Collections as data: Acceso computacional a colecciones digitales

Collections as data: Computational access to digital collections

Gustavo Candela

Candela, Gustavo (2024). "Collections as data: Acceso computacional a colecciones digitales". *Anuario ThinkEPI*, v. 18, e18e06.

https://doi.org/10.3145/thinkepi.2024.e18a06

Publicado en *IweTel* el 19 de marzo de 2024

Gustavo Candela
https://www.directorioexit.info/ficha6986
https://orcid.org/0000-0001-6122-0777
Universidad de Alicante
gcandela@ua.es

Resumen: Durante décadas, las instituciones de patrimonio cultural han explorado nuevas formas de hacer accesibles sus colecciones digitales para la comunidad investigadora. La iniciativa *Collections as data* promueve la publicación de colecciones digitales que facilitan el acceso computacional. La comunidad ha publicado buenas prácticas y ejemplos de proyectos para facilitar su adopción. En esta nota, se introduce la iniciativa *Collections as data* en el marco de las instituciones de patrimonio cultural junto con los proyectos más representativos, buenas prácticas y los nuevos desafíos.

Palabras clave: *Collections as Data*; Acceso computacional; Bibliotecas; GLAM.

Abstract: During the last years, Cultural Heritage institutions have been exploring new ways to make available their digital collections. Collections as data is an initiative that fosters the publication of digital collections suitable for computational use. The research community has made available best practice and examples of projects to facilitate the adoption of the Collections as Data principles. This work presents the initiative Collections as data and introduces relevant projects based on Cultural Heritage institutions, best practice and new challenges.

Keywords: Collections as Data; Computational access; Digital libraries; GLAM

1. Introducción

Desde hace décadas, las instituciones de patrimonio cultural, y en concreto las instituciones GLAM (del inglés *Galleries, Libraries, Archives and Museums*), han explorado diversas formas para hacer sus colecciones accesibles para la comunidad. Los avances tecnológicos recientes en materia de Inteligencia Artificial han motivado la adaptación de las instituciones de patrimonio cultural a las nuevas necesidades de la comunidad investigadora que requiere formas más flexibles y eficientes para acceder al contenido de las colecciones digitales. Las comunidades de ámbito internacional

como la *International GLAM Labs community* y *AI-4LAM* promueven la reutilización, el acceso computacional y el uso de la Inteligencia Artificial en el marco de las instituciones de patrimonio cultural (**Mahey** *et al.*, 2019). En paralelo se han creado recientemente infraestructuras de datos que promueven la ciencia abierta como, por ejemplo, la *European Open Science Cloud*.

Desde hace décadas, las instituciones de patrimonio cultural han explorado diversas formas para hacer sus colecciones más accesibles para la comunidad

En este nuevo contexto, la iniciativa *Collections as data* promueve la publicación de colecciones digitales que facilitan el acceso computacional (**Padilla** *et al.*, 2019). Entre los tipos de trabajos basados en el acceso computacional se incluye, por ejemplo, la minería de textos, la visión por computador, el aprendizaje automático o la visualización de datos. Atendiendo al uso creciente de este tipo de colecciones como fuente de datos para el entrenamiento en Inteligencia Artificial, la declaración *Vancouver Statement* sugiere una serie de principios que fomentan el uso responsable de las colecciones digitales publicadas como *Collections as data* (**Padilla** *et al.*, 2023). Desde su aparición, numerosas propuestas de diferentes partes del mundo han explorado su adopción (**Chambers** *et al.*, 2023). Sin embargo, publicar colecciones como datos puede convertirse en una tarea compleja, así como conllevar riesgos en cuanto al uso inadecuado y la utilización de métodos carentes de transparencia.

2. Publicación de *Collections as data*: buenas prácticas

La forma en la que las colecciones digitales se publican difiere entre unas instituciones y otras debido a diferentes factores como pueden ser el tipo de contenido, las técnicas utilizadas, el tamaño de los contenidos o los recursos disponibles. Sin embargo, existe una serie de requisitos que resultan fundamentales a la hora de facilitar su reutilización como la licencia, la documentación sobre el proceso de creación o la estructura del contenido proporcionado. De forma adicional, la publicación de ejemplos de reutilización basados en una colección digital, ya sea en forma de prototipos sencillos o código reproducible, puede estimular a la comunidad a la hora de reutilizar los contenidos. En este sentido, las instituciones han comenzado a publicar proyectos basados en *Jupyter Notebooks* que combinan código reproducible y documentación para introducir cómo reutilizar las colecciones digitales (**Candela**; **Chambers**; **Sherratt**, 2023).

La adopción de *Collections as data* por parte de las instituciones puede resultar una tarea compleja, especialmente para las instituciones de menor tamaño y con menos recursos. Documentar las colecciones digitales no es una labor sencilla teniendo en cuenta la diversidad de los tipos de contenido e instituciones (**Alkemade** *et al.*, 2023). En este sentido, la *International GLAM Labs Community* identificó la necesidad de crear una guía de buenas prácticas para publicar *Collections as data.* Como resultado, y tras un proceso de edición colaborativo, se obtuvo una lista de verificación (*checklist* en inglés) con un conjunto de pasos que a su vez fueron refinados en una serie de seminarios online posteriores (**Candela** *et al.*, 2023). La lista incluye tareas que cubren diferentes aspectos como, por ejemplo, proporcionar una licencia abierta que permita la reutilización, incluir información sobre cómo citar la colección digital o proporcionar documentación adicional sobre el proceso de creación. Con el objetivo de automatizar el proceso de publicación de *Collections as data*, la lista de verificación ha sido transformada recientemente en un flujo de trabajo que además proporciona una lista de ejemplos con información adicional para cada uno de los pasos a modo de introducción (**Candela**; **Chambers**; **Irollo**, 2024). Además, recientemente se ha creado un grupo de interés

La iniciativa *Collections as Data* promueve la publicación de colecciones digitales que facilitan el acceso computacional y su uso responsable

con el objetivo de promocionar la publicación de *Collections as data* en el marco de la iniciativa internacional *Research Data Alliance* (*Research Data Alliance*, 2024). Este tipo de iniciativas fomenta la colaboración de forma abierta y activa entre los miembros de las distintas comunidades para compartir el conocimiento y desarrollar proyectos relacionados con *Collections as data*.

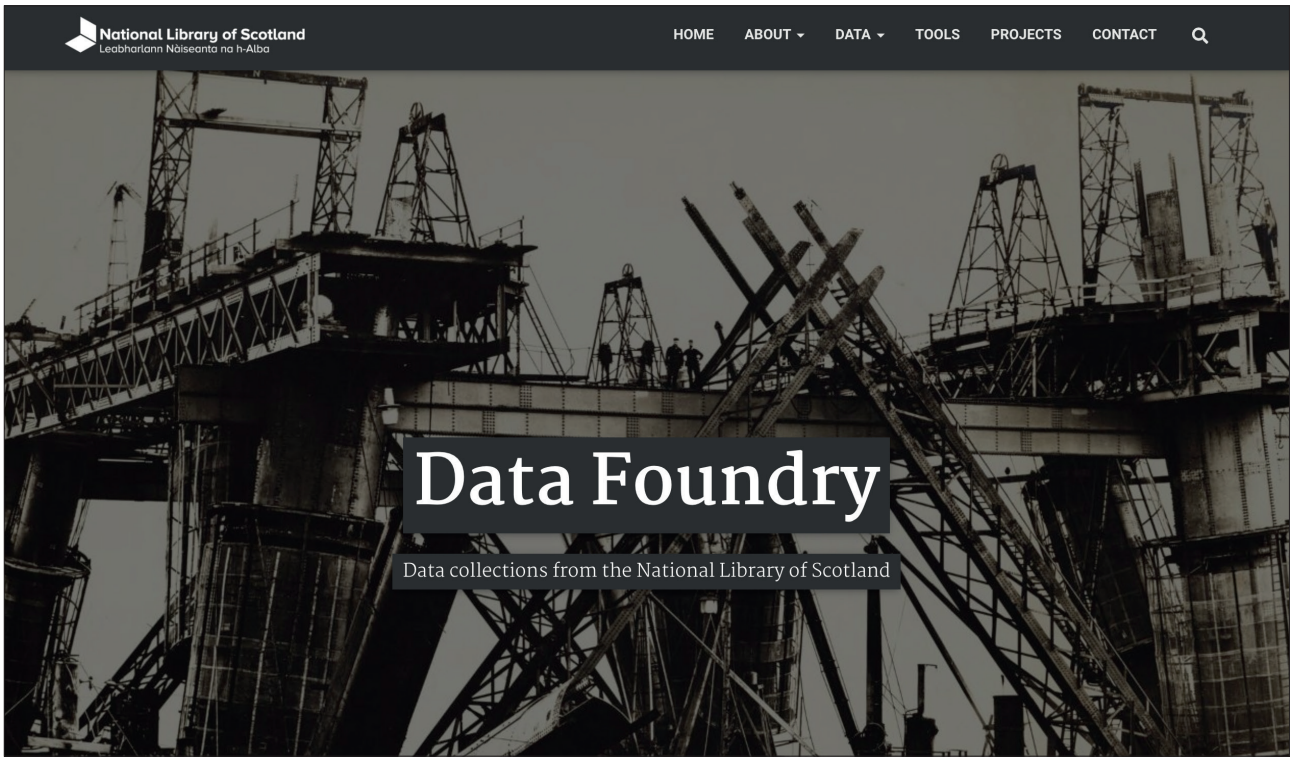
Entre los ejemplos más representativos de *Collections as data* se encuentra el *Data Foundry* de la *National Library of Scotland* que publica colecciones digitales con licencias abiertas que permiten su reutilización

3. *Collections as data*: ejemplos de proyectos

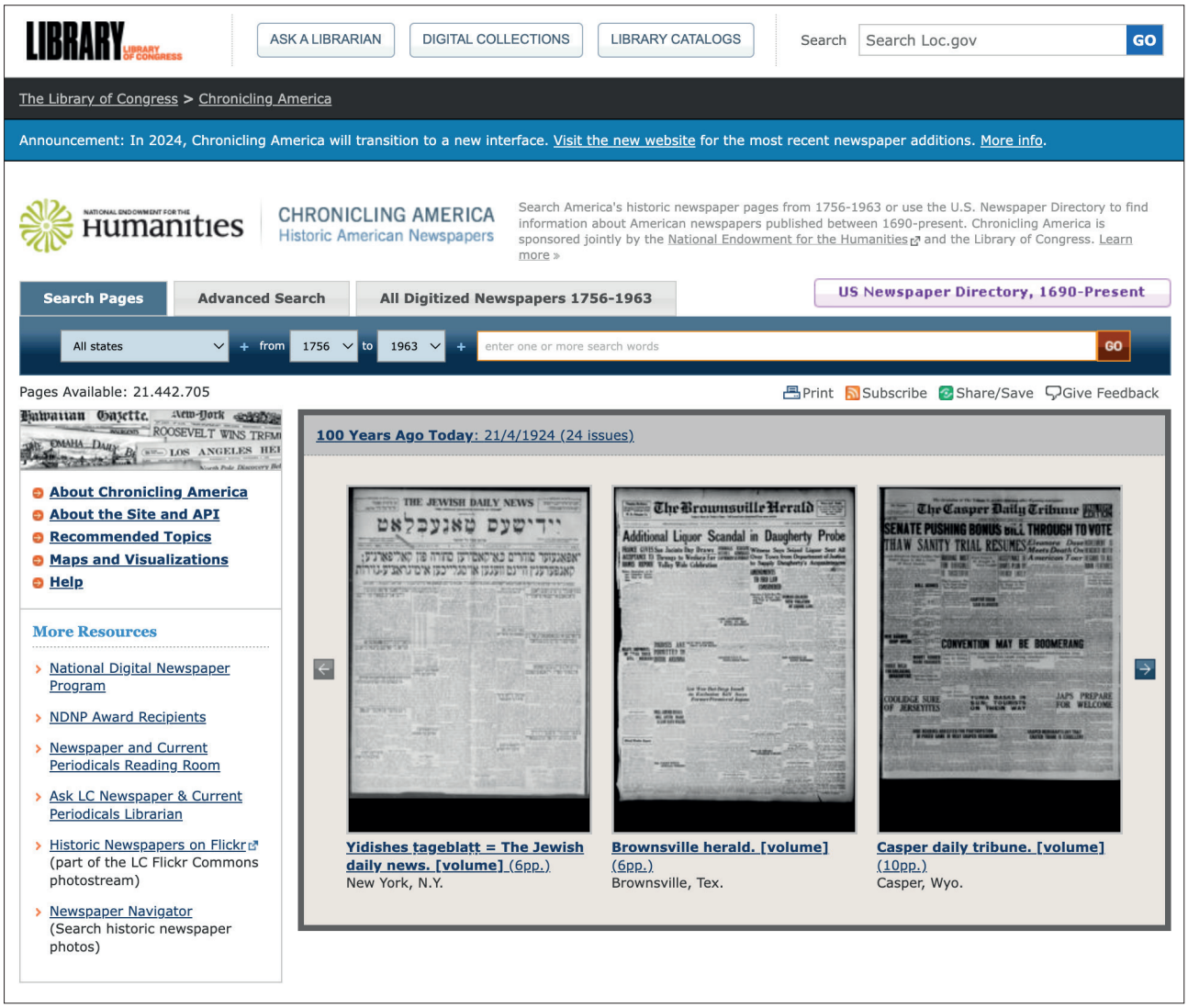
Actualmente, existe una gran diversidad de proyectos que proporcionan colecciones digitales que aplican los principios de *Collections as data*. En esta sección se introducen brevemente a modo de ejemplo algunos de ellos.

El *Data Foundry* de la *National Library of Scotland* es uno de los ejemplos más representativos de proyectos que publican *Collections as data* (**Ames**; **Lewis**, 2020). Las colecciones digitales disponibles en este proyecto incluyen contenido de diferente tipo como, por ejemplo, metadatos, texto obtenido a través de métodos de reconocimiento óptico de caracteres OCR (en inglés, *Optical Character Recognition*) que en algunos casos se ha corregido manualmente, imágenes y mapas. Todas las colecciones incluyen información relacionada con el tipo y formato del contenido proporcionado, cómo citar la colección, la licencia de uso, así como ejemplos de reutilización basados en *Jupyter Notebooks*.
https://data.nls.uk

Otro ejemplo relevante es la *Bibliothèque Nationale du Luxembourg* que facilita su colección de periódicos históricos a través de diferentes formatos y tamaños. Además, existen numerosas instituciones GLAM que facilitan sus colecciones digitales como *Collections as data* entre las que se encuentra la *Library of Congress* y su iniciativa *Chronicling America* de periódicos históricos, el *Research*



https://data.nls.uk



https://chroniclingamerica.loc.gov

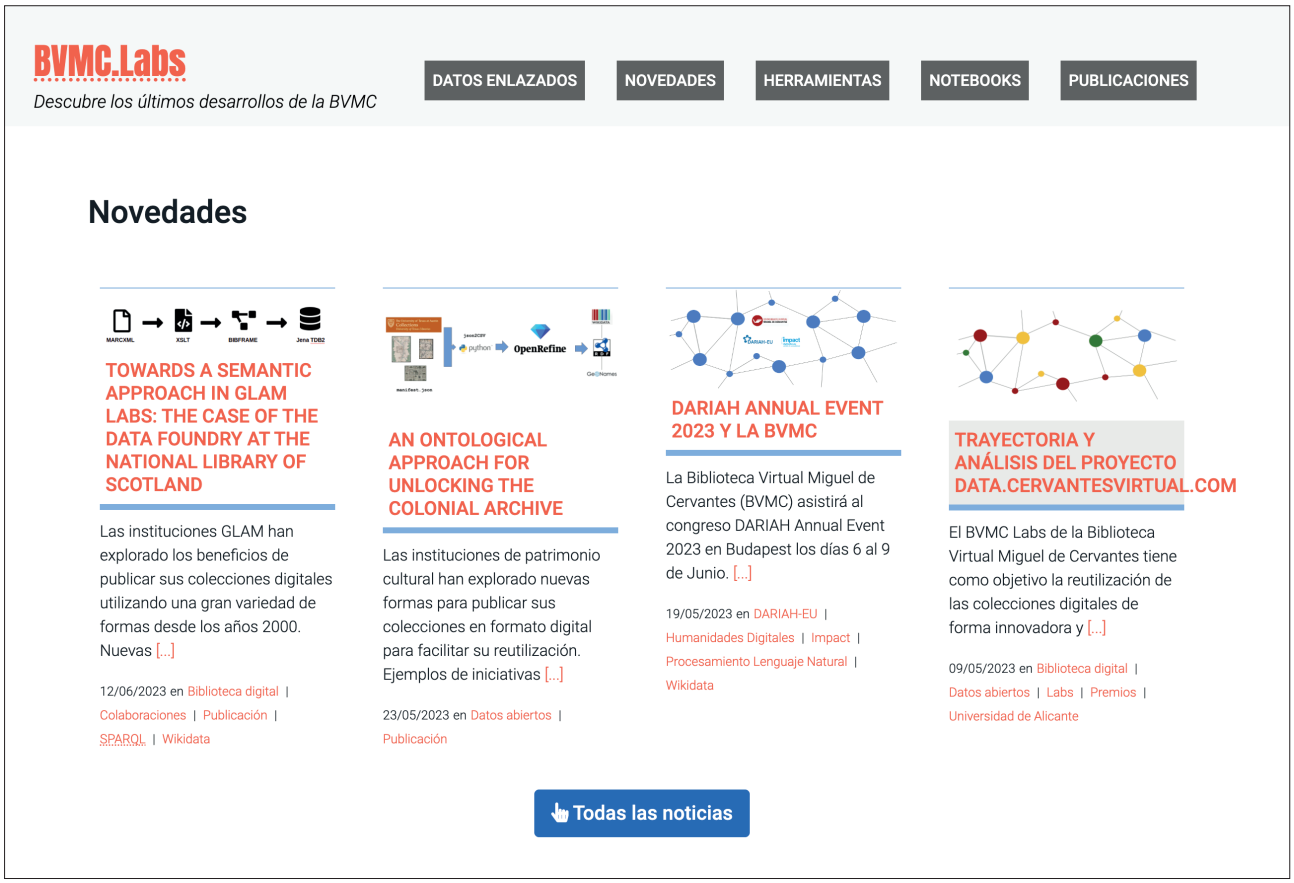
Repository de la *British Library* que proporciona acceso a sus colecciones digitales, el *Rijksmuseum* y las iniciativas de tipo Lab de la *Biblioteca Nacional de España* y la *Biblioteca Virtual Miguel de Cervantes*.

https://chroniclingamerica.loc.gov
https://bl.iro.bl.uk
https://data.rijksmuseum.nl
https://bnelab.bne.es
https://data.cervantesvirtual.com

4. Desafíos y oportunidades

Actualmente se plantean numerosos desafíos a los que se enfrentan las instituciones de patrimonio cultural. En primer lugar, el uso y aplicación de métodos carentes de transparencia por parte de ciertas instituciones relacionadas con la Inteligencia Artificial. En este sentido, la *KB Nationale Bibliotheek* (Biblioteca Nacional de Holanda) optó recientemente por limitar el acceso para determinado tipo de usuarios (*Koninklijke*

La adopción de *Collections as data* por parte de las instituciones puede resultar una tarea compleja, especialmente para las instituciones de menor tamaño y con menos recursos



https://data.cervantesvirtual.com

Bibliotheek, 2024). Además, resulta fundamental disponer de una infraestructura tecnológica que garantice el funcionamiento y la seguridad (*British Library*, 2024). A la hora de publicar colecciones, es esencial considerar aspectos relacionados con la ética y la responsabilidad para garantizar el beneficio colectivo tal y como se describe en los principios CREA para la gobernanza de datos indígenas (**Carrol** *et al.*, 2020). En ese sentido, las instituciones de patrimonio cultural han de adaptarse en los próximos años al nuevo contexto y necesidades de la comunidad investigadora (*Research Libraries UK*, 2020).

La publicación de colecciones que facilitan el acceso computacional puede resultar compleja debido a diferentes factores. Sin embargo, las instituciones se pueden beneficiar de las buenas prácticas y guías recientemente publicadas por la comunidad investigadora. En los próximos años infraestructuras de datos como la *European Open Science Cloud* van a jugar un papel fundamental en materia de estandarización y ciencia abierta.

5. Referencias

Alkemade, Henk; **Claeyssens, Steven**; **Colavizza, Giovanni**; **Freire, Nuno**; **Lehmann, Jörg**; **Neudecker, Clemens**; **Osti, Giulia**; **Van-Strien, Daniel** (2023). "Datasheets for digital cultural heritage datasets". *Journal of open humanities data*, v. 9 n. 1 pp. 17.
https://doi.org/10.5334/johd.124

Ames, Sarah; **Lewis, Stuart** (2020). "Disrupting the library: Digital scholarship and big data at the National Library of Scotland". *Big data & society*, v. 7, n. 2.
https://doi.org/10.1177/2053951720970576

British Library (2024). *Learning lessons from the cyber-attack, British Library cyber incident review.*
https://www.bl.uk/home/british-library-cyber-incident-review-8-march-2024.pdf

Candela, Gustavo; **Chambers, Sally**; **Sherratt, Tim** (2023). "An approach to assess the quality of Jupyter projects published by GLAM institutions". *Journal of the Association for Information Science and Technology*, v. 74, n. 13, pp. 1550-1564.
https://doi.org/10.1002/asi.24835

Candela, Gustavo; **Gabriëls, Nele**; **Chambers, Sally**; **Dobreva, Milena**; **Ames, Sarah**; **Ferriter, Meghan**; **Fitzgerald, Neil**; **Harbo, Victor**; **Hofmann, Katrine**; **Holownia, Olga**; **Irollo, Alba**; **Mahey, Mahendra**; **Manchester, Eileen**; **Pham, Thuy-An**; **Potter, Abigail**; **Van-Keer, Ellen** (2023). "A checklist to publish collections as data in GLAM institutions". *Global knowledge, memory and communication*.
https://doi.org/10.1108/GKMC-06-2023-0195

Candela, Gustavo; **Chambers, Sally**; **Irollo, Alba** (2024). "A workflow to publish Collections as Data: the case of Cultural Heritage data spaces". *SSH Open marketplace.*
https://marketplace.sshopencloud.eu/workflow/l3JvP6

Carroll, Stephanie-Russo; **Garba, Ibrahim**; **Figueroa-Rodríguez, Oscar L.**; **Holbrook, Jarita**; **Lovett, Raymond**; **Materechera, Simeon**; **Parsons, Mark**; **Raseroka, Kay**; **Rodriguez-Lonebear, Desi**; **Rowe, Robyn**; **Sara, Rodrigo**; **D. Walker, Jennifer**; **Anderson, Jane**; **Hudson, Maui** (2020). "The CARE Principles for indigenous data governance". *Data science journal*, v. 19, n. 1, pp. 43.
https://doi.org/10.5334/dsj-2020-043

Chambers, Sally; **Walsh, Melanie**; **Caswell, Michelle**; **Harder, Geoff**; **Okumura, Mercedes**; **Corrin, Julia**; **Baeza-Ventura, Gabriela**; **Antonijevic, Smiljana**; **Knazook, Beth**; **Narlock, Mikala**; **Bailey, Jefferson**; **Neudecker, Clemens**; **Downie, J. Stephen**; **Layne-Worthey, Glen**; **Van-Strien, Daniel**; **Irollo, Alba**; **Whitmire, Amanda**; **Lee, James**; **Berry, Dorothy**; **Del-Rio-Riande, Gimena**; **Bordalejo, Barbara**; **Buckland, Amy**; **Vollmer, Timothy**; **McLellan, Robert**; **Henley, Amanda**; **Talboom**, **Leontien**; **Nekesa, Wyne**; **Milligan, Ian**; **Owens, Trevor**; **Loxton, Duncan**; **Perez, Paul-Jason**; **Scheltjens, Saskia**; **Claeyssens, Steven**; **Pham, Kim**; **Russey-Roke, Elizabeth**; **Jordan, Kari L.**; **Tsang, Martin**; **Allen, Laurie**; **Weber, Chela-Scott**; **Hawkins, Kevin**; **Cox, Andrew**; **Evangelestia-Dougherty, Tamar**; **Pena, Patricia**; **Ng, Yvonne**; **Garcia-Merchant, Linda**; **Candela, Gustavo**; **Ranade, Sonia**; **Tindall, Alexis**; **Riley, Jenn**; **Becker, Devin**; **Lar-Son, Kayla**; **Varela, Miguel-Escobar**; **Schallier, Wouter**; **Leigland, Linn**; **Warren, Margaret**; **Abner, Kayla**; **Maemura, Emily**; **Hamilton, Summer**; **Ridge, Mia** (2023). *Declaraciones de posturas -> Collections as Data: State of the Field and Future Directions* (ES). Zenodo.
https://doi.org/10.5281/zenodo.10412779

Koninklijke Bibliotheek (2024). "KB restricts access to collections for training commercial AI". *KB Nationale Bibliotheek.*
https://www.kb.nl/en/nieuws/kb-restricts-access-to-collections-for-training-commercial-ai

Mahey, Mahendra; **Al-Abdulla, Aisha**; **Ames, Sarah**; **Bray, Paula**; **Candela, Gustavo**; **Chambers, Sally**; **Derven, Caleb**; **Dobreva-McPherson, Milena**; **Gasser, Katrine**; **Karner, Stefan**; **Kokegei, Kristy**; **Laursen, Ditte**; **Potter, Abigail**; **Straube, Armin**; **Wagner, Sophie-Carolin**; **Wilms, Lotte** (2019). *Open a GLAM Lab*. Doha: Book Sprint. ISBN: 978 16 4606 142 6

Padilla, Thomas; **Allen, Laurie**; **Frost, Hannah**; **Potvin, Sarah**; **Russey-Roke, Elizabeth**; **Varner, Stewart** (2019). *Final report. Always already computational: Collections as data* (Versión 1). Zenodo.
https://doi.org/10.5281/zenodo.3152935

Padilla, Thomas; **Scates-Kettler, Hannah**; **Varner, Stewart**; **Shorish, Yasmeen** (2023). *Vancouver statement on collections as data* (ES). Zenodo.
https://doi.org/10.5281/zenodo.8341571

Research Data Alliance (2024). *Collections as Data IG Charter (transitioning from Archives and Records Professionals for Research Data IG).*
https://archive.rd-alliance.org/comment/31727

Research Libraries UK (2020). *A manifesto for the digital shift in research libraries*.
https://www.rluk.ac.uk/digital-shift-manifesto/