<https://data.cervantesvirtual.com/datos-enlazados>



¿Cómo proporcionar documentación?

Ejemplos de consultas SPARQL

Estos ejemplos de consultas se pueden ejecutar en el [punto de acceso SPARQL](#). La BYMC dispone de un [tutorial de introducción a SPARQL](#). Para crear una aplicación que reutilice el repositorio de datos abiertos es posible utilizar la siguiente dirección <http://data.cervantesvirtual.com/bvmc-loc/repositories/data>.

OBRAS DEL AUTOR MIGUEL DE CERVANTES SAAVEDRA

Esta sentencia SPARQL devuelve las obras del autor Miguel de Cervantes Saavedra. Los autores se identifican con una URL del tipo <http://data.cervantesvirtual.com/person/identificador>, en concreto al autor Miguel de Cervantes le corresponde la URL <http://data.cervantesvirtual.com/person/40>.

```
PREFIX rdac: <http://rdaregistry.info/Elements/c/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT DISTINCT ?rol ?m ?label
WHERE {
  <http://data.cervantesvirtual.com/person/40> ?rol ?m .
  ?m a rdac:Manifestation .
  ?m rdfs:label ?label
}
```

<https://data.cervantesvirtual.com/help>

- [Obras de autor](#)
- [Idiomas](#)
- [Fechas de publicación](#)
- [Ediciones y traducciones](#)
- [Publicaciones que son continuación de otras obras](#)
- [Adaptaciones](#)
- [Roles con manifestación](#)
- [Entidad corporativa](#)

[Inicio](#) / Datos enlazados

<https://data.cervantesvirtual.com/datos-enlazados>

Datos enlazados en la Biblioteca Virtual Miguel de Cervantes

INTRODUCCIÓN

El catálogo de la [Biblioteca Virtual Miguel de Cervantes \(BYMC\)](#) contiene más de 300 000 registros que fueron creados originalmente a partir del estándar [MARC21](#). Siguiendo el [plan estratégico](#), el catálogo se actualizó con el [modelo conceptual FRBR](#) promovido por el [IFLA](#) y recibió el premio [SPIRL](#) de [Stanford Libraries](#). El contenido de la base de datos ha sido migrado automáticamente a tripletas [RDF](#) utilizando el vocabulario [RDA](#) para describir las entidades, así como sus propiedades y relaciones.

data.cervantesvirtual.com ha sido premiado con el [Premio Aporta 2017](#) por su grado de innovación, utilidad e impacto.

El proyecto tiene como objetivo ser una guía para otras instituciones que deseen publicar sus datos en abierto, así como también adoptar las últimas tendencias promovidas por instituciones referentes en el ámbito de las bibliotecas como la [IFLA](#) y la [OCLC](#). Además, los resultados de este proyecto son utilizados como material docente en la [Universidad de Alicante](#).



Semantic Web - Interoperability, Usability, Applicability *an IDS Press Journal*

About Calls Blog Issues Under Review Reviewed For Authors For Reviewers Scientometrics FAQ

Login

Username or e-mail *
gcandela@ua.es

Password *

Create new account
Request new password

Log in

Editorial Board

Migration of a library catalogue into RDA linked open data

Submitted by Gustavo Candela on 09/01/2016 - 03:49
Tracking #: 1453-2665

Authors:
Gustavo Candela
Pilar Escobar
Rafael Carrasco
Manuel Marco-Such

Responsible editor:
Christoph Schlieder



5

Share examples of use to demonstrate how the dataset can be reused

¿Cómo proporcionar ejemplos de uso?

Jupyter Notebooks

Esta sección introduce varios proyectos basados en Jupyter Notebooks y el acceso computacional a colecciones digitales realizados en el marco de la [Biblioteca Virtual Miguel de Cervantes](#) e inspirados por la [Comunidad Internacional GLAM Labs](#), el libro [Open a GLAM Lab. Collections as Data](#) y el proyecto [GLAM Workbench](#). Este trabajo ha sido premiado por la [Biblioteca Británica](#) y presentado en varias charlas y conferencias como [Research Libraries UK \(RLUK\)](#), [European Library Automation Group](#) y la [Universidad de Lancaster](#).

Todas las colecciones están publicadas en abierto, utilizan el lenguaje de programación [Python](#) y pueden ser ejecutadas en la nube sin necesidad de instalar ningún software, únicamente utilizando un navegador. Los notebooks han sido creados con la herramienta [Jupyter](#).

Con el objetivo de facilitar la reutilización y el acceso, los notebooks han sido clasificados según el nivel de dificultad [introducción](#) o [avanzado](#) teniendo en cuenta las librerías de código utilizadas y la complejidad para reproducir los resultados.



```
Setting a period 1
1 period = 2010, 2011
2 pages = s.get_page(1)
3 top = pd.DataFrame.from_dict(pages, orient='index').reset_index()
4 print(top)
5 top.columns = ['YEAR', 'AGE']
6 top = top.sort_values('AGE', ascending=False).head(10)
7 top['TOP_PAGE'] = top['TOP_PAGE'].get(0)
8 top['TOP_PAGE'] = top['TOP_PAGE'].get(0)
9 # mostrar la página
10 print(top[['TOP_PAGE', 'TOP_PAGE', 'TOP_PAGE', 'TOP_PAGE', 'TOP_PAGE', 'TOP_PAGE', 'TOP_PAGE', 'TOP_PAGE', 'TOP_PAGE', 'TOP_PAGE']])
```

GLAM JUPYTER NOTEBOOKS
Colección premiada por la Biblioteca

REUTILIZAR COLECCIONES DIGITALES: GLAM LABS

IDENTIFICAR TÉRMINOS EMERGENTES

Compartir Pantalla: Linked Open Data

Proyecto presentado en la actividad ["Compartir pantalla"](#) con la [Asociación de Bibliotecas Digitales de España](#)

[launch binder](#)

DOI: [10.5281/zenodo.10123566](#)



Exploring collections as data with Jupyter Notebooks

<https://glamlabs.io/>



Humanidades Digitales Hispánicas



BIBLIOTECA VIRTUAL MIGUEL DE CERVANTES



Universitat d'Alacant
Universidad de Alicante

<https://data.cervantesvirtual.com/notebooks>

<https://github.com/hibernator11/hdh-compartir-pantalla-2023>



7

Include machine-readable metadata about the content provided in the dataset

¿Qué vocabulario se utiliza para describir los metadatos?

RDA Registry Elements Values Data Tools About GitHub Project RDA Toolkit

RDA Registry

- RDA Registry (Home)
- Elements (RDA element sets)
 - Classes
 - Agent properties
 - Expression properties
 - Item properties
 - Manifestation properties
 - Nomen properties
 - Place properties
 - Timespan properties
 - Work properties
 - RDA Entity properties
 - Unconstrained properties
 - RDA/ONIX Framework elements
- Values (value vocabularies)

Welcome to the RDA Registry!

The RDA Registry contains **linked data and** of the entities, elements, and terminologies a Committee (RSC).

For details of the latest release see [Release 5](#)

Downloads

[v5.0.19 \(zip\)](#)

[v5.0.19 \(tar.gz\)](#)

Contacts

The RDA Registry is maintained by the RSC in Reference.

<https://www.rdaregistry.info/>



8

Use an existing collaborative-edition platform to include the information about the dataset

¿Entidad en Wikidata?



LOD data.cervantesvirtual.com (Q111396572)

Linked Open Data from Biblioteca Virtual Miguel de Cervantes

edit

[In more languages](#)

[Configure](#)

Language	Label	Description	Also known as
English	LOD data.cervantesvirtual.com	Linked Open Data from Biblioteca Virtual Miguel de Cervantes	
Spanish	LOD data.cervantesvirtual.com	Proyecto de datos abiertos de la Biblioteca Virtual Miguel de Cervantes	
Catalan	No label defined	No description defined	
Galician	No label defined	No description defined	

Statements

instance of open data portal edit

[▼ 0 references](#)

[+ add reference](#)

[+ add value](#)

<https://www.wikidata.org/wiki/Q111396572>





CENTRO BIBLIOTECA VIRTUAL
MIGUEL DE CERVANTES

Workflow steps (10)

9

Provide the dataset by means of an existing API

SPARQL Query Editor Extensions: cxml save to dsv sponge User: SPARQL

Default Data Set Name (Graph IRI)

Query Text

```
select distinct ?Concept where {{{ a ?Concept}} LIMIT 100
```

Results Format

<https://data.cervantesvirtual.com/sparql>



CENTRO BIBLIOTECA VIRTUAL
MIGUEL DE CERVANTES

Workflow steps (10)

10

Create a website to present and describe the dataset to encourage its reuse

GitHub pages?

[Inicio](#) / Datos enlazados

Datos enlazados en la Biblioteca Virtual Miguel de Cervantes

INTRODUCCIÓN

El catálogo de la [Biblioteca Virtual Miguel de Cervantes \(BVMC\)](#) contiene más de 300 000 registros que fueron creados originalmente a partir del estándar [MARC21](#). Siguiendo el [plan estratégico](#), el catálogo se actualizó con el [modelo conceptual FRBR](#) promovido por el [IFLA](#) y recibió el premio [SPIRL](#) de [Stanford Libraries](#). El contenido de la base de datos ha sido migrado automáticamente a tripletas [RDF](#) utilizando el vocabulario [RDA](#) para describir las entidades, así como sus propiedades y relaciones.

[data.cervantesvirtual.com](#) ha sido premiado con el [Premio Aporta 2017](#) por su grado de innovación, utilidad e impacto.

El proyecto tiene como objetivo ser una guía para otras instituciones que deseen publicar sus datos en abierto, así como también adoptar las últimas tendencias promovidas por instituciones referentes en el ámbito de las bibliotecas como la [IFLA](#) y la [OCLC](#). Además, los resultados de este proyecto son utilizados como material docente en la [Universidad de Alicante](#).



GitHub Pages

Websites for you and your projects.

<https://pages.github.com/>

<https://data.cervantesvirtual.com>



Resumen de la evaluación

Workflow steps (10)

- 1 Provide a clear license and terms of use allowing reuse of the dataset without restrictions [Expand](#)
- 2 Provide a suggested citation for the dataset so reusers are aware of how to cite it [Expand](#)
- 3 Include documentation about the dataset [Expand](#)
- ~~4 Use a public platform to make available the dataset for the public~~ [Expand](#)
- 5 Share examples of use to demonstrate how the dataset can be reused [Expand](#)
- ~~6 Think about a structure for the dataset for a better understanding of how to reuse the content~~ [Expand](#)
- 7 Include machine-readable metadata about the content provided in the dataset [Expand](#)
- 8 Use an existing collaborative-edition platform to include the information about the dataset [Expand](#)
- 9 Provide the dataset by means of an existing API [Expand](#)
- 10 Create a website to present and describe the dataset to encourage its reuse [Expand](#)

¿cita sugerida?

¿Zenodo (DOI), GitHub,
institutional
repositories?

No aplica

¿Dudas o preguntas?



CENTRO BIBLIOTECA VIRTUAL
MIGUEL DE CERVANTES

<https://data.cervantesvirtual.com>

Actividad

¿Cómo usar el workflow propuesto?

¿Dificultades? ¿Por dónde empezar?

Propuestas del público

Uso de Jupyter Notebooks y Wikidata

Compartir Pantalla: Linked Open Data y SPARQL

Proyecto presentado en la actividad "[Compartir pantalla](#)" con la [Asociación de Humanidades Digitales Hispánicas](#).

 launch  binder

DOI [10.5281/zenodo.10123566](https://doi.org/10.5281/zenodo.10123566)



<https://github.com/hibernator11/hdh-compartir-pantalla-2023>



UNIVERSIDAD
COMPLUTENSE
MADRID



Desafíos

- Las instituciones GLAM pueden jugar un papel relevante en IA
 - <https://www.rluk.ac.uk/digital-shift-manifesto/>
- Adopción de Collections as data por parte de las instituciones más pequeñas
 - Publicar el texto, no solo metadatos
 - Diferentes versiones de la colección digital
 - Uso de licencias abiertas
 - Ética y términos de uso ([CARE principles](#))



Desafíos



Diversidad en el uso de vocabularios controlados para definir los metadatos

Table 1. Overview of LOD repositories made available by CH institutions


Institution	URL	Vocabulary
Biblioteca Virtual Miguel de Cervantes	https://data.cervantesvirtual.com/sparql	RDA
British Library	https://bl.natbib-lod.org	BIBFRAME
Europeana	https://sparql.europeana.eu	EDM
Getty Research Institute	https://data.getty.edu/vocab/sparql	CIDOC-CRM
Library of Congress	https://id.loc.gov	BIBFRAME
National Library of Finland	https://data.nationallibrary.fi/bib/sparql	Schema.org
National Library of France	https://data.bnf.fr/sparql	FRBR
National Library of Scotland	https://data.nls.uk/tools/jupyter-notebooks/semantic-web/	BIBFRAME, schema.org
National Library of Spain	https://datos.bne.es/sparql	FRBR
National Library of the Netherlands	http://data.bibliotheken.nl/sparql	Schema.org
Rijksmuseum	https://data.rijksmuseum.nl	EDM
Smithsonian American Art Museum	http://edan.si.edu/saam/sparql	CIDOC-CRM
Zeri Photo Archive	http://data.fondazionezeri.unibo.it/	CIDOC-CRM
Yale Collections Discovery	https://lux.collections.yale.edu	CIDOC-CRM

Desafíos

Métodos para evaluar la **calidad de datos** en instituciones GLAM



```
<book> {  
  rdf:type [schema:Book] ;  
  schema:inLanguage xsd:string *; SHEx  
  rdfs:label xsd:string +;  
  schema:numberOfPages xsd:integer ?;  
  schema:sameAs IRI ?;  
  schema:isbn xsd:string *;  
  schema:author IRI *;  
  schema:description xsd:string *  
}
```



A Shape Expression approach for assessing the quality of Linked Open Data in libraries

<https://doi.org/10.3233/SW-210441>



Evaluating the quality of linked open data in digital libraries

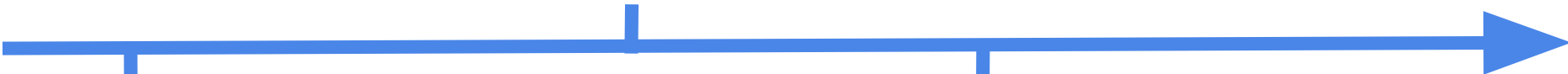
<https://doi.org/10.1177/0165551520930951>



[sheXer](https://www.sheXer.org)

An automatic data quality approach to assess semantic data from cultural heritage institutions

<https://doi.org/10.1002/asi.24761>



Desafíos

Reutilización - Jupyter Notebooks

An approach to **assess the quality** of Jupyter projects published by GLAM institutions

- A method to assess Jupyter notebook projects
- Results of the evaluation of Jupyter Notebooks projects
- Best practice and guidelines



Dimension	Criterion
Understandability	Using literate programming features
	Including additional documentation and guidelines
	Naming of the notebooks
	Storing cell output
	Audience/intended use
Provisioning of metadata	
Availability	License
Efficiency	Size
Traceability	Versioning
Portability	Providing dependencies
Recoverability	Providing citation information
	Last run date
Credibility	Trustworthiness on project level

Candela, G., Chambers, S., & Sherratt, T. (2023). An approach to assess the quality of Jupyter projects published by GLAM institutions. *Journal of the Association for Information Science and Technology*, 1–15. <https://doi.org/10.1002/asi.24835>

Desafíos

Calidad OCR



Hemeroteca Digital

Biblioteca Nacional de España



Xlilill. 4s3.
llflrit S 10 lic jllili0 lic 1 S 34.
I 1) C II flil'1.0S.
precia dc lc inicrlcion en Mcillevado cl pcriúo leo i
b lct vciattcc tnicritvrcl. Pcr nn mcc co rc.
LA
A..
Prcio dc lc inicricíon en llii pro-, vinclet franco d
Par nn mcc sg rs
ESTE OEBIODIt O SALE TODOS LOS DIAS KSCEPTO LOS LUÍXX
ctl Rea\onza Xagür1 un s Á cctc tic D p p crri llisin
ADI'El\TErCIA.
Hnbieiidose tiignedo S. llf. Ió Iteittu Gobcrttedoru
ste cu eonoeímietto tle lss señtores suscritores g d
tssntos dísnepusurmi su beuenoíetteío ul períotheo sun



Impact White Paper

Desafíos

Inteligencia Artificial



AI companies do not ask permission to import these data and are not transparent about how these data are used. *KB management team member Martijn Kleppe*

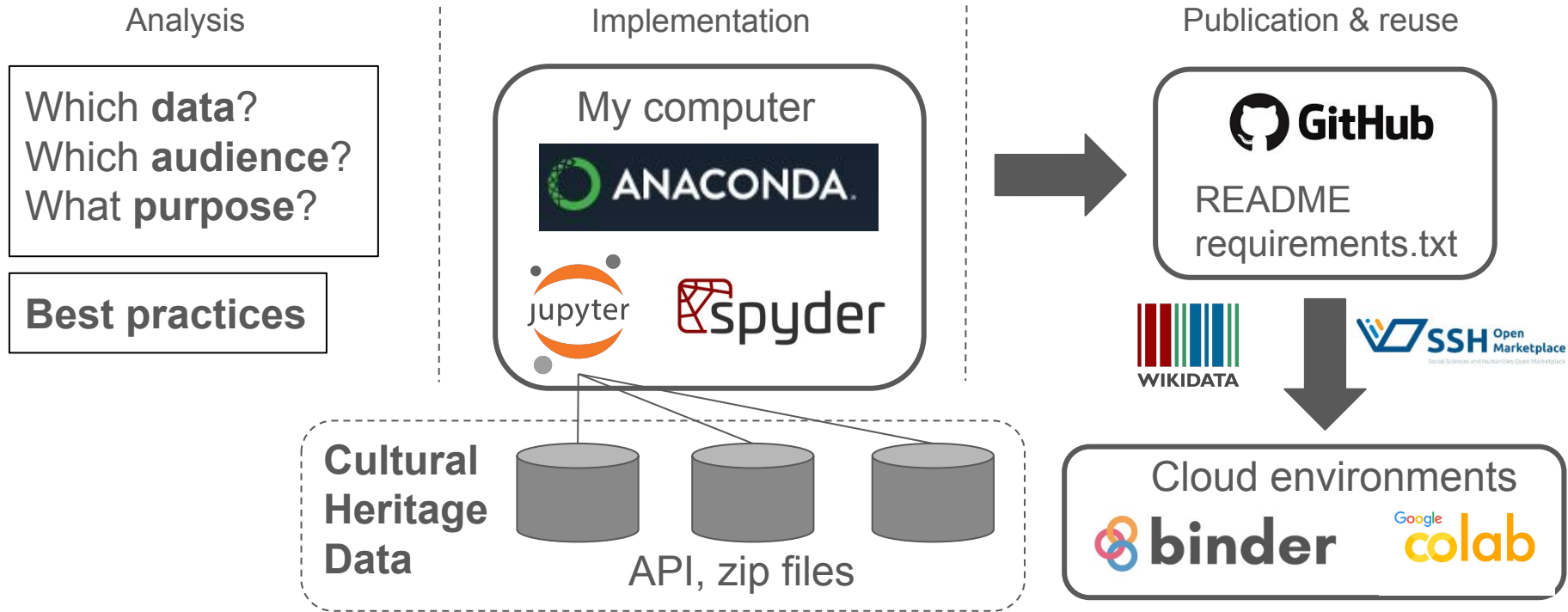
KB restricts access to collections for training commercial AI

The KB does not want commercial companies to use digital resources without permission for training AI. That goes against the AI principles established by the KB. The KB has since taken measures to restrict this use and will be releasing a statement on this today.

<https://www.kb.nl/en/nieuws/kb-restricts-access-to-collections-for-training-commercial-ai>

Trabajo futuro

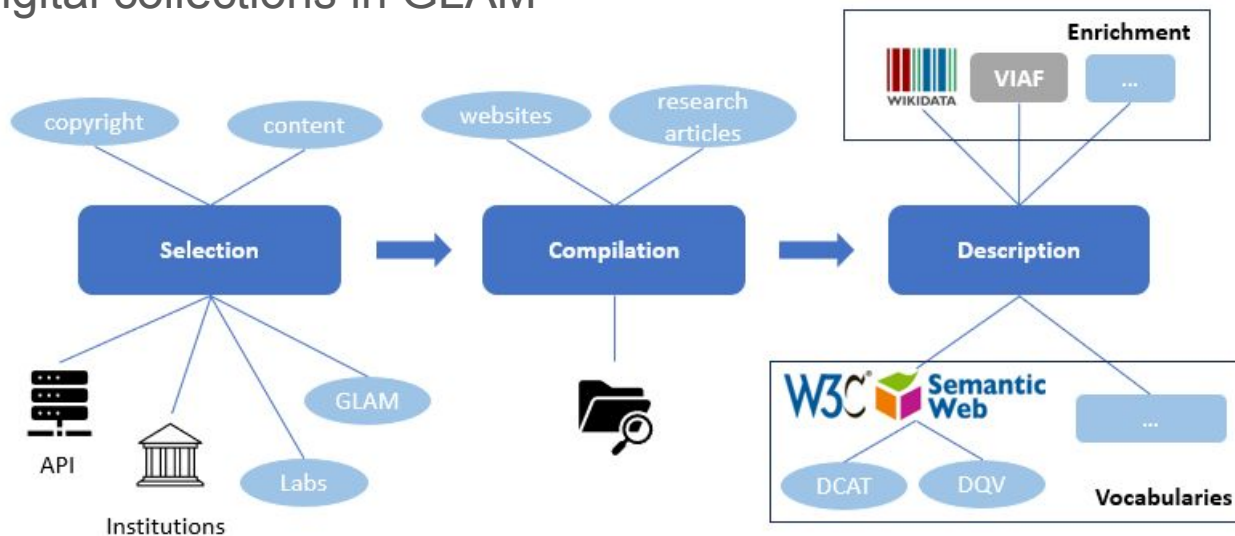
Creating a Jupyter Notebook collection from scratch



Trabajo futuro

Uso de vocabularios controlados para describir catálogos

A **Linked Open Data** framework based on **Data Catalog Vocabulary** to describe digital collections in GLAM



Categories of metadata

- Distribution
- Quality
- Provenance
- Examples of use
- Composition
- Terms of use
- Transparency
- Collecting process
- Preprocessing

<https://github.com/hibernator11/dcat-glam-catalog>

DARIAH Annual Event 2024

June 18, 2024 - June 21, 2024

**Workflows: Digital Methods for Reproducible
Research Practices in the Arts and Humanities**



DARIAH Annual Event 2024
June 18-21, Lisbon, Portugal



<https://annualevent.dariah.eu/>



Looking for learning resources?

DARIAH-CAMPUS is a discovery framework and hosting platform for learning resources.

Browse content

Choose a category to browse or search above



Resources

Learn about different topics
with online resources
provided by DARIAH



Events

Missed a face-to-face
DARIAH event? Check out
what happened



Pathfinders

Collections of external
resources curated by the
DARIAH team



OSCARS

Open Science Clusters' Action
for Research & Society



ENVRI
Community



ESCAPE

LIFE
SCIENCE
RI



panOSC
Open Science Cluster



SSHOpenCluster
Social Sciences & Humanities Resources

Call for Open Science Projects & Services

**NOW
OPEN**

Ciencia abierta, datasets, EOSC

<https://oscars-project.eu/oscars-first-open-call>

Estancias en instituciones europeas

<https://atrium-research.eu/>



ATRIUM

Facilitating access to digital research
infrastructures and advancing frontier knowledge
in the arts and humanities — across disciplines,
languages and media.



Referencias

- [Asociación Humanidades Digitales Hispánicas y data.cervantesvirtual.com](#)
- [Collections as Data: State of the Field and Future Directions](#)
- [Impact Centre of competence: Sharing and Sustaining Digitisation Knowledge](#)
- [International GLAM Labs Community](#)
- Lancaster University: [Unlocking the Colonial Archive](#)
 - <https://github.com/hibernator11/UCA-relacionesgeograficas>
- [National Library of Scotland](#)
 - <https://data.nls.uk/projects/the-national-librarians-research-fellowship-in-digital-scholarship-2022-23/>
 - <https://data.nls.uk/tools/jupyter-notebooks/semantic-web/>
- [OCLC Research](#)
- Tim Sherratt. [GLAM Workbench](#)
- <https://www.kbr.be/en/projects/data-kbr-be>
- <https://web.ua.es/es/actualidad-universitaria/2023/junio2023/12-16/la-ua-acoge-a-los-impulsores-de-la-red-estrategica-clariah-es.html>

Referencias

- Mahey, M. et al. Open a GLAM Lab. International GLAM Labs Community, Book Sprint, 2019. <https://doi.org/10.21428/16ac48ec.f54af6ae>
- Candela, G., Chambers, S., & Sherratt, T. (2023). An approach to assess the quality of Jupyter projects published by GLAM institutions. Journal of the Association for Information Science and Technology, 1–15. <https://doi.org/10.1002/asi.24835>
- Candela, G. An automatic data quality approach to assess semantic data from cultural heritage institutions. J. Assoc. Inf. Sci. Technol. 74(7): 866-878 (2023). <https://doi.org/10.1002/asi.24761>
- Candela, G., Escobar, P., Carrasco, R. and Marco-Such, M. Evaluating the quality of linked open data in digital libraries. J. Inf. Sci. 48(1): 21-43 (2022). <https://doi.org/10.1177/0165551520930951>
- Candela, G. Towards a semantic approach in GLAM Labs: The case of the Data Foundry at the National Library of Scotland. Journal of Information Science. (2023) Online first. <https://doi.org/10.1177/01655515231174386>
- Candela, G., Pereda, J., Sáez, D., Escobar, P., Sánchez, A., Villa-Torres, A., Palacios, A., McDonough, K. and Murrieta-Flores, P. 2023. An ontological approach for unlocking the Colonial Archive. J. Comput. Cult. Herit. Just Accepted (April 2023). <https://doi.org/10.1145/3594727>
- Candela, G., Gabriëls, N., Chambers, S., Dobрева, M., Ames, S., Ferriter, M., Fitzgerald, N., Harbo, V., Hofmann, K., Holownia, O., Irollo, A., Mahey, M., Manchester, E., Pham, T.-A., Potter, A. and Van Keer, E. (2023), "A checklist to publish collections as data in GLAM institutions", Global Knowledge, Memory and Communication, Vol. ahead-of-print No. ahead-of-print. <https://doi.org/10.1108/GKMC-06-2023-0195>

CLARIAH-CM: Workflows para la publicación de colecciones como datos

¡Muchas gracias!

