Collections as Data: Getting Started

A LIBER Digital Scholarship & Data Science Topic Guide for Library Professionals

Gustavo Candela

2024-06-04

A quick overview of the Collections as Data movement and what it can do for your library, along with some practical advice on how to take steps to transform your own library collections.

Introduction

GLAM (Galleries, Libraries, Archives and Museums) institutions have been making their content available to the public and researchers in a variety of formats for different purposes for decades. From maps, images, text, historical newspapers or postcards, digitising collections and displaying our items online are at the core of making this information accessible for research and enjoyment. But recent advances in and expansion of the use of AI and Machine Learning in library work and library research have transformed the way in which we and our users expect to access, view, use and search our collections. As the demand for computationally accessible cultural heritage collections continues to grow, and we're being asked to provide our collections in the form of datasets for instance, how do we go about this practically?

In this new context, where GLAM institutions now find themselves playing a leading role as data providers and data curators, the Collections as Data concept emerged as a new approach, guidance and framework to support responsible development and computational use of digital cultural heritage collections. By computational use we mean the application of computational techniques such as natural language processing (NLP), computer vision, and more to the analysis, exploration and reuse at scale of digitised cultural heritage content. For more practical use cases see our AI and Machine Learning in Libraries topic guide.

Three particularly useful outputs from the Collections as Data community of practice are:

- Fifty things (2018)

"Want to support collections as data at your institution, but not sure how to begin? Drawing on what we learned from engaging with practitioners and researchers throughout the Always Already Computational project, the project team compiled a list of 50 Things you can do to get started. 50 Things is intended to open eyes, stimulate conversation, encourage stepping back, generate ideas, and surface new possibilities. If any of that gets traction, then perhaps you can make the case for investing in collections as data at your institution in a meaningful, if not systematic, way."

- Vancouver Statement on Collections as data (2023)

"The Vancouver Statement suggests a set of principles for thinking through questions that collections-as-data work produces, as part of an expanding global, interprofessional, and interdisciplinary effort to empower memory, knowledge, and data stewards (for example, practitioners and scholars) who aim to support responsible development and computational use of collections as data. This stewardship role only grows in importance as artificial intelligence applications, trained on vast amounts of data, including collections as data, impact our lives ever more pervasively."

- A Checklist to Publish Collections as Data in GLAM Institutions (2023)

Developed as part of the collaborative work of the International GLAM Labs community, the checklist covers different aspects such as providing a suggested citation, including documentation about the datasets (for example, README files or tutorials) or sharing examples of use (for example, prototypes or Jupyter Notebooks).

Relevance to the Library Sector (Case Studies/Use Cases)

Many institutions have started to adopt the collections as data principles to varying degrees, often combining publication of digital collections suitable for computational use, with a lab offering technical support to use them. You might notice that the examples below are all slightly different implementations of the principles though as each institution will necessarily need to decide for themselves, based on their individual goals and constraints (such as lack of resources, staff or IT skills) how to practically adopt the Collections as Data principles.

- Data Foundry at the National Library of Scotland provides datasets in the form of downloadable files. Each of them includes a website with transparent information about the creation and curation of the content. It also provides examples of use and prototypes in the form of Jupyter Notebooks.
- National Library of Luxembourg provides a digital collection based on Historical newspapers. It follows an innovative approach by providing different datasets in terms of the size and the content according to the purpose of the reuse (for example "getting started" or "Big Data").
- British Library Labs provides a selection of datasets made available under open licences to experiment.
- DATA-KBR-BE is a project that aims at facilitating data-level access to the KBR (Royal Library of Belgium) digitised and born-digital collections for digital humanities research.

- Biblioteca Virtual Miguel de Cervantes (BVMC Labs) is a digital library that published its bibliographic catalogue in the form of Linked Open Data. It also provides examples of use by means of Jupyter Notebooks.

More advanced approaches are focused on the concept of data spaces, which are new cloud environments in which data can be shared for research use, while also allowing data suppliers to retain rights and control over the data. This can be useful for contexts in which the access of the data is restricted (for examples, geographic region, licensed content). A workflow to publish Collections as Data: the case of Cultural Heritage data spaces gives more information on this topic in the context of the European common data space for cultural heritage.

It is important to note that institutions need to carefully consider how and for what purpose they want to publish their digital content. For instance, the National Library of the Netherlands has recently restricted access to collections for training commercial AI. See our *DS Topic Guide* on Copyright and Licensing for more on that.

Hands-on activity/self-guided tutorial(s)

One simple activity you could do at your institution is to gather a group of like-minded colleagues together and have a read through Fifty things, Vancouver Statement on Collections as data and/or A Checklist to Publish Collections as Data in GLAM Institutions and have a look at some of the case studies above, and as a group consider the questions:

- Are there some activities suggested that are already underway in your institution?
- Can you identify one or two simple things that your institution could do now in order to start using these principles?
- What would you like "collections as data" to look like at your institution?

If you'd like to get a sense of how people are reusing collections published as data, the GLAM Workbench and the new computational access section of the International GLAM Labs Community website both have tutorials that can walk you through using cultural heritage datasets in a variety of ways.

Recommended Reading/Viewing

If you are interested in reading more about the collections as data concept we highly recommend the Zotero | Groups > collections as data - projects, initiatives, readings, tools, datasets which is an open bibliography collaboratively maintained by the community of practice of projects, readings, initiatives, tools, and datasets that are in some way or another related to collections as data.

Finding Communities of Practice

Joining the International GLAM Labs Community is a great way to get started on the path of opening your collections up as data as well as the Collections as Data - Google Group as both communities are very active and have a wide range of experience and expertise to share on this topic. LIBER's Working Groups address important areas of work for the research library community) address important areas of work for the research library community such as Data Science and Research Data Management.Some international organisations also have special interest groups such as the Research Data Alliance (RDA) which hosts the Collections as Data IG.