



HAL
open science

Integrating Data Stewardship into the Research Lifecycle: A PARSEC Approach

Shelley Stall, Romain David, Rorie Edmunds, Laurence Mabile, Jeaneth Machicao, Yasuhiro Murayama, Margaret O'Brien, Pedro Luiz Pizzigatti Correa, Alison Specht, Lesley Wyborn

► To cite this version:

Shelley Stall, Romain David, Rorie Edmunds, Laurence Mabile, Jeaneth Machicao, et al.. Integrating Data Stewardship into the Research Lifecycle: A PARSEC Approach. AGU Fall Meeting 2020 (#AGU2020), Dec 2020, online everywhere, United Kingdom. , 2022. hal-03740770

HAL Id: hal-03740770

<https://hal.science/hal-03740770v1>

Submitted on 14 Aug 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

Integrating Data Stewardship into the Research Lifecycle: A PARSEC Approach

PARSEC

Integrating Data Stewardship into the Research Lifecycle: A PARSEC Approach

STALL, Shelley (1), DAVID, Romain (2), EDMUNDS, Rorie (3), MABILE, Laurence (4), MACHICAO, Jeaneth (5), MURAYAMA, Yasuhiro (6), O'BRIEN, Margaret (7), PIZZIGATTI, CORREA Pedro (5), SPECHT, Alison (8), VELLEINICH, Danton Ferreira (5), WYBORN, Lesley (9,10)

1- American Geophysical Union, 2- European Research Infrastructure on Highly Pathogenic Agents, 3- World Data System, 4- University of Toulouse 5- University of São Paulo, 6- National Institute of Information and Communications Technology, 7- University of California, Santa Barbara, 8- The University of Queensland, 9- National Computing Infrastructure, 10- Australian National University


What is Data and Software Stewardship?

Researchers are encountering stronger mandates from journals and funders to make their data and software as open (as possible) and preserved in a manner that supports discovery, access, interoperability, and optimizes reuse. These characteristics embody the FAIR principles.

Data Stewardship: An organizational plan of the roles and responsibilities of those overseeing the management of data across all stages of the data lifecycle, including its preservation. A large research project may involve several data stewards as the data moves from stage to stage across the lifecycle.

Source: Original Research Data

Data Stewardship - The Five-Legged Sheep



From the book *Engaging Researchers with Data Management: The Cookbook* **Martine Pronk**, the head of Academic Services of the Utrecht University (UU) Library explains that "[t]he 'five-legged sheep' is an old Dutch expression, referring to somebody who needs to be unreasonably versatile. And sadly, this is what is expected from the twenty-first-century researcher. Transparent and well-documented data management is one of many tasks that researchers now need to add to their already heavy workload."

The PARSEC team is working hard to make these stewardship tasks easier for all researchers.

Data Stewardship entails determining and implementing practices during the research effort on:


1. How datasets are selected for use by the team.
2. How data are created by the team.
3. How data are protected during the project. This includes any sensitive or restricted access datasets.
4. The quality standards for the data and how they

Software Stewardship - The Alpaca




If Data Stewardship can be seen as a sheep, then Software Stewardship looks more like an alpaca. Still producing wool, but a different animal all together. Though similar in concept, software

What PI's need to know - The Sheepdog Role



As the Principal Investigator (PI), you are responsible for the progress and

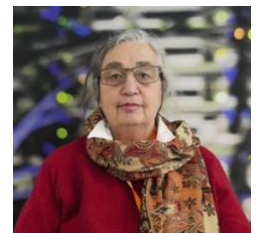
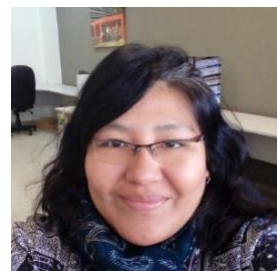
Why are stewardship practices valuable to me as a researcher?





STALL, Shelley (1), DAVID, Romain (2), EDMUNDS, Rorie (3), MABILE, Laurence (4), MACHICAO, Jeaneth (5), MURAYAMA, Yasuhiro (6), O'BRIEN, Margaret (7), PIZZIGATTI, CORREA Pedro (5), SPECHT, Alison (8), VELLEINICH, Danton Ferreira (5), WYBORN, Lesley (9,10)

1- American Geophysical Union, 2- European Research Infrastructure on Highly Pathogenic Agents, 3- World Data System, 4- University of Toulouse 5- University of São Paulo, 6- National Institute of Information and Communications Technology, 7- University of California, Santa Barbara, 8- The University of Queensland, 9- National Computing Infrastructure, 10- Australian National University



WHAT IS DATA AND SOFTWARE STEWARDSHIP?

Researchers are encountering stronger mandates from journals and funders to make their data and software as open (as possible) and preserved in a manner that supports discovery, access, interoperability, and optimizes reuse. These characteristics embody the FAIR principles.

Data Stewardship: An organizational plan of the roles and responsibilities of those overseeing the management of data across all stages of the data lifecycle, including its preservation. A large research project may involve several data stewards as the data moves from stage to stage across the lifecycle.

Source: Original Research Data Canada Glossary, <https://www.rdc-drc.ca/glossary/original-rdc-glossary/>

Software Stewardship: **Software Stewardship** has a similar definition to Data Stewardship, i.e. to define an organizational plan of the roles and responsibilities of those overseeing the management of software across all stages of the lifecycle. Though the nature of software and its lifecycle are somewhat different than data, the need for stewardship to align with funder and publisher policies is similar.

WHY ARE STEWARDSHIP PRACTICES VALUABLE TO ME AS A RESEARCHER?



As a researcher, it can be difficult to understand why Data and Software Stewardship is important. It often happens that in the long term, data and software are more valuable than the research results themselves.

Like caring for sheep and alpaca, the care we take in the well-being of these animals results in high-quality wool that can be used by either ourselves or others.

Data and software also need to be cared for in order to protect and manage them during the research effort, as well

as preserve them as an important scholarly output supporting transparency, integrity, and reproducibility.

The time we spend on stewardship for data and software results in:

1. Well-understood data and software that we can reference long after the paper is published.
2. A smoother experience when submitting your paper to a journal with easy to cite data and software.
3. Compliance with your funder and publisher in sharing your data and software.
4. Improved transparency and integrity of our research.
5. New scientific insights to be gained from the reuse or repurposing of our data and software for applications that had not previously been considered to help unravel the environmental grand challenges of today and those of tomorrow that are yet to be identified.

DATA STEWARDSHIP - THE FIVE-LEGGED SHEEP



From the book *Engaging Researchers with Data Management: The Cookbook*: **Martine Pronk**, the head of Academic Services of the Utrecht University (UU) Library explains that “[t]he **‘five-legged sheep’** is an old Dutch expression, referring to somebody who needs to be unreasonably versatile. And sadly, this is what is expected from the twenty-first-century researcher. Transparent and well-documented data management is one of many tasks that researchers now need to add to their already heavy workload.”

The PARSEC team is working hard to make these stewardship tasks easier for all researchers.

Data Stewardship entails determining and implementing practices **during the research effort** on:

1. How datasets are selected for use by the team.
2. How data are created by the team.
3. How data are protected during the project. This includes any sensitive or restricted access datasets.
4. The quality standards for the data and how they are implemented.
5. How the provenance, transformations, aggregations, of datasets are tracked by the team.
6. The actions each team member takes periodically to ensure they follow the practices of the team around data stewardship.
7. The actions the team leadership takes to ensure each team member understands their role in data stewardship and follows the agreed upon practices and expectation.

And practices for planning and preparing **for the time of dataset preservation** on:

- a. How the preservation repository is selected.
- b. Contacting the preservation repository to include any additional guidance in this practice.
- c. The licensing of the datasets to be as open as possible.
- d. The format of the preserved dataset and if any transformation is needed prior to depositing.
- e. The time needed to prepare for preservation, for both the research team and the repository.
- f. Any associated costs with preservation that need to be identified in the budget if they have not already.
- g. The necessary documentation to understand the dataset and how that is gathered during the project.
- h. The relationships (links) the dataset has with the project (Grant ID, Institution ROR), data creators (ORCIDs), Publications (DOIs), Primary datasets from which this one is derived (persistent identifiers), software (DOI or other PID), and other relevant digital output.

See the PARSEC Data and Digital Output Management Plan for processes to help you with these tasks.

SOFTWARE STEWARDSHIP - THE ALPACA



If Data Stewardship can be seen as a sheep, then Software Stewardship looks more like an alpaca. Still producing wool, but a different animal all together. Though similar in concept, software documentation, management, and preservation are unique skills for Software Stewardship. The hurdle is providing these techniques early in the career of the researcher so that they can integrate them into the research lifecycle in order to fully benefit.

Very commonly researchers feel their software is too messy to share. Or they find navigating the stark contrast between the iterative evolution of a software package and the need for a persistent, preserved version to be connected with a

publication as being difficult.

In PARSEC, we have identified the processes, management, and documentation for us to use as our work progresses. We are confident this will help when we get ready to share and publish our work.

Software Stewardship entails determining and implementing practices during the research effort on:

1. How software is developed by the team with selection of language, tools, and environment.
2. How software configuration and version management is performed during the project.
3. The quality standards for the software and how they are implemented.
4. How the software is documented both internally to the code and externally.
5. The actions each team member takes periodically to ensure they follow the practices of the team around software stewardship.
6. The actions the team leadership takes to ensure each team member understands their role in software stewardship and follows the agreed upon practices and expectation.

And practices for planning and preparing for the time of software preservation on:

1. How the preservation repository is selected.
2. The licensing of the software to be as open as possible.
3. The format of the software to avoid (as much as possible) proprietary formats.
4. The time needed to prepare for preservation, for both the research team and the repository.
5. Any associated costs with preservation that need to be identified in the budget if they have not already.
6. The necessary documentation to understand the software and how that is prepared during the project (e.g., readme file, CodeMeta, Cf file).
7. The relationships (links) the software has with the project (Grant ID, Institution ROR), data creators (ORCID), Publications (DOIs), datasets (persistent identifiers), executable environment, and other relevant digital output.

See the PARSEC Data and Digital Output Management Plan for processes to help you with these tasks.

WHAT PIS NEED TO KNOW - THE SHEEPDOG ROLE



As the Principal Investigator (PI), you are responsible for the progress and accomplishments of the entire research team. In order to comply with the data and software sharing requirements of your funder and publisher you need to incorporate a set of processes and decisions that need to be made as you begin and conduct your research. Your team needs to support these decisions and help ensure the proper stewardship is being conducted throughout the entire project.

Things to do:

1. Make decisions on data and software storage during your project that best support you as you progress and when you are ready to preserve your datasets and software.
2. Make decisions on data and software preservation that you will use prior to publishing your research or even during critical milestones.
3. Contact the preservation repositories to ensure the methods you chose for managing your data and software will make it easier for deposit. This includes not only the format of your data, but information (metadata) to document as you go. You don't want to guess at this later if you forget to capture it.
4. Make sure you are following your institution requirements, as well as your funder(s), country, and your selected publisher. Also, take into account requirements from researchers you are collaborating with from other organizations or countries.
5. If you are collecting data from countries that require permission to do so, or even to take the data outside the country, make sure you begin this process early enough to deal with necessary steps.
6. And FINALLY, make sure your team is following the practices you have put into place through tools like checklists.

See the PARSEC Data and Digital Output Management Plan and checklist for processes to help you with these tasks.



Figure 1: PARSEC PIs Nicolas Mouquet, Shelley Stall, Alison Specht, David Mouillot

Data and Digital Output Management Plan and
Workbook for the Belmont Forum

Collaborative Research Action (CRA) Science-driven
e-Infrastructure Innovation (SEI) for the Enhancement
of Transnational, Interdisciplinary and Transdisciplinary
Data Use in Environmental Change

**Project Building New Tools for Data Sharing and Reuse through a Transnational
Investigation of the Socioeconomic Impacts of Protected Areas (PARSEC)**

June 2020



Figure 2: The front page of the PARSEC Data and Digital Output Management Plan and Workbook

ABSTRACT

Earth, space, and environmental science researchers have to be proficient in the expertise of their discipline, as well as the technology of instrumentation, analysis tools, writing, creating visualizations and countless other skills necessary to further their careers. The very active movement of ensuring our scientific data and software are well managed and preserved require the further development of data (and software) stewardship techniques that were likely self-taught, or possibly learned from colleagues.



Journals and funders now require that our scientific data (and software) be as open (as possible) and preserved in a manner that supports discovery, access, interoperability, and optimizes reuse. These characteristics embody the FAIR principles. Researchers typically are faced with complying with these new data (and software) preservation requirements at the time of publication. This tends to be out-of-synch with their research work and makes it difficult and time consuming to comply. Plus it can slow down the process of publication.

In our work on the PARSEC project, funded by the Belmont Forum, we are leveraging the current research process and adding in the necessary stewardship tasks, when they are most optimal and least disruptive, to capture the necessary information, provenance, decision made about the data, workflows and software necessary to adequately preserve these research objects allowing for better transparency of the process, support integrity, and optimize reuse. In this talk we will share our approach and the tools for others to incorporate the technique.



Figure 3: The Herd

The PARSEC project funded by the Belmont Forum, Collaborative Research Action on Science-Driven e-Infrastructures Innovation.

REFERENCES

Belmont Forum, Data and Digital Outputs Management Plan (DDOMP) Guide, <https://bfe-inf.github.io/toolkit/ddomp.html>

Clare, C., Cruz, M., Papadopoulou, E., Savage, J., Teperek, M., Wang, Y., Witkowska, I., & Yeomans, J. (2019) *Engaging Researchers with Data Management: The Cookbook*. Cambridge, UK: Open Book Publishers. <https://doi.org/10.11647/OBP.0185>

Cruz, M.J., Kurapati, S., Turkyilmaz-van der Velden, Y., 2018. The Role of Data Stewardship in Software Sustainability and Reproducibility, in: 2018 IEEE 14th International Conference on E-Science (e-Science). Presented at the 2018 IEEE 14th International Conference on e-Science (e-Science), IEEE, Amsterdam, pp. 1–8. <https://doi.org/10.1109/eScience.2018.00009>

David, R., Mabile, L., Specht, A., Stryeck, S., Thomsen, M., Yahia, M., Jonquet, C., Dollé, L., Jacob, D., Bailo, D., Bravo, E., Gachet, S., Gunderman, H., Hollebecq, J.-E., Ioannidis, V., Le Bras, Y., Lerigoleur, E., Cambon-Thomsen, A. and Alliance – SHaring Reward and Credit (SHARC) Interest Group, T.R.D. (2020) FAIRness Literacy: The Achilles' Heel of Applying FAIR Principles. *Data Science Journal*, 19(1), p.32. DOI: <http://doi.org/10.5334/dsj-2020-032>

Stall, S., Specht, A., Corrêa, P.L.P., David, R., Edmunds, R., Mabile, L., Machicao J., O'Brien M., & Wyborn, L. (2020). PARSEC Data and Digital Output Management Plan and Workbook. Zenodo. [10.5281/zenodo.3891426](https://doi.org/10.5281/zenodo.3891426)

Stall, S., & Specht, A. (2020, July). Data and Digital Output Management Plan (DDOMP), JpGU M-GI36 Open Science in Progress. Presented at the Japan Geoscience Union (JpGU), Chiba, Japan held virtually: Zenodo. <http://doi.org/10.5281/zenodo.3942688>

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3, 160018 (2016). <https://doi.org/10.1038/sdata.2016.18>