Commonplace

An Opportunity to Do Better, Together

Daniella Lowenberg

Published on: Jul 09, 2020

DOI: 10.21428/6ffd8432.d497e780

License: [Creative Commons Attribution 4.0 International License \(CC-BY 4.0\)](https://creativecommons.org/licenses/by/4.0/)

I am often questioned about [my perspective](#) on how data sharing has changed because of the pandemic. I argue that data sharing has not changed, but the pandemic highlights not only how important data sharing is (like other crises have, for instance, the climate crisis) but how it spotlights larger issues in our data sharing social and technical infrastructure. I want to take a moment to parse out how we have an opportunity to do better as we re-build.

Research processes at the start and end of the 20th century have evolved in terms of our abilities to experiment with and understand pathogens and human immunology, but much of Dr. Salk's below quote reflecting on his work with the polio vaccine stands to illustrate how complex the scientific process is and how reliant it is on fine tuned *responsiveness*.

"I would picture myself as a virus... for example, and try to sense what it would be like... I would also imagine myself as the immune system, and I would try to reconstruct what I would do as an immune system engaged in combating a virus... When I had played through a series of such scenarios on a particular problem and had acquired new insights, I would design laboratory experiments accordingly" — Jonas Salk

As infrastructure providers, publishers, educators, it's important for us to consider our response, so far, and going forward, calibrating our expectations according to the scientific process without assumptions that we could or have created in silo about how best to support scientists and data sharing right now.

Ethical behavior takes time

We are living in an unprecedented time. Unlike previous pandemics such as the Spanish flu and even the more recent onset of the AIDS crisis, we now live in the era of "big data" and we have the ability to rapidly disseminate results. We have supportive infrastructure in place to educate researchers on data sharing and we have sturdy technical frameworks for disseminating and hosting these data. And while speed and quantity can be seen as wins, there is a balance with quality. Importantly, there is a real need to understand pitfalls ahead if we solely focus on rushing or false advertising, misaligned with bringing value to scientific research publishing.

Ethical behavior that maintains confidence in the science community takes time. *Ethical behavior does not just mean doing good, it means doing better*

science. There are disciplines of research where we will see rapid results, and we've already seen the benefits of this like the COVID19 sequences uploaded and analyzed that have allowed for a better [understanding of transmission patterns](#). But many results we will not see for a while, and we know that high quality data collection takes time. Lab resourcing may vary within and across institutions and resourcing weighs in on how fast data collection and analyses can happen. Types of data collected also vary in timing. For example, mouse model studies take time to infect the animal, implement proper time constraints, etc.

This is all important to note because the cost of recovery from sloppy scientific practices is too high. Light will be shown on the reputation of the scientific community during this pandemic if experimentation or data analyses are rushed or are not performed at a high caliber. As [supporters](#) of the scientific community and processes, repositories and publishers have a role in this as well. Sloppy publishing practices, not accounting for proper curation and re-use, or publishing data that should not be public, reflects on our roles promoting responsible and open science.

Misaligned priorities, shifting focus

Considering these realities, I am concerned about the types of questions routinely posed in the scholarly communications space — pitting repositories against each other, driving competitive attitudes over topics misaligned with the realities of current research or researcher needs:

- *Immediate publication*. There was an instant reaction to favor and associate success with repositories that host many COVID19 datasets, without considering the time it takes for data collection & analyses. This reaction also did not consider the value or time dependencies for data curation and peer review.
- *Quantity over quality of publications*. By rewarding repositories that have more content, with an emphasis on quantity of materials published, we see over simplifications of what data are, broadly referring to all research outputs (including supporting PDF files, preprints, slides) as “research data” to inflate numbers. If we want to elevate research data as a “first class output” we should value it as its own entity as we do for articles.
- *Fancy portals for discovery*. Many questions and surveys have been about resources put into flashy portals at individual repositories to convene COVID19 data. These efforts have ignored the reality of how researchers find and use research data (at aggregate searching interfaces, science Twitter, through article linking, etc.) and

these efforts could have and could be focused on optimizing metadata discovery across the web.

Not only are these prompts misaligned with good science, they are misaligned with another area that deserves focus in the repository world: resourcing. With labs closing down and new experimentations paused, people have more time to consider their generated data and/or publicly available data. Repositories, across the scientific disciplines, have seen an increase in data submissions. This means that aligning with and supporting researchers right now includes considering proper support for community-run, researcher-adopted repositories, regardless of their focus or capacity to host COVID19 data, and thinking about long term sustainability for these resources that scientists rely on. This also includes thinking about how best to include the libraries and educators who are responding to and adjusting to new roles themselves. I shared some examples and stories of these responses in a recent [virtual presentation](#) at the NIH.

Building together, building for the future

While this pandemic may be once in a century, we need to make sure we are building for the future and not rushing to accommodate needs as they are dynamic and changing daily. Not only does this mean building flexible workflows but thinking about all disciplines of research. We should be building with the mindset that we are in this for the long-haul — we are going to see the production of COVID19 related data for years to come, and if trends go the way we would like, we should see increased data re-usage continually. We need to consider all stakeholders that are involved in publishing or re-using research data, accounting for their values up front, so as to scale, *together*, appropriately.

Publishers, preprint servers, repositories, and institutions, know that it is important to prioritize the curation and publication of all materials related to COVID19. But this rapid publication also requires us to not rush and publish flawed (threatening scientific reputation, inhibiting our understanding) or harmful (threatening participants and patients involved) data. Again, if our purpose for data sharing is to maximize the utility of research, we should coordinate to ensure all published research outputs are appropriate for publication: ethically sound and usable. This requires a large shift in our current practices and emphases.

Considering how the research landscape has been impacted, labs are trying to stay productive and want to lend their expertise, translating current research to COVID19

to get involved. More researchers are entering the space of working with personally identifying information (PII) who have limited experience working with human information. For example, researchers have told me that their labs are pivoting from working with mouse models to human studies or adapting their previous drug development studies. Researchers need to be mindful when working with human information, and this training and education requires support from funding bodies and institutions who can provide clear guidance on de-identification procedures, anonymization techniques, etc. There is often a focus on repositories gate keeping these data, solely responsible for proper publication, but this is largely ineffective to educate or effect change at the end of the road. We need to figure out how to better educate researchers upstream, adjusting to and resourcing for these trends across labs globally, so the larger research supporting landscape can more efficiently publish quality data and analyses.

And so I believe....

Fragmented approaches, misaligned with the goal of advancing scientific discovery, will fail to meet or evolve rapidly enough to support the changing ecosystem. The research process involves too many stakeholders and touch points to assume one repository or one publisher will, alone, effect widespread change. Institutions, funders, publishers, and scientists all have a role in promoting the rapid and responsible dissemination of research. *For years we've talked about collaborating, but for many, the natural instinct is to compete and be the "best" instead of thinking about what's best for science.* There is no quick answer for *how* to work together, but a start would be to consider resources researchers are increasingly in need of, across disciplines, and communally shift our focus to supporting those.

As research supporters, driven by scientific discovery, it is essential that we are guided by our alignment with researchers and are agile enough to respond to an ever-changing landscape. This means focusing on how best to support data-driven research, as opposed to applauding fancy features and marketing, while simultaneously guarding researchers from understanding the cost of publishing bad science, disincentivizing proper data practices like curation. As research evolves, we can respond accordingly by working together in open, sustainable ways, to advance the amount of re-usable data available for evidence based discovery. And if we work separately or try to seek competitive advantage, we misalign with good science, missing an opportunity to leverage our goals and values at a rare time where people are listening.