"What About Model Data?"
Best Practices for Preservation and Replicability

User Guidance

Document Purpose: Information for understanding and using rubric and use case materials

Current Rubric Worksheet links:
https://modeldatarcn.github.io/rubrics-worksheets/Descriptor-classifications-worksheet-v2.0.pdf
https://modeldatarcn.github.io/rubrics-worksheets/Descriptor-classifications-worksheet-v2.0.xlsx

Document Organization:
1. Project Overview
2. Data Production versus Knowledge Production
3. Rubric Instructions
4. Guidance for Data Production Projects
5. Use Case Instructions

1. Project Overview

There is strong agreement across the sciences that replicable[1] workflows are needed for computational modeling. Open and replicable workflows not only strengthen public confidence in the sciences, but also result in more efficient community science. However, recent efforts to standardize data sharing and preservation guidelines within research institutions, professional societies, and academic publishers make clear that the scientific community does not know what to do about data produced as output from computational models. To date, the rule for replicability is to "preserve all the data", but simulation data can be prohibitively large, particularly in a field like atmospheric science. The massive size of the simulation outputs, as well as the large computational cost to produce these outputs, makes this not only a problem of replicability, but also a "big data" problem. The ultimate goal of the EarthCube Research Coordination Network (RCN) project "'What About Model Data?' Determining Best Practices for Preservation and Replicability" is to provide simulation data management best practices to the community, including publishers and funding agencies.

2. Data Production versus Knowledge Production

The majority of research involving simulations is knowledge production not data production[2]. In other words, the primary goal of most projects involving computer simulations is to increase scientific knowledge and the simulations are used as a tool to that end. Data production

projects (e.g., CMIP), in contrast, are motivated by scientific questions, but the primary goal is to provide a dataset that multiple users can access to investigate those scientific questions. While most researchers that produce simulation output would welcome more use of their output products, and many end users would welcome more data availability, the reality is that we are producing far more simulation output than can be reasonably stored in repositories. Knowledge production research should preserve minimal output in repositories. Further guidance in determining whether a given project is knowledge or data production is provided in the Rubric Instructions.

3. Rubric Instructions

The rubric is built to help researchers make decisions about what simulation output needs to be shared via a repository, i.e. made accessible and preserved a sufficient time to satisfy the requirements of publishers and funding agencies. Ultimately, these decisions are based on the goal of all community members (researchers, publishers, end users) to communicate knowledge in a sustainable way. Note this rubric is not meant to dictate what a researcher or research group keeps on their own local storage.

The rubric is a list of simulation/experiment descriptors, organized into themes. To use the rubric, consider a specific simulation workflow and review all of the descriptors, selecting a Class (1, 2, or 3) that best fits your workflow for each descriptor. If you find that your workflow could be a Class 1 or 3 based on which descriptor you are looking at, you can separate the workflow into logical sections and then classify each section separately.

Once you have selected classes for each descriptor, multiply classifications of 2 or 3 by the recommended weighting shown in the “suggested weighting” column. Certain researchers may have compelling reasons to adjust the weighting based on the goals of a specific project, but testing thus far has shown these recommended weights to be useful for most projects.

Finally, the total weighted score will indicate what to deposit into a repository: few outputs, some outputs, or the majority of outputs (see score ranges at the bottom of the rubric). Further guidance on specifically what outputs to save are provided in the Use Case Example document. See instructions for using that document in section 5 below.

4. Guidance for Data Production Projects

If your workflow scored high in the first theme area of the rubric, Community Commitment, your research may be a data production project, regardless of your total score. Data production projects are designed to have significant products preserved with wide accessibility. These projects hopefully included an appropriate budget to support anticipated data preservation and community data access needs.

Although outputs preserved are often broadly dictated by the research proposal for data production research, the rubric and use case examples may still be useful. For example, additional decision-making may be needed regarding what specific outputs should be preserved

and shared. This project can also be helpful when communicating with end users about why not all the outputs can be made available.

5. Use Case Instructions

The purpose of the Use Case Example document is to give examples of other projects in a particular rubric score range. Specifically, what data are being preserved and why? As all projects are unique, there is no one solution for all projects, even when rubric scores are identical. Instead, the Use Case document is meant to assist a researcher in making thoughtful and informed decisions on what data preservation is necessary for replicable science. Please note that the Use Case Example document is a first draft version.

While each project is unique, certain data should be preserved for all projects. The following components should be preserved unless they are already publicly available from another provider:

- simulation code
- initialization data
- simulation setup (e.g., parameterization selection)
- pre-processing code
- post-processing code

Note that while we encourage full preservation for all above categories, several challenges have been identified in workshop discussions, including proprietary data/code and giant non-accessible datasets (e.g., operational forecasts). Examples of these are included in the Use Case document. The challenge of sharing code also includes the continuing concerns that code is not cited properly, such that sharing code can be detrimental to researchers getting credit for their work. This last issue is a challenge, but not a barrier, and we must continue to work towards better code/data citation practices (and recognition of such in job evaluations) to achieve our goals of open, efficient, and replicable science.

With the above components saved for all projects, the Use Case Example document can then be used to make decisions on preserving simulation output. As discussed in section 2, most projects are Knowledge Production research, and as such should be saving little to no raw simulation data in repositories and focus on 2-D derived fields that help communicate to future researchers the environmental state or other information important for building similar studies in the future. Particularly for highly nonlinear case studies, the goal is not exact reproducibility, but rather enough output to understand the environmental state that forced, and the impacts of, the features being investigated.

[1]National Academies of Sciences, Engineering, and Medicine. 2019. Reproducibility and Replicability in Science. Washington, DC: The National Academies Press. https://doi.org/10.17226/25303.

[2]Baker, K. S., and Mayernik, M. S. 2020. Disentangling knowledge production and data production. *Ecosphere* 11(7):e03191. 10.1002/ecs2.3191

[3]See https://zenodo.org/record/3482769#.X-zbd-lKifU and https://zenodo.org/record/3479199#.X-zbU-lKifU