# Data Management Plan

## Introduction

Information Management (IM) is an integral component of the CAP LTER Program with the overarching goals of: (1) supporting data collection; (2) archiving well-structured and -documented research data in long-term data repositories for the benefit of the scientific community, decision makers, and public; (3) enabling and promoting dataset discovery and access; and (4) providing leadership and education on sound data management. CAP maintains high standards for data archival and documentation to ensure the quality of the scientific data and metadata produced. We employ a diverse array of tools and technologies to make those resources discoverable and useful. The Information Manager stays abreast of developing technologies and approaches to IM and incorporates beneficial innovations that support the goals listed above. The Information Manager works with CAP scientists, students, and staff in a variety of capacities to address data management throughout the knowledge-generating enterprise—from research design to data publication. CAP is an active contributor to LTER-Network IM and adheres to all NSF and LTER Network data policies.

## IM resources

*Personnel*: The CAP Information Manager (S. Earl) is supported at 0.5 FTE by CAP as the Information Manager, and 0.5 FTE by CAP’s host institution, the Julie Ann Wrigley Global Institute of Sustainability (GIOS) at Arizona State University (ASU). Earl brings an excellent combination of ecological research experience and data management skills to the project, and the bulk of his time is devoted to CAP’s data management needs. Earl is the central point of contact and primary overseer of information resources for the program. He works closely with GIOS programmers, web application developers, and the Director of Informatics and Technology to deploy and maintain the hardware, software, communications, and web resources needed to support the high-quality information management and technology essential to the success of the Program.

*Infrastructure*: CAP leverages a suite of technologies and resources to meet its IM requirements (Table 1). GIOS provides the technical resources CAP requires, including comprehensive computing solutions based on Amazon Web Services (AWS), and file storage on Windows network shares purchased through ASU’s University Technology Office. AWS provides web and database services running on virtual Linux servers ranging in size from t2.micro to t2.large. These resources host the CAP website, other web resources (e.g., personnel database, CAP’s equipment reservation system, web-based data-entry applications), and centralized databases (MySQL, PostgreSQL). Data and metadata files, accessible via the CAP website data catalog, and file-based research data that are not stored in one of the database systems reside on AWS storage (S3). Other documents (e.g., images, equipment manuals, machine output) are stored on a Windows network share, Dropbox, or Google Drive, dependent on project needs. ASU has agreements with Dropbox and Alphabet for large-volume cloud storage, and CAP is transitioning to store most legacy documents on Dropbox. These secure storage systems provide resilient storage for all project files while still ensuring CAP personnel and external collaborators have access to research materials. All networked systems and web applications are password-protected, and ASU performs regular security sweeps to identify vulnerabilities or suspect behavior.

The local network supports high-bandwidth connectivity with high-speed internet access through ASU’s connections to the Internet2 backbone. GIOS addresses long-term data protection through regular technology transfers to maintain current standards for hardware and software. This strategy minimizes the risk of data loss through media or format obsolescence. GIOS also provides advanced video- and tele-conferencing facilities for CAP investigators.

|  |  |  |
| --- | --- | --- |
| **system** | **application** | **implemented service** |
| AWS Linux servers | WordPress | CAP website, including CAP data catalog |
| miscellaneous web services | CAP personnel database; publications database; data-entry applications; CAP equipment reservations |
| databases (Postgres, MySQL) | long-term research and experiment data |
| AWS | Simple Storage Service (S3) | data and metadata files accessible through CAP data catalog; file-based long-term experiment and research data |
| Dropbox |  | document storage |
| Windows network share |  | document storage |

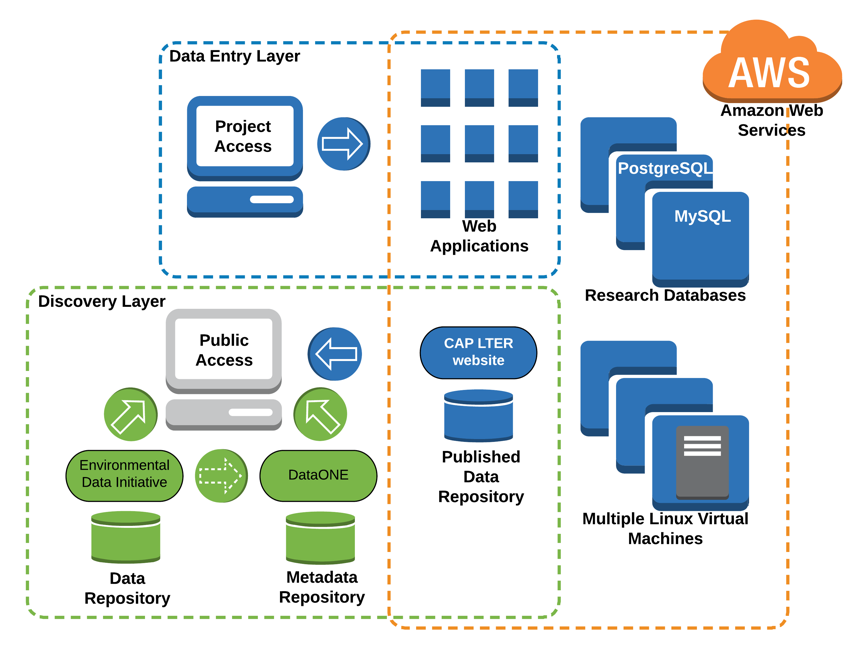
## Table 1. CAP information management technical resources and their application

## IM system

*Systems architecture*: The CAP website is the primary access point for project information and resources. Here, users are able to view personnel information, explore CAP research themes and projects, search out education and outreach resources, and find information about many other aspects of CAP. The website is presented by a WordPress content management system multisite installation. CAP bibliographic details reside in a MySQL database, which is searchable by title, author, journal, or keyword using a simple search interface. This search interface removes the need to enter search arguments in multiple search fields. Past proposals, reports, and annual All Scientist Meeting posters are also accessible through the website. Other documents, including journal articles, images, and white papers in digital format are stored in structured file systems on a Windows network share or Dropbox, and are accessible to all CAP investigators. The CAP equipment reservation system, and data-entry and -upload tools are custom applications that are also available through the project website.

CAP archives a wide range of datasets, including long-term observational and experimental data, one-off project data (e.g., from student research), and supported third-party data (e.g., remotely-sensed imagery, USGS digital elevation data, climate and hydrology data). Data entities span many types, including spatial (raster, vector), tabular, text, and imagery. Dataset keywords are harvested from the metadata files and stored in a MySQL database to support CAP data catalog searches. Dataset metadata resulting from catalog searches are dynamically generated from the metadata XML file and an XSL stylesheet. CAP’s long-term monitoring and experiment data are stored in relational database systems (MySQL, PostgreSQL) with periodic (ca. annual) snapshots made publicly available. Workflows and code-based tools for processing and publishing CAP’s long-term data, as well as other tools to support CAP information management, such as code for web-based data-entry applications, are housed in the ‘CAPLTER’ GitHub organization. Git repositories document code and dataset updates or additions, providing a valuable, detailed audit trail of both content and metadata changes.

*Support for data collection and quality assurance*: CAP employs a combination of tools and workflows to facilitate data collection, processing, transfer, and storage, including automated data streams from sensors, field and laboratory data entry, and uploading to centralized databases (Figure 1). Autonomous systems capture sensor data streams and transfer them into permanent storage after undergoing quality-control routines. Data from field and laboratory observations are recorded via web-based data-entry tools developed with Ruby-on-Rails or Shiny. These tools are tuned to optimize workflow efficiency and quality control at the time of entry. CAP has recently had success employing tablets that use form-based data-entry interfaces for field data collection; we will continue to develop this technology. Data generated from analytical analyses are uploaded to databases using web-entry tools, or processed with scripts (R, Ruby) for efficient transfer from the laboratory to data repositories while applying appropriate quality control. Many analytical workflows employ barcodes on samples, which greatly increases processing efficiency and minimizes data recording and transfer errors.



**Figure 1.** CAP information management data collection, publishing, and access workflow.

*Data publishing workflow*: Whereas there is a near-continuous flow of data from CAP’s long-term observational and experimental efforts, data from individual research projects (e.g., student research) are typically provided upon completion of the project. Revisions of CAP’s long-term data are released at approximately yearly intervals, and individual research projects are published within two years of submission, as per the LTER Network Data Access Policy, or sooner if in conjunction with publication of associated journal articles. Investigators submit datasets using metadata forms that are available on the CAP website along with submission instructions. The Information Manager works with data providers to address data and metadata issues in order to produce high-quality, well-documented datasets. To maximize discoverability and compatibility with other LTER data, dataset keywords are mapped as closely as possible to the LTER Controlled Vocabulary, and measurement units are aligned with the LTER Unit Dictionary or detailed according to LTER Best Practices in cases where a unit is not in the dictionary. The Information Manager works with a wide array of data types (e.g., tabular, spatial, imagery), with these data types often commingled in dataset packages. This approach aids ease-of-use by eliminating the need for data users to download multiple datasets to obtain, for example, tabular and spatial components of a project.

Metadata, stored as XML files, are encoded in the LTER standard Ecological Metadata Language (EML), which is generated using a combination of the EML package for R (rOpenSci) and custom R scripts. The integrity of CAP data and metadata is maintained through careful review, and evaluation using the quality-control checks within the Provenance Aware Synthesis Tracking Architecture (PASTA+) system that ingests data into the Environmental Data Initiative (EDI) data repository. Data and metadata files are uploaded into AWS, and select project metadata (i.e., author(s), title, keywords, LTER core area, interdisciplinary research team, and abstract) are added to a MySQL database that enables dataset discovery and access through the CAP data catalog. At the same time, metadata are submitted to the EDI for inclusion in the EDI data repository (PASTA+ uses the file locations encoded in the metadata file to harvest data files from AWS to the EDI data repository).

*Data discovery and access*: CAP makes data available in accordance with the NSF Proposal & Award Policies & Procedures Guide (PAPPG 2018), the Directorate for Biological Sciences Updated Information about the Data Management Plan Required for Full Proposals, and the LTER Network Data Access Policy. CAP has adopted the Type I and Type II data designations detailed in the LTER Network Data Access Policy. In accordance with this policy, all Type I data are made publicly accessible in the CAP LTER data catalog and EDI data portal within two years of data collection. Only copyright-protected, third-party data; and human-subject data deemed sensitive by an Institutional Review Board are not public. These Type II data are available by request to the project investigator or Information Manager.

CAP data are discoverable and accessible through multiple avenues, specifically the CAP website data catalog, the EDI data portal, and DataONE through EDI’s membership (Figure 1). CAP has contributed 188 unique datasets to the EDI data repository (see Dataset Table). Other relevant documentation, such as metadata, publications, posters, and protocols are also available through the CAP website. Long-term observational and experimental data collected after the revision available in the catalogs are available to investigators upon request. The CAP website provides additional context and background to our long-term monitoring through overview vignettes. Restricted data resources, such as satellite imagery and copyright-protected publications, are available to CAP researchers via secure network storage or through the Information Manager.

## IM integration with research

The Information Manager is a member of the CAP Management Team (see Program Management Plan) and contributes to CAP events and efforts. The Information Manager works closely with CAP staff, investigators, and students to guide data management design and implementation. Research efforts require careful planning, and investigators work with the Information Manager before starting new projects to design effective data management approaches, especially for new long-term observational efforts or experiments. This forethought to data management, and the close collaboration among the Information Manager, other CAP managers, and researchers ensures the quality of research data, minimizes the back-end effort required to make the resulting data available for public consumption, and helps to smooth transitions when unexpected changes in methodology or project parameters are implemented. The Information Manager may also work directly with investigators on the analysis and interpretation of project data, and contribute to scientific publications (examples include Hale et al. 2014; 2015; Volo et al. 2014; Banville et al. 2017). For leveraged projects involving additional funding, the Information Manager may work with investigators to provide guidance on the development of data management plans. At the close of projects, the Information Manager works with investigators to help format data and develop metadata for publication in the CAP data catalog and the EDI data portal.

## IM education and outreach

The CAP Information Manager makes an extensive effort to educate CAP investigators and staff about data management resources and expectations, and strives to keep data management highly visible within the project. Many interactions with students are at the time of data publication, and CAP’s internal funding mechanisms (e.g., CAP Grad Grants) are structured to encourage students to submit their data and metadata upon completion of their research. However, there are many venues through which the Information Manager interacts with students and other project investigators earlier in the research process, including data management introductions and overviews given by the Information Manager at the annual All Scientists Meeting, at the annual fall welcome for new participants, and through numerous invited seminars. CAP students also have the opportunity to receive structured guidance by taking a research data management methods course, offered through ASU’s School of Sustainability, taught by the current (S. Earl) and former (P. Tarrant) CAP Information Managers.

## Network participation

CAP has and will continue to contribute significantly to Network-wide IM activities and leadership. CAP participates in all relevant Network-wide databases, complies with IM review criteria, and implements Network-wide information management approaches that maintain the highest degree of Network-wide interoperability. The Information Manager (S. Earl) served as a Tiger Team member for the EDI’s PASTA+ development, is a member of the LTER Information Management Committee’s (IMC) Executive Committee, chaired an IMC working group that revamped LTER IM training guidelines and resources, serves on an IMC working group tasked with revising the IMC’s terms-of-reference (renamed to bylaws), serves on the EML Congruence Checker (ECC) committee (that prioritizes quality control checks for the PASTA+ system), and is an information manager on the LTER Network Communications Office’s Soil Synthesis working group. The CAP Information Manager has also worked with other LTER Information Managers and members of the EDI to prototype approaches to employing the PASTA+ Application Programming Interface (API) to construct a custom data catalog by harvesting data from the EDI data repository. This work was demonstrated at the 2017 meeting of the Earth Sciences Information Partnership (ESIP). Relevant work stemming from CAP IM efforts has been presented at national and international conferences, in training forums, and has been published in peer-reviewed journals (e.g., Aguilar et al. 2010, Jones and Gries 2010) as well as the IMC newsletter Databits (e.g., Raub 2014).

## Notable future initiatives

Though CAP’s robust and effective approach to information management meets or exceeds all Network standards and expectations, the Information Manager strives continually to identify ways to improve the system. There are several areas in which we are planning improvements during CAP IV, including:

1. *Information Management Advisory Committee(IMAC)*: The Information Manager works closely with CAP investigators to implement effective data management. While valuable insight is garnered through those interactions, the interactions themselves are unstructured and we lack a formal mechanism for feedback. As part of CAP IV, CAP will form an IMAC consisting of a four-member, rotating group of CAP staff, managers, investigators, and students. The goal of this committee will be to critically review CAP’s information management resources and policies, from a user perspective, and offer suggestions to the CAP Executive Committee that will increase data discoverability and usability. The IMAC will be a welcome addition to the CAP program infrastructure when it is convened in summer 2018.
2. *Data catalog consolidation*: As detailed in this Data Management Plan, our current approach to data cataloging and dissemination includes maintaining data and metadata files accessible through the CAP website data catalog, and uploading metadata and data to the EDI data repository. CAP’s current infrastructure requires this parallel approach to meet the NSF requirement that data are deposited in a national, public data repository (the EDI in this case) and, at the same time, to showcase CAP data products and provide convenient access to CAP data to project investigators, decision makers, and the public through the CAP data catalog. CAP will modify its infrastructure such that the local data catalog will be generated from CAP datasets housed in the EDI data repository via the PASTA+ API. Specifically, we will deploy a version of the PASTA+ architecture along with Apache Tomcat on an AWS server to retrieve CAP datasets from the EDI repository, and make them available through a new catalog interface still available on the CAP website. This consolidation will reduce the overhead of maintaining two separate but identical data catalogs while meeting the NSF data-sharing requirements. The new infrastructure will also provide the Information Manager with greater flexibility to generate links to data collections (e.g., such as datasets pertaining to particular research topics). Such customization is currently constrained by the limitations of the search functionality offered by WordPress, which powers the search feature of the CAP data catalog. We look forward to the streamlined workflow and improved functionality when this new system is implemented in summer 2018.