IJDC | *Brief Report*

Researchers and Research Data: Improving and Incentivising Sharing and Archiving

Minna Ventsel Beth Montague-Hellen

The Francis Crick Institute

Abstract

There has been a lot of discussion within the scientific community around the issues of reproducibility in research, with questions being raised about the integrity of research due to failure to reproduce or confirm the findings of some of the studies. Researchers need to adhere to the FAIR (findable, accessible, interoperable, and reusable) principles to contribute to collaborative and open science, but these open data principles can also support reproducibility and data integrity. This article uses observations from data sharing and research integrity related activities, undertaken by a Research Integrity and Data Specialist at the Francis Crick Institute, to discuss potential reasons behind a slow uptake of FAIR data practices. We then suggest solutions undertaken at the Francis Crick Institute to improve the integrity of research from a data perspective, which can be followed by other institutes and universities. One major solution discussed is the implementation of a data archive system at the Francis Crick Institute to ensure long term data integrity, comply with our funders' data management requirements, and safeguard our researchers against any potential research integrity allegations in the future.

Submitted 30 May 2024 ~ *Revision accepted* 12 January 2025

Correspondence should be addressed to Minna Ventsel. Email: minna.ventsel@crick.ac.uk

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

http://dx.doi.org/10.2218/ijdc.v19i1.983
DOI: 10.2218/ijdc.v19i1.983

Introduction

Being able to rely on the validity and reliability of results in research is crucially important. Particularly so when those research results are built upon to create policy, develop medicines or advance our understanding of the world. One method of verifying the reliability of results is to ensure that they are reproducible.

Reproducibility in research refers to the ability to replicate results using the original researcher's materials and procedures (House of Commons report, 2023). As new knowledge is built on already published results, the ability to trust research findings is enhanced by ensuring the reproducibility of every step in the chain. Openness of data, code, and methodologies is key for testing reproducibility and ensuring that the data is as trusted as possible.

The galvanising phrase "reproducibility crisis" (Baker, 2016), which has become a hot topic in recent years, highlights the gravity of the situation in which many research findings are difficult or impossible to reproduce. This in turn can lead to ineffective, or even dangerous, interventions and other applications of research findings (Diaba-Nuhoho & Amponsah-Offeh, 2021). However, some researchers believe this term is hyperbolic and counterproductive, making the issue too big and intimidating to address (Springer Nature, 2021).

Whether or not it is a crisis, there is definitely room for improvement (Munafò et al., 2017). According to Baker (2016), 70% of researchers had "tried and failed to reproduce another scientist's experiments." This finding demonstrates the need for clearer and more transparent research practices. While research irreproducibility could be caused by a myriad of reasons (e.g., poor experimental design, selective reporting), the lack of sufficient data, code, description of methods, and accompanying data documentation is a common cause. In addition to improving reproducibility of research and therefore improving the trust of researchers in the results, sharing data has also been found to increase public trust in science (Rosman et al., 2022; Wingen, Berkessel & Englich, 2019). This is especially relevant at present when many in research and higher education sectors are actively fighting against misinformation and where there is a growing scepticism about science in the general public (Nyhan, Porter & Wood, 2022).

Data sharing challenges for reproducibility

The main challenges for reproducibility include lack of access to the required data, metadata, and documentation, and there are some known reasons behind this (Hahnel et al.,~~The State of Open Data~~, 2023). For instance, collating the data and making it usable for others is time consuming. It is easier to state "request from author" in the manuscript to satisfy the existence of a data availability statement – a section often mandated by journals. However, it is quite rare to actually get your hands on the data when requested (Gabelica, Bojčić & Puljak, 2022). Gabelica et al. found that among manuscripts where authors indicated a willingness to share data on request, only 6.8% provided the requested data. Compounding this finding, it has been suggested that the availability of data declines over time, where the odds of a dataset still existing falls by 17% per year (Vines et al., 2014). In one study the loss of data was also stated to be the second most common reason for declining data sharing by researchers (Tedersoo et al., 2021). Even more worryingly, when an Editor-in-Chief of Molecular Brain requested raw data from 41 authors with manuscripts categorised as "Revise before review," 21 of them withdrew without providing any data, and 19 out of the remaining 20 manuscripts were rejected due to insufficient raw data, resulting in just one acceptance (Miyakawa, 2020). Miyakawa suggests that raw data may not have existed in some of these cases.It could also be that the authors did not want to spend time collating the data and withdrew their manuscript as a result.

Researchers may also believe that sharing data in the supplementary materials section of the manuscript is sufficient. A commentary by Santos et al. (2005) discussed how supplementary and raw data in research articles should be deposited in a repository instead of in a supplementary materials section at the end of the manuscript, and how journals should adopt policies to change this habit. However, 19 years later, while sharing in repositories has increased, supplementary materials are still omnipresent in research articles and depending on the journal, can be the most popular place to include data (Colavizza et al., 2020). The data in supplementary materials is not considered FAIR (findable, accessible, interoperable, and reusable) because data are presented in a range of, sometimes unsuitable, file formats. There are also no standards for data organisation, no metadata, and the dataset does not get a permanent identifier, making the data uncitable and more perishable. Although supplementary data is not FAIR, and is sometimes behind a paywall and hence not even open, some researchers may believe that data in supplementary materials counts as sharing because data availability statements in many journal submission forms allow a statement pointing to supplementary materials section: e.g. “The minimum dataset that is needed to interpret and verify research is provided in this paper and the supplementary material” (Federer et al., 2018). One reason for this could be due to contradicting or unclear guidance provided by either funders, publishers or research centres (Montague-Hellen & Montague-Hellen, 2023). In other cases, differing opinions across nationalities and sectors may lead organisations to inform their researchers that supplementary information is an acceptable way to share data if no suitable subject specific repository exists (National Institute of Health, 2023).

Lack of motivation due to researchers being unclear about the benefits of sharing can also inhibit sharing of data (Hahnel et al., 2023). Researchers may not be aware that sharing data contributes to improving the reliability of scientific knowledge, and can also benefit them directly, e.g., by attracting more citations (Colavizza et al., 2020). It also cannot be assumed that all researchers are familiar with repositories and data sharing in general (Stuart, 2018). Being aware of the open science principles and following them is strongly influenced by the laboratory culture of the researcher and the advice they get from their supervisors (Zuiderwijk, Shinde & Jeng, 2020). The role of journals is also important in setting clear data sharing policies, and providing guidance on their platforms to maximise data availability and reusability (Vasilevsky et al., 2017).

Open data initiatives at the Francis Crick Institute

The Francis Crick Institute is a biomedical research centre that was opened in 2016. The key values of this Institute are to be bold, collegial and open. As a result, several steps have been taken to increase and encourage FAIR data sharing as quickly as possible, and through this, research integrity.

Hiring of specialist staff

The initial and crucial step in the project to promote and prioritise open data and research integrity was the hiring of new specialist staff to oversee the project. Within the Library and Information Services team a research data steward was hired. The steward oversees any data sharing related activities with a particular emphasis on improving research integrity. The tasks in this role include reviewing data deposits to the institutional repository, offering advice on data management, and promoting data sharing and archiving.

At the same time the Francis Crick Institute also hired a research integrity officer within the Research Management team who oversees other areas of research integrity such as policy, as well as research integrity breaches. However, this article is primarily concerned with the work which has been carried out based on research integrity expertise in the library.

Developing guidelines for data deposits

Even if data is openly shared, it's not always reproducible or reusable due to poorly described metadata or lacking data documentation. To combat this issue, all datasets deposited in Crick's institutional repository Figshare are reviewed by the data steward. To standardise the process of reviewing, we developed a set of clear guidelines to improve deposits. Metadata which will no longer be accepted includes:

- Unspecific deposit titles, e.g. "Data file", "Dataset", or "Results". There tens of thousands of deposits with identical titles in repositories. We are also discouraging the use of the manuscript title as the sole title of a dataset. The best practice is to give an informative name to the deposit that describes the data, e.g., "Figure 2 electron microscope images with cells of interest". The minimum requirement is a title such as "Data associated with <Paper>"

- Uninformative/nonsense file names, e.g. "pgp.0329483.a001" or "Notes". These file names tell us nothing of what they contain. If another researcher wants to use this data for their study, it might take a while to identify what each of the files contains if file names are not clear. We encourage more informative file names if possible, e.g. "Figure 4A_histogram_raw_data.xlsx" or "20230712_ProjectA_Ex1Test2_v03.csv". If there are too many files to rename (e.g., hundreds of files generated by instruments), the researchers are encouraged to include documentation to explain the logic behind the file names

- Not linking the deposit to the manuscript. There are plenty of data deposits in repositories that include no indication of which manuscript the data supports, making the data difficult to use for those who find the dataset before they have seen the manuscript. Datasets need to include a DOI link to the related manuscript, or at least a citation of the manuscript in the description section if the manuscript has not been published yet

- Uninformative deposit descriptions, e.g. "Results of image analysis". Researchers should provide an informative description section within the deposit, especially if no documentation has been included to describe the data

- Datasets with no documentation or limited documentation, e.g. not explaining excel table variables, or generally not providing any information that would help another researcher to reproduce the study. We encourage adding a text document with the deposit that includes any relevant information about the data to increase reusability

Although the number of deposits is still small, we have seen an increase in deposits that include datasets, software, figures, media, posters, and presentations from 8 (Feb 2022–Feb 2023) to 66 deposits (Feb 2023–Feb 2024). The quality of deposits in Crick's Figshare repository also appear to be improving and becoming more FAIR, although this is based on administrator perceptions and has not been marked against any standardised criteria. We believe that reviewing deposits before publication, as well as data management training, has helped to nudge researchers toward transparent and FAIR data sharing and prevent poor quality data deposits. However, future qualitative research into researchers' experiences would help us to evidence this.

Provision of training and guidance

In the 2023/24 academic year compulsory training in Open Research and Research Integrity was introduced for first year PhDs that includes guidance on good data management and data sharing practices. Whilst PhD students at the Francis Crick Institute can access training from their degree awarding institutions, this was the first time that training had been integrated within their on-site training portfolio, signalling the importance of data sharing to the institution.

New guidance was also developed for the intranet pages with clear instruction on what was considered acceptable levels of open data and metadata at the Francis Crick Institute, together with materials that take the researcher through the whole depositing process in the institutional repository. General pages describing the institutional data repository, how to share research data and how to archive data have been the most popular pages for researchers to visit within the newly created research data management and sharing guidance section. Visitor metrics of these webpages have been encouraging.

Encouraging asking for help

The Francis Crick Institute prioritises supporting others in producing world-class science and in upholding our value of collegiality. The Library and Information Services team supports this through encouraging researchers to ask for support with data sharing, and in making this process as easy as possible.

Researchers submitting papers at the Institute are asked to inform the library team by filling out a form providing details about their recently submitted manuscript. To bolster support for data sharing, this existing form has been edited to include questions about whether a data availability statement has been included in the manuscript and whether data has been shared. A new option has been enabled to allow the researcher to ask for help with depositing data or writing a data availability statement. This support has been used by some of the Francis Crick researchers to improve their knowledge in open data best practice, leading to new data deposits in Crick's Figshare repository. We have also observed an increase in data availability statements in journal articles by Crick researchers from 82% in 2022 to 89% in 2024.

Personal outreach

We have carried out personal outreach after each manuscript submission where data sharing had not already occurred to the standard that we expect. This has led to a small increase in data deposits in our institutional repository.

Reaching out to researchers also helps to highlight the existence of staff at our Institute who are able to help the researchers with data related questions as researchers might not be aware that such support exists. Frequently we find that researchers are either too busy, or do not consider, searching for professional services support on their own. By directly reaching out rather than waiting for them to come to us we remove this barrier ensuring that across the Institute all researchers have access to the same support system.

Motivating researchers

Motivating researchers to share their data can also increase the number of data deposits. On our intranet pages for data sharing, we have highlighted several benefits of data sharing, such as increased citations, as found by Colavizza et al. (2020), or that besides helping to further science, sharing in a repository can also facilitate discovery of their research through metadata. We have also highlighted the benefit of having a citable DOI when reaching out to researchers about depositing data. These benefits are repeated in our training sessions, talks and outreach materials, ensuring that researchers know that we are not simply asking them to share their data for altruistic reasons.

Data Archiving Initiative at the Francis Crick Institute

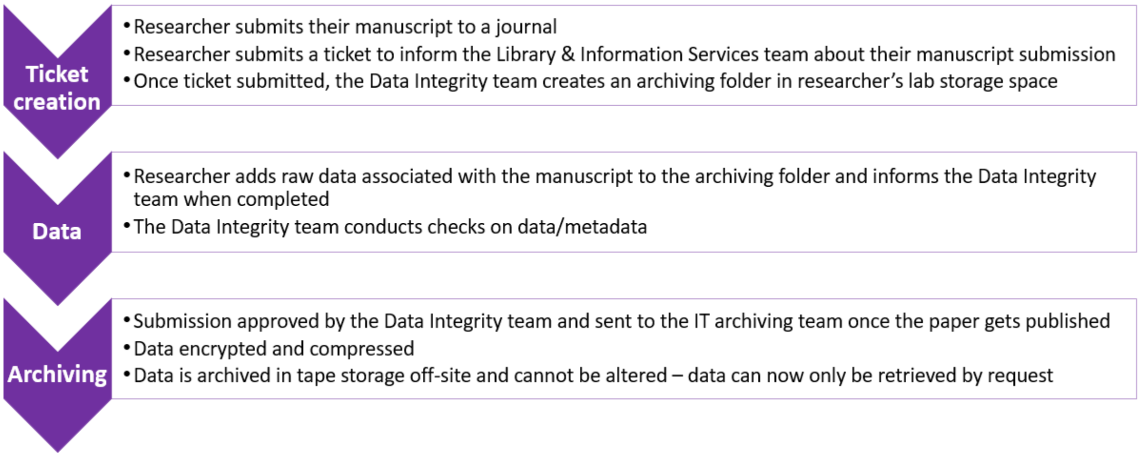
Although the Francis Crick Institute encouraged open data sharing as much as possible, there are cases where this is simply not possible. Both clinical and commercial data is created at the Crick, and even though some external databases have secure and robust systems for housing this data e.g., the European Genome-phenome Archive (EGA), this is not available for all types of data.

To facilitate secure retention of this data, even when it cannot be open, the decision was made by senior leaders that all data should be stored safely with our offsite data storage system, ensuring that all papers published by the Institute were supported by data that was as FAIR as possible, i.e., data in the storage system is findable through the ticket number attached to each archive submission and through article data availability statements; accessible via a request to the IT to retrieve it from the storage system; interoperable by encouraging the use of standard and open data types and the inclusion of detailed and standard metadata files; and reusable as the datasets are accompanied by documentation to provide any further details not already described in the manuscript. All researchers at the Crick who publish papers are required to archive all their raw data associated with their publications in line with funder expectations. This internal data archiving system was created in collaboration between the library, IT Operations and the research integrity staff. The archive was modelled on the data archive at the CRUK Scotland Institute. The internal archive differs from an openly accessible repository in that although both are used for storing data post publication, the archive is a closed and secure system allowing for the retention of all raw data, not just the anonymised and non-proprietary data which can be shared openly. While this archive is a closed internal storage system and may seem like a step in an opposite direction having just been discussing open data sharing, there are several reasons why a closed archive can be beneficial.

Not all research data can be made immediately open in a data repository. This can be due to the sensitive or personal nature of the data, commercial constraints, or the size of the data files. Data catalogues can be created to advertise the existence of this data, but where it is inappropriate to store it in a repository, even with an embargo, it can be a difficult task to ensure that data is stored securely and with enough metadata to be checked or reused if the need arises. It is expected that archiving will also indirectly increase data deposits in public repositories, because as researchers will already have their data organised for archiving, it will be quick and easy to add the dataset to the institutional repository. To encourage data sharing after archiving, researchers will be followed up with an email to suggest additionally depositing data to the institutional repository.

The data archive is looked after by the data integrity team, which is a sub-team of the research integrity team and only deals with data related activities. The archiving process itself is straightforward and is initiated when the researcher informs the data integrity team of their manuscript submission (see Figure 1). An archiving folder is created for the researcher within their laboratory storage space into which they will add their data, following a suggested folder hierarchy, e.g., folders for raw data associated with Figure 1, Figure 2, Table 1 and so forth.

Once the researcher has finished copying their data into the archiving folder, the data integrity team conducts checks on data and metadata, ensuring that all the necessary data is there, the file and folder names are informative, and that documentation is included to provide any necessary details about the data and make it reusable. Once the archive submission is approved by the reviewer, data is archived in tape storage off-site and cannot be altered.



Ticket creation

- Researcher submits their manuscript to a journal
- Researcher submits a ticket to inform the Library & Information Services team about their manuscript submission
- Once ticket submitted, the Data Integrity team creates an archiving folder in researcher's lab storage space

Data

- Researcher adds raw data associated with the manuscript to the archiving folder and informs the Data Integrity team when completed
- The Data Integrity team conducts checks on data/metadata

Archiving

- Submission approved by the Data Integrity team and sent to the IT archiving team once the paper gets published
- Data encrypted and compressed
- Data is archived in tape storage off-site and cannot be altered – data can now only be retrieved by request

Figure 1. A process flow describing the steps needed to archive data associated with a published manuscript

Planning a new system for storing research data involved a number of detailed decisions. These were regularly discussed with stakeholders who were experts in the research process, research integrity, data storage, and data sharing. Some of the topics discussed included similar topics that would come up when discussing open repositories:

Required metadata

The first question to ask was – what information will be useful for anyone looking at the dataset in the future? We used the Dublin Core metadata standard for guidance and selected the most essential metadata fields (e.g., title, creator, subject, data type) to not overwhelm researchers with too many extra questions on the form while maintaining compliance with the archiving policy. We also added a few other metadata fields that we found necessary to capture for data that is going to be stored away for a long period of time, such as secondary person of contact, and whether the data has been deposited in any repository.

Minimum requirements

Over and about the requirements for metadata, it was important to consider what was considered 'good enough' and what would require further work. It is often said that "the enemy of the good is the perfect" and we did not want to require so much time and effort that the researchers refused to engage with the process. We therefore aimed to strike the right balance between not overburdening researchers during the archiving process with excessive requirements, while also ensuring that data that gets archived is reproducible and reusable if there was ever a need for the research integrity officers to review the data in the future, or if the Crick wished to reuse the data under the terms that it was originally collected.

When deciding the minimum requirements, we thought about file names, folder structure and documentation. Essentially the same elements that we require for open data shared on the institutional Figshare repository. As it is essential to have a standardised review process to maintain consistency in the quality of data that will be archived, we created a minimum archiving standards document (DOI: https://doi.org/10.25418/crick.26217920) to be used as a point of guidance for the reviewer when checking each data archive submission.

Reducing duplication of effort

It is important that when researchers have already shared their data openly and well, that they are not penalised by being asked to repeat the process for an internal system. This would likely have the opposite effect intended, reducing either use of the archive, or engagement with data sharing practices.

When researchers have added some of their data to a recognised repository, they can provide an identifier or a DOI of that dataset instead of having to duplicate their data in the archiving folder. To determine which repositories are trustworthy, we created a list that builds on that provided by the journal Scientific Data (Scientific Data, 2024).

Communications and training

For an organisation-wide initiative to work, people have to know about it, and understand both how and why they should contribute to it. A full communications programme was planned for the six months prior to the launch of the archive, including emails to research group leaders, drop-in sessions, items in newsletters and on display screens throughout the Institute, and a programme of training events to walk researchers through the process. We also created a step-by-step archiving guidance document which can be accessed by Crick researchers whenever needed, and provided both email and Slack routes for researchers to ask for further support.

Conclusion

The slow uptake of research data sharing practices and data documentation has led to research findings that are often not reproducible or reusable. This can decrease public trust in science, but also within the research community. Institutions, universities, funders, and journals all have a responsibility to find ways to nudge researchers to share data, e.g., by reviewing their data sharing policies, in order to contribute to open science.

At the Francis Crick Institute, we have identified several opportunities to increase sharing and make datasets more FAIR, such as reviewing datasets before deposit in the institutional repository, creating guidance materials, or making it easier to ask for support with depositing data. We believe these steps have led and will continue leading to increased and higher quality datasets.

Acknowledgements

We would like to acknowledge the work of Dr James Turner and Eleanor Adams who have been integral to pushing forward research integrity at the Francis Crick Institute. We would also like to thank John Roche, Narkkya Arulselvam and Andrew Jordan for their work in creating the research integrity archive, and Catherine Winchester from the CRUK Scotland Institute for offering advice on the archive.

This work was supported by the Francis Crick Institute which receives its core funding from Cancer Research UK (CC1034), the UK Medical Research Council (CC1034), and the Wellcome Trust (CC1034).

References

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature, 533*, 452-454. doi: 10.1038/533452a

Colavizza, G., Hrynaszkiewicz, I., Staden, I., Whitaker, K. J., & McGillivray, B. (2020). The citation advantage of linking publications to research data. *PLOS ONE*, *15*(4), e0230416. doi: 10.1371/journal.pone.0230416

Diaba-Nuhoho, P., & Amponsah-Offeh, M. (2021). Reproducibility and research integrity: the role of scientists and institutions. *BMC Research Notes, 14*, 451. doi: 10.1186/s13104-021-05875-3

Federer, L. M., Belter, C. W., Joubert, D. J., Livinski, A., Lu, Y-L., Snyders, L. N., & Thompson, H. (2018). Data sharing in PLOS ONE: An analysis of Data Availability Statements. *PLOS ONE. 13*(5): e0194768. doi: 10.1371/journal.pone.0194768

Gabelica, M., Bojčić, R., & Puljak, L. (2022). Many researchers were not compliant with their published data sharing statement: a mixed-methods study. *Journal of Clinical Epidemiology, 150*, 33–41. doi: 10.1016/j.jclinepi.2022.05.019

Hahnel, M., Smith, G., Scaplehorn, N., Schoenenberger, H., & Day, L. (2023). *The State of Open Data 2023.* Retrieved from https://doi.org/10.6084/m9.figshare.24428194.v1

House of Commons Science, Innovation and Technology Committee (2023). *Reproducibility and Research Integrity – Report Summary.* Retrieved from https://committees.parliament.uk/publications/39343/documents/194466/default/

Miyakawa, T. (2020). No raw data, no science: another possible source of the reproducibility crisis. *Molecular Brain, 13*, 24. doi: 10.1186/s13041-020-0552-2

Montague-Hellen, B., & Montague-Hellen, K. (2023). Publishers, funders and institutions: who is supporting UKRI-funded researchers to share data? *Insights*, *36*, 4. doi: 10.1629/uksg.602

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., du Sert, N. P., Simonsohn, U., Wagenmakers, E-J., Ware, J. J., & Ioannidis, J P. A. (2017). A manifesto for reproducible science. *Natire Human Behaviour, 1*, 0021. doi: 10.1038/s41562-016-0021

National Institute of Health (2023). *Selecting a Data Repository.* Retrieved from https://sharing.nih.gov/data-management-and-sharing-policy/sharing-scientific-data/selecting-a-data-repository. Accessed 21/05/2024

Nyhan, B., Porter, E., & Wood, T. J. (2022). Time and skeptical opinion content erode the effects of science coverage on climate beliefs and attitudes. *Proceedings of the National Academy of Sciences*, *119*(26), e2122069119. doi: 10.1073/pnas.2122069119

Rosman, T., Bosnjak, M., Silber, H., Koßmann, J., & Heycke, T. (2022). Open science and public trust in science: Results from two studies. *Public Understanding of Science, 31*(8), 1046–1062. doi: 10.1177/09636625221100686

Stuart, D., Baynes, G., Hrynaszkiewicz, I., Allin, K., Penny, D., Lucraft, M., & Astell, M. (2018). Practical challenges for researchers in data sharing. *Springer Nature*. Retrieved from https://doi.org/10.6084/m9.figshare.5975011.v1

Santos, C., Blake, J., & States, D. J. (2005). Supplementary data need to be kept in public repositories. *Nature, 438*, 738. doi: 10.1038/438738a

Scientific Data (2024). Data Repository Guidance. Retrieved from https://www.nature.com/sdata/policies/repositories. Accessed 29/05/2024

Springer Nature (2021). Written Evidence Submitted by Springer Nature (RRE0047). Retrieved from https://committees.parliament.uk/writtenevidence/39684/html/. Accessed 29/05/2024

Tedersoo, L., Küngas, R., Oras, E., Köster, K., Eenmaa, H., Leijen, Ä., Pedaste, M., Raju, M., Astapova, A., Lukner, H., Kogermann, K., & Sepp, T. (2021). Data sharing practices and data availability upon request differ across scientific disciplines, *Scientific Data* 8, 192. doi: 10.1038/s41597-021-00981-0

Vasilevsky, N. A., Minnier, J., Haendel, M. A., & Champieux, R. E. (2017). Reproducible and reusable research: are journal data sharing policies meeting the mark?. *PeerJ, 5*, e3208. doi: 10.7717/peerj.3208

Vines, T. H., Albert, A. Y., Andrew, R. L., Débarre, F., Bock, D. G., Franklin, M. T., Gilbert, K. J., Moore, J., Renaut, S., & Rennison, D. J. (2014). The Availability of Research Data Declines Rapidly with Article Age. *Current Biology*, *24*(1), 94–97. doi: 10.1016/j.cub.2013.11.014

Wingen, T., Berkessel, J. B., & Englich, B. (2019). No replication, No trust? How low replicability influences trust in Psychology. *Social Psychological & Personality Science, 11*(4), 454–463. doi: 10.1177/1948550619877412

Zuiderwijk, A., Shinde R., & Jeng, W. (2020). *What drives and inhibits researchers to share and use open research data? A systematic literature review to analyze factors influencing open research data adoption. PLOS ONE, 15*(9): e0239283. doi: 10.1371/journal.pone.0239283