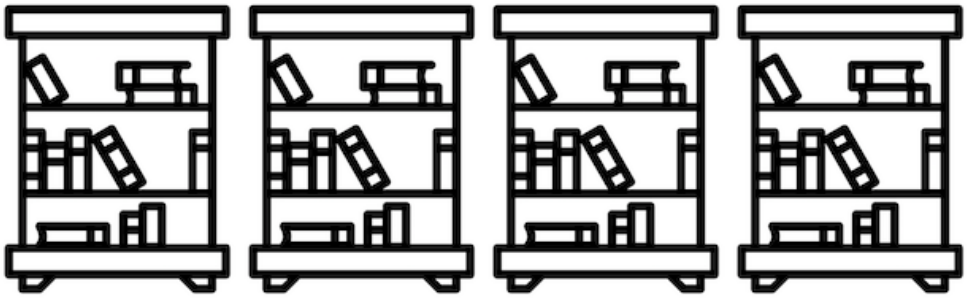
Always Already Computational: Collections as Data



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

---

50 Things

October 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.

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.

Our best advice: start simple and engage others in the process. You may find some activities listed here are already underway!

1. Know how optical character recognition (OCR) output is produced in your digitization workflows. What software is used? What formats are created? What levels of accuracy are produced? Where is it stored? Is it available for user download?

2. Create an inventory of full-text collections managed by your institution. Document rights status, license status, discoverability, and downloadability. Ask the question: are we offering optimal access for computational use of the full-text? How can we make it better?

3. Migrating a legacy digital collection to a new system or platform? Take the opportunity to make the content accessible to researchers that have computational projects in mind.

4. Interview the archivist, librarian, or curator responsible for a digital collection to document data provenance and decisions made in the course of collection processing and digitization. Work to make this information publicly available.

5. Inventory your data holdings. Just make a simple list. And then commit to keeping it up to date, and watch it grow.

6. Add new fields to the collection management database to indicate and describe data components.

7. Survey your digital collections to identify characteristics -- good metadata, open access, good OCR, high usage, relevance to a high-profile academic program or research area at the institution -- which lend themselves to high impact as data.

8. Recognize and identify the things you need to do differently than have been done for physical collection objects.

9. Find out if your digital collection database or access platform has an API available for querying by the public. If it does not, see if it is possible to develop one. If it does, determine if it is actively used. If it is actively used, see if you can reach out to users and ask about their usage!

10. Talk to a colleague responsible for systems that provide networked access to digital collections about possible approaches to facilitate download of collection data in bulk.

11. Add a terms of use to your archival finding aids.

12. Read the language of your organization's collection deed of gift or purchase agreement to evaluate whether it allows for providing access to collection content in the form of data.

13. Review your digital collections metadata and evaluate the rights statements and license statements in terms of consistency and clarity. Are you able to adopt rightsstatements.org?

14. Socialize Collections as Data as something that can be supported by units and staff across the library. Identify some champions across the organization and people who have skills or position to do the work.

15. Talk to people responsible for research data management to encourage planning for data preservation and other considerations that make it possible for others to reuse the data in the future.

16. Review your institution's mission statement or strategic plan documentation, and consider if and how Collections as Data activities are aligned with and support it.

17. Share sample projects with community partners to give them an idea of how their collections can be used and be relevant to new ways of conducting scholarship.

18. Network with people who work with data and have the skills or knowledge you need to get your work done.

19. Identify barriers and limitations to what services you can offer support, and talk with colleagues about creative, feasible solutions to overcome them.

20. Publish or present on "Wikidata for librarians," including case studies of libraries working with Wikidata to expand discovery of collections.

21. Read up on IIIF (for example, check out this useful tutorial) and determine what hurdles to implementation exist at your institution. Then talk to relevant folks about what it would take to overcome them.

22. Read the resources in the Always Already Computational project's Zotero library.

23. Develop a workshop focused on the use of data in abd about collections; shop it to department faculty and incorporate it into research orientations for faculty and students.

24. Mentor a liaison interested in learning a data science skill who is well positioned to identify datasets and data support needs amongst their researchers.

25. Conduct user testing of your library's main discovery environment, with the goal of understanding how easy or hard it is for a researcher to find the available data collections.

26. Develop a portal page with a site map specifically for discovering collections at your institution available for computational use and related support services.

27. Begin tracking demand for and use of data in and about your collections.

28. For a collection that cannot be made available openly on the web, investigate if your organization is able to support mediated access to the data, such as through an offline or encrypted workstation.

29. Prepare and provide datasets that are intentionally useful, in terms of size and complexity, for teaching in semester- or quarter-long classes.

30. For classes that draw directly on library collections and generate data, ask the students to submit their data products back to library, through the institutional repository. Normalize the process of giving back and augmenting the collections with data. This may work particularly well for collections that are institutionally or regionally focused.

31. Identify a faculty member who does computational analysis for their own research and find a way to transfer or replicate the tools and approaches they use to apply them to a library collections-as-data use case.

32. If you offer an API to your repository, evaluate the public-facing documentation to see if it is clear, current, accurate, and discoverable by researchers.

33. Publish documentation about how to find, use, and interpret collections as data in multiple places including blogs, README files, and LibGuides.

34. A dataset should always be accompanied by a README plain text file that documents basic, important information about the data. Make READMEs part of your data documentation practice. Develop one or more template to that can be used by librarians and researchers.

35. Make an effort to make existing OCR output generated from past scanned text collections projects more available for computational analysis, such as through bulk download.

36. When planning your next digitization project, incorporate additional steps for preparing content files, OCR or transcription text, and metadata for bulk access. Document the key issues and

decision points you encounter as you evolve and expand your digitization workflows.

37. Talk to colleagues involved in taking in deposits to your institutional repository or research data repository about a process for encouraging and accepting contributions back from users of data in your collections.

38. Gain the support of administration by following and supporting the work of third-party research groups like OCLC that help bolster and highlight the trends in the development of collections as data.

39. Provide a resource that shows a data user how to cite a dataset, and that shows a data creator how to format a preferred citation for an original dataset and a derivative dataset.

40. Ask a subject specialist at your institution if faculty or students are requesting data about or derived from library collections.

41. Take a public services librarian, curator, or archivist out for coffee to talk about collections as data. Ask what they are hearing from faculty, students, and other users of collections about computational use and which collections have potential for taking action to lower barriers to computational use.

42. Investigate how your library is collecting, managing, and making email archives accessible. Consider whether a collections as data approach will serve your institution's goals.

43. Start small. Start with a research question, and choose projects that have promise to be generalizable for use by future scholars such that the investment is worth the level of commitment. No one-offs!

44. Start with a prototype or proof of concept. It's fine if your Collections as Data project does not integrate with institutional repository or formalized infrastructure.

45. Collaborate with subject specialists or instruction librarians to ask scholars about interest in computational data in and about collections. Compile their ideas to make a case, and build a team for the next opportunity to pursue one of them.

46. Be thoughtful and strategic about allocating scarce resources to collection digitization projects. Consider prioritizing projects that produce outcomes that are reusable (derivative datasets) and repeatable (processes, tools, workflows) that can benefit your department and your users again and again.

47. Explore what it would take for your organization to contribute subject data to Wikidata, drawing on a local collection and then incorporating the Wikidata links into your local discovery environment.

48. Test how data gathered in a crowdsourcing project can be associated with the existing source object data and can also serve as stand-alone dataset.

49. Use your favorite search engine to find information about APIs provided by museums and read about the various ways that data about museum collections can be analyzed to discover new insights.

50. Keep tabs on the projects emerging in the Collections as Data: Part to Whole project, funded by the Mellon Foundation. They are bound to point a way forward for us all!