Jodi Allison-Bunnell

The Puzzle of Large-Scale Digital Collections

Have We Reached an Inflection Point?

Since the debut of digital collections[1] from libraries, archives, and other cultural heritage institutions in the mid-1990s, we've searched for solutions to make those collections easily available to researchers. Aggregations and subject-based portals emerged as part of those solutions, with enthusiastic support from federal granting agencies, states, and foundations. Some (California Digital Library, Mountain West Digital Library) have adapted and persisted over time, some are present but less robust (Western Waters), and others are long gone (Colorado Digitization Program, Washington Women's Heritage). After a quarter-century of investment in digital collections at and across institutions in the United States, we clearly struggle to find sustainable and effective solutions. It's a fiendishly difficult problem in the absence of other options, such as a federally supported national digital collections program. Significantly, the Digital Public Library of America (DPLA) announced publicly on April 3, 2024, that it was seeking a new organizational home for its cultural heritage aggregation program after concluding that it could not sustain the program in its current form.[2] With this announcement, and the debut last summer of JSTOR's Shared Collections, it's useful to reflect on whether this new service represents a significant progression in this space, or if it's more likely that any cultural heritage aggregation in the United States will continue to struggle.[3]

Shared Collections allows institutions either to have JSTOR harvest their digital collections of documents, photos, and other special collections from a local Digital Asset Management System, or to create and share those same collections through JSTOR's collection management tool.[4] The cost for either harvesting or hosting is modest, but participants can also add other services.

In some ways, Shared Collections looks a lot like the DPLA. Both aggregate unique digital content from libraries, archives, and museums across America. Founded in 2013, DPLA currently includes almost 50 million photographs, documents, and audio and video recordings from a wide spectrum of contributing institutions, from the smallest historical societies to the Library of Congress and many portions of the Smithsonian Institution.[5] DPLA has been an important part of efforts to develop access to digital collections. Its network of state and regional hubs, many of which are supported by Library Services and Technology Act (LSTA) funds from the Institute for Museum and Library Services, provide essential infrastructure, training, and communities of practice for far-flung practitioners.

DPLA has also struggled with three key issues: financial stability, metadata inconsistency, and raison d'etre. To determine if Shared Collections represents a significant advance, it's useful to compare the two offerings on these issues.

Jodi Allison-Bunnell is head of Archives and Special Collections and senior archivist at the Montana State University Library, email: jodi.allisonbunnell@montana.edu.

Funded largely by grants from its inception, DPLA has struggled to transition to more sustainable funding. It received about $7 million from federal granting agencies between 2014 and 2018, plus significant grants from major foundations over the full ten years of its existence.[6] Since then, additional foundation funding and a re-focus on ebooks, along with a fee-based model for its hubs, has kept its cultural heritage aggregation afloat. But not by very much. I calculated from a review of DPLA's forms 990 that the hub fees cover only about 13% of its operating costs. Also, not all hubs have been unable to continue annual fees or have chosen not to do so.[7] Without more support from participating institutions or other sources, this is obviously unsustainable. Over five years ago, Roger Schonfeld of Ithaka S + R posited that it may have been a mistake for DPLA to try and exist as a standalone organization.[8] That DPLA itself has concluded this and launched a search for a new home suggests that his statement was correct.

JSTOR is part of ITHAKA, a financially stable organization that also maintains Artstor, Portico, and Ithaka S + R, its research operation.[9] The organization has a long history of developing cost distribution models to sustain its services. Those models are used to determine cost distribution for other services in libraries and archives, including ArchivesSpace (which is, in turn, one of the few sustainability success stories in the cultural heritage sector). In its presentations on the financial model for Shared Collections, JSTOR predicted that after a start-up phase of major development and promoting adoption, the program would reach a financial break-even point in about 2025. By being part of an established organization, Shared Collections also has lower overhead costs. There's another advantage to a JSTOR service that I observed in over a decade at a regional academic library consortium: an addition to a bill from an existing vendor or partnership is more easily supported by a resource allocator than one coming from a new or unfamiliar organization.

The inconsistent metadata for digital collections presents the second major challenge for any aggregator. With standards like Dublin Core (a lowest-common-denominator but very accessible metadata scheme), a passion for institution-specific branding and customization, and few or no systems that encourage standards-compliant metadata, digital collections inevitably have a mass of metadata that does not gracefully co-mingle in a shared search and retrieval system. DPLA has some minimal requirements for metadata ingested through its system of hubs.[10] Despite those requirements, DPLA has faced many challenges with metadata consistency that have required remediation either at the originating institutions, at the hubs, or at DPLA central.

As part of facing those challenges, DPLA has had significant success in advancing the cause of consistency and standards compliance. The organization's work on standardized rights statements has been quite transformative. That work, led in the United States by Emily Gore, was a response to the mind-bending variation in the Dublin Core Rights field. As DPLA worked with Europeana and other partners, they identified more than 87,000 unique rights statements in the Rights field in DPLA, most of which were confusing, inaccurate, or not about rights at all.[11] This is a prime example of what only becomes evident in aggregations! Implementing standardized rights statements has been groundbreaking, difficult, and immensely important. From my experience leading the work to develop two hubs for the Northwest, the work on standardized rights statements motivated us to develop and deliver training on determining copyright status that made the task approachable for even the smallest institutions.[12] Ultimately, the leadership that DPLA has exercised on this and many other problems in this space are of lasting importance.[13]

Can JSTOR overcome issues of metadata inconsistency? It's possible that they might in an organization with robust technical capabilities, over 200 employees, and with the promise of artificial intelligence (AI) tools.[14] JSTOR is beginning Shared Collections only requiring a Resource Type in contributed metadata. It offers participating institutions guidelines that strongly encourage nine other fields and make the impact of institutional decisions on search clear.[15] When I first read these ultra-light requirements, I was concerned that the absence of requirements would replicate the same issues that DPLA has faced. JSTOR is considering making copies of submitted metadata and enhancing them with AI tools. Current work focuses on identities and keywords but will include other properties in the future. The original records would not be modified, but the search infrastructure could work primarily with the enhanced records that have the regularity that the originals may lack.[16] This approach—which both overcomes the limitations of inconsistent metadata and also applies AI tools at a scale not possible for many institutions—could be a significant advance in this field if JSTOR is able to implement it.

JSTOR and DPLA have different positions in the information landscape. JSTOR is a destination site for millions of academic researchers from high school upward who depend on it for resources that are otherwise inaccessible to them. (According to Bruce Heterick, Senior Vice President, Open Collections & Infrastructure at ITHAKA, JSTOR is consistently among the top three used sites in the academic libraries he has visited over the course of his career.) Making unique materials available alongside published materials allows "accidental" discovery of unique materials that may enrich a project or inquiry but that the researcher didn't think to search for specifically. DPLA promised to be a destination site as well. Many hoped for that outcome, but also questioned whether their infrastructure could really deliver that.[17] DPLA is certainly a destination for K-12 teachers who use its well-curated Primary Source Sets, of selected sources from DPLA and others with a teaching guide. With its commitment to shareable metadata, DPLA was also originally designed to be a platform for building apps—a promise that didn't pan out. (Though who can forget the Historical Cats app.)

Last, the technical model for each organization is significantly different. When it was created, DPLA decided to harvest metadata from a system of state and regional hubs and a few very large institutions, and to not host the digital objects themselves. In doing so, they avoided the need to build a national-level digital asset management system and established a network of local and regional sources for training, support, and infrastructure. That was a brilliant decision in many ways: The hubs structure kept DPLA from having to manage relationships with hundreds of contributing institutions. But as search engines evolved, that decision came to negatively affect harvesting and indexing by search engines. Search engines have little interest in metadata-only records; they want to deliver their customers directly to digital objects. In 2021, Montana State University partnered with DPLA to determine if, by hosting display objects, DPLA could improve search engine rankings.[18] The results were a clear "yes," as reported by Kenning Arlitsch and Michael Della Bitta at the Council for Networked Information meeting in 2021.[19] Making that move, however, would have been a fundamental shift in DPLA's technical and governance infrastructure. With a strong tradition of institutional control over digital cultural heritage, giving up that control by having traffic directed away from the institution—even with increased ease for researchers—was untenable when DPLA was created and may remain so.

Shared Collections either harvests or hosts the original objects from each collection. As expected from Arlitsch and Della Bitta's research—and admittedly based on a review of my institution's collections only—the search results from Shared Collections are very highly ranked (even above those of our locally hosted collections where we have taken great pains with search engine optimization).

While Shared Collections appears to represent a significant advance, the jury will be out for some time. The fundamental issues facing DPLA and Shared Collections are simply difficult, and the struggles with them have little or nothing to do with the skills or intentions of the capable people of both organizations. It is both a tough economic problem and an outcome of what we might call "rugged individualism in heritage collections": while shared descriptive efforts have been in place for books for more than a century, many standards for heritage collections have emerged since 2000. It's a symptom of under-investment in cultural heritage in the United States. We may look with admiration at Europeana, Trove, and the national libraries of Europe, which exist because there is national (and in the case of Europeana, multi-national), centralized, and sustained investment in cultural heritage. In the United States, we are left with much more piecemeal efforts that leave organizations struggling from the very bottom (small historical societies and public libraries) to the top (national-level aggregations). In in cultural heritage work, we have a bad habit of focusing overmuch on local customization, functionality, and appearance while working with metadata. Any efforts to combat that are an advance in the cause of making digital collections genuinely discoverable.

So, let's hope that JSTOR is onto something great. Let's hope that all the good that DPLA has done can be sustained by finding an organization that is able to take on its mission financially and technically.[20] Let's hope that other major efforts to aggregate cultural heritage (the National Finding Aid Network and Social Networks and Archival Context are the most important ones) can be sustained. The US desperately needs a national-level infrastructure and approach to digital collections and cultural heritage, and we can all be grateful for and supportive of the efforts of all the organizations and individuals who are working to make that a reality.

Notes

[1] I am using the term "digital collection" in the sense of the definition "a logical grouping of related digital content that is organized by collection-level metadata. All digital content items (digitized and born digital) are capable of existing within a digital collection." Library of Congress, Digital Collections Management Compendium, Glossary, accessed December 8, 2023, https://www.loc.gov/programs/digital-collections-management/about-this-program/glossary/.

[2] Dominic Byrd-McDevitt, "Applications Open to Find a New Home for America's Digital Heritage," Digital Public Library of America, April 3, 2024, https://dp.la/news/applications-open-now-a-new-home-for-americas-digital-heritage?mc_cid=0a71256a7e&mc_eid=0e17877a45.

[3] JSTOR, "Amplify the Reach of Your Collections with JSTOR," accessed April 5, 2024, https://about.jstor.org/whats-in-jstor/infrastructure/share/.

[4] JSTOR, "Why Share Your Content with Collection Loader?," accessed April 5, 2024, https://about.jstor.org/l/load/.

[5] According to the Digital Public Library of America, https://dp.la/, accessed April 5, 2024.

[6] J. Allison-Bunnell, "Finding Aid Aggregation at a Crossroads," UC Office of the President: California Digital Library, 2019, https://escholarship.org/uc/item/5sp13112.

[7] Jodi Allison-Bunnell, "Finding Aid Aggregation: Toward a Robust Future," *The American Archivist* 85, no. 2 (2022): 556–86, https://scholarworks.montana.edu/xmlui/handle/1/17556. An example of a hub that has been unable to pay fees is the California Digital Library; the Big Sky Country Digital Network decided to withdraw in 2021 because the value proposition for the hub fee was insufficient.

[8] Roger Schonfeld, "Learning Lessons from DPLA," *The Scholarly Kitchen*, November 13, 2018, https://scholarlykitchen.sspnet.org/2018/11/13/learning-lessons-from-dpla/.

[9] ITHAKA, "What We Do," accessed April 5, 2024, https://www.ithaka.org/#what-we-do.

[10] Its required properties are a title, URL, data provider, and rights statement. It encourages use of nine other properties and includes controlled vocabularies and authority sources in the appendixes. Digital Public Library of America, Metadata Application Profile v. 5.0, December 7, 2017, https://drive.google.com/file/d/1fJEWhnYy5Ch7_ef_-V48-FAViA72OieG/view?usp=sharing.

[11] Mark Matienzo, "Rights Statements in Digital Object Aggregators," Digital Library Federation Forum, October 28, 2018, https://matienzo.org/storage/2014/2014Oct-DLF-Rights.pdf.

[12] Northwest Digital Heritage, which serves organizations in Washington and Oregon, and the Orbis Cascade Alliance's hub (which the organization decided not to maintain).

[13] The author serves as co-chair of the Rights Statements Working Group of DPLA, which is currently working on issues around indigenous materials.

[14] JSTOR LinkedIn profile, accessed April 5, 2024, https://www.linkedin.com/company/jstor/about/.

[15] "Guide to JSTOR Search for Shared Collections Contributors," https://support.contributors.jstor.org/hc/en-us/articles/360058878154-Guide-to-JSTOR-Search-for-Shared-Collections-Contributors; Shared Collections Metadata, https://support.contributors.jstor.org/hc/en-us/articles/360044658434-Shared-Collections-Metadata. The nine "strongly recommended" fields are identifier, title, contributor, date(s), description, subject(s), rights/license, holding institution, and canonical URL.

[16] Paraphrased from conversation between Bruce Heterick and the author, January 5, 2024; information enhanced from conversation between Jason Przbylski and Lenny Adler (both of JSTOR) and the author, April 23, 2024.

[17] See Kenning Arlitsch and Patrick O'Brien, "Our Relationship with Internet Search Engines," Re:Thinking, Council on Library and Information Resources, March 21, 2013, https://www.clir.org/2013/03/our-relationship-with-internet-search-engines.

[18] Michael Della Bitta, "Improving Access and Discovery," DPLA News, January 29, 2021, https://dp.la/news/improving-access-and-discovery.

[19] Kenning Arlitsch and Michael Della Bitta, "Is It Time to Give the Digital Public Library of America Our Digital Objects?," CNI Fall 2021 Project Briefings, last updated July 25, 2022, https://www.cni.org/topics/special-collections/is-it-time-to-give-the-digital-public-library-of-america-our-digital-objects.

[20] Digital Public Library of America, "Applications Open to Find a New Home for America's Digital Heritage," accessed April 5, 2024, http://dpla.wpengine.com/wp-content/uploads/2024/04/EOI-FAQ-4.2.24.pdf?mc_cid=0a71256a7e&mc_eid=0e17877a45.