# Data Management Plan

## Overview

Information Management (IM) is an integral component of the CAP research program with the overarching goals of: 1) supporting data acquisition; 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, adhering to all NSF and LTER Network data policies, to ensure the quality of the scientific data and metadata produced. CAP employs a diverse array of tools and technologies to make data resources discoverable and useful, including encoding rich metadata using the Ecological Metadata Language (EML) schema and archiving data in the Environmental Data Initiative (EDI) data repository. The Information Manager stays abreast of developing technologies and approaches to IM, and incorporates beneficial innovations into a continually evolving and improving IM system that supports the aforementioned IM goals. The Information Manager works with CAP investigators and staff in a variety of capacities to address data management throughout the knowledge-generating enterprise—from research design to data publication to education. CAP is an active contributor to IM at the LTER Network level and broader data-management community.

## Resources

*Personnel*: The CAP Information Manager (Stevan Earl) is supported at 0.5 FTE by CAP as the Information Manager, and 0.5 FTE as a Research Data Manager for the Knowledge Enterprise (the SRO) at Arizona State University (ASU, see the Facilities, Equipment, and other Resources statement for details). Dr. Earl brings a combination of ecological research experience and data-management knowledge to the project, and most of his time is devoted to addressing CAP’s data-management needs. He is the central point of contact and primary overseer of information resources for the program.

*Infrastructure*: CAP leverages a diverse suite of technologies, and physical and digital resources to meet the program’s IM requirements. The Information Manager works closely with ASU Knowledge Enterprise staff who provide web, application, and database services hosted on virtual Linux servers using Amazon Web Services (AWS) or ASU cloud ecosystems. These resources host the CAP website, other web resources (e.g., personnel and citation databases, CAP’s equipment reservation system, web-based data-entry applications), and centralized databases (MySQL, PostgreSQL) with appropriate access control, security, and recovery. ASU has an institutional agreement with Dropbox, which provides CAP staff and investigators with unlimited storage; this is used for document storage and as a collaboration tool. All networked systems and web applications are password-protected, and ASU performs regular security sweeps to identify vulnerabilities or suspect behavior. CAP organizational GitHub and GitLab accounts house all project-related code and informatics documentation. The local network supports high-bandwidth connectivity with high-speed internet access through ASU’s connections to the Internet2 backbone.

## Data policies

CAP is committed to maximizing the availability of its research products, and makes project data available in accordance with the NSF Proposal & Award Policies & Procedures Guide (PAPPG), 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, typically within two years or sooner of data collection. Type I data are released under the Creative Commons CC0 License that affords the widest possible reuse of data. Only copyright-protected, third-party, or human-subject data deemed sensitive by an Institutional Review Board are not public (i.e., Type II). Though not publicly accessible, Type II data are typically available by request to the project investigator or Information Manager.

## Integration with and support of research

Information management is an integral component of CAP research, and CAP IM is dedicated to support all phases of the data life cycle (outlined by DataONE): *Plan*, *Collect*, *Assure*, *Describe*, *Preserve*, *Discover*, *Integrate*, and *Analyze*. IM resources are available to CAP investigators through all stages of the research process from planning through dissemination, and the Information Manager works closely with CAP leadership and investigators to maximize the success and efficacy of urban ecological research through sound data management.

*Planning*: The Information Manager is a member of the CAP Management Team (see the Project Management Plan) and contributes to overall planning and management of CAP administration and research. 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 experiments and observational efforts. 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 protocols 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; McPhillips et al. 2019). 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 an appropriate data repository.

*Support for data collection and quality assurance*: CAP employs a suite of tools and workflows to facilitate data collection, processing, transfer, and storage of data generated by CAP’s long-term observational and experimental (core) programs. Most field data are collected with pre-formatted field sheets or tablets. Data collected with tablets are uploaded to CAP databases with scripted (R code) workflows. Observational data (e.g., bird surveys) recorded on field data sheets are double-entered into the CAP databases via web-based data entry applications developed with R (Shiny). These tools are tuned to optimize workflow efficiency and quality control at the time of entry. Data generated from our analytical laboratory (Goldwater Environmental Lab) undergo rigorous quality control at the time of analysis. These data and data from sensor platforms (e.g., micrometeorological stations) are uploaded to databases using web-entry tools (Shiny) or processed with scripted (R) workflows for efficient transfer of data to CAP databases while applying additional quality-control measures. Many analytical workflows employ barcodes on samples, which greatly increases processing efficiency and minimizes data recording and transfer errors. All source materials (e.g., scanned field data sheets, sensor downloads) are archived in Dropbox for redundancy and future accessibility. All scripted tools and workflows are documented in GitHub or GitLab.

For novel datasets (e.g., from student research), investigators submit data and metadata with 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 to produce high-quality, well-documented datasets with the goal of maximizing the potential reuse of the data. All datasets are processed with scripted workflows to ensure complete traceability of data processing. In addition to the published dataset, all materials contributed as part of the data submission, and processing scripts and documentation are archived in Dropbox.

*Data documentation*: Metadata, stored as XML files, are encoded in the newest version (2.2.0) of the Ecological Metadata Language (EML) schema. Dataset EML metadata are generated using a suite of publicly-accessible R packages (capeml, capemlGIS) developed by CAP. An additional R package (gioseml) facilitates pulling investigator details from the Global Institute of Sustainability and Innovation (GIOSI) database. The integrity of CAP 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 EDI data repository. To maximize discoverability and interoperability with other ecological 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 otherwise detailed according to LTER Best Practices. The Information Manager works with a wide array of data types (e.g., tabular, spatial, imagery), and these data types are often commingled in dataset packages. This approach aids ease-of-use by eliminating the need for users to download multiple datasets to obtain, for example, tabular and spatial components of project data.

*Preservation and discovery*: By default, CAP datasets are archived in the EDI data repository, but they are discoverable and accessible through several venues. The primary access point for CAP data is the data catalog on the CAP website. The catalog is populated from CAP datasets in EDI through a deployment of the PASTA+ architecture along with Apache Tomcat on an AWS server that retrieves datasets from the EDI repository and makes them available through the catalog interface on the CAP website. The CAP website provides additional context and background for CAP’s long-term data through overview vignettes. As CAP datasets are archived in the EDI data repository, they are discoverable and accessible through the EDI Data Portal, and through DataONE by extension of EDI’s participation as a DataONE member node.

## Education and outreach

The Information Manager makes an extensive effort to educate CAP investigators, staff , and the ASU community in general about data management resources and expectations, and works to keep data management highly visible within CAP. 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 CAP annual All Scientists Meeting, at the annual fall welcome for new participants, and through invited seminars. CAP students are encouraged also to take a course taught by Earl on research data management best practices offered through the ASU School of Sustainability.

## Contributions to LTER Network and other research communities

CAP has and will continue to demonstrate its strong commitment to contributing to informatics within the LTER Network and the data management community in general. The Information Manager participates in all Network information meetings and activities, and recently concluded a five-year term as a member of the LTER Information Management Executive Committee (IM Exec), serving as co-chair of the committee for the past three years. The Information Manager contributes to other community-wide efforts (e.g., is a contributing author of Ecological Metadata Language version 2.2.0), participates in and presents at numerous scientific conferences, and contributes to scientific and informatics publications. Earl also served as the embedded data manager for an LTER synthesis working group that addressed soil organic matter dynamics; this effort not only resulted in numerous scientific publications, but dovetailed with the establishment of an Earth Science Information Partnership (ESIP) cluster focused on soil ecology semantics and informatics for which Earl is a coordinating member. Additionally, Earl is a contributing member of a recently funded NSF Science and Technology Center (the Science and Technologies for Phosphorus Sustainability, or STEPS, project), serving as a liaison between CAP and STEPS.

## Milestones and future initiatives

Though CAP’s robust and effective approach to IM meets or exceeds all Network standards and expectations, the Information Manager strives continually to improve CAP IM. Implementing a much improved data catalog is an example of an IM improvement since the last proposal, and there are several areas in which CAP is planning improvements during CAP V:

*Improved discoverability of data through the CAP website*: Since the last proposal, CAP has implemented a new approach to providing a catalog of data on the CAP website. Whereas the previous approach required maintaining local copies of dataset metadata and data, the new approach employs RESTful web services to access dataset holdings in EDI. This streamlined approach is considerably more efficient than maintaining two parallel catalogs of identical data that draw from different sources. It allows CAP to harness the excellent search capabilities provided by the EDI, and eliminates potential versioning conflicts that could arise from drawing like data from different systems.

*Improved access and interpretability of data*: As detailed throughout this data management plan, making project data available to CAP investigators and the scientific community is a fundamental goal of CAP IM. However, while CAP IM is meeting this goal in a technical sense, as was highlighted in CAP’s 2020 mid-term review, that data are accessible in a public repository such as EDI does not necessarily translate to actionable data. That is, most CAP data are by nature academic and, even with rich metadata, can be difficult to interpret and reuse (the ultimate goal) by an investigator not directly involved with the data collection. Improving not just the accessibility but the interpretability of CAP data is an important goal for CAP V, which will beaddressed, in part, by presenting interactive visualizations of signature CAP datasets on the website. We envision that such visualizations will be useful exploratory tools for not only the scientific community but for K-12 students and citizen scientists as well.

*Expanded venues for data management training*: The graduate-level research data management course that Earl teaches through the ASU School of Sustainability has been a resounding success, earning strong reviews in each of the five years it has been taught. However, graduate students must maintain a careful balance among commitments, which often puts a semester-long course out of reach. As a result, Earl is working with the ASU Research Data Management Office and ASU Library to develop shorter, workshop-style (sensu Software Carpentry) educational module to provide CAP and other graduate students essential data management training that is available to them any time.