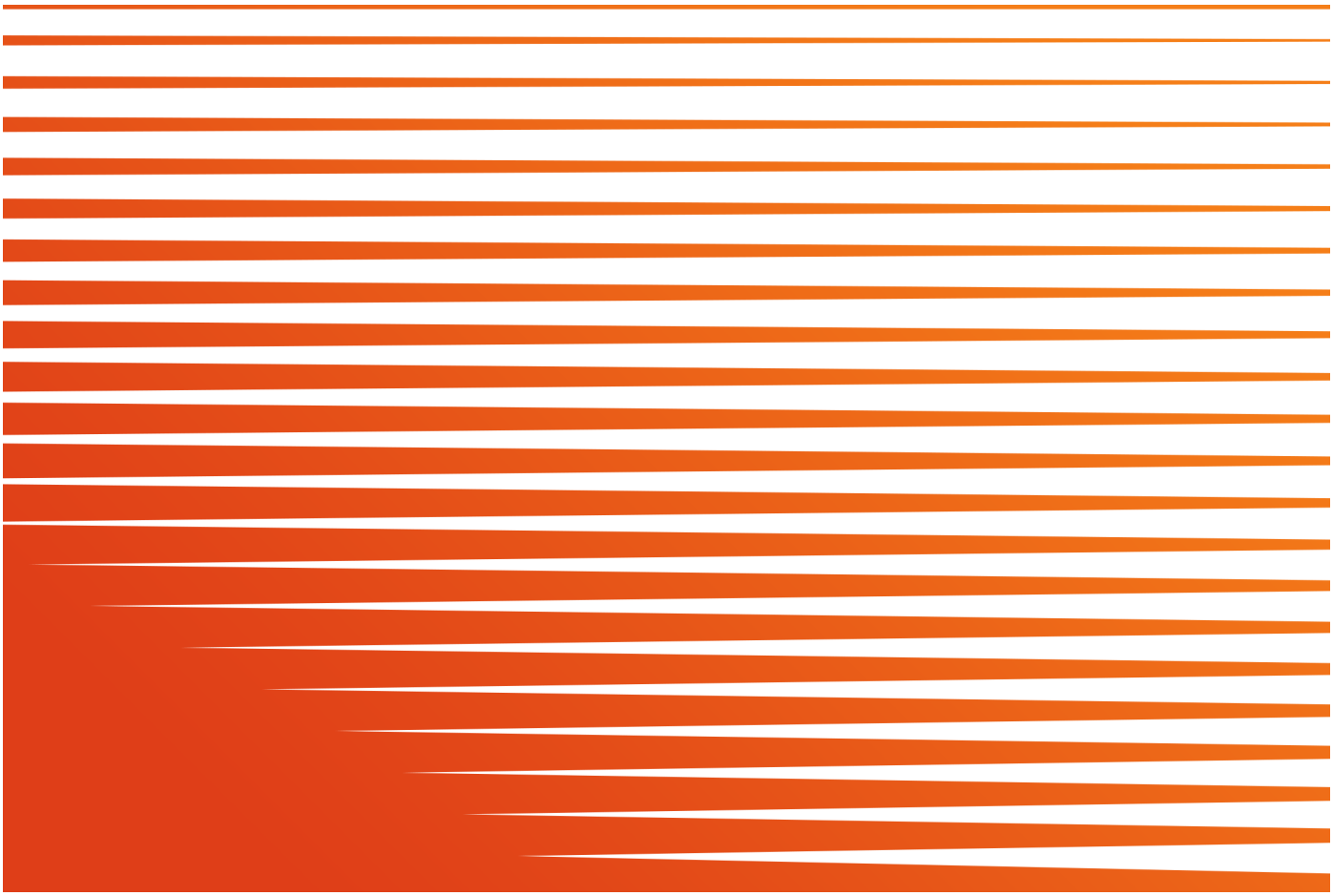
Are the Humanities Ready for Data Sharing?

Dylan Ruediger
Ruby MacDougall



ITHAKA S+R

Ithaka S+R provides research and strategic guidance to help the academic and cultural communities serve the public good and navigate economic, demographic, and technological change. Ithaka S+R is part of ITHAKA, a not-for-profit with a mission to improve access to knowledge and education for people around the world. We believe education is key to the wellbeing of individuals and society, and we work to make it more effective and affordable.

We are interested in disseminating this brief as widely as possible. Please contact us with any questions about using the report: research@ithaka.org.

Introduction

On January 11, the White House Office of Science and Technology Policy (OSTP) declared 2023 the "year of open science" and announced that federal agencies will spend the year promoting public access to research publications and data. While primarily focused on the hard sciences, the January 11 announcement includes a bullet point from the National Endowment for the Humanities, which emphasizes the NEH's commitment to "addressing issues of accessibility and usability, and designing equitable, open, replicable, and sustainable" research in the humanities.[1]

The announcement reiterates the Biden administration's commitment to leveraging federal funding to foster open sharing of data and other research outputs and builds on last fall's OSTP memorandum revising federal regulations regarding public access to the results of federally funded research. Signed by then-OSTP interim director Alondra Nelson, the "Nelson memo" requires all publications and supporting data produced with federal funds be made freely and publicly available without an embargo period and points towards future mandates that would require all data generated with federal funds (not just data associated with publications) to be made public. The Nelson memo is not the first federal policy to address data sharing and open access, but it is the first to apply to not only large funders such as the NSF and NIH, but to smaller ones such as the NEH. While the NEH funds only a tiny percentage of research and publications in the humanities, its inclusion in the Nelson memo and in the "year of open science" is clear evidence that humanists—who have largely existed on the margins of major trends towards mandatory data sharing that are transforming research practices and scholarly communication in other fields—must now consider their place in this policy landscape.[2]

> Humanists—who have largely existed on the margins of major trends towards mandatory data sharing that are transforming research practices and scholarly communication in other fields—must now consider their place in this policy landscape.

It is not yet clear how the NEH will define data for the purposes of compliance with the Nelson memo, but the requirement that they do so should stimulate conversation about data sharing in the humanities. When should the evidence humanists collect be considered data? How might humanists adopt STEM-oriented norms around data sharing, and what might humanists bring to the table that would help other fields improve their data sharing practices?

[1] "FACT SHEET: Biden-Harris Administration Announces New Actions to Advance Open and Equitable Research - OSTP," The White House, 11 January 2023, https://www.whitehouse.gov/ostp/news-updates/2023/01/11/fact-sheet-biden-harris-administration-announces-new-actions-to-advance-open-and-equitable-research/.

[2] For reactions to the Nelson memo that include consideration of the humanities, see: Ann Michael Clarke Todd A. Carpenter, Angela Cochran, Lisa Janicke Hinchliffe, Karin Wulf, Michael, "Ask The Chefs: OSTP Policy Part II," *The Scholarly Kitchen*, 31 August 2022, https://scholarlykitchen.sspnet.org/2022/08/31/ask-the-chefs-ostp-policy-ii/; Dylan Ruediger, "The Outlook for Data Sharing in Light of the Nelson Memo," *The Scholarly Kitchen*, 8 September 2022, https://scholarlykitchen.sspnet.org/2022/09/06/guest-post-the-outlook-for-data-sharing-in-light-of-the-nelson-memo/.

With the notable exception of digital humanists (still marginal and siloed off from the field as a whole), these questions will be unfamiliar to many humanists, few of whom have incorporated data sharing into their research and publication workflows. Ithaka S+R's 2021 national faculty survey found that fewer than 20 percent of humanists report sharing or depositing research data or datasets as a regular part of their research practice, a number that lags well behind researchers in other fields.[3] Even among participants at a forum on data sharing hosted by the *Journal of Open Humanities Data*—presumably a group with well above average interest in data sharing—only 47 percent of respondents had previously deposited research data into a repository.[4]

Why is data sharing so rare in the humanities?

Organizations such as the Research Data Alliance (RDA), European Federations of Academies of Sciences and Humanities (ALLEA), and Digital Research Infrastructure for the Arts and Humanities (DARIAH) have examined the financial and technical barriers to data sharing in the humanities.[5] Notable among these barriers is that the dispersed network of domain repositories that the NSF and NIH, for example, have funded, has no analog in the humanities. As we have documented over the past several years, domain repositories play a critical role in facilitating data sharing and reuse and often serve as hubs for "data communities"—fluid, interdisciplinary, networks of scholars with overlapping research interests. These data communities play an essential role in fostering vigorous, voluntary sharing and reuse of research data in the sciences.[6]

Cultural barriers play a significant role in limiting data sharing between humanists as well. Some of these, such as promotion and tenure standards that prioritize publication over data deposit, are familiar across academic disciplines. Humanists are more distinctive for framing their research as based on sources rather than data, which can make the entire concept of "data sharing" seem alien.[7] Digital humanists—who make up only a small fraction of humanities

[3] Melissa Blankstein, "Ithaka S+R US Faculty Survey 2021," *Ithaka S+R*, 14 July 2022, https://doi.org/10.18665/sr.316896.

[4] Journal of Open Humanities Data, *Q&A*, YouTube, 16 December 2020, https://www.youtube.com/watch?v=UyyYS1dNXZ4.

[5] Natalie Harrower, Maciej Maryl, Timea Biro, Beat Immenhauser, & ALLEA Working Group E-Humanities, "Sustainable and FAIR Data Sharing in the Humanities: Recommendations of the ALLEA Working Group E-Humanities," Digital Repository of Ireland, https://doi.org/10.7486/DRI.tq582c863; Lindsay Porier, "Scales of Data Sharing Challenges in the Digital Humanities," https://www.rd-alliance.org/system/files/documents/PosterPresentations.com-70CMx100CM-Pro.pdf; René van Horik, "The Research Data Alliance and the Humanities," Zenodo, July 2019, https://doi.org/10.5281/zenodo.3355145; Toma Tasovac, Sally Chambers, Erzsébet Tóth-Czifra, "Cultural Heritage Data from a Humanities Research Perspective: A DARIAH Position Paper," Digital Research Infrastructure for the Arts and Humanities, 2020, https://hal.science/hal-02961317/document.

[6] Dylan Ruediger, Ruby MacDougall, Danielle Miriam Cooper, Jake Carlson, Joel Herndon, and Lisa Johnston, "Leveraging Data Communities to Advance Open Science: Findings from an Incubation Workshop Series," *Ithaka S+R*, 9 August 2022, https://doi.org/10.18665/sr.317145; Danielle Miriam Cooper and Rebecca Springer, "Data Communities: A New Model for Supporting STEM Data Sharing," *Ithaka S+R*, 13 May 2019, https://doi.org/10.18665/sr.311396.

[7] Danielle Miriam Cooper, Roger C. Schonfeld, Richard Adams, Matthew Baker, Nisa Bakkalbasi, John G. Bales, Rebekah Bedard, et al., "Supporting the Changing Research Practices of Religious Studies Scholars," *Ithaka S+R*, 8 February 2017, https://doi.org/10.18665/sr.294119; Danielle Miriam Cooper, Cate Mahoney, Rebecca Springer, Robert Behra, Ian G. Beilin, Guy Burak, Margaret Burri, et al., "Supporting Research in Languages and Literature," *Ithaka S+R*, 9 September 2020, https://doi.org/10.18665/sr.313810; Erzsébet Tóth-Czifra, "The Risk of Losing the Thick Description," in *Digital Technology and the Practices of Humanities Research*, ed. Jennifer Edmund (Open Book Publishers, 2020), https://www.openbookpublishers.com/product/1108; Jennifer L. Thoegersen, "'Yeah, I Guess That's Data': Data Practices and

researchers and whose methods are viewed with ambiguity and even hostility from some more traditionally oriented colleagues—are more likely to engage with structured data in the sense familiar to scientists than their colleagues and have made significant investments in identifying best practices for data sharing in the humanities.[8] However, even among digital humanists, there is no clear consensus on what data sharing means, how to implement it, or how to define its success.

To get a sense of trends in data sharing within the humanities, we conducted semi-structured interviews with key personnel at several humanities projects with strong data components. The interviews focused on identifying where and how they planned to share their research data, how they imagined it might be used by others, and their perspective on barriers and opportunities to data sharing in the humanities. The research agendas, skills, and perspectives of the people we spoke with are not representative of most humanities-oriented research. However, the interviews provide important insight into the thinking of humanists who are already working across the cultural divide around data that separate the humanities from most other academic disciplines. We use them here as a springboard for consideration of what humanities data is, how to access and preserve it, and how it fits into the larger goals of creating an open research culture.

> One key perspective that humanists can bring to larger debates about data sharing and open access research outputs is their uniquely well-developed infrastructure for the public sharing of knowledge creation.

This issue brief will ultimately suggest that one key perspective that humanists can bring to larger debates about data sharing and open access research outputs is their uniquely well-developed infrastructure for the public sharing of knowledge creation, exemplified in the many public humanities initiatives that are a highly visible and vibrant part of humanities scholarship.[9] The formats, methods, and goals of public humanities initiatives vary widely, but

---

Conceptions among Humanities Faculty," *Portal: Libraries and the Academy* 18, no. 3 (2018): 491–504, https://doi.org/10.1353/pla.2018.0030; Jon Treadway, Mark Hahnel, Sabina Leonelli, Dan Penny, David Groenewegen, Nobuko Miyairi, Kazuhiro Hayashi, Daniel O'Donnell, Digital Science, and Daniel Hook, "The State of Open Data Report," *Digital Science*, 25 October 2016, https://doi.org/10.6084/m9.figshare.4036398.v1; Christof Schöch, "Big? Smart? Clean? Messy? Data in the Humanities," *Journal of Digital Humanities* 2, no. 3 (Summer 2013), 2-13, http://journalofdigitalhumanities.org/2-3/big-smart-clean-messy-data-in-the-humanities/.

[8] Shane Hawkins, ed., *Access and Control in Digital Humanities* (New York: Routledge, 2021); Matthew K. Gold and Lauren F. Klein, "Digital Humanities: The Expanded Field" in *Debates in the Digital Humanities 2016,* eds. Matthew K. Gold and Lauren F. Klein (Minneapolis: University of Minnesota Press, 2016), https://dhdebates.gc.cuny.edu/read/untitled/section/14b686b2-bdda-417f-b603-96ae8fbbfd0f; Barbara McGillivray et al., "The Challenges and Prospects of the Intersection of Humanities and Data Science: A White Paper from The Alan Turing Institute," *The Alan Turing Institute*, 2020, https://doi.org/10.6084/M9.FIGSHARE.12732164. For perceptions of digital humanists, see Danielle Miriam Cooper, Cate Mahoney, Rebecca Springer, Robert Behra, Ian G. Beilin, Guy Burak, Margaret Burri, et al., "Supporting Research in Languages and Literature," *Ithaka S+R*, 9 September 2020, https://doi.org/10.18665/sr.313810.

[9] Public humanities projects focus on building bridges between academic humanities scholars and the broader public. The National Humanities Alliance (NHA) has compiled a comprehensive database of more than 2,000 public humanities projects as part of their Humanities for All program. This database provides valuable insight into the broad scope and wide reach of public humanities projects. The NHA has also documented the impact of public humanities projects in its Community Case Studies initiative, a project

many involve collaboration between academics, students, and diverse community partners.[10] Historically, the public humanities have been distinguished by their focus on placing "place communities, or other public audiences, at their core."[11] Many recent public humanities projects emphasize community-driven, collaborative data generation efforts, in which knowledge is co-created *with* community participants not *for* the community.[12] These models of community generated data and co-created knowledge production can help inform and shape dynamic, equitable, and process-oriented data sharing across the academic sector. They may also hold the key to articulating an ethics of reuse within the humanities that is also valuable to researchers in other fields and point the way towards a data humanism.

Interviews

We conducted interviews with the following individuals representing three ongoing projects and one completed high profile digital humanities project. We are grateful for their generosity and insights.

Amanda Regan of Mapping the Gay Guides

Mapping the Gay Guides (https://www.mappingthegayguides.org/), launched in 2020, geocodes map locations listed in Bob Damron Address Book, a guide to gay bars and queer friendly spaces in twentieth century America, in an effort to understand queer geographies. The project began with an initial dataset of locations from the Damron Guides from 1965-1980 that were focused on the Southeastern United States. It continued to transcribe and map locations from across the US from 1965-1980 and has recently received funding from the National Endowment for the Humanities to digitize, transcribe, and geolocate the Damron Guides from 1981-2000. By associating geographical coordinates with each location mentioned within the Damron Guides, MGG provides an interface for visualizing the growth of queer spaces between 1965 and 1980 (eventually 2000). Directed by Dr. Amanda Regan and Dr. Eric Gonzaba, the project works with a team of graduate student researchers and interns to digitize and process data from the Damron Guides. The project's advisory board includes faculty members, librarians, and the president of Damron Books. The project's website includes historical vignettes and analysis derived from the data and written by student researchers, and an interactive map. Data for the project, and some associated code, is available on GitHub.

---

that showcases humanities work taking place within three distinct communities and the variety of issues with which these projects engage. For more on Humanities for All, see: https://humanitiesforall.org/; for more on Community Case Studies, see: https://www.nhalliance.org/community_case_studies.

[10] Daniel Fisher, "A Typology of the Publicly Engaged Humanities," Humanities for All, accessed 2 May 2022, https://humanitiesforall.org/essays/five-types-of-publicly-engaged-humanities-work-in-u-s-higher-education.

[11] Sheila Brennan, "Public, First," in *Debates in the Digital Humanities*, eds. Matthew K. Gold and Lauren F. Klein (Minneapolis: University of Minnesota Press, 2016), https://dhdebates.gc.cuny.edu/read/untitled/section/11b9805a-a8e0-42e3-9a1c-fad46e4b78e5.

[12] Susan Smulyan, ed., *Doing Public Humanities* (Routledge, 2020), https://www.routledge.com/Doing-Public-Humanities/Smulyan/p/book/9780367500177. In STEM fields, citizen science projects share similar goals.

Matt Jansen of On the Books: Jim Crow and the Algorithms of Resistance

On the Books: Jim Crow and the Algorithms of Resistance (https://onthebooks.lib.unc.edu/) started in 2017 as a reference question to the UNC Chapel Hill library from a high school teacher about Jim Crow laws and led to building a "collections as data" project. It uses text mining and machine learning to explore the language of Jim Crow and racially based legislation signed into law in North Carolina between Reconstruction and the Civil Rights Movement (1866-1967). Taking inspiration from Safiya Noble's *Algorithms of Oppression: How Search Engines Reinforce Racism*,[13] On the Books coined the phrase "algorithms of resistance" to describe its methods to identify racist language in legal documents, thereby helping expose the wide-ranging effects of Jim/Jane Crow on the American South. On the Books produces free K-12 lesson plans and educational resources for how to engage with their dataset. The interdisciplinary project team, largely based in the UNC library, includes experts in African American history, special collections, digital research, data analysis, data visualization, graduate student research assistants, and student library workers. Like Mapping the Gay Guides, On the Books makes its data, code, and Open Educational Resources page open and available on GitHub. On the Books has been funded by the Mellon Foundation as part of the first cohort for Collections as Data: Part to Whole and the ARL Venture Fund. We interviewed Matt Jansen, co-PI and technical lead of On the Books, and a data analysis librarian at UNC Chapel Hill.

Sayeed Choudhury of Black Beyond Data

Black Beyond Data: Computational Humanities and Social Sciences Laboratory for Black Digital Humanities is a collaborative, interdisciplinary project based at Johns Hopkins University that brings together three data-intensive humanities projects: Life X Code: DH against enclosure, The Black Press Research Collective, and the Risk and Racism Data Project. In 2021, Black Beyond Data received a Mellon Foundation grant that built upon a previous planning grant, and since its inception, the project has been developing data literacy resources, hosting reading groups, and growing a critical mass of scholars, teachers, students, and community members to use digital humanities data and methods against racial injustice by highlighting the stories of the African American community through their own voices. Connecting the fields of digital humanities, Black studies, and data and computation, Black Beyond Data is fostering collaborations across institutions, disciplines, and organizations with the goal of using innovative methods and technologies to better understand how data and visualizations can be analyzed and used to center Black humanity. We interviewed Sayeed Choudhury, now at Carnegie Mellon University, who continues to contribute to the project in his current capacity and previously contributed as the associate dean for research data management and director of the Digital Research and Curation Center at Johns Hopkins University.

[13] Safiya Umoja Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism* (New York: NYU Press, February 2018), https://nyupress.org/9781479837243/algorithms-of-oppression/.

Jessica Otis of Six Degrees of Francis Bacon

Six Degrees of Francis Bacon (http://www.sixdegreesoffrancisbacon.com/) is a digital reconstruction of early modern English social networks based on text mining 62 million words in the Oxford Dictionary of National Biography. To demonstrate how linked connections spread ideas and knowledge in a visual way, the project created a website that traces the social networks of figures like Bacon, Shakespeare, Ann Boleyn, and more than 6,000 others. With an underlying infrastructure that supports crowdsourcing, the website allows professional researchers, students, and amateurs from all over the world to view and contribute to the project by collaboratively expanding, revising, curating, and critiquing the data. The project's data is available for download on the Six Degrees of Francis Bacon website and archived at the Folger Shakespeare Library. Launched in 2013, and the oldest of the projects featured in this report, Six Degrees of Francis Bacon was supported by grants from NEH, Google, and Carnegie Mellon University among others. We interviewed Jessica Otis, co-PI of the project, and Assistant Professor of History and the Director of Public Projects at the Roy Rosenzweig Center for History and New Media.

What is humanities data?

While the small but growing number of humanists who work with structured data often self-publish their datasets on project websites, they have little expectation that the data itself will be widely used or cited. There are important exceptions to this trend such as the Slave Voyages Database[14], but even very high-profile digital projects such as Six Degrees of Francis Bacon have seen little use of their underlying data, as most users interact with the project's visualizations instead. Some projects, including On the Books, are experimenting with offering small research grants to incentivize reuse, but all of our interviewees agreed that organic reuse of datasets is rare in the humanities. While the data infrastructure in the humanities is undeveloped—perhaps most decisively due to a lack of domain repositories—the biggest hurdle to data sharing in the humanities may be a systematic lack of understanding about how to reuse structured data as a primary source. Within theories of open research, the value of data sharing is rooted in the transparency it brings to the research process by allowing validation and reproducibility and the discoveries enabled by reuse of deposited data.[15] What validation and reproducibility means in most humanities contexts, which depend on interpretative rather than experimental methods, is unclear.[16] And thus far, as several of our interviewees pointed out, few humanists have the skills to ask new research questions of existing datasets.

[14] "Explore the Origins and Forced Relocations of Enslaved Africans Across the Atlantic World," Slave Voyages, Rice University, https://www.slavevoyages.org/.

[15] Our use of the term "open research" is deliberate. As Paul Arthur and Lydia Hearn have pointed out, the more commonly used term "open science" is but one part of the broader issue of "open research." As they point out, the term "open science" reflects the centrality of STEM fields within debates about open research practices. "Open research" creates space for the humanities and their distinct ways of knowing within these conversations. Paul Longley Arthur and Lydia Hearn, "Toward Open Research: A Narrative Review of the Challenges and Opportunities for Open Humanities," *Journal of Communication* 71, no. 5 (October 1, 2021): 827–53, https://doi.org/10.1093/joc/jqab028.

[16] Rik Peels, "Replicability and Replication in the Humanities," *Research Integrity and Peer Review* 4, no. 1 (January 9, 2019): 2, https://doi.org/10.1186/s41073-018-0060-4; https://www.nature.com/articles/d41586-018-05845-z; Sarah de Rijcke and Bart Penders, "Resist Calls for Replicability in the Humanities," *Nature*, 1 August 2018, https://www.tandfonline.com/doi/abs/10.1080/24694452.2020.1806024; Daniel Suit and Peter Kedron, "Reproducibility and

Given these circumstances, it is not surprising that, to date, there are no obvious examples of a mature data community in the humanities. One necessary step towards creating a data sharing culture in the humanities is to build consensus within research communities about the *purpose* of data sharing in their field. When and how do reproducibility and replicability matter in the humanities? Are those the appropriate conceptual frameworks for articulating the value of transparency and integrity in humanistic research? What is the relationship between reproducibility and reinterpretation? Likewise, what does reuse mean in the humanities? What skills and perspectives would humanists need to cultivate to build disciplinary capacities for vigorous reuse of data?

Lurking behind these questions is another that is perhaps even more fundamental. What is the threshold at which evidence becomes data? Most humanities scholars consider their research as based on sources rather than data, an assessment that is often a fair one.[17] Humanist methodologies and ways of knowing are rooted in their ability to contextualize cultural materials, an approach that often requires close attention to nuance and foregrounds complexity more than it seeks to resolve it. As Miriam Posner has described it, traditional humanist methodologies are not focused on "extracting features in order to analyze them." Instead, Posner notes, humanist interpretation is oriented towards "trying to dive into it, like a pool, and understand it from within."[18] Our interviewees agreed with Posner, noting that while sources are readily understood as reusable, few humanists use them to generate the abstracted, highly structured, and decontextualized datasets that increase their mobility and facilitate computation. Humanists seldom transform unstructured data into structured data in the ways familiar to researchers in other fields: indeed, many humanistic research questions would be hindered by such a transformation.

> Humanists seldom transform unstructured data into structured data in the ways familiar to researchers in other fields: indeed, many humanistic research questions would be hindered by such a transformation.

The Nelson memo fails to shed light on the ambiguities of the distinction between sources and data by repeatedly referring to data as "scientific data," providing little guidance to help the NEH define humanistic data. As we have seen, this will not be easy. The NEH should respect the

---

Replicability in the Context of the Contested Identities of Geography," *Annals of the American Association of Geographers* 111, no. 5 (13 October, 2020,) https://link.springer.com/article/10.1007/s13194-019-0269-1; Stephan Guttinger, "The Limits of Replicability," *European Journal for Philosophy of Science* 10, no. 10 (2020) https://muse.jhu.edu/article/698844; Grant Wythoff, ""On Method in the Humanities," *Configurations* 26, no. 3 (2018): 289-295, https://muse.jhu.edu/article/698844; Maria Rahal, Hanjo Hamann, Hilmar Brohmer, and Florian Pethig, "Sharing the Recipe: Reproducibility and Replicability in Research Across Disciplines," *Research Ideas and Outcomes* 8: e89980, https://doi.org/10.3897/rio.8.e89980.

[17] Jennifer L. Thoegersen, "'Yeah, I Guess That's Data': Data Practices and Conceptions among Humanities Faculty," *Portal: Libraries and the Academy* 18, no. 3 (2018): 491–504, https://doi.org/10.1353/pla.2018.0030; Stephen Marche, "Literature Is Not Data: Against Digital Humanities," *Los Angeles Review of Books*, 28 October 2012, https://lareviewofbooks.org/article/literature-is-not-data-against-digital-humanities/.

[18] Miriam Posner, "Humanities Data: A Necessary Contradiction," *Miriam Posner's Blog*, 25 June 2015, https://miriamposner.com/blog/humanities-data-a-necessary-contradiction/; Ruth Ahnert et al., *The Network Turn: Changing Perspectives in the Humanities* (Cambridge: Cambridge University Press, 2021), 51–52.

important distinction between sources and data, a distinction that is central to humanistic ways of knowing. The easiest path forward for the NEH would likely be to define data narrowly and decide that only quantitative materials and structured data will be subject to deposit requirements. Such an interpretation would reinforce the sense that data is largely an issue for digital humanists, in the process reinforcing the lines separating most humanistic scholarship from that of other disciplines. The NEH should use this opportunity to push humanists to think more expansively about the research materials and methodologies they use and nudge the humanities into better alignment with the main currents of academic research as a whole.

The Nelson memo provides the language to do so. It specifies that data includes "the recorded factual material commonly accepted in the scientific community as of sufficient quality to validate and replicate research findings," terms which do not treat data as a synonym for structured, quantitative, material. This definition could be read as including a wide range of material that humanists use to support their conclusions, such as interview transcripts, textual corpora, or archival records. Issues of copyright and privacy aside, there is no consensus within the humanities about which of these materials count as "data," nor articulate the conditions under which they might be considered as such. The Nelson memo provides a unique opportunity to push these issues into the spotlight and to help build consensus with humanities disciplines about what data means. These are hard but necessary conversations. If humanists do not articulate community norms and practices around data sharing and reuse, they risk having those norms set for them.

Data preservation and discoverability

In many important respects, the data challenges humanists face are similar to those researchers in any field face. These include a lack of time, expertise, and funding to fully integrate research data management across the lifecycle of a project as well as promotion and tenure standards that disincentivize rigorous data management and curation.[19] Prominent scholars have recently argued that because most data-intensive humanities projects rely on the labor of very small teams or even of single individuals, their work should be exempted from policies and requirements designed to ensure that project datasets can conform to high standards of reproducibility.[20] Clearly, humanities scholars face a challenging funding landscape, though the

[19] Dylan Ruediger and Danielle Miriam Cooper, "Big Data Infrastructure at the Crossroads: Support Needs and Challenges for Universities," *Ithaka S+R*, 1 December 2021, https://doi.org/10.18665/sr.316121; Dylan Ruediger, Ruby MacDougall, Danielle Miriam Cooper, Jake Carlson, Joel Herndon, and Lisa Johnston, "Leveraging Data Communities to Advance Open Science: Findings from an Incubation Workshop Series," *Ithaka S+R*, 9 August 2022, https://doi.org/10.18665/sr.317145; Carol Tenopir et al., "Data Sharing by Scientists: Practices and Perceptions," PLOS ONE 6, no. 6 (29 June 2011): e21101, https://doi.org/10.1371/journal.pone.0021101; Laia Pujol Priego, Jonathan Wareham, and Angelo Kenneth S. Romasanta, "The Puzzle of Sharing Scientific Data," *Industry and Innovation* 29, no. 2 (7 February 2022): 219–50, https://doi.org/10.1080/13662716.2022.2033178; Natasha J. Gownaris et al., "Barriers to Full Participation in the Open Science Life Cycle among Early Career Researchers," *Data Science Journal* 21, no. 1 (19 January 2022): 2, https://doi.org/10.5334/dsj-2022-002; Greg Tananbaum and Michael M. Crow, "We Must Tear Down the Barriers That Impede Scientific Progress," *Scientific American*, 18 December 2020, https://www.scientificamerican.com/article/we-must-tear-down-the-barriers-that-impede-scientific-progress/.

[20] Ruth Ahnert et al., *The Network Turn: Changing Perspectives in the Humanities* (Cambridge: Cambridge University Press, 2021), 97.

relatively small size of even large data projects in the humanities compared to data-intensive research in many other fields makes it difficult to generalize about funding levels relative to the total number of observations. Across disciplines, the costs of data curation and preservation are poorly understood, though forthcoming work from the Data Curation Network and ACRL promise to shed light on the topic.[21]

Humanists also face daunting skills gaps to conducting data-intensive research, which requires methods and tools that are outside the mainstream of humanistic research and training. One of the individuals we interviewed noted that this creates practical problems—they had observed that many humanities datasets were poorly structured, overly narrow in scope, and too often stored in file formats that discouraged reuse. Another interviewee mentioned that humanists often skimped on metadata, documentation, and methodologies for data collection, a "black box" problem that was exacerbated by a shortage of peer-reviewers qualified to assess the quality of "what is going on underneath the hood." Humanists on the whole, said several interviewees, are unfamiliar with quantitative methods. When they launch digital projects, they have a tendency to be simultaneously distrustful of numbers and too quick to take them at face value. Researchers just beginning to work with structured data would be wise to use the highly collaborative model of many digital humanities projects, which often include substantial contributions from programmers, data scientists, and especially librarians, as models.[22]

But perhaps the biggest infrastructure challenge is the lack of visible places to deposit data. Discoverability emerged throughout the interviews as a major barrier to data reuse. In the absence of domain repositories, many humanities datasets end up posted on individual websites. This leaves them highly vulnerable to disappearing as those websites become obsolete or are shut down entirely, a chronic problem in the digital humanities.[23] Moreover, the highly dispersed nature of individual project websites creates formidable discovery challenges. Supplemental deposit into institutional repositories or, as in the case of Six Degrees of Francis Bacon, with independent research libraries, mitigates preservation risks but does little to

[21] "DCN collaborates with Association of Research Libraries (ARL) on National Science Foundation Award," Data Curation Network, 22 July 2021, https://datacurationnetwork.org/2021/07/22/dcn-collaborates-with-association-of-research-libraries-arl-on-national-science-foundation-award/.

[22] Susan Hockey, "Digital Humanities in the Age of the Internet: Reaching Out to Other Communities," in *Collaborative Research in the Digital Humanities*, eds. Willard Mccarty and Marilyn Deegan (London: Routledge, May 2016) https://www.taylorfrancis.com/chapters/edit/10.4324/9781315572659-5/collaborative-research-digital-humanities-willard-mccarty; Gabriele Griffin and Matt Steven Hayler, "Collaboration in Digital Humanities Research – Persisting Silences," *Digital Humanities Quarterly* 12, no. 1 (2018), http://www.digitalhumanities.org/dhq/vol/12/1/000351/000351.html; Bethany Nowviskie, "Evaluating Collaborative Digital Scholarship (or, Where Credit is Due)," *Journal of Digital Humanities* 1, no.4 (Fall 2012) http://journalofdigitalhumanities.org/1-4/evaluating-collaborative-digital-scholarship-by-bethany-nowviskie/; Anne B. McGrail, Angel David Nieves, and Siobhan Senier, eds. *People, Practice, Power: Digital Humanities outside the Center* (Minnesota: University of Minnesota Press, 2021), https://www.upress.umn.edu/book-division/books/people-practice-power; Shannon Lucky and Craig Harkema, "Back to Basics: Supporting Digital Humanities and Community Collaboration Using the Core Strength of the Academic Library," *Digital Library Perspectives* 34, no. 3 (19 September 2018), https://www.emerald.com/insight/content/doi/10.1108/DLP-03-2018-0009/full/html.

[23] Christine Barats, Valérie Schafer, and Andreas Fickers, "Fading Away... The Challenge of Sustainability in Digital Studies," *Digital Humanities Quarterly* 14, no. 3 (September 25, 2020) http://www.digitalhumanities.org/dhq/vol/14/3/000484/000484.html.

improve discoverability, which one interviewee called the "fundamental problem of data sharing in the humanities."

Domain repositories would go a long way towards solving preservation and discovery problems. Their ability to make data visible to those researchers who are most likely to understand and reuse research data is why the NSF and NIH direct researchers to deposit data in them whenever possible and why both agencies continue to fund the development of new domain repositories.[24] However, domain repositories are expensive to create and maintain. They also work best when they can serve as hubs for data sharing with specific data communities, with overlapping research interests and sense of purpose. How to effectively translate the concept of a domain repository into the humanities is not clear. Humanists are as siloed as researchers in STEM fields and are also much less likely to be working on data-intensive projects. Scoping a humanist data community would require deep research into promising topics and communities, comparative research into lessons from domain repositories in STEM or social science fields and exploration of challenges faced by the Humanities Commons and the small number of other experiments with building repositories in this space. It would also require partnership with a funder or funders willing to underwrite further experiments. Though the payoff would potentially be significant, such an investment is unlikely given the overall state of funding in the humanities and current priorities of major funders. For this reason, data humanists would benefit from developing strategies to maximize the value of generalist repositories such as Dryad or Zenodo, or hybrids like the Humanities Commons, which seems to be repositioning itself as a general purpose repository.

Contributing to open research cultures

Humanists have much to learn about data deposit and reuse from researchers in other fields, many of whom still struggle with challenges that are common to data sharing regardless of disciplinary affiliation. But they can also contribute to the development of open research practices across disciplines. Two people we interviewed had backgrounds outside the humanities—one in data science, the other in engineering. Both described the value that humanistic ways of thinking can bring to developing scientific cultures of data sharing and curation. As one interviewee remarked, humanists are well positioned to help researchers across fields to understand how larger social forces are embedded in and shape research data, to highlight the importance of "context and interpretation" to data analysis, and to foreground the experiences and perspectives of marginalized groups.

Some digital and public humanists are also articulating ethical and methodological frameworks to embrace community co-creation of data, in which researchers collaborate with community

[24] Dylan Ruediger, Ruby MacDougall, Danielle Miriam Cooper, Jake Carlson, Joel Herndon, and Lisa Johnston, "Leveraging Data Communities to Advance Open Science: Findings from an Incubation Workshop Series," *Ithaka S+R*, 9 August 2022, https://doi.org/10.18665/sr.317145; K. Lehnert, L. Profeta, A. Johansson, and L. Song, "Best Practices: The Value and Dilemma of Domain Repositories," EGU General Assembly 2020, Online, 4 - 8 May 2020, EGU2020-22533, https://doi.org/10.5194/egusphere-egu2020-22533; J Boté and M. Térmens, "Reusing Data: Technical and Ethical Challenges," *DESIDOC Journal of Library & Information Technology* 39, no. 6 (2019), 329–337, https://doi.org/10.14429/djlit.39.06.14807.

groups as users, stakeholders, and generators of data.[25] Often driven by social justice imperatives, many of these projects focus on the intersection of data and structural racism. For example, the Notes on Creating Livable Black Futures project, a collaboration between Dr. Stacie McCormick of Texas Christian University and the Aifya Center, a community reproductive health organization, uses community storytelling as the foundation for data collection and research to inform reproductive health policy in Texas.[26] Likewise, the Black Beyond Data project includes substantial community partnership aimed at allowing marginalized black communities in Baltimore to "take control of their own data," and conduct data-driven research on their communities.[27] The emphasis on community engagement with both data collection and with defining the project's research agenda pushes these projects beyond outreach models of community engagement, and provides a model for future work at the intersection of data and the humanities.

Such projects have the potential to drastically reshape the boundaries of what is conventionally understood as "data sharing," a term that most often refers to the transfer of data between highly specialized groups of academic researchers. In them, we can see the articulation of a concept of sharing that emphasizes the affective and empathetic connotations of the word, reconceptualizing sharing as a communal process that gains strength through mutual exchange and reuse, crossing in and out of academic communities.

Looking forward

Data is likely to continue to be a problematic concept for the humanities. Deeply rooted ways of knowing and standards for evidence have made them less likely to turn to the quantitative and computational methods prevalent in the majority of disciplines. The standards and principles of open research have developed with little concern for the humanities, and many humanists, in turn, have seen little reason to engage with or participate in data sharing.[28] However limited its direct effects may be, the Nelson memo should stimulate humanists to reconsider their distance from the principles and practices of open research and develop new ways to engage with the new scholarly infrastructure.

---

[25] Kim Gallon, "Making a Case for the Black Digital Humanities" in *Debates in the Digital Humanities 2016*, eds. Matthew K. Gold and Lauren F. Klein (Minnesota: University of Minnesota Press, 2016), https://dhdebates.gc.cuny.edu/read/untitled/section/fa10e2e1-0c3d-4519-a958-d823aac989eb; Catherine D'Ignazio and Laura F. Klein, *Data Feminism* (The MIT Press, 2020); Susan Smulyan, ed., *Doing Public Humanities* (New York: Routledge, 21 July 2020); D. H. Mutibwa, A. Hess, and T. Jackson, "Strokes of Serendipity: Community Co-curation and Engagement with Digital Heritage," *Convergence* 26, no. 1 (2020), 157–177, https://doi.org/10.1177/1354856518772030.

[26] Stacie McCormick, "Notes on Creating Livable Black Future," Humanities for All, December 2022, https://humanitiesforall.org/projects/notes-on-creating-livable-black-futures.

[27] Julia Scharper, "Black Beyond Data," *Arts & Sciences Magazine*, 2 June 2022, https://hub.jhu.edu/2022/06/02/black-beyond-data-jessica-marie-johnson/.

[28] Erzsébet Tóth-Czifra, "The Risk of Losing the Thick Description," in *Digital Technology and the Practices of Humanities Research*, ed. Jennifer Edmund (Open Book Publishers, 2020), https://www.openbookpublishers.com/product/1108; Christine Barats, Valérie Schafer, and Andreas Fickers, "Fading Away... The Challenge of Sustainability in Digital Studies," *Digital Humanities Quarterly* 014, no. 3 (September 25, 2020), http://www.digitalhumanities.org/dhq/vol/14/3/000484/000484.html.

Even those humanists who consider their sources to be data encounter serious barriers to sharing research data.[29] Humanists work with evidence that is too unruly for abstraction, too slippery for ontologies, and so unique as to require its bespoke solutions and infrastructure. They worry that their data will be misused or misinterpreted. Copyright and privacy issues make sharing difficult. They have no professional incentives that encourage them to do so. They lack the skills and time to properly manage, curate, and document their evidence—and, as one interviewee noted, they think of data as a means rather than an end.

Without erasing the distinctiveness of the humanities or denying their frequent struggle to secure financial and technical support, all of these challenges are familiar to researchers in STEM fields, decades of sustained investment from the NSF, NIH, and other funders notwithstanding. Humanists have much to learn from these fields, even if the blurry but essential distinction between sources and data means that any solutions will need to be adapted rather than adopted wholesale. And they will likely find that tangible changes to researchers' practices will be incremental.

The animating principles of democratizing access to knowledge and upholding integrity in the research process that drive the open research movement are well aligned with the ethics of many humanists. Advocates of FAIR data have invested considerable effort developing sophisticated frameworks for balancing these broad principles with other important values and accommodating the needs of individual disciplines; these frameworks provide an entry point for adapting open research to the humanities.[30] Humanists may be able to return the favor by turning their skills at understanding the institutional, social, and cultural context of knowledge production to advocate for equitable, inclusive, and just ways of conceptualizing the fairness and openness that "plac[e] the human and human society at the center of debates" around data ethics.[31]

Changing researchers' practices depends on creating a sense of benefit and impact. Within the contexts of data sharing, the two most important motivators have been mandates from funders and journals, and the likelihood that datasets will be cited or reused.[32] Even in well-resourced fields, the economics of funder and journal mandates are contentious and a path to equitable sustainability unclear: in the humanities, where most research is largely self-funded, existing mechanisms for covering the costs associated with data deposit seem out of reach and unlikely

[29] Toma Tasovac, Sally Chambers, Erzsébet Tóth-Czifra, "Cultural Heritage Data from a Humanities Research Perspective: A DARIAH Position Paper," Digital Research Infrastructure for the Arts and Humanities, 2020, https://hal.science/hal-02961317/document; Barbara McGillivray et al., "The Challenges and Prospects of the Intersection of Humanities and Data Science: A White Paper from The Alan Turing Institute," The Alan Turing Institute, 2020, https://doi.org/10.6084/M9.FIGSHARE.12732164; Rebecca Grant "Reusable, FAIR Humanities Data: Creating Practical Guidance for Authors at Routledge Open Research," Zenodo, 2022, https://zenodo.org/record/6645166#.Y73i5nbMJPY.

[30] Annika Jacobson et al., "FAIR Principles: Interpretations and Implementation Considerations," *Data Intelligence* 2 (2020): 10-29.

[31] ALLEA, "Sustainable and FAIR Data Sharing in the Humanities" (Berlin, 2020), https://allea.org/portfolio-item/sustainable-and-fair-data-sharing-in-the-humanities/; Barbara McGillivray et al., "The Challenges and Prospects of the Intersection of Humanities and Data Science: A White Paper from The Alan Turing Institute," The Alan Turing Institute, 2020, https://doi.org/10.6084/M9.FIGSHARE.12732164.

[32] Gregory Goodey, et al., "The State of Open Data 2022," *Digital Science*, 14 October 2022, https://doi.org/10.6084/m9.figshare.21276984.v4.

to drive significant change. At present, reuse and citation of datasets in monographs and journal articles is also rare.

> Changing researchers' practices depends on creating a sense of benefit and impact. Within the contexts of data sharing, the two most important motivators have been mandates from funders and journals, and the likelihood that datasets will be cited or reused.

Perhaps the most promising frameworks for rethinking data in the humanities have one foot outside the academy. The new wave of data-driven projects share some characteristics with digital humanities projects and kinship with decades of work in the public humanities, yet they push humanistic research in new directions. By tackling new subjects and experimenting with methodological approaches that focus on co-generation of research agendas, as well as co-creation and co-ownership of data with community partners, they offer new ways to think about what open research, sharing, and reuse *could* mean in and beyond the humanities.

The "year of open science" and the mandates of the Nelson memo's both seek to increase public access to research data and foster collaboration and data sharing among researchers. These are important goals, though it is worth pointing out that in the memo, the public's role in research is largely passive: they fund it through taxes and consume it by reading scholarly articles. The value of open access to data is focused on its usefulness to other researchers, whether in academia, the government, or in industry, not to the public *per se*.

These kinds of researcher-to-researcher data sharing communities have not flourished in the humanities, partially because the infrastructure of domain repositories and other resources they depend on are beyond the means of humanities funders to create and sustain. What humanists are beginning to do is expand who the members of a data community can be by investing in co-creation of data designed to reach out to and originate from outside of academia and disrupt the passivity of the public within the open research framework. Humanists do not have a monopoly on community-engaged research: participatory action research methodologies and citizen science projects, among others, have similar orientations. However, humanists who are seeking an entry point into the largely alien world of open research and data sharing can draw on these projects to build frameworks for data sharing, reuse, and community in the humanities.