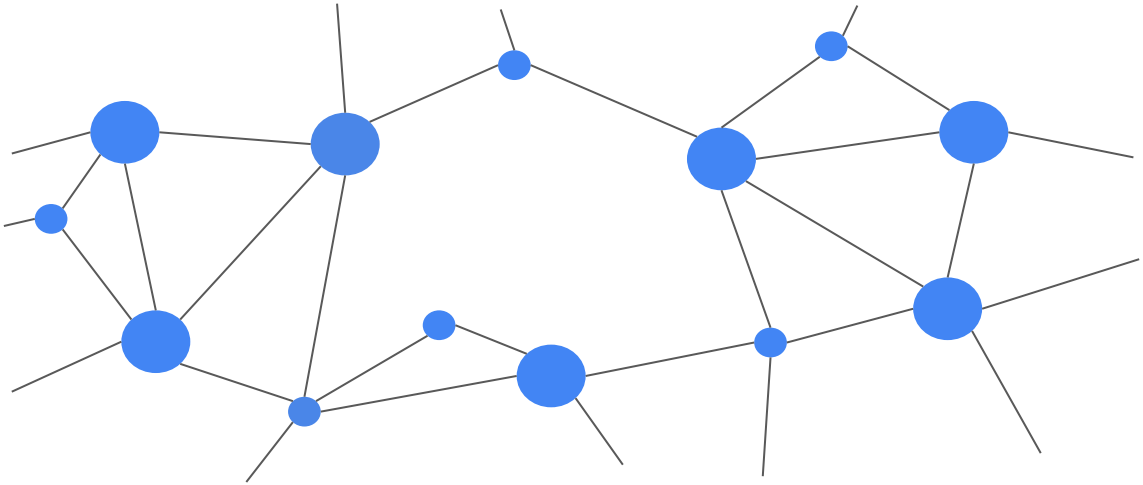
CLARIAH-EUS: Europako ikerketa azpiegiturekin lotuta egongo den euskararako ikerketa azpiegitura eraikitzen

Acceso computacional a colecciones digitales en el contexto GLAM: Balance y nuevos retos

23 de noviembre de 2023

Gustavo Candela



CENTRO BIBLIOTECA VIRTUAL MIGUEL DE CERVANTES

CLARIAH-EUS

Universitat d'Alacant
Universidad de Alicante

CLARIAH-EUS: Europako ikerketa azpiegiturekin lotuta egongo den euskararako ikerketa azpiegitura eraikiz

Acceso computacional a colecciones digitales en el contexto GLAM: Balance y nuevos retos

23 de noviembre de 2023

With many thanks to …

Unidad de Digitalización, Jesús Pradells, Manuel Marco-Such, Rafael Carrasco, Manuel Bravo, Juan Carlos García, Pilar Escobar, Dolores Sáez, David Quintela, Ester Serna, Delia De Sayas, María Elena Sáez, Borja Navarro, Mahendra Mahey, Sally Chambers, Sarah Ames, Abbey Potter, Katrine Hofmann, Milena Dobreva, Alba Irollo, Olga Holownia, Tim Sherratt, Nele Gabriëls, Thomas Padilla, Yasmeen Shorish, Neil Fitzgerald, Rossitza Atanassova, Isabel Martínez, Jesse de Does, Tomasz Parkola, Katrien Depuydt, David Abián, Tomás Saorín, Manuela Rodríguez, Apostolos Antonacopoulos, Ines Vodopivec, Annemieke Romein, Mikel Iruskieta, Dolores Romero, Juan Trujillo, Rania Osman José Luis Vicedo, National Library of Scotland, International GLAM Labs Community, INTELE, Universidad de Alicante, Impact Centre of Competence, Collections as Data, Code4lib Editorial Committee, Wikimedia España, Universidad de Murcia, German Rigau, Patricia Murrieta-Flores, Javier Pereda

CENTRO BIBLIOTECA VIRTUAL MIGUEL DE CERVANTES

CLARIAH-EUS

Universitat d'Alacant
Universidad de Alicante

Index

- About me
- Introduction - BVMC steps
- Collections as data & examples
  - Publication
  - Reuse
  - Data quality
  - Catalogue description
- Challenges and future work
- References

CENTRO BIBLIOTECA VIRTUAL
MIGUEL DE CERVANTES

About me

Gustavo Candela

- PhD in Computer Science
- Developer at the BVMC 2010-2023
- Lecturer in Computer Science
- @gus_candela
- Research articles
- gcandela@ua.es

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

https://www.ua.es/

Introduction

The Biblioteca Virtual Miguel de Cervantes has been recently transformed into a **Digital Humanities centre** aiming at promoting research, sharing knowledge as well as designing and developing new tools.

https://www.cervantesvirtual.com/

Introduction

The centre is based at the **University of Alicante** in Spain.

It comprises several departments with **25 people**:

- Cataloguing
- Collections
- IT - Labs
- Digitisation
- Management

https://www.cervantesvirtual.com/

It also includes **researchers** of different academic backgrounds such as Literature, Digital Humanities, History and Computer Science.

Introduction

Engaging with Digital Humanities researchers in Spain

The BVMC is a member of **CLARIAH-ES** and has participated in the original **INTELE** Spanish network.

Introduction

DARIAH-EU
Digital Research Infrastructure for the Arts and Humanities

CLARIN

A research infrastructure that enhances and supports digitally enabled research and teaching across the Arts and Humanities

With thanks to German Rigau, Manuel Marco and Borja Navarro

CLARIAH-ES

https://www.clariah.es/

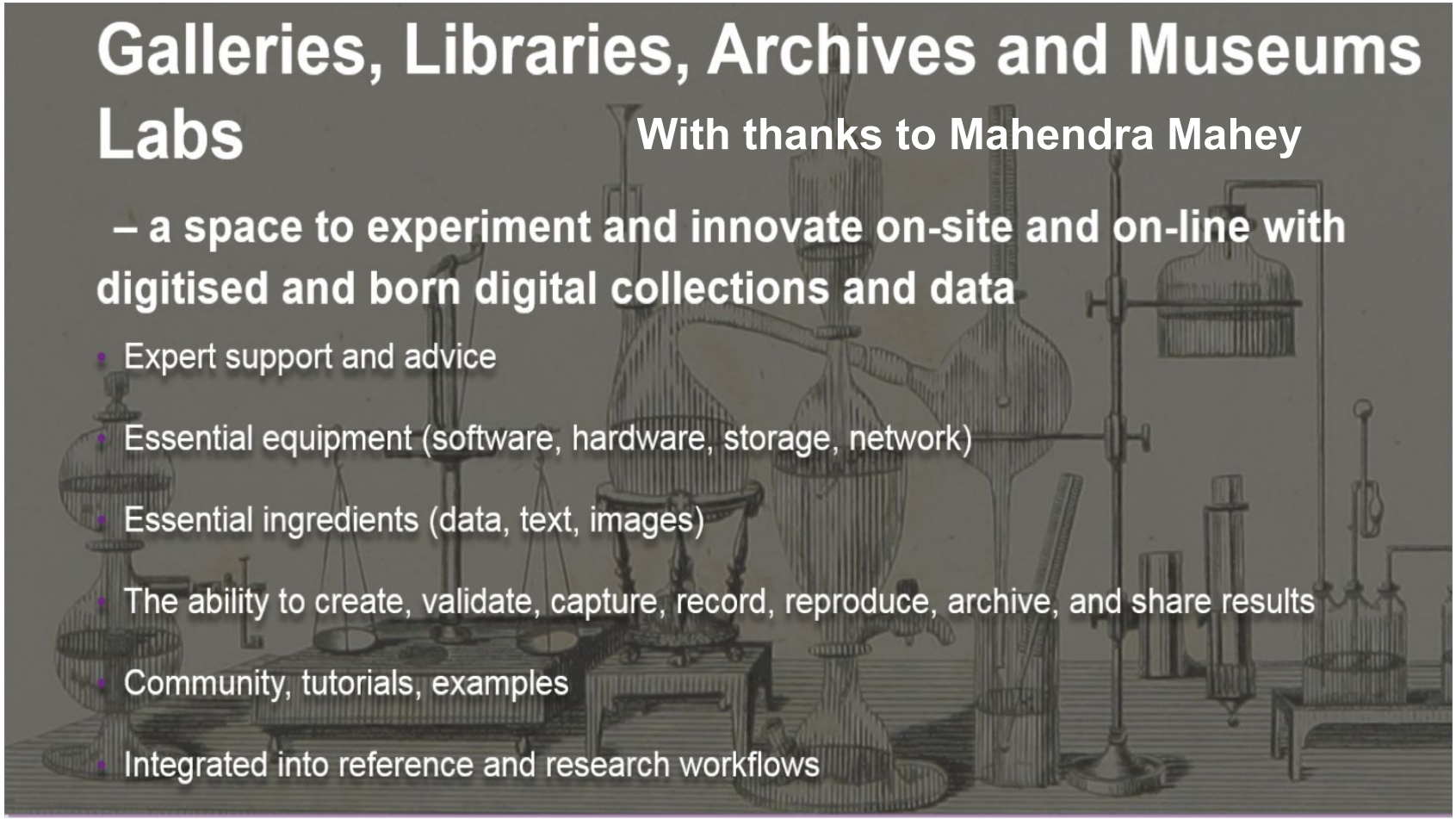
UJA. Universidad de Jaén
UNED
Universitas Complutensis Matritensis
Universitat d'Alacant
Universidad de Alicante
Universidad del País Vasco
Euskal Herriko Unibertsitatea
USC Universidade de Santiago de Compostela
BNE
CSIC Consejo Superior de Investigaciones Científicas
BSC Barcelona Supercomputing Center
Centro Nacional de Supercomputación
ULPGC Universidad de Las Palmas de Gran Canaria

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



Introduction

The **BVMC Labs (data.cervantesvirtual.com)** aims at reusing digital collections in innovative and creative ways

- Tools and prototypes
- Research publications
- Tutorials
- Conferences
- Examples of use

BIBLIOTECA VIRTUAL MIGUEL DE CERVANTES

OPEN A GLAM LAB

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

COLLECTIONS AS DATA

https://collectionsasdata.github.io/

Introduction

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
- …

slack

Jisc email subscription

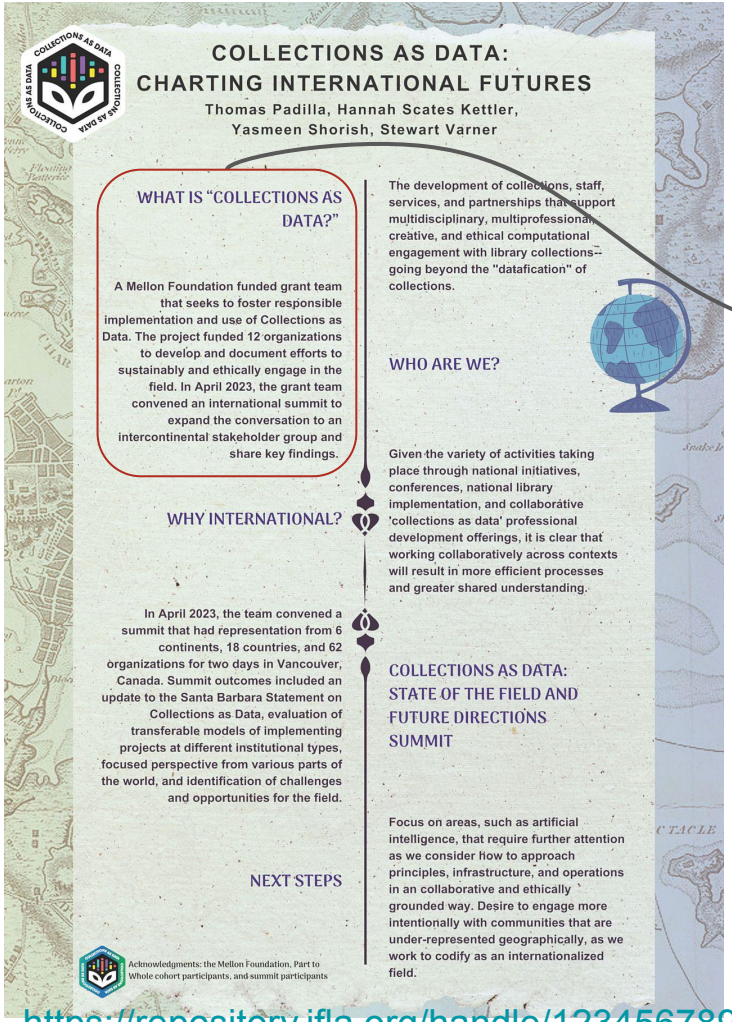
Introduction

Steps

New platform

2015 2016 2017 2018 2019 2020 2021 2022 2023

WIKIDATA



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

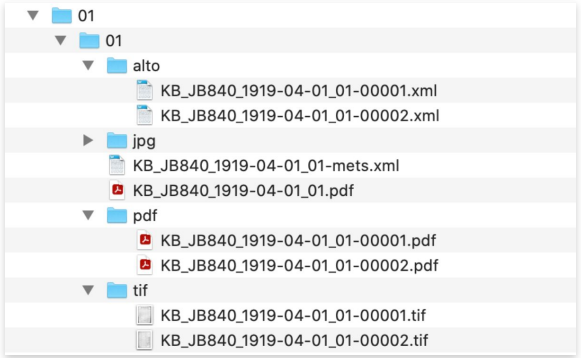
https://repository.ifla.org/handle/123456789/3085

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

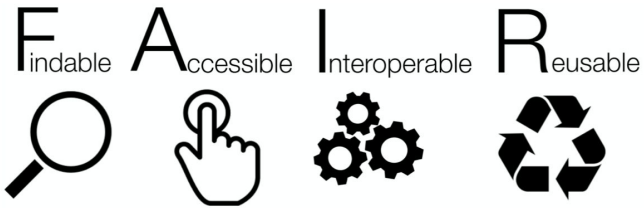
‘Collections as Data’ or ‘Data-level access’

Collections as Data

Providing data-level access to digital collections is a primary challenge for undertaking digital humanities research. In the United States, the flagship initiatives, ‘Always Already Computational: Collections as Data’ and ‘Collections as Data: Part to Whole’, define ‘*Collections as Data*’ as a “conceptual orientation to collections that renders them as ordered information, stored digitally, so that they are inherently amenable to computation”. The initiative was established to document, exchange experience and share knowledge to encourage cultural heritage institutions to implement ‘collections as data’ in their own institutions. **DATA-KBR-BE will kick-start the implementation of ‘Collections As Data’ in Belgium.**



https://collectionsasdata.github.io



Providing access to the underlying files of digitised cultural heritage resources to facilitate data analysis by means of tools and methods developed in the field of digital humanities

https://www.kbr.be/en/projects/data-kbr-be/

With thanks to Sally Chambers

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

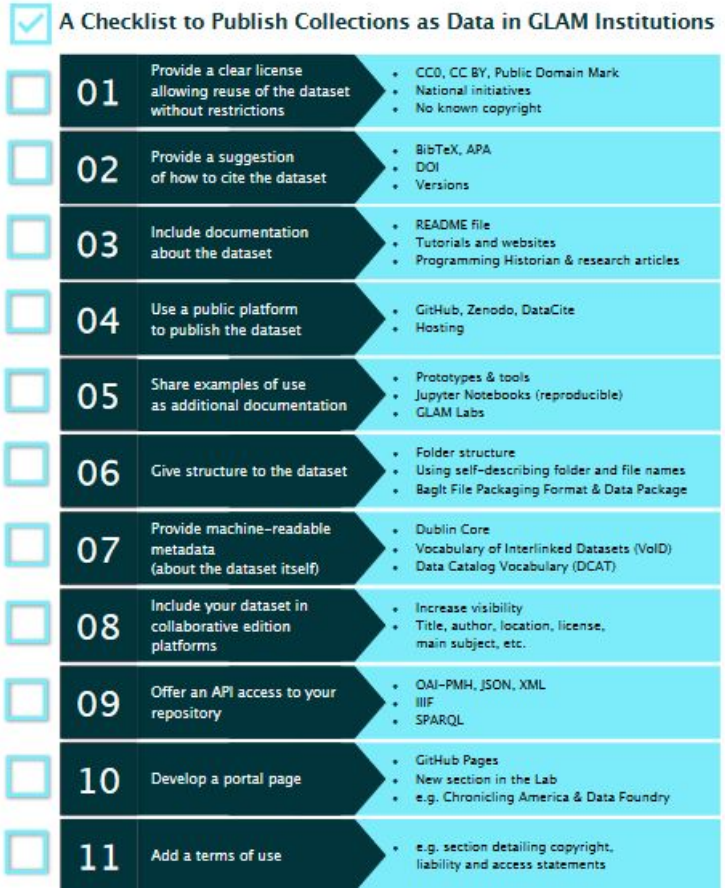
Vancouver Statement on Collections-as-Data
(translations to Spanish, Arabic, French, etc.)

Collections as data

Publication

A checklist to make available digital collections suitable for computational use

International GLAM Labs Community

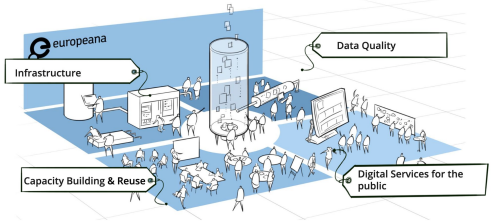


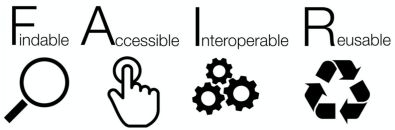
https://doi.org/10.48550/arXiv.2304.02603

Global Knowledge, Memory and Communication

A *Collections as Data* Workflow on the SSH Open Marketplace

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





With thanks to Sally Chambers

Reuse

Jupyter Notebooks

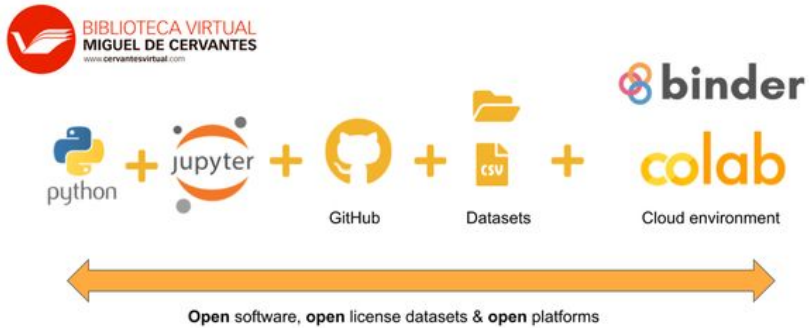
Inspired by:

International GLAM Labs Community

GLAM Workbench

Awards received:

British Library Labs Awards Symposium 2020



https://glamlabs.io/computational-access-to-digital-collections/

Candela, G., Sáez, M. D., Escobar, P., & Marco-Such, M. (2022). Reusing digital collections from GLAM institutions. *Journal of Information Science*, *48*(2), 251-267. https://doi.org/10.1177/0165551520950246

Candela, G., Sáez, M. D., Escobar, P., & Marco-Such, M, "Reutilizar colecciones digitales: GLAM Labs", *Programming Historian en español* 5 (2021), https://doi.org/10.46430/phes0054.

Reuse

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

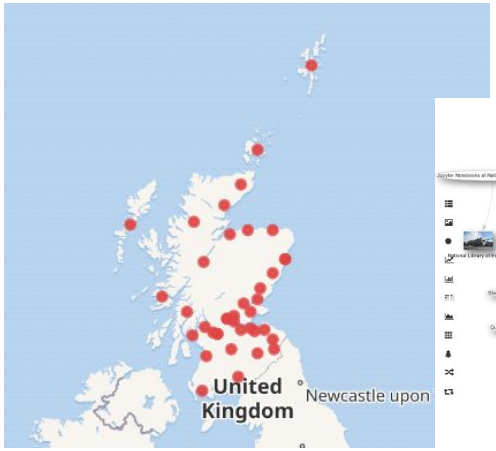
Candela, G., Chambers,S., & Sherratt, T. (2023). An approach to assess the quality of Jupyter projects published by GLAM institutions. Journal of the Association for Information Science and Technology, 1–15. https://doi.org/10.1002/asi.24835

Reuse

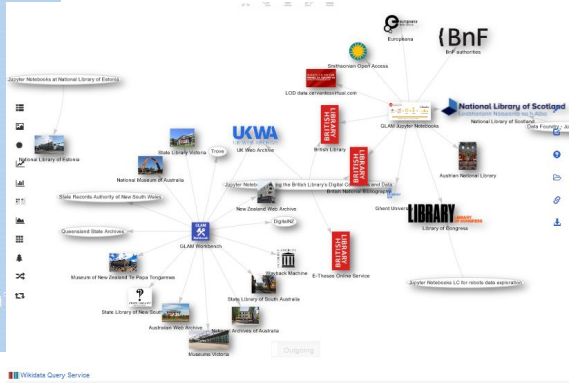
Introduction to SPARQL visualisations

GLAM Labs members

BVMC authors



NLS - Moving Image Archive



Jupyter Notebooks projects

Humanidades Digitales Hispánicas

WIKIDATA

https://humanidadesdigitaleshispanicas.es/compartir-pantalla-con-gustavo-candela-14-de-noviembre-de-2023/

Data quality

europeana pro

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

LOD repositories in GLAM

SAAM

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

Data Foundry
Data collections from the National Library of Scotland

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

LIBRARY LABS
LIBRARY OF CONGRESS

https://id.loc.gov

KUNGL. BIBLIOTEKET
National Library of Sweden

https://libris.kb.se/sparql

16 40
THE NATIONAL LIBRARY OF FINLAND

https://data.nationallibrary.fi

BRITISH LIBRARY

https://bnb.data.bl.uk

DATOS·BNE·es

https://datos.bne.es/

BnF Data

https://data.bnf.fr/

KB LAB

https://data.bibliotheken.nl

CENTRO BIBLIOTECA VIRTUAL MIGUEL DE CERVANTES

https://data.cervantesvirtual.com/

Bibliothèque nationale du Luxembourg
Open Data

https://data.bnl.lu/

DEUTSCHE NATIONAL BIBLIOTHEK

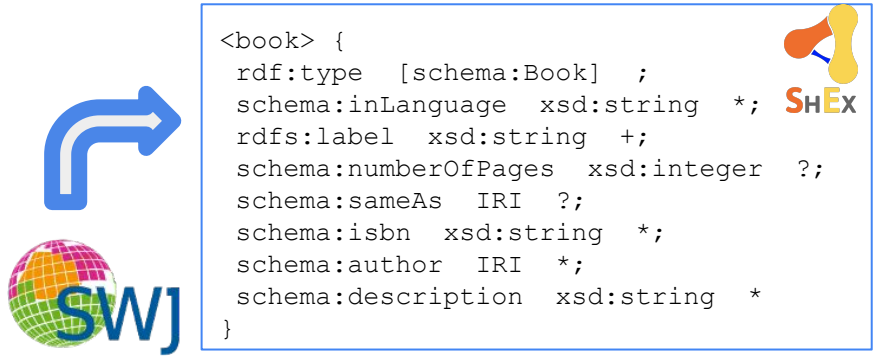
https://www.dnb.de/EN/lds

Österreichische Nationalbibliothek

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

Data quality

Approaches and methods to assess the quality of LOD in GLAM institutions



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

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

?

W3C SPARQL

Evaluating the quality of linked open data in digital libraries

https://doi.org/10.1177/0165551520930951

JASIST JOURNAL OF THE ASSOCIATION FOR INFORMATION SCIENCE AND TECHNOLOGY

SHEX sheXer

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

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

Data quality

Apply the methodology to other research projects

https://data.nls.uk/projects/the-national-librarians-research-fellowship-in-digital-scholarship-2022-23/

Towards a semantic approach in GLAM Labs: The case of the Data Foundry at the National Library of Scotland

https://doi.org/10.1177/01655515231174386

National Library of Scotland
Leabharlann Naiseanta na h-Alba

JASIST
JOURNAL OF THE ASSOCIATION FOR INFORMATION SCIENCE AND TECHNOLOGY

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

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

Publication of metadata as LOD

An ontological approach for unlocking the Colonial Archive

https://doi.org/10.1145/3594727

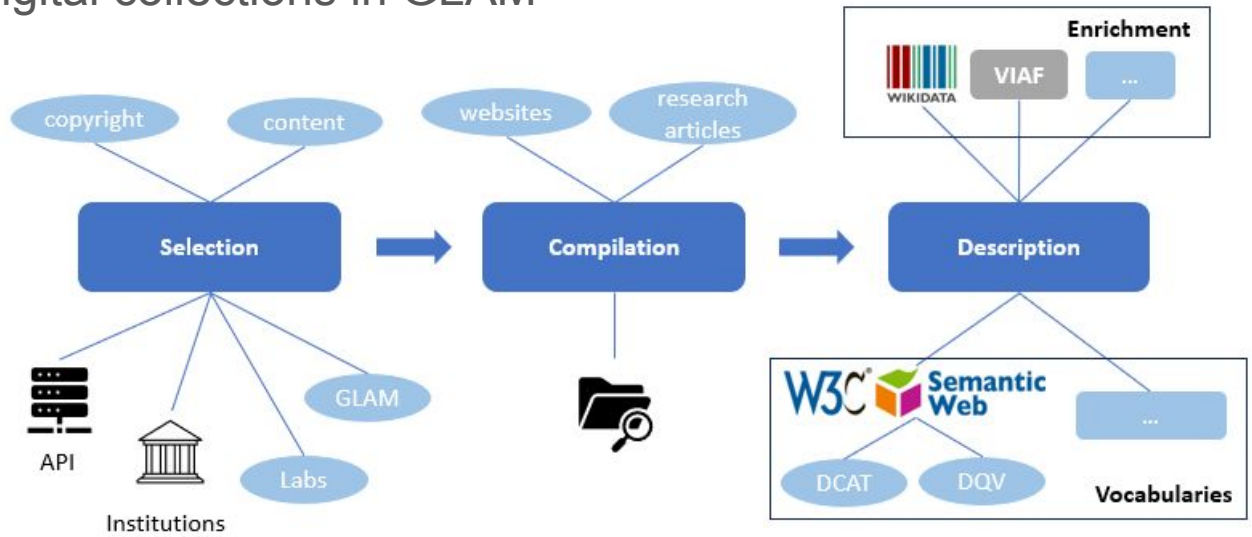
Lancaster University

https://unlockingarchives.com/team/

Collections as data

Catalogue description

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

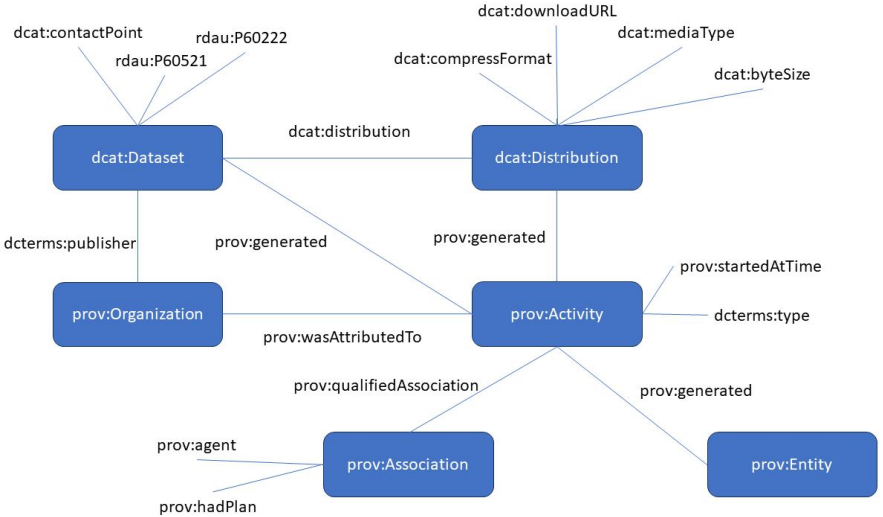
Collections as data

Catalogue description

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

This is the list of examples provided by GLAM institutions that haven been transformed into the Data Catalog Vocabulary. The information to describe the datasets has been retrieved from websites, research articles and projects. In some cases, the DCAT model has been used only too some extent due to the lack of information provided by the institutions regarding the datasets. The entire collection of datasets is available in this file.

- British Library - Free dataset downloads (Alexander the Great and Shakespeare)
- Digital Library of the Caribbean - Panama American - Aruba Esso News
- Europeana - Downloads (Theater Posters and National Heritage Institute Bucharest)
- Harvard Art Museums API
- Library of Congress - Chronicling America
- Metropolitan Museum of Art - Collection API
- National Library of France - Mandragore
- National Library of Luxembourg - Historical Newspapers
- National Library of Scotland - Moving Image Archive - A Medical History of British India
- National Library of the Netherlands - Dutch Novels 1800-2000
- Rijksmuseum - Actors - Thesauri
- South Australian Museum - Minerals Collection
- Zeri Photo Archive - Zeri&LOD



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

Challenges

- GLAM institutions can play a leading role in Artificial Intelligence
  - https://www.rluk.ac.uk/digital-shift-manifesto/
- Adoption of Collection as data by small and medium-sized GLAM institutions
  - Making available full content
  - Open licenses
  - Ethics and terms of use (CARE principles)
- Improving OCR quality
  - Impact White Paper
- Diversity in the controlled vocabularies used to describe metadata as LOD in GLAM

impact
digitisation.eu
centre of competence

RLUK Research Libraries UK

Future work

- Towards a data space for Cultural Heritage
  - Europeana Metis (data aggregation workflow)
  - https://share3d.eu/
- Systematic review of Wikidata in GLAM
  - New roles in GLAM: Wikidata librarian/curator
- Reusing digital collections
  - NER training and poetry
  - Measuring diversity of data and metadata in digital libraries
- Best practices, curricula, guidelines, tutorials such as Programming Historian

DARIAH-CAMPUS

Universitat d'Alacant
Universidad de Alicante

https://pro.europeana.eu/files/Europeana_Professional/Publications/data_space_annual_report_2022_2023.pdf

References

- 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
- https://www.dariah.eu/2023/09/04/spain-joins-dariah-as-full-member/

References

- 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., Sáez, M. and Marco-Such, M. A Shape Expression approach for assessing the quality of Linked Open Data in libraries. Semantic Web 14(2): 159-179 (2023). https://doi.org/10.3233/SW-210441
- 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., Dobreva, 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-EUS: Europako ikerketa azpiegiturekin lotuta egongo den euskararako ikerketa azpiegitura eraikitzen

Acceso computacional a colecciones digitales en el contexto GLAM: Balance y nuevos retos

23 de noviembre de 2023

Eskerrik Asko

Gustavo Candela
gcandela@ua.es
@gus_candela

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

CENTRO BIBLIOTECA VIRTUAL MIGUEL DE CERVANTES

CLARIAH-EUS

Universitat d'Alacant
Universidad de Alicante