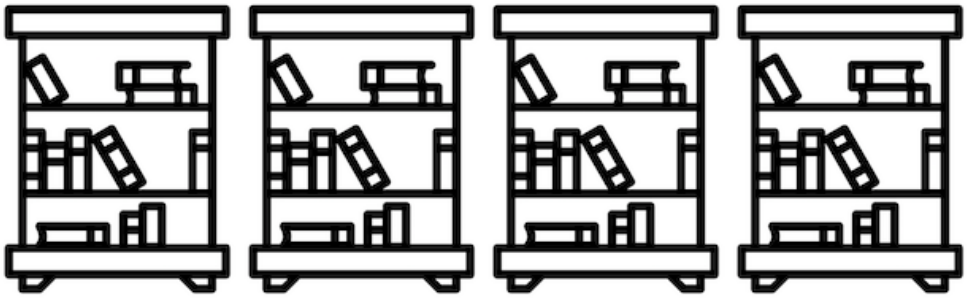
Always Already Computational: Collections as Data



Thomas Padilla (PI)
Laurie Allen (Co-PI)
Hannah Frost (Co-I)
Sarah Potvin (Co-I)
Elizabeth Russey Roke (Co-I)
Stewart Varner (Co-I)

The Santa Barbara Statement on Collections as Data

May 2018

This publication is part of the Collections as Data Framework hosted at https://osf.io/mx6uk/.

This project was made possible by the Institute of Museum and Library Services (LG-73-16-0096-16). The views, findings, conclusions, or recommendations expressed in this publication do not necessarily represent those of the Institute of Museum and Library Services or author host institutions.

The Santa Barbara Statement on Collections as Data

May 2018

The Santa Barbara Statement on Collections as Data was written by the Institute of Museum and Library Services supported Always Already Computational: Collections as Data project team. The first version was based on the collaborative work of participants at the first Collections as Data National Forum (UC Santa Barbara, March 1-3 2017). After its release, the team gathered comments from the Hypothesis web annotation tool and sought additional feedback across a series of conversations and workshops (April 2017 - April 2018). The current version of the statement was revised based on that community feedback, especially the close, directed feedback provided by workshop participants at the Digital Library Federation Forum 2017.

What are "collections as data"? Who are they for? Why are they needed? What values guide their development? The Santa Barbara Statement on Collections as Data poses these questions and suggests a set of principles for thinking through them, as part of a community effort to empower cultural heritage institutions to think of collections as data and consequently to explore what might be possible if cultural heritage seen in this light was more readily open to computation.

The concept of collections as data emerges at – and is grounded by – a particular moment in the recent history of cultural heritage institutions. For decades, cultural heritage institutions have been building digital collections. Simultaneously, researchers have drawn upon computational means to ask questions and look for patterns. This work goes under a wide variety of names including but not limited to text mining, data visualization, mapping, image analysis, audio analysis, and network analysis. With notable exceptions like the Hathitrust Research Center, the National Library of the Netherlands Data Services & APIs, the Library of Congress' Chronicling America, and the British Library, cultural heritage institutions have rarely built digital collections or designed access with the aim to support computational use. Thinking about collections as data signals an intention to change that, and efforts like the Library of Congress' Collections as Data: Stewardship and Use Models to Enhance Access and the multinational Digging into Data suggest that a broader community shift intentionally scoped to institutions large and small comes at an opportune time.

While the specifics of how to develop and provide access to collections as data will vary, any digital material can potentially be made available as data that are amenable to computational use. Use and reuse is encouraged by openly licensed data in non-proprietary formats made accessible via a range of access mechanisms that are designed to meet specific community needs.

Ethical concerns are integral to collections as data. Collections as data should make a commitment to openness. At the same time, care must be taken to comply with legal requirements, cultural norms, and the values of vulnerable groups. The scale of some collections may also obfuscate what is hidden

or missing in the histories they are perceived to represent. Cultural heritage institutions must be mindful of these absences and plan to work against their repetition. Documentation should be informed by archival principles and emergent reproducibility practice to ensure that users have the information they need to work with collections responsibly.

Principles

1. **Collections as data development aims to encourage computational use of digitized and born digital collections.** By conceiving of, packaging, and making collections available as data, cultural heritage institutions work to expand the set of possible opportunities for engaging with collections.

2. **Collections as data stewards are guided by ongoing ethical commitments.** These commitments work against historic and contemporary inequities represented in collection scope, description, access, and use. Commitments should be formally documented and made publicly available. Commitment details will vary across communities served by collections but will share common cause in seeking to address the needs of the vulnerable. Collection stewards aim to respect the rights and needs of the communities who create content that constitute collections, those who are represented in collections, as well as the communities that use them.

3. **Collections as data stewards aim to lower barriers to use.** A range of accessible instructional materials and documentation should be developed to support collections as data use. These materials should be scoped to varying levels of technical expertise. Materials should also be scoped to a range of disciplinary, professional, creative, artistic, and educational contexts. Furthermore the community should be motivated and encouraged to build and share tools and infrastructure to facilitate use of collections as data.

4. **Collections as data designed for everyone serve no one.** Specific needs inform collections as data development. These needs may be commonly held by particular user communities. Rather than assuming these needs or imagining these communities, stewards should be intentional about who their collections are designed for, work to lower the barriers to use for the people in those communities, and continue to assess these needs over time. Where resources permit, multiple approaches to data development and access are encouraged.

5. **Shared documentation helps others find a path to doing the work.** For example, collections as data work can entail decisions about selection, description, conversion cleaning, formatting, and delivery mechanisms or platforms that enable discovery and provide access. In order for a range of individuals and institutions to engage collections as data work, it must be possible to locate documentation that demonstrates how and why the work is done. Documentation must also attest to the history of how the collection has been treated over time. While no documentation can be fully comprehensive, incomplete or in-progress

documentation is better than no documentation. Examples of documentation include human and machine readable metadata schemas, data sheets, workflows, application profiles, deeds of gift, and codebooks. Documentation should be publicly accessible by default.

6. **Collections as Data should be made openly accessible by default, except in cases where ethical or legal obligations preclude it**. Terms of use for collections as data must be made explicit and should align with community-based practices such as RightsStatements.org and standard licenses such as Creative Commons, Open Data Commons, and Traditional Knowledge licenses.

7. **Collections as data development values interoperability.** Interoperability entails alignment with emerging and/or established community standards and infrastructure and eases integration with centralized as well as distributed infrastructure. This approach facilitates collections as data discovery, access, use and preservation.

8. **Collections as data stewards work transparently in order to develop trustworthy, long-lived collections.** Trustworthiness depends upon efforts to ensure and publicly document the technical integrity of the data as well as its provenance. It also requires that data stewards acknowledge absences and areas of uncertainty within the collection as data. Trustworthy collections as data should include open, robust metadata, and should be under the care of stewards and institutions committed to their preservation.

9. **Data as well as the data that describe those data are considered in scope.** For example, images and the metadata, finding aids, and/or catalogs that describe them are equally in scope. Data resulting from the analysis of those data are also in scope.

10. **The development of collections as data is an ongoing process and does not necessarily conclude with a final version.** Work in progress status can be seen as a virtue when iteration is geared toward developing productive collaborations and integrations between new and existing technologies, workflows, and service models. The ongoing development of collections as data can impact staffing models, workflows, and technical infrastructure.