Data Management Procedures
Using the NPRB Research Workspace

Prepared by Axiom Data Science
on behalf of NPRB
August 2017

1. Purpose

The North Pacific Research Board (NPRB)'s vision is to build a clear understanding of the Gulf of Alaska, Bering Sea/Aleutian Islands, and Chukchi/Beaufort Seas that enables effective management and sustainable use of marine resources. The data generated by NPRB funded projects are key assets to realizing this vision that must be managed correctly to support management decision-making, scientific integrity, and enhanced information availability. This document provides a data management framework for the NPRB Core Program with defined roles and responsibilities, and procedures for the collection, quality, storage, maintenance, and dissemination of project data. Procedures may be followed at any time during the preparation of your dataset, but are most useful when considered at the onset of project planning and implemented during data collection. The intent of this framework is to improve the accessibility and long-term usability of NPRB funded data.

2. NPRB Metadata and Data Policy

The NPRB metadata and data transfer policy is as follows:

a. For projects involving data collection or generation, NPRB requires the transfer of all data and its associated metadata ,
b. For any third party datasets used in the NPRB funded project, only the transfer of the metadata associated with the third party data is required,
c. If third party data is modified for use in the NPRB funded project, the metadata associated with the third party data is required in addition to the modified dataset and the associated metadata,
d. For modeling projects transfer of data inputs, modeling code and modeling output to NPRB is required along with associated metadata.

Any questions regarding metadata and data transfer should be directed to the NPRB Program Manager. Metadata and data records must be submitted with the final programmatic report.

The metadata from all NPRB Core Program funded projects are made available to the public generally within six months after the project is completed. Data files are also archived with NPRB when projects are complete, but do not become public for two years (researchers may make special requests for a longer archive period).

3. Definitions

For the purposes of this document, the following definitions apply:

Data: Data are distinct units of information, such as numbers, model code base, research outputs, usually formatted in a specific way stored within a database or file and suitable for processing by a computer.

Data storage: Data files uploaded and stored in the Research Workspace, a secure, web-based scientific collaboration and data management tool used to centralize and share NPRB project information. Data files are stored for long-term access exclusively by

NPRB and collaborators on your project.

Data transfer: Data files that are submitted to NPRB either by upload to the Research Workspace or other means, such as email or physical copy (e.g. hard drive).

Data shared with the public: Final project data and data products made available publicly through the NPRB Project Search catalog following a two year embargo. The Project Search & Database tool provides the public with scientific information about funded projects in the Core Program dating back to 2002.

Archived: Final data and data products are stored in a data repository for long-term preservation. At the time of writing, the Research Workspace supports an automated pathway to submit NPRB project data to the Research Workspace Member Node in the DataONE network where it is discoverable and accessible through the DataONE Search catalog. The final data package is assigned a digital object identifier (DOI) or accession number to facilitate the citation of project data.

4. Roles and Responsibilities

North Pacific Research Board- NPRB is responsible for managing the NPRB Core Program projects and for making policies related to project data. They also oversee the submission and maintenance of data, and the dissemination of final data and data products through the NPRB Project Search catalog and long-term data repositories.

Axiom Data Science- Axiom is contracted to the NPRB Core Program for data standards and management services. Axiom responds to requests by NPRB and is the primary contact for PIs relative to data structure and metadata documentation. Axiom has a responsibility to ensure that final project data and metadata content meet quality and format standards for long-term preservation. Axiom also maintains the NPRB Research Workspace for project data storage and transfer to DataONE for archiving, and the NPRB Project Search catalog for public data access. Axiom will respond to PI requests for technical assistance using these tools.

Project Principal Investigators (PIs)- The project PIs are responsible for data collection, analysis, metadata generation, and the delivery of final data, data products, and metadata to NPRB according to the project MOU. The PIs are also responsible for ensuring data quality, completeness, and integrity relative to the scientific standard within the respective project discipline throughout data creation and maintenance.

Data Users- Data users are primarily NPRB stakeholders, including the marine science community. Data users access data content using the NPRB Project Search catalog and may reuse any publicly available data without constraint, including for study replication, statistical analyses, data mining, proposal development, and data reporting.

5. NPRB Core Projects Data Management Procedures

All projects shall follow the procedures listed below for managing and submitting final data and data products. Further, it is recommended to follow the common data management guidelines for the Research Workspace. The rationale for these procedures is to save PIs time at project close-out by ensuring data is well-organized and documented and, ultimately, to increase access by the broader community to research outputs.

5.1 Data Management Plan

Data management plans are the best way to ensure that your data are well-organized, managed, and prepared for preservation into the future. Completing a data management form is highly recommended to document the planned research effort, the expected outputs, and the plan for documenting and archiving your data. This form will be reviewed by Axiom Data Science to identify any potential problems that could be obstacles to long-term preservation and sharing. Axiom will communicate with the PI about any highlighted issues allowing for consideration of solutions.

For projects involving model or analytical workflows, at the request of NPRB, Axiom may contact the PIs early in the project to understand the nature of the data and recommend an appropriate data management structure. This information would be formalized into a data management plan by Axiom and shared with the project PI to help define the final data, data products, and documentation that should be generated.

5.2 Data Storage & Security

In recent years, NPRB has updated many of its program guidelines and protocols. A major component of this evolution is the migration of Core Program projects to an online portal, called the NPRB Research Workspace. The Research Workspace is a web-based scientific collaboration and data management tool used to secure and centralize project information, generate metadata, and ultimately select final data files to be published openly in the NPRB Project Search catalog and national data repositories. The Workspace enables PIs to capture and retain the entire legacy of their project, while providing NPRB with a real-time and transparent view of project status.

All final project data and metadata shall be submitted using the Research Workspace: https://workspace.nprb.org/login. PIs are required to log into the Workspace and create a user account. User accounts are associated with a secure project within the Workspace that is maintained behind a password-protected firewall and is accessible only to your project collaborators. Project information is automatically populated into the Workspace from the NPRB proposal submission system, including name and contact information, project title, NPRB project number, project abstract, purpose, and keywords. The PIs and co-PIs listed in the project proposal are given access to the project, though additional collaborators can be added to your project in the Workspace upon request to the NPRB Program Manager.

5.3 Data Submission

Under the Core Program MOU, the PI(s) agree to transfer all data and metadata to NPRB at the completion of the project using the Research Workspace. Final data and metadata shall be submitted prior to the final report.

While the minimum requirement is for final data to be stored in the Workspace, PIs are encouraged to use the Workspace throughout the life of the project. The Workspace is an effective tool to help PIs centralize project information, securely store data files, and share project data with collaborators.

Detailed information on how to use the Workspace can be found in the help documentation.

5.4 Data Organization

Project information stored in the NPRB Research Workspace shall adhere to the below data guidelines. Refer to Axiom's documentation for more detailed recommendations about best practices for managing your data.

Folder Structure

Folders are important for breaking down project files into smaller, easier-to-manage and identifiable units. Your project in the Research Workspace contains three pre-existing folders for storing information: Admin, Data, and Media Products. Files related to these broad topics shall be stored in the respective folders.

PIs are encouraged to build upon the 'Data' folder structure to help you stay organized and easily retrieve your data files. Refer to the best practices documentation for guidance on creating and naming new folders.

Finalized files shall be stored in separate and clearly-marked folder(s) from intermediary or raw data files. This will expedite the final review and close-out for your project.

Folder and File Naming

How you name folders and files added to your projects will have an impact on you and your collaborator's ability to find and understand the project's data. Naming consistently and descriptively will help users identify records at a glance, and will help to facilitate the storage and retrieval of data. Final data files shall follow these naming guidelines to ensure they are consistently formatted and informative.

5.5 Data Formats

For data-based projects, final data files shall be stored in non-proprietary formats to help ensure they are useable, open, and readable into the future. Refer to the best practices documentation for data formats appropriate for long-term preservation.

For model-based projects, project PIs shall follow these recommendations found in Appendix A.

5.6 Data Quality

Beyond scientific quality assurance, basic quality reviews shall be performed to your data throughout its lifecycle, from collection through submission of final data files in the Research Workspace. The data quality procedures used during your project should be described in the metadata documentation to indicate to future users the quality and accuracy of your data. Refer to the best practices documentation for the data quality guidelines.

5.7 Prepare Metadata and Data Documentation

Metadata

Metadata are required for all final data files or data products submitted to NPRB. Metadata must be in a standards-compliant format suitable for long-term archive. The Research Workspace includes an integrated metadata editor to generate FGDC-endorsed ISO 19110 and 19115-2 standards metadata. Use of this editor is highly encouraged to ease the publication and archive of data through the NPRB Project Search catalog.

A new metadata record should be created using the NPRB Core Program template available through Workspace metadata editor. Follow these steps for how to copy a whole record from a metadata template. The name of the existing template record to be used in step 2 is NPRB Core Program template. New content about your project data can be added to your metadata record after the template is copied.

Depending on the project, more than one metadata record may be required to sufficiently describe the data. Metadata should be created at the final data folder to describe the file or files contained within the folder. The intent of folder metadata is to reduce the burden upon the creator for constructing the metadata and resource archive.

- Depending on the nature of the data, groups of files of the same format or sharing similar characteristics or methods can be documented by a single metadata record. Examples that require one metadata record include: single data collection methods resulting in one data file; a single instrument or sensor type (e.g. a glider); or single data collection methods repeated at more than one location resulting in one or more data files.

- If there are project datasets that contain distinct characteristics or were generated using different methods, then more than one metadata record should be created to describe each unique dataset. Examples that require more than one metadata record are projects organized by chapters; projects using more than one instrument or sensor type (e.g. moorings, gliders, etc); or projects having generated more than one distinct model or data product.

Refer to the Best Practices for Scientific Metadata document for guidance in assembling the final data package and creating scientific metadata using the metadata editor. This document provides field-by-field guidance on how to write high-quality metadata.

If you have questions about how to structure your metadata record(s) relative to your project data, please contact Axiom Data Science at metadata@axiomdatascience.com.

Data Documentation

Beyond standardized metadata, additional documentation about your dataset may be useful to further describe the actions taken to the data. Examples of data documentation include standard operations procedures, field notes, QA/QC manuals, and model readme files.

For model-based projects, refer to Appendix A for the additional data documentation that should be generated for your project.

5.8 Progress and Final Reports

All progress and final reports shall be submitted to NPRB using the integrated reporting tool in the Research Workspace. Refer to the NPRB report guidelines in the Workspace for report submission specifications.

5.9 Final Data and Metadata Quality Review

Once submitted, Axiom Data Science will perform reviews of the metadata record(s) to help ensure accuracy, consistency, and completeness of the metadata content. Any recommended edits or additions to the metadata will be communicated directly from Axiom Data Science to the PI. After which, Axiom will perform the final review of any modified metadata prior to approving the final data and metadata for project close-out.

5.10 NPRB Project Catalog and Archive

At the end of the two-year embargo period, final project data may be archived by NPRB into a preservation-oriented data center. At the time of writing, the Research Workspace supports an automated pathway to submit data to DataONE through its member node. In the future, data archive and/or replication from DataONE to NCEI may be supported.

At the end of the two-year period, submission of final data to the repository may be made by NPRB with technical assistance from Axiom Data Science. Prior to submission, the project PI would be notified about the pending submission. At which time, any updates to the dataset or metadata content may be made or requested by the PI, including: updating PI contact information; any new changes to the data content; and referencing project-related publications.

5.11 Maintenance and Updates to Project Data and Metadata

Publicly-available and archived metadata are living documents that need regular review and maintenance. Routine reviews to the technical metadata structure will be made by Axiom Data

Science. It is the responsibility of the PI to notify NPRB of any substantial changes to the dataset or metadata to ensure currency, accuracy, and completeness. Changes may include updating of the data contents, contact information, or publications. Axiom Data Science will work with NPRB to reflect these changes within the published or archived metadata records.

6. Technical Support

Project PIs are responsible for reading and adhering to the principles and guidelines written or referenced in this document. For additional questions on using the Research Workspace or creating metadata for your project, contact Axiom Data Science at metadata@axiomdatascience.com. Questions asked early in the project can save time and frustration when preparing your final dataset and metadata documentation!

6.1 Resolving Data Issues

Any user of publicly-available or archived data may question the accuracy of any data element. The user is responsible for helping to correct the problem by supplying as much detailed information as possible about the nature of the problem to NPRB. NPRB will respond to questions about the accuracy of data, and work with the project PI and/or Axiom Data Science, as necessary, to correct inconsistencies in the published or archived resource.

Data Management Procedures
Using the NPRB Research Workspace

Appendix A: Guidance for Modelers

1. Introduction

This appendix describes the guidelines for archiving data and creating metadata for model-based projects funded by NPRB. Many of the projects funded by NPRB could be described as “data-centric” in that they involve the collection of original data, either in the field or through laboratory analysis. However, some projects focus on the creation, refinement, or application of a model, rather than the collection and analysis of data. These model-based projects have different considerations than data-centric projects when it comes to deciding what work products to archive and how to describe those work products.

NPRB’s goals for model documentation include the following:

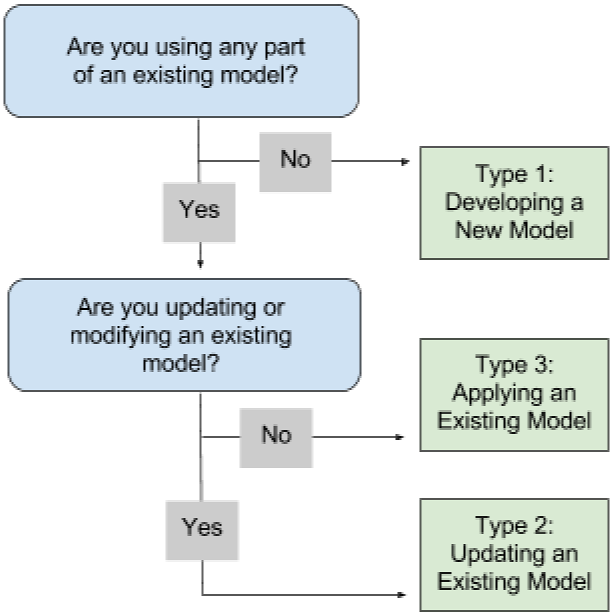
1. Models, model-based projects, and model derived products should have documentation sufficient to allow someone with the appropriate skills and knowledge to generate comparable results using a similar method.
2. Allow NPRB funded efforts and resulting research products need to be found and understood in the project catalog, and in other publicly available archives and metadata directories.
3. Make it easy for NPRB funded scientists to understand and meet NPRB expectations for project data management.

1.1 What is a Model?

Depending on the context, the term “model” can mean many different things. For the purposes of this document, we use “model” to mean a computational tool developed by a researcher (or a team of researchers) to address a particular problem or answer a particular question. In this context, creating a “model” involves the development of computer code that accepts parameters and inputs and returns a set of outputs. These outputs are often used to predict the future behavior of a natural system, such as how nutrients may move with ocean currents or how many salmon will return to a given stream. While this definition may not cover all model-based projects funded by NPRB, it does describe the most common cases. If you have questions about whether or not your project is a model-based project and subject to these guidelines, please contact the NPRB Core Program Manager to discuss.

2. Which Type of Modeler Are You?

By looking over past projects and engaging in discussion with modelers, Axiom Data Science has identified three general types of model-based projects that are typically funded by NPRB. The decision tree below will help you figure out which type of project you’re working on and which of the guidelines below will apply to you.


Are you using any part of an existing model?
No
Yes
Type 1: Developing a New Model
Are you updating or modifying an existing model?
Type 3: Applying an Existing Model
No
Yes
Type 2: Updating an Existing Model

2.1 Summary of Model Guidelines by Type

Model Metadata Matrix			
	Type 1 Standalone model	Type 2 Updated model	Type 3 Applied model
Inputs	●	●	●
Codebase	●	●	
Outputs			●

2.2 What If I'm None of the Above?

If you feel like your project doesn't clearly fit into one of the types described above, contact the NPRB Core Program Manager. Axiom Data Science and NPRB will work with you to figure out your archiving and documentation needs.

3. Guidance for Type 1 Projects (New Models) and Type 2 Projects (Updated Models)

If you're developing a new model or updating an existing model as the primary work product of your NPRB-funded project, the archiving and metadata requirements will be the same. Follow the guidelines below when you're ready to archive and document your projects.

3.1 Archiving requirements

The final product of Type 1 and Type 2 projects will include the model's inputs and codebase:

- Inputs may include data collected for the project, source datasets collected or curated by others, and any model parameters required to run the model.
- The model's codebase should include the actual model code, as well as any supplemental documentation needed to obtain, install, and run the model. If the model code will be maintained in a publicly-accessible repository such as Github, the code need not be also archived in the Research Workspace, but a link to the repository must be provided in the metadata record.

3.2 Metadata requirements

Most model-based projects will be described by one metadata record that covers your model inputs and codebase. This will typically be a folder-level record for the "Data" folder within your project listing in the Research Workspace. For large projects with multiple models, or other unique projects, more than one record may be required. If you don't know how many metadata records to generate for your project, contact the NPRB Program Manager to discuss.

In addition to the metadata fields that describe your model in a general way (e.g., keywords, spatial extent, constraints), several of the fields within your metadata record will need to deascribe your model inputs and codebase more specifically. The list below describes how those sections of your metadata record should be used to describe these components of your model:

- Resource Overview
    - Basic Overview
        - Title
            - Project in general
                - The title of your metadata record should include general information about the project, including its subject, location, and time period.
        - Abstract
            - Project in general
                - Your metadata Abstract should include general information about the project, such as its subject, location, and time period.
            - Inputs

-
  -
    -
      -
        - Additionally, your metadata Abstract should include a 1-2 sentence summary of any source data or inputs used by your model. If applicable, include the citation for the source data. If your inputs come from data collected as part of this project (i.e., funded by NPRB under this project number), note that here.
      - Codebase
        - Your metadata Abstract should also include a 1-2 sentence summary of your model code, including the computer language it's written in (including version) and information needed to install and run the code. Code files themselves should contain descriptive comments to help future users understand your model's steps. If applicable, include a link to the publicly-available repository (e.g., Github) where your model code can be obtained.
- Resource Content
  - Data Table(s)
    - Data Table and Attributes
      - Inputs
        - Include a full description of the attributes of any tabular source data used as inputs for your model (e.g., CSV files). For each column of the table, be sure to include the column's header (Attribute Code), a plain-English name (Attribute Name), a narrative description (Attribute Definition), the type of data in the column (e.g., numeric, text; Data Type), its units, where applicable (Unit), and a list of values for categorical data (Possible Values).
      - Outputs
        - Include a full description of the attributes of any tabular data generated as outputs from your model. For each column of the table, be sure to include the column's header (Attribute Code), a plain-English name (Attribute Name), a narrative description (Attribute Definition), the type of data in the column (e.g., numeric, text; Data Type), its units, where applicable (Unit), and a list of values for categorical data (Possible Values).
- Methods
  - Resource Lineage
    - Lineage Statement
      - Codebase
        - Include a 1-2 sentence summary of any relevant information about the the development of your model.

  - Original Environment
    - Inputs
      - Note the file format(s) of your input files (e.g., CSV, ASCII).
    - Codebase
      - Note any programming language(s) and version(s) used to write your model (e.g., Python 3.5).
    - Outputs
      - Note the file format(s) of your model’s output files (e.g., NetCDF).
- Process Steps
  - Process Step Description
    - Codebase
      - Use these fields to provide a narrative description of what your model does. Include a separate Process Step Description field for each major step followed by your model when it runs (you can add as many of these fields as needed in the Research Workspace’s metadata editor). Aim to provide future re-users of your model with a high-level, conceptual overview of how your model operates.
- Source Data
  - Source Title
    - Inputs
      - Provide a descriptive title for your source data, including general information about the project such as its subject, location, and time period. If possible, use the same title as the source data’s citation.
  - Source Description
    - Inputs
      - Provide a brief description of the source data.
  - Source Citation
    - Inputs
      - Provide the original citation as specified by the source data’s publisher.
  - Processing
    - Inputs
      - Describe how any source data were prepared before for model ingestion.
- Data Quality Reports
  - Attribute Accuracy
    - Inputs and Outputs

        - Describe the accuracy of the values and include a description of any tests used to verify that accuracy.
  - Data Consistency Report
    - Inputs and Outputs
      - Describe any quality control measures that were taken to assess the data consistency of either your inputs or outputs, including a description of any specific tests you used.
  - Completeness Report
    - Inputs and Outputs
      - Include information about any data omissions, inclusion or exclusion criteria, or other rules used on your model inputs or outputs.
  - Usability Report
    - Inputs and Outputs
      - Include information about the degree to which your inputs or outputs adhere to a specific set of conventions or user requirements (e.g., climate and forecast conventions).

4. Guidance for Type 3 Projects: Applied Models

If the primary purpose of your project is to apply an existing model in a new way, follow the guidelines below when you're ready to archive and document your projects.

4.1 Archiving requirements

The final product of Type 3 projects will include the model's inputs and ouputs:

- Inputs may include data collected for the project, source datasets collected or curated by others, and any model parameters required to run the model.
- Outputs may include hypothetical data, either in structured formats such as NetCDF or in custom formats, as well as any diagnostic information about the model run that produced the output.

4.2 Metadata requirements

Most model-based projects will be described by one metadata record that covers your model inputs and outputs. This will typically be a folder-level record for the "Data" folder within your project listing in the Research Workspace. For large projects with multiple models, or other unique projects, more than one record may be required. If you don't know how many metadata records to generate for your project, contact the NPRB Program Manager to discuss.

In addition to the metadata fields that describe your model in a general way (e.g., keywords, spatial extent, constraints), several of the fields within your metadata record will need to describe your model inputs and outputs more specifically. The list below describes how those sections of your metadata record should be used to describe these components of your model:

- Resource Overview
    - Basic Overview
        - Title
            - Project in general
                - The title of your metadata record should include general information about the project, including its subject, location, and time period.
        - Abstract
            - Project in general
                - Your metadata Abstract should include general information about the project, such as its subject, location, and time period.
            - Inputs
                - Additionally, your metadata Abstract should include a 1-2 sentence summary of any source data or inputs used by your model. If applicable, include the citation for the source data. If your inputs come from data collected as part of this project (i.e., funded by NPRB under this project number), note that here.
            - Outputs
                - Finally, your metadata Abstract should include a 1-2 sentence summary of your model output files, including a mention of the file formats (e.g., NetCDF) and any conventions (e.g, climate and forecast conventions) used. If your output files are hosted outside the Research Workspace, include a link to those files here.
- Resource Content
    - Data Table(s)
        - Data Table and Attributes
            - Inputs
                - Include a full description of the attributes of any tabular source data used as inputs for your model (e.g., CSV files). For each column of the table, be sure to include the column’s header (Attribute Code), a plain-English name (Attribute Name), a narrative description (Attribute Definition), the type of data in the column (e.g., numeric, text; Data Type), its units, where applicable (Unit), and a list of values for categorical data (Possible Values).
            - Outputs
                - Include a full description of the attributes of any tabular data generated as outputs from your model. For each

column of the table, be sure to include the column’s header (Attribute Code), a plain-English name (Attribute Name), a narrative description (Attribute Definition), the type of data in the column (e.g., numeric, text; Data Type), its units, where applicable (Unit), and a list of values for categorical data (Possible Values).

- Methods
  - Resource Lineage
    - Lineage Statement
      - Inputs
        - Include a 1-2 sentence summary describing the origin of any source data used in your project.
      - Original Environment
        - Inputs
          - Note the file format(s) of your input files (e.g., CSV, ASCII).
        - Outputs
          - Note the file format(s) of your model’s output files (e.g., NetCDF).
  - Process Steps
    - Process Step Description
      - Inputs and Outputs
        - Use this field to provide a general summary of how the model you’re applying processes its inputs to produce the outputs.
  - Source Data
    - Source Title
      - Inputs
        - Provide a descriptive title for your source data, including general information about the project such as its subject, location, and time period. If possible, use the same title as the source data’s citation.
    - Source Description
      - Inputs
        - Provide a brief description of the source data.
    - Source Citation
      - Inputs
        - Provide the original citation as specified by the source data’s publisher.
  - Processing
    - Inputs

      - Describe how any source data were prepared before for model ingestion.
  - Data Quality Reports
    - Attribute Accuracy
      - Inputs and Outputs
        - Describe the accuracy of the values and include a description of any tests used to verify that accuracy.
    - Data Consistency Report
      - Inputs and Outputs
        - Describe any quality control measures that were taken to assess the data consistency of either your inputs or outputs, including a description of any specific tests you used.
    - Completeness Report
      - Inputs and Outputs
        - Include information about any data omissions, inclusion or exclusion criteria, or other rules used on your model inputs or outputs.
    - Usability Report
      - Inputs and Outputs
        - Include information about the degree to which your inputs or outputs adhere to a specific set of conventions or user requirements (e.g., climate and forecast conventions).