Hubbard Brook LTER Data Management Plan

Introduction

The overarching goal of Hubbard Brook (HBR) Information Management (IM) is to ensure that the research data products collected through NSF-LTER funding are preserved and openly available to support the work of HBR investigators, the broader scientific community, educators, resource managers, policy makers, and the public. This is achieved through 1) documentation and preservation of research data, 2) organization, management, and distribution of non-data assets – publications, educational material, public outreach materials, physical samples, 3) development and distribution of training materials on best practices for the collection, documentation, and publishing of research data, and 4) involvement with the broader LTER IM community committees and activities.

The Data Lifecycle

HBR Information Management works with investigators throughout the data lifecycle to provide expertise in data collection, quality control, documentation, archiving, retrieval, and analyses. The LTER IM community has developed, and continually refines, best practices for data curation, guidelines for site-level information management, and website content. This section describes HBR IM work throughout the data lifecycle.

Plan - Research at HBR is conducted by a geographically dispersed group of cooperating scientists from more than twenty institutions. The data collected by these scientists include core data sets that comprise the long-term monitoring program, as well as data from shorter-term studies. All experiments and data collections at HBR originate with a proposal to the Research Approval Committee (RAC). The proposal submission form contains questions about anticipated data collection, and upon approval, new project teams are provided with information regarding our goal to curate and preserve all data, regardless of funding source, and to make these data publicly accessible. In addition to the RAC chair, the IM receives all proposals as they are submitted, and can respond to the chair with data questions and/or concerns prior to the RAC panel review.

Ongoing long-term data at HBR includes meteorology, hydrology, biogeochemistry, vegetation, birds, insects, and more. The HBR IM works with the teams collecting these data to implement best practices to manage data from collection to repository. Coordination at this level ensures that data are standardized throughout the HBR data catalog, and that requirements for final repository submissions are clear to individual investigators and are incorporated into data management workflows at the earliest stages.

Data management training for both new and seasoned HBR researchers is provided by IM through meeting presentations, one-on-one consultations, and several weekly blocks of virtual office-hours. A series of short HBR-specific data management videos is currently in development. The goals of the training program are to establish familiarity with the entire data lifecycle at HBR, provide resources for investigators to manage their data within their own labs, and to provide instructions on a clear path to submitting data to the Environmental Data Initiative repository (EDI).

Collect - Data collection occurs year-round at HBR,with methods that include manually recorded observations (on paper and/or tablet), sensors streaming environmental data from the field to on-site (via radio telemetry) and off-site (via cellular) base stations, field-based audio and camera data, and unmanned aerial vehicle sensor platforms. The long-term climate and hydrology data, used in some form by most HBR researchers, have been collected by the USDA Forest Service since 1955. With the transition in the last decade from paper charts to digital sensors, these data are now accessible in near-realtime, through the hourly mirroring of data from the USFS to the HBR IM, and to research partners. These raw, provisional data are used to monitor current conditions, to trigger specialized sampling in response to meteorological and hydrological conditions, and to maintain a level of data QC in near-realtime (allowing for immediate response to instrument malfunction).

Assure **-** Data collected at HBR are diverse and often highly specialized. Individual researchers are responsible for developing their own data management protocols (e.g., database design, QA/QC, data backup) prior to submission to the centralized data catalog. The IM is available to provide guidance and assistance through all phases of research to ensure the integrity and safety of the data and metadata. Once data and metadata have been provided, the IM reviews submissions (including a review of data formatting, QA/QC checks, EDI data-metadata congruency checks, etc.) prior to uploading data to the data catalog.

Describe, Preserve, and Discover - To maximize discovery and usability of HBR data, datasets are prepared following the Best Practices for Dataset Metadata in Ecological Metadata Language (Gries et al. 2021) and submitted to EDI. These best practices use keywords aligned with controlled vocabularies (LTER, USDA Forest

Service, and ISO 19115), fully describe all data table attributes, reference funding sources, and include temporal, geographic, and taxonomic coverages. All published data are released under the Creative Commons 'CC BY – Attribution' license, providing open data access with credit to the data originator. Dataset documentation and licensing, combined with features and services of the EDI repository, help to align our data with FAIR (Findable, Accessible, Interoperable, Reusable) data practices (Wilkinson 2016). EDI further publishes these data through the DataONE federation. At both EDI and DataONE, HBR data can be discovered via a sophisticated search interface and manually or programmatically downloaded. The EDI repository fully supports immutability and strong versioning of datasets with all older versions being available for auditing and reproducibility. Each dataset version is provided with its own unique Digital Object Identifier (DOI), which is registered by DataCite and resolvable through the International DOI Foundation (IDF). In addition to data discovery at the HBR website, EDI data portal, and DataONE, Data in EDI are also more widely discoverable through other data search engines (e.g., Google Dataset Search and DataCite) by the inclusion of rich schema.org structured metadata.

Integrate and Analyze - Data collected at HBR are integrated into research workflows, viewed by stakeholders and the public, and used in K-12, post-secondary, and graduate programs. HBR researchers are developing analysis workflows that tie directly to data in the EDI repository, a practice that adds robust traceability to research products. Access to data is also available through apps such as the HBWaTER public portal, which supports interactive and dynamic queries and displays of long-term stream chemistry data. This public interface reads data directly from EDI, providing updates in sync with the most recent data publication.

Leveraging resources from the Environmental Data Initiative (EDI)

Over the past five years of the current funding cycle, IM tasks at HBR have been transformed through the ability to leverage tools and resources developed by EDI. Where HBR IM was once structured on software and web services developed in-house, we have now adopted many EDI and LTER network-level IM resources and services. This has improved the volume and timeliness of repository data submissions, and our ability to meet the guidelines for FAIR data curation (Wilkinson 2016).

EDI resources leveraged for HBR IM work include workflows that aid in reporting (data download reports, data citation statistics, dataset alignment with FAIR guidelines), the PASTA+ API to generate a local data catalog to display all HBR data, and the EML Assemblyline R package for developing EDI-ready data packages. HBR realizes cost-savings through the use of these resources and benefits by using resources and workflows with a wider user base and strong developer and community support.

IM resources

Personnel - Since 2012, the HBR Information Manager position has been held by Mary Martin, based at the University of New Hampshire in the Earth Systems Research Center (ESRC). With HBR-LTER funding supporting nearly 100 investigators at more than 20 institutions, many individuals contribute to HBR IM through efforts in individual research labs. A number of core HBR datasets are collected and managed by cooperating entities, each with in-house information management teams. Most notable here are 1) the USDA Forest Service, Northern Research Station, collecting core hydrology and meteorology data used throughout HBR research since the mid-1950s and 2) The Hubbard Brook Watershed Ecosystem Record (HB-WaTER) team, funded by NSF Long Term Research in Environmental Biology (LTREB) collects and curates the long-term stream and precipitation chemistry data. The HBR-IM works closely with the USFS and HBWaTER teams to ensure that all long- and short-term data are prepared and preserved seamlessly across the entire HBR community. The Hubbard Brook Research Foundation (HBRF) contributes to the development and maintenance of the website, with HBRF communication specialists focusing on content that synthesizes research findings for land managers policy makers, and the public. HBRF also provides an extensive collection of K-12 classroom data exercises available on the HBR website.

Computing and storage infrastructure - The HBR IM work at UNH (University of New Hampshire) is supported by a department-level Service Level Agreement (SLA) with the UNH Research Computing Center (RCC), which covers systems administration, hardware and software support, system security, and backups. THe UNH RCC support for IM desktop and server systems includes weekly operating system patches, generation of log files for auditing, and monitoring by an intrusion detection system. Data backups consist of incremental daily and weekly full backups, with both on-site and redundant cloud-based storage. In addition to general IT support, RCC can also provide personnel for as-needed project support. Computing software and resources are shown in Table 1.

Cloud services - HBR uses several secure cloud-based platforms to store and support the distribution of information (Table 1). The website is migrating to WordPress (Bluehost: mid-2022), the bibliography is managed and served publicly with Zotero (local storage in Zotero desktop and browser-based access by mirroring to zotero.org), reporting and registration forms are developed in JotForm, and at the pre-repository stage, data and work products are shared on BOX and OneDrive/SharePoint. Identifying and using cloud services where appropriate was recommended in the last mid-term review, and has been beneficial to overall site IM in terms of economical use of resources and time.

Table 1. Features of HBR Information Management System

Feature	**Details, software, resources**
Website	html, css, php, xslt, javascript, apache, Piwigo, Drupal migration to WordPress in 2022
Bibliography	Zotero (desktop and cloud), bibutils
Data Catalog	Local website access to HBR EDI content uses the PASTA+ API
Metadata	EML Assemblyline (developed by EDI), custom R scripts
Computer Hardware	Dell PowerEdge R510, desktop and laptop linux systems.
Backup	BackupPC for daily, weekly and monthly backups; rsnapshot for hourly backups; on and off-site backup storage
Data management	R, R-shiny, Trello, Git, LibreOffice, MS Office/OneDrive/SharePoint, JotForm, QGIS, Postgres, MySQL, BOX

HBR Website

The HBR website is the primary means by which HBR information is disseminated. Website development follows the network-developed Guidelines for LTER Web Site Design and Content (version 2.0; 2018). All updates to the website undergo initial testing and review on a developmental website. Full website content backups occur daily and are stored on the hosting platform and a second cloud service. Website use is tracked using Google Analytics. Over the past three years, the website has averaged 500-800 visits per month. The website provides access to:

Personnel database - A searchable personnel database, is maintained on the HBR website where each individual investigator, graduate student, postdoc, and staff member is listed in a directory, and has an individual page. HBR publications for each individual researcher are displayed dynamically on each page, through a query to the Zotero bibliography. Updates to personnel pages can be made at any time and reminders are sent out twice annually to ensure that the information is current.

Current Research – A description of current research activities is available to inform the research community and public about research initiatives and preliminary findings. Updates reflect changes in the development and scope of current research. Each project links to relevant publications, data, fact sheets and research briefs.

Research Synthesis – A multi-chapter online 'book' describing key findings throughout the research history of HBR is featured on the website. Each chapter is maintained by one or more HBR topic specialists. To date, there are 17 chapters, including one chapter focusing on a step-by-step data analysis exercise. All chapters link to underlying available data and pose several questions suitable for future investigation. In 2022, we will begin to convert static graphs within these chapters to dynamic interactive graphs that reflect the most recent data available in the EDI repository.

Photo archive - The website has a searchable archive of digital images that are frequently used in publications, presentations, and textbooks. Many of the historical HBR photographs and slides have been scanned at high resolution to ensure that these irreplaceable images are preserved. The online photo gallery uses a local installation of Piwigo. This full-featured, open-source photo management software allows for photo upload, tagging, search, and for user accounts with varying permission levels. An Epson V800 slide scanner is available

for general HBR use, and several thousand 35mm slides have recently been scanned at high resolution for addition to this digital collection.

Education and Outreach Material – The Hubbard Brook Research Foundation (HBRF) provides website content focusing on the synthesis of research findings for land managers, policy makers, and the public, and provides an extensive collection of classroom data exercises for K-12. Outreach content includes research briefings, fact sheets, stakeholder roundtable reports/products, and access to speakers and a zoom-a-scientist program.

Publications - The publications from the Hubbard Brook Experimental Forest date back to 1955, and include more than 2,400 books, journal articles, conference presentations, and theses. New publications are identified through self-reporting by investigators, a monthly newsletter, annual reports, and Google Scholar alerts. Citations are managed locally with Zotero. This open-source bibliography management software harvests citations, citation metrics, and associated pdfs through a browser, and exports to standard reference management file exchange formats. All HBR publications are uploaded to the central LTER bibliography managed by the LTER Network Office.

Data catalog - The HBR data catalog is displayed on the website as a custom data portal developed through the PASTA+ API. A complete inventory of data is available in a Supplemental Document to this proposal. The data catalog contains 270 data sets ranging from single year studies to long-term data collections. More than 20 data packages contain data collected for more than 50 years, and another 30 cover a timespan of more than 20 years. During this past funding cycle, we have appended data to our core data sets and incorporated more than 130 new data sets. We have also restructured several long-term data sets to better serve the user community. For example, data collected at 5-year intervals had been packaged in separate data sets for each collection cycle, with varying formats. These have now been harmonized and restructured into single datasets to enhance usability. Statistics for data downloads from EDI (for individual data packages and as HBR-wide summaries) are generated using a programmatic reporting application and show a data package download rate of 650 per month over the past three years.

Sample and Document Archives

For more than 30 years, HBR has maintained a commitment to the permanent storage of physical samples collected at the site (e.g., streamwater, precipitation, vegetation, soil). A dedicated building on-site serves as the archive facility, and now houses approximately 100,000 samples. Samples are preserved, barcoded, and cataloged with associated metadata in a MYSQL database; a process that ensures the discoverability and access to samples for future research. A sample archive subsampling policy has been developed to 1) maintain the chemical integrity of the samples; 2) preserve sample volume for future analysis; 3) document the use of the samples, and any resulting changes; 4) inform principal investigators of interest in sample use; and 5) acknowledge the appropriate funding sources for their original collection. Requests for analysis of archive samples (e.g., isotopic analyses, heavy metals) are received regularly, and have resulted in at least 37 publications (soils n=8 publications; water n=15; forest floor n=11; plant material n=3).

During this current funding cycle, LTER funds were used to 1) improve physical storage efficiency through sample reorganization and increased space efficiency, and 2) enhance sample discovery through a restructuring of sample collections and the associated database used for sample management and public-facing search capabilities. A physical sample and data entry station in the archive building incorporates bar-code scanning and direct scale-to-computer entry of sample weights and has facilitated the addition of 35,000 new samples to the archive in the past 5 years.

A new activity over the past five years has been the development of a document archive. Research at HBR, like many LTER sites, crosses the “analog-to-digital divide”, which creates a vulnerability that key documents and information (e.g., data sheets, field notebooks, strip charts, instrument manuals) could be lost. We have had a series of meetings to discuss the need for, purpose and uses of a document archive at HBR. Funds have been raised from private sources (~$100K) for construction of a building at HBR and we have organized a committee that will set the policies and practices of the document archive. Over the next few years the building will be constructed and we will begin to move materials stored elsewhere into the archive.

Network participation

IM Martin actively contributes to the LTER Network’s IM activities in several capacities. In 2021, she was elected to a 3-year term on the executive committee of the LTER Information Manager Committee (IM-exec), and she serves on the EDI advisory committee. She has also participated in LTER/EDI working groups addressing data

package design for special cases (Gries 2021), semantic annotation, and best practices for units. HBR is also participating in the migration of high resolution meteorological and hydrological data to a harmonized 'next generation' data package design (formerly the climDB and hydroDB databases). HBR also attends and participates in monthly and annual LTER-IM meetings, and regularly scheduled EDI webinars and 'town hall' meetings.

Work Plan Over the Next Six Years:

ANNUALLY

- Review and update data management tasks, tools, and practices to align with network standards.
- Update all HBR signature datasets and develop new data collections as submitted.
- Publish all updated and new datasets to EDI.
- Offer training to HBR investigators, graduate students, and staff on the data submission process, tools for data quality control, best practices for data citation, and new repository features.
- Attend and present IM updates at all HBR Quarterly Project Meetings.
- Contribute 'Data Report' to HBR monthly newsletter.
- Attend monthly and annual LTER Information Management Committee meetings and participate in LTER/EDI working groups.

YEARS 1-4

- Modify the HBR data package workflow to streamline common metadata content into EML Assemblyline
- Develop workflows for the contribution of climate and hydrology data to the next-generation harmonized data package format replacing the climDB and hydroDB database.
- Add metadata content to older closed datasets, where enhancements are now enabled through the use of EML2.2.
- Evaluate all dataset metadata for the addition of appropriate semantic annotation, implementing the annotations additions, and publishing these revisions to EDI.
- Continue an initiative to convert static figures in the online research synthesis chapters to live, interactive graphs that read data directly from EDI.
- Continue an ongoing initiative to identify and preserve data assets falling into 'Special Cases' as described in Gries et al. (2021) (e.g., models, UAV data, images, and audio files).
- Identify high value data assets contained in the developing Document Archive collection and prioritize conversion to digital format and submission to EDI.
- Publish GIS base layers in ArcGIS online, enabling data access through web services.
- Continue development of a video tutorial series on data access, use, and submission.

YEARS 5-6

Continuing work on all the above, the last two years of the funding cycle focus on completion of all tasks listed above, responding to input from the mid-term review, and preparing activity reports for the next proposal.

References:

Environmental Data Initiative. Best Practices for Dataset Metadata in Ecological Metadata Language (EML Best Practices V3). 2017. https://environmentaldatainitiative.files.wordpress.com/2017/11/emlbestpractices-v3.pdf

Gries, C., S. Beaulieu, R.F. Brown, S. Elmendorf, H. Garritt, G. Gastil-Buhl, H. Hsieh, L. Kui, M. Martin, G. Maurer, A.T. Nguyen, J.H. Porter, A. Sapp, M. Servilla, and T.L. Whiteaker. 2021. Data Package Design for Special Cases ver 1. Environmental Data Initiative. https://doi.org/10.6073/pasta/9d4c803578c3fbcb45fc23f13124d052 (Accessed 2022-03-17).

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18