Repository Guidelines

Rorie Edmunds,, Alison Specht,, Shelley Stall,, Romain David, Laurence Mabile,, Margaret O'Brien,, Yasuhiro Murayama,, Pedro Correa,, Paulo Machicao, Jeaneth University of Sao, Nobuko Miyairi,

1 Introduction – PARSEC Survey Results

When PARSEC Data Synthesis team members were asked whether they had 'published' their data, 13 of the 19 respondents said they had. The respondents answering 'yes', were then asked the open, follow-up question of 'how <they> chose the repository for <their> data'. This elicited the most responses of all the questions with nearly everyone adding a comment.

Four people said that they published through a laboratory-hosted web server (or similar), with another two only publishing as a data supplement/paper. Three people typically use generalist repositories, such as Dryad. Two people seem to use the portals for data repositories to upload their data, and one had built their own repository. Finally, one person said they simply go with whatever is easiest and cheapest!

It is clear from the above that there were different understandings among team members as to exactly what 'published' means in the sense stated in the survey, and whether there was a full recognition of why the follow-up concerning repositories led naturally from the initial question on the publishing of data. It is emphasized in the following comment added by a person who had not published their data, 'I actually have data I would have liked to publish but never spent the time to understand how It should be done…'

The survey questions were deliberately put in such a way to highlight the importance that digital research outputs[1] (simply termed 'data' from hereon) should not just be made available *somewhere* for example, in a data supplement to a journal paper. Rather, it is becoming vital to scientific funders—and importantly in the case of PARSEC, is mandated by the funders that make up the Belmont Forum—for data to be stored (and ideally managed) for the long-term in a reliable data repository. What is meant by a 'reliable data repository' will be discussed below.

2 What is a Repository and Why is its Selection Important to You?

From the CASRAI Research Data Management Glossary[2], repositories are organizations that 'preserve, manage, and provide access to many types of digital materials in a variety of formats. Materials in online repositories are curated to enable search, discovery, and reuse. There must be sufficient control for the digital material to be authentic, reliable, accessible and usable on a continuing basis'. Put simply, repositories take responsibility for the stewardship (management) of data and ensure that they are held in an appropriate environment for appropriate periods of time. From this definition, it should be apparent that temporary storage solutions or shared environments such as a Google Drive or Dropbox are NOT repositories, and neither is your laboratory webserver!

That your research data are valuable should hopefully be beyond dispute. But, they are not only valuable to you, they are valuable to others too. The proper management and storage of your data are vital to ensure their safeguarding such that they remain accessible, unchanged over time, and of as high quality as possible. Their use and reuse are not only important from a scientific perspective—in terms of the accountability, transparency and verifiability of science—but also from an economic one. Funders increasingly recognize this, and are mandating the data they fund (your data) be openly shared and be preserved for the long-term such that their investment in the data's production is not wasted and they achieve the highest return on that investment. The role of the repository in the data lifecycle should not be underestimated—a good repository will not only increase the adherence of your data to the FAIR (Findable, Accessible, Interoperable, Reusable) Principles, it will ensure that your data remain FAIR over time—and therefore selection of the 'right' repository is essential.

[1] Although we traditionally talk about 'data' as *the* research output, there is a growing understanding that all digital objects (code, models, workflows,...)—and even physical samples—enabling the verifiability and reproducibility of research should be preserved in a repository.

[2] The CASRAI website will close on 23 June 2022. The Committee on Data of the International Science Council (CODATA) is taking responsibility for the CASRAI Research Data Management Glossary, which will now be known as the Research Data Management Terminology.

3 Types of Repository

Most repositories fall into one of two main categories: domain or generalist. Most of what follows on repository selection focusses more heavily on domain repositories, since they are more specialist, and thus more likely to fulfil both the common functions you would want from a repository, as well as any specific needs you may have within your research field(s).

3.1 Domain Repositories

A domain repository—sometimes known as a 'subject-based' repository—will specialize in a specific research field or data type. It usually has a well-defined group of users at which its data and services are aimed, its 'Designated Community'. In many cases, domain repositories have a national or regional remit, or at least are publicly funded, and thus you will be able to deposit your data (and access others data) free of charge. They may also be part of a wider network of similar national repositories or be subject to international agreements regarding data sharing and management, which can ensure a wider pool of expertise and guarantees that multiple mirrored copies of your data exist.

3.2 Generalist Repositories

A generalist repository is a generic, multi-subject repository. Typical examples include institutional repositories serving research performing organizations such as a university library, open access repositories such as Zenodo or Dryad, and technical service providers such as Figshare. The user community of a generalist repository will be very broad and may even be the general public at large. Because of this, and since you may be a (paying) 'client' generalist repositories will often rely on data depositors to manage their own data. Many do not offer services beyond simple archiving—static, long-term preservation—although an institutional repository (or a paid service contract) may include curation expertise to help with (for instance) basic metadata.

4 Benefits of Storing Research Data in a Repository

There are many advantages to you as both a data producer and data user if you and your peers choose to preserve data in a repository. Of course, not all repositories are created equal, and these potential benefits are only realized by selecting a repository that does its job correctly, as described in the next section.

If you are a…

Data Producer/Depositor	Data User
✓ Your Data Management Plan is fulfilled (i.e., satisfies funders/Open Data requirements).	✓ You can easily discover data.
✓ The initial investment of collecting your data is preserved.	✓ You can easily understand your access and usage rights
✓ You have the satisfaction that your data are being stewarded correctly and remain useful and meaningful.	✓ You can reuse/repurpose data without the costs of collection/production.
✓ Your data are looked after long term, even if the data service discontinues.	✓ You can verify (and thus build on) others results, accelerating scientific knowledge.
✓ The ease of discovery of your data is increased.	✓ You can cite peers, knowing that the data will still exist into the future.
✓ Publication, reuse or repurposing, and citation[1] is facilitated for your data.	✓ You have the satisfaction that the data are original/uncorrupted, and that any changes are recorded (provenance).
✓ Recognized expertise is available to assist you with technicalities.	✓ (Re)Use of the data is made easier through full/appropriate metadata in an international or community standard.
✓ It can be ensured that any necessary/wanted conditions on access and use, as well as licensing, are adhered to. (N.B. This is especially important for sensitive data.)	✓ Ability to give feedback to the data producer/holder.

5 What Can Happen if You Make the Wrong Choice

Conversely, if you happen to make the wrong choice of repository, then one loses the benefits listed above and things can quickly go wrong. For example,

- ✗ The repository's system and processes do not operate according to its stated objectives and specifications.
- ✗ Data may be incomplete or include unintended modifications.
- ✗ Data may not actually contain what it is claimed they contain.
- ✗ Access to data and services is not guaranteed, whether that be for technical reasons, or because of licensing issues, or otherwise.
- ✗ Data and services are not usable, again for whatever reason.

6 Things for You to Think About when Selecting a Repository

Depending on who you ask, the essential and/or desirable characteristics for selecting a repository may be quite different: each set of stakeholders within the research endeavour may have their own viewpoint. In particular a scientific publisher or even individual journals, may have a set of criteria or a list of recommendations for repository selection. Such criteria or lists are typically very inconsistent across publishers[3], both in terms of what they contain and how rigidly they are applied. They may also differ quite largely to what is required by your funder or institution.

While it might be questioned whether it is within the remit of journals to point authors to repositories (they do so because they are asked by authors) ultimately the correct execution of your Data Management Plan should take precedence, and how you reconcile any conflicts is beyond the scope of this guidance. Instead, we try to give here a more general overview of what you might look for when selecting a repository.

One important point of note is that if your institution requires you to put your research outputs in its repository, this does not preclude you from also putting a copy elsewhere for more guaranteed long-term management and preservation. However, you must ensure that only one Digital Object Identifier is assigned to your data and that all access and usage licences are consistent. The same is also true if you deposit your data in any other generalist repository first.

7 Trustworthy Data Repositories & Certification

Ideally, when selecting a repository, you should look for what is termed a 'Trustworthy Data Repository' (TDR; sometimes 'Trustworthy Digital Repository'). TDRs are typically accredited against a certification standard in which the repository goes through a review process by an independent third party. A hierarchy of three main certification standards currently exists:

- Core Level: CoreTrustSeal
- Extended Level: DIN 31644/nestor Seal
- Formal Level: ISO 16363

These standards increase in the number of criteria that must be satisfied and the depth of detail they examine, but all look at organizational structure (financial, staffing, and legal aspects, etc.), the management of data (workflows are in place and documented for quality control, discovery, access, reuse, etc.), as well as technical infrastructure (systems appropriateness, operation, and security).

The standards are all based on the Open Archival Information System (OAIS) Reference Model. While it is not important to read through this very technical document aimed at data management specialists, it is of value to you to have some understanding of what it contains. The main points have been distilled down into the much more digestible form in the TRUST Principles for Digital

[3] The Enabling Fair Data project has been successful in ensuring a level of agreement among a number of publishers. A group of publishers (and others) connected with the FAIRsharing repository registry also developed a position paper on this topic.

Repositories, which offer the following guidance for maintaining the trustworthiness of data repositories responsible for the stewardship of research data.

Principle	Guidance for Repositories
Transparency	To be transparent about specific repository services and data holdings that are verifiable by publicly accessible evidence.
Responsibility	To be responsible for ensuring the authenticity and integrity of data holdings and for the reliability and persistence of its service.
User Focus	To ensure that the data management norms and expectations of target user communities are met.
Sustainability	To sustain services and preserve data holdings for the long-term.
Technology	To provide infrastructure and capabilities to support secure, persistent, and reliable services.

8 What a Repository Should Make Clear to You as a Depositor and/or User

The following list is adapted from the CoreTrustSeal Trustworthy Data Repositories Requirements. It summarizes and puts into context the characteristic that the CoreTrustSeal certification expects a repository should have and be transparent[4] about in order to serve its Designated Community well. It may well be that some of these aspects are more or less important to you according to whether you are a depositor or user, and their importance may change from project-to-project. However, all aspects should be considered in terms of whether the repository matches your expectations/needs.

A repository should tell you:

- **Its role in preserving and providing access to data** and that this role is fulfilled through having
  - Sufficient funding and resources (possibly) backed by a recognized host organization.
  - Adequate staff in terms of numbers and expertise, and that their knowledge is kept up-to-date (i.e., through training).
- **How it ensures that data stay valuable and relevant both scientifically and technologically**, including whether/how it obtains advice from experts within its community. It should operate using reliable and appropriate hardware and software to perform its functions and meet user needs.
- **What will happen to the data should it need to stop operation temporarily, or even permanently**. This includes what security procedures are in place to counter natural, human, and technical threats, and protect all aspects of the repository and its users. Access and preservation should continue uninterrupted.
- **How (1) ethical norms are adhered to, (2) sensitive data are managed, and (3) licences for data access and use are maintained**. It should also state all conditions for data use and the consequences for noncompliance.
- **Its processes for deciding whether data will be accepted for deposit, how data will be managed, and how often data are re-evaluated to ensure they are relevant and**

[4] A repository should have publicly available, and easily discoverable and accessible, documentation on its website that gives information on many—if not all—of the items listed here.

understandable to users. It should have a list of preferred data formats, which should align with those actually used by users, as well as assisting in adherence to these formats.

- **The responsibilities of the repository, data depositors, and users for deposited data**. These responsibilities should be stipulated in a depositor agreement (contract) that outlines the transfer of responsibility to the repository.
- **The level of curation it performs**. Namely, what it does to ensure data are fit for purpose and available for discovery and reuse (e.g., error checking of the content or enriching the metadata).
- **How it ensures that data and metadata are not unintentionally changed.** Intentional changes should also be fully documented so that users can understand the differences from the original.
- **How it evaluates the completeness and quality of data and metadata**. Enough information should be provided to enable well-informed decisions about research value and suitability for use.
- **Whether a searchable catalogue (data portal) of its holdings is provided and this enables evaluation, and that it has a citation method to give credit and attribution to those who created the data**. In particular, use of persistent identifiers (PIDs) ensures that data can be accessed and referenced into the future, and these should be integrated into the workflow.

9 Where You Can Look for Suitable Repositories

If you wish to store your data in a generalist repository, this Generalist Repository Comparison Chart lists the main repositories and gives information under various categories to assist in selection.

There are two main registries where you can search for domain repositories that match your needs.

- re3data (registry of research data repositories): Managed under the auspices of DataCite, this registry contains well over 2500 entries. It has a very useful and easy way of visualizing important aspects of the repositories via a set of symbols that show you if:
  1. Additional information is available.
  2. Data are open access.
  3. Information is given on terms of use/licensing.
  4. PIDs are used.
  5. It is certified.
  6. It provides a data policy.

Another useful feature of re3data is the Repository Finder front-end developed to search for repositories that are most likely to meet the FAIR Principles—initially as part of the Enabling FAIR Data project for the Earth, Space, and Environmental Sciences, and then extended to all domains as part of the FAIRsFAIR project.

FAIRsharing: Run by the Data Readiness Group at the University of Oxford, FAIRsharing started out as a project promoting minimum reporting guidelines for biological and biomedical investigations, before becoming the BioSharing portal focussing on the Life Sciences, and finally extending to all disciplines in its current guise. Unlike re3data, which solely lists repositories, FAIRsharing interlinks records on standards (for identifying, reporting, and citing data and metadata), repositories, and journal publisher and funder data policies. It does not highlight the certification of repositories, but does allow filtering according to whether a repository:

1. Is ‘recommended’ within a domain or topic area (see footnote 4).
2. Has a publication attached to it.
3. Is maintained by a particular organization.
4. Has a full and up-to-date record in the registry.

Although certification is highlighted in at least re3data, for ease of reference, you can also find all of the CoreTrustSeal-certified repositories on the List of Repositories and Search Repositories pages of the CoreTrustSeal website. The pages also include those previously certified by the Data Seal of Approval or the World Data System of the International Science Council (WDS).

The WDS is a member-based community of TDRs. In addition to being certified at the core level, WDS Members also adhere to the WDS Data Sharing Principles, which advocate that data, metadata, products, and information should be shared:

- Fully and openly, subject to laws and ethical norms.
- At minimum time delay and free of charge.
- Responsibly, to ensure that their authenticity, quality, and integrity are preserved, respect for the data source is maintained, and appropriate attribution is given.
- On the least restrictive basis possible if data are sensitive.

The repository pages on the CoreTrustSeal website highlight which of the certified repositories are WDS Members. However, it is potentially easier to find that information here on the WDS website.

Acknowledgement

This research is product of the PARSEC group funded by the Belmont Forum as part of its Collaborative Research Action (CRA) on Science-Driven e-Infrastructures Innovation (SEI), supported by the French National Research Agency (ANR), the São Paulo Research Foundation, FAPESP (Brazil), the Japan Science and Technology Agency, the National Science Foundation (USA), and the synthesis centre CESAB of the French Foundation for Research on Biodiversity.

You can cite this document as follows

Edmunds, Rorie, Specht, Alison, Stall, Shelley, David, Romain, Mabile, Laurence, O'Brien, Margaret, Murayama, Yasuhiro, Correa, Pedro, Machicao, Jeaneth, & Miyairi, Nobuko. (2022). Repository Guidelines. Zenodo. https://doi.org/10.5281/zenodo.6542493