How to stop re-inventing the wheel: a data management case study

Jen Thomas, Swiss Polar Institute (Switzerland), jenny.thomas@epfl.ch
Marco Alba, EMODnet Physics (Italy), marco.alba@ettsolutions.com
Eric Bouillet, Swiss Data Science Center (Switzerland), eric.bouillet@epfl.ch
Antonio Novellino, EMODnet Physics (Italy), antonio.novellino@ettsolutions.com
Carles Pina Estany, Swiss Polar Institute (Switzerland), carles.pinaestany@epfl.ch
Michele Volpi, Swiss Data Science Center (Switzerland), michele.volpi@sdsc.ethz.ch

Introduction

Over the past three years, the Swiss Polar Institute (SPI) has adopted and integrated a number of existing tools and services to manage data from a recent expedition in order to assist scientists in making them openly accessible. SPI aimed for secure data storage, well-documented, discoverable and citable datasets and recording of dataset provenance.

Here we put into context the data to be managed, explain the use of tools chosen to implement the data management workflow, as well as notable successes and challenges of putting this into practice.

The data in question

Initial data to be managed resulted from the Antarctic Circumnavigation Expedition[1] (ACE). Data and metadata were managed during the expedition using **Django**, **Python** and **MySQL**[2]; tools that allowed rapid development of a relational database to record metadata, monitor backup of files, and produce simple reports or displays of the data collected. Almost 30,000 samples and over 20 TB of data were recorded, resulting in around 200 datasets[3] from a range of natural science disciplines.

Selection of tools and services

Rclone[4], an open-source tool, is used to transfer raw data files to the object storage and perform consistency checks on data files.

SPI was able to adopt **RENKU**[5] as a tool for recording the provenance of datasets through a collaboration (ACE-DATA: Delivering Added value To Antarctica[6]) between SPI, the Paul Scherrer Institute and Swiss Data Science Centre. This open platform addresses the issue of data versioning and ensures reproducibility by capturing what was done to the data, by whom and when.

Zenodo[7] offers long-term archiving of published datasets and was chosen for ACE dataset publication because citation and dataset re-use metrics are facilitated using a Digital Object Identifier (DOI).

EMODnet Physics[8] incorporates already-published data from Zenodo into its well-established repository, displaying it through a user-friendly map-based portal as well as adding value to the data through visualisations. Data are described using machine-readable metadata schemas from **Frictionless Data**[9] and validated by goodtables-py[10] prior to publication in Zenodo. In addition to a Github repository with Github actions to provide this information for already-published datasets, this provides machine-readable metadata to EMODnet Physics.

Facilitating the discovery of data from ACE is important to ensure that it reaches its potential for maximum re-use and can be incorporated into further regional and discipline-specific studies. **RENKU**, **Zenodo** and **EMODnet Physics** all contribute to this aim: **RENKU** exposes data-derived insights to a perhaps unforeseen audience; text-based search of **Zenodo** allows discovery and through their use of schema.org[11] and adoption of the DataCite Metadata Schema[12], also opens up the data to potential re-use through other repositories such as Mendeley Data and Google dataset search; and the

EMODnet Physics map-based portal displays the data to more domain-specific users as well as those interested in the geographic region through the SOOS map[13].

Internally, SPI uses **Python** as the language of choice for writing utility scripts that bring some parts of this workflow together. **Github** is used to manage, document and finally publish releases of the code through its link with **Zenodo**. Whilst dataset documentation is always maintained alongside the data, project and data management documentation is written in **Github** and a **C4Science** wiki.

	rclone	MySQL	Django	Python	Gitlab	Github	c4science	Frictionless Data	RENKU	Zenodo	EMODnet Physics
Project management and documentation						■	■				
Code management and publication						■				■	
Data file management	■	■	■	■	■				■		
Data provenance				■	■				■		
Data description								■	■	■	
Data publication						■			■	■	■
Data discovery						■		■	■	■	■
Data visualisation				■							■

Table 1. Overview of tools and services, and the roles they play within data management at the Swiss Polar Institute.

Can we continue to "reduce, reuse and recycle" software tools and services in this world of plenty?

The adopted tools meet the needs of the SPI for managing current and future datasets. Researchers, institutes, data managers and others are always looking to existing tools to meet their needs. Use of the open-source technologies described here, is in line with the principles of open science such as reproducibility and transparency, setting an example for requiring open access to datasets.

The most pressing challenge is the lack of integration between these tools and services. As the number of datasets grows, it is becoming more evident that connecting each tool with custom-developed software to form an automated and continuous workflow, would greatly improve reliability, robustness and scalability.

We will continue to look for existing tools to adopt that meet the needs of what we require, but cannot shy away from building something new if it would be more suitable. In short, it is possible to stop reinventing the wheel, adopt existing software tools and services, collaborate and integrate.

References

[1] Walton, D.W.H and J. Thomas. (2018). Cruise Report - Antarctic Circumnavigation Expedition (ACE) 20th December 2016 - 19th March 2017 (Version 1.0). Zenodo. doi: 10.5281/zenodo.1443511

[2] Pina Estany, C and J. Thomas. (2019). Swiss-Polar-Institute/science-cruise-data-management v0.1.0 (Version 0.1.0). Zenodo. doi: 10.5281/zenodo.3360649

[3] https://zenodo.org/communities/spi-ace

[4] Craig-Wood, Nick. (2020). Rclone. https://rclone.org/

[5] Swiss Data Science Center. RENKU. https://datascience.ch/renku/

[6] Thomas, J., S. Landwehr, M. Volpi and J. Schmale. (2019). ACE-DATA: Antarctic Circumnavigation Expedition: Delivering Added value To Antarctica (Version 1.0). Zenodo. doi: 10.5281/zenodo.2587954

[7] Zenodo. (2009-2017). CERN. https://zenodo.org

[8] EMODnet Physics project. European Marine Observation and Data Network. www.emodnet-physics.eu/map

[9] Fowler, D., Barratt, J. and Walsh, P. (2018) 'Frictionless Data: Making Research Data Quality Visible', International Journal of Digital Curation, 12(2), pp. 274–285. doi: 10.2218/ijdc.v12i2.577

[10] Frictionless Data. goodtables-py. Available at: https://github.com/frictionlessdata/goodtables-py

[11] schema.org. (2020). https://schema.org

[12] DataCite Metadata Working Group. (2019). DataCite Metadata Schema for the Publication and Citation of Research Data. Version 4.3. DataCite e.V. doi: 10.14454/f2wp-s162

[13] SOOSmap. Southern Ocean Observing System. (2020). www.soosmap.aq