Thomas G. Padilla

Collections as data

Implications for enclosure

In recent years a growing amount of interest has been dedicated to *collections as data*.[1] A collections as data paradigm seeks to foster an expanded set of research, pedagogical, and artistic potential predicated on the computational use of cultural heritage collections. Collections as data raises the question of what it might mean to treat digitized and born digital collections as data rather than simple surrogates of physical objects or static representations of digital experience.

Examples of work pursuing this question are growing. Project AIDA uses machine learning to identify and subset poetry from the pages of nearly 200 years of digitized historic newspapers.[2] Archives Unleashed works to develop a cloud-based environment that enables computational analysis of web archive collections.[3] The HathiTrust Research Center continues efforts to support computational analysis of collections and is joined by independent initiatives at institutions like the University of Miami, Carnegie Museum of Art, and the Massachusetts Institute of Technology.[4] The University of California Libraries system plans to address collections as data as a Shared Content Leadership Group 2017/18 priority.[5] The Institute of Museum and Library Services-supported Always Already Computational: Collections as Data (AAC) is wholly focused on collections as data advocacy and resource development.[6]

While conceiving of AAC, there was some initial debate about whether it should seek partnerships with for-profit publishers and content providers. The team resolved that it would be more productive to focus on sparking forward momentum among nonprofit cultural heritage organizations. We made a commitment to openness, aligning our project activity with a corresponding community of practice, including libraries, archives, and museums within the United States and beyond. Project deliverables like *The Santa Barbara Statement on Collections as Data* are not just openly available, they are openly annotatable.[7] In making commitments of this kind, we aspire to live in accordance with project values. As others have noted, the product of work guided by these values run the risk of enclosure by for-profit actors.[8]

As we reflect on the generative potential of collections as data, we must also consider the threat of enclosure. This line of thinking spans collections, infrastructure, and, ultimately, you and me.

The reflections that follow are offered independent of AAC and are not an official representation of the University of Nevada-Las Vegas Libraries.

Thomas G. Padilla is principal investigator, Collections as Data, and visiting digital research services librarian at the University of Nevada-Las Vegas, email: thomas.padilla@unlv.edu

On collections

. . . many of us have been party to enclosing that which could have been open.

Increasingly, researchers request *access at scale* to large collections of in-copyright and/or licensed materials in order to conduct various forms of computational research (e.g., text mining, data mining, machine learning). The route to use of these resources is often beset by exorbitant costs, opaque delivery timelines, technical underdevelopment, and terms of use that fly in the face of reproducible research.[9] When faced with criticism, vendors will often raise the challenge of multitudinous legacy content provider agreements.

On the face of it, this challenge might foster some patience. Who would envy the task of renegotiating use rights for content produced by some of the most storied publishers of news and contemporary culture? However, upon further investigation we find that many nonprofit cultural heritage organizations commingle with this content provider group.

Many of us have been party to enclosing that which could have been open. Royalties are commonplace. Some have asserted that commercial partnerships offer a legitimate route to increasing collection access. The assertion is followed by an observation that most public institutions do not receive their total operating budget from public funds. Consequently, it is argued that commercial partnerships are necessary to fulfill public benefit, where possible, channeling derived profit toward support of the commons.[10] We must ask whether this strain of commercial collaboration and the unfavorable context it flourishes in are worth sustaining. Is it worth boosting cost of admission to an enclosed garden that weakens the library community *and* inhibits emerging forms of research?

On collections use

. . . work to guide, and indeed, in some cases inhibit the use of collections as data altogether.

In light of factors inhibiting in-copyright or licensed collection use, some have declared that, "the right to read is the right to mine."[11] The initial context—digitized scientific information—is admirable, though the ethical dimensions of the declaration are troubled as they combine with the expanded scope of collections as data. Systems like Mukurtu and Traditional Knowledge (TK) Labels were expressly developed to inhibit unbridled collection use predicated on the legacy and ongoing operation of colonial appropriation.[12] As libraries and archives ramp up collections as data development, it is imperative that they critically engage with the question of ethical use vis à vis the proliferation of right to mine perspectives. It is crucial that we work to guide, and, in some cases, inhibit the use of collections as data altogether. To bring this imperative into focus, we might consider library and archive efforts to develop web and social media collections.

Bergis Jules has noted the rise of social media data services that market to law enforcement.[13] Private security firms have sought access to datasets that document the Ferguson, Missouri, protests that arose following the killing of Michael Brown. What role might our institutions play in providing access to collections that hold the potential to harm communities? What might happen if we allow a role in stewarding ethically grounded use of collections to be enclosed by entities that value capital over social good?

On infrastructure

. . . the scope of research questions are demarcated by the resources of for-profit gardens strewn with transmogrified open source tools.

For some time, academic publishers have been expanding business strategy to accommodate enclosure of scholarly infrastructure. As Tom Cramer noted in a Spring 2018 Coalition for Networked Information presentation, emerging verticals are enclosing core components of the research process—discovery, reference management, social networking, profiling, publications/citations, evaluation, funding opportunities, and digital repositories.[14] Seen from the

vantage point of collections as data, we might discern an additional component of scholarly infrastructure—infrastructure that seeks to enable computational research. Publishers and vendors appear to be trending away from allowing libraries to acquire and provide access to collections as data on their own terms, in line with their values, supported by community-owned infrastructure. Instead, publishers and vendors are retaining their collections and developing application programming interfaces and web-based environments on top of them.

In prioritizing this approach, publishers and vendors create another point of library lock-in and further drain library budgets. Access points to data and analytical functions are by design enclosed, lacking interoperability with other data sources. Consequently, the scope of research questions are demarcated by the resources of for-profit gardens strewn with transmogrified open source tools. Furthermore, by retaining control of programmatic interaction, publishers and vendors have the capacity to monitor and monetize researcher queries to their data. This situation is antithetical to academic freedom and a point of risk that should be mitigated, given a tradition of governmental efforts to surveil user activity that shows no sign of abating.[15]

In lieu of a corporate incentive to act differently, cultural heritage organizations are called to think expansively about the development of community-owned infrastructure that enables computational research. This work must be grounded by a capacious sense of variation in institutional resources and missions. The work of the HathiTrust Research Center is admirable, but we would be mistaken to assume the feasibility of a one-ring-to-rule them all solution. We need many rings.

On you and me

. . . discussions of infrastructure tend to elide discussions of people.

As various components of scholarly infrastructure have been enclosed, little attention has been paid to the potential enclosure of you and me. To some extent this gap is a product of how discussions of infrastructure tend to elide discussions of people. In response to the threat of enclosure, SPARC has prioritized "research and development efforts on new economic and organizational models for the collective provisioning of open resources and infrastructure;" David Lewis has raised the notion of a 2.5% library budget commitment to organizations and projects that contribute to common digital infrastructure—a variant of an argument previously advanced by Leslie Chan; and the "COAR Next Generations Repositories" report articulates a series of steps for advancing the capabilities of open scholarly infrastructure.[16] Nary a mention of explicit threats to librarian roles. When infrastructure and people are raised to the same level of consideration, an enclosure threat to librarian roles comes into clearer focus.

Threats approach the crystalline when we consider the manner in which publishers and content vendors are leveraging infrastructure to compete directly with the viability of librarian roles. For example, consider publisher and content vendor development of web-based environments that enable basic forms of text analysis and visualization and the rise of data curation services stacked on top of for-profit repository infrastructure. [17] In the best cases, these services come in the guise of a thing that is supposed to help librarians. However that claim is weighed, these services come at the cost of fully realizing our roles as educators, consultants, partners, and advocates.

Rather than bolstering our relevancy by working through the various components of the research process, we outsource the potential of locally held strengths. We must resist commodification of digital scholarship, data curation, and research data management roles across our libraries. Given the complex data-oriented challenges of the contemporary information environment, we must ask ourselves whether we are willing to allow emerging facets of librarian

expertise to be enclosed. Finally, we must cultivate a sense of for-profit market strategy that equally addresses infrastructure and the human resources that give those infrastructures meaning in the world. Data-oriented roles in libraries have long been troubled by organizational integration. Despite ongoing challenges, we must not allow for-profit actors to leverage cracks in our efforts. Rather, we should work to ensure that those cracks are the places where the light gets in.

Notes

1. Thomas Padilla, "On a Collections as Data Imperative," Library of Congress, last modified February 15, 2017, https://escholarship.org/uc/item/9881c8sv.

2. Elizabeth Lorang and Leen-Kiat Soh, "Image Analysis for Archival Discovery (Aida)," accessed April 15, 2018, http://projectaida.org/.

3. Milligan, Ian, Nick Ruest, and Jeremy Lin, "The Archives Unleashed Project," accessed April 15, 2018, http://archivesunleashed.org/.

4. Paige Morgan, Elliot Williams, and Laura Capell, "La Gaceta de la Habana," accessed April 15, 2018, https://collectionsasdata.github.io/facet7/; David Newbury and Daniel Fowler, "Carnegie Museum of Art Collection Data," accessed April 15, 2018, https://collectionsasdata.github.io/facet2/; Richard Rodgers, "MIT Libraries Text and Data Mining," accessed April 15, 2018, https://collectionsasdata.github.io/facet1/.

5. "2017/2018 SCLG Plans & Priorities for 2017/2018 Based on the University of California Library Collection: Content for the 21st Century and Beyond," University of California, last modified September 29, 2017, http://libraries.universityofcalifornia.edu/groups/files/sclg/docs/SCLG_2017_2018%20Plan.pdf.

6. "Always Already Computational: Collections as Data," accessed April 10, 2018, https://collectionsasdata.github.io/.

7. Thomas Padilla, Laurie Allen, Stewart Varner, Sarah Potvin, Elizabeth Russey Roke, and Hannah Frost, "Santa Barbara Statement on Collections as Data," Always Already Computational Collections as Data, accessed April 10, 2018, https://collectionsasdata.github.io/statement.

8. Heather Joseph, "Open In Action—Bridging the Information Divide," last modified December 10, 2017, www.caul.edu.au/content/upload/files/cairss/repository-infrastructure2017sparc-cni-presentation.pdf.; Geoffrey Bilder, Jennifer Lin, and Cameron Neylon, "Principles for Open Scholarly Infrastructures-V1," *Figshare,* 2015, https://doi.org/10.6084/m9.figshare.1314859.

9. Stodden, Victoria., et al., "Enhancing Reproducibility for Computational Methods," Science 354, no. 6317 (December 9, 2016): 1240–41. https://doi.org/10.1126/science.aah6168.

10. Richard Fyffe and Beth Forrest Warner, "Where the Giants Stand: Protecting the Public Domain in Digitization Contracts with Commercial Partners," *Journal of Library Administration* 42, no. 3–4 (May 31, 2005): 83–102, https://doi.org/10.1300/J111v42n03_06.

11. Peter Murray-Rust, "The Right to Read Is the Right to Mine," *Open Knowledge International* (blog), June 1, 2012, https://blog.okfn.org/2012/06/01/the-right-to-read-is-the-right-to-mine/.

12. "Mukurtu," Mukurtu CMS, accessed April 12, 2018, http://mukurtu.org/; "Traditional Knowledge (TK) Labels," Local Contexts, accessed April 12, 2018, http://localcontexts.org/tk-labels/.

13. Bergis Jules, "Surveillance and Social Media Archiving," *Medium—Documenting DocNow* (blog), October 4, 2016, https://news.docnow.io/surveillance-and-social-media-archiving-7ea21b77b807.

14. Keith Webster, "Nice Depiction of Verticals (or Silos) of Research Information Management Systems by @tcramer of @StanfordLibs," accessed April 12, 2018, https://twitter.com/CMKeithW/status/984551536406683648.

15. Robert D. McFadden, "F.B.I. in New York Asks Librarians' Aid In Reporting

on Spies," *The New York Times,* September 18, 1987, sec. N.Y. / Region, www.nytimes.com/1987/09/18/nyregion/fbi-in-new-york-asks-librarians-aid-in-reporting-on-spies.html; Dustin Volz, "IBM Urged to Avoid Working on 'Extreme Vetting' of U.S. Immigrants," *Reuters,* November 16, 2017, https://www.reuters.com/article/us-ibm-immigration/rights-groups-pressure-ibm-to-renounce-interest-in-trumps-extreme-vetting-idUSKBN1DG1VT.

16. "2018 SPARC Program Plan," SPARC, accessed April 14, 2018, https://sparcopen.org/who-we-are/program-plan/; David W. Lewis, "The 2.5% Commitment," Working Paper, September 11, 2017, https://doi.org/10.7912/C2JD29; Leslie Chan, "Transforming Global Knowledge Exchange: Reframing the Roles of the Libraries," last modified August 9, 2010, https://www.slideshare.net/lesliechan/chan-pre-ifla2010; "Next Generation Repositories," COAR, accessed April 14, 2018, https://www.coar-repositories.org/activities/advocacy-leadership/working-group-next-generation-repositories/.

17. "Gale Digital Scholar Lab," GALE, accessed April 18, 2018, https://www.gale.com/primary-sources/digital-scholarship; "Springer Nature Research Data Support," *Springer Nature,* accessed April 18, 2018, https://www.springernature.com/gp/authors/research-data-policy/.

("2018 top trends . . . " continues from page 293)

at Virginia Tech," Report, Virginia Tech, June 22, 2017, http://bit.ly/2FccOwu.

45. Alexandre Ribas Semeler, Adilson Luiz Pinto, and Helen Beatriz Frota Rozados, "Data science in data Librarianship: Core competencies of a Data Librarian," *Journal of Librarianship and Information Science* (November 2017): 1–10.

46. Andrew Weiss, *Big Data Shocks: An Introduction to Big Data for Librarians and Information Professionals* (London: Rowman and Littlefield, 2018).

47. Michael Zeoli, "Trends in Academic Library Acquisitions," presentation at the Charlotte Initiative Symposium, held at the Charleston Conference, November 6–10, 2017.

48. Stephanie J. Spratt, Gabrielle Wiersma, Rhonda Glazier, and Denise Pan, "Exploring the Evidence in Evidence-Based Acquisition," *The Serials Librarian* 72, nos. 1-4 (2017): 183–89.

49. David W. Lewis, "The 2.5% Commitment," IUPUI Scholar Works, September 11, 2017, http://hdl.handle.net/1805/14063.

50. David W. Lewis, Lori Goetsch, Diane Graves, and Mike Roy, "Funding Community Controlled Open Infrastructure for Scholarly Communication," *College and Research Libraries News* 79, no. 3 (March 2018): 133–36.

51. Lisa Macklin and Chris Palazzolo, "Open Access Collection Development Policy at Emory," ASERL Webinar, June 6, 2016, http://bit.ly/2FkmN6p.

52. Kristin Antelman, "Leveraging the Growth of Open Access in Library Collection Decision-Making," http://bit.ly/2FOpICk.

53. "Austrian Open Access Agreement with Publisher Wiley," accessed April 8, 2018, http://bit.ly/2tdgfSg.

54. See https://www.hathitrust.org/shared_print_program.

55. "The Future of the Academic Library Print Collection: A Space for Engagement," Arizona State University White Paper, http://bit.ly/2Fa50vx.

56. Lisa M. Rose-Wiles and John P. Irwin, "An Old Horse Revised?: In-House Use of Print Books at Seton Hall University," *The Journal of Academic Librarianship* 42, no. 3 (2016): 207–2014.