A matter of integrity: Can improved curation efforts prevent the next data sharing disaster?

Wider openness and access to data may be a necessary first step for scientific and social innovation, but as the controversial release of OK Cupid data highlights, open data efforts must also consider the quality and reproducibility of this data. What would it take for data curation to routinely consider quality and reproducibility as standard practice? **Limor Peer** *suggests some future directions to ensure data quality, consistency, and integrity.*

In May 2016 a toxic dataset with detailed records of 70,000 OK Cupid users was deposited online by a researcher from the University of Aarhus in Denmark with ostensible disregard for scientific and ethical norms around open data. One observer, Oliver Keyes, formerly with the Wikimedia Foundation, said about the incident, "The rise of convenient code and data platforms has been amazing, and helped us be more transparent about our methods – but it's also pushed the responsibility of ensuring that releasing the data does not pose harm back on to the researcher. As this incident shows a dozen times over, that's not something we can rely on." And another raised the question of whether "even transparent science may need some gatekeeping."

In 2013 I argued that, when it comes to science, data repositories have a special responsibility for data quality, especially as it relates to reproducible research. This responsibility rests on the understanding that, (a) sharing data is necessary but not sufficient for future reuse, (b) ensuring that data is "intelligently open" is crucial, and (c) incorporating a data review process is feasible. If data sharing platforms of all types took steps to ensure data quality as a matter of course, the OK Cupid release may have been flagged and stopped in its tracks.

Image credit Pixabay polaroid public domain

Increasingly, data repositories accept a variety of materials, mostly driven by the fact that data and software are increasingly inextricable. Code may be integral to using the data because it actually creates or models the data, because it is necessary for reading ASCII data into statistical software packages, or because it is used to analyze, interpret, and visualize the data. So when aiming to vouch for the quality of the data we have to take into account code as well. Code review is an established practice in computer science, and data repositories would be advised to heed the Recomputation Manifesto which states, among other things, that “tools and repositories can help recomputation become standard.”

Will curation for quality and reproducibility become routinized?

Max Weber, a German sociologist, philosopher, and political economist wrote about the inevitable trend toward rationalization, systemization, and routinization. The basic idea is that routinized procedures become attractive when people and organizations realize how they can contribute to achieving valued goals. While Weber is well known for his critique of these processes, it is worth noting that the basic tendency can be found in many aspects of human activity. We can already see evidence of routinization of the research data ecosystem in the realm of open science. I think we will see the inevitable routinization of curation practices as quality and reproducibility similarly become a valued goal.

Here are some reflections about what it would take for curation for quality and reproducibility to become inevitable:

1. Wide recognition that curating for quality and reproducibility is a valued goal

As Ann Green and I noted last year, there is reason to be optimistic. Goals and practices often develop together, and in mid-2016 data repositories and the scholarly digital infrastructure have matured, scientific societies and journals have evolved, and there is some evidence that curation for quality and reproducibility is on the rise. As journals increasingly require deposit of materials underlying publications, third party groups are beginning to offer review of these materials as a service. For example, the Odum Institute reviews all files for the American Journal of Political Science (AJPS) and Cornell Institute for Social and Economic Research offers review as a fee-for-service for Cornell researchers. In response to well-publicized problems with data sets, some organizations are promoting the use of badges.

For example, Open Science Framework is working to establish standards for open practices by introducing three badges (for data, materials, and pre-registration) and recommends the development of a “reproducibility badge” in the future, and the Center for Open Science published guidelines for transparency. The disciplines and their professional organizations are getting more involved. For example, computer scientists worry about replicability, code quality, and code failure, and devise approaches to overcome obstacles involved in creating reproducible computational research. The Association for Computing Machinery is developing procedures for artifact evaluation. Institutions, having a stake in their researchers’ output, could do more to ensure quality and reproducibility. The visibility of scientific reproducibility in scholarly publications and in the popular press makes this an exciting time to be working on these issues and provides an opportunity to make progress.

2. Finding the right balance between algorithms and people

Organizations that engage in curation for quality and reproducibility have established workflows and tools (e.g., the Inter-University Consortium for Political and Social Research ; the UK Data Archive). Routinization dictates that these will be continuously refined and increasingly automated and shared. The use of APIs such as ClamAV can go a long way toward this goal. At ISPS, we are developing a curation workflow tool that will automate as many curation tasks as possible, technically integrate these curation tasks, and do so using a structured but flexible workflow. Increasing use of algorithms can make the process more efficient and consistent.

That said, I don’t think we want to be 100% algorithmic. Even Facebook and Google – the ultimate algorithmic companies – acknowledge human judgment plays a role in deciding what information to display. Human experience and discretion may still be necessary in certain contexts, such as research involving human subjects. Algorithms can be helpful in flagging instances that require further review by humans.

3. Creating meaningful partnerships between researchers and data professionals

More automation of curation and review processes will not only reduce cost for organizations that care about quality but potentially enable researchers to meaningfully engage with these processes by integrating them with active research workflows and thus making them easy to use (see notion of "sheer curation"). In principle, researchers have an interest in curation for quality and reproducibility if only for the reason of having "a better relationship with your future self"; data analysis is essentially computer programming and it can and should be reproducible also for the benefit of cumulative knowledge.

Partnerships among researchers and data professionals can primarily be achieved in two ways. First, situating data professionals and curators upstream in the research process, potentially as embedded personnel can help support the data-focused aspects of the research life cycle for the benefit of the researcher and facilitate research transparency, open access, and data ethics for the benefit of the community. In this scenario, repositories are closer to the research process and out of the library, and there is less distinction between active and archive. Credit-giving and attribution are important here: Repositories taking on active curation tool could be considered contributors. Second, many research activities are now conducted on and via computers – the data are digital, data collection utilizes computer-assisted techniques, and scripts are applied to clean and analyze the data. All this can and should be logged, tracked, and easily reproduced. Making use of rapidly-evolving computational methods that capture (and potentially preserve) the entire research workflow, with tools such as Knitr, Sweave, Docker, ReproZip can contribute to overall more reliable reproducibility workflows as well as reduce the human effort required.

I continue to think that it is incumbent upon all data sharing platforms, and not just established data archives and repositories, to vouch for (or at least send a signal about) the quality of the data they hold. Curation for openness and access is a necessary first step, but we have to also think about quality and reproducibility. It's a matter of integrity.

This piece is cross-posted on Yale's ISPS blog.

Note: This article gives the views of the author, and not the position of the LSE Impact blog, nor of the London School of Economics. Please review our Comments Policy if you have any concerns on posting a comment below.

About the Author

Limor Peer is Associate Director for Research at the Institution for Social and Policy Studies (ISPS) at Yale University. She oversees research infrastructure and process at ISPS, including the Field Experiment Initiative, which encourages field experimentation and interdisciplinary collaboration in the social sciences at Yale. In this capacity, she has led the creation of a specialized research data repository (the ISPS Data Archive) and is currently involved in campus-wide efforts relating to research data sharing and preservation. At ISPS, Peer also heads the website team, and is responsible for research-related content on the site.