Supplementary Document – Data Management Plan

Introduction: The ARC-LTER Information Management (IM) supports and enhances long-term research by integrating data management from the planning to the collecting and analysis stages through to the publication of datasets and papers. Here we describe that sequence in terms of the responsibilities, planning, design, protocols, and formatting of our data management, and describe our interactions with related databases and finally the changes to IM we will make in this renewal proposal.

General information about the ARC-LTER project is provided on our web site, including site descriptions, past proposals and other documents, a site bibliography including publications from the project, educational opportunities, contact information for site personnel, and links to related sites.

Planning: Careful planning at the research design stage is required to ensure that any single set of measurements is easily linked to other measurements; typically, this includes working closely with collaborating projects to ensure that their research on LTER sites and experiments or with our long-term datasets is optimally integrated. In addition, to maximize data access both within the project and to other researchers, all datasets are available for downloading from online data portals as soon as data are error checked (usually before 2 years). We follow the LTER Network's IM Policy and NSF's Proposal & Award Policies & Procedures Guide for data archiving. Datasets are rarely embargoed and then only with well documented justification for exceptions and approval by the lead PI. Our data access and use policy is under the license: CC BY 4.0 Attribution.

Planning also includes careful review of the overall IM system, which we do at our annual ARC-LTER science meeting and at our semi-annual ARC Executive Committee meetings where we discuss current, specific needs or ways of improving the IM system and data accessibility. At the annual meeting we focus in particular on the upcoming field season and on planning our research for optimum integration of diverse datasets. All project personnel including postdocs, graduate students, and occasional REU students participate in these discussions. Because the ARC-LTER is spread across many institutions this is a valuable time for review of the IM system. Any new IM procedures or changes are incorporated into the system by the ARC-IM manager.

Responsibilities: A Senior Research Associate, Jim Laundre, is the overall project information manager (ARC-IM) with responsibility for overseeing the integrity of the ARC information system. He maintains the ARC-LTER web site, dataset catalog, and oversees the data submission workflow including archiving to the NSF repositories EDI (Environmental Data Initiative) and ADC (Arctic Data Center) (see Fig 1). Laundre attends the LTER Network Information Manager's video teleconferences and meetings and makes sure we stay up to date and compatible with Network data standards.

Specific IM responsibilities are distributed to each PI who is most familiar with the collection methods, data quality checks, and analyses (e.g., chemical or biological). Thus each PI oversees the data management for their part of the project to ensure that the processing stream of collection - analysis - verification - archiving is seamless, follows ARC-LTER standards, and is consistent from year to year. They also ensure that the technicians and students in their group understand and follow the LTER data management procedures and that the final data products are sent to the ARC-IM.

In response to the mid-term site review we began development of a series of training resources and guidance for better understanding of our IM system, contributing to the system, implementing the LTER Network data practices and policies, and using the LTER, EDI, and ADC data repositories to support research. The target audience for these resources is students, postdocs, and new PIs, but we recognize that these resources may be useful for other LTER sites or NSF projects. Implementing

this training of project personnel also will provide a level of redundancy on IM activities (i.e., the ARC-IM is not the only person who can operate the system) especially in submission and curation of datasets, without compromising the ultimate responsibilities of the PIs and ARC-IM for our data management.

Design and Processing: As stated in the previous section it is the responsibility of each investigator to submit datasets that conform to the LTER network data and metadata policies. An Excel workbook is available as a template with standardized sheets for data and metadata. Comments to guide the users who are submitting a dataset are used extensively throughout the metadata sheet. Data validation lists are used to create drop-down lists for units, sites, and data types. Once a file is submitted a basic check is made to ensure that the data and metadata conform to the LTER standards, e.g., LTER controlled vocabulary, core area keywords, and grant numbers. For researchers who do not use Excel, a word document is also available for entering metadata with the data being submitted as comma delimited ASCII. The Word Metadata Template we use is from EDI, and it has similar fields as the Excel template.

When the data and metadata checks are completed an EML metadata file is created and uploaded to the repository of the EDI (Fig. 1). Our previous system design used the DEIMS Drupal 7 content management software to provide MySQL data management, an ARC web site that hosted ARC documents, images, presentations, data, metadata, and a bibliography, and generated the files in EML format for uploading datasets to the EDI data repository. With Drupal 7 support ended we are moving to Drupal 9 to maintain our databases and web site. However, we have moved to using an R script (available from ARC-LTER GitHub) to parse the metadata from our Excel workbook template. This script creates the text template files required by EDI's EMLAssemplyline R script and then runs the EMLAssemplyline script to produce an EML file for uploading to the EDI portal. The Drupal 9 ARC web site will still provide information on the ARC-LTER project, including research sites, keywords, past documents, bibliography, and a dataset catalog. However, the data catalog will link to public data repositories where the full metadata and data can be downloaded. For datasets submitted in a Word template the ARC-IM will manually parse out the information needed to run the EMLAssemblyline R script.

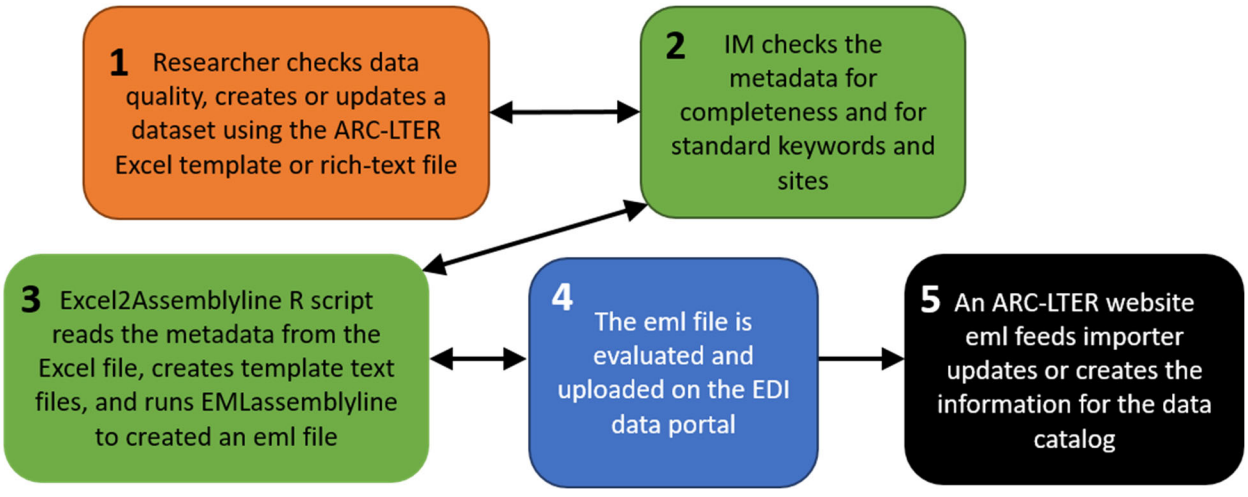

Figure 1. *The ARC-LTER dataset workflow.*

Computing and storage infrastructure: A virtual Linux server, maintained by the Marine Biological Lab (MBL), is used for the ARC Drupal web site. Backups of the Drupal MySQL database and files are made every night with weekly and monthly backups retained for 1 month and 1 year, respectively. The entire server is imaged and backed up within MBL's Veeam backups and disaster recovery platform every evening. Local backup of working data files at PI institutions occurs similarly. For backup and sharing of working documents (e.g., manuscripts, protocols) we use local servers, Google Docs, or Microsoft OneDrive as needed. The ARC bibliography is maintained in Zotero open-source online reference software which shares a group library with the LTER network Zotero library. Similar infrastructure is available at Columbia University.

Availability of Datasets: Datasets of the ARC-LTER are available from either our web site data catalog or the EDI data portal. Because EDI is a member node of DataONE our datasets are also available through DataONE's search page. We only ask those submitting that the datasets are properly cited and that NSF and the ARC-LTER are acknowledged in published papers and other data-related products. Data from the large-scale experiments and from routine monitoring are available online as soon as the data are checked for quality, and where necessary transformed for presentation in standard units and scales. Many datasets, such as weather observations, stream flow, and data that require little post-collection processing or chemical analyses, are available within a year of collection. Other data, particularly from samples requiring intensive chemical analysis in our home laboratories, may take up to two years before they appear online. Collaborating projects can and often do contribute their datasets to our online database, where we highlight them on our webpage, and if required these datasets can be replicated to the NSF ADC. No other ancillary data, software, or tools are necessary to access these data. No exceptional arrangements or access limitations (i.e., ethical, privacy, intellectual property, or copyright issues) are anticipated.

Dataset format, versioning, and quality control: As introduced above, PIs, technicians, and students who collect the data are responsible for data analysis, quality control (QAQC), and documentation. This ensures that the data are checked and documented by those most familiar with the research. While investigators may use any software for their own data entry and analysis, we expect that all documentation and datasets that are submitted conform to the required ARC-LTER and LTER Network formats. Scripts are used to check metadata and data for compliance with our protocols first by the senior research associates on the project and then by the ARC-IM (Fig. 1).

Versions. We annotate each data file or source code file with a version number. If data are altered in any data file at a later time, the version number is advanced, a note is made in the metadata file, and the updated file is transferred to the online data repository. Version control along with DOI identifiers are used to uniquely identify all data products that generate multiple versions as a result of reprocessing.

Quality control. ARC-LTER guidelines include checking data for errors starting with sample labeling in the field, initial data entry in the field, downloaded data from instruments in the field, analyses in the lab including quality-assurance quality-control (QAQC) procedures for machine operation and data outliers, and integration from a machine output into a final file format before transferring to a data archive. Data quality is checked initially by the person performing the analysis, then the lab manager, and finally by the PI before it is released to the ARC-IM for further processing and archiving (Fig. 1).

Toolik Field Station Environmental Monitoring Program: The ARC-LTER and its precursor projects have maintained an environmental monitoring program at Toolik Lake since 1975, including basic weather data (beginning in 1988) as well as stream and lake observations. In September 2006 the Toolik Field Station (TFS, operated by University of Alaska, Fairbanks) assumed responsibility

for maintenance and data management of the main Toolik weather station, which ARC-LTER had been supporting since 1988. Toolik Field Station weather data are available from TFS Environmental Data Center and from ADC. The ARC-LTER project is still responsible for collection and management of weather and other data collected from experimental plots and as part of LTER research. The TFS Environmental Data Center has additional observational components including plant phenology, snow cover, bird observations, and other year-round observations of weather and natural history that cannot be made by LTER personnel who are not year-round residents.

Geographic Information Systems, Mapping, and Remote Sensing: Geographic information from the Toolik Lake region is extensive, detailed, and linked to several key global and regional databases. Because much of this information system was developed with funding independent from the ARC-LTER project, we have focused our efforts on ensuring access to this valuable database and on optimizing its usability for the needs of our research and our collaborators.

(1) The Circumpolar Geobotanical Atlas, developed by Dr. Donald Walker and colleagues at the Alaska Geobotany Center, University of Alaska, features a nested, hierarchical series of maps of arctic ecosystems at scales ranging from 1:10 (1 m^2) to 1:7,500,000 (the entire Arctic), with multiple data layers at each scale including vegetation, soils, hydrology, topography, glacial geology, permafrost, NDVI, and other variables. Much of the development of this hierarchical system is based on original work done by Walker and colleagues at Toolik Lake and Imnavait Creek, with multilayer maps of these areas at 1:10, 1:500 (1 km^2), 1:5000 (25 km^2), and of the Kuparuk River basin at 1:25,000 and 1:250,000 scales.

(2) The Toolik Field Station GIS and Remote Sensing (TSF-GIS) service was developed with support from NSF Office of Polar Programs to help manage and support research based at TFS including ARC-LTER research. This GIS is maintained by a full-time GIS and Remote Sensing Manager and includes a multilayer GIS based largely on the Geobotanical Atlas data described above, combined with land ownership information, roads and pipelines, and disturbances. Particularly important for our purposes is a detailed map of research sites including all of the LTER experimental plots and sample locations in the upper Kuparuk region. We routinely use this database to guide new researchers at Toolik in collecting samples or establishing new experiments or research plots. The GIS includes a map of Inupiaq place names with annotations of historic use of the land by the Inupiaq people, along with a dictionary of plant and animal names and common words.

Newly developed is a *Research Plots Dashboard* where you can "review current and historic Toolik research project information, explore where data were collected over the years, and discover what was learned." Selecting individual points on the map produces a pop-up window providing hyperlinks to associated Principal Investigator(s) and project publications, and links to project data (when available). ARC-LTER worked closely with TSF-GIS to provide the links to our datasets for this Dashboard.

Another service offered by TFS-GIS is an Unmanned Aerial System (UAS) for RGB and NDVI imagery of study sites near Toolik. High resolution Digital Surface Models can also be created from the UAS imagery. We routinely request this service in support of our research, and our feedback to GIS managers helps to improve the products and services. The data products from these requests reside at TFS-GIS and are available to other researchers.

Anticipated changes for 2022-2028:

- We will move our website to Drupal 9 (a development site is currently in testing). The website is currently hosted at MBL but will move to Columbia University within the next year. At that time the site's content will be updated to improve information and data discovery. For example, we will use EDI's utilities to help build the data catalog and provide links to the full metadata and data. Drupal content types will continue to be used to provide sorting by keywords, investigators, research sites, and projects, in addition to providing a MySQL database for researcher information, controlled vocabularies, research sites and projects. Funds are available to facilitate this change and those listed below, and our current ARC-IM Laundre will be supported on the new grant and will be instrumental in the transition from MBL to Columbia.
- We will continue organizing and consolidating current datasets and make available older "legacy" datasets on the website. This review and update process will be guided by the FAIR Data Principles (Findable, Accessible, Interoperable, and Reusable). Many of the older datasets have outdated usage rights and keywords, and they will be updated to the current CC BY 4.0 and to keywords in the LTER controlled vocabulary.
- Our metrics for data usage will be improved by entering journal citations for data packages on the EDI Data Repository. Although older publications often do not specify what datasets were used, more recent papers do include data citations.
- We will continue to develop and expand our approach to processing data and metadata. R scripts are currently used to parse the Excel metadata template files and run EDI's EMLAssembleline R script to create EML files. To date we have developed R scripts to QAQC and analyze multi-year datasets and for display on ARC-LTER web site, and these improvements will be continued in the new grant.
- The ARC-LTER GitHub is currently hosting ARC-LTER R scripts and model information repositories. In the new grant we will add an ARC data handbook, website code, and other software or models developed at the ARC-LTER.
- We will continue our development of R scripts to QAQC and analyze multi-year datasets, and expand this useful tool for ARC researchers and collaborators.
- Many collaborating projects choose to or are required to use other data repositories (e.g., ADC, GenBank). Although these repositories are searchable through DataONE, it can be a challenge to uniquely identify ARC-LTER supported datasets. To address this challenge we will create EDI metadata for linked data packages (Data Package Best Practices). In addition, we will test methods of using keywords and grant numbers with a DataONE Custom search portal to identify ARC-LTER data across multiple data repositories. Currently TFS is using a DataOne Custom search portal. Finally, we will collaborate with TFS to ensure that our datasets are part of their search process.
- We have participated strongly in LTER IM network activities in our past projects, and we will continue that participation. New and continuing projects include: Hydromet (harmonized database of meteorological and hydrologic data in CUAHSI ODM CSV), EDI Unit Dictionary, contributions to EMLAssembleline GitHub, AKDatUMA (Alaska Data for Undergraduate Educational Modules), and the R package lterdatasampler.