Data Management Plan Template: **Advanced Research Computing**

Abstract

Advanced Research Computing (ARC) provides researchers with digital technology, infrastructure and expertise to help them solve research problems that are either too large or too complex to undertake by other means. It includes access to both computational and storage resources, such as multi-core and many-core high performance computing (HPC or supercomputers) systems, distributed high-throughput computing (HTC) environments, large-scale data analysis frameworks (e.g., Hadoop, Spark), visualization and data analysis systems, large-memory systems, data storage, and cloud systems. This template is intended for researchers whose research cannot be conducted on a traditional computer but has to rely on one or more of the advanced research computing resources mentioned above. ARC-based research occurs in a wide range of fields including genomics, molecular dynamics, bioinformatics, neuroscience, biochemistry, quantum chemistry, structural mechanics, astrophysics, energy economics, climate change, machine learning, artificial intelligence, and the humanities.

Administrative Details

Template Author(s): Qian Zhang, University of Waterloo, National Digital Research Infrastructure Organization (NDRIO)

Published: March 2, 2021

DOI: 10.5281/zenodo.4573539

Contact: Portage Network - portage@carl-abrc.ca, portagenetwork.ca

Version:

Version	Date	Changes
1.0	2021-03-02	Formatted for inaugural publication.

Phase 1: Data Preparation

Data Collection

What types of data, metadata, and scripts will you collect, create, link to, acquire, record, or generate through the proposed research project?

Examples of data types may include text, numeric (ASCII, binary), images, audio, video, tabular data, spatial data, experimental, observational, and simulation/modelling data, instrumentation data, codes, software and algorithms, and any other materials that may be produced in the course of the project.

Data Organization

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

Proprietary file formats that require specialized software or hardware are not recommended, but may be necessary for certain data collection or analysis methods. Using open file formats or industry-standard formats (e.g. those widely used by a given community) is preferred whenever possible. Read more about recommended file formats at UBC Library or UK Data Service.

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?

It is important to keep track of different copies and versions of files, files held in different formats or locations, and any information cross-referenced between files.

Logical file structures, informative naming conventions, and clear indications of file versions all contribute to better use of your data during and after your research project. These practices will help ensure that you and your research team are using the appropriate version of your data, and will minimize confusion regarding copies on different computers, on different media, in different formats, and/or in different locations. Read more about file naming and version control at UBC Library or UK Data Service.

Phase 2: Active Research (Data) Management

Documentation and Metadata

What information will be needed for the data to be read and interpreted correctly?

Typically, good documentation includes high-level information about the study as well as data-level descriptions of the content. It may also include other contextual information required to make the data usable by other researchers, such as: your research methodology, definitions of variables, vocabularies, classification systems, units of measurement, assumptions made, formats and file types of the data, a description of the data capture and collection methods, provenance of various data sources (original source of data, and how the data have been transformed), explanation of data analysis performed (including syntax files), the associated script(s), and annotation of relevant software.

Some types of documentation typically provided for research data and software include:

- README file, codebook, or data dictionary
- Electronic lab notebooks such as Jupyter Notebook

If you are using a metadata standard and/or tools to document and describe your data, please list them here. Explain the rationale for the selection of these standards.

There are many general and domain-specific metadata standards that can be used to manage research data. These machine-readable, openly-accessible standards are often based on language-independent data formats such as XML, RDF, and JSON, which enables the effective exchange of information between users and systems. Existing, accepted community standards should be used wherever possible, including when recording intermediate results.

Where community standards are absent or inadequate, this should be documented along with any proposed solutions or remedies. You may wish to use this DMP Template to propose alternate strategies that will facilitate metadata interoperability in your field.

Example: There are a wide variety of metadata standards available to choose from, and you can learn more about these options at UK Digital Curation Centre's Disciplinary Metadata, FAIRsharing standards, RDA Metadata Standards Directory, Seeing Standards: A Visualization of the Metadata Universe.

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

Consider how you will capture information during the project and where it will be recorded to ensure the accuracy, consistency, and completeness of your documentation. Often, resources you've already created can contribute to this (e.g. publications, websites, progress reports, etc.). It is useful to consult regularly with members of the research team to capture potential changes in data collection or processing that need to be reflected in the documentation. Individual roles and workflows should include gathering, creating or maintaining data documentation as a key element.

Advanced Research Computing (ARC)-Related Facilities and Other Resources

Please identify the facilities to be used (laboratory, computer, office, clinical and other) and/or list the organizational resources available to perform the proposed research. If appropriate, indicate the capacity, pertinent capabilities, relative proximity and extent of availability of the resources to the research project.

ARC resources usually contain both computational resources and data storage resources. Please describe only those ARC resources that are directly applicable to the proposed work. Include existing resources and any external resources that may be made available.

What will be the primary production computing platform(s) (e.g., compute clusters, virtual clusters)?

What are the technical details of each of the computational resources?

It is important to document the technical details of all the computational and data storage resources, and associated systems and software environments you plan to use to perform the simulations and analysis proposed in this research project.

You may wish to provide the following information:

- Startup allocation limit
- System architecture: System component and configuration
    - CPU nodes
    - GPU nodes
    - Large memory nodes
- Performance: e.g., FLOPs, benchmark
- Associated systems software environment
- Supported application software
- Data transfer
- Storage

What large-scale data analysis framework (and associated technical specifications) will be used?

Examples of data analysis frameworks include:

- Hadoop
- Spark

What software tools will be utilized and/or developed for the proposed research?

Examples of software tools include:

- High-performance compilers, debuggers, analyzers, editors
- Locally developed custom libraries and application packages for software development

What metadata/documentation do you need to provide for others to use your software?

(Re)using code/software requires, at minimum, information about both the environment and expected input/output. Log all parameter values, including when setting random seeds to predetermined values, and make note of the requirements of the computational environment (software dependencies, etc.) Track your software development with versioning control systems, such as GitHub Bitbucket, GitLab, etc.

If your research and/or software are built upon others' software code, it is good practice to acknowledge and cite the software you use in the same fashion as you cite papers to both identify the software and to give credit to its developers.

For more information on proper software documentation and citation practices, see: Ten simple rules for documenting scientific software and Software Citation Principles.

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?

Storage-space estimates should take into account requirements for file versioning, backups, and growth over time, particularly if you are collecting data over a long period (e.g. several months or years). Similarly, a long-term storage plan is necessary if you intend to retain your data after the research project.

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

Data may be stored using optical or magnetic media, which can be removable (e.g. DVD and USB drives), fixed (e.g. desktop or laptop hard drives), or networked (e.g. networked drives or cloud-based servers). Each storage method has benefits and drawbacks that should be considered when determining the most appropriate solution.

The risk of losing data due to human error, natural disasters, or other mishaps can be mitigated by following the 3-2-1 backup rule: Have at least three copies of your data; store the copies on two different media; keep one backup copy offsite. Further information on storage and backup practices is available from the University of Sheffield Library and the UK Data Service.

What are the technical details of each of the storage and file systems you will use during the active management of the research project?

Technical detail example:

- Quota

Examples of systems:

- High-performance archival/storage storage
- Database
- Web server
- Data transfer
- Cloud platforms, such as Amazon Web Services, Microsoft Azure, Open Science Framework (OSF), Dropbox, Box, Google Drive, One Drive

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

An ideal solution is one that facilitates cooperation and ensures data security, yet is able to be adopted by users with minimal training. Transmitting data between locations or within research teams can be challenging for data management infrastructure. Relying on email for data transfer is not a robust or secure solution.

Third-party commercial file sharing services (such as Google Drive and Dropbox) facilitate file exchange, but they are not necessarily permanent or secure, and the servers are often located outside Canada.

What do you estimate the overall cost of managing your data will be?

This estimate should incorporate data management costs incurred during the project as well as those required for ongoing support after the project is finished. Consider costs associated with data purchase, data curation, and providing long-term access to the data. For ARC projects, charges for computing time, also called Service Units (SU), and the cost of specialized or proprietary software should also be taken into consideration.

Some funding agencies state explicitly that they will provide support to meet the cost of preparing data for deposit in a repository. These costs could include: technical aspects of data management, training requirements, file storage & backup, etc. OpenAIRE has a useful tool for estimating costs for RDM.

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.

Your data management plan has identified important data activities in your project. Identify who will be responsible -- individuals or organizations -- for carrying out these parts of your data management plan. This could also include the time frame associated with these staff responsibilities and any training needed to prepare staff for these duties.

Ensure Portability and Reproducibility of Results

What will you do to ensure portability and reproducibility of your results?

A computationally reproducible research package will include:

- Primary data (and documentation) collected and used in analysis
- Secondary data (and documentation) collected and used in analysis
- Primary data output result(s) (and documentation) produced by analysis
- Secondary data output result(s) (and documentation) produced by analysis
- Software program(s) (and documentation) for computing published results
- Dependencies for software program(s) for replicating published results
- Research Software documentation and implementation details
- Computational research workflow and provenance information
- Published article(s)

All information above should be accessible to both designated users and reusers.

(Re)using code/software requires knowledge of two main aspects at minimum: environment and expected input/output. With sufficient information provided, computational results can be reproduced. Sometimes, a minimum working example will be helpful.

Example: Container solutions, such as docker and singularity, can replicate the exact computational environment for others to run. For more information, these Ten Simple Rules for Writing Dockerfiles for Reproducible Data Science and Ten Simple Rules for Reproducible Computational Research may be helpful.

Phase 3: Data Protection

Ethics and Legal Compliance

Will your proposed research include any sensitive, private, confidential, or other legally protected information or data?

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

Consider where, how, and to whom sensitive data with acknowledged long-term value should be made available, and how long it should be archived. Decisions should align with your institutional Research Ethics Board requirements.

Methods used to share data will be dependent on the type, size, complexity and degree of sensitivity of data. For instance, sensitive data should never be shared via email or cloud storage services such as Dropbox. Outline any problems anticipated in sharing data, along with causes and possible measures to mitigate these. Problems may include: confidentiality, lack of consent agreements, or concerns about Intellectual Property Rights, among others.

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

Obtaining the appropriate consent from research participants is an important step in assuring your Research Ethics Board that data may be shared with researchers outside of your project. The consent statement may identify certain conditions clarifying the uses of the data by other researchers. For example, it may stipulate that the data will only be shared for non-profit research purposes or that the data will not be linked with personally identified data from other sources. Read more about data security: UK Data Service.

You may need to anonymize or de-identify your data before you can share it. Read more about these processes at UBC Library, UK Data Service, or Image Data Sharing for Biomedical Research—Meeting HIPAA Requirements for De-identification.

Under what licence do you plan to release your data?

Licenses stipulate how your data may be used. Funding agencies and/or data repositories may have end-user license requirements in place; if not, they may still be able to guide you in the selection of a license. Once selected, please include a copy of your end-user license with your Data Management Plan. Note that only the intellectual property rights holder(s) can issue a license, so it is crucial to clarify who owns those rights.

Example: There are several types of standard licenses available to researchers, such as the Creative Commons licenses and the Open Data Commons licenses. For most datasets it is easier to use a standard license rather than to devise a custom-made one. Even if you choose to make your data part of the public domain, it is preferable to make this explicit by using a license such as Creative Commons' CC0. More about data licensing: Digital Curation Centre.

Under what licence do you plan to release your software?

By providing a licence for your software, you grant others certain freedoms, and define what they are allowed to do with your code. Free and open software licences typically allow someone else to use, study, improve and share your code. You can licence all the software you write, including scripts and macros you develop on proprietary platforms. For more information, see Choose an Open Source License or open source licenses options at the Open Source Initiative.

Please be aware that software is typically protected by copyright that is often held by the institution rather than the developer. Ensure you understand what rights you have to share your software before choosing a license.

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

Before you copy, (re-)use, modify, build on, or (re-)distribute others' data and code, or engage in the production of derivatives, be sure to check, read, understand and follow any legal licensing agreements. The actions you can take, including whether you can publish or redistribute derivative research products, may depend on terms of the original license.

If your research data and/or software are built upon others' data and software publications, it is good practice to acknowledge and cite the corresponding data and software you use in the same fashion as you cite papers to both identify the software and to give credit to its developers. Some good resources for developing citations are the Software Citation Principles (Smith et al., 2016), DataCite - Cite Your Data, and Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data.

Compliance with privacy legislation and laws that may restrict the sharing of some data should be discussed with your institution's privacy officer or data librarian, if possible. Research Ethics Boards are also central to the research process and a valuable resource. Include in your documentation a description concerning ownership, licensing, and intellectual property rights of the data. Terms of reuse must be clearly stated, in line with the relevant legal and ethical requirements where applicable (e.g., subject consent, permissions, restrictions, etc.).

Phase 4: Sharing and Preserving

Sharing and Preservation

What will be the potential impact of the data within the immediate field and in other fields, and any broader societal impact?

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

One of the best ways to refer other researchers to your deposited datasets is to cite them the same way you cite other types of publications. The Digital Curation Centre provides a detailed guide on data citation. You may also wish to consult the DataCite citation recommendations.

Some repositories also create links from datasets to their associated papers, increasing the visibility of the publications. If possible, cross-reference or link out to all publications, code and data. Choose a repository that will assign a persistent identifier (such as a DOI) to your dataset to ensure stable access to the dataset.

Other sharing possibilities include: data registries, indexes, word-of-mouth, and publications. For more information, see Key Elements to Consider in Preparing a Data Sharing Plan Under NIH Extramural Support.

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

Consider which data are necessary to validate (support or verify) your research findings, and which must be shared to meet institutional or funding requirements. This may include data and code used in analyses or to create charts, figures, images, etc. Certain data may need to be restricted because of confidentiality, privacy, or intellectual property considerations and should be described below.

Wherever possible, share your data in preservation-friendly file formats. Some data formats are optimal for the long-term preservation of data. For example, non-proprietary file formats, such as text ('.txt') and comma-separated ('.csv'), are considered preservation-friendly. The UK Data Service provides a useful table of file formats for various types of data. Keep in mind that preservation-friendly files converted from one format to another may lose information (e.g. converting from an uncompressed TIFF file to a compressed JPG file), so changes to file formats should be documented.

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

Data retention should be considered early in the research lifecycle. Data-retention decisions can be driven by external policies (e.g. funding agencies, journal publishers), or by an understanding of the enduring value of a given set of data. The need to preserve data in the short-term (i.e. for peer-verification purposes) or long-term (for data of lasting value), will influence the choice of data repository or archive. A helpful analogy is to think of creating a 'living will' for the data, that is, a plan describing how future researchers will have continued access to the data.

It is important to verify whether or not the data repository you have selected will support the terms of use or licenses you wish to apply to your data and code. Consult the repository's own terms of use and preservation policies for more information. For help finding an appropriate repository, contact your institution's library or reach out to the Portage DMP Coordinator at support@portagenetwork.ca.

Example: The general-purpose repositories for data sharing in Canada are the Federated Research Data Repository (FRDR) and Scholars Portal Dataverse. You can search for discipline-specific repositories on re3data.org or by using DataCite's Repository Finder tool.

What software code will you make available, and where?

Making the software (or source code) you developed accessible is essential for others to understand your work. It allows others to check for errors in the software, to reproduce your work, and ultimately, to build upon your work. Consider using a code-sharing platform such as GitHub Bitbucket, or GitLab. If you would like to archive your code and receive a DOI, GitHub is integrated with Zenodo as a repository option.

At a minimum, if using third-party software (proprietary or otherwise), researchers should share and make available the source code (e.g., analysis scripts) used for analysis (even if they do not have the intellectual property rights to share the software platform or application itself).

Describe your software sustainability plan.

After the software has been delivered, used and recognized by a sufficiently large group of users, will you allocate both human and financial resources to support the regular maintenance of the software, for activities such as debugging, continuous improvement, documentation and training?