Removing Barriers to Reproducible Research in Archaeology

Emma Karoune[1] and Esther Plomp[2]

[1]The Alan Turing Institute & Historic England, ORCID: 0000-0002-6576-6053, ekaroune@turing.ac.uk

[2]Delft University of Technology, Faculty of Applied Sciences, ORCID: 0000-0003-3625-1357, e.plomp@tudelft.nl

Authors in alphabetical order

Abstract

Reproducible research is being implemented at different speeds in different disciplines, and Archaeology is at the start of this journey. Reproducibility is the practice of reanalysing data by taking the same steps and producing the same or similar results. Enabling reproducibility is an important step to ensure research quality and validate interpretations. There are currently many barriers to moving towards reproducible research such as upskilling researchers in the practices, software and infrastructure needed to do reproducible research and concerns relating to opening up research such as how to share sensitive data.

In this article, we seek to introduce reproducible research in an understandable manner so that archaeologists can learn where and how to start improving the reproducibility of their research. We describe what reproducible archaeological research can look like and suggest three different computational skill levels of reproducible workflows with examples. Finally, in an extensive appendix, we address common questions about reproducible research to remove the stigma about these issues and suggest ways to overcome them.

Lay summary

Reproducible research (*Reproducible research is when data can be reanalysed taking the same steps and producing the same or similar result*) is being implemented at different speeds in different disciplines, and Archaeology, as a discipline that sits at the intersection of the sciences and humanities, is at the start of this journey. Enabling reproducibility of your work by others is an important step in ensuring research quality. There are currently many barriers to moving towards reproducible research such as upskilling researchers in the practices, software and infrastructure needed to do reproducible research and also the need to address how we can, as a discipline, deal with issues like sensitive data.

In this article, we seek to introduce reproducible research in an understandable manner so that archaeological researchers can learn where and how to start with this approach. We describe what reproducible archaeological research can look like and suggest three different computational skill levels of constructing reproducible research workflows (a research workflow is the different parts of a research lifecycle such as data collection, data analysis, data archiving, etc, and making all stages reproducible by using a history tracking system (version control) and transparent documation). Finally, in an extensive appendix, we address common questions about reproducible research to remove the stigma about these issues and suggest ways to overcome them.

Introduction

The move towards reproducible research has been accelerating in recent years in all research disciplines. Developments such as the UNESCO recommendation on open science are driving forward open science practices including reproducibility (UNESCO 2021). The adoption of open science practices has been happening even faster since the COVID-19 pandemic because researchers have had to work out how to conduct research in distributed teams and move research activities online. These online research activities have adopted the collaborative and computational methods common in open science communities, pushing this approach further into the mainstream of research.

Nevertheless, there is still a long way to go for all archaeological research to be reproducible and there are many barriers that archaeological researchers face when trying to implement reproducible research (Carney & Davies 2020, Marwick 2017, Marwick *et al*. 2017, Strupler 2021, Strupler & Wilkinson 2017). Often researchers do not know where to start as reproducible research is currently not common practice in archaeology, and is not actively taught in educational programmes. In this article, we are therefore seeking to remove some of the barriers to reproducible research by explaining what we mean by ***reproducible*** (with terms that are bold and italicised in the main text explicitly described in the glossary), describing why reproducible research is important for archaeological research, giving some examples of what ***reproducible workflows*** look like, and answering common concerns and questions about reproducible research (**Appendix A**).

We are also proposing that researchers take a small-steps approach to implementing a reproducible workflow: start by applying open science practices to one aspect of your research and then keep adding another skill or practice.

Conducting reproducible research involves learning knowledge and skills about many different open science practices and this can take time. By taking small steps in your learning of new skills, reproducible research and open science practices seem less daunting and archaeologists can gradually move towards fully reproducible workflows (**Figure 1**).

Figure 1: Taking incremental steps to improve your reproducible workflow will help you to increase your skills in transparently sharing your research. The Turing Way project illustration by Scriberia. Used under a CC-BY 4.0 licence. DOI: [10.5281/zenodo.3332807](https://doi.org/10.5281/zenodo.3332807).

What is reproducible research?

Reproducible research is when data from the original study can be reanalysed taking the same steps and producing the same or similar results (**Figure 2**). This can only be achieved with a transparent record of the research, also known as a reproducible workflow. Therefore, the data, methods, and analysis have to be made available to allow other researchers to review and reproduce the study. This increases the quality of research as it can be validated and reused more easily. Research that is not reproducible, or not shared in a transparent manner, is not representative of the full extent of the research that has been conducted. It is much like **Figure 3**: when we just look at publications and presentations, we are only able to see the tip of the ‘Research Iceberg’, and we may not understand the entire nature of the conducted research. Next to reproducible research the term computational reproducibility is used more specifically for obtaining the same results despite different hardware or compiler set ups (see for example Marwick *et al*. 2018 and Strupler & Wilkinson 2017).

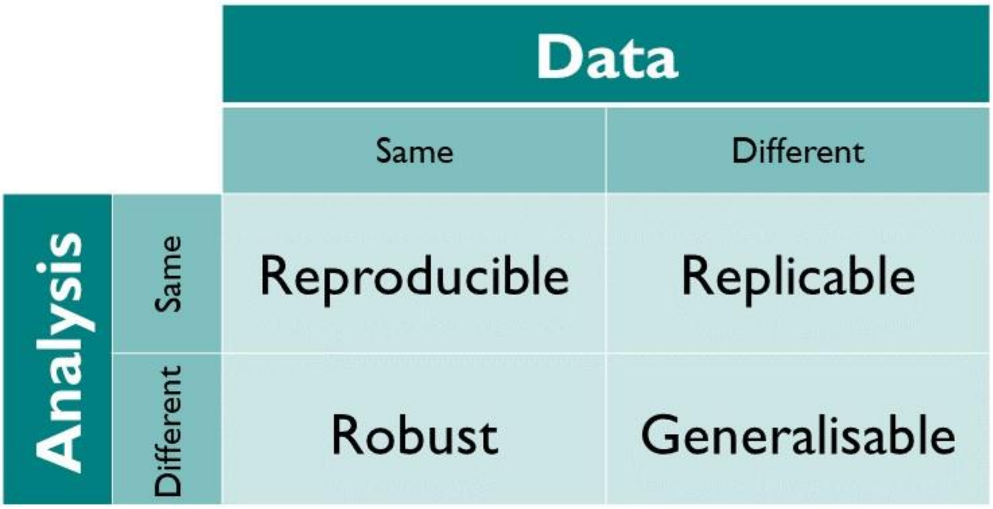

Figure 2: Reproducibility and replicability terminology explained. Image by The Turing Way Community (2021) under CC-BY 4.0.

Figure 3: The Research Iceberg, where only the article, ***preprint***, and presentations on research are visible. The components of the research on which these visible outputs are based remain invisible (research questions, methods, data, mistakes and corrections, discussions, community consultation, documentation and ideas).

Reproducibility is distinct from ***replication*** (or repeatability **Figure 2**), where a study is conducted independently using the same analysis from the first one to produce different data that produces the same or similar answers. In archaeology, direct replication of results (from the same samples) is very unlikely due to the limited availability of remains to investigate. As Strupler (2021) suggests, replication of archaeological investigations does take place by returning to earlier excavated sites and carrying out further investigations, re-analysing museum collections, or revisiting earlier publications. A result is ***robust*** when a dataset is analysed using different analysis approaches that provide similar answers. ***Replicable*** and robust findings then allow us to establish ***generalisable*** results, where the result is not dependent on a particular dataset or specific workflow (The Turing Way Community 2021).

To learn more about the differences between reproducibility and replication see Graham and Huffer (2020). See Marwick *et al.* 2020b for more details on how to organise replications as a part of undergraduate courses.

Why is reproducible archaeology important?

Archaeology is the scientific study of the materials of past human life and activities (Daniel 2021). Archaeological research is extremely varied involving many different sub-disciplines and crossing the humanities and the sciences. It produces many types of data, both quantitative and qualitative, and as a discipline we are just coming to grips with what this means in terms of open data sharing and other transparent practices that enable reproducible research (Marwick *et al*. 2017).

There are several reasons for moving towards reproducible research in archaeology: 1) the limited remains available for study (limited by the destructive nature of archaeological research, financial, location and ownership limitations); 2) equal access to knowledge generated by these remains; and 3) the sensitive remains that we study.

The majority of archaeological research involves the destruction of materials (Harris 2006) - whether this is during excavations or scientific investigations. The data and ***metadata*** [***paradata***] collected during excavations is often all that is left of the in-situ archaeological remains. We use the stratigraphic method to record information about archaeological sites. The artefacts and ecofacts removed from archaeological sites are changed during the process of our studies through sampling, cleaning, conservation and analysis. Hence, we need to implement ways of working to preserve the data and metadata of these processes in the most sustainable manner possible to allow future generations to reuse this information for reinterpretation of archaeological remains. Kansa & Kansa (2021) very rightly suggest that broadening data literacy skills in archaeology will result in realising the full potential of archaeological data such as data reuse across projects and large-scale data integrations. We must therefore concentrate on facilitating reuse of physical and digital artefacts, data and metadata, with as much care as we do with recording sites stratigraphically to preserve the archaeological record.

Compounding this destructive methodology is the finite remains that we study. Archaeological excavations are limited by the amount of funding for archaeological research and the limited locations that can be excavated. Many excavations happen as part of rescue or commercial work, which limits the time allowed for excavations and often the areas on archaeological sites that can be excavated. Therefore, we don’t get to excavate the whole surface of archaeological sites and the process of excavation requires destruction of the specific locations that we do excavate.

The artefacts and ecofacts that we sample are altered or destroyed through analysis and often only studied in a limited way - limited by restraints on money for the specialists' time and also limited by restricting the number of people who can study the material. Often only one or very few specialists examine each type of material from one site. Consequently, it is of paramount importance that our research is reproducible to enable (re)assessments of archaeological research.

We therefore also need reproducible research practices to ensure equitable access to archaeological research. Transparent recording makes research more accessible to anyone, allowing them to participate in the research process. Transparent recording also allows credit to be given fairly for the work that is done in the whole research project. To move to a more sustainable and inclusive future for archaeological research, we need to move away from the idea of sole ownership of research kept on our local computers that only benefits ourselves or few researchers. We must move to a more altruistic way of working for collective benefit by opening up our data (when possible) and processes for increased validation and reuse.

A third reason to move to reproducible ways of working is that some types of archaeological data and research focuses are sensitive. For example, studies involving human skeletal remains and also excavations conducted on sites belonging to Indigenous groups. We therefore need to consider carefully who owns the remains and the data we produce from these studies (Carroll *et al*. 2020). We need to consider questions such as who should have access to these resources for research and also how they are best preserved in the long term. It is also imperative to work out how the physical artefacts and the digital outputs can be stored to make them accessible to the appropriate audiences and for sustainable future use.
Sensitive data does not preclude reproducibility. In fact, it is more important to establish validation processes as there may be limitations with sharing data.

What does reproducible research look like?

Most published research articles are currently not reproducible - they are not transparent records of research. They are stand-alone papers that contain brief methods, limited data and are mostly filled with interpretations and discussions of results. Validation through peer-review and the reuse potential of these pieces of research is therefore rather limited.

The differences between stand-alone articles and articles that contain the details for full reproducibility can be found in Peng's (2011: figure 1) reproducibility spectrum (**Figure 3**). This diagram shows the addition of data, code and ***computational environment*** to the paper to move towards full reproducibility of the research. In fact, more detail is needed than stated in Peng's spectrum because the full methodological details (protocols) used for data collection would be required for replication of any experimental work included in the article (**Figure 4**). These methodological details could also be called ***metadata*** or ***paradata.*** Large meta-analysis studies need computational reproducibility to enable merging and reuse of datasets as well as studies that want to reuse the same methods for additional analysis of samples from the same or similar archaeological sites.

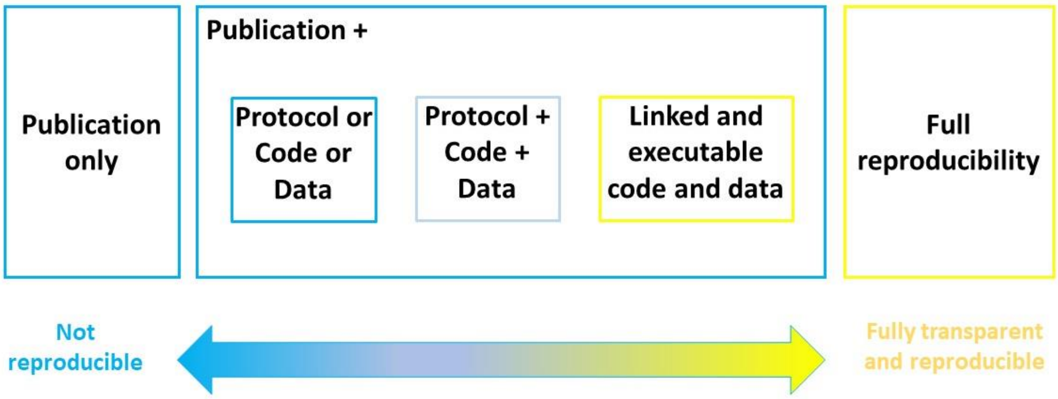

Figure 4: An adapted reproducible spectrum (Peng 2011) with the addition of protocols.

It is recommended practice, for greatest sustainability and findability, to deposit these files (data, code and methods) in an open ***repository*** (such as Zenodo, Open Science Framework, or Figshare - see **Appendix A** for more information on how to make your data accessible). If you are using GitHub, you can link your account with a repository for archiving. Using open repositories will give you a ***Digital Object Identifier*** (DOI) for your files as a whole, or for each research output, depending on where and how you choose to archive. It is then important to use your DOI(s) to write a ***data*** and code ***availability statement*** at the end of your article - this links your article with the rest of your research outputs. Thanks to the DOI assigned to your research outputs the transparency of the research record is improved and benefits such as increased visibility and citation are obtained (Piwowar *et al*. 2007; Piwowar & Vision 2013; Christensen *et a*l. 2019; Colavizza *et al.* 2020).

Computational tools can facilitate transparency of research by: 1) enabling ***version control***, and 2) using open source software for analysis. Version control is a systematic approach to record changes made in a file, or set of files, over time. It creates a history of the changes made to the file(s) that can be transparently reported. Version control can be achieved simply by using naming conventions, such as file-v0.1 and file-v0.2, to name your files. You can also use software such as Google Drive that automatically tracks the history of your files. There are more advanced version control systems such as GitHub or GitLab, which use the computer code Git. These computational tools create a much more detailed history of your research files that can even be used to assign credit for each individual researcher's work during the project. Please see The Turing Way for an example of how contributors can be recognised.

Open source software is software released under a license in which re-users have the rights to use, study, change, and distribute the software and its source code to anyone and for any purpose. Commonly used open source software languages are ***R*** and ***Python***. When analysis scripts are written in R and Python, all of the steps taken in the data analysis are transparent and traceable and can be shared with others. Other researchers are able to reuse the code for their own needs and it could potentially allow others to reproduce your analysis, if accompanied with other research outputs (method, data and computational environment). ***Proprietary software*** (software that requires a paid license to be able to use it) may in some cases be more user friendly, but using these tools prohibits the examination and reproduction of methods if not accompanied with written documentation of analysis steps, due to the inability to examine the analysis code. Furthermore, others may not have access to the paid software that you have used (see Nust & Pebesma 2020 for a more detailed discussion).

Although advanced version control systems and open-source software help you to create a transparent reproducible workflow, they often have a steep learning curve creating a barrier to some researchers. However, you don't have to have advanced computational skills to achieve reproducibility and there are many levels of reproducible workflows. We describe three different ways to create a reproducible workflow here, listed in order of least computational skill to most computational skill (**Table 1**). Following one of the skill levels of reproducible workflows proposed here will produce a transparent record of your research that you can publish linked to your research article. This creates a fully reproducible research article.

Table 1: Three levels of reproducible workflows based on computational skills (least skilled to most skilled).

Method	Needed	Computational skill required	Examples	Tools
1. Transparent recording	- Documentation of data collection and analysis steps - Raw data - Analysis output file	Yes, basic (non-coding)	Karoune (2021, 2022); Strupler & Wilkinson (2017)	Excel, Google docs and sheets, SPSS, Repository
2. Research Compendium	- Documentation (README) - Data	Yes, intermediate	Plomp (2021)	GitHub, GitLab, R, Repository

	- Code			
3. Executable Article	-Documentation (README) -Data -Code -Computational environment	Yes, advanced	Wang & Marwick (2020)	GitHub, GitLab, R, Binder, Repository

1. **Transparent recording of all sampling, laboratory methods, data and analysis through documentation.**

Transparent recording requires the least computational skills but produces a full transparent record of what you have done. It does not include any computational code, as the analysis steps you take could all be written down in a simple document and linked to an open dataset. This means you can use any type of analysis software such as Excel, Google sheets, SPSS, etc. Just remember to write down all the analysis steps that you took in a document in a way that another person could understand and reproduce what you have done.

Files to include with your article can be deposited in a ***repository*** and referred to in the text of the article, the data availability statement and the references, such as:

- Document file that has clearly written data collection methods (sampling and laboratory methods) and analysis steps.
- Raw data file - csv format is the best for reuse.
- Analysis output file - SPSS output file or analysis version of Excel file.

You can version control all of your work using a file naming system or choose a software that contains a simple history tracking system such as using Google Docs and sheets. This will help you to document your data collection and analysis steps fully.

Examples of transparent recording from articles:

- Karoune, E. (2022) Assessing open science practices in phytolith research - linked to a research compendium for Assessing open science practices in phytolith research. https://doi.org/10.17605/OSF.IO/9WA2F.
- Record of fieldwork project - Project Panormos including data and documentation (Strupler & Wilkinson 2017).

2. **A research compendium linked to your article**

A research compendium contains extensive documentation about the methodologies used, code files, details of the computational environment and raw data files. A set of folders can be set up from the beginning of your research project and continually added to throughout the project.

Files and folders to include in your research compendium, deposited in a repository and linked to using a DOI in your article (in the text and in the data/code availability statement):

- README file - contains clearly written data collection methods (sampling and laboratory methods), information about the computational environment.
- Data - raw data, cleaned data, analysis data.
- Code - scripts used to analyse your data.

You might also want to include an outputs folder for the final article tables and figures.

Examples of research compendium folder structures:

- Project Tier.
- Research compendium chapter from *The Turing Way*.

Example of research compendia in an archaeological article:

- Plomp, E. (2021a). Neodymium isotopes in modern human dental enamel: An exploratory dataset for human provenancing. Data in Brief, 38. DOI: https://doi.org/10.1016/j.dib.2021.107375.
    - Link to research compendium - https://doi.org/10.5281/zenodo.5150521 (code) & https://doi.org/10.48530/isoarch.2021.011 (data)

Plomp (2021a) provided a detailed description of the dataset in the article (Plomp 2021b), with links to the dataset on the disciplinary specific data repository, IsoArcH (Salesse *et al*. 2018), and scripts used in data analysis are publicly available on GitHub/Zenodo. The dataset on IsoArcH is available in .xlsx format and includes more detailed geographical information of the samples (latitude, longitude, altitude and distance from sea) as well as a .ris file containing the relevant research articles (Plomp 2021b). The figures in the data article were produced using R, and the scripts (with documentation and installation instructions) are shared on GitHub and archived on Zenodo (Plomp & Peterson 2021).

3. An executable research compendium

In an executable research compendium, the figures are enabled to be reproduced using your data making it easy for others, such as peer reviewers, to reproduce your results.

The files and folders to include with your article are the same as with a research compendium but you need to package them up to run the code:

- README file - contains data collection methods (sampling and laboratory methods), information about the computational environment.
- Data - raw data, cleaned data, analysis data.
- Code - scripts used to analyse your data.
- ***Container*** - using a tool such as ***Binder***.

Again, you would deposit these files in a repository and then add the DOI link within your article's data and code availability statement (and in the text of the methods section if needed).

Here are some links to using Binder:

- Gibson, Sarah. (2021, December 8). From Zero to Binder. AGU Fall Meeting, New Orleans, LA, USA and Online. Zenodo. https://doi.org/10.5281/zenodo.5767616.
- Using Binder with R Studio.

Example of executable research compendium:

- Wang, L.-Y., Marwick, B., (2020). Standardization of ceramic shape: A case study of Iron Age pottery from northeastern Taiwan. *Journal of Archaeological Science: Reports 33,* https://doi.org/10.1016/j.jasrep.2020.102554.
  - GitHub repository - with Binder badge

Confronting your barriers to starting implementing reproducible workflow

To be able to start with one of these steps in setting up your reproducible workflow you may still have questions or need more information. We have provided a glossary for keywords used in this article and **Appendix A** is a compilation of answers to frequently asked questions about working reproducibly.

Below follow two of the frequently asked questions about reproducible workflows to get you started:

1. How do I decide if I should publish my data and/or code openly?
2. Where do I start training myself in open science skills and reproducibility?

How do I decide if I should publish my data and/or code openly?

There may be several reasons that you cannot share your data or code publicly. The data you work with may belong to a community you are collaborating with, you may be dealing with personal data, sharing the data may have consequences on biodiversity, you might not be sure if you have any data to begin with, you may not have the rights to share the data or software, or you may be concerned about people ‘scooping’ your results.

I collaborate with a community

To ensure that you do not harm the community that the data belongs to, it is important to follow the CARE principles. The CARE principles facilitate Indigenous control in data governance and reuse, promoting equitable participation (Carroll et al. 2020). They address historical inequities and ensure that value from Indigenous data is created in a way that is grounded in Indigenous worldviews and by creating opportunities for Indigenous Peoples.

- ‘**C**ollective benefit’ for Indigenous Peoples must be facilitated when Indigenous data is used, to achieve inclusive and equitable innovation, as well as to improve governance and citizen engagement.
- ‘**A**uthority to control’ and govern data is the right of Indigenous People.
- ‘**R**esponsibility’ is achieved through nurturing respectful relationships with Indigenous peoples when working with their data.
- ‘**E**thics’ in data practices is representation and participation of Indigenous Peoples, who must be the ones to assess benefits, harms, and potential future uses based on community values and ethics.

The CARE principles require engagement with people and purpose to address the cultural, ethical, legal, and social dimensions associated with the intended uses of the dataset (Carroll *et al*. 2020; 2021, see also Marwick *et al*. 2020a). The CARE principles address issues of relevance for many populations (such as privacy, future use, reuse, stewardship) and can be used as a standard in crafting policies on data acquired about communities or populations (Carroll *et al.* 2020).

I work with personal data

The CARE principles are also aligned with privacy laws, which can also place requirements on the public sharing of personal data. This may be less relevant for archaeological remains, but can play a role in more recent cases or when your research is based on interviews such as ethnographic studies. These privacy laws differ per country and it is important to check which laws apply. If you are based at a larger institution there are generally experts available that can provide advice.

When following the CARE principles, or privacy laws, it may not always be possible to make the data publicly available, which could hamper reproducibility. The CARE principles and privacy laws should be prioritised in these cases but this does not mean you should not try to work reproducibly. There are alternative methods to fully open data that you could take: restricting data access by providing private repository links, providing access to synthetic data (synthetic data is a fake dataset produced to have the same qualities as your real dataset and therefore would produce similar results using your analysis - see Shannon and Walker 2018; for a case study in geographic research), or anonymising/generalising datasets by erasing personal/location data. Sharing part of your data or a dataset that is very similar to the original allows others to understand, evaluate and verify the used methods.

I work with sensitive location data

It might be harmful to share certain types of locational data and you should weigh the risks versus the benefits of sharing these types of data. Freely releasing GIS coordinates online as part of your dataset could potentially help looters and illegal excavators find sites (Strupler & Wilkinson 2017). This could lead to destruction of archaeological sites. Location data can always be omitted from a dataset if you think this is a potential problem.

The biodiversity community has similar potential problems with sharing the location data of endangered plant and animal species. However, the majority of this community feel there is more benefit using open data as its future reuse could lead to greater conservation opportunities, promote community engagement and reduce duplication of survey efforts (Tulloch *et al.* 2018).

I work with qualitative or theoretical data

If your research is more theoretically focused or based on other resources you may not have a dataset to share. Reproducibility may not directly translate to qualitative data given the unique importance of interpretation and subjective nature of qualitative data collection (Tsai *et al*. 2016). Instead, you can focus on providing information about the context of these resources and make your publications and/or books openly available.

I do not have the rights to share the data/code

When reusing the materials that others have created, or when you are using ***proprietary software*** or hardware, it is important to check if you have the right to share the resulting data and code. It may not always be possible to share your results if license restrictions are in place (see Appendix A: 'What about licenses/copyright?'). In these cases you should be as transparent as possible about the procedures or processes followed and about the limitations of making your outputs available. In the long term you can consider moving away from proprietary software, if possible, towards open source software such as R or Python so that you can make your code publicly available.

What if people will 'scoop' me?

You may wonder what will happen to your data once it is openly available and fear that someone will use the data for their next publication. This is something which has not yet been reported and there are several reasons for this. Generally, when you share your data through a repository, there is a timestamp associated with the work (similar to ***preprints*** or published articles). With ***version control*** on platforms such as GitHub it is even clearer who contributed what to the work as there are timestamps and records of all contributions. As you are the expert of the data and/or code, it will also be easier for others to collaborate with you instead of trying to reinvent the wheel themselves, so it's a good idea to make your contact details available to enable collaborators to contact you. Making your data available sets you up for these collaborations because your work is more easy to find and having access to the data/code facilitates collaboration.

Where do I start training myself in open science skills and reproducibility?

Upskilling yourself can be time consuming so take it a step at a time and remember it does not have to be costly. There are lots of free and open educational resources for you to use.

Start by looking in these places:

1. **Your own institution**: Investigate what courses your own institution offers. This could be through your own department, student services, a research software engineering group or library services.
2. **Open science online courses to work through at your own speed**: Most online courses are not specific to archaeology, but focus on general skills or knowledge for open science that can then be applied to your own research. There are dedicated open science online training platforms that have courses you can work through at your own speed such as FOSTER, Open Science MOOC and Open Scholarship Knowledge Base.
3. **Attend an online course or workshop:** This has the benefit of providing you with training materials but also an instructor you can ask for help. For computational skills courses, The Carpentries runs lots of different courses on data, library and software skills. There are also many open science focused workshops some of which are archaeology focused such as recent efforts by the Association for Environmental Archaeology that ran an open science focused conference and a workshop (Karoune *et al*. 2021) and a workshop on Reproducible Archaeology held at Durham University (Clarke *et al*. 2021).
4. **Apply to join a training programme:** For a more in-depth training experience, you could join an open science training and mentoring programme such as Open Life Science or Open Hardware Makers. These programmes are a mix of seminars, hands on training and mentorship to allow you to gain the skills and support to start or complete an open science focused project related to your own research. There have been a number of archaeological projects within the Open Life Science programme such as "FAIR Phytoliths" and "Intellectual Property, Indigenous Knowledges, and the Rise of Open Data in Australian Environmental Archaeology".

5. **Join a community or association** - There are a number of archaeological associations focused on this way of working such as Computer Applications and Quantitative Methods in Archaeology. There are also online communities such as Rchaeology (https://rchaeology.github.io/). The Software Sustainability Institute (SSI) is a large community of Research Software Engineers and researchers that use software. They run a fellowship programme for those using computational methods in their research and also offer lots of great resources for those wanting to learn computational skills. Examples of SSI blogs for beginner codes are: Resources for using spreadsheets in research and moving to other tools and Training resources for researchers who want to code.

More free educational resources:

- Teaching Reproducible Collaborative Data Analysis to Undergraduates Using Compendia https://osf.io/zpcn4/
- Introduction to R Programming for Historical Archaeologists (https://github.com/DAACS-Research-Consortium/DAACS-Open-AcademyThe)
- Tidyverse for Archaeologists - A Guide for Beginners
- There are also many free e-books on R - such as Big Book of R and R for Data Science.
- Quantitative Methods in Archaeology Using R - this one is not free!
- *Geocomputation with R*

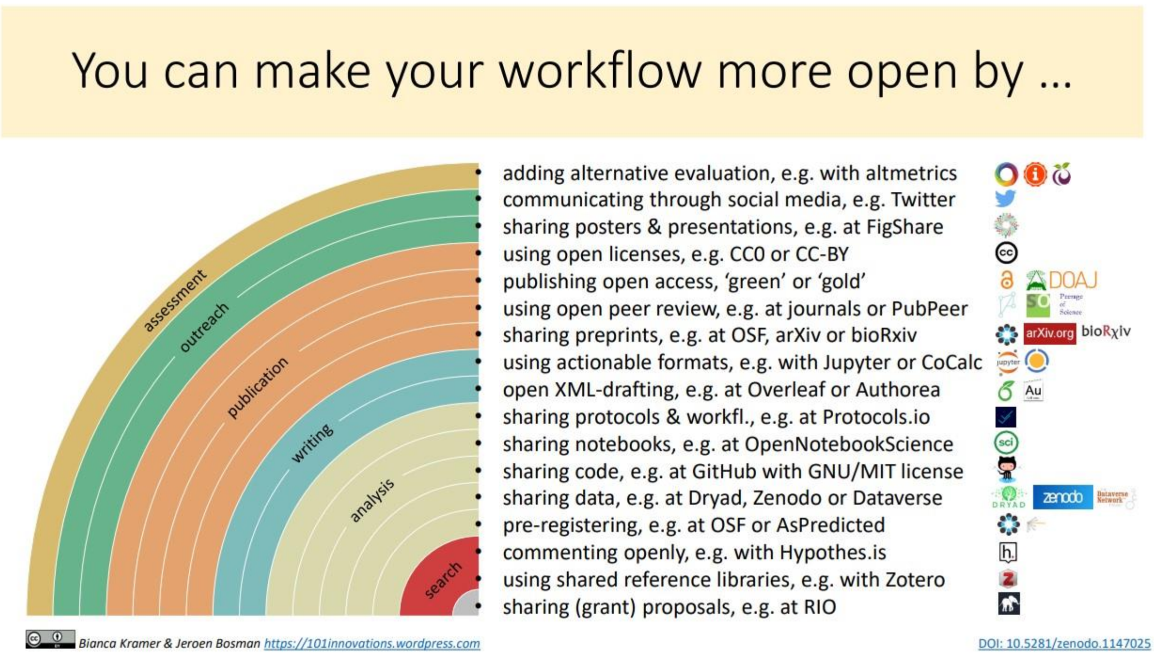

Figure 5: The rainbow of Open Science practices by Kramer & Bosman 2018.

You could take a look at the rainbow of open science practices to get some ideas (Kramer & Bosman 2018, **Figure 5**).

Conclusions

We hope we have motivated you to start your reproducibility journey and we have also managed to remove some barriers that previously prevented you from starting. Remember that you do not have to start with a fully computational reproducible workflow as done in the executable research compendium. The most important thing is to start making your materials available in a transparent manner, which can be achieved by transparent recording and documentation. Each time you have obtained more experience with making your research available in a more transparent way, you can then take a further step to improve the computational reproducibility of your work.

Data and Code availability statement

There is no new data or code used for this article.

Conflict of Interest

Emma is on the managing board of Peer Community in Archaeology. Esther is the Secretary General and Open Research Ambassador of IsoArcH.

Acknowledgements

Thanks to *The Turing Way* community. Many thanks to Yo Yehudi for her work on indexing references on data science and to Ben Marwick for his list of reproducible archaeological articles. Thank you to Bjorn Bartholdy and Celine Kerfant for comments on the drafts of this article.

Emma's FAIR Phytolith Project is funded through EOSC-Life from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824087.

References

Borer, E.T. *et al.* (2009) 'Some Simple Guidelines for Effective Data Management', *Bulletin of the Ecological Society of America*, 90(2), pp. 205–214. doi:10.1890/0012-9623-90.2.205.

Briney, K., Coates, H. and Goben, A. (2020) 'Foundational Practices of Research Data Management', *Research Ideas and Outcomes*, 6, p. e56508. doi:10.3897/rio.6.e56508.

Broman, K.W. and Woo, K.H. (2018) 'Data Organization in Spreadsheets', *The American Statistician*, 72(1), pp. 2–10. doi:10.1080/00031305.2017.1375989.

Brown, S. and et al. (2021) *Identifying ZooMS Spectra (mammals) using mMass v1*. preprint. doi:10.17504/protocols.io.bzscp6aw.

Buchanan, E.M. *et al.* (2021) 'Getting Started Creating Data Dictionaries: How to Create a Shareable Data Set', *Advances in Methods and Practices in Psychological Science*, 4(1), p. 251524592092800. doi:10.1177/2515245920928007.

Carney, M. and Davies, B. (2020) 'Agent-Based Modeling, Scientific Reproducibility, and Taphonomy: A Successful Model Implementation Case Study', *Journal of Computer Applications in Archaeology*, 3(1), pp. 182–196. doi:10.5334/jcaa.52.

Carroll, S.R. *et al.* (2020) 'The CARE Principles for Indigenous Data Governance', *Data Science Journal*, 19, p. 43. doi:10.5334/dsj-2020-043.

Carroll, S.R. *et al.* (2021) 'Operationalizing the CARE and FAIR Principles for Indigenous data futures', *Scientific Data*, 8(1), p. 108. doi:10.1038/s41597-021-00892-0.

Cerasoni, J.N. (2021a) 'Vectorial application for the illustration of archaeological lithic artefacts using the "Stone Tools Illustrations with Vector Art" (STIVA) Method', *PLOS ONE*. Edited by P.F. Biehl, 16(5), p. e0251466. doi:10.1371/journal.pone.0251466.

Cerasoni, J.N. (2021b) 'Stone Tools Illustrations with Vector Art: The 'STIVA' Method v2'. doi:10.17504/protocols.io.bubqnsmw.

Christensen, G. *et al.* (2019) 'A study of the impact of data sharing on article citations using journal policies as a natural experiment', *PLOS ONE*. Edited by F. Naudet, 14(12), p. e0225883. doi:10.1371/journal.pone.0225883.

Clarke, Alison *et al.* (2021) 'Reproducible Research in Archaeology'. doi:10.5281/ZENODO.5615561.

Colavizza, G. *et al.* (2020) 'The citation advantage of linking publications to research data', *PLOS ONE*. Edited by J.M. Wicherts, 15(4), p. e0230416. doi:10.1371/journal.pone.0230416.

Daniel, G.E. (2021) 'Archaeology', *Encyclopedia Britannia*. Available at: https://www.britannica.com/science/archaeology (Accessed: 1 January 2022).

Earnaud *et al.* (2022) *earnaud/MetaShARK-v2: Winter production 2022*. Zenodo. doi:10.5281/ZENODO.3648148.

Ebersole, C.R., Axt, J.R. and Nosek, B.A. (2016) ‘Scientists’ Reputations Are Based on Getting It Right, Not Being Right’, *PLOS Biology*, 14(5), p. e1002460. doi:10.1371/journal.pbio.1002460.

FAIRsharing Team (1987) ‘FAIRsharing record for: Thesaurus of Geographic Names’. FAIRsharing. doi:10.25504/FAIRSHARING.1413B5.

FAIRsharing Team (2018a) ‘FAIRsharing record for: Art and Archaeology Vocabulary’. FAIRsharing. doi:10.25504/FAIRSHARING.NIVBZ9.

FAIRsharing Team (2018b) ‘FAIRsharing record for: FISH Archaeological Sciences Thesaurus’. FAIRsharing. doi:10.25504/FAIRSHARING.4YRCBC.

FAIRsharing Team (2018c) ‘FAIRsharing record for: Monument Inventory DAta Standard Heritage’. FAIRsharing. doi:10.25504/FAIRSHARING.Q0HGQ.

FAIRsharing Team (2022) ‘FAIRsharing record for: CARARE Metadata Schema’. FAIRsharing. doi:10.25504/FAIRSHARING.BA7C93.

Falcucci, A. (2022) *MicroStone: Exploring the capabilities of the Artec Micro in scanning stone tools v1*. preprint. doi:10.17504/protocols.io.81wgb6781lpk/v1.

Fitzpatrick, K. (2020) ‘Not All Networks: Toward Open, Sustainable Research Communities’, in Eve, M.P. and Gray, J. (eds) *Reassembling Scholarly Communications: Histories, Infrastructures, and Global Politics of Open Access*. The MIT Press, p. 0. doi:10.7551/mitpress/11885.003.0035.

Fuchs, S. and Kuusniemi, M.E. (2018) ‘Making a research project understandable - Guide for data documentation’. Zenodo. Available at: https://doi.org/10.5281/zenodo.1914401.

Gibson, Sarah (2021) ‘From Zero to Binder’. doi:10.5281/ZENODO.5767616.

Göldner, D., Alexandros Karakostis, F. and Falcucci, A. (2022) *StyroStone: A protocol for scanning and extracting three-dimensional meshes of stone artefacts using Micro-CT scanners v2*. preprint. doi:10.17504/protocols.io.4r3l24d9qg1y/v2.

Graham, S. and Huffer, D. (2020) ‘Reproducibility, Replicability, and Revisiting the Insta-Dead and the Human Remains Trade’, *Internet Archaeology* [Preprint]. doi:10.11141/ia.55.11.

de Haas, T. and van Leusen, M. (2020) ‘FAIR survey: Improving documentation and archiving practices in archaeological field survey through CIDOC CRM.’, *Fasti Online Documents and Research*, 12. Available at: https://research.rug.nl/en/publications/fair-survey-improving-documentation-and-archiving-practices-in-ar.

Harris, E.C. (2006) ‘Archaeology and the Ethics of Scientific Destruction’, in Archer, S.N. and Bartoy, K.M. (eds) *Between Dirt and Discussion*. Boston, MA: Springer US, pp. 141–150. doi:10.1007/978-0-387-34219-1_7.

Hart, E.M. *et al.* (2016) ‘Ten Simple Rules for Digital Data Storage’, *PLOS Computational Biology*. Edited by S. Markel, 12(10), p. e1005097. doi:10.1371/journal.pcbi.1005097.

Hrynaszkiewicz, I. (2019) ‘Publishers’ Responsibilities in Promoting Data Quality and Reproducibility’, in Bespalov, A., Michel, M.C., and Steckler, T. (eds) *Good Research Practice in Non-Clinical Pharmacology and Biomedicine*. Cham: Springer International Publishing (Handbook of Experimental Pharmacology), pp. 319–348. doi:10.1007/164_2019_290.

Kansa, E. and Kansa, S.W. (2021) ‘Digital Data and Data Literacy in Archaeology Now and in the New Decade’, *Advances in Archaeological Practice*, 9(1), pp. 81–85. doi:10.1017/aap.2020.55.

Kansa, E., Kansa, S.W. and Goldstein, L. (2013) ‘On Ethics, Sustainability, and Open Access in Archaeology’, *The SAA Archaeological Record*, 13(4), pp. 15–22.

Karoune, E. (2021) ‘Research compendium for Assessing open science practices in phytolith research 2021’. Open Science Framework. Available at: https://doi.org/10.17605/OSF.IO/9WA2F.

Karoune, E. (2022) ‘Assessing Open Science Practices in Phytolith Research’, *Open Quaternary*, 8, p. 3. doi:10.5334/oq.88.

Karoune, Emma *et al.* (2021) ‘Open Science Skills Workshop for the Association for Environmental Archaeology’. doi:10.5281/ZENODO.5717154.

Kim, M. *et al.* (2018) ‘Data Scientists in Software Teams: State of the Art and Challenges’, *IEEE Transactions on Software Engineering*, 44(11), pp. 1024–1038. doi:10.1109/TSE.2017.2754374.

Krafczyk, M.S. *et al.* (2021) ‘Learning from reproducing computational results: introducing three principles and the *Reproduction Package*’, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 379(2197), p. rsta.2020.0069, 20200069. doi:10.1098/rsta.2020.0069.

Kramer, B. and Bosman, J. (2018) ‘Rainbow Of Open Science Practices’. doi:10.5281/ZENODO.1147025.

Krystalli, A. (2021) *Reproducible research data and project management in R*. Available at: https://annakrystalli.me/rrresearchACCE20/.

Kwon, D. (2022) ‘ResearchGate dealt a blow in copyright lawsuit’, *Nature*, 603(7901), pp. 375–376. doi:10.1038/d41586-022-00513-9.

Lamprecht, A.-L. *et al.* (2020) ‘Towards FAIR principles for research software’, *Data Science*. Edited by P. Groth, P. Groth, and M. Dumontier, 3(1), pp. 37–59. doi:10.3233/DS-190026.

Langham-Putrow, A., Bakker, C. and Riegelman, A. (2021) ‘Is the open access citation advantage real? A systematic review of the citation of open access and subscription-based articles’, *PLOS ONE*. Edited by S. Lozano, 16(6), p. e0253129. doi:10.1371/journal.pone.0253129.

Markowetz, F. (2015) ‘Five selfish reasons to work reproducibly’, *Genome Biology*, 16(1), p. 274. doi:10.1186/s13059-015-0850-7.

Marwick, B. (2017) ‘Computational Reproducibility in Archaeological Research: Basic Principles and a Case Study of Their Implementation’, *Journal of Archaeological Method and Theory*, 24(2), pp. 424–450. doi:10.1007/s10816-015-9272-9.

Marwick, B. *et al.* (2020) ‘How to Use Replication Assignments for Teaching Integrity in Empirical Archaeology’, *Advances in Archaeological Practice*, 8(1), pp. 78–86. doi:10.1017/aap.2019.38.

Marwick, B., Boettiger, C. and Mullen, L. (2018) ‘Packaging Data Analytical Work Reproducibly Using R (and Friends)’, *The American Statistician*, 72(1), pp. 80–88. doi:10.1080/00031305.2017.1375986.

Marwick, B. and et al. (2017) ‘Open Science in Archaeology’, *The SAA Archaeological Record*, 17(4), pp. 8–14.

Marwick, B., Pham, T.S. and Ko, M.S. (2020) ‘Over-research and ethics dumping in international archaeology | Nghiên cứu mang tính lối mòn và sự tha hóa về mặt đạo đức trong khảo cổ học quốc tế | နိုင်ငံတကာရှေးဟောင်းသုတေသနပညာရပ်မှ မဆီလျော်သော သုတေသနများနှင့် မသင့်လျော်သော ကျင့်ဝတ်များ’, *SPAFA Journal*, 4. doi:10.26721/spafajournal.v4i0.625.

Matzig, D.N. (2021) *outlineR: Artefact Processing and Extraction Protocol v1*. preprint. doi:10.17504/protocols.io.bygaptse.

Navarro, D. (2021) ‘Project structure - part 1’. Available at: https://www.youtube.com/watch?v=u6MiDFvAs9w&list=PLRPB0ZzEYegPiBteC2dRn95TX9YefYFyy&index=3.

Nüst, D. and Pebesma, E. (2021) ‘Practical Reproducibility in Geography and Geosciences’, *Annals of the American Association of Geographers*, 111(5), pp. 1300–1310. doi:10.1080/24694452.2020.1806028.

Orfanou, E. *et al.* (2020) ‘Minimally-invasive sampling of pars petrosa (os temporale) for ancient DNA extraction v2’. doi:10.17504/protocols.io.bqd8ms9w.

Peng, R.D. (2011) ‘Reproducible Research in Computational Science’, *Science*, 334(6060), pp. 1226–1227. doi:10.1126/science.1213847.

Perez-Riverol, Y. *et al.* (2016) ‘Ten Simple Rules for Taking Advantage of Git and GitHub’, *PLOS Computational Biology*. Edited by S. Markel, 12(7), p. e1004947. doi:10.1371/journal.pcbi.1004947.

Piwowar, H.A., Day, R.S. and Fridsma, D.B. (2007) ‘Sharing Detailed Research Data Is Associated with Increased Citation Rate’, *PLoS ONE*. Edited by J. Ioannidis, 2(3), p. e308. doi:10.1371/journal.pone.0000308.

Piwowar, H.A. and Vision, T.J. (2013) ‘Data reuse and the open data citation advantage’, *PeerJ*, 1, p. e175. doi:10.7717/peerj.175.

Plomp, E. (2021a) ‘Neodymium isotopes in modern human dental enamel: An exploratory dataset for human provenancing’, *Data in Brief*, 38, p. 107375. doi:10.1016/j.dib.2021.107375.

Plomp, E. (2021b) ‘Neodymium isotopes in modern human dental enamel: an exploratory dataset’. IsoArcH. doi:10.48530/ISOARCH.2021.011.

Plomp, E. and Peterson, J.C. (2021) *EstherPlomp/Figures-Nd-data*. Zenodo. doi:10.5281/ZENODO.5150521.

Plomp, E. Smeets, R., Koornneef, J. and Davies, G. (2019) ‘Chromatographic separation of neodymium isotopes in human dental enamel for Thermal Ionisation Mass Spectrometry (TIMS) analysis v1’. doi:10.17504/protocols.io.xzmfp46.

Plomp, E., Smeets, R. and Davies, G. (2020) ‘Chromatographic separation of strontium isotopes in human dental enamel for Thermal Ionisation Mass Spectrometry (TIMS) analysis v1’. doi:10.17504/protocols.io.37dgri6.

Ram, K. (2019) ‘A guide to making your data analysis more reproducible.’ *Rstudio Conference 2019*. Available at: https://github.com/karthik/rstudio2019 (Accessed: 1 May 2022).

Reimer, C.B. *et al.* (2019) ‘Open Up – the Mission Statement of the Control of Impulsive Action (Ctrl-ImpAct) Lab on Open Science’, *Psychologica Belgica*, 59(1), p. 321. doi:10.5334/pb.494.

Ross, S.A. and Ballsun-Stanton, B. (2021) *Introducing Preregistration of Research Design to Archaeology*. preprint. SocArXiv. doi:10.31235/osf.io/sbwcq.

Sabin, S. and A Fellows Yates, J. (2020) ‘Dental Calculus Field-Sampling Protocol (Sabin version) v2’. doi:10.17504/protocols.io.bqecmtaw.

Salesse, K. *et al.* (2018) ‘IsoArcH.eu: An open-access and collaborative isotope database for bioarchaeological samples from the Graeco-Roman world and its margins’, *Journal of Archaeological Science: Reports*, 19, pp. 1050–1055. doi:10.1016/j.jasrep.2017.07.030.

Sandve, G.K. *et al.* (2013) ‘Ten Simple Rules for Reproducible Computational Research’, *PLoS Computational Biology*. Edited by P.E. Bourne, 9(10), p. e1003285. doi:10.1371/journal.pcbi.1003285.

Schönbrodt, Felix (2022) ‘Academic job offers that mentioned open science’. doi:10.17605/OSF.IO/7JBNT.

Shannon, J. and Walker, K. (2018) ‘Opening GIScience: A process-based approach’, *International Journal of Geographical Information Science*, 32(10), pp. 1911–1926. doi:10.1080/13658816.2018.1464167.

Somer, J. (2018) ‘The Scientific Paper is Obsolete’, *The Atlantic*, 5 April. Available at: https://www.theatlantic.com/science/archive/2018/04/the-scientific-paper-is-obsolete/556676/.

Stark, P.B. (2018) ‘Before reproducibility must come preproducibility’, *Nature*, 557(7707), pp. 613–613. doi:10.1038/d41586-018-05256-0.

Strand, J.F. (2021) *Error Tight: Exercises for Lab Groups to Prevent Research Mistakes*. preprint. PsyArXiv. doi:10.31234/osf.io/rsn5y.

Strupler, N. (2021) ‘Re-discovering Archaeological Discoveries. Experiments with reproducing archaeological survey analysis’, *Internet Archaeology* [Preprint]. doi:10.11141/ia.56.6.

Strupler, N. and Wilkinson, T.C. (2017) ‘Reproducibility in the Field: Transparency, Version Control and Collaboration on the Project Panormos Survey’, *Open Archaeology*, 3(1). doi:10.1515/opar-2017-0019.

Tang, Y., Niccolo Cerasoni, J. and Yuko Hallett, E. (2022) *High Resolution "DIY" Photogrammetry - 'HRP' Protocol v2*. preprint. doi:10.17504/protocols.io.b53xq8pn.

Thaler, U. and Gneisinger, W. (2021) *Workflow for wooden contact samples in use-wear experiments with bronze axe replicas (MAP-protocol_B) v1*. preprint. doi:10.17504/protocols.io.bv2gn8bw.

The Turing Way Community (2021) *The Turing Way: A handbook for reproducible, ethical and collaborative research*. Zenodo. doi:10.5281/ZENODO.6533831.

Tsai, A.C. *et al.* (2016) ‘Promises and pitfalls of data sharing in qualitative research’, *Social Science & Medicine*, 169, pp. 191–198. doi:10.1016/j.socscimed.2016.08.004.

Tulloch, A.I.T. *et al.* (2018) ‘A decision tree for assessing the risks and benefits of publishing biodiversity data’, *Nature Ecology & Evolution*, 2(8), pp. 1209–1217. doi:10.1038/s41559-018-0608-1.

Turner, T.R. and Mulligan, C.J. (2019) ‘Data sharing in biological anthropology: Guiding principles and best practices’, *American Journal of Physical Anthropology*, 170(1), pp. 3–4. doi:10.1002/ajpa.23909.

UNESCO (2021) ‘UNESCO Recommendation on Open Science’. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000379949.locale=en.

Wang, L.-Y. and Marwick, B. (2020) ‘Standardization of ceramic shape: A case study of Iron Age pottery from northeastern Taiwan’, *Journal of Archaeological Science: Reports*, 33, p. 102554. doi:10.1016/j.jasrep.2020.102554.

Warinner, C., Velsko, I. and A Fellows Yates, J. (2020) ‘Dental Calculus Field-Sampling Protocol (Warinner Version) v1’. doi:10.17504/protocols.io.7hphj5n.

Wickham, H. (2014) ‘Tidy Data’, *Journal of Statistical Software*, 59(10). doi:10.18637/jss.v059.i10.

Wilkin, S. *et al.* (2021) ‘SP3 (Single-Pot, Solid-Phase, Sample-Preperation) Protein Extraction for Dental Calculus v1’. doi:10.17504/protocols.io.bfgrjjv6.

Wilkinson, M.D. *et al.* (2016) ‘The FAIR Guiding Principles for scientific data management and stewardship’, *Scientific Data*, 3(1), p. 160018. doi:10.1038/sdata.2016.18.

Ye, H. (2020) ‘Data Organization in Spreadsheets’. doi:10.5281/ZENODO.3892183.

Glossary of terms used in this paper (some definitions are adapted from *The Turing Way* Glossary):

- **Binder -** The Binder Project is a software project to package and share interactive, reproducible environments. A *Binder* or "Binder-ready repository" is a code repository that contains both code and content to run, and configuration files for the environment needed to run it.
- **Computational environment -** Features of a computer which can impact the behaviour of work done on it, such as its operating system, what software it has installed, and what versions of software packages are installed.
- **Container -** A container is a standard unit of software that packages up code and all its dependencies so the application runs quickly and reliably from one computing environment to another.
- **Data availability statement -** A data availability statement (also sometimes called a ‘data access statement’) tells the reader where the research data associated with a paper is available, and under what conditions the data can be accessed. They also include links using a DOI (where applicable) to the data set, code and other documentation.
- **Digital Object Identifier -** A digital object identifier (DOI) is a persistent identifier or handle used to identify objects uniquely, standardized by the International Organization for Standardization (ISO). An implementation of the Handle System, DOIs are in wide use mainly to identify academic, professional, and government information, such as journal articles, research reports, data sets, and official publications. However, they also have been used to identify other types of information resources, such as commercial videos.
- **Generalisable** - Combining replicable and robust findings allow us to form generalisable results. Note that running an analysis on a different software implementation and with a different dataset does not provide generalised results. There will be many more steps to know how well the work applies to all the different aspects of the research question. Generalisation is an important step towards understanding that the result is not dependent on a particular dataset nor a particular version of the analysis pipeline.
- **Gold open access -** the publisher makes all articles and related content available for free immediately on the journal's website. In such publications, articles are licensed for sharing and reuse via creative commons licenses or similar. An article processing charge (APC) is paid by the authors.
- **Green open access -** Independently from publication by a publisher, the author posts the work to a website controlled by the author, the research institution that funded or hosted the work, or to an independent central open repository, where people can download the work without paying. This can be a pre-print (version of article prior to peer preview) or post-print (version that has been peer reviewed). This is free for the author.

- **Metadata -** the data/information about the data. This can include information about who collected the data and when, and also the methods used for data collection.
- **Paradata -** Paradata of a data set or survey are data about the process by which the data were collected.
- **Persistent Identifier** - A long-lived method for identifying a resource that is unique, and widely understandable by a community. This includes ORCIDs as an identifier of researchers and digital object identifiers (DOI) as identifiers of research objects.
- **Postprint** - is the version of an article that incorporated changes from the peer review process, but does not yet have publication formatting or layout applied. It is usually uploaded by the authors to a public or institutional server where it is available openly.
- **Preprint** - is a version of an article that precedes formal peer review and publication in a peer-reviewed journal. Like postprints, authors generally upload this version of the article themselves using a public/institutional server where it is available openly.
- **Preregistration** - is the practice of registering the research design of the research project before it is conducted. This includes details of hypotheses, methods and proposed analysis steps. For more details see the Wikipedia page on preregistration.
- **Proprietary software** - is software that requires a paid license to be able to use it and it is closed-source (the code behind the software and the code that you produce in your analysis is not available to see).
- **Python** - is a high-level, interpreted, general-purpose programming language. Its design philosophy emphasises code readability with the use of significant indentation.
- **R -** is a programming language for statistical computing and graphics supported by the R Core Team and the R Foundation for Statistical Computing.
- **Registered report** - is a type of publication that is written before the research is conducted and includes the research question/s, methodology and proposed analysis steps. It is then peer reviewed prior to data collection.
- **Replicable/Replication** - A result is replicable when the same analysis performed on different datasets produces qualitatively similar answers.
- **Repository -** A long-lived place on the internet where resources (be they data, software, publications or anything else) can be stored and accessed. This keyword is often shortened to ‘repo’.
- **Reproducible** - A result is reproducible when the same analysis steps performed on the same dataset consistently produces the same answer.
- **Reproducible workflow -** a transparent record of the research that includes data, methods, and analysis to allow other researchers to review, reproduce and replicate the study.

- **Robust** - A result is robust when the same dataset is subjected to different analysis workflows to answer the same research question and a qualitatively similar or identical answer is produced. Robust results show that the work is not dependent on the specificities of the programming language chosen to perform the analysis.
- **Version Control** - is a systematic approach to record changes made in a file, or set of files, over time.

Appendix A: Frequently asked questions about reproducible research in archaeology

You probably have many questions about different aspects of reproducible research. Therefore, we want to discuss the most frequently asked questions that we hear from archaeologists about reproducibility to try to remove barriers and help you make progress along your reproducible research journey. The easiest way to use this Appendix is to go to the question that is currently on your mind:

How do I share data to make it more accessible to others?

To make your data accessible and reusable you should share your data according to the FAIR principles. The FAIR principles (Wilkinson *et al.* 2016; Lamprecht et al. 2020) facilitate the reproducibility of the research undertaken. The principles recommend that scientific data and software are:

- '**F**indable' thanks to their ***persistent identifier*** that is assigned to the dataset via a data repository or through a data article.
- '**A**ccessible' so that the data and metadata can be examined. Note that for data to be Accessible it does not necessarily need to be open: if only the metadata about the dataset is available, the data is still considered to follow the FAIR principles.
- '**I**nteroperable' so that data can be analysed and integrated with other data through the use of common vocabulary and formats.
- '**R**eusable' data is appropriately documented and licensed. A license defines what others may or may not do with your data. Open licenses, such as those of the Creative Commons or the Open Data Commons, allow others to reuse the data without limiting restrictions (see for more detail: What about licenses/copyright below).

Figure S1: The FAIR principles. The Turing Way project illustration by Scriberia. Used under a CC-BY 4.0 licence. DOI: 10.5281/zenodo.3332807.

When choosing where to disseminate your data or code you can choose between two routes: 1) choose one platform or 2) use multiple platforms based on their different functionalities and link the persistent identifiers in the documentation. For example, you could share your data and code on Zenodo and your research protocol on protocols.io. Both Zenodo and protocols.io allow you to add the persistent identifiers to other research outputs in the metadata, making it easy for others to find the related outputs. Note that it is not recommended to share the same outputs multiple times on different platforms, as it will be difficult for reusers to interpret which version they should use and cite.

How do I clean up the data and code before sharing this publicly?

Before you share your data or code you want to make sure that the dataset is complete and that variables are explained (**Figure S2**). Similarly, for code it will be needed to remove unnecessary parts and make sure functions and variables are adequately documented.

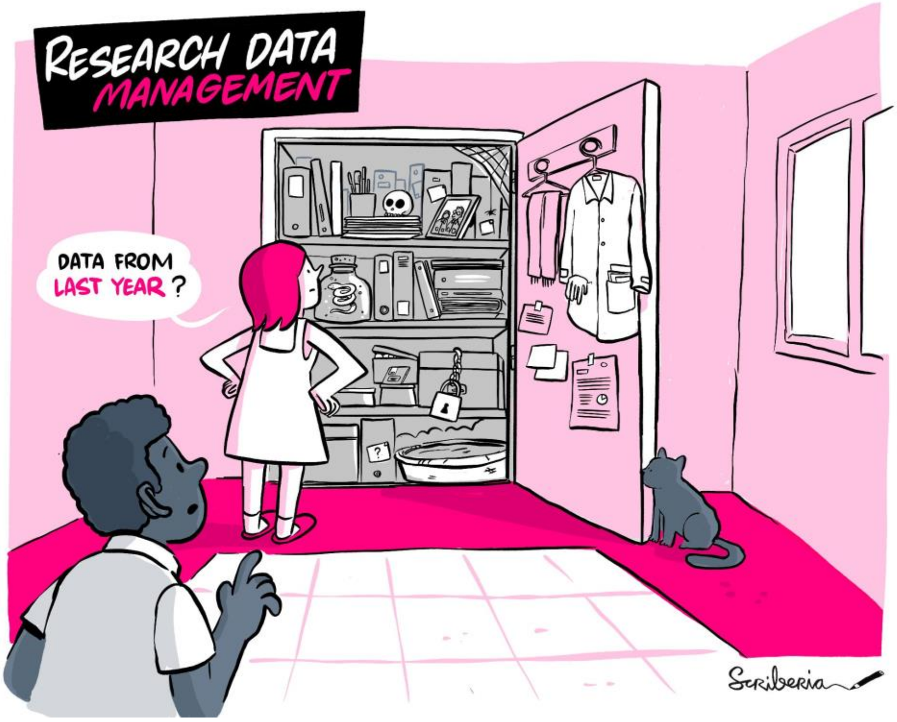

Figure S2: Cleaning up your data and code using research Data Management practices is recommended before sharing your data and code. The Turing Way project illustration by Scriberia. Used under a CC-BY 4.0 licence. DOI: 10.5281/zenodo.3332807.

For both data and code, it can help to have a colleague or collaborator review your work (see Reimer *et al.* 2019 for an example on how to set this up). They can provide you with feedback on the readability and completeness, and reproduce your results. Any feedback on where your collaborators get stuck or struggle will benefit the outputs that you will eventually share with a wider public.

There are several resources that delve deeper in how you can structure and document your data (Borer *et al*. 2009; Briney *et al.* 2020; Hart *et al*. 2016; Fuchs and Kuusniemi 2018) or code (Sandve *et al*. 2013; Ram 2019). Some of them go deeper into the specifics of a programming language, such as R (Wickham 2014; Krystalli 2021; Navarro 2021).

It can be helpful to have a folder structure set up and explained in a README file if your dataset/code is very complex. For folder structure examples, see templates set up by Nikola Vukovic, Chelsea Beck and Barbara Vreede. You can structure folders based on the person that has generated the data/folder, chronologically (month, year, sessions), per project, or based on analysis method/equipment/type of data.

In data management it is important to stay consistent, avoid leaving empty values (use NA instead) as it is not always clear what an empty cell actually means (no value, a value of zero, not measured?). If you use consistent file naming it is easier for you to find your files (see Jenny Bryan’s work and Caltech’s guide). For example, you can include the date in the format YYYMMDD in your file name so that your files order chronologically. This also makes it easier to see if you have any duplicate files. In spreadsheets, put as little information as possible in a single cell and only one observation per row (Broman and Woo 2018). You can share additional information in a README file or in a data dictionary (Buchanan *et al.* 2021) or code book that describes the spreadsheet and any cleaning steps you took. In your data, avoid formatting to describe the data (colours, font, bolding). Instead, add additional cells for the information that this formatting should be conveying. You can also use data validation to avoid errors. Excel and OpenRefine have several options that you can use. For more spreadsheet tips see the Carpentries curriculum on spreadsheets for ecologists and social scientists, Hao Ye's work (Ye 2020) and information on The Turing Way.

To manage your code it can be helpful to use Git/GitHub to keep better track of any modifications made (and by whom) (Perez-Riverol *et al*. 2016). If you share your research software from the start you will also structure it differently and more readable to others than you would if you would if you kept it closed. For software it should be clear what language and environment you are using, and if there are any dependencies and/or packages needed to process the data in a similar fashion as the analysis conducted for the study. See 'Make sure that your code is in a sharable state' and Krafczyk *et al.* (2021, p5-11) for more details about how to ensure your code is ready to be shared.

Add a README file to your dataset or your software repository. README text files should describe the methods used for data collection and analysis and include data/software-specific information (parameters, variables, column headings, symbols used, etc.). See Make a README for more information on why README files are important and how you can set up your own. You can use README files from existing projects and datasets as examples or inspiration (for example for data and a general and archaeological example for code).

How do I share my research methods more openly?

Research methods are the processes that generate research data. Using different methods, or adapting certain steps of a procedure, can affect the resulting research data. To increase the reproducibility of your work it is therefore crucial to make methods more openly available. Methods can include wet lab protocols, software analyses, strategies for surveys (see Strupler & Wilkinson 2017) and may involve various types of equipment. Methods shared on platforms such as protocols.io can facilitate reuse of the data or the method you used, as these platforms allow anyone to set up a copy of the method (forking).

Examples are

- Article by Cerasoni (2021a) and accompanying protocol (Cerasoni 2021b) on **stone tool** illustrations and Matzig (2021) on an R-package for artefact processing.
- Protocol by Thaler and Gneisinger (2021) on **use-wear experiments**
- Protocols by Brown *et al.* (2021) on **ZooMS Spectra**
- Protocols by Plomp *et al.* (2019 and 2020) on **isotope** analysis (neodymium and strontium respectively).
- Protocols on **dental calculus** sampling by Warinner *et al.* (2020), Sabin and Fellows Yates (2020), Wilkin *et al.* (2021).
- Protocols on **3D models** by Tang *et al*. (2022), Falcucci (2022) and Göldner *et al*. (2022).
- Protocol on **DNA** sampling by Orfanou et al. (2020).

What is metadata?

Metadata is information about the data. These could range from your notes about data collection and processing to the information that you are required to fill in when you deposit data in a data repository. The last type of metadata is machine readable and will facilitate data discovery (see FAIR). Most data repositories, such as Zenodo and Figshare, will use standardised schemes of these information fields (such as Dublin Core). Standardised metadata, or a metadata standard, will enhance the interoperability of information as similar descriptions are used which should make it easier to integrate data. The integration of studies would allow archaeologists to address research questions on a larger scale. You can start small by searching for metadata standards using FAIRsharing.org or start discussions in your subfield about how to standardise data documentation.

To our knowledge, archaeology has these specific metadata standards:

- CIDOC CRM for **field surveys** (de Haas and van Leusen 2020)
- Monument Inventory DAta Standard Heritage (MIDAS Heritage), for recording **heritage** information on buildings, archaeological sites, shipwrecks, parks and gardens, battlefields, areas of interest and artefacts (FAIRsharing Team 2018c).
- Art and Archaeology Vocabulary employed for indexing bibliographical records for the "**Art and Archaeology**" FRANCIS database (FAIRsharing Team 2018a).
- FISH Archaeological Sciences Thesaurus (FISH-AST) for recording **techniques, recovery methods and materials** (FAIRsharing Team 2018b).
- CARARE Metadata Schema for an **organisation's online collections**, heritage assets and their digital resources (FAIRsharing Team 2022).
- Thesaurus of Geographic Names (TGN) terminology that focuses on recording names, relationships, place types, dates, notes, and coordinates for

current and historical cities, nations, empires, **archaeological sites**, lost settlements, and physical features (FAIRsharing Team 1987).
- MetaShARK for **ecological data** (Earnaud *et al*. 2021).

Other metadata standards that could be useful are:
- The RFC-3339 or ISO 8601 standards, which specify the order in which **dates** are written: YYYY-MM-DD.
- ISO 19115 for **geographic** information.

To learn more about Metadata, visit the Archaeology Data Services website.

What about licenses and copyright?

Licenses govern what someone else can do with data and software that you share. The various licenses have different criteria about what is allowed when the data/software is reused, and there are different types of licenses available for data and software.

- For **data** the Creative Commons Licenses or Open Data Licenses are most often used. For example, the CC-BY license for data requires that the reuser provides attribution for data re-use through, for example, citation.
- For **software** the Choose a License website provides an overview of the available licenses. An often used license for software is the MIT license, that similarly to the CC-BY licence, requires attribution for reuse.

For both data and software it is important to follow the license requirements. Sometimes these requirements are in conflict, or incompatible. Incompatible licenses can get especially complex when you want to reuse software created by others. This makes combining datasets or software difficult, which is something to keep in mind when you choose a more restrictive license for your outputs. The fewer restrictions a license has, the easier it is for others to reuse your work (for data CC0 or CC-BY, for software MIT). If you are unsure whether you are complying with license requirements, check if your institution provides any advice on this. Generally this type of support is available from the Library or a copyright support desk.

Isn’t reproducible archaeology more expensive?

It is a misconception that working with an open science approach is more expensive. This idea of higher cost stems from the well known high costs of ***gold open access*** journal articles and also dedicated archaeological data repositories being commercial businesses that charge for data deposition. See this blog for more information about this misconception - Getting started with open repositories - part 1 - what you might think.

In fact, everything that you would want to do openly with your research can be done for free using free open-source software, free tools and apps such as GitHub and Google Drive, and free open repositories such as Zenodo, Open Science Framework, Figshare and Dataverse.

- **Depositing data and other research outputs:** There is a wide choice of free and open repositories for depositing data and other research outputs. This might be through your own institution or one of the large public infrastructure repositories such as Zenodo, Open Science Framework, Dataverse or Figshare.
- **Software for open analysis:** To use the R coding language for analysis, you can use Rstudio. It is free to download and there are many packages that allow you to do the types of statistical analysis, which you would have done in expensive proprietary software such as SPSS.
- **Publishing open access:** You can make your articles open access for free using the green or diamond open access route. ***Green open access*** is where you deposit a version of your article (not the final formatted version that will be in the journal but a preprint or postprint version) on an open repository such as a ***preprint*** server (some examples are arXiv, bioRxiv, or EarthArXiv) or one of the open repositories mentioned above. This can be done at no cost to you or the reader. The majority of journals allow you to do this, but do be careful to read the journal's guidelines on doing this (see details of these policies on Sherpa Romeo).
    - You can also use diamond open access, which is free for authors to publish and free for readers type of open access that some journals offer such as those paid for by societies, associations or communities such as Peer Community in Archaeology.
- **Version control for open reproducible workflows:** For simple version control, you can use Google Drive. There are free advanced version control tools that you can use based on Git - GitHub or GitLab. An alternative to Git is Subversion - also a free and open-source software.

What if people misinterpret my data or find a mistake?

To avoid misinterpretation of the data you should provide sufficient information about your dataset and all the data required for appropriate reuse (**Figure S3**). You can also list your contact details in the documentation or readme file so that reusers can contact you with questions or concerns. You could, for example, set up an ORCID, a ***persistent identifier*** for researchers that you own and control, with your contact details to ensure that reusers are able to find you. ORCIDs are particularly beneficial if you have a common name or if you expect to switch between institutions in the future.

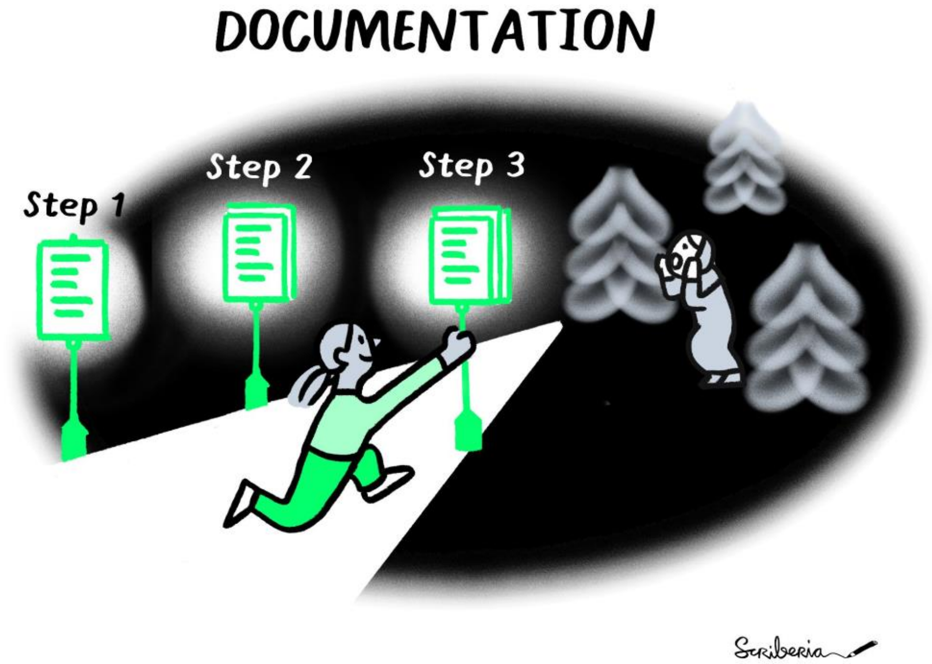

Figure S3: Clearly written and available documentation will allow others to follow the steps you took in the research process, preventing misinterpretation. The Turing Way project illustration by Scriberia. Used under a CC-BY 4.0 licence. DOI: 10.5281/zenodo.3332807.

As you are only human it is entirely possible that there is a mistake in your data or script. Keep in mind that if anyone would find an error in your data that this means that your dataset is engaging and relevant (Strupler 2021). To prevent errors you can use the guidance from Error Tight to set up a workflow in the lab that makes it more likely that mistakes made in the lab are caught early (Strand 2021). You can also minimise mistakes in your own research outputs by asking someone from your lab to check your data or code before making it more widely available, for example, by trying to reproduce your work (a co-pilot, see Reimer *et al*. 2019).

Even after close scrutiny by a colleague it could be that someone discovers a mistake after you shared the data or code publicly. Most data repositories allow you to upload a new version of the data/code where you can explain in the documentation what has changed in this new version and why. Correcting this mistake may save the re-users of your data and code, and yourself, a lot of time and may increase the trustworthiness of your data and code as you facilitate the self-correcting nature of the scientific process. Research shows that improving the original work can have a beneficial effect on your reputation (Ebersole *et al.* 2016).

Is archaeology suitable for preregistration?

A ***preregistration*** is a document in which the research design, and sometimes hypotheses, is specified before research is carried out. This could also be done through a ***registered report***. Preregistering your research may structure your data collection, management and analysis which can result in more robust research, reusable datasets and reduce the time spent managing problems and data cleaning on a more ad hoc manner (Ross and Ballsun-Stanton 2021). Ross and Ballsun-Stanton (2021) argue that preregistration is beneficial for archaeologists. Preregistration encourages a more thoughtful approach to research design, better management of biases through making approaches and assumptions more explicit, and it encourages good practices in research transparency (Ross and Ballsun-Stanton 2021). Good practice around archaeological preregistration is still emerging, but Ross and Ballsun-Stanton (2021) offer some helpful pointers.

My supervisor won't let me work reproducibility, how do I convince them?

There are several strong arguments to make for moving to a reproducible research workflow (**Figure S4**). Many funders are now requiring more open practices. The UK Research and Innovation and the European Research Council both have policies requiring immediate open access publishing through ***Gold or Green Open access*** for all grant holders. These publications must be linked to all research outputs to validate research. This means that your supervisor will have to start opening up their work to some extent and it would be good to learn how to do this well now.

Similarly to changing funding requirements, the importance of the published research articles is likely to change in the upcoming years. Several individuals have already called the stand-alone scientific paper outdated (Marwick *et al.* 2017), obsolete (Somers 2018), or dead (Robert Terry during the second UNESCO Conference on Open Science - Link to video). While the scientific paper has not yet died, the journals have requirements that your work should fulfil before it will be published. Increasingly, this includes making the underlying data and code available (Hrynaszkiewicz 2019), see for example the American Journal of Physical Anthropology requirements (Turner & Mulligan 2019). Even if journals do not have these requirements, it may be that your reviewers ask to see the underlying code and data (Stark 2018). Sharing the data/code during the peer review process may thus result in improvements of your work or faster acceptance as the reviewer does not have to wait for access (Markowetz 2015).

Having a reproducible workflow, which is transparent and open, has greater research impact. This has now been proven in a number of ways. Open access publications are known to have a citation advantage over publications behind paywalls (Langham-Putrow *et al*. 2021). It has also been found that linking open data to your article increases citations significantly (Piwowar *et al*. 2007; Piwowar & Vision 2013; Christensen *et al*. 2019; Colavizza *et al*. 2020).

You could also consider publishing more articles by writing a data paper or software paper for your project. This would give you credit for the extra work that you are doing to produce a reproducible workflow and also increase the overall outputs of the project therefore increasing the impact.

Moving to reproducible workflows is going to take time and it will help to talk about the benefits within your research group to encourage others to follow your example. Find allies within your department or other people within your subfield that do work reproducible to convince the supervisor that this is a good thing.

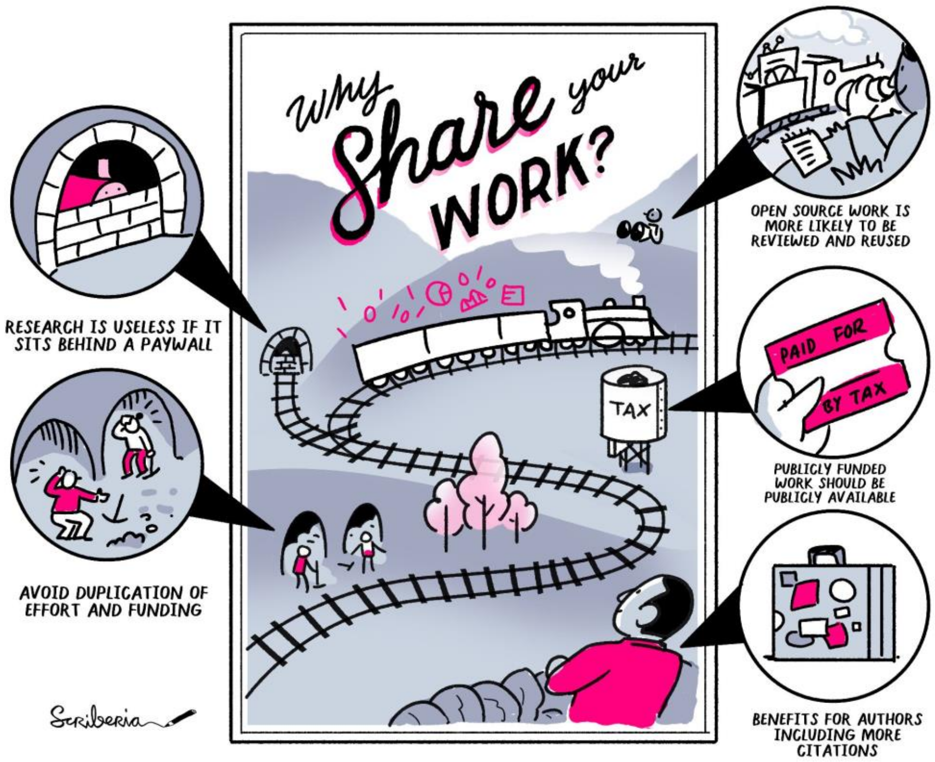

Figure S4: Benefits of sharing your work openly. Research is useless if it is not accessible and sitting behind a paywall. Through sharing your work you can avoid duplication of effort and waste of funding. Publicly funded work should be publicly available as it is paid for by taxpayers. Open source work is more likely to be reviewed and reused and can generate more citations. The Turing Way project illustration by Scriberia. Used under a CC-BY 4.0 licence. DOI: 10.5281/zenodo.3332807.

Will reproducible research be taken into account when looking for a next job?

Academic institutes are changing the focus of research evaluations, moving away from the impact factor of articles to a more broader evaluation that also takes into account education, open science practices and leadership. Examples are the [TRIPLE model at Utrecht University] in the Netherlands.

Making your work openly available will help build your reputation for being an honest and careful researcher (Markowetz 2015). Experience with Open Science practises is also increasingly asked for in vacancies (Schönbrodt *et al.* 2021).

Increasingly funding bodies are asking about data and software management and the sharing of these research outputs. Moving towards sharing these outputs will therefore outweigh the costs in the long term by increasing your chances for funding and by improving your sharing workflows earlier rather than later.

Next to improving your chances on the academic job market, open data and code can also be useful in positions elsewhere, such as in industry where the demand for computational skills is high (Anaconda The State of Data Science 2020; Kim *et al*. 2018).

Do platforms like SciHub, ResearchGate, Academia.edu count as Open Access?

Platforms such as SciHub, ResearchGate, and Academia.edu do not count as sustainable Open Access. SciHub, while providing access to research more widely, is not a legal platform and is hosted by a single individual. This makes long term sustainability questionable, and the founder, Alexandra Elbakyan, is dealing with multiple lawsuits.

Academia.edu is not an educationally-affiliated organisation and instead monetising scholarly outputs. By agreeing to their privacy policy Academia.edu is furthermore able to sell your information to other companies (Tóth Czifra 2020). ResearchGate has been subjected to lawsuits that determined that the platform is responsible for copyright infringement, which can result in the removal of the papers that they made openly available (Kwon 2022). ResearchGate and Academia.edu are also not open about their business and sustainability models, or interoperable with other services (Fitzpatrick 2020).

While Academia.edu and ResearchGate are good for advertising your research and networking like other social media platforms, you might be illegally sharing copyrighted work through these platforms. If your article has a CC-BY-NC-ND license, you are not allowed to share it on Academia.edu and ResearchGate as these are commercial platforms which are excluded by the NC part of the license (Non-Commercial). This can be circumvented by choosing a CC-BY license so that you are allowed to share it on these platforms, as you retain the rights to your work and there are no commercial reuse restrictions.

You can also share your work via a ***preprint*** or ***postprint*** version under an open license through more sustainable solutions such as data repositories and preprint servers. Institutions can also play a role here by retaining control of the infrastructures that provide access to research outputs.

An example of scholarly communities retaining control of all the infrastructure involved in making research available is the Peer Community in Archaeology platform and IsoArcH database (Salesse *et al*. 2018). The Peer Community in Archaeology are openly reviewing and recommending preprints therefore increasing the transparency of quality control processes. Disciplinary specific repositories such as IsoArcH (for bioarchaeological isotope data) increase the impact of datasets, as they are curated by specialists and accompanied by the relevant metadata, which makes the data more reusable.

If you would like to learn more about Open Access in archaeology, read the article by Kansa *et al*. (2013).