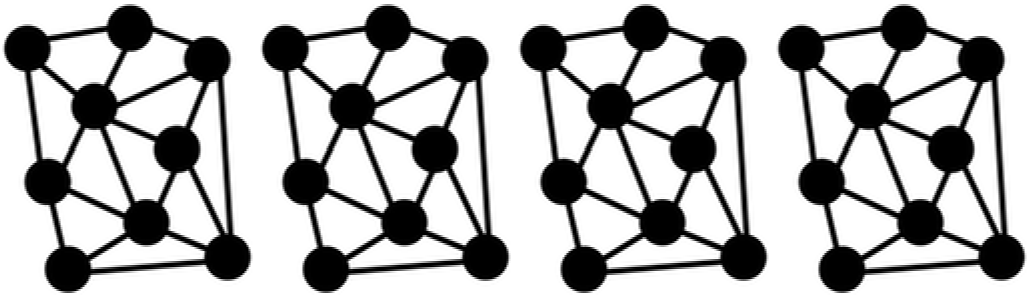
Collections as Data: Part to Whole

Final Report



Thomas Padilla
Hannah Scates Kettler
Yasmeen Shorish

—

This project was made possible by the Mellon Foundation.

Introduction

As a concept, community, and area of practice, collections-as-data has many origin stories. The diffuse nature of origin presents a benefit to the present insofar as the collections-as-data community is flexible enough to support convergent effort across roles, organizations, and countries that advance responsible development and computational use of memory organization collections.[1] Ongoing evolution of the work is abundant, considering the level of changes to memory organization staffing, workflow, infrastructure, strategy, as well as a proliferation of multinational community formation. In many respects, collections-as-data is a kind of "boundary object" - a concept first advanced by Susan Leigh Starr in her work with James Griesemer, *'Translations' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology*:

> *Boundary objects are objects which are both plastic enough to adapt to local needs and constraints of the several parties employing them, yet robust enough to maintain a common identity across sites. They are weakly structured in common use, and become strongly structured in individual-site use. They may be abstract or concrete. They have different meanings in different social worlds but their structure is common enough to more than one world to make them recognizable, a means of translation. The creation and management of boundary objects is key in developing and maintaining coherence across intersecting social worlds.*[2]

In an effort to maintain, "coherence across intersecting social worlds", a series of efforts have gathered and distilled high level principles that guide collections-as-data work in abstract and concrete terms - the *Santa Barbara Statement on Collections as Data (2017)* and *Vancouver Statement on Collections-as-Data (2023)*. Each statement of principles was developed in cycles of synchronous and asynchronous contributions from an intentionally diverse set of professional and disciplinary communities. Key to both statements is the advance of a fundamental understanding that not all collections have or should have an inevitable expression as data. Some collections should not be made openly accessible because open access would pose harms to communities represented in the collections (e.g., social media data collections documenting social protest under authoritarian regimes) or because it is culturally inappropriate to have those collections be open to all. Other collections-as-data development efforts must critically contend with a history of knowledge

[1] Memory organization refers to libraries, museums, archives, and/or other organizations that document and preserve knowledge production (e.g., history, science, art, government). Sometimes referred to as "cultural heritage institutions"

[2] Star, Susan Leigh, and James R. Griesemer. "Institutional Ecology, `Translations' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39." *Social Studies of Science* 19, no. 3 (August 1, 1989): 387–420. https://doi.org/10.1177/030631289019003001.

extraction exacted by libraries, archives, and museums on minoritized communities. The *CARE Principles*[3], advanced by the Global Indigenous Data Alliance, provide a place to begin the conversation with communities represented in collections of that kind. Work on reparative archives by Lae'l Hughes Watkins suggests a critical path for memory organizations invested in collections-as-data work to, " … repair their holdings and develop a holistic approach to disrupting homogeneous histories through acquisition, advocacy, and utilization of collections and challenging …" predominant representation of histories.[4] Rather than perceiving these factors negatively, collections-as-data practitioners are motivated by them insofar as they present an opportunity to repair the past and strengthen the future.

Weakly structured and strongly structured, abstract and concrete, locally situated and broadly framed, collections-as-data is a boundary object that works to encourage diverse forms of collaboration within and across communities that support responsible development and computational use of memory organization collections. In what follows, we, the authors, share lessons learned from our work on the Mellon Foundation supported *Collections as Data: Part to Whole*, with a focus on opportunities for growth of the work at local and global scale.

Cohorts and Models

Collections as Data: Part to Whole has its roots in a preceding effort called *Always Already Computational: Collections as Data*.[5] [6] Began in 2017, *Always Already Computational* helped foster diverse community awareness with respect to the challenges and opportunities of responsibly developing and providing access to collections-as-data. Concurrent to *Always Already Computational* activities an increasingly international field produced national collections-as-data strategies, state-based collections-as-data strategies, numerous library-wide strategic plans that prioritized collections-as-data work, collections-as-data conferences and conference presentations, journal articles, and a number of job postings focused on collections-as-data work. As *Always Already Computational* came to a close, community

[3] Carroll, S.R., Garba, I., Figueroa-Rodríguez, O.L., Holbrook, J., Lovett, R., Materechera, S., Parsons, M., Raseroka, K., Rodriguez-Lonebear, D., Rowe, R., Sara, R., Walker, J.D., Anderson, J. and Hudson, M., 2020. The CARE Principles for Indigenous Data Governance. *Data Science Journal*, 19(1), p.43.DOI: https://doi.org/10.5334/dsj-2020-043

[4] Hughes-Watkins, Lae'l. "Moving Toward a Reparative Archive: A Roadmap for a Holistic Approach to Disrupting Homogenous Histories in Academic Repositories and Creating Inclusive Spaces for Marginalized Voices." *Journal of Contemporary Archival Studies* 5, no. 1 (May 16, 2018). https://elischolar.library.yale.edu/jcas/vol5/iss1/6.

[5] Always Already Computational: Collections as Data. https://collectionsasdata.github.io/.

[6] Stewart Varner is an additional Co-Investigator on *Collections as Data: Part to Whole*. He has been fundamental to the effort overall, contributing to project design, cohort guidance, and project promotion. He was also Co-Investigator on *Always Already Computational: Collections as Data.*

feedback helped determine the objectives of a succeeding effort designed to support further collections-as-data community growth. The Mellon Foundation supported *Collections as Data: Part to Whole* aimed to explore 3 primary questions:

- What organizational models support sustainable collections-as-data development and access?
- What organizational models support sustainable collections-as-data use by researchers, artists, fellow memory organization practitioners, and more?
- How can organizations create collections as data in a manner that demonstrates commitment to developing and implementing processes for addressing complex ethical issues inherent to engagement with cultural heritage data, and the needs of marginalized and underrepresented communities?

Part to Whole supported the development of answers to these questions through subgrants to large and small organizations throughout the United States, paired with a cohort development program.[7] The product of subgrantee experimentation was intended to be helpful to institutions of varying size, budget, communities served, and mission. After circulating a call for proposals, applications were evaluated based on the following criteria:

1. Use model demonstrates:
    - Innovative, cross-division and/or cross-departmental formation of positions and services that support the use of collections as data
    - Local sustainability
    - Ready potential for adaptability by other institutions
2. Implementation model demonstrates:
    - Innovative, cross-division and/or cross-departmental formation of positions and services needed to implement collections as data
    - Local sustainability
    - Ready potential for adaptability by other institutions
3. Proposed collections-as-data evidence:
    - significant research value
    - the perspectives of underrepresented and/or oppressed groups
4. The proposed project demonstrates commitment to developing and implementing processes for addressing complex ethical issues inherent to engagement with cultural heritage data, and the needs of marginalized and underrepresented communities.

[7] Laurie Allen was a part of the initial project team and was involved in the CFP and first cohort of subgrantees. Due to a position change, she stepped away from the project team. In 2021, during the preparation for the second cohort, Yasmeen Shorish joined the team for the remainder of the project.

5. The proposed project evidences knowledge of complementary collections, standards, and initiatives in the library field and scholarly disciplines that speak to project goals.
6. The proposed project utilizes open source technologies that aim for interoperability (where appropriate) with a broader open scholarly communication infrastructure.

Part to Whole required teams to include one senior administrator who could allocate resources as necessary, a disciplinary scholar who could provide insight into how collections could be used for research and education purposes, and one project lead who would coordinate the work.

Over the course of *Part to Whole*, two cohorts totalling twelve teams were selected:

Cohort 1:

- "On the Books: Jim Crow and Algorithms of Resistance"
  University of North Carolina Chapel Hill
  María R. Estorino, Amanda Henley, Matt Jansen, Lorin Bruckner, Sarah Carrier, William Sturkey.

- "Uncovering Health History: Transcribing and Publishing Early Twentieth-Century Tuberculosis Patient Records as Data"
  University of Denver
  Kim Pham, Kevin Clair, Jack Maness, Jeanne Abrams, Fernando Reyes, Jeff Rynhart, Alice Tarrant.

- "Collections as Data: Redefining Creators, Users, and Stewards of the Charles “Teenie” Harris Photographic Archival Collection"
  Carnegie Museum of Art
  Dominique Luster, Charlene Foggie-Barnett, Ed Motznik, Samantha Ticknor.

- "The Native American Educational Services College Digital Library Project"
  Northwestern University
  Josh Honn, Kelly Wisecup, John Dorr, Dorene Wiese, Melanie Cloud, Allison Conner.

- "From Collection Records to Data Layers: A Critical Experiment in Collaborative Practice"
  University of Pittsburgh
  Tyrica Terry Kapral, Aaron Brenner, Matthew J. Lavin, Gesina Philips.

Cohort 2:

- "LGBTQ+ Audio Archive Mining Project"
  University of Wisconsin-Milwaukee
  Ann Hanlon, co-Project Lead, Daniel Siercks, Cary Costello, Marcy Bidney, Shiraz Bhathena, Jie Chen, Karl Holten, Ling Meng, Constance Dewitt, Syeda Ashrafi

- "Images as Data: Processing, Exploration and Discovery at Scale"
  Harvard University
  Carol Chiodo, Taylor Arnold, Lauren Tilton, Ardys Kozbial

- "...And 25 of our closest friends: The Louisiana Digital Library as Community-Focused Data"
  LSU Libraries
  Sophia Ziegler, Gina Costello, Leah Powell, Elizabeth Joan Kelly

- "Surfacing hidden water data: Water, people, displacement in Southern California"
  The Claremont Colleges Library
  Jeanine Finn, Jessica Dávila, Char Miller

- "dLOC as Data: A Thematic Approach to Caribbean Newspapers"
  Digital Library of the Caribbean at Florida International University
  Miguel Asencio, Jamie Rogers, Perry Collins, Hadassah St. Hubert

- "Using Newspapers as Data for Collaborative Pedagogy: A Multidisciplinary Interrogation of the Borderlands in Undergraduate Classrooms"
  University of Arizona Libraries
  Mary Feeney, Anita Huizar-Hernández, Sarah Shreeves, Erika Castaño, Celeste González de Bustamante, Marya McQuirter, Katherine Morrissey, Jeff Oliver, Cristina Ramírez, Verónica Reyes-Escudero, Megan Senseney

The COVID-19 pandemic was a major disruption to cohort teams, causing changes in staffing, resources, and organizational priority that necessitated multiple no-cost extensions to their grants. Despite these challenges all 12 cohort teams completed their projects. Cohort 1 and 2 project deliverables are available in the *Collections as Data OSF Community.*[8]

Cohort Themes

[8] Collections as Data: Part to Whole, https://osf.io/r9n3s/

The following key themes arose from cohort experimentation:

No collections-as-data about us without us[9]

Communities represented in collection-as-data must be engaged (e.g., advisory boards, collections-as-data team composition, calls for input) in the process of collections-as-data creation and use. Memory organizations have erred in the past when they failed to seek input from the communities represented in their collections. Prioritizing community involvement will not only show respect and help to build trust but it will also result in more generative research and education possibilities.

Collections-as-data work should be supported by organizational structure

Collections-as-data work has the potential to impact nearly all areas of an organization. This is especially the case as contemporary knowledge production is increasingly accomplished via digital means. The most effective collections-as-data initiatives incorporate expertise and input from information technology, technical services, and digital collections, as well as insight from subject specialists and substantive support from senior administration.

Good documentation is crucial

Collection-as-data work involves a broad range of stakeholders during creation, maintenance, and use. It is imperative that decisions and rationales are well-documented. It may be that future collections-as-data stewards will need to perform major updates and the more they know about why a collection was built the way it was, the more likely they will be successful in maintaining intent and purpose of the effort. Good documentation is crucial for collections-as-data users as well, insofar as documentation speaks to questions of motivation, scope, and completeness - all essential factors for assessing possible use of collections as data.

Community of practice and skill sharing are essential

Collection-as-data work will continue to change as technologies and researcher needs evolve. Recognizing a state of constant change, both cohorts expressed a desire to establish a repository for documentation as well as a mechanism to support an ongoing community of practice for collections-as-data. The COVID-19 pandemic induced isolation that spanned much of *Part to Whole* cohort activity underscored the need for community.

[9] Pfeifer, Whitney. "From 'Nothing About Us Without Us' to 'Nothing Without Us.'" Text, March 28, 2022. https://www.ndi.org/our-stories/nothing-about-us-without-us-nothing-without-us.

Vancouver Summit

As *Part to Whole* came to a close the project team considered how to best build upon cohort experiences and connect with concurrent, internationally distributed collections-as-data work. It was agreed that convening an international working summit would be the most generative approach. *Collections as Data: State of the Field and Future Directions*, was held April 24-25, 2023 in Vancouver, British Columbia at Internet Archive Canada. *Part to Whole* encouraged participation through direct invitations to stakeholders and a broad call for participation (CFP). Given international proliferation of the work since *Always Already Computational,* it was important to have representation from many regions of the world and a range of organizations and groups. The CFP was promoted extensively to help ensure that diverse world views and experience were represented at the event.[10]

Attendees (see Appendix A) from 60 organizations gathered for two days in Vancouver, Canada. Approximately two-thirds of the attendees were invited, leveraging professional networks, and the remainder were selected via the CFP. The project team hoped that the Canadian location would be more welcoming to international participants, which appeared to be partially true as several participants remarked on the choice. Unfortunately, the event did coincide with the strike action of Canadian Federal workers responsible for visa and passport services, which resulted in visa complications for at least one participant causing them to cancel their travel plans. The event was not conducted in a hybrid format because summit objectives relied on sustained participant interaction and dedicated reflection and writing time.

Part to Whole adopted an instructional design approach to the working event. The project team identified objectives for every section of each day and considered how best to scaffold the events to achieve those objectives. The project team spent time considering the range of participant expertise and background knowledge and designed seating arrangements to maximize sharing and connection. The first day opened with presentations by the project team followed by activities to allow individuals to highlight their positionality in collections-as-data work before they embarked on the discussion of cohort outcomes. The following topics were used as prompts: phases of maturity, cohort model transferability, and cultural context reflection. The cultural context reflections were completed individually, while the other activities were done in groups. The second day was more activity-driven. Tables worked

[10] Myrna E. Morales, Stacie Williams, 2021. "Moving toward Transformative Librarianship: Naming and Identifying Epistemic Supremacy", Knowledge Justice: Disrupting Library and Information Studies through Critical Race Theory, Sofia Y. Leung, Jorge R. López-McKnight https://doi.org/10.7551/mitpress/11969.001.0001

together on a speculative futures worksheet and potential revisions to the *Santa Barbara Statement*, while individuals created action plans and completed a final reflection.

The project team worked to create an inclusive and comfortable experience for event attendees. Prior to the event, the project team sent a participant survey asking for preferred name usage, pronouns, dietary concerns, and mobility or audio/visual considerations. Based on the responses, the project team made arrangements to ensure that the space and menu was responsive to survey results. The project team worked to limit their use of colloquialisms and jargon, while acknowledging that presentations would only be in English.

Observations

Model Assessment

A goal of *Part to Whole* was to establish collections-as-data implementation and use models viable in a range of institutional contexts. The summit provided an opportunity for a diverse group of administrators, practitioners, and researchers to assess these models. Each table read a cohort team report and assessed model transferability. In general, participants felt that many of the projects contained elements of a model but were more akin to proof-of-concepts rather than readily transferable models. Teams that were the most successful in integrating their projects into established workflows or position scope resonated more as models than those that did not. This feedback has been helpful, as it helps inform whether a “model” approach is the best approach for advancing collections-as-data work. Given the thoughtful insights shared by summit participants, it may be that sharing of practices and consideration of principles is a more effective way to demonstrate and sustain collections-as-data effort.

International Scope

The Vancouver Summit was designed to gather as representative a group of perspectives on collections-as-data as possible. Each attendee was asked to describe what they saw as challenges and opportunities in an internationalized collections-as-data field.

Responses coalesced around the following themes:

> **Socio-cultural:**
> Challenges - political tensions and trust across nations/contexts; multinational rights/licensing differences, and gatekeeping (particularly within higher education); cultural biases, siloes and hierarchies and making the cultural context of collections and locales visible; threats to autonomy/sovereignty, and a concern that focusing on internationalization could minimize important local, smaller efforts; differences in AI and data regulation.

Opportunities - greater adoption of FAIR and CARE principles, and building a community of practice and distributed governance; internationalization as a means to reduce ethnocentrism; building shared understanding, repatriation, expanding ways to knowing.

Technical:
Challenges - variable digitization practices across countries/institutions.

Opportunities - increase shared infrastructure; greater use of datasheets for collections as data, better OCR, and linking projects/collections across locales.

Practical:
Challenges - differences in collections-as-data vocabulary where the same word can mean different things in different languages; the dominance of English for tools like OCR and AI; uneven data fluency and gaps in teaching materials; open data availability.

Opportunities - increase shared practices; develop more multilingual resources.

Organizational:
Challenges - organizational capacity.

Opportunities - increase shared infrastructure (which could have positive environmental impact); creation of solidarity among traditionally marginalized groups and the confrontation of colonial heritage.

Resourcing:
Challenges - financial limitations and pressure to defer to richer locales/institutions; travel (both costs and logistics).

Opportunities - increase shared resources; create regional collections-as-data hubs.

Collaboration:
Challenges - institutional risk aversion limiting new collaborations.

Opportunities - model collections-as-data sustainability with each successful international collaboration; tackling global challenges, increase visibility of the work.

Some of these challenges and opportunities are not unique to collections-as-data, but rather to international collaborative efforts writ large. The points raised related to tool and infrastructure development, rights and licensing, open data, implementing CARE and FAIR principles, contending with the legacy of colonization and ongoing heritage work, data fluency, and connecting projects and objects across locales warrants attention from a global community. Addressing these challenges and opportunities provides motivation to redress long standing barriers to international collaboration.

Outcomes

Vancouver Statement on Collections-as-Data

In order to guide future collections-as-data community work, summit participants expressed a need for an update to the *Santa Barbara Statement on Collections as Data*. For background, the *Santa Barbara Statement on Collections as Data* was the result of the first *Always Already Computational: Collections as Data* National Forum, held March 1-3 2017 at the University of California, Santa Barbara. That forum was attended by 28 information professionals and disciplinary scholars from a wide variety of institutions. The Santa Barbara Statement became a critical resource to help ground approaches to doing collections-as-data work.[11] It has been downloaded more than 700 times and has been referenced frequently across scholarly and professional literature and in working institutional documentation.[12] [13] [14]

Drawing on perspectives shared by participant exercises at the Vancouver Summit, the *Part to Whole* project team drafted the *Vancouver Statement on Collections-as-Data*. The drafting process was further informed by post-summit participant feedback as well as a period of open community feedback to create space for voices that were not present at the summit. The final statement provides an updated emphasis on ethics, sustainable labor, community relations, non-open data, climate impact, artificial intelligence, and organizational structure. The Vancouver Statement raises topics that should support ongoing maturation of collections-as-data community work.

[11] A full list of attendees to the first forum can be found here: https://collectionsasdata.github.io/partners/

[12] Wittmann, R., Neatrour, A., Cummings, R., & Myntti, J. (2019). From Digital Library to Open Datasets: Embracing a "Collections as Data" Framework. Information Technology and Libraries, 38(4), 49–61. https://doi.org/10.6017/ital.v38i4.11101

[13] Mirza, Rafia. “Collections as Data; ML Literacies in Libraries”. ASIST. July 17, 2022. https://www.asist.org/2022/07/17/collections-as-data-ml-literacies-in-libraries/.

[14]Klerk, Taylor de. “Ethics in Archives: Decisions in Digital Archiving.” https://www.lib.ncsu.edu/news/special-collections/ethics-in-archives:-decisions-in-digital-archiving.

Position Statements

All participants provided a position statement in advance of the summit that responded to the following prompt:

> *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: Part to Whole (see grant narrative and sub grantee final reports) … We strongly welcome bridging, divergence, and provocation in your position statements. Is there something concrete or conceptual we are missing? Are there questions and communities we aren't currently considering? This is an opportunity to highlight aspects of your experience that relate to collections as data and will help stage interaction at the face-to-face meeting as we collectively consider the state of the field and future directions together. As a whole, these position statements will form a collective resource to be shared publicly with any community interested in collections as data.*

At the time of this white paper writing, the position statements have been downloaded nearly 800 times.[15] Position statements can be found in Related Resources.

Summit participants were asked to identify shared challenges expressed in their position papers. Synthesis of that activity produced the following shared challenges:

- Transitioning collections-as-data work to "business as usual," rather than a special project
- Need for robust community connections (users and creators); "Nothing about us, without us" ethos for community-involved collections
- Further exploring the impact of AI (artificial intelligence) and LLM (large language model) on collections-as-data
- Navigating ethics, sovereignty, and contextualization
- Addressing climate impact of computational work
- Supporting linguistic diversity in collections-as-data work
- Fostering lower technical barriers to engage with collections-as-data work

[15] "Position Statements -> Collections as Data: State of the Field and Future Directions." https://doi.org/10.5281/zenodo.7897735.

Moving Forward

Sustain Potential, Mitigate Harm

Collections as Data: Part to Whole and the effort that preceded it *Always Already Computational: Collections as Data* have consistently focused on encouraging responsible computational use of memory organization collections. That focus has had a critical orientation from the beginning, informed by the work of scholars and practitioners like Gabrielle Foreman, Safiya Noble, Timnit Gebru, and innumerable others. Focus on responsibility in the work has been seen as an opportunity to sustain the abundantly diverse potential of computational work (e.g., research, education, creativity) with memory organization collections while mitigating the potential for harm (e.g., amplification of bias, avoiding service quid pro quos that conflict with mission alignment). Practitioners in the collections-as-data community have resisted the emphasis on speed so often prioritized in "computational" areas of engagement in favor of deliberative, historically conscious pace.

As the field matures, practitioners continue to engage with critical issues, some new and some enduring. For example, how should varied languages, norms, and tellings of history come together to inform collections-as-data work? How can practitioners best deal with memory organization histories of knowledge extraction? What opportunities and challenges does artificial intelligence present to memory organization work? In order to best answer these questions, collections-as-data practitioners need to learn about and from each other. This need may be best addressed through the formation and maintenance of community venues for sharing experiences, good documentation, humane strategies for managing fluctuation in staff resourcing, and negotiating organizational priorities in times of significant change - the last point being especially pressing as the world emerges from a global pandemic into financial recession and politically-fraught uncertainties.

Resurgent nationalist movements would have memory organization practitioners diminish their role and impact, set to pursue resolution of challenges and realization of opportunities within the most unimaginative sense of community. Nationalism artificially circumscribes the creativity memory organizations need in order to fulfill commitments to user communities. Issues of data sovereignty, community autonomy, extraction, commodification, colonial gaze[16], linguistic hegemony, and epistemological justice are held in common across national borders and must be contended with on a transnational level. Throughout the world, self-identified and allied collections-as-data practitioners advance critical practice within the constraints of local conditions and histories. Solutions generated within these contexts stand to benefit a broader

[16] "Gaze, Colonial ." International Encyclopedia of the Social Sciences. . Encyclopedia.com. 18 Oct. 2023 <https://www.encyclopedia.com>.

community of practitioners. In order to truly realize the potential of collections-as-data, it is essential that the community has the means to centralize and decentralize effort at local and global scales.

The need to operate at local and global scales becomes all the more important given increasingly pervasive implementation of artificial intelligence driven applications that make use of collections as data. Memory organizations need to know how various national contexts are addressing artificial intelligence in their work as a matter of policy and law. Sustained collaboration will be needed to determine what specific agency memory organizations have to impact policy and law in this space. The potential for corporations or other entities to use collections as data for training data - with or without memory organization consent - has implications for how, when, and in some cases whether or not collections as data should be produced at all.

Of course, AI has the potential to benefit memory organizations. For example, AI can strengthen collection access (e.g., improved optical character recognition (OCR), metadata generation) and help scale research services.[17] [18] Through automation, AI can help free up time for more high touch research and education support. Collections-as-data practitioners are well-suited to help realize the potential of AI. The collections-as-data community should be a key partner in AI assessment and implementation.[19] [20] [21]

Ideally, all collections-as-data effort is backed by “sustainable human infrastructure”.[22] A sustainable human infrastructure operates like an adaptive system where the removal or transition of one component in the system is not a catastrophic event. Rather, disruptions are absorbed, compensated for, and reformed by the infrastructure. To be clear, this is not to suggest that any one person is a cog or replaceable in collection-as-data work but rather that the infrastructure is in support of the human who is conducting said work. Sustainable human infrastructure encourages and supports development of sharable documentation, multilingual

[17] Hegghammer, Thomas. “OCR with Tesseract, Amazon Textract, and Google Document AI: A Benchmarking Experiment.” Journal of Computational Social Science 5, no. 1 (May 1, 2022): 861–82. https://doi.org/10.1007/s42001-021-00149-1.

[18] Corrado, Edward M. “Artificial Intelligence: The Possibilities for Metadata Creation.” Technical Services Quarterly 38, no. 4 (October 2, 2021): 395–405. https://doi.org/10.1080/07317131.2021.1973797.

[19] Responsible AI, https://www.lib.montana.edu/responsible-ai/

[20] Padilla, Thomas. “Responsible Operations: Data Science, Machine Learning, and AI in Libraries.” Accessed September 9, 2020. https://doi.org/10.25333/XK7Z-9G97.

[21] Rakova, Bogdana, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. “Where Responsible AI Meets Reality: Practitioner Perspectives on Enablers for Shifting Organizational Practices.” Proceedings of the ACM on Human-Computer Interaction 5, no. CSCW1 (April 22, 2021): 7:1-7:23. https://doi.org/10.1145/3449081.

[22] Cooper, Robert, and Michael Foster. “Sociotechnical Systems.” American Psychologist, vol. 26, no. 5, 1971, pp. 467–74, https://doi.org/10.1037/h0031539.

project processes and outcomes, opportunities to share technical expertise across institutions, and discussion of approaches to organizational design that support collections-as-data work. This infrastructure contributes to, and benefits from, a global community of practice.

The sum total of *Collections as Data: Part to Whole* activities suggests that the future of collections-as data work is best advanced through community mechanisms that facilitate sharing of experience across a wildly diverse set of professional and disciplinary roles on a global level. However, that mode of engagement faces fundamental resourcing challenges. In exceptional cases, organizations are able to facilitate collections-as-data effort at the multinational level (such as DARIAH and Europeana), yet they are ultimately bound at the supranational level with financial resources committed nearly exclusively to European Union (EU) member states. The United Kingdom's exit from the European Union had immediate impacts on UK memory organizations' continued ability to access financial resources that support EU member state organization collaboration. In the United States, as in many countries, government funding for memory organization work is often programmatically scoped on a national basis with little room for resourcing collaboration with memory organizations in other countries. In countries like Canada and Mexico, the funding environment for memory organizations is limited, constraining the ability to collaborate on a global basis.

Despite these challenges, collections-as-data practitioners have communicated intent to continue fostering community on a global scale. The *Collections as Data: Part to Whole* team is actively seeking partners that can resource and share the responsibility of collections-as-data community growth and impact on a global level. Preliminary conversations have been had with potential partners in the European Union, the United States, and Argentina. If your organization or community is interested in partnering on any part of the future of the collections-as-data field, as described above, please reach out to the *Collections as Data: Part to Whole* project team. On a similar note, if there is a way the project team can support your collections-as-data community effort, please reach out.

At times, the challenge and opportunity of collections-as-data can feel overwhelming. The implications of the work impact nearly every aspect of memory organization operations. An evermore digital cultural record suggests that collection-as-data work is imperative to organizational purpose.[23] The best way to tackle that imperative is together.

Sustain potential, mitigate harm.

[23] Padilla, Thomas. "On a Collections as Data Imperative," February 15, 2017. https://escholarship.org/uc/item/9881c8sv#author.

Let that be the clarion call.

1. Related Resources
    a. Cohort reports
    b. 2023 Position Statements
2. Appendix A
    a. Participant list
3. Glossary
    a. AI

Artificial Intelligence) is a field of study and practice that combines methods and areas of focus that include, but are not limited to, natural language processing, machine learning, computer vision, robotics, philosophy, mathematics, neuroscience, psychology, computer engineering, and linguistics in order to create “intelligent machines.”

b. Collections-as-data

Collections-as-data, written with hyphens, refers to the concept, community, and area of practice that focuses on responsible development and computational use of collections as data.

c. Cultural heritage

Cultural heritage is an admittedly imprecise term often used to describe the sector that includes Galleries, Libraries, Archives, and Museums (GLAM).

d. Datafication

Datification is the process of creating data that are machine actionable and thus amenable to computational analysis. It differs from digitization in that digital items are sometimes produced simply to serve as a digital surrogate and not organized, described, or made accessible in such a way that would facilitate computational analysis.

e. Data sovereignty

Data sovereignty refers to a group or individual’s agency over their data, including the collection, storage, and interpretation of those data[24]. Claims for data sovereignty may include the ability to control what data is shared, how it is shared, and with whom it is shared. The concept is particularly important in considering data of Indigenous and/or exploited populations.

f. Data Steward

A data steward is a person who is responsible for the development and maintenance of collections as data. Data stewards are often information professionals in GLAM institutions but the rise of community archiving has meant that community members are increasingly taking on these roles.

g. Digitization

[24] Smith, D.E. (2016). Governing data and data for governance: the everyday practice of Indigenous sovereignty, Ch. 7 in Indigenous Data Sovereignty: Toward an Agenda

Digitization is the process of reproducing analog items in digital formats. Examples of digitization include scanning pages of a manuscript to produce .tiff images, 3D scanning, and producing mp3 files from vinyl records.

h. GLAM

GLAM is an acronym for Galleries, Libraries, Archives, and Museums. Sometimes it is helpful to refer to these institutions as a group because they often face similar challenges related to creating, describing, providing access to, and supporting the use of digital collections.

i. LLM

LLM stands for Large Language Model. LLMs are trained on large amounts of text and used to produce text output. ChatGPT is currently the most popular LLM but it is not the only one. Collections of texts that are made available as data may be used to train LLMs.