Data Management Plan Template: **Water Quality Research**

Abstract

This template provides guidance to researchers who are collecting or generating water quality data. Highly Qualified Personnel (HQP) and students are encouraged to consult with their advisors or principal investigators while completing this template.

Administrative Details

Template Author(s): Bhaleka Persaud, Water Institute, University of Waterloo; Krysha Dukacz, Global Water Futures Programme, McMaster University; Patrick LeClair, Lindsay Day, The Gordon Foundation; Laleh Moradi, Amber Peterson, Global Institute for Water Security and Global Water Futures Programme, University of Saskatchewan; Gopal Chandra Saha, Wilfrid Laurier University

Support: Global Water Futures Programme, Water Institute, The Gordon Foundation, Global Institute for Water Security, University of Waterloo, McMaster University, University of Saskatchewan, Wilfrid Laurier University

Published: April 16, 2021

DOI: 10.5281/zenodo.4697621

Contact: Portage Network - portage@engagedri.ca, portagenetwork.ca

Version:

Version	Date	Changes
1.0	2021-04-09	Formatted for inaugural publication.
1.1	2021-04-16	Attribution adjusted.

Data Collection

Why are you collecting or generating your data?

Describe the purpose or goal of this project. Is the data collection for a specific study or part of a long-term collection effort? What is the relationship between the data you are collecting and any existing data? Note existing data structure and procedures if building on previous work.

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

Types of data you may create or capture could include: geospatial layers (shapefiles; may include observations, models, remote sensing, etc.); tabular observational data; field, laboratory, or experimental data; numerical model input data and outputs from numerical models; images; photographs; video.

Please also describe the tools and methods that you will use to collect or generate the data. Outline the procedures that must be followed when using these tools to ensure consistent data collection or generation. If possible, include any sampling procedures or modelling techniques you will use to collect or generate your data.

Where are you collecting or generating your data (i.e., study area)? Include as appropriate the spatial boundaries, water source type and watershed name.

Try to use pre-existing collection standards, such as the CCME's Protocols for Water Quality Sampling in Canada, whenever possible.

If you will set up monitoring station(s) to continuously collect or sample water quality data, please review resources such as the World Meteorological Organization's technical report on Water Quality Monitoring and CCME's Protocols for Water Quality Guidelines in Canada.

Are you using third party data? If so, describe the source of the data including the owner, database or repository DOIs or accession numbers.

Include full references or links to the data when possible. Identify any license or use restrictions and ensure that you understand the policies for permitted use, redistribution and derived products.

Ethics and Legal Compliance

Does your project include sensitive data?

Sensitive data is data that carries some risk with its collection. Examples of sensitive data include:

- Data governed by third party agreements, contracts or legislation.
- Personal information, health related data, biological samples, etc.
- Indigenous knowledge, and data collected from Indigenous peoples, or Indigenous lands, water and ice.
- Data collected with an industry partner.
- Location information of species at risk.
- Data collected on private property, for example where identification of contaminated wells could impact property values or stigmatize residents or land owners.

Additional sensitivity assessment can be made using data classification matrices such as the University of Saskatchewan Data Classification guidance.

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

Methods used to share data will be dependent on the type, size, complexity and degree of sensitivity of data. 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.

Decisions should align with Research Ethics Board requirements. If you are collecting water quality data from Indigenous communities, please also review resources such as the Tri-Council Policy Statement (TCPS2) - Chapter 9: Research Involving the First Nations, Inuit and Métis Peoples of Canada, the First Nations Principles of OCAP, the CARE Principles of Indigenous Data Governance, the National Inuit Strategy on Research (NISR), and Negotiating Research Relationships with Inuit Communities as appropriate.

Restrictions can be imposed by limiting physical access to storage devices, placing data on computers with no access to the Internet, through password protection, and by encrypting files. Sensitive data should never be shared via email or cloud storage services such as Dropbox. Read more about data security here: UK Data Service.

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

Consider where, how, and to whom sensitive data with acknowledged long-term value should be made available, and for how long it should be archived. If you must restrict some data from sharing, consider making the metadata (information about the dataset) available in a public metadata catalogue.

Obtaining the appropriate consent from research participants is an important step in assuring Research Ethics Boards that the data may be shared with researchers outside your project. The consent statement may identify certain conditions clarifying the uses of your 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. It is important to consider how the data you are collecting may contribute to future research prior to obtaining research ethics approval since consent will dictate how the data can be used in the immediate study and in perpetuity.

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

Compliance with privacy legislation and laws that may impose content restrictions in the data should be discussed with your institution's privacy officer or research services office. Research Ethics Boards are also central to the research process.

Describe ownership, licensing, and any intellectual property rights associated with the data. Check your institution's policies for additional guidance in these areas. The University of Waterloo has a good example of an institutional policy on Intellectual Property Rights.

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.).

File Management

What file formats will your data be in? Will these formats allow for data re-use, sharing and long-term access to the data? If not, how will you convert these into interoperable formats?

Data should be collected and stored using machine readable, non-proprietary formats, such as .csv, .json, or .tiff. Proprietary file formats requiring specialized software or hardware to use are not recommended, but may be necessary for certain data collection or instrument analysis methods. If a proprietary format *must* be used, can it be converted to an open format or accessed using free and open source tools?

Using open file formats or industry-standard formats (e.g. those widely used by a given community) is preferred whenever possible. Read more about file formats: UBC Library, USGS, DataONE, 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?

File names should:

- Clearly identify the name of the project (Unique identifier, Project Abbreviation), timeline (use the ISO date standard) and type of data in the file;
- Avoid use of special characters, such as $ % ^ & # | :, to prevent errors and use an underscore (_) or dash (-) rather than spaces;
- Be as concise as possible. Some instruments may limit you to specific characters. Check with your labs for any naming Standard Operating Procedures (SOPs) or requirements.

Data Structure:

- It is important to keep track of different copies or versions of files, files held in different formats or locations, and information cross-referenced between files. This process is called 'version control'. For versioning, the format of vMajor.Minor.Patch should be used (i.e. v1.1.0);
- 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 minimize confusion regarding copies on different computers and/or on different media.

Read more about file naming and version control: UBC Library or UK Data Service.

Documentation and Metadata

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

Typically, good documentation includes information about the study, data-level descriptions and any other contextual information required to make the data usable by other researchers. Elements to document, as applicable, include: research methodology, variable definitions, vocabularies, classification systems, units of measurement, assumptions made, format and file type of the data, and details of who has worked on the project and performed each task, etc.

A readme file describing your formatting, naming conventions and procedures can be used to promote use and facilitate adherence to data policies. For instance, describe the names or naming process used for your study sites.

Verify the spelling of study site names using the Canadian Geographical Names Database.

If your data will be collected on Indigenous lands, ensure your naming scheme follows the naming conventions determined by the community.

How will you describe samples collected?

Include descriptions of sampling procedures and hardware or software used for data collection, including make, model and version where applicable. **Sample and replicate labels** should have a consistent format (sample number, name, field site, date of collection, analysis requested and preservatives added, if applicable) and a corresponding document with descriptions of any codes or short forms used.

For examples and guidelines, see the CCME Protocols Manual for Water Quality Sampling in Canada (taxonomy example p. 11, general guidelines p. 32-33).

Consistency, relevance and cost-efficiency are key factors of sample collection and depend on the scope of the project. For practical considerations, see “4.3 Step 3. Optimizing Data Collection and Data Quality” in the CCME Guidance Manual for Optimizing Water Quality Monitoring Program Design.

How will you analyze and interpret the water quality data?

It is important to have a standardized data analysis procedure and metadata detailing both the collection and analysis of the data. For guidance see “4.4 Step 4. Data Analysis, Interpretation and Evaluation” in the CCME Guidance Manual for Optimizing Water Quality Monitoring Program Design.

If you will collect or analyze data as part of a wider program, such as the Canadian Aquatic Biomonitoring Network, include links to the appropriate guidance documents.

What kind of Quality Assurance/Quality Control procedures are you planning to do?

Include documentation about what QA/QC procedures will be performed. For guidance see “1.3 QUALITY ASSURANCE/CONTROL IN SAMPLING” in the CCME Integrated guidance manual of sampling protocols for water quality monitoring in Canada or the Quality-Control Design for Surface-Water Sampling in the National Water-Quality Network from the USGS.

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

Consider how you will capture this information and where it will be recorded, ideally in advance of data collection and analysis, to ensure accuracy, consistency, and completeness of the documentation. Writing guidelines or instructions for the documentation process will enhance adoption and consistency among contributors. Often, resources you have already created can contribute to this (e.g., laboratory Standard Operating Procedures (SOPs), recommended textbooks, publications, websites, progress reports, etc.).

It is useful to consult regularly with the 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 data documentation as a key element. Researchers should audit their documentation at a specific time interval (e.g., bi-weekly) to ensure documentation is created properly and information is captured consistently throughout the project.

List any metadata standard(s) and/or tools you will use to document and describe your data:

Metadata describes a dataset and provides vital information such as owner, description, keywords, etc. that allow data to be shared and discovered effectively. Researchers are encouraged to adopt commonly used and interoperable metadata schemas (general or domain-specific), which focus on the exchange of data via open spatial standards. Dataset documentation should be provided in a standard, machine readable, openly-accessible format to enable the effective exchange of information between users and systems.

Examples:

Water Quality metadata standards:

- DS-WQX: A schema designed for data entered into the DataStream repository based off of the US EPA's WQX standard.
- WQX: Data model designed by the US EPA and USGS for upload to their water quality exchange portal.

Ecological metadata standards:

- Ecological Metadata Language (EML): A comprehensive vocabulary and a readable XML markup syntax for documenting research data.

Geographic metadata standards:

- ISO 19115: International Standards Organisation's schema for describing geographic information and services.

Read more about metadata standards: UK Digital Curation Centre's Disciplinary Metadata.

If the metadata standard will be modified, please explain how you will modify the standard to meet your needs.

Deviation from existing metadata standards should only occur when necessary. If this is the case, please document these deviations so that others can recreate your process.

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

Once a standard has been chosen, it is important that data collectors have the necessary tools to properly create or capture the metadata. Audits of collected metadata should occur at specific time intervals (e.g., bi-weekly) to ensure metadata is created properly and captured consistently throughout the project.

Some tips for ensuring good metadata collection are:

- Provide support documentation and routine metadata training to data collectors.
- Provide data collectors with thorough data collection tools (e.g., field or lab sheets) so they are able to capture the necessary information.
- Samples or notes recorded in a field or lab book should be scanned or photographed daily to prevent lost data.

Storage and Backup

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. 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?

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. 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 such as Compute Canada). Each storage method has benefits and drawbacks that should be considered when determining the most appropriate solution.

Raw data should be preserved and never altered. Some options for preserving raw data are storing on a read-only drive or archiving the raw, unprocessed data. The preservation of raw data should be included in the data collection process and backup procedures.

Examples of further information on storage and backup practices are available from the University of Toronto and the UK Data Service.

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

An ideal shared data management solution facilitates collaboration, ensures data security and is easily adopted by users with minimal training. Tools such as the Globus file transfer system, currently in use by many academic institutions, allows data to be securely transmitted between researchers or to centralized project data storage. 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 are often located outside Canada. Additional resources such as Open Science Framework and Compute Canada are also recommended options for collaborations.

If your data will be collected on Indigenous lands, how will you share data with community members throughout the project? Please contact librarians at your institution to determine if there is support available to develop the best solution for your research project.

Responsibilities and Resources

Who will be responsible for managing this project's data during and after the project, and for what major data management tasks will they be responsible?

Describe the roles and responsibilities of all parties with respect to the management of the data. Consider the following:

- If there are multiple investigators involved, what are the data management responsibilities of each investigator?
- If data will be collected by students, clarify the student role versus the principal investigator role, and identify who will hold theIntellectual Property rights.
- Will training be required to perform the data collection or data management tasks? If so, how will this training be administered and recorded?
- Who will be the primary person responsible for ensuring compliance with the Data Management Plan during all stages of the data lifecycle?
- Include the time frame associated with these staff responsibilities and any training needed to prepare staff for these duties.

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?

Indicate a succession strategy for these data in the event that one or more people responsible for the data leaves (e.g., a student leaving after graduation). Describe the process to be followed in the event that the Principal Investigator leaves the project. In some instances, a co-investigator or the department or division overseeing this research will assume responsibility.

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

This estimate should incorporate data management costs incurred during the project and those required for longer-term support for the data when the project is finished, such as the cost of preparing your data for deposit and repository fees. Some funding agencies state explicitly the support that they will provide to meet the cost of preparing data for deposit. This might include technical aspects of data management, training requirements, file storage & backup and contributions of non-project staff. Can you leverage existing resources, such as Compute Canada Resources, University Library Data Services, etc. to support implementation of your DMP?

To help assess costs, OpenAIRE's 'estimating costs RDM tool' may be useful.

Sharing, Reuse and Preservation

In general, data collected using public funds should be preserved for future discovery and reuse. As you develop your data sharing strategy you will want to consider the following:

Do you, your institution or collaborators have an existing data sharing strategy?

Use all or parts of existing strategies to meet your requirements.

Are there restrictions on sharing due to ethics or legal constraints?

In these instances, it is critical to assess whether data can or should be shared. It is necessary to comply with:

- Data treatment protocols established by the Research Ethics Board (REB) process, including data collection consent, privacy considerations, and potential expectations of data destruction;
- Data-sharing agreements or contracts. Data that you source or derive from a third party may only be shared in accordance with the original data sharing agreements or licenses;
- Any relevant legislation.

Note: If raw or identifiable data cannot be shared, is it possible to share aggregated data, or to de-identify your data, or coarsen location information for sharing? If data cannot be shared, consider publishing descriptive metadata (data about the data), documentation and contact information that will allow others to discover your work.

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

Think about what data needs to be shared to meet institutional or funding requirements, and what data may be restricted because of confidentiality, privacy, or intellectual property considerations.

- **Raw data** are unprocessed data directly obtained from sampling, field instruments, laboratory instruments, modelling, simulation, or survey. Sharing raw data is valuable because it enables researchers to evaluate new processing techniques and analyse disparate datasets in the same way.
- **Processed data** results from some manipulation of the raw data in order to eliminate errors or outliers, to prepare the data for analysis or preservation, to derive new variables, or to de-identify the sampling locations. Processing steps need to be well described.
- **Analyzed data** are the results of qualitative, statistical, or mathematical analysis of the processed data. They may be presented as graphs, charts or statistical tables.
- **Final data** are processed data that have undergone a review process to ensure data quality and, if needed, have been converted into a preservation-friendly format.

Will you deposit your data for long-term preservation and access at the end of your research project? Please indicate any repositories that you will use.

Data repositories help maintain scientific data over time and support data discovery, reuse, citation, and quality. Researchers are encouraged to deposit data in leading "domain-specific" repositories, especially those that are FAIR-aligned, whenever possible.

Domain-specific repositories for water quality data include:

- DataStream: A free, open-access platform for storing, visualizing, and sharing water quality data in Canada.
- Canadian Aquatic Biomonitoring Network (CABIN): A national biomonitoring program developed by Environment and Climate Change Canada that provides a standardized sampling protocol and a recommended assessment approach for assessing aquatic ecosystem condition.

General Repositories:

- Federated Research Data Repository: A scalable federated platform for digital research data management (RDM) and discovery.
- Institution-specific Dataverse: Please contact your institution's library to see if this is a possibility.

Other resources:

- The Repository Finder Tool developed by DataCite. This tool queries the re3data registry of research data repositories.
- Coalition for Publishing Data in the Earth and Space Sciences: Enabling FAIR Data – FAQs.

What steps will you take to ensure your data is prepared for preservation?

Consider using preservation-friendly file formats. For example, non-proprietary formats, such as text (.txt) and comma-separated (.csv), are considered preservation-friendly. For guidance, please see UBC Library, USGS, DataONE, or UK Data Service. Keep in mind that files converted from one format to another may lose information (e.g., converting from an uncompressed TIFF file to a compressed JPG file degrades image quality), so changes to file formats should be documented.

Some data you collect may be deemed sensitive and require unique preservation techniques, including anonymization. Be sure to note what this data is and how you will preserve it to ensure it is used appropriately. Read more about anonymization: UBC Library or UK Data Service.

What type of end-user license will you use for your data?

Licenses dictate how your data can be used. Funding agencies and data repositories may have end-user license requirements in place; if not, they may still be able to guide you in choosing or developing an appropriate license. Once determined, please consider including 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.

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: UK DCC.

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

Possibilities include: data registries, repositories, indexes, word-of-mouth, publications. How will the data be accessed (Web service, ftp, etc.)? 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. Some repositories also create links from datasets to their associated papers, increasing the visibility of the publications. Read more at the National Institutes of Health's Key Elements to Consider in Preparing a Data Sharing Plan Under NIH Extramural Support.

Contact librarians at your institution for assistance in making your dataset visible and easily accessible, or reach out to the Portage DMP Coordinator at support@portagenetwork.ca.