

The life-changing magic of sharing your data

Laurence Hunt^{1,2}

¹ Wellcome Centre for Integrative Neuroimaging, Department of Psychiatry, University of Oxford

² Max Planck-UCL Centre for Computational Psychiatry and Ageing Research, University College London

Email: laurence.hunt@psych.ox.ac.uk

The benefits of data sharing to the scientific community are widely agreed upon. But does data sharing also benefit individual scientists? I argue that data sharing may carry tangible benefits to one's own research that can outweigh any potential associated costs.

It is increasingly expected that scientists not only publish results from their research but also freely share the raw data and analysis pipelines leading to those results. Data sharing is widely considered beneficial by other scientists, journals, funding agencies, and by society as a whole.

It remains less certain whether similar benefits are conferred upon the scientists who share the data¹. Does the ambitious scientist really want to spend their time tidying data and scripts to freely share with their competitors? Mightn't their current findings be undermined, or their future discoveries be scooped? Couldn't this time be better spent advancing their own career, running new experiments and publishing yet more papers?

In this commentary, I argue that data sharing is in fact beneficial to even the most avaricious and self-interested scientists, as well as those who are more munificent and public-spirited in nature. Data sharing may actually lead to the advancement of one's own career, accelerate the pace of one's own scientific discoveries, and increase the impact of one's own research output. Although there are legitimate concerns that must be carefully considered by the scientific community as data sharing is increasingly mandated, there are also concrete benefits among those who share data as well as growing enthusiasm for data sharing.

Lessons from the past

Imperatives to share data in psychology and cognitive neuroscience are not new. In 1999, for example, inspiration was drawn from efforts in genetics, proteomics and X-ray crystallography to launch a major new initiative to share human functional MRI data online: the fMRI data centre (fMRIDC)². Despite significant NSF funding, the fMRIDC quickly suffered blow back from the scientific community upon launch³. It was argued that fMRI was not mature enough for such efforts to be useful; or that unlike other fields, fMRI experimental design meant that each dataset was only optimised to address a single question. There were initial proposals that journals would demand fMRIDC data deposition at the time of publication, but these were mostly rescinded. While over a

hundred studies were ultimately deposited, and several new discoveries were made via reuse of fMRIDC data², the withdrawal of NSF funding in 2006 ultimately led to the centre's closure. Worse, long-term storage was not planned for in the absence of funding: studies deposited on fMRIDC can no longer be readily accessed online.

Such tales provide both caution and instructive lessons for current repositories: to ensure their long-term stability is secure, and their scientific contributions well documented. Successor projects to fMRIDC, such as the International Neuroimaging Data-Sharing Initiative (INDI)⁴ and OpenNeuro.org (formerly OpenfMRI.org) are now backed by more stable long-term infrastructure. Their scientific impact has also started to be quantified. It has been estimated that data reuse from OpenfMRI has saved the US taxpayer \$878,400, for example⁵. A recent analysis identified 913 publications to date that had reused data from INDI, of which 295 had received 10 or more citations (i10-index) and 66 had received 66 or more citations (H-index)⁴. This demonstrates the increasing impact of shared data on the field.

This evidence points towards concrete benefits for scientific progress overall. But if data sharing mandates are to be well received by the community, it is important to consider the views and concerns of those asked to share their data, as well as the greater scientific good.

Benefits to individuals?

The intention to share data appears widespread these days, a notion borne out in a small, informal survey among faculty members of Psychology departments at eight UK Universities (Fig. 1; see also ref. ⁶). A majority of colleagues at these universities now intend to share most of their data (Fig. 1a), but they do so for diverse reasons (Fig. 1b/Fig. 2). Some of the strongest motivations for data-sharing (Fig. 2a) are “other-regarding”, in that the primary beneficiary is the scientific community rather than the individual. There is, for example, the prominent drive towards improving reproducibility, a particularly pressing issue in psychology research at present⁷. Ensuring that data is archived for other researchers to reuse or reanalyse is also a salient concern (Fig. 2a). These other-regarding motivations are undeniably important, but they may lead to a perception that data-sharing is biased towards being motivated by ‘sticks’, and not enough by ‘carrots’. In fact, data sharing can also lead to a wide range of benefits to the individual scientist (Fig. 1c).

Firstly, sharing data with others means that it is likely to be better documented, and will have better long-term stability on a repository than stored locally on a hard disk. Crucially, this means that data will be stably archived for your *own* future reuse. The time invested preparing data and documentation for repositories is a potential drawback, and considered as such by survey respondents (see Fig. 2b). But reuse becomes especially important once former students and postdocs have left the lab. The time invested preparing data and documentation for repositories is then quickly recouped as new lab members arrive. They can rapidly get up to speed on past studies and can try out new ideas, simply by downloading the previously deposited data and analysis scripts. Even if all the key analyses have already been performed, well-documented datasets can provide an invaluable training opportunity for new students. These students will typically obtain a much deeper understanding of a lab's previous work by analysing the original raw data, rather than solely reading the resulting papers.

Secondly, sharing data means that your work can have a greater impact than that obtained from publication alone. By sharing the entire analysis code that takes you to a final conclusion from a given dataset, it becomes possible for other researchers to reverse-engineer your conclusions in a way that may not always be possible from examining the methods section of a paper, despite the authors' best intentions. This means that other researchers understand the conclusions more fully. This may also translate into increased citations, especially if a "data paper" is published to allow credit to be attributed for data reuse. More importantly, however, this deeper understanding allows others to make use of your work and analysis pipeline in designing their own experiments. This can often give rise to unexpected new collaborations that might arise from having shared the data – and this is my personal experience.

It is also becoming increasingly appreciated that a training in data sharing can prove helpful in one's future career (Fig. 1b/Fig. 2a). From a purely self-regarding perspective, data sharing may improve your chances of success when applying to postdoctoral positions or fellowship schemes that place emphasis on data sharing. Certain institutions, such as the Charité in Berlin, have begun to ask applicants to provide evidence of their commitments to open science when applying for faculty positions. Such requests may become more common as the open science movement grows. In the absence of strong incentives, these self-regarding motivations are probably still perceived as slightly less important than other-regarding motivations among the

community. To strengthen motivations for data sharing, funding agencies should perhaps consider whether a successful track record, such as the community widely reusing an applicants' data, should be weighed equally in grant applications against a list of high-profile publications.

Researchers' concerns

While enthusiasm for data sharing appears to be growing, researchers may also have legitimate concerns that curb this enthusiasm⁸. One major reason for potentially not sharing is that data may be confidential or problematic to anonymise. Members of the community have commented that institutional support to navigate issues surrounding legality and confidentiality is sometimes lacking. Nearly half of all survey respondents rated this issue as being a potentially 'very important' issue that may limit willingness to share (Fig. 2b). This may be particularly salient in the United Kingdom at present, following the introduction of the European Union General Data Protection Regulation (GDPR) in May 2018. In the UK and other EU countries, GDPR has meant a change in the law such that 'pseudonymised' data (that is, where a key code held only by the researcher could still be used to decode the subjects' identities if needed) is now considered 'personal' data⁹. Only data where even the researcher would no longer be able to identify individuals would be considered 'fully anonymised', non-personal data. An approach that may work in practice is simply to destroy the key used to identify subjects when the study is complete.

These issues vary, of course, across different countries, studies and institutions. In practice, such issues are best addressed at the outset of a study, when first submitting an ethics application. Fortunately, online resources are increasingly available for researchers to help to navigate these concerns (see <https://open-brain-consent.readthedocs.io/> for an example from neuroimaging).

Another major concern is the potential for being scooped on one's own future publications. This is perhaps particularly pertinent to researchers who invest several years collecting a rare dataset, with the expectation of this data yielding multiple publications. If data has to be shared alongside the first publication, sharing could backfire. While there appear to be relatively few attested cases of such scooping taking place, it is important to acknowledge that this concern is widespread among the community and one that often comes up in discussions both offline and online.

Alarmingly, on occasion, researchers who publicly voice this concern have been vilified on social media and other forums for 'hiding' their data, even if they have a track record of sharing previous datasets¹⁰. Such public shaming must be strongly discouraged if the open science community does not want to become perceived as a self-appointed police force, doling out punishment to researchers who dare to express dissenting views. Researchers who raise concerns about open science should be listened to and debated with, but not personally attacked or humiliated.

Further potential concerns may include that other researchers may perform poorly designed studies with the shared data; take up limited author resources with questions; or pursue projects that the researcher disagrees with. But there is little evidence that this actually occurs, or when it does, has tangible negative effects that outweigh the benefits.

Conclusions

Data sharing can yield benefits to oneself –new collaborations, new papers, and better long-term storage for subsequent reanalysis – but these factors still appear less prominent in researchers' minds than concerns about science as a whole. There is widespread enthusiasm for data sharing, and evidence that it makes a scientific impact. But it would now make sense to ensure the rewards become even more tangible for those who have a sustained track record of data sharing. It is also important that institutions and funding agencies ensure that these scientists are well supported when it comes to dealing with uncertainty concerning the legality of sharing certain datasets.

References

- 1 Gewin, V. Data sharing: An open mind on open data. *Nature* **529**, 117-119 (2016).
- 2 Van Horn, J. D. & Gazzaniga, M. S. Why share data? Lessons learned from the fMRIDC. *Neuroimage* **82**, 677-682, doi:10.1016/j.neuroimage.2012.11.010 (2013).
- 3 A debate over fMRI data sharing. *Nat Neurosci* **3**, 845-846, doi:10.1038/78728 (2000).
- 4 Milham, M. P. *et al.* Assessment of the impact of shared brain imaging data on the scientific literature. *Nat Commun* **9**, 2818, doi:10.1038/s41467-018-04976-1 (2018).
- 5 Gorgolewski, K. J., Wheeler, K., Halchenko, Y. O., Poline, J.-B. & Poldrack, R. A. The impact of shared data in neuroimaging: the case of OpenfMRI.org [version 1; not peer reviewed]. . *F1000Research* **4**, 299 (poster), doi:10.7490/f1000research.1110040.1 (2015).
- 6 Houtkoop, B. L. *et al.* Data Sharing in Psychology: A Survey on Barriers and Preconditions. *Advances in Methods and Practices in Psychological Science* **1**, 70-85, doi:10.1177/2515245917751886 (2018).
- 7 Collaboration, O. S. Estimating the reproducibility of psychological science. *Science* **349**, aac4716, doi:10.1126/science.aac4716 (2015).
- 8 Bishop, D. V. Open research practices: unintended consequences and suggestions for averting them. (Commentary on the Peer Reviewers' Openness Initiative). *R Soc Open Sci* **3**, 160109, doi:10.1098/rsos.160109 (2016).
- 9 Rumbold, J. M. & Pierscione, B. The Effect of the General Data Protection Regulation on Medical Research. *J Med Internet Res* **19**, e47, doi:10.2196/jmir.7108 (2017).
- 10 Barron, D. *How freely should scientists share their data?*, <<https://blogs.scientificamerican.com/observations/how-freely-should-scientists-share-their-data/>> (2018).

Figure Legends

Figure 1. Increasing propensity to share data among psychology researchers at eight UK universities. To get better insight into colleagues' attitudes towards data sharing, I contacted all faculty members of Psychology departments at eight UK Universities (Birmingham, Cambridge, Cardiff, Edinburgh, Oxford, St. Andrews, UCL and York). I provide the anonymous survey responses from the 165 respondents (of 502 contacted) as supplementary data. **(a)** Frequency of sharing past and future data. **(b)** Perceived importance of data-sharing for scientific progress in the field, and career success. **(c)** Benefits observed among those who have previously shared data.

Figure 2. Motivations and concerns when sharing data among psychology researchers at eight UK universities. **(a)** Potential motivations for sharing data. **(b)** Potential concerns/reasons for not sharing data. Plots are sorted by % of survey respondents rating each reason as 'very important'. All items listed in the survey are shown in figures 1 and 2, with the exception of free-text responses to two questions, which are provided as Supplementary Information online alongside raw data and analysis scripts.

Data Availability

Raw anonymised data from the survey are available for download as supplementary data (**Hunt_data_sharing_supplementary_material.zip**).

Code Availability

MATLAB analysis scripts to reproduce figures from the survey are available for download as supplementary data (**Hunt_data_sharing_supplementary_material.zip**).

Competing interests statement

The author declares no competing interests.