Data Management Plan for Computational reproducibility in High-Performance Computing (HPC) (v. 1.0)

Abstract

This exemplar DMP was created to document development and implementation of practices in managing and curating both a dataset of input and output data from a numerical simulation project involving several team members, and the accompanying software for use/reuse of the data. This exemplar demonstrates how a DMP may be developed that thoroughly describes software, file types, and metadata documentation, as well as demonstrating how the DMP interacts with a larger project management plan.

Administrative Details

Project Name:

Computational reproducibility in High-Performance Computing (HPC) - a use case study in relativistic astrophysics (v. 1.0)

Principal Investigator / Researcher:

Qian Zhang

Project Data Contact:

Qian Zhang, PhD

Description:

This is a case study of computational reproducibility in High Performance Computing - a use case in relativistic astrophysics and gravitational physics. The goal of this project is, with the use of an open, community-driven software platform of core computational tools, to develop and implement practices in managing and curating both dataset and scientific software for reuse.

Institution:

University of Waterloo

Data Management Plan for Computational reproducibility in High-Performance Computing (HPC) (v. 1.0)

Data Collection

What types of data will you collect, create, link to, acquire and/or record?

Since this is a numerical simulation project, the input and output of the numerical model will be the input and output data of this project, respectively. And both are compatible data formats and types that comply with the numerical model Einstein Toolkit[1] (ET). Specifically, the input file is stored in the parameter file (**.par**), which describes the grid structure information, modules (named "thorns" in ET) to be used, initial and boundary conditions, etc. of the simulation run. The output that is produced from running the simulation is simulation/modeling datasets.

What file formats will your data be collected in? Will these formats allow for data re-use, sharing and long-term access to the data?

The input data is stored in the **.par** file . The simulation output files are stored in the **.asc** files (a type of space-separated-value ASCII file). Although both file formats are specific to the Einstein Toolkit, they can be read and opened by a text editor, rather than requiring specialized software to use. Therefore these file formats allow for data re-use, sharing and long-term access to the data.

Including links in your DMP to external tools and software programs helps direct users to more information about which specific features of those tools will support the management of your data. Don't assume that all of these tools are necessarily familiar to those reviewing your DMP.

Define your file types/extensions, and indicate where appropriate which programs can read these files, and whether they are open or proprietary.

[1] Einstein Toolkit (http://einsteintoolkit.org/): is a collaborative community-driven computational infrastructure to advance and support high-performance computing (HPC) research in relativistic astrophysics and gravitational physics.

Data Management Plan for Computational reproducibility in High-Performance Computing (HPC) (v. 1.0)

What conventions and procedures will you use to structure, name and version-control your files to help you and others better understand how your data are organized?

The complete source code, documentation and tools included in the ET are distributed under open-source licenses. The ET itself maintains a version control system[2] (Most thorns stored in Subversion (SVN), some in Git) with open access. Thus obtaining the ET (ET_2019_10 "Mayer"[3], released 2019-10-25) directly from its repositories[4] requires SVN and Git tools.

For file naming, ET has developed a well-structured and consistent standard that follows the general file naming convention guidelines.

ET is an active and ongoing tool that is designed to be well-organized to have its own directory hierarchy structure for its input and output datasets. For each simulation run, all that is needed is to provide a name for that particular run, such that all the output files will be automatically created, stored and saved under that centralized (rather than distributed) directory with the name that is provided before the simulation. Therefore, duplicated simulation names will not be allowed.

If using a software program that automates file naming, file structure and versioning, may be sufficient to refer readers to the documentation for this software. If such a program is not being used, you may need to provide specific examples.

[2] http://svn.einsteintoolkit.org/
[3] http://einsteintoolkit.org/about/releases/ET_2019_10_announcement.html
[4] https://bitbucket.org/einsteintoolkit/einsteintoolkit/src/ET_2019_10/

Documentation and Metadata

What documentation will be needed for the data to be read and interpreted correctly in the future?

All simulation output data are self-explanatory, meaning there are a couple of lines of comments at the very beginning within each file, containing such as a brief description about the data, creation date, parameter file used, and the meaning of each column, etc.

A documentation for describing the input parameter file (.**par**) will be created.

A documentation of how to download, compile, install and run the numerical model and simulation will be provided.

A report of the post-processing data analysis incorporating data and code will be generated.

Note that documentation is provided to describe both input and output data.

How will you make sure that documentation is created or captured consistently throughout your project?

The JIRA Agile[5] project management tool will be used throughout the active project management to ensure the smooth and flexible coordination within the project team. The project stakeholders review progress and follow up on the issues and tickets in active sprints on a regular basis. All discussions, decisions and ideas are well documented in the backlog. where moving tickets back and forth between different project queues is fairly easy and flexible. An active sprint is often classified as to-do list, in-progress list, and done list, etc. On the other hand, GitHub is a very helpful tool in the sense of code sharing and versioning control for software development.

The PI will also create a shared Google Drive among team members for each member to contribute their ideas, opinions and research progress. We will create a detailed timeline and action items for the reproducibility workflow and then follow them.

5 https://www.atlassian.com/software/jira/agile#:~:text=Jira%20Software%20is%20an%20agile,projects%20from%20a%20single%20tool.

Regular group meetings will be scheduled with collaborators to ensure that we keep track of notes, discussion and decisions from each meeting. In addition, due to the reproducibility essence of this project, the PI will intentionally capture all the communications (email, online chats, conversations, etc.), activity logs, brainstorming records and computational workflows, also within Google Drive.

This section is really about project management, and should be consistent with the overall project management process. Particularly on larger projects, different team members may not interact regularly, and may be working on discrete portions of the project, so it is important to establish clear and explicit procedures. This documentation is also useful in the event of turnover among the project team.

If you are using a metadata standard and/or tools to document and describe your data, please list here.

Metadata creation and documentation are described in the sections above.

There is no suitable metadata standard available at this point.

Storage and Backup

How and where will your data be stored and backed up during your research project?

The simulation will make use of high-performance computing (HPC) resources that are available through the Extreme Science and Discovery Environment (XSEDE) program. The two computational clusters from the XSEDE ecosystem are: Stampede2's Skylake (SKX)[6] is one of the Texas Advanced Computing Center (TACC), University of Texas at Austin's flagship supercomputers; Comet[7] is a dedicated XSEDE cluster designed by Dell and San Diego Supercomputer Center (SDSC). Both clusters will provide a scratch space or filesystem for temporary and large-sized simulation data that are produced by the ET. Until we are ready to

[6] https://portal.xsede.org/tacc-stampede2
[7] https://portal.xsede.org/sdsc-comet

disseminate and share the data and codes, all datasets and software will reside in the clusters, maintained by the cluster's IT team.

What are the anticipated storage requirements for your project, in terms of storage space (in megabytes, gigabytes, terabytes, etc.) and the length of time you will be storing it?

For each simulation, depending on the resolution of the grid structure, gigabytes to terabytes of storage will be needed.

Currently, we plan to run the simulation using Stampede2's Skylake (SKX) cluster at TACC and Comet cluster at SDSC, the storage of which is 30 PB and 7 PB, respectively. According to the XSEDE Allocations Info & Policies, “All allocations for XSEDE resources are made for a 12-month period. PIs can continue their activities in subsequent years through annual renewal requests (Section 3.5.1)”. It is worth mentioning that this storage and computational resources are for the active project management phase, rather than a long-term storage or preservation purpose.

After that, we will pursue reproducibility from Amazon AWS. We have submitted an application for the on-demand AWS Cloud Credits for Research through the AWS Programs for Research and Education. If the proposal is approved and the cloud credits are granted, we will test our reproducibility use case through this program. If not, we will use part of our research grant to purchase such computational and associated storage resources, approximately the same as we have requested from XSEDE. The corresponding storage and duration will depend on AWS service’s policy.

How will the research team and other collaborators access, modify, and contribute data throughout the project?

Each team member will get their own computational resource allocations on the clusters (SKX at TACC and Comet at SDSC) under the same project, so they can work independently on the simulation. The project team will use both JIRA Agile and the shared Google Drive folder created by the PI to share, compare and discuss the results, and debug and solve any problems during the reproducible research.

Preservation

Where will you deposit your data for long-term preservation and access at the end of your research project?

The project team will give priority to the institutional data repository - Illinois Data Bank[8], or the Federated Research Data Repository (FRDR)[9], since they offer long-term sustainable data curation and preservation services. Both repositories also mint DOIs, unique permanent identifiers that allow for stable long-term discovery of data, and improve the citation and discoverability of the data.

Indicate how you will ensure your data is preservation ready. Consider preservation-friendly file formats, ensuring file integrity, anonymization and de-identification, inclusion of supporting documentation.

Both **.par** and **.asc** are preservation-friendly file formats that can be read by a text editor.

Sharing and Reuse

What data will you be sharing and in what form? (e.g. raw, processed, analyzed, final).

Simulation data sharing will be part of the research topics of this project. For big data simulation, it will be very expensive to store the whole simulation output package, the long-term value of which is yet unpredictable. If we could provide the exact version number ("Mayer" release in this project) of the software and associated source code we use, and the input and/or parameter file (**.par** file in this project), plus adequate documentation of computational workflows, intermediate analyses and results, and relevant metadata information, it might be sufficient for others to reproduce the scientific results. We will monitor data growth and modify plans for sharing and preservation as the project proceeds.

[8] https://databank.illinois.edu/
[9] https://www.frdr-dfdr.ca/repo/: FRDR is a collaboration between Portage Network, Compute Canada (CC) and the Canadian Association of Research Libraries (CARL) to provide a scalable federated platform for digital research data management (RDM) and discovery.

In principle, the software code developed within the team, together with all intermediate data analyses and results will also be shared for others to use.

For the long term, all files within the shared Google Drive folder will be transferred to Open Science Framework (OSF)[10].

Have you considered what type of end-user license to include with your data?

For the software code developed by the project team, we plan to choose an open source license[11].

For data sharing including the pre-/post-processing analysis, we will pick one license from either a Creative Commons[12] or Open Data Commons[13]. This decision will be made in consultation with local experts on data sharing, or in accordance with the licenses available through our chosen repository.

What steps will be taken to help the research community know that your data exists?

When sharing data and software, a unique Digital Object Identifier (DOI) will be generated, assigned and associated to each research object. This facilitates discovery and accurate citation by providing a link that remains stable and persistent over time.

[10] https://osf.io/
[11] https://choosealicense.com
[12] https://creativecommons.org/
[13] https://choosealicense.com/

Responsibilities and Resources

Identify who will be responsible for managing this project's data during and after the project and the major data management tasks for which they will be responsible.

This will be an ongoing project. So the PI and the project manager will be responsible for managing the project data within and beyond the lifetime of this research use case.

How will responsibilities for managing data activities be handled if substantive changes happen in the personnel overseeing the project's data, including a change of Principal Investigator?

The Co-PI and the project manager will take over the full responsibilities of data management if substantial changes happen.

What resources will you require to implement your data management plan? What do you estimate the overall cost for data management to be?

The project team will monitor our use of HPC computational resources allocations from XSEDE, to ensure we remain within storage limits. Should we exceed these limits, we may need to purchase additional storage from Amazon Web resources, which will be monitored using the same process.

Ethics and Legal Compliance

If your research project includes sensitive data, how will you ensure that it is securely managed and accessible only to approved members of the project?

Not applicable – there is no sensitive data associated with this project.

If applicable, what strategies will you undertake to address secondary uses of sensitive data?

Not applicable - no sensitive data associated with this project.

How will you manage legal, ethical, and intellectual property issues?

The project team will ensure that software and data are used in accordance with any licensing requirements; for example, the Einstein Toolkit; and give appropriate credit through citation.

The copyright of all research objects produced from this project will be shared between both universities, i.e., the University of Waterloo and the University of Illinois at Urbana-Champaign.