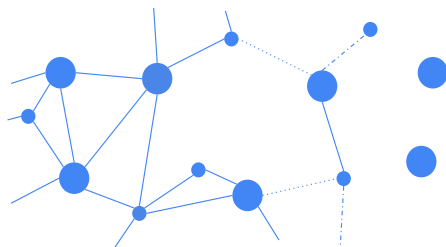


Wikidata and GLAM: Enriching Digital Collections

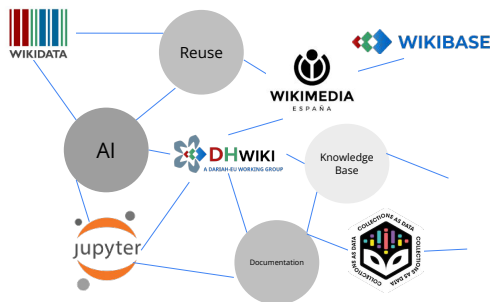
Gustavo Candela

Mini-symposium "Open Innovation for Digital Heritage"
7th July 2026



Content

- About me
- Introduction
- Examples & use cases
- Future work



About me

Gustavo Candela

- PhD
- Developer at Biblioteca Virtual Miguel de Cervantes 2010-2023
- Lecturer in CS
 - Joint Research committee DARIAH
 - Technical Assessment Committee Echoes
 - Wikimedia Spain
 - DHwiki DARIAH WG
- gcandela@ua.es



Introduction

1. GLAM Datasets
2. Collections as data

Dataset examples



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



<https://bnelab.bne.es>



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



<https://id.loc.gov>



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



Bibliothèque nationale du Luxembourg
Open Data

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



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



<https://data.bibliotheek.nl>



<https://api.moma.org/>



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



<https://labs.kb.dk/>



<https://data.nationallibrary.fi>



British Library Labs
projects



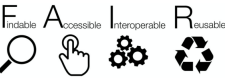
<https://labs.onb.ac.at>



Getty Provenance Index

Collections as data

With many thanks to
Thomas Padilla



Focused on data



Complement FAIR

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

<https://www.gida-global.org/care>

Findable Rich metadata and permanent identifiers	Accessible Machine-readable metadata
Interoperable Metadata described with vocabularies	Reusable Licence and documentation
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

Publishing Collections as data

Publication

A checklist to make available digital collections suitable for computational use



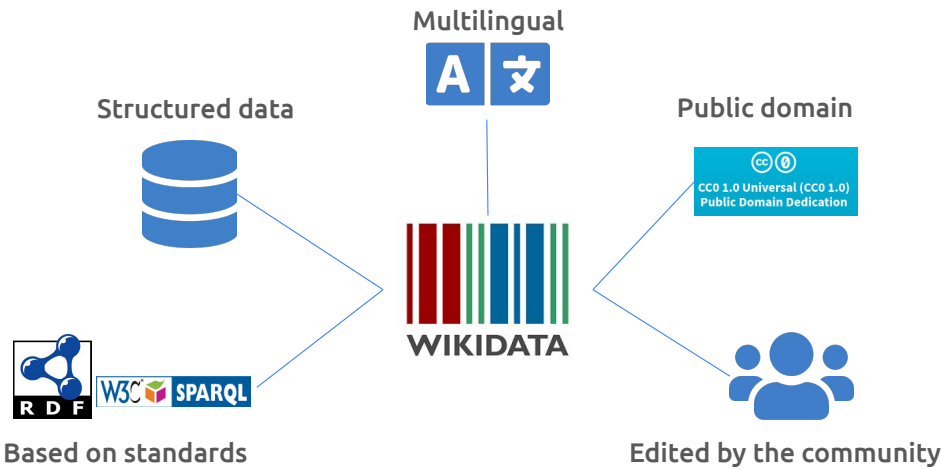
<https://doi.org/10.48550/arXiv.2304.02603>

[Global Knowledge. Memory and Communication](#)

☒ A Checklist to Publish Collections as Data in GLAM Institutions

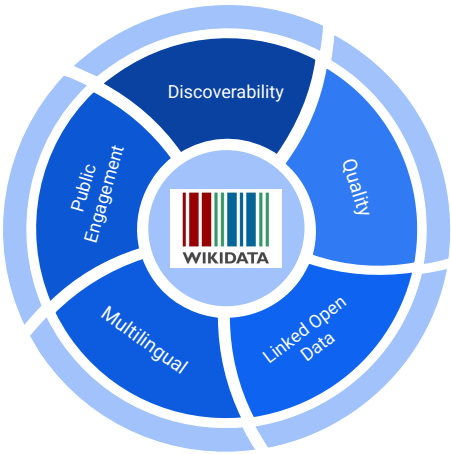
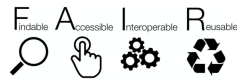
☐ 01	Provide a clear license allowing reuse of the dataset without restrictions	• CC0, CC-BY, Public Domain Mark • National initiatives • No known copyright
☐ 02	Provide a suggestion of how to cite the dataset	• BitText, APA • DOI • Versions
☐ 03	Include documentation about the dataset	• README file • Tutorials and webinars • Programming Historian & research articles
☐ 04	Use a public platform to publish the dataset	• GitHub, Zenodo, DataCite • Hosting
☐ 05	Share examples of use as additional documentation	• Prototypes & tools • Jupyter Notebooks (reproducible) • GLAM Labs
☐ 06	Give structure to the dataset	• Folder structure • Using self-describing folder and file names • BagIt File Packaging Format & Data Package
☐ 07	Provide machine-readable metadata (about the dataset itself)	• Dublin Core • Vocabulary of Interlinked Datasets (VoID) • Data Catalog Vocabulary (DCAT)
☐ 08	Include your dataset in collaborative edition platforms	• Increase visibility • Title, author, location, license, main subject, etc.
☐ 09	Offer an API access to your repository	• OAI-PMH, JSON, XML, RDF • SPARQL
☐ 10	Develop a portal page	• GitHub Pages • New section in the Lab e.g. <i>Chronising America & Data Foundry</i>
☐ 11	Add a terms of use	• e.g. section detailing copyright, liability and access statements

Wikidata



Wikidata in GLAM - Benefits

Why?

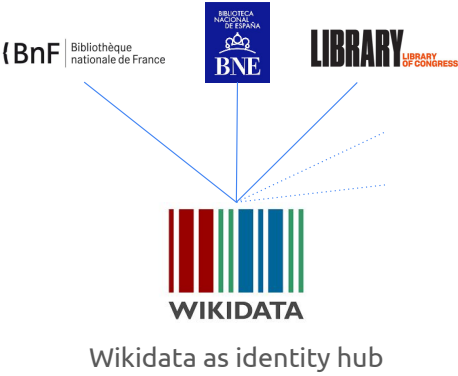


GLAM in Wikidata

How?

Properties

Property	Description
P5361	BNB person ID
P268	Bibliothèque nationale de France ID
P950	Biblioteca Nacional de España ID
P227	German National Library
P244	Library of Congress authority ID



Gustavo Candela, Mirjam Cuper, Olga Holownia, Nele Gabriëls, Milena Dobрева, Mahendra Mahey: A Systematic Review of Wikidata in GLAM Institutions: a Labs Approach. TPDL (2) 2024: 34-50. https://doi.org/10.1007/978-3-031-72440-4_4

GLAM in Wikidata

→ ↻

www.wikidata.org/wiki/Q5682

↗ In more languages:

compare

Language	Label	Description	Also known as
default for all languages	Miguel de Cervantes Saavedra		Cervantes Miguel de Cervantes
English	Miguel de Cervantes	Spanish novelist, poet, and playwright (1547-1616)	Miguel de Cervantes y Saavedra Miguel de Cervantes Cortinas Miguel de Cervantes y Corti... Miguel de Cervantes Saavedra Miguel de Cervantes-Saavedra
Spanish	Miguel de Cervantes	escritor español	Miguel de Cervantes
Bulgarian	Мигел де Сервантес	испански писател	
Russian	Мигель де Сервантес	испанский писатель	Мигель де Сервантес: Casa...

All entered languages

Statements

instance of

human

+ 1 reference

image



<https://www.wikidata.org/wiki/Q5682>

How?



→ ↻

www.wikidata.org/wiki/Q5682

BookBrainz author ID

204c5290-daad-4add-bc89-5f22d3d1761b

+ 0 references

Bridgeman artist ID

19903

+ 0 references

British Museum person or institution ID

22378

+ 0 references

BVLarramendi ID

FILA20140548965

+ 0 references

BVMC person ID

40

+ 0 references

Case studies

- 1. Examples
- 2. Challenges



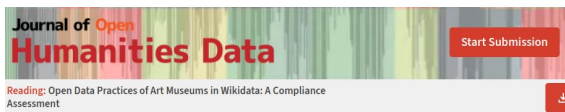
Wikidata in GLAM



[Seoane, CC BY-SA 4.0](#)

<https://wikimedia.es/abiertas-las-inscripciones-para-wikidata-en-el-contexto-qlam-ii-patrimonio-conectado-en-el-museo-del-prado/>
<https://wikimedia.es/patrimonio-conectado-wikidata-impulsa-el-ecosistema-qlam-desde-el-museo-del-prado/>

Data quality in museums



Collection: [Wikidata Across the Humanities: Datasets, Methodologies, Reuse](#)

Discussion Paper

Open Data Practices of Art Museums
in Wikidata: A Compliance Assessment

Meltem Dişli , Gustavo Candela , Silvia Gutiérrez , Giovanna Fontenelle 

<https://doi.org/10.5334/johd.438>

Data quality criteria

- Understandability
- Availability
- Federated
- Reusability
- Timeliness
- Enrichment
- File repository (Wikimedia Commons)

```
SELECT distinct ?museum ?museumLabel ?property ?propertyLabel (count(?item) as ?total)
WHERE {
  ?museum wdt:P31/wdt:P279* wd:Q207694 .
  ?museum wdt:P1687 ?property .
  ?property wdt:P31 wd:Q44847669 .
  BIND(URI(REPLACE(STR(?pr STR(wd:), STR(wdt:))) AS ?id)
  ?item ?id ?value
  SERVICE wikibase:label {
    ceParam wikibase:language "en". }
}
group by ?museum ?museumLa
having(?total>5000)
order by desc(?total)

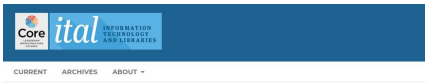
LIMIT 100 OFFSET 0
```

<https://w.wiki/FZGn>

Figure 1

SPARQL Query for retrieving Art museums with at least 5,000 records in Wikidata (<https://w.wiki/FZGn>).

Exploring the use of SPARQL federated queries



HOME / ARCHIVES / VOL. 44 NO. 4 (2025) / Articles

From Linked Open Data to Collections as Data
A Reproducible Framework Using Federated Queries

Meltem Digi
Hacettepe University, Department of Information Management
<https://doi.org/10.5860/ital.v44i4.17432>

Giulia Osti
University College Dublin
<https://orcid.org/0000-0001-8786-8885>

Gustavo Candela
University of Jülich
<https://doi.org/10.5860/ital.v44i4.17432>

Richard Zijdemann
International Institute of Social History
<https://doi.org/10.5860/ital.v44i4.17432>

DOI: <https://doi.org/10.5860/ital.v44i4.17432>

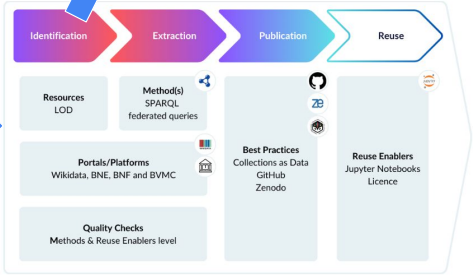
PDF

PUBLISHED
2023-12-15

HOW TO CITE

Digi, M., Osti, G., Candela, G., & Zijdemann, R. (2023). From Linked Open Data to Collections as Data: A Reproducible Framework Using Federated Queries. *Information Technology and Libraries*, 42(4). <https://doi.org/10.5860/ital.v44i4.17432>

<https://doi.org/10.5860/ital.v44i4.17432>



Avoiding data silos....

Wikidata and SOLID

Giving control to our data



Wikidata Query Service

Ejemplos Ayuda Más herramientas Constructor de consultas

```
1 PREFIX dcterms: <http://purl.org/dc/terms/>
2 PREFIX rdarerelationships: <http://rdvocab.info/RDARelationshipsWEMI/>
3 PREFIX rdagrouplelements: <http://rdvocab.info/Elements/>
4
5 SELECT ?author ?expression ?title ?edition ?placeOfPublication ?yearOfPublication WHERE {
6   wd:Q158526 wdt:P268 ?id
7   BIND(uri(concat(concat("http://data.bnf.fr/ark:/12148/cb", ?id), "#about")) as ?author)
8   SERVICE <http://data.bnf.fr/sparql> {
9     ?expression <http://id.loc.gov/vocabulary/relators/aut> ?author .
10    OPTIONAL { ?expression dcterms:language ?langCode .}
11    OPTIONAL { ?manifestation dcterms:publisher ?edition .}
12    ?manifestation rdarerelationships:expressionManifested ?expression .
13    ?manifestation dcterms:title ?title .
14    ?manifestation dcterms:date ?yearOfPublication .
15    OPTIONAL { ?manifestation rdagrouplelements:placeOfPublication ?placeOfPublication .}
16  }
17  SERVICE wikibase:label { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],mul,en" . }
18 }
19
20 Limit 100
```

Table

37 resultados en 39454 ms

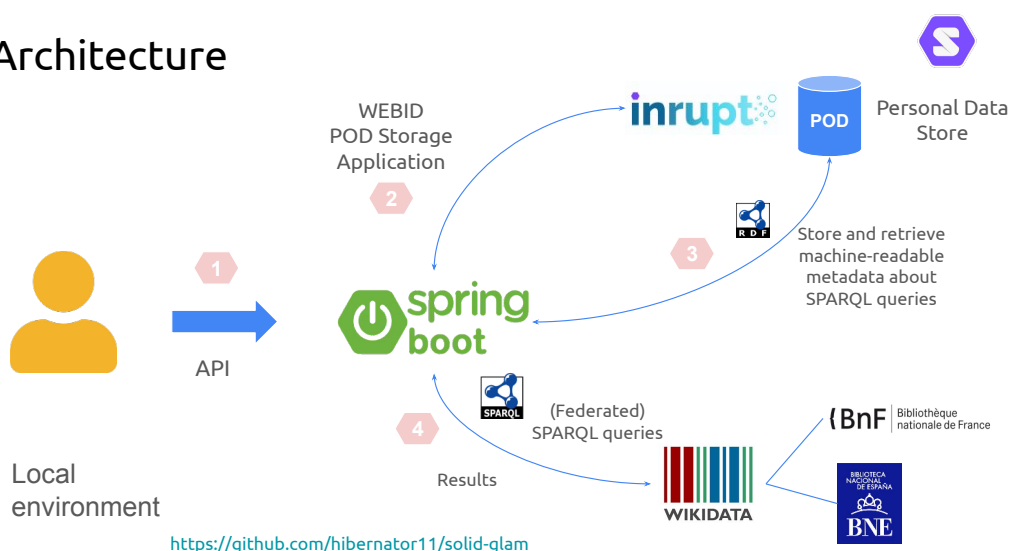
Código

author	expression	title	edition	placeOfPublication	yearOfPublication
<http://data.bnf.fr/ark:/12148/cb/122754921#about>	<http://data.bnf.fr/ark:/12148/cb/34302929#Expression>	Programas de histología e histología normales y anatomía patológica	Madrid : Ministerio de educación y ciencia , 1978	Madrid	1978

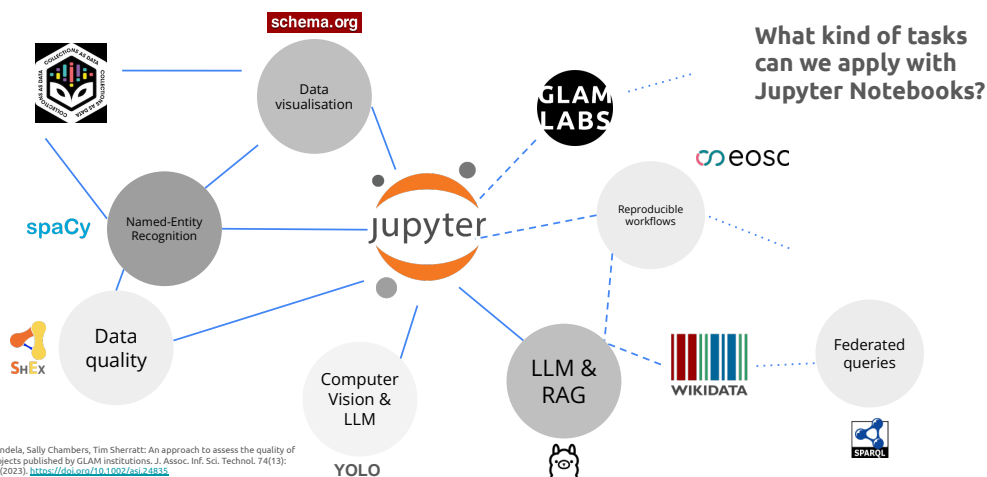
Creating a Solid POD (Personal Data Store) to store a [federated SPARQL query to retrieve all the works from the National Library of France of the author Ramón y Cajal](#)

<https://solidproject.org/>

Architecture



Reproducible data access and reuse: Jupyter Notebooks



Gustavo Candela, Sally Chambers, Tim Sherratt: An approach to assess the quality of
Jupyter projects published by GLAM institutions. *J. Assoc. Inf. Sci. Technol.* 74(13):
1550-1564 (2023). <https://doi.org/10.1002/asi.24835>

European Open Science Cloud - EOSC



European Open Science Cloud - EU Node

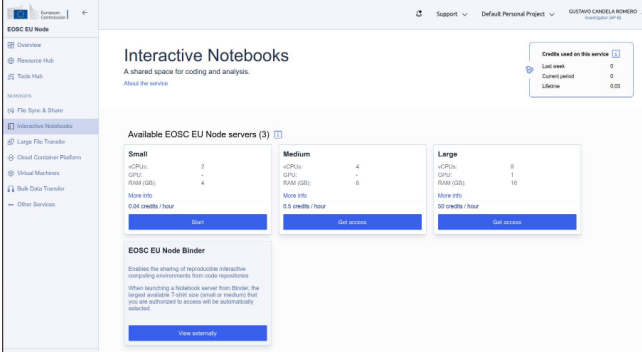
Home | About | Services | Resource Hub | Support | Contributors | News

Home > Our Services > Interactive Notebooks

Interactive Notebooks
A shared space for coding and analysis.

[Get Access](#)

<https://open-science-cloud.ec.europa.eu/services/interactive-notebooks>



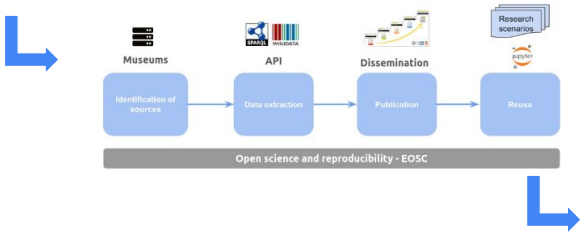
The screenshot shows the 'Interactive Notebooks' page of the EOSC EU Node. It includes a sidebar with navigation options like Overview, Resource Hub, and Tools Hub. The main content area displays 'Available EOSC EU Node servers' with three options: Small, Medium, and Large. Each option lists specifications like vCPUs, GPU, RAM, and storage, along with a 'Get access' button. A 'Credits used on this service' summary is shown in the top right corner.

Extracting collections as data from museums

Article
Preserving Digital Cultural Heritage: A Wikidata Approach to Extract Multimodal Collections as Data in Museums

Gustavo Candela^{1,2}, C. Annemieke Romein^{3,4,5}, Julie M. Birkholz^{6,7}

Journal Digital Heritage:
Advances in Digital Heritage
Preservation and Open Science



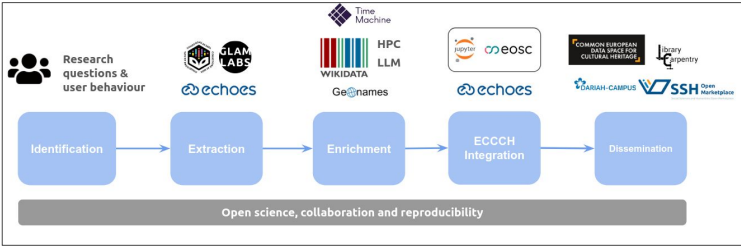
- Scenarios of reuse**
- Federated queries
 - Computer vision
 - RAG search engine
 - LLM chatbot

<https://doi.org/10.20944/preprints202606.1901.v1>

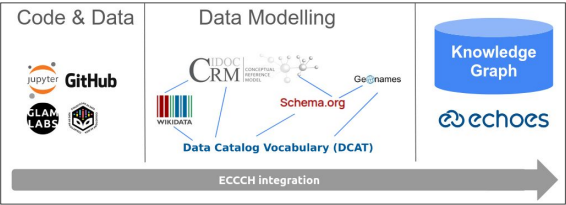
Embedding Jupyter Notebooks in the European Cultural Heritage Cloud

5 use cases

- Using and reusing notebooks in high-performance computing environments
- Application of notebooks in digital twins
- AI preservation
- Open Science for notebooks
- Sustainability for notebooks

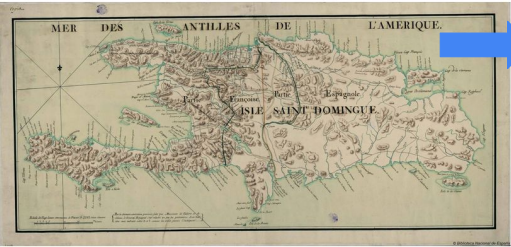


Gustavo Candela, Milena Dobrova, Henk Alkemade, Olga Holownia, Mahendra Mahey, Sarah Ames, Karen Renaud, Ines Vodopivec, Benjamin Charles Germain Lee, Thomas Padilla, Steven Claeysens, Isto Huvila, Beth Knazook, and Sander Münster. 2026. A Use Case Lens on Digital Cultural Heritage. J. Comput. Cult. Herit. Just Accepted (July 2026). <https://doi.org/10.1145/3828548>



LLM & Computer Vision

- <https://bncdigital.bne.es/bd/es/viewer?id=a908cb5-3fb1-4b6c-840d-ecdf9d422198&page=15>
- <https://bncdigital.bne.es/bd/es/viewer?id=a908cb5-3fb1-4b6c-840d-ecdf9d422198&page=41>



As an example, we will use ollama-ocr. We need to download the vision model using the following command:

- `ollama pull llama3.2-vision:11b`

Gustavo Candela, Milena Dobrova, Henk Alkemade, Olga Holownia, Mahendra Mahey, Sarah Ames, Karen Renaud, Ines Vodopivec, Benjamin Charles Germain Lee, Thomas Padilla, Steven Claeysens, Isto Huvila, Beth Knazook, and Sander Münster. 2026. A Use Case Lens on Digital Cultural Heritage. J. Comput. Cult. Herit. Just Accepted (July 2026). <https://doi.org/10.1145/3828548>

```
Now we can use the vision model to extract the text

11) from llama_ocr import OllamaProcessor

# Initialize OCR processor
ocr = OllamaProcessor(model_name="llama3.2-vision:11b") # You can use any vision model available on Ollama

ocr = OllamaProcessor(model_name="llama3.2")
ocr = OllamaProcessor(model_name="llama3.2")

# Process an image
image_path = "img/isle-saint-domingue.jpg"
format_type = "text" # Options: markdown, text, json, structured, key-value
print(result)

Using default prompt: Extract all text content from this image in an "exactly as it appears", without modification, summarization, or omission.
Format the output as requested.

Our headers (e.g., ##, ###) "exactly as they appear in the image"
Preserve original line(s), "exactly as they are"
- Maintain all text formatting (bold, italics, underline) exactly as seen
- Do not add, interpret, or restructure any content"

**Title:** A map of the Caribbean, likely from the 18th century.

**Map Description:**
- The map is divided into two sections, with the top section showing the northern part of the Caribbean and the bottom section showing the southern part.
- The map is oriented with north at the top.
- The map is labeled in French, with place names and geographical features written in a decorative font.
- The map features a mix of hand-drawn and printed elements, including:
  - Printed place names and geographical features.
  - Printed illustrations of ships and other nautical symbols.
  - A key or legend explaining the symbols and abbreviations used on the map.
- A note about the map's source and authorship.
- A note about the map's title and projection.
```

Could we improve the description of maps in institutions using LLM? And create machine-readable metadata?

Challenges



- Do we store all the data in Wikidata? Who have the skills to use Wikidata at large scale? Do we use Wikibase instead?
- Copyrighted material
- Local environment VS Cloud services
- New Wikimedia tools to come...
- Integration with existing EU data infrastructure

Future work

1. Wikibase in GLAM
2. Additional use cases based on the Spanish Civil War & WWII
3. Exploring archives in Chile

zenodo

Search records...

Communities

My dashboard

Log in

Sign up

Published July 1, 2026 | Version v2

Preprint

Open

The Wikibase Ecosystem in DH and GLAM

Lindemann, David¹; Candela, Gustavo^{2,3}; Marchetti, Alessandro^{4,5}; Pellizzari di San Girolamo, Camillo Carlo⁶; Olea, Ismael^{7,3}; Varvantakis, Christos⁸; Assis, Tiago⁹; Molinhe de Almeida, Vera⁹; Obregón Sierra, Ángel^{10,3}; Salgado, Ana⁹; Santiago Faria, Alice¹¹; Schöch, Christof¹²; Trigo, Luis⁹

Show affiliations

DARIAH-EU Working Group DHwiki presents this White Paper about Wikibase and the Wikibase Ecosystem as a digest of its permanent activity in providing documentation about MediaWiki-based software solutions to researchers. The output of the Working group is accessible on its Wikibase instance at <https://dhwiki.wikibase.cloud>. As diverse as the Working Group itself is the target group of this report: DH researchers, GLAM institutions, contributors to the software platforms provided by Wikimedia, and software developers. When starting to engage with Wikidata and Wikibase, we have seen ourselves sometimes lost in a diverse landscape of disperse documentation resources with different scope and aimed at differently skilled readers. With this contribution, we hope to provide a comfortable pathway towards adopting Wikibase in FAIR Linked Data publishing and collaborative data curation workflows.

Files

<https://doi.org/10.5281/zenodo.18077408>

dhwiki_transformations_accepted.pdf

178 VIEWS

198 DOWNLOADS

Show more details

Versions

Version v2	Jul 1, 2026
10.5281/zenodo.21110147	
Version v1	Dec 28, 2025
10.5281/zenodo.18077409	

View all 2 versions

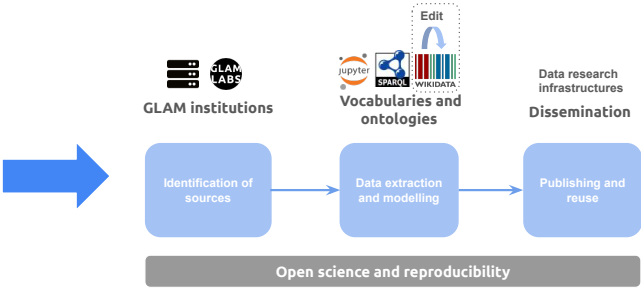
Cite all versions? You can cite all versions by using the DOI 10.5281/zenodo.18077408. This DOI represents all versions, and will always resolve to the latest one. [Read more.](#)



Extracting collections as data & Spanish Civil War

Uses cases

- Refugee ships on which Spanish exiles embarked during the Spanish Exile and the Spanish Civil War
- Artists from the period of the Spanish Republic and Spanish Civil War
- French refugee camps, which concentrated the Spanish Republican exiles of 1939
- Bibliographic records related to the Spanish Civil War



[Guide to the Spanish Exile of 1939 in the State Archives](#)



<https://github.com/hibernator11/Spanish-Civil-War-KGs>

Article published soon! Journal on Computing and Cultural Heritage

Many thanks to Meltem Dişli, Paul Spence, Cezary Rosiński and Arkadiusz Margraf

Spain & World War II

Can we map this information to Wikidata? Can we describe the historical events?



Canfranc International Railway Station

Albert LeLay, chief of French customs in Canfranc, helped refugees and assist Allied intelligence and Resistance networks

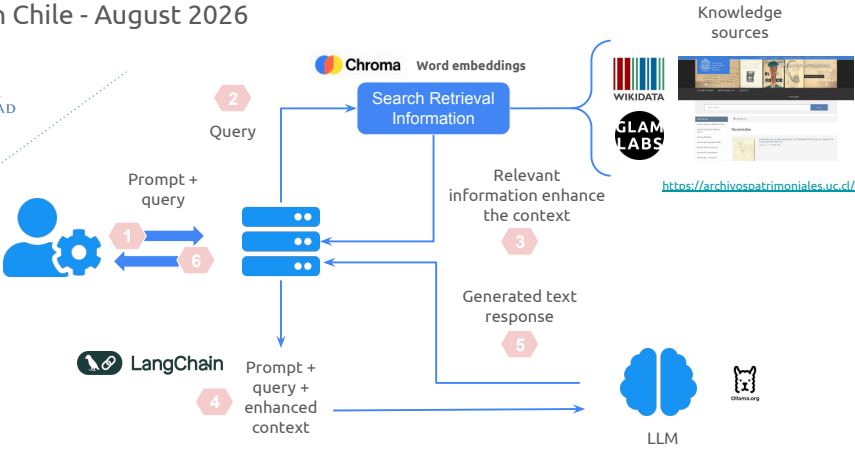


Exploring archives in Chile: A Collections as data approach

Fellowship in Chile - August 2026



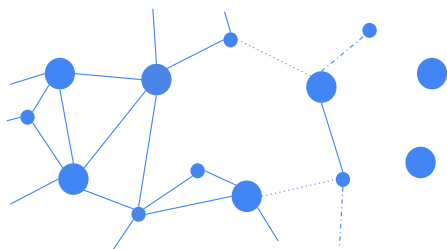
PONTIFICIA
UNIVERSIDAD
CATÓLICA
DE CHILE



Wikidata and GLAM: Enriching Digital Collections

Thank you!

gcandela@ua.es



"Inspiration exists, but it has to find you working." Pablo Picasso

