

## On a Collections as Data Imperative

---

Thomas Padilla, Humanities Data Curator, University of California Santa Barbara

*At the back of any subject is the threat  
of knowing nothing of certainty at all.*

*The threat of knowing nothing is not the same  
as uncertainty, which is the presence of alternatives.*

*A Maxwell, Candor is the Brightest Shield*

Libraries support individuals working through the many facets of complexity that constitute the human condition. The collections as data conversation is an extension of this tradition - provision of the means for meaning making. Disposition toward the work is unadorned, grounded by engagement with community need and vested in the challenges and opportunities latent in the traces of human action gathered, described, preserved, and provided access to. Typically, these traces are called collections. What might be gained by thinking of the digital objects that comprise them as data? Within this question lies the potential of a collections as data imperative.

To see collections as data begins with reframing all digital objects as data. Data are defined as ordered information, stored digitally, that are amenable to computation. Wax cylinders, reel to reel tape, vellum manuscripts, websites, masterworks, musical scores, social media, code, and software in digital collections are brought onto the same field of consideration. The value of such a shift can be explored in part by asking how thinking about an object as data multiplies and/or extends the questions that can be asked. For example, if the notion of a single digitized text is shifted from a surrogate of a bound paper object to consider the possibility latent in a form that is computationally processable at the level of thousands or even millions of texts, a move is made toward meaning making that engages affordances unique to data. When a tweet, website, or work of electronic literature is examined, what interlocking standards and structures can be discerned beneath the representation seen on the screen? How might these standards and structures be traversed computationally to gain a better understanding of networks of meaning built daily across time, space, and language? When art is explored as data what computational tools are available to aid eyes and minds in the assessment of hue, saturation, light, darkness, intention and style?

Meaning making with collections as data is not solely a consideration of whether a computer can be used to process, visualize, and mine them. An orientation to collections as data is about cultivating perception that pushes past the surface of the things that inhabit digital environments. A Word document is not just a document, a website is not just a projection on a screen, and a tweet is much more than 140 characters.

To realize this is to regain a kind of agency that has become increasingly elusive. This orientation forces reckoning with the composite nature of digital experience. Awareness of composition highlights human intention humming below the surface of seemingly mundane interactions. Taking Twitter data as an example, they appear to be simple 140 character utterances. Yet digging deeper bears the realization that a 140 character transmission is a small portion of a dataset that is captured each time a tweet occurs.

Geolocation, timestamping, links to webpages, links to images, language, and a wide array of data that records relationships between Twitter users is captured and counted discreetly *ad infinitum*. This mass of data serves purposes that are often not readily apparent. When thinking of using Twitter as merely sharing a brief message, or perhaps a paper note, individual ability to have material purchase on the world is done a disservice. In the meantime, power and control resides in the hands of those that take a data first, representation second mentality - namely corporations, governments, researchers, and increasingly as Bergis Jules has noted - law enforcement agencies. In simple terms, a collections as data imperative entails developing the means to help all members of society, across all classes and backgrounds, working within the academy and outside of it to engage critically with the traces of human activity we collect in the fullest manner possible, native to the complexity of their form, and critically attuned to the possibilities *and* perils that come with their use.

### **Collections as Data: Stewardship and Use Models to Enhance Access**

In September 2016, the Library of Congress' National Digital Initiatives (NDI) team brought together a diverse group of experts from academic institutions, non-profits, independent research groups, Federal agencies, private funding organizations, public libraries, and industry to highlight collections as data work and discuss possible roles that libraries can play. Day 1 featured a rich group of talks that highlighted work with Library of Congress collections, a broader field of data use, nuanced ethical issues, and approaches to developing and sustaining communities around the use of collections as data. Day 2 hosted a series of focus groups geared toward generating recommendations for how the Library of Congress might expand its role in this space through infrastructure development, programmatic innovation, and intra-governmental and extra-governmental partnerships. Aspects of the recommendations could be adopted singly or in collaboration by institutions differing in size and resources. Three conceptual frames inspired by the event are provided to help guide thinking through how libraries can participate in a collections as data imperative:

**Generativity** – to increase meaning making capacity

**Legibility** – to document and convey provenance and possibility

**Creativity** – to empower experimentation

#### ***Generativity***

Libraries foster the conditions for meaning making to occur. Success in this effort is predicated on potential latent in collections, infrastructure and spaces - both virtual and physical - that provide access, and librarian ability to help diverse communities navigate the possibilities that arise during use. Collections as data entails thinking about ways to increase meaning making capacity by making collections more amenable to use across an expanded set of methods and tools, typically but not exclusively computational in nature. Increasing amenability entails reconsidering approaches to collection form, description, discovery, and access - but perhaps more importantly it entails vigilant attention to *who the work is done for*. It is typical when considering this question to think of “target audiences”. To think about partnerships is more productive.

During the *Collections as Data* symposium Matthew Weber, Kate Zwaard, and Elizabeth Lorang spoke to the value of academic partnerships. For Weber, comparing disciplinary perspectives with a historian during a web archive hackathon hosted at the Library of Congress helped remove a “black box” from data. For Zwaard, Library of Congress staff collaboration with researchers at the event provided an opportunity to leverage subject area and collection expertise. In turn, library staff gained exposure to demands that emerging research places on library data. With *Text and/as Image*, Elizabeth Lorang shared a librarian and computer science faculty collaboration that explores the use of image recognition techniques to identify and extract poetry from page images of historical text in the Library of Congress’ *Chronicling America* collection. This brief sampling of partnerships demonstrates reciprocal benefits that are possible in collections as data work. Partnership afforded enhancement of the meaning making capacity of collections for academic researchers in the near term, for the library to steward that meaning making, for the library to gain a sense of where services can be improved, and for research interests to be aligned in such a way that techniques are developed to enhance collections for a wide set of communities. For example, Lorang’s work with identification of poetry in *Chronicling America* could likely be applied to textual collections held by other cultural heritage institutions. Consider how powerful it could be to provide public access to all of the poetry embedded in digitized historical text that libraries, archives, and museums have produced over the past 20 years.

With *Documenting the Now*, Bergis Jules’ modeled a form of partnership that extends to data creators. Documenting the Now develops open source tools that enable collection, analysis, and sharing of Twitter data. The project was spurred by the death of Michael Brown in Ferguson, Missouri. From the outset the project team has consistently focused on the human dimensions of the data they are collecting. That focus has led to partnerships with a diverse cast of community representatives, in an attempt to do collections as data work that elevates the primacy of source community need over a normative commitment to unfettered access to data. This type of partnership is simultaneously about respecting and safeguarding humans that can be readily identified and possibly persecuted through the data they leave behind, while trying to balance preservation and future use of an important testament to American life in a time of great discord. Documenting the Now provides a model for the Library of Congress to consider as it continues to work toward providing access to its own Twitter data collection and other collections that bear similar concerns.

### ***Legibility***

To make collections as data usable, the processes by which they are established must be made legible. These data are the product of design decisions whose purposes are typically not available for a user to consider. Lack of availability can be traced to a predominant understanding of digital collection use that does not address the needs of users who desire to work with collections computationally. The result is the presentation of seamless digital collections that aim to support interactions with objects rather than with the data that comprise those objects. Libraries do not often provide access to the scripts that generate collection derivatives, access to processes for cleaning or subsetting data, access to custom schema that have been used, indications of how representative digital holdings are relative to overall holdings, nor is the quality of data typically indicated. Libraries do not typically expose why some collections have been made available and others have not. Libraries do not typically identify the library staff personally responsible for modifying, describing, and creating collections – a dimension of provenance that must be

accessed in order to determine data ability to support a research claim. Collections as data possibility is contingent on integrity vouchsafed by expanded documentation practice. Work in this area can be informed by efforts to cultivate data practices that encourage reproducible research. Victoria Stodden provides a vital reference point in the recent *Science* article, “Enhancing Reproducibility, for Computational Methods”, which argues for ready availability of code, data, and workflows.

Attempts at creating legible collections are aided by becoming more familiar with collections as data use. During her presentation, Harriett Green introduced *Digging Deeper, Reaching Further: Libraries Empowering Users to Mine the HathiTrust Digital Library Resources*. This project aims to help librarians expand their role from “content provider” to partner in data driven research. With a three-year award from the Institute of Museum and Library Services, Green and her team are developing a program that will teach librarians how to work with textual data using text mining methods and tools. The aim of this work is to empower librarians with knowledge of data and computational methods that will help to position them as partners with researchers in exploring the potential that collections as data provide. While Green’s project is focused on use of HathiTrust, its utilization of data and infrastructure can inform similar efforts at other institutions.

With the *Synergies among Digital Humanities and African American History and Culture (AADHum)* initiative, Trevor Muñoz presented a nuanced view on developing community that can deeply inform development of legible data. Muñoz cautioned against thinking about communities as narrow expert groups, familiar with certain methods, acculturated in such a way that they will find, learn, and collaborate with each other on collections as data projects. Instead, Muñoz advocated for thinking deeply about how to develop community in a manner that doesn’t ignore a multiplicity of community memberships - a campus community, a regional community, a State community. If communities are to be developed that make use of collections as data, diversity of identity and motivation must be considered in concert with the methods and tools commonly employed to work with collections as data. When Muñoz engages in community development he does not presume to know all of the ways a set of individuals might wish to engage data. In doing so he makes room for learning how to develop data that aligns with community need.

### ***Creativity***

Pursuing a collections as data imperative requires creative thinking. Creative thinking requires safety and protection. In order for this work to occur it will be necessary to create space administratively and programmatically that encourages a wide range of experimentation. These experiments should be grounded only by their commitment to serving as the means to empower library staff as they consider refocusing, redefining, and extending aspects of their roles and responsibilities to increase the resonance of their commitments. To the extent that it is possible, experimentation should be built in as a core institutional activity across a broad range of areas - reference services, collections, outreach, preservation, repository development, digitization, web archiving, metadata, digital initiatives - all have a role to play. While some experiments will be internal, it is essential that creativity extend to thinking about ways to build partnerships with communities outside of the library - to invite them in to experiment with collections as data and the evolving services that are developed to support their use.

Day 2 of *Collections as Data: Stewardship and Use Models to Enhance Access* generated recommendations meant to help guide the Library of Congress' initial collections as data experimentation. Recommendations were provided by individuals from institutions and organizations that include but are not limited to the Massachusetts Institute of Technology, Pinboard, the Office of Creative Research, the University of California Santa Barbara, National Endowment for the Humanities, Amherst College, University of Illinois at Urbana Champaign, Yale University, the Smithsonian Center for Learning and Digital Access, University of Maryland College Park, Zooniverse, WGBH Media & Archives, the Mellon Foundation, and the Schlesinger Library at Harvard University. While specifically geared toward surfacing opportunities for the Library of Congress, aspects of the recommendations can be taken up by other institutions.

Ethically grounded approaches to creating, acquiring, developing, and using data ran throughout *Collections as Data: Stewardship and Use Models to Enhance Access*. In his talk on day 1, Jer Thorp asked what it was like “to live in data”. Thorp illustrated a combination of systems, infrastructure, and cultivated social norms that produce readymade versions of the self from data individuals produce. Often unwitting data producers are in turn impacted by these representations. Thorp provided concrete examples: erroneous determinations of community depression in a high school based on tweets and explorations into personhood as constituted by targeted advertisements. Thorp also articulated alternative possibilities for “living in data” that foreground individual agency. With an app called *Floodwatch*, individuals can track how advertisers are targeting them and choose whether or not to contribute that data for research purposes. In his talk, Maciej Ceglowski warned against overdependence on algorithmic approaches to working with data. For certain purposes Ceglowski pointed to the benefits these approaches could generate for working through digital collections to identify objects in images. Yet he warned against their potential to become “money laundering for bias” without sustained and nuanced human intervention. Melissa Parham argued that every act of data collection is an act of erasure – either implied by what is collected or not collected, or explicit via collection normalization.

Where data ethics are foundational to collections as data, a turn toward beginnings is possible.<sup>1</sup> Three high-level recommendations guide an initial course – (1) form data partnerships, (2) foster data engagement, (3) iterate toward providing collections as data. These recommendations are derived from Day 2 and informed by presentations on Day 1. Data partnerships are used to broaden and deepen

---

<sup>1</sup> *A natural first recommendation could chart engagement with the ethical dimensions of collections as data work. A focus of this kind would be problematic. The problem can be illuminated by thinking on efforts that bear similar tensions. For example, much has been written on the perils of prioritizing diversity by programmatically segmenting it, as a step in hiring, or in track at an academic conference. Segmentation appears to be a marker of recognizing a problem and according a space for that problem to be discussed under host endorsement. Yet the terms of engagement presented by the framing undermine the goal. By “according space”, this kind of approach often has the effect of moving conversations to their “appropriate place”, an isolated room in a conference venue, or in organizational terms rendered as a check box to be ticked in the process of recruitment. While often well-meaning, segmentation is a poor substitute for translating the problem into a set of values that are systematically incorporated and materially implemented throughout the culture of an entire organization. In collections as data work, data ethics should aspire to this latter form of integration rather than relegation to a type of activity to be completed. This kind of work is never finished.*

collections as data work. Fostering data engagement is a development and outreach activity. Iterating toward providing collections as data celebrates and endorses an inherently experimental endeavor.

### *Form Data Partnerships*

Data partnerships support efforts to broaden and deepen collections as data work. Partnerships should be internal as well as external. Internal partnerships enable sustained library staff exploration of primary as well as ancillary work that holds potential to drive collections as data forward. Internal partnerships should be intentionally designed to cross-pollinate library skills, dispositions, and community connections throughout an organization. Additional funding and release time should be provided to support these partnerships. Library staff secondment to any organization that stands to provide an experience that benefits collections as data work should be supported. External data partnerships are intended to draw a diversity of perspectives into the library. External candidates for partnership should be drawn from a wide pool of possible partners – journalists, activists, academics, technologists, librarians, archivists, and individuals organizing initiatives like Black Girls Code and PyLadies. Onsite as well as virtual residencies should be accommodated. Long and short term commitments should be supported. All labor should be compensated at or above market norms and shared publicly in traditional and nontraditional formats. Labs and/or focused units (e.g. research and development, digital scholarship) should be developed and resourced in such a way that they are able to effectively catalyze sociotechnical collections as data work. Labs and/or focused units should encourage collaborative effort in pursuing primary and ancillary collections as data research. Capacious scope for collections as data effort is meant to work against a perception of specialization that renders the endeavor irrelevant to other functional components of a library. In order to catalyze the work, labs and/or focused units need to provide onsite as well as virtual access to data and the infrastructure to engage in analysis of those data. Hosting more events akin to the Chronicling America data challenge provide a prime opportunity to encourage innovative use of data that informs improvement of collections, infrastructure, and associated services.

- Create internal fellowships that enable primary as well as ancillary exploration of collections as data
- Promote cross pollination of library skills, dispositions, and community connections
- Encourage secondment to any organization that stands to benefit collections as data work
- Form diverse external partnerships: journalists, activists, academics, technologists, librarians, archivists, and individuals organizing initiatives like Black Girls Code and PyLadies
- Initiate long/short term residencies
- Support onsite/virtual residencies
- Compensate residency and partnership labor at or above market norms
- Share product of partnerships in traditional and nontraditional formats
- Create labs and/or focused units (e.g. research and development, digital scholarship) that aim to catalyze sociotechnical collections as data research
- Host data challenges that encourage innovative use of data and inform improvement of collections, infrastructure, and associated services

### *Foster Data Engagement*

Data engagement fosters development of methods, software, and infrastructure that encourage collections as data use. Libraries should foster development of methodological approaches to computational analysis of data. Work akin to Elizabeth Lorang's efforts to leverage image recognition techniques to identify and extract poetic content in digitized historic newspaper pages provides an example worthy of emulation. Libraries should foster development of open source software to enable computational analysis of data. In this effort libraries can be informed by existing engagement with open source development manifest in projects like Hydra. Libraries should think about ways to extend and/or create new infrastructure and services to facilitate computational analysis of collections as data. In this work libraries can be informed by the Hathitrust Research Center, though solutions in this space are needed that map to a range of local compute, staffing, and financial resources. In an academic context this effort may entail exploration of cloud compute support and/or approaches to developing services based on stable rather than one-off utilization of campus computing clusters. Across all of these efforts it will be important to maintain a focus on lowering barriers to use. Understanding how to lower barriers to use is predicated on knowledge of community need. Reaching this understanding can be arrived at in part through collaboration with library staff that have direct contact with communities that use collections and/or are likely to use collections with greater frequency in the near future (e.g. historian use of web archives). Understanding of use is crucially reinforced by sustained outreach activity that aims to learn from and invest in a range of perspectives on collections as data.

- Foster development of methodological approaches to computational analysis of data
- Foster development of open source software to enable computational analysis of data
- Foster development of infrastructure and services to support computational analysis of data
- Focus on lowering barrier of entry for working with collections as data
- Work with library staff that have direct contact with communities that use collections to inform collections as data work (e.g. subject experts)
- Engage in sustained outreach activity that aims to learn from and invest in a range of perspectives on collections as data

### *Iterate Toward Providing Collections as Data*

Iterating toward providing collections as data celebrates and endorses an inherently experimental endeavor. No consensus for this work currently exists. Solutions are contingent on community need, infrastructure, staffing, and financial resources. Variation in availability of resources presents the conditions for a range of solutions to arise. Broadly, the Library of Congress forum suggests focus on collections as data access, form, and quality.

**Access** entails development of mechanisms that provision collections as data. Generally, collections as data access mechanisms seek to enable "bulk" downloading of digital collections. In a basic implementation they provide the ability to download multiple collections, a single collection, or components of collections. Provisioning this kind of access is a relatively low barrier implementation. Libraries can readily provide links to zipped collections or expose static collection directories that can be accessed using open source programs like wget or Rsync. More advanced implementations provide the

ability to parametrically scope downloads by the metadata ascribed to collections and/or features of the data themselves. This type of ability can be served through development of an application programming interface (API). Provisioning access via API should be informed by steps peers have taken to encourage use of APIs by users with a range of technical expertise. For example, Europeana and the University of British Columbia provide access to an API query builder that helps users become familiar with how to request data from an API.

- Provision access to data via zipped collections and static collection directories
- Provision access to data via application programming interface (API)
- Develop access mechanisms that accommodate a range of technical expertise

**Form** entails development of collections as data that represent a span of content types instantiated in formats that are likely to be usable. Predominantly, users interested in computational work with digital library collections are constrained by infrastructure and services focused on provisioning text data. Moving forward, data type availability should be expanded through additional investment in web archives, social media, moving image, image, audio, and geospatial collections. As this portfolio expands it will be important to consider which formats data should be made available in, in light of ready compatibility with commonly used methods and tools.

- Expand availability of non-textual collections as data
- Determine data format requirements of commonly used methods and tools
- Provide data in a variety of formats

**Quality** entails development of collections as data that can readily support claims that are made with them. Objections to claims predicated on analysis of text data in digital library collections can often be traced to concerns about the relative fidelity of plain text data generated via optical character recognition (OCR) to paper page. Libraries should continue to think creatively about how they might improve the quality of data generated by processes of this kind. Improvements can be gained via in-house library research and development. Improvements can also be encouraged through collaborations with funding agencies that aim to support national data quality improvement competitions. Where possible transnational partnerships should be explored. In particular, partnering with efforts in the United Kingdom and the European Union to improve Handwritten Text Recognition (HTR). Opportunities for improving quality should also be explored via crowdsourcing partnerships with organizations like the Smithsonian and Zooniverse. For example, the Smithsonian could help a larger community by circulating non-Smithsonian collections into their transcription center platform.

- Invest in research and development that improves quality of data (e.g. improve OCR)
- Partner with funding agencies to support national data quality improvement competitions
- Engage in transnational partnerships to improve data quality generated by experimental methods (e.g. partnering with organizations in the United Kingdom and the European Union on Handwritten Text Recognition (HTR) technology)
- Work in collaboration with organizations like the Smithsonian and Zooniverse to enable crowdsourced enhancement of collections

This paper advanced a series of frames and recommendations to help guide the Library of Congress and a broader library community as they consider how and where they might contribute to a collections as data imperative. Collections as data experimentation should be taken on in the truest sense possible. Failures weighed with consideration trace paths to discovery. While libraries can draw strength from tradition, normative assumptions underlying the way work is approached will need to be critically and continuously reexamined. Admittedly, the ground that lies ahead is uncertain, yet uncertainty is not the same as knowing nothing - uncertainty indicates the presence of alternatives.

## Further Reading

- Arroyo-Ramirez, Elvia. "Invisible Defaults and Perceived Limitations: Processing the Juan Gelman Files," October 30, 2016. <https://medium.com/on-archivy/invisible-defaults-and-perceived-limitations-processing-the-juan-gelman-files-4187fdd36759>.
- Bowker, Geoffrey C., and Susan Leigh Star. *Sorting Things out: Classification and Its Consequences*. 1. paperback ed. Inside Technology. Cambridge, Mass.: MIT Press, 2000.
- boyd, dana, and Kate Crawford. "Six Provocations for Big Data." SSRN Scholarly Paper. Rochester, NY: Social Science Research Network, September 21, 2011. <https://papers.ssrn.com/abstract=1926431>.
- boyd, danah, Emily Keller, and Bonnie Tijjerina. "Supporting Ethical Data Research: An Exploratory Study of Emerging Issues in Big Data and Technical Research." *Data & Society*, September 19, 2016. <http://datasociety.net/output/supporting-ethical-data-research-an-exploratory-study-of-emerging-issues-in-big-data-and-technical-research/>.
- Burrows, Toby. "Sharing Humanities Data for E-Research: Conceptual and Technical Issues." Custom Book Centre, 2011. <http://ses.library.usyd.edu.au:80/handle/2123/7938>.
- Carlson, Jake, and Lisa Johnston. *Data Information Literacy: Librarians, Data, and the Education of a New Generation of Researchers*, 2015. <http://public.eblib.com/choice/PublicFullRecord.aspx?p=2039088>.
- Ceglowski, Maciej. "Deep-Fried Data." Accessed November 21, 2016. [http://idlewords.com/talks/deep\\_fried\\_data.htm](http://idlewords.com/talks/deep_fried_data.htm).
- Faniel, Ixchel M., and Ann Zimmerman. "Beyond the Data Deluge: A Research Agenda for Large-Scale Data Sharing and Reuse." *International Journal of Digital Curation* 6, no. 1 (August 3, 2011): 58–69. doi:10.2218/ijdc.v6i1.172.
- Gibbs, Fred, and Owens, Trevor. "The Hermeneutics of Data and Historical Writing." In *Writing History in the Digital Age*, edited by Nawrotzki, Kristen and Dougherty, Jack. Digital Humanities. Ann Arbor, MI: University of Michigan Press, 2013.
- Gitelman, Lisa, ed. *"Raw Data" Is an Oxymoron*. Infrastructures Series. Cambridge, Massachusetts: The MIT Press, 2013.
- Jules, Bergis. "Surveillance and Social Media Archiving." *Documenting DocNow*, October 4, 2016. <https://news.docnow.io/surveillance-and-social-media-archiving-7ea21b77b807>.
- Ibid. "Confronting Our Failure of Care Around the Legacies of Marginalized People in the Archives," November 11, 2016, <https://medium.com/on-archivy/confronting-our-failure-of-care-around-the-legacies-of-marginalized-people-in-the-archives-dc4180397280>.
- Leonard, Peter. "Mining Large Datasets for the Humanities." Lyon, France, 2014. <http://library.ifla.org/930/>.
- Manus, Susan. "Digging Up the Recent Past: An Interview With Doug Reside." *The Signal: Digital Preservation*, n.d. <http://blogs.loc.gov/digitalpreservation/2012/08/digging-up-the-recent-past-an-interview-with-doug-reside/>.

- Mattern, Shannon. "Public In/Formation." *Places Journal*, November 15, 2016. doi:10.22269/161115.
- Maxwell, Andrew, and Ugly Duckling Presse. *Candor Is the Brightest Shield*, 2014.
- Mueller, Martin. "Morgenstern's Spectacles or the Importance of Not-Reading | Scalable Reading." Accessed December 29, 2014. <https://scalablereading.northwestern.edu/2013/01/21/morgensterns-spectacles-or-the-importance-of-not-reading/>.
- Noble, Safiya Umoja. 2012. Missed Connections: What Search Engines Say About Women. *Bitch Magazine*, 54.
- Ibid. "Google Search: Hyper-Visibility as a Means of Rendering Black Women and Girls Invisible." *InVisible Culture: An Electronic Journal for Visual Culture*, no. 19. Accessed January 17, 2017. <http://ivc.lib.rochester.edu/google-search-hyper-visibility-as-a-means-of-rendering-black-women-and-girls-invisible/>.
- Nowviskie, Bethany. "A Skunk in the Library." *Bethany Nowviskie*, June 28, 2011. <http://nowviskie.org/2011/a-skunk-in-the-library/>.
- Ibid. "Speculative Collections," *Bethany Nowviskie*, October 27, 2016, <http://nowviskie.org/2016/speculative-collections/>.
- Padilla, Thomas. "Humanities Data in the Library: Integrity, Form, Access." *D-Lib Magazine* 22, no. 3/4 (March 2016). doi:10.1045/march2016-padilla.
- Parrish, Allison. "Programming Is Forgetting: Toward a New Hacker Ethic." *Open Transcripts*. <http://opentranscripts.org/transcript/programming-forgetting-new-hacker-ethic/>.
- Rawson, Katie, and Trevor Munoz. "Against Cleaning," July 6, 2016.
- Risam, Roopika. "Beyond the Margins: Intersectionality and the Digital Humanities" 9, no. 2 <http://www.digitalhumanities.org/dhq/vol/9/2/000208/000208.html>.
- Sadler, Bess, and Chris Bourg. "Feminism and the Future of Library Discovery." *Code4lib Journal*, no. 28 (April 15, 2015). <http://journal.code4lib.org/articles/10425>.
- Stodden, Victoria, and Sheila Miguez. "Best Practices for Computational Science: Software Infrastructure and Environments for Reproducible and Extensible Research." *Journal of Open Research Software* 2, no. 1 (July 9, 2014). doi:10.5334/jors.ay.
- Ibid., M. McNutt, D. H. Bailey, E. Deelman, Y. Gil, B. Hanson, M. A. Heroux, J. P. A. Ioannidis, and M. Taufer. "Enhancing Reproducibility for Computational Methods." *Science* 354, no. 6317 (December 9, 2016): 1240–41. doi:10.1126/science.aah6168.
- Verhoeven, D. "As Luck Would Have It: Serendipity and Solace in Digital Research Infrastructure." *Feminist Media Histories* 2, no. 1 (January 1, 2016): 7–28. doi:10.1525/fmh.2016.2.1.7.
- Zwaard, Kate. "NDI Talk at Collections as Data | The Signal." Webpage, October 12, 2016. [//blogs.loc.gov/thesignal/2016/10/ndi-talk-at-collections-as-data/](http://blogs.loc.gov/thesignal/2016/10/ndi-talk-at-collections-as-data/).