ELEPHANT

IN THE LAB

OPINION

Open Data can be advanced by the COVID-19 pandemic, but will still require a comprehensive approach

Title	Open Data can be advanced by the COVID-19 pandemic, but will still require a comprehensive approach
Author	Evgeny Bobrov
Author affiliation	QUEST Center, Berlin Institute of Health (BIH), Berlin, Germany
Author bio	Dr. Evgeny Bobrov supports biomedical researchers with managing and sharing their data. He offers consulting and training, and develops strategies for the provision of research data management (RDM) services. One of his core interests is the diversification of RDM services, to cater for the needs of both basic and clinical researchers within the same institution.
Author social links	Evgeny Bobrov: ORCID – Twitter
Date published	30 March 2020
DOI	10.5281/zenodo.3732948
Cite as (APA)	Bobrov, E. (2020). Open Data can be advanced by the COVID-19 pandemic, but will still require a comprehensive approach. Elephant in the Lab. DOI: 10.5281/zenodo.3732948

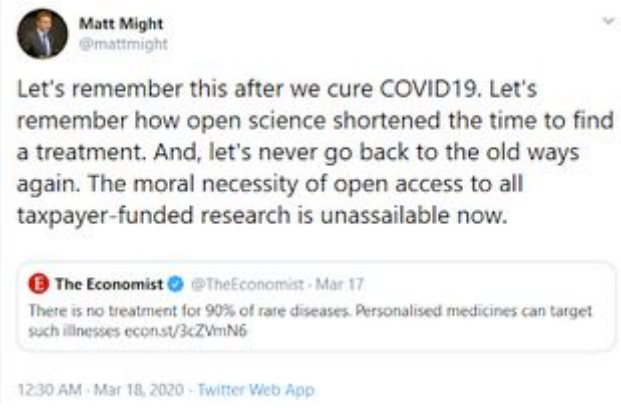
Sharing research data openly is becoming more common, but only slowly. Here, I will discuss whether the corona pandemic will accelerate the adoption of open data as a common academic practice.

The corona pandemic requires swift reactions, which in turn makes the sharing of all available knowledge on the virus as quickly and as freely as possible an imperative. The more reliable information epidemiologists, virologists, and emergency physicians can use for their work to slow the spread, develop vaccines etc., the better their work will be. Actors in the academic system have swiftly responded, and especially journals have made large strides towards openness by

nearly universally granting open access to articles on corona-related research (including publishing giants Elsevier, Springer Nature, and Wiley). The Allen Institute for AI has collated a massive set of articles on related research, and this collection is the basis for an analysis challenge to create text and data mining tools for answering corona-related questions. Similarly, datasets are shared at an unprecedented speed, as documented by the daily updated overview of datasets and clinical trials provided by Dimensions, a research information system owned by Digital Science. Open data have already allowed Johns Hopkins University to create a dashboard to track cases, as well as use of open source software Nextstrain to track virus evolution. And this might be just the beginning, given calls like a Nature editorial which urges governments and their advisors to embrace openness. Open data, including data on other corona virus infections shared before the crisis, are already contributing to the equally unprecedented speed at which potential treatments are developed and trials initiated. For example, data shared on the Gene Expression Omnibus database have contributed to developing a new drug repurposing methodology. If data sharing increases further, ideally including the multitude of clinical trials already registered (178 as of 27.3.2020 on ClinicalTrials.gov alone), open data can be expected to contribute even more to fighting covid-19 in the near future.

Expectations that data sharing will be accelerated

This speed and amount of collaboration is impressive, and has naturally caused open science advocates to see this as an impulse to bring academic cooperation, openness, and transparency to a new level. Exemplary is the rhetoric in a tweet by Matt Might, director of the Hugh Kaul Precision Medicine Institute at University of Alabama and former White House Executive Office strategist:



Open access to research articles is already a very widely adopted practice, and with quickly increasing rates of open access publishing, as well as an abundance of funder, learned society, and institutional policies in support of it, the tide has already turned towards open access. This is also underscored by how widely preprints on bioRxiv and medRxiv have been deployed as methods to rapidly communicate corona-related research results, with 791 articles posted as of

27.3.2020, showcasing the potential of preprints. The corona crisis might add further impetus to the open access movement, but this will be to an already rolling wave towards a widely accepted goal. Clearly there are still substantial questions regarding the effectiveness and efficiency of the gold open access model, as opposed to both 'diamond (or 'platinum') publishing and self-archiving with post-publication peer review. Nevertheless, the difficulties seem small compared to those on how to implement other open science practices which so far have much less footing. Calls to grant open access to all taxpayer-funded research importantly include open access to research data or simply 'open data'[1]. Top-down initiatives include a 2015 White House directive to US-agencies with a budget over 100M$, which has, however, not been implemented yet. At the same time, multiple bottom-up initiatives argue for open data, from large-scale projects like the Open Science Framework to local networks like the LMU Munich's Open Science Center. In agreement with these calls, as data sharing slowly becomes more common, its value starts to be verifiable, as e.g. by the analysis of Milham et al. (2018) quantifying both the scientific impact generated and the money saved by sharing and reusing brain imaging data. Over the last weeks, these earlier threads in support of open data have been combined with current successes and promises of corona data sharing, and a new narrative developed which suggests that

(i) the corona pandemic convincingly shows the power of data sharing,

(ii) this will cause researchers to be more aware of the value of sharing data in general,

(iii) this in turn will motivate researchers to share specifically their own data,

(iv) for which they will reconsider concerns and supersede obstacles to data sharing.

However, I will argue that this narrative is based on several questionable assumptions, and thus such a course of events is not to be expected, at least in the short-term and as a direct consequence.

Expectations of increased data sharing might not materialize rapidly

Let us take a look back to 2016. The year before had seen the start of a Zika virus outbreak in South and Central America, which led to the Statement on Data Sharing in Public Health Emergencies in February 2016. It states:

> "We [...] call for all research data gathered during the Zika virus outbreak, and future public health emergencies, to be made available as

[1] This article remains deliberately imprecise with the terms "open data" and "data sharing". Although technically speaking open data means completely free access to data for anyone and any purpose, it is understood that for some types of data this is not possible. However, restricted data sharing is conceptually and organizationally very closely linked, and to connect to current discourse, shaped by the framework of open science, such restricted sharing is here included as part of a more loosely defined "open data" practice.

> rapidly and openly as possible. The arguments for sharing data, and the consequences of not doing so, have been thrown into stark relief by the Ebola and Zika outbreaks."

The document's signatories include publishers and journals, and, importantly, large funders. The funders supporting the statement were indeed amongst the largest there are: NIH and NSF (both US), UK Medical Research Council, the German Research Foundation, and the Chinese Academy of Sciences. The exhortation of this widely endorsed statement was largely the same as can be found now in comments prompted by the corona crisis. But how much has an effect, if present at all within the communities immediately addressed, spilt over to science (or at least biomedical research) as a whole? Surely the last years have seen a surge of projects supporting the sharing of data in a reusable way, including the large infrastructures European Open Science Cloud (EOSC) and German National Research Data Infrastructure (NFDI), international standardization within e.g. the Research Data Alliance, and a multitude of smaller-scale repositories, platforms, tools, and standards. Indeed, the FAIR Data Principles as the most common framework for the reusable sharing of data were published in 2016. However, data sharing still remains rather an exception than the norm in most research communities. A survey commissioned by the European Open Science Monitor to a consortium including Elsevier ("Open data: The researcher perspective") has found that only 15% of researchers shared their data in repositories in 2017, while it is only repository data deposition that can reliably ensure data findability and reusability. Indeed, even these 15% might have been inflated, given both self-selection and self-report biases. Also, the Open Science Monitor reports (with unclear source) that in 2018 only 11% of researchers shared their data "directly" (a term bound to create confusion) with other researchers not collaborating with or personally known to them. While data sharing is slowly gaining traction, concerns persist, as both the "Researcher Perspective" survey and Digital Science's The State of Open Data 2019 survey show.

Thus, the question here is not whether we will have more data sharing in the future - we will. The question is whether the corona pandemic will in itself have a considerable effect on data sharing beyond its immediate disciplinary reach.

As stated before, the second assumption behind an expected effect of the pandemic on data sharing is that researchers will be more aware of the value of sharing data in general. Already this seems improbable in many cases, as it is easy and sometimes maybe even tempting to draw a line between research fields where data can reasonably be expected to be used for immediate improvements to health and environment, and those where such use is much more distant and abstract. In addition, it remains to be seen whether the corona-related data sharing efforts will even be widely perceived within the researcher community. At that, it should be considered that the amount of researchers immediately involved with corona-related research like virologists and epidemiologists is of course only a very tiny fraction of all researchers. Thus, the most persuasive first-hand accounts of the potential of reused data will only reach another tiny fraction in their immediate surroundings.

But even assuming that the value assigned to data sharing overall will increase, will researchers apply that to their own datasets? Surely many will, but as preregistered research from Houtkoop et al. (2018) suggests, the value attributed to data sharing for science as a whole is consistently higher than the value attributed to sharing one's own datasets, which might indeed be seen as an example of the tragedy of the commons. Thus, one can expect many researchers to accept the general importance of sharing data and at the same time fail to apply this practice to their own datasets.

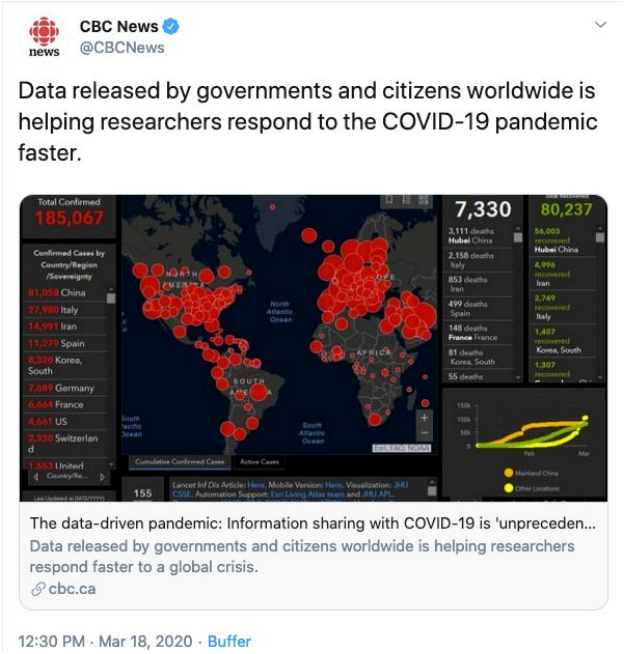
And of course whether data will actually be shared will strongly be impacted by knowledge, concerns, obstacles, and motivations, as evidenced in surveys (in addition to the aforementioned see Fecher et al., 2017) and other lines of work (Laine 2017, Pronk et al., 2015). An important obstacle to data sharing is that in the current environment it is often perceived as an "on-top" task, which is time-consuming to do - and even more so to do right - but is not sufficiently rewarded, and this seems improbable to change quickly. The perceived transactional cost of data curation and sharing will be augmented by a corona rebound effect, in which work will have piled up, and once researchers can return to normality, pressure will be higher than ever to conduct experiments, publish, submit reports etc. This will sideline non-prioritized tasks. It would in principle be conceivable that the push to share more data would start already while researchers are bound to do home office and barred from much of their regular work, but this seems equally improbable - there is always enough analysis and writing to catch up on, and unless the severe work restrictions carry on into 2020 on a world-wide scale, this does not amount to a disruption of regular workflows. However, as a side-note, it would be an interesting research question to pursue whether the increase in home office work fueled the reuse of open datasets.

Lastly, even if concerns and prioritization issues are overcome, several concrete obstacles remain. Some of them are legal (and in part ethical), with uncertainties regarding the sharing of personalized data being the major obstacle in biomedical research. Other concerns, especially regarding intellectual property and commercialization, are also prominent in that regard. And even if data can be shared without such concerns, the knowledge to curate them in a way which makes them truly reusable is not readily available. Indeed, many of the necessary platforms, standards, and tools are not even developed or at least sufficiently consolidated yet. The bottom-line is that in many research fields, those researchers who want to "do it right", are largely left with complex processes, yielding data sharing outcomes of unclear future usefulness.

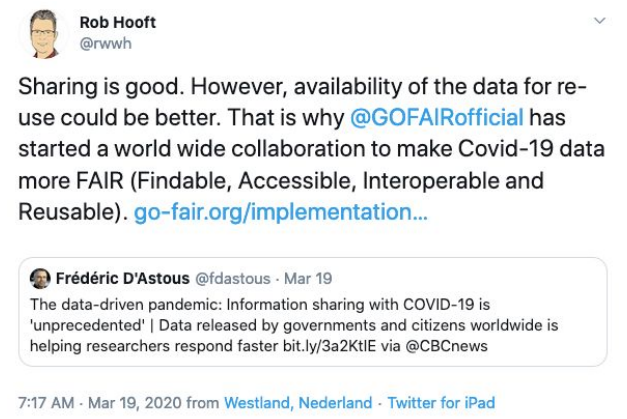
Lessons learned, and lessons known already

It is thus imperative to turn to funders and institutions, and ultimately to politics, to ensure processes and infrastructures which would allow for low-threshold, efficient and "reuseful" data sharing. At a heightened speed, this very same discussion on data sharing, and how to make data reusable, is already taking place within the wider corona research community. Amidst wide

acclaim to the Johns Hopkins dashboard, featuring the newly-coined term "data-driven pandemic",



the need to share data in a "FAIR" way has been highlighted,



leading to the founding of the Virus Outbreak Data Network within the GO FAIR Initiative. It can only be hoped that under the pressure of the pandemic, the different stakeholders will be able to overcome obstacles quickly and data sharing standards and practices will be taken up universally. Such a pace and degree of mobilization might be unfeasible and indeed undesirable as a general mode of work in "peace times". However, the many efforts to develop data sharing infrastructures and standards, which are already under way, need to receive more attention. By prioritizing their work and widening support by research institutions and learned societies, hurdles to FAIR data sharing can be reduced at a quicker pace than observed so far. The lessons learned during the corona crisis can surely contribute to this, and corona researchers might well become sought-after consultants later in other research fields. They will have a lot to say about the

potential (as well as the limitations) of disrupting regular processes to severely cut down the time needed to publish results, agree on standards, and implement sharing practices.

Importantly, processes and infrastructures for FAIR data sharing must include human support infrastructure. Large infrastructures like EOSC are critical, but as much are local research data management units, which support researchers with data curation tasks. As Barend Mons, one of the leading authors of the FAIR principles, has recently argued, 5% of research funding should go into data curation. At the same time, he reiterated his argument that it would be highly inefficient to expect researchers to take over increasingly complex data curation tasks ("Data Stewardship for Open Science"). While the 5% number might be contentious, it is clear that a world of machine-readable, linked data is only to be had with sufficient funding for professional data curators or stewards of some sort and with placing this as high on the agenda of funders and institutions as it is on the agenda of the European Commission's president. While research funder mandates for data sharing are in principle laudable, they need to be firmly in line with what is currently feasible for researchers. Where they are not, it can lead to data shared in a "reuseless" way, or even undermine the credibility of these efforts. It needs to be ensured that on both systemic and local institutional level researcher data sharing efforts are facilitated, rewarded and aligned with competing frameworks, resolving conflicts between openness, data protection, and commercial use. This will require the much wider recognition of data curation and sharing as a valuable service to the community, and datasets as research outputs in their own right. It will also require an agreement on robust and accessible guidelines for what is or is not legally admissible and institutionally desirable with certain types of sensitive data, thus freeing researchers from legal uncertainty and removing contradictory obligations.

As researchers, research managers, support staff, and others in the system, we have to accept that pushing data sharing to a new level will require a large, concerted, and costly effort and even then will take many years to accomplish. Quickening the pace is desirable and feasible, but this will not lead to quick solutions still. How much data sharing efforts will contribute to lessening the impact of the corona pandemic further remains to be seen. If they continue to do so, as is to be expected, this should be yet another argument used to drive efforts in data management and sharing. This will require improving existing infrastructures, solving complex legal and ethical questions, agreeing on universal standards, and many more. We should proceed with it thoroughly, but steadily, so that future health emergencies will not have to repeat the call to share data yet another time. Everything will already be in place.