Making data meaningful: Guidelines for good quality open data

Andrea S. Towse[1], David A. Ellis[2] & John N. Towse[1]

1. Lancaster University

2. University of Bath

In the most recent editorial for the *The Journal of Social Psychology (JSP)*, Grahe (2021) set out and justified a new journal policy: publishing papers now requires authors to make available all data on which claims are based. This places the journal amongst a growing group of forward-thinking psychology journals that mandate open data for research outputs[1]. It is clear that the editorial team hopes to raise the credibility and usefulness of research in the journal, as well as the discipline, through increased research transparency.

As this editorial appeared, we had a paper accepted for publication in *Behavior Research Methods* (Towse et al., 2020) that reported empirical data on open data practices across psychology. Between 2014 and 2017, we found that public data sharing was uncommon (less than 4% of empirical papers; see Hardwicke et al., 2020, for similar data across the social sciences). We also observed that when data was publicly shared, the majority of datasets were incomplete and had limited reusability. Nearly half were at risk of being orphaned due to the lack of a permanent link between data and research paper (for similar dataset quality issues, see also Hardwicke et al., 2018; Roche et al., 2015). Although the time period for our study already might appear distant there is evidence that, despite researchers being encouraged to include open data, the inclusion and quality of datasets remains disappointingly low. For example, of 5,905 published articles on COVID19, only

[1] See for example, a recent editorial in Behavior Research Methods has also mandated that authors to share materials, data and code as part of any submission (Brysbaert et al., 2020).

13.6% shared their data and only 1.2% shared their data in non-proprietary format such as .csv (Lucas-Dominguez et al., 2021).

This commentary represents a natural and complementary alliance between the ambition of *JSP*'s open data policy and the reality of how data sharing often takes place. We share with *JSP* the belief that usable and open data is good for social psychology and supports effective knowledge exchange within and beyond academia. For this to happen, we must have not just more open data, but open data that is of a sufficient quality to support repeated use and replication (Towse et al, 2020). Moreover, it is becoming clear that researchers across science are seeking guidance, training and standards for open data provision (Roche et al., 2021; Soeharjono & Roche, 2021). With this in mind, we outline several simple steps and point towards a set of freely available resources that can help make datasets more valuable and impactful. Specifically, we explain how to make data meaningful; easily findable, accessible, complete and understandable. We have provided a simple checklist (Table 1) and useful resources (Appendix A) based on our recommendations, these can also be found on the project page for this article (https:doi.org/10.17605/OSF.IO/NZ5WS). While we have focused mostly on sharing quantitative data, much of what has been discussed remains relevant to qualitative research (for an in-depth discussion of qualitative data sharing, see DuBois et al, 2018).

Findable

The best way to ensure readers can obtain data is by storing it in a third-party data repository. Data repositories (see for example, OSF, http://osf.io) provide independent, stable and safe online storage where multiple files (not limited to data, but also materials, preregistrations, preprints and publications) can be uploaded and managed. A digital object identifier (DOI) is provided which can be referenced in related publication(s). This simple

link provides a quick, easy and enduring route to finding relevant data. It is citable making it is easy to claim ownership of the data, credit authors and track use. Additionally, settings can be public or private and include files that are read-only or editable. *JSP* now require an Open-Ended Registration which can be provided by the OSF. Not to be confused with pre-registration, Open-Ended Registration creates a frozen snapshot of a project. All project files are thereby locked and time stamped. This preserves the data files at the point of publication of the associated research paper. We describe the step-by-step process for doing this within the accompanying OSF project (https:doi.org/10.17605/OSF.IO/NZ5WS).

In reality, Towse et al. (2020) found numerous problematic instances where data files were stored on an individual author's website, on a lab website, or on a publisher's own website. But web addresses change, researchers retire or move institution and journals change publisher. In these cases, URLs linking to data cited in research papers became obsolete and, in some cases, data were irretrievable. Similarly, whilst many authors claim with good faith that 'data is available on request', we know that this too can be problematic (Savage & Vickers, 2009; Vanpaemel et al., 2015; Vines et al., 2014). None of these options are permitted when submitting work to *JSP*, indeed *JSP* requires authors to use a data repository (see Grahe, 2021, adapted from Nosek et al.'s (2020) Transparency and Openness Promotion Guidelines). Each paper requires a public data availability statement which provides a link to any data. Such clear data availability statements allow data to easily reach a wider audience. A DOI can also be usefully placed in a results or methods section of a paper, with submissions to *JSP* now requiring a data availability statement in the methods section. However, it is important to note that DOIs can easily be missed if not carefully positioned. Towse et al. (2020) found occasions when links to data in footnotes do not appear in every version of a paper (e.g., in a PDF but not a web version).

Accessible

An obvious but important accessibility issue is the ability to open the data file itself. Some types of data files will be difficult to access due to license restrictions of proprietary software (sometimes software versioning even prevents backwards compatibility, so old files may not be accessible even with a current software license). Saving data in a simple non-proprietary format (e.g., .txt or .csv) provides the flexibility to allow access for all. In addition, it is easy to store *multiple* files in online data repositories. Given that some software enables the storage of rich information that is easy to manipulate, there are clear advantages for interested parties who do have access to the licensed software. Therefore, we recommend that an author uploads the data both in its original proprietary format and also in a separate, more accessible file format.

Some additional details to take into consideration, which can further enhance data accessibility, relate to file layout and language. Simple, but important details like ensuring columns are lined up under the correct headings are quick to check and rectify where necessary, yet vitally important for the reader to understand the data. Similarly, the language used in the data file should match the language used in the published article. Since *JSP* is published in English, an English dataset (or with suitable translations) means all data will be accessible.

Complete

To ensure datasets make valuable contributions to science, all essential data need to be shared. In other words, the minimum requirement should be for all data to be provided at a level that allows for any reported analysis to be verified (analytic reproducibility, see for example Hardwicke et al., 2018). However, data sharing can be far more functional when looking beyond this minimal threshold. The best-case scenario would involve the sharing of

all raw data. Many datasets require some level of processing before analyses (e.g., an average response time across trials) in such cases, as noted above, it is possible to upload multiple files where raw data files can sit alongside processed data files (e.g., provide the average response time variable but also the trial-by-trial data from which this is derived, or the total score from a validated scale but also the item responses) affording the greatest transparency and opportunity for re-analysis.

Imagine a future researcher uncovers a new and theoretically motivated way of scoring a psychological scale from individual responses that materially improves its psychometric properties and wants to understand whether this might reveal some hitherto hidden aspect of a dataset. Or consider the utility of a new statistical model that can address data questions that that were simply not fathomable when the work was originally published (see for example, Orben & Przybylski, 2019, Shaw et al., in press). Future use of deposited data in these cases may depend critically on the decisions made to share data at the most detailed level possible.

Unsurprisingly, common problems encountered with the completeness of a data file include missing participants, missing participant details, missing conditions and missing values in certain cells. Of course, there are often valid reasons why there may be a missing value in a data set (e.g., no response from a participant) but these need to be easily identified and explained in a *Results* section or a readme file (a separate document, saved in a non-proprietary format alongside the data, which provides essential details not in the data file itself, e.g., expands abbreviations, provides information on units, explains missing data or data transformations etc.). These explanations should also mention software specific values generated in missing cells (e.g., 'NA' in R or '99' in SPSS) not only is this information useful for those looking at the data file but is also useful to note in case that convention

changes over time. Relatedly, Towse et al. (2020) noted that as well as missing participants, there were datasets that included too many participants. Within psychology, participants are commonly removed before analysis for legitimate reasons (e.g., failure to follow instructions), and providing data from excluded participants is a potentially good, transparent practice. Yet, if it is not possible to identify which participants are removed from the analyses then the data file becomes much less usable. Therefore, researchers need to make clear, either as part of additional documentation, or within marked-up analysis code how data has been processed and prepared before analysis. This becomes even more important when variables are generated following significant processing or during the development of novel methodologies (e.g., Andrews et al., 2015). Of course, there may also be occasions that it is important to omit key elements of the data. We address these in sections below on sensitive data and anonymity.

Well-described

To have an impact on the research community through data sharing, simply providing or explaining why data has been omitted is not enough. The data needs to have clear metadata, that is well described within the data file or in a separate readme file. Although a column labeled, for example, *t22_b* might make sense to the authors, it requires the reader to make a set of inferences that might not always be correct. Files, columns, rows and any other material should all be clearly labelled complete with measurement units and any other key information[2]. Indeed, some variables may need further information to explain how they have been calculated (e.g., a mathematical transformation of another variable).

[2] It is important to keep in mind that although some proprietary software allows for metadata to be stored within the file (e.g., SPSS), it is not always possible to access this information without the software itself. This links to the previous point recommending that a version of all files in a non-proprietary and therefore accessible format should be stored in a repository

Ideally, a data file will be described to a standard that it can be understood as a stand-alone document. One simple and effective way to achieve well described data is to have a separate readme file archived alongside the main dataset that … (see for example, Lynott et al., 2019). Including the annotated or commented analysis code can also be highly valuable because it helps the mapping between data structure and reported results, as well as making explicit data transformations, data wrangling, and data processing operations.

Further considerations

It would be nice to think that all that is needed is a carefully curated data file uploaded to a repository. In many cases this is true. Sometimes, though, there are more complex considerations. In the following sections, we provide thoughts on sharing sensitive data, ethics and the ongoing challenges associated with new and emerging forms of data.

Sharing sensitive data

JSP publishes important research that involves, in the majority of cases, data that should be straightforward to share in full. However, we recognize that social psychology can also involve more sensitive data where there may be commercial, legal, medical, ethical or other constraints over the sharing or use of data (e.g., Joel et al., 2018). There may also be cases where, because of the nature of the research or how it is funded, other companies, organizations or individuals may exercise the right to control the data. Simply talking to some of these parties about the importance of open data and how it can be done sensitively and/or anonymously may alleviate their concerns. The time at which these conversations take part may be key. If data sharing discussions occur at the start, there is an intention to share data openly and it can be a point of negotiation with the funding body or research participants.

Accordingly, start from the default position that there are usually ways to share sensitive data. It is possible to use synthetic data whereby an artificial dataset is created that mimics the properties of the raw data (Quintana, 2020). Use of a mediating process is another potential solution. As Joel et al. (2018) point out, creating a "Shiny App" allows researchers to specify and run analyses on a dataset without having direct access to it (see for a recent example, Shaw et al., in press). Datasets can be licensed so that researchers can specify how the data are used (e.g., Carroll, 2015). Alternatively, restricting data access or requiring end-user agreements mitigate the potential harm of inappropriate use, however this of course qualifies the notion of *public* data-sharing. Even partial data sharing, where only less sensitive elements are shared whilst shielding other elements, is better than no data sharing. Identifying data as sensitive should be the start of a reflection or conversation about data sharing, not the point of abandonment.

Ethics and Emerging Forms of Data

Ethics forms should always state how data will be stored and shared and it is straightforward to include a statement about open data. Evidence suggests that participants are not discouraged from taking part in a study when informed that their anonymized data will be shared openly (Eberlen et al., 2019). We suggest including a copy of the ethics form in a project repository to provide everyone accessing the archive clarity about permissions. Clear licensing of the dataset will help users understand relevant rights and restrictions.

There are cases where, despite anonymizing aspects of the data, individuals may still be identified from the content of the data (e.g., geolocation data, indirect identifiers). In some interconnected settings, the privacy of one individual can also be affected by the decisions of others, giving rise to *interdependent privacy* (Olteanu et al., 2016). In these instances, researchers may be reluctant to openly share the data in order to protect their

participants. As recommended above when discussing sensitive data, it may be possible to share sections of data that obscure any identifying features. However, this is an area that continues to evolve especially when it comes to new and emerging forms of data. For example, the power of individual or combinations of new variables often only becomes clear long after the data has been collected. In these instances of intensive data collection that involves new digital systems or devices, some researchers have suggested that participants should be able to remove their data from researchers' repositories at any point (even after is has been collected) (Dennis et al., 2019a). This raises a number of challenges concerning data access for researchers (who would need access to data via differential privacy systems that ensured participants were not identifiable) while simultaneously allowing participants to access their own data at any time and make decisions concerning any future use.

Different solutions have implications for open science practices including transparency and ease of replication as data sets like these could change at any time after publication (Dennis et al., 2019b). This makes version control important. Similar to software development where features are added or removed, if a dataset is updated (e.g., additional columns added, new data or corrections), then it is important to know when this happened and what exactly has changed because it can affect aspects of future analyses. This can be done, for example, within an OSF project page. Researchers can upload an amended file with the same name in the same location and this will replace but not overwrite the previously stored file - which therefore remains accessible. We recommend that any file changes should be explained in a separate readme file.

Ideally, social psychology should be at the heart of opportunities and challenges that emerge from existing as well as new and emerging forms of data. Our recommendations point towards a future where the field could become more impactful by addressing these

challenges that are important within and beyond psychological science. For example, increasingly diverse data sets that include psychologically relevant variables are being championed across central government and can support wider knowledge exchange with effective policy collaboration and data analysis[3].

Conclusion

Authors had previously been encouraged to provide materials and data by a variety of journals including *JSP* (Grahe, 2014). This also allowed authors to earn badges that rewarded open scholarship. The new open data policy introduced by *JSP (Grahe, 2021)* sets the scene for a more trustworthy and progressive research environment. Our recommendations for high-quality datasets scaffold this process, ensuring that the efforts to include data are not wasted and the datasets have the desired impact. Of course, the issues highlighted, and recommendations provided are not unique to social psychology, but remain essential for all scientific progress. These are summarized in a useful checklist (Table 1).

We recognize that researchers' entry point may vary and we are not able to provide definitive guidelines for every aspect of open data. However, our advice can be useful for both those preparing to share data for an already written paper and those planning in advance for how best to incorporate data sharing into a new research project. Many of the suggestions outlined here are good practice for researchers irrespective of data sharing requirements. *JSP* are simply requiring authors to share this information publicly. However, we also acknowledge that these recommendations are not exhaustive, they are simple and concise based on current observations in psychology and beyond. The accompanying

[3] https://re.ukri.org/knowledge-exchange/knowledge-exchange-framework/

resources are therefore flagged “at the current point in time” and whilst some things should be constant, we expect other things to change given the dynamic nature of data sharing processes. Other issues will become apparent with subsequent successes. It’s very tempting to think of preparing open data as a static process (use rules A, B, and C) however we need to keep in mind that the way data might be re-used or processed can change very quickly (Lazer et al., 2020). We will endeavour to add helpful resources to our project page in the future (https:doi.org/10.17605/OSF.IO/NZ5WS). As members of a growing and supportive community we believe in the usefulness of open data for the advancement of science. Indeed, we are reminded of this in a recent paper arguing how open and collaborative practices are vital for knowledge generation (Ellemers, 2021). To do this well, we need open datasets to be high quality.

References

Andrews, S., Ellis, D. A., Shaw, H., & Piwek, L. (2015). Beyond self-report: tools to compare estimated and real-world smartphone use. *PLOS ONE*, *10*(10), e0139004.

Brysbaert, M., Bakk, Z., Buchanan, E. M., Drieghe, D., Frey, A., Kim, E., ... & Yap, M. (2021). Into a new decade. *Behavior Research Methods, 53*, 1-3.

Carroll, M. W. (2015). Sharing Research Data and Intellectual Property Law: A Primer. *PLOS Biology 13*(8): e1002235.https://doi.org/10.1371/journal.pbio.1002235

Delhove, M. & Greitemeyer, T. (2021). Violent media use and aggression: Two longitudinal network studies. *The Journal of Social Psychology*, DOI: 10.1080/00224545.2021.1896465

Dennis, S., Yim, H., Garrett, P., Sreekumar, V., & Stone, B. (2019a). A system for collecting and analyzing experience-sampling data. *Behavior Research Methods, 51*(4), 1824–1838.

Dennis, S., Garrett, P., Yim, H., Hamm, J., Osth, A. F., Sreekumar, V., & Stone, B. (2019b). Privacy versus open science. *Behavior Research Methods*, *51*(4), 1839-1848.

DuBois, J. M., Strait, M., & Walsh, H. (2018). Is it time to share qualitative research data? Qualitative Psychology, 5(3), 380-393. http://dx.doi.org/10.1037/qup0000076

Eberlen, J. C., Nicaise, E., Leveaux, S., Mora, Y. L. & Klein, O. (2019). Psychometrics Anonymous: Does a Transparent Data Sharing Policy Affect Data Collection? *Psychologica Belgica, 59*(1), 373–392. http://doi.org/10.5334/pb.503

Ellemers, N. (2021), Science as collaborative knowledge generation. *British Journal of Social Psychology*, 60, 1-28. https://doi.org/10.1111/bjso.12430

Grahe, J. E. (2014). Announcing open science badges and reaching for the sky. *The Journal of Social Psychology,* 158(1), 1. https://doi.org/10.1080/00224545.2018.1416272

Grahe, J. (2021). The necessity of data transparency to publish. *The Journal of Social Psychology,* 161(1), 1-4. https://doi.org/10.1080/00224545.2020.1847950

Hardwicke, T. E., Wallach, J. D., Kidwell, M. C., Bendixen, T., Crüwell, S., & Ioannidis, J. P. A. (2020). An empirical assessment of transparency and reproducibility-related research practices in the social sciences (2014-2017). Royal Society Open Science. https://doi.org/10.1098/rsos.190806

Hardwicke, T. E., Mathur, M. B., MacDonald, K., Nilsonne, G., Banks, G. C., Kidwell, M. C., Hofelich Mohr, A., Clayton, E., Yoon, E. J., Henry Tessler, M., Lenne, R. L., Altman, S., Long, B., & Frank, M. C. (2018). Data availability, reusability, and analytic reproducibility: Evaluating the impact of a mandatory open data policy at the journal Cognition. Royal Society Open Science. https://doi.org/10.1098/rsos.180448

Joel, S., Eastwick, P. W., & Finkel, E. J. (2018). Open Sharing of Data on Close Relationships and Other Sensitive Social Psychological Topics: Challenges, Tools, and Future Directions. *Advances in Methods and Practices in Psychological Science*, 86–94. https://doi.org/10.1177/2515245917744281

Lazer, D. M., Pentland, A., Watts, D. J., Aral, S., Athey, S., Contractor, N., … & Wagner, C. (2020). Computational social science: Obstacles and opportunities. Science, 369(6507), 1060-1062.

Lucas-Dominguez, R., Alonso-Arroyo, A., Vidal-Infer, A., & Aleixandre-Benavent, R. (2021). The sharing of research data facing the COVID-19 pandemic. *Scientometrics*, 1-16.

Lynott, D., Walsh, M., McEnery, T., Connell, L., Cross, L., & O'Brien, K. (2019). Are you what you read? Predicting implicit attitudes to immigration based on linguistic distributional cues from newspaper readership; a pre-registered study. *Frontiers in Psychology* DOI: 10.3389/fpsyg.2019.00842

Nosek, B. A. , Alter, G. , Banks, G. C. , Borsboom, D. , Bowman, S. D. , Breckler, S. J. , … DeHaven, A. C. (2020, November 1). Transparency and openness promotion (TOP) guidelines . Retrieved from osf.io/9f6gx

Olteanu, A. M., Huguenin, K., Shokri, R., Humbert, M., & Hubaux, J. P. (2016). Quantifying interdependent privacy risks with location data. *IEEE Transactions on Mobile Computing*, *16*(3), 829-842.

Orben, A., Przybylski, A.K. The association between adolescent well-being and digital technology use. Nat Hum Behav 3, 173–182 (2019). https://doi.org/10.1038/s41562-018-0506-1

Quintana, D. S. (2020). A synthetic dataset primer for the biobehavioural sciences to promote reproducibility and hypothesis generation. *eLife* 9: e53275 https://doi.org/10.7554/eLife.53275

Roche, D., Berberi, I., Dhane, F., lauzon, F., Soeharjono, S., Dakin, R., & Binning, S. (2021, May 18). The quality of open datasets shared by researchers in ecology and evolution is moderately repeatable and slow to change. https://doi.org/10.32942/osf.io/d63js

Roche, D. G., Kruuk, L. E. B., Lanfear, R. & Binning, S. A. (2015). Public Data Archiving in Ecology and Evolution: How Well Are We Doing? *PLoS Biology, 13*(11), e1002295. https://doi.org/10.1371/journal.pbio.1002295

Savage, C. J. & Vickers, A. J. (2009). Empirical study of data sharing by authors publishing in PLoS journals. *PloS one*, *4*(9), e7078.

Shaw, H., Taylor, P., Ellis, D. A., & Conchie, S. (2021, March 26). Behavioral consistency in the digital age. https://doi.org/10.31234/osf.io/r5wtn

Soeharjono, S. & Roche, D. G. (2021). Reported individual costs and benefits of sharing open data among Canadian academic faculty in ecology and evolution. BioScience. biab024. http://dx.doi.org/10.1093/biosci/biab024

Towse, J. N., Ellis, D. A., & Towse, A. S. (2020). Opening Pandora's Box: Peeking inside Psychology's data sharing practices, and seven recommendations for change. *Behavior Research Methods.* Advance online publication: https://doi.org/10.3758/s13428-020-01486-1

Vanpaemel, W., Vermorgen, M., Deriemaecker, L., & Storms, G. (2015). Are we wasting a good crisis? The Availability of Psychological Research Data after the Storm. *Collabra, 1(1)*, Art. 3. https://doi.org/10.1525/collabra.13

Vines, T. H., Albert, A. Y., Andrew, R. L., Débarre, F., Bock, D. G., Franklin, M. T., … Rennison, D. J. (2014). The availability of research data declines rapidly with article age. *Current biology, 24(1)*, 94–97.

Table 1. Checklist for good quality data files

Is your data file:

- findable?
    - Stored in a permanent repository
        - Settings allow public, read-only access
        - Preferably in OSF with Open-Ended Registration
    - Accompanied with a clear data accessibility statement in the methods section
        - Include the DOI for the data

- accessible?
    - Saved in a non-proprietary format (e.g., .txt)
    - Written in the same language as the as the publishing language of the journal
    - Accompanied with a consent form (to ensure data is being used in line with participants approval)

- complete?
    - All raw data provided
    - All participants (ensuring that any participants removed from the analysis are easily identifiable)
    - All variables
    - Missing cells are accounted for and easily identified
    - Analysis code (annotated if necessary)

- well described?
    - Files labelled clearly
    - Metadata included (in same file or separate readme file)
        - Columns/rows labelled
        - Clear measurement units
        - Further explanation where necessary
    - Version details (if/when files have been updated and how)

Appendix A. Further resources. Links active at the time of writing. This list will be updated online as more resources are developed: https:doi.org/10.17605/OSF.IO/NZ5WS

Findable

Registry of Research Data Repositories:
https://www.re3data.org

An overview of OSF:
Foster, E., & Deardorff, A. (2017). Open Science Framework (OSF). *Journal of the Medical Library Association, 105*(2), 203–206. doi:https://doi.org/10.5195/jmla.2017.88

Complete

Additional analysis of how to treat missing data:
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0182362

Providing analysis code:
Blischak JD, Carbonetto P and Stephens M. Creating and sharing reproducible research code the workflowr way. *F1000Research* 2019, **8**:1749
(https://doi.org/10.12688/f1000research.20843.1)

Well described

R package 'codebook'
https://rubenarslan.github.io/codebook/articles/codebook_tutorial.html
https://doi.org/10.1177/2515245919838783

creating a codebook in SPSS
https://libguides.library.kent.edu/spss/codebooks
https://stats.idre.ucla.edu/spss/modules/labeling-and-documenting-data/

Collating data for a meta-analysis - dataMaid
https://cran.r-project.org/web/packages/dataMaid/dataMaid.pdf

Addressing data sensitivity issues

An introduction to data anonymization
https://www.ukdataservice.ac.uk/manage-data/legal-ethical/anonymisation/qualitative.aspx

Anonymization in R
http://psychbrief.com/anonymous-data-r/
https://bookdown.org/martin_monkman/DataScienceResources_book/anonymity-and-confidentiality.html

synthetic data, using R synthpop
https://www.synthpop.org.uk/get-started.html

synthetic data for SPSS using GRD
https://www.tqmp.org/RegularArticles/vol10-2/p080/p080.pdf