Plan Overview

*A Data Management Plan created using DMPTool*

**DMP ID:** [https://doi.org/10.48321/D14776D470](https://doi.org/10.48321/D14776D470)

**Title:** Hakai Standalone Instrument Data Pipeline

**Creator:** Jessy Barrette - **ORCID:** [0000-0002-0422-2589](https://orcid.org/0000-0002-0422-2589)

**Affiliation:** Hakai Institute

**Data Manager:** Jessy Barrette

**Project Administrator:** Jessy Barrette

**Funder:** Hakai Institute

**Template:** Hakai Institute Data Management Plan

**Project abstract:**

This data management plan outlines Hakai Institute's comprehensive approach to managing data from standalone instruments deployed at specific locations for defined periods.

Metadata pertaining to deployment and site specifics is meticulously recorded and maintained within dedicated dataset repositories. Upon retrieval, instruments undergo regular on-site maintenance to ensure data integrity. Instrument data is harvested, uploaded to project-specific GitHub repositories in the original manufacturer-specific format, and subjected to rigorous validation processes using GitHub Actions. Issues identified trigger aren't merged until corrective actions are made by the repository maintainer. The issues are listed through a pull request, and ensures metadata and data completeness.

Harvested data is automatically parsed and integrated into the Hakai database using the ocean-data-parser Python package, with initial quality control checks performed to uphold data integrity. Manual quality control procedures by data analysts further validate data accuracy within the Hakai database, incorporating revisions and flags as needed.
Processed data is then made accessible through the Hakai Sensor Network page and/or the Hakai ERDDAP, facilitating broader access and utilization for research and decision-making purposes.

**Start date:** 01-01-2012

**Last modified:** 07-05-2024

**Copyright information:**

The above plan creator(s) have agreed that others may use as much of the text of this plan as they would like in their own plans, and customize it as necessary. You do not need to credit the creator(s) as the source of the language used, but using any of the plan's text does not imply that the creator(s) endorse, or have any relationship to, your project or proposal.
Hakai Standalone Instrument Data Pipeline

Data Management Plan

Adapt, edit, add and remove sections below to complete your Data Management Plan as you see fit.

This plan outlines the systematic approach Hakai Institute employs to manage data from standalone instruments, ensuring rigorous data validation, integration, and accessibility to support research and decision-making processes effectively.

Deployment of Standalone Instruments

- **Location and Duration:** Instruments are deployed at specific locations for defined periods to collect data continuously.
- **Metadata:** Deployment and site-specific metadata is recorded and maintained within the dataset repositories, instrument-log and station-log.
- **Maintenance Procedures:** Regular maintenance is conducted on-site when retrieving instruments to ensure proper functionality and data integrity.

Data Retrieval, Submission and Archiving

- **Data Upload:** Instrument data is harvested and uploaded to project-specific GitHub Dataset repositories in the original manufacturer-specific format.
- **Organization:** Each project has dedicated GitHub repositories for storing datasets, maintaining clarity and project-specific context.
- **Validation Process:** Upon data submission, GitHub Actions execute tests to validate the completeness and appropriateness of metadata related to instrument deployment.
- **GitHub Action Feedback:** If issues are identified, GitHub Actions generate a pull request (PR) listing encountered issues for resolution.

Data Integration

- **Automatic Upload:** Once all issues are addressed, the latest changes to the dataset repository main branch is merged to a production branch. Changes to the production branch triggers a remote server which runs the Hakai Data Repository Harvester. Which is applying the following steps:
  - Update its local copy of the repository
  - Detect new and modified files
  - Parse the manufacturer proprietary format
  - Join the new data with the instrument- and station-specific metadata
  - Applies a series of automated quality control tests using the ioos_qc package
  - If applicable, apply manual flags available for this dataset within the sn_qc schema of the Hakai database.
  - Upload the resulting data to the Hakai database within the sn_sa schema and platform specific table.
  - Bin resulting data to 5 minutes and 1 hour intervals to be used by the Hakai Sensor Network interface.
- **Data Parsing Tool:** The ocean-data-parser Python package is utilized for data harvesting into the Hakai EIMS database, supporting various standard manufacturer data formats.
- **Quality Control Procedures:** Initial data quality control (QC) checks are performed during data harvesting to ensure data integrity.

**Data Revision**

- **Manual Quality Control:** Data analysts conduct additional manual QC procedures within the Hakai database to validate data accuracy and completeness.
- **Revision Integration:** Manual revision and flags are integrated within the Hakai database and merged to the dataset by individual record.

**Data Access**

- Processed data is made publicly accessible through the Hakai Sensor Network page and/or the Hakai ERDDAP, ensuring data visibility and usability.
- 5 minute and hourly binned data is made available via the Hakai Sensor Network page.