Data curation as collective action during COVID-19

Kalpana Shankar, Wei Jeng, Andrea Thomer, Nicholas Weber, Ayoung Yoon

Abstract

In this commentary, the authors, an international group data curation researchers and educators, reflect on some of the challenges and opportunities for data curation in the wake of the COVID-19 pandemic. We focus on some topics of particular interest to the information science community: data infrastructures for scholarly communication and research, the politicization of data curation and visualization for public-facing “dashboards,” and human subjects research and policies. We conclude with some areas of opportunity and need, including broader and richer data curation education in the information schools, the establishment of better data management policy implementations by research funders, the award of formal academic credit for data curation activities and data sharing, and engagement in cooperative action around data ethics and security.

1 INTRODUCTION

Since emerging as a global public health crisis in January 2020, the COVID-19 pandemic has changed nearly everything about our world and our lives—including our collective understanding of the vital role that data curation plays in scientific research, particularly viral outbreaks. Data curation, as both a field of study in information science and professions and as a practice among various fields of research, enables the efficient, reliable, and trustworthy dissemination of key evidence needed for decision making. Topics that are central to data curation, such as the development and application of data and metadata standards, the organization of data for discovery and re-use, and open science and data sharing have taken on new urgency. Meanwhile, activities that were peripheral to what some consider the scope of the field—topics such as peer review, research evaluation, research infrastructure development, privacy/security, to name a few—are now clearly core to the use of data in managing COVID-19.

In light of these seismic shifts, we take this opportunity to consider how COVID-19 has troubled our current assumptions and understandings of data curation—both its practice and its teaching. We offer this commentary as a group of data curation researchers, but also as a committee convened by the iCaucus in 2019 to define and map data curation as a field of study currently being taught in iSchools. In writing this piece we acknowledge we are setting aside many of broader and serious issues that COVID-19 has raised for scholarly communication and research practice: rushed peer review, murky public-private partnerships, questionable authorship practices, and the double-edged sword posed by preprint servers (Majumder & Mandl, 2020; Mandavilli, 2020). We focus instead on a few arenas of collective action: (a) the distribution of data curation labor and the need for collaboration among multiple institutions; (b) the role that data curation plays in communicating data to the public via dashboards and

This is the author's manuscript of the article published in final edited form as:

Shankar, K., Jeng, W., Thomer, A., Weber, N., & Yoon, A. (2021). Data curation as collective action during COVID-19. Journal of the Association for Information Science and Technology, 72(3), 280-284. https://doi.org/10.1002/asi.24406

similar visualizations; and (c) the challenges posed for human subjects protections, privacy, surveillance, and potential for data misuse and the role of data curation experts in addressing such challenges. We conclude by outlining some emergent challenges our research and teaching must address to ensure that Information Science is prepared to contribute substantially to the practice of data curation during this period of social and scientific uncertainty.

2 COOPERATIVE CURATION DURING COVID-19

Contemporary science depends upon the development of sophisticated infrastructures for collecting, managing, and analyzing data. These activities, the technologies, and practices that support them, and the outputs they generate require purposeful curation (Palmer et al., 2013).

Transparency, accountability, and access to empirical research results have always been central to a scientific ethos of reliable knowledge production. It is only through cooperation and collective action that expert knowledge is effectively produced, evaluated, and disseminated. It should not take a pandemic to underscore how fundamental these communal activities are to the production of scientific knowledge. Yet, like all infrastructure, the human and technical systems that facilitate knowledge production remain invisible until they break down.

The COVID-19 pandemic has made the fragility of our infrastructures frighteningly clear yet has simultaneously highlighted the lifesaving importance of those infrastructures and collective action within them in times of crisis. Examples include:

- Outdated government systems that depend on technologies like the fax machine which cannot easily scale when a surge of unemployment insurance requests flood the system.
- An internet that is susceptible to bandwidth shortages and network constraints. This is a result of developing international governance of internet provision that is optimized for the delivery of videos and other streaming media entertainment services but are poorly equipped to handle these same demands for public education, health care, and mutual aid in a fully online environment.
- Inequitable access to hardware, software, and broadband for conducting business and education in a completely online environment.
- Marketplace solutions for basic activities, like teleconferencing, indicate our need to take seriously and act expediently on the implications of developing data collection and privacy policies which can protect consumers from surveillance capitalism (Soltani, Calo, & Bergstrom, 2020).

Despite these infrastructural shortcomings, there are also myriad examples of collective action in science and scholarship that are truly impressive and, in many ways, lifesaving. One of the shining examples among these cooperative activities is the preparation, preservation, sharing, and reuse of research data. This is not by accident. The human and technical infrastructures that facilitate the sharing of valuable research data have been diligently developed for nearly three decades. During COVID-19, data curators are putting these infrastructures to work in not just meeting emergent demands of a research community, but accelerating the effectiveness of resource sharing and distributed work at a distance. This form of cooperative curation through existing infrastructures has numerous positive examples:

- Nexstrain is an open-source pathogen tracking platform developed by researchers at the University of Washington and Fred Hutchinson Cancer Research Center. During COVID-19's outbreak NextStrain has been used to both compare and display sequenced genomes of the novel coronavirus strain by curating data shared by different research agencies throughout the world. In addition to this analysis and data sharing curators are also lending their expertise to translate documentation and shared resources into numerous languages so that decision makers can access and make use of this valuable information.
- The United States Centers for Disease Control (CDC) COVID Data Tracker is publishing official statistics about infection rates and hospitalization on a live basis, while the Johns Hopkins University (JHU) Coronavirus Resource Center is curating and publishing data reported by the Red Cross and other international agencies to create a holistic view of not just infection rates, but also comparative statistics about where the virus is being contained or spiraling out of control.

There are also numerous examples of citizen science and journalism filling needed gaps in government data coverage and communication:

- The New York Times has embarked on a massive data curation endeavor to publish an Excess Deaths tracker which aggregates federal, state, and county level data to show the impact of COVID-19's spread on a national infrastructure. They have also been diligently recording the methodology behind this curation, including epidemiological processes for reporting confirmatory laboratory testing, and probable cases based on county-level reports.
- The COVID Tracking Project housed by Atlantic Magazine is aggregating state-level data in the United States and creating visualizations that depict the racial inequality in infection rates. This data curation goes beyond simply reporting but provides a context for how viral infections are being distributed through a diverse population.
- Avi Schiffman, a teenager in Mercer Island, Washington created one of the most heavily accessed COVID-19 data tracking websites (Schiffman, 2020) by integrating openly available data reported from countries around the world. He simultaneously showed that curation of available data was possible in a rather short time span, and pushed counties, states, and countries around the world to make their data more interpretable.

3 DATA CURATION AND THE PUBLIC SPHERE

In addition to demonstrating the importance of data curation is to the production of knowledge, the above examples also show data curation's importance in communicating science to the public sphere. Although the popularity of dashboards and other data tracking websites is often ascribed to their power as information visualizations, they are only as effective and trustworthy as the data curation that goes into them. Each of the examples above is dependent upon careful, consistent, and transparent data aggregation, normalization, and publishing workflows. The trustworthiness of these sites is largely directly tied to the curatorial work that goes into them, as well as the transparency regarding that work. For instance, both the CDC and JHU provide clear provenance information detailing how data are collected, normalized, and accessible for reuse, which has greatly increased the reusability of the dashboards and their data. Conversely, the trustworthiness of some other dashboards has been called into question when the chain of provenance between data and dashboard is broken. Consider the now notorious case of the state of Florida's COVID-19 dashboard; its lead. On May 18, 2020 data scientist Rebekah Jones was fired when she allegedly refused to manipulate data to make the rate of infection seem lower, and therefore hasten reopening plans. Since her firing, Jones has gone on to create her own dashboard of Florida case data, in which data are openly published in a transparent manner (Wamsley, 2020)—however, Florida's official dashboard remains controversial. Similarly, the state of Georgia's Department of Public Health has also come under fire for publishing confusing or downright misleading figures that intentionally obscure rising infection rates (Carr, 2020). And most recently and worryingly, the Trump administration's decision to re-route all hospital data from a functioning CDC data system into their own, bespoke spoke system has created confusion and led to delays. These cases reaffirm the importance of tracking and communicating data provenance—for example, the processing, selection, and transformation of the data going into these dashboards to the public, and how obscuring—or even casting uncertainty on—chains of provenance can profoundly spread misinformation and disorder. Furthermore, they highlight the role that issues of curatorial control play in communicating data to the public: who gets to decide what data are “good”? Who has final say over curatorial decisions? Who has control over curatorial workflows? Opaque data curation policies and practices run the risk of reducing public trust in data, and therefore in the institutions and public health messages represented by these data.

4 HUMAN SUBJECT PROTECTION AND DATA PRIVACY

The diverse data sources that need to be curated for COVID-19 research and public policy can include not just traditional health data (e.g., clinical trial data) but also data from personal devices (e.g., via contact tracing apps) and aggregated business transactions (e.g., credit card usage; Sapiezynski, Stopczynski, Gatej, & Lehmann, 2015; Lampos & Cristianini, 2010). When effectively combined these data sources allow for the digitized tracking of human mobility, modeling of viral spread, and prediction of outbreaks and other trends. However, the collection and sharing of this data also raise serious concerns regarding privacy infringement, surveillance, and the risk of misuse—all of which may persist after the major health threats of the pandemic subsides.

To respond to possible data privacy infringement concerns, many governments have enacted such legislative remedies as adopting existing emergency laws (e.g., Israel, Singapore, South Korea, Taiwan), amending existing laws (e.g., Germany, France, Norway), or proposing new laws from scratch (e.g., Italy; Organisation for Economic Co-operation and Development (OECD), 2020). While these laws allow governments to wield extraordinary powers during states of emergency, such laws are often vague in scope and lack proper oversight mechanisms (Soltani et al., 2020). As a result, some have taken initial steps to limit the authorization of data collection and processing. For example, the EU parliament (2020) recommends adding sunset clauses to contact-tracing apps so that these apps cannot be used during post-pandemic times. Norway's health authority has also decided to move away from using contact-tracing apps and has deleted data collected so far.

As lawmakers and privacy enforcement authorities worldwide are considering the balance between public health and data privacy, for example, the principle of proportionality (Ienca & Vayena, 2020), researchers in the field of data curation will need to pay close attention to the ethics and legitimacy of government regulations around COVID-19 data during and after the pandemic—and intervene when necessary. For example, in Ireland, academic researchers with expertise in data and visualization and civil society groups were effective in shaping and evaluating the contact tracing application developed and subsequently deployed by the Irish Health Service Executive (Irish Council of Civil Liberties, 2020).

5 CHALLENGES FOR AN UNCERTAIN FUTURE

For the scientific community in general and data curation in particular to evolve for the better we need to address long-standing challenges of scale, scope, and impacts in our practices. Above, we have outlined how critical issues of data curation are to tackling the COVID-19 crisis. Data curation training needs to incorporate these topics in meaningful ways—not just signaling their importance to students, but actually equipping students with the skills necessary to work in broad, sometimes unexpected collaborations, support responsible data management and use/reuse, deal with the challenges of data provenance at scale, and make and support ethical decision-making about the proper use and reuse of sensitive human subjects data.

To give one example, responsible data management and curation often begin with effective data management plans which will continue to gain in importance. We argue that treating data management plans as aspirational promises is simply intractable. Funders need to weigh the costs of data management plan implementation and fund curation work to actually achieve these lofty goals. Data management plans were an important first step for scientific research institutions to encourage more responsible data management and sharing and they will continue to play a role in ensuring the effective management of research data for ongoing research. But funding agencies have been reluctant to enforce these plans, in part, because forcing researchers to actually execute or follow through on data management plans requires additional, substantive, financial support. Students also need to be prepared to step into settings like government agencies where data management directives are encouraged, but strict adherence to best practices are still being sorted out. There is an incredible opportunity for Information Science to produce valued contributors to the massive data curation enterprise that has emerged during COVID-19, but that will likely require a step-wise procedure—faculty and instructors incorporating data management trainings from within and outside of federally funded research agencies, as well as quickly moving and rapidly evolving data publication in the public sector.

We need to credit and acknowledge activities related to curating data, as well as developing, maintaining, and supporting research software and hardware. We need to not just mention these activities in the acknowledgment sections of academic publications (Weber & Thomer, 2014) but give equal weight in tenure, promotion, and hiring practices to all of the activities necessary to cooperatively produce new knowledge. This requires a radical change in not just the way that we reward curation activities but fundamentally shifting our expectations for reliably educating and sustaining a research and development workforce that can perform curation tasks at scale. [Corrections added on 19 September 2020, after first online publication: the reference citation in this paragraph has been updated.]

While governments have adopted mobile technology and data analytics to fight the pandemic, for example, border control, case identification, and contact tracing, the threat of data breach and abrogations of personal data protection have also evolved. Data curation professionals and scholars need to contend with the balance of data protection and public health and how their roles engage this balance and advocate for technologies available to ensure privacy-by-design in the stages of data collection, processing, and sharing.

All of this is possible. COVID-19 is not a great equalizer, but it is a reckoning with the past—with our failures, our neglect, our willingness to accept known inequities and inefficiencies in the scientific research ecosystem. But this moment also brings to light so much of our collective capacity to quickly generate important research and engage in cooperative sharing of valuable data. Some scientists argue that science is “better” for COVID-19 having accelerated the sharing of some data, the development of infrastructures, and open access publication (Callender, 2020). However, caution is warranted. We saw similar developments during the Ebola and SARS outbreaks that were not sustained (Delaunay, Khan, Tatay, & Liu, 2016); it remains to be seen whether current changes are sustainable.

We do know that our previous practices in conducting and supporting scientific research are going to be fundamentally different from this point forward. As Ravetz (2020) wrote, for a pandemic recovery, we do not need a “new normal” for science; we need a “post-normal” science that acknowledges that social and political factors are constitutive of science and response and not a response or even a barrier to it (not a new viewpoint, granted). We argue that the same is true for our research, teaching, and understanding, and implementation of data curation.

REFERENCES

- Callender, C. (2020, July 11). COVID-19 is making science better. *Issues in Science and Technology*. Retrieved from https://issues.org/covid-19-is-making-science-better/#.Xv9CHI3vlb4.link

- Carr, N. (2020, July 23). Public health experts call out confusing COVID-19 data maps; DPH set to make changes [WSB-TV]. Retrieved from https://www.wsbtv.com/news/georgia/public-health-experts-call-out-confusing-covid-19-data-maps-dph-set-make-changes/KZPBOLBG2BG2TAMI2N4CQGWESA/

- Delaunay, S., Khan, P., Tatay, M., & Liu, J. (2016). Knowledge sharing during public health emergencies: From global call to effective implementation. *Bulletin of the World Health Organization*, **94**(4), 236– 236A.

- European Parliament. (2020, June 5). *COVID-19 tracing apps: Ensuring privacy and data protection*. Retrieved from https://www.europarl.europa.eu/news/en/headlines/society/20200429STO78174/covid-19-tracing-apps-ensuring-privacy-and-data-protection

- Ienca, M., & Vayena, E. (2020). On the responsible use of digital data to tackle the COVID-19 pandemic. *Nature Medicine*, **26**(4), 463– 464.

- Irish Council for Civil Liberties. (2020, July 2). *Experts issue pre-release report card on the HSE COVID-19 tracker app*. Retrieved from https://www.iccl.ie/2020/experts-issue-pre-release-report-card-on-the-hse-covid-19-tracker-app/

- Lampos, V., & Cristianini, N. (2010). Tracking the flu pandemic by monitoring the social web. In *Second international workshop on cognitive information processing* (pp. 411– 416). Elba, Italy: IEEE.

- Majumder, M. S., & Mandl, K. D. (2020). Early in the epidemic: Impact of preprints on global discourse about COVID-19 transmissibility. *The Lancet Global Health*, **8**(5), e627– e630.

- Mandavilli, A. (2020, June 18). Scientists take aim at another coronavirus study in a major journal. *The New York Times*. Retrieved from https://www.nytimes.com/2020/06/18/health/coronavirus-retractions-studies.html?fbclid=IwAR3Cl8LG9eq4pLqtcvobBUrUApbR-4jrqVY0aL8VoPhi4SFGLvqTbK6ZdII

- Organisation for Economic Co-operation and Development. (2020). *OECD policy responses to coronavirus (COVID-19)*. Retrieved from http://www.oecd.org/coronavirus/policy-responses/ensuring-data-%20privacy-as-we-battle-covid-19-36c2f31e/#section-d1e69

- Palmer, C.L, Weber, N.M., Muñoz, T, & Renear, A.H. (2013). Foundations of Data Curation: The Pedagogy and Practice of Purposeful Work with Research Data. *Archive Journal*. (3) http://www.archivejournal.net/essays/foundations-of-data-curation-the-pedagogy-and-practice-of-purposeful-work-with-research-data/.

- Ravetz, J. R. (2020, June 19). Science for a proper recovery: Post-normal, not new normal. *Issues in Science and Technology*. Retrieved from https://issues.org/post-normal-science-for-pandemic-recovery/

- Sapiezynski, P., Stopczynski, A., Gatej, R., & Lehmann, S. (2015). Tracking human mobility using wifi signals. *PLoS One*, **10**(7), e0130824.

- Schiffman, A. (2020). *nCov2019.live*. Retrieved from https://ncov2019.live/data

- Soltani, A., Calo, R., & Bergstrom, C. (2020, April 27). *Contact tracing apps are not a solution to COVID-19 crisis*. Brookings TechStream. Retrieved from https://www.brookings.edu/techstream/inaccurate-and-insecure-why-contact-tracing-apps-could-be-a-disaster/

- Wamsley, L. (2020, June 14). Fired Florida data scientist launches a coronavirus dashboard of her own [National Public Radio]. Retrieved from https://www.npr.org/2020/06/14/876584284/fired-florida-data-scientist-launches-a-coronavirus-dashboard-of-her-own

- Weber, N. M., & Thomer, A. K. (2014). Paratexts and Documentary Practices: Text Mining Authorship and Acknowledgment from a Bioinformatics Corpus. In N. Desrochers & D. Apollon (Eds.), *Examining Paratextual Theory and its Applications in Digital Culture* (pp. 84–109). Hershey, PA: IGI Global. [Correction added on 19 September 2020, after online publication: This reference has been changed.]