Managing Idaho's Landscapes for Ecosystem Services (MILES)
Idaho EPSCoR Research Data Policy
Version 8.0 (October 12, 2015)

Introduction
The following policy statements have been developed and refined for the Idaho NSF EPSCoR Managing Idaho's Landscapes for Ecosystem Services (MILES) grant IIA-1301792. The National Science Foundation (NSF) requires establishment and enforcement of such policies as part of the new data management requirements of the Idaho EPSCoR program. These policies are intended to supplement standard legal constraints (such as Institutional Review Board [IRB] requirements); they are not intended to supplant or create conflicts with other institutional and legal requirements. These policies will be reviewed annually by the Cyberinfrastructure (CI) Working Group[1] in consultation with the MILES Executive Leadership Team (ELT)[2]. Recommended policy revisions will be submitted to The MILES ELT for final approval.

As benefactors and contributors in the EPSCoR Program we recognize that we are a team of scientists and administrators who have a collective obligation to maximize the utility of the granted resources through 1) active communication between project leaders, teams, the CI Working Group, and CI staff at the universities; 2) through collaboration, to more efficiently allocate and share resources (*e.g.* personnel, equipment, and software); and 3) recognition that raw data and subsequent data products generated internal to EPSCoR projects are a common good and priority should be given to provide mechanisms for access and sharing to the research community and the public as rapidly as possible according to the polices outlined below (*see also* the Idaho EPSCoR MILES Data Management Plan (IE-MDMP)) .

The new paradigm for research data management is for rapid and open access. The policy statements in this document set a trajectory towards this end for Idaho NSF EPSCoR MILES. These policies will require revision and vigilance before rapid and open access to data become the new norm in the collaborating institutions. We are committed to this pathway. However, one or several institution's new policies will not by themselves usher in a new data management paradigm and neither will "rapid and open access" to all research data occur without substantial new cyberinfrastructure investments to data management hardware, software, tools, personnel, etc.

Definitions:
Research Data
For the purpose of this document we have defined "research data" according to the definition proposed by the National Academy of Sciences, National Academy of Engineering, and Institute of Medicine in their 2009 publication "Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age." They define research data as "information used in scientific, engineering, and medical research as inputs to generate research conclusions." About research data in the publication it states:

> *It includes textual information, numeric information, instrumental readouts, equations, statistics, images (whether fixed or moving), diagrams, and audio recordings. It includes raw data, processed data, published data, and archived data. It includes the data generated by experiments, by models and simulations, and by observations of natural phenomena at specific times and locations. It includes data gathered specifically for research as well as information gathered for other purposes that is then used in research. It includes data stored on a wide variety of media, including magnetic and optical media.*
> National Academy of Sciences, National Academy of Engineering, and Institute of Medicine 2009, page 22

[1] The current CI Working Group consists of Nancy Glenn, Luke Sheneman, Donna Delparte, and Marisa Guarinello
[2] The MILES Executive Leadership Team is currently Shawn Benner, David Rodgers, John Anderson

It is the goal of the State of Idaho NSF EPSCoR program to make all such data available for common access and reanalysis by other researchers.

Types of Research Data

We define ***raw data*** as data as it was collected with no alterations. Examples include the dissolved oxygen readings from an *in* situ monitoring instrument that are stored on the instrument and downloaded as a complete file, the point cloud file generated in LiDAR collection, and the unabridged responses to a survey instrument. *Raw data* are rarely used "as is" in subsequent analyses the researcher uses to generate research conclusions.

We define ***processed data*** as the version of the data that is used in subsequent analyses the researcher uses to generate research conclusions. These may also be referred to as 'cleaned' data –the initial process of data quality assurance checks. Some examples of the way raw data are 'cleaned' include applying correction factors for biofouling on water quality monitoring instruments, correcting for cloud cover in remote sensing data, removing outliers deemed to be errors from LiDAR point cloud files, and removing sensitive information from survey results.

We define ***derived data products*** as data products that result from the analysis of processed data (including existing, available published data created by others). For example, the resulting dataset of a spatial analysis and model that uses existing elevation, precipitation, and land cover data to generate values of erosion potential would be considered a derived data product. Another example would be the calculated phosphorus load to a lake, derived and calculated from measurements of other variables (nutrients, flow, water levels). A third example would be trailheads classified for recreationists' behavior based on the aggregation of interview results. Often intermediate data products are created in the process of creating these derived data products; MILES researchers are not responsible for making these available, although their existence should be identified in the processes steps or methodology recorded for the final derived data products. For the purposes of MILES, only derived data products that answer specific research questions and will be part of peer-reviewed publications or used for outreach efforts need to be documented and made accessible.

Research Products

We define ***research products*** as those unique intellectual contributions to the accumulation of knowledge under the MILES project that cannot be defined as research data (as defined above). These include models, tools, scripts, code, and visualization products. Only those products that are new contributions should be shared as MILES research products. An existing tool that is used by a MILES researcher to generate data would not meet this definition. For example, an existing ArcToolbox geoprocessing script that was used to generate a derived data product would not fit in our definition of a research product, but a collection of newly written Python scripts packaged as a new ArcToolbox would.

Within the above definition, we define ***models*** as those models that are new, unique modeling environments created under the MILES grant. These may include existing models that have been calibrated or refined for use with data and parameters specific to the study area or research question or they may include entirely new modeling environments. This definition is not meant to include statistical modeling routines (e.g., linear regression, cluster analysis, etc.) that can be performed using any number of methods and applications.

Documentation and Sharing

We adopt the following definitions from Hampton et al. 2015[3]:

[3] Hampton, S. E., S. S. Anderson, S. C. Bagby, C. Gries, X. Han, E. M. Hart, M. B. Jones, W. C. Lenhardt, A. MacDonald,W. K. Michener, J. Mudge, A. Pourmokhtarian, M. P. Schildhauer, K.H.Woo, and N. Zimmerman. 2015. The Tao of open science for ecology. Ecosphere 6(7):120. http://dx.doi.org/10.1890/ES14-00402.1

code repository: an accessible, central place where computer code is stored to facilitate the collection, manipulation, analysis, or display of data

data repository: an accessible, central place where accumulated files containing collected information are permanently stored; typically these house multiple sets of databases and/or files

open access: providing free and unrestricted access to research products, especially journal articles and white papers—to be read, downloaded, distributed, reanalyzed, or used for any other legal purpose— while affording authors control over the integrity of their work and the right to be acknowledged and cited (adapted from the Budapest Open Access Initiative definition, Chan et al. 2002[4])

open data: data that can be freely used, reused, and redistributed without restrictions beyond a requirement for attribution and share-alike (Molloy 2011[5])

open source: computer code (software) that is available for free distribution and re-use, with source code unobscured, and explicit acknowledgement of the right to create derived works by modifying the code (Gacek and Arief 2004[6])

We define ***data embargo*** as the process by which access to data is restricted for a certain amount of time before it is made available as open data. This process is initiated when complete metadata for the data is published and the data are stored in the restricted section of the data repository. This process will allow MILES participants to meet the requirements of sharing data in a timely manner while also meeting the researcher's need to have the first right of publication.

Data Management Plan

Whereas, this document sets forth the policies to be followed by participants in the Idaho NSF EPSCoR MILES grant, the Idaho MILES Data Management Plan (IE-MDMP) goes into more procedural detail and addresses individual's roles and responsibilities. The IE-MDMP describes in detail the organization of the cyberinfrastructure team, appropriate metadata standards and metadata management, quality assurance and quality control (QA / QC), matches data and research product types with approved repositories, and a discussion of supporting tools, technologies, and resources for investigators to manage data and meet the demands of this IE-Research Data Policy (IE-RDP). Additionally, the IE-MDMP document includes details on application, use, and integration of other efforts to address research data management regionally and sustainably.

The CI Working Group and CI staff in consultation and cooperation with the MILES ELT establishes the list of approved repositories for EPSCoR funded data in the IE-MDMP. Approved repositories may be 1) only publicly accessible, 2) with limited access, or 3) a combination of both.
Approved repositories will have established mechanisms in place to make research data and products easily discoverable and free of charge. Local and regional initiatives from participating universities and other professionally sanctioned repositories exist as well as do many useful CI technologies and tools. The CI Working Group / CI staff / ELT may review and either approve or reject repositories, CI technologies, and tools recommended by project researchers.

RESEARCH DATA POLICY STATEMENTS

[4] Chan, L., et al. 2002. Budapest Open Access Initiative. http: / /www.opensocietyfoundations .org/ openaccess/read
[5] Molloy, J. C. 2011. The Open Knowledge Foundation: open data means better science. PLoS Biol 9:e1001195.
[6] Gacek, C., and B. Arief. 2004. The many meanings of open source. IEEE Software 21:34–40.

1. Communications and Consultation Policy Statement

NSF EPSCoR MILES PIs, Co-PIs, ELT, and the CI-Working Group have a responsibility to routinely communicate and consult in order to 1) assess and record the status and effectiveness of CI expenditures (*e.g.* hardware, software, tools, and personnel); 2) coordinate and develop strategies that maximize the cumulative effect of CI expenditures and promote sustainability; and 3) define appropriate repositories for Idaho EPSCoR data and research products.

2. Researcher Accountability and Additional Responsibilities Policy Statement

Researchers are expected to generate a broad array of scientifically robust high-quality data suitable for further use. Individual researchers are responsible for data management throughout the data life-cycle, documenting data lineage and properly protecting the data. Researchers who generate data that require additional protections, such as research that is subject to IRB review, are responsible for communicating and requesting any additional assistance needed from the CI Working Group in proper curation of these data. It is the responsibility of individual researchers to contribute their raw, processed, and derived data products, along with valid standards-based metadata, in a timely manner as explicitly described below. The institutional leads from the ELT are responsible for monitoring and following up with individual researchers who need to provide data. The State EPSCoR Director has the responsibility for enforcement. Any barriers[7] will be clarified and documented by the institutional lead and approved or denied by the EPSCoR Director. If the EPSCoR Director is unable to resolve the issue, the EPSCoR Director will communicate the problem to the appropriate Vice President of Research at the institution where the researcher resides.

There are existing primary and secondary data available from State/Federal agencies and prior EPSCoR projects, most of which are available from various data outlets (listed in IE-MDMP). Idaho EPSCoR recognizes the value of such data. Researchers are encouraged to explore and make use of the available data before committing to primary data collection effort.

3. Data Formats and Metadata Requirements

The format of deposited data, raw, processed and derived, should conform to open, non-proprietary and documented standards, where applicable. When data and data products cannot be converted to such formats after a reasonable effort, researchers will provide sufficient documentation about the data format and software needed to access the data. For example, Excel spreadsheets should be converted to comma separate files that are not bound to proprietary software and are compatible with many data analysis routines (e.g., R).

All research data and research data products must be documented and cataloged with appropriate, complete metadata. Shared data without adequate, corresponding metadata has limited intrinsic value and is incomplete. IE-MDMP will maintain a list of Idaho EPSCoR approved metadata standards.

Shared data should be documented with one of these metadata standards. Researchers are encouraged to facilitate the adoption of international standards (ISO 191***) for geospatial metadata, in compliance with FGDC recommendation on transition to ISO metadata standards. The ELT must approve the use of alternative metadata standards. Upon request, the CI Staff will provide necessary documentation, coordination, and training for effective use of these metadata standards.

[7] See Appendix – Potential Barriers to Sharing Data and Procedures for Policy Compliance.

4. Timeliness of Data and Research Products Contribution Policy Statement

Research data and research products created by MILES participants, as well as other supporting materials, should be shared within a reasonable time and in accordance with data sharing policies from the National Science Foundation (NSF) (the National Science Foundation Award & Administration Guide (AAG) Chapter VI.D.4[8]). Research data and research products created by MILES participants will be shared with two audiences: 1) other MILES participants and 2) the general public. Sharing with the former is instrumental to successful collaboration within this large statewide project and sharing with the latter makes data and knowledge available to the taxpayers that fund the work through NSF and is instrumental in meeting the MILES goal to "… establish the infrastructure to provide science-based decision support needed to sustainability management Idaho's resources." Providing data and research products to both these audiences is in accordance with the open science principles this policy embraces.

Details on when and how research data and products should be shared with both audiences are provided in the IE-MDMP, along with additional guidelines for graduate students and post-doctoral scholars. For all types of data and research products, researchers should make every effort to document these data with appropriate metadata in approved repositories early in the process; it is appropriate and useful to provide metadata before the data are ready for public distribution. Additionally, researchers may request a data embargo when they submit data to an approved repository and also request a DOI. At the time metadata describing the data and the data itself are uploaded to one of the approved public repositories (or a valid link to the location of the data is provided as an online resource for the data, such as data offered up as web service rather than file download), the requirements for sharing publicly within the time frame specified in the IE-MDMP are met, regardless of if an embargo is requested. Using the **data embargo** process and DOIs are the best ways for researchers to meet the MILES policy requirements for sharing data and also to protect the first right to publish manuscripts using that data.

5. Researcher Expectation of Infrastructure Policy Statement

Researchers have a right to expect 1) clear procedures for data cataloging and depositing their data; 2) unrestricted access to edit and modify their data or records about their data (*i.e.*, metadata); 3) the availability of tools for cataloging and depositing their data, or in the absence of tools, staff who can assist them with data curation; 4) a secure and well maintained repository for their data; 5) adequate mechanisms for data discovery and access by others; 6) that they, as creators of data, have first rights to analyze and publish those data; and 7) to be credited for their data contributions[9].

6. Research Data Citation Policy Statement

After data have been deposited and cataloged in an approved repository (as defined in the IE-MDMP), users of the data should appropriately acknowledge the National Science Foundation, Idaho EPSCoR and the individual investigators responsible for the data set. Any use of data provided by the Idaho EPSCoR must acknowledge Idaho EPSCoR and the funding source(s) that contributed to the collection of the data. Any restrictions on the usage of the data shall be clearly stated and obvious to the potential data user. All derived data products built from external data sources will cite the external data sources and collectors in the metadata for the new derived product.

---

[8] The National Science Foundation Proposal Award Policies and Procedures Guide: http://www.nsf.gov/pubs/policydocs/pappguide/nsf14001/nsf14_1.pdf

[9] See also the data sharing policy outlined by the National Science Foundation at http://www.nsf.gov/geo/ear/EAR_data_policy_204.pdf.

Until systems are integrated, individual repositories may set their own citation policies. These policies should be easily discoverable and honored by data users. If such policy is not available, data users should follow the standard guidelines available at http://www.datacite.org/whycitedata and properly attribute the data creator. In addition, the appropriate statement to be used when acknowledging data that was generated under support from Idaho EPSCoR is: "… data were provided by (Name, University Affiliation) through the support of the NSF Idaho EPSCoR Program and by the National Science Foundation under award number IIA-1301792."

Compliance with Federal and State Privacy Regulations

- **Protection of Personal Financial Information:** The Federal Trade Commission's Safeguard Rule and the Financial Services Modernization Act of 1999, also known as the Gramm-Leach-Bliley Act (GLBA) requires institutions of higher education to implement administrative, technical, and physical safeguards for certain types of nonpublic personal financial information. Therefore, all deposited research data shall be separated from any data with nonpublic personal, financial, or confidential content.
- **Personal Medical Information:** All Idaho EPSCoR researchers must comply with all applicable Federal and State regulations concerning the privacy and security of personal medical information, including but not limited to the Health Insurance Portability and Accountability Act (HIPAA) and the Health Information Technology for Economic and Clinical Health Act (HITECH).
- **Personal and Sensitive Socioeconomic information:** All researchers will adhere to Federal OMB Circular A-110 guidelines on not recording personal and sensitive socioeconomic information, when such disclosure would constitute a clearly unwarranted invasion of personal privacy, such as information that could be used to identify a particular person in a research study.
- **Sensitive Biological and Ecological information:** All researchers will adhere to federal and state wildlife laws and policies concerning release of sensitive information such as the precise location species have been observed. For example, locations of endangered species or other data protected by the Endangered Species Act will not be shared and similar restrictions may apply to data maintained by states in natural heritage databases.

7. Intellectual Property and Ownership Policy Statements

Individual researcher, within the policy of their relevant institutions, will maintain the ownership of copyright and intellectual property of data and data products. Researchers and their universities are responsible for ensuring compliance with the state and federal policy and laws. Idaho EPSCoR will not be liable for any legal action taken against the researcher for copyright or intellectual property right infringement related to the deposited data or data products.

Researchers are encouraged to have a citation record created with a Digital Object Identifiers (DOI) for their data. The Northwest Knowledge Network at the University of Idaho, and potentially other approved repositories, has the ability to issue DOIs for stored datasets. Researchers are encouraged to request DOIs for their novel and high-quality datasets to ensure the visibility and permanence to the dataset. Assigning DOIs will allow users to cite research data and products in the same manner as the journal articles, giving appropriate credits and recognizing the time and effort of the researchers in creating and managing the dataset.

8. Consideration of Additional Policies

Finally, we are sensitive to the potential effect that these policy statements may have throughout academia. We are particularly aware of potential deleterious consequences to junior academic research faculty if the data they collect are not adequately cited and if there is not institutional consideration given to these contributions with respect to tenure and promotion. To promote the new paradigm for

rapid and open access to research data being advocated by NSF, the need for additional institutional policies are currently in progress. For example, the VPRs of the universities signed an MOU in 2011 for Coordinated CI and Data Management and a statewide CI Advisory Council meets regularly to facilitate coordination among institutions.

APPENDIX -- Potential Barriers to Sharing Data and Procedures for Policy Compliance

A detailed survey of cyberinfrastructure users within the RII community and experiences of Idaho researchers in Environmental Observatory networks have identified the following *potential* barriers to the contribution of shared data resources. This is not meant to be a comprehensive list. Special cases that fall outside the scope of this list should be brought before the ELT.

- *Type I. Inadequate resources to upload and provide QA/QC for data*

  A pervasive problem has been the underestimation of the time required to assemble and upload data and associated metadata. This is particularly true for legacy data, which may be in paper form, or require highly qualified researchers to screen data, identify what is important, assess the quality of data, and decide how to report the information.

- *Type II. Lack of training in data preparation*

  Several emerging data management systems, such as HIS, require training in uploading data, development of scripts to accelerate this process, and learning correct terminology.

- *Type III. Sensor Failure or inability to capture required information*

  The risk for this type of data loss is highest when using experimental sensor technology or when intended surrogates do not correlate well with the required parameter.

- *Type IV. Refusal to share data or failure to deliver data in a timely fashion*

  This can arise for many reasons including personality conflicts, concerns about data being published by third parties before being published by the research team that collected the data, and time/resource constraints.

The barriers described above are not unique to Idaho EPSCoR participants, but represent the broader research community. Solutions should be general in nature and can apply to other research data programs in the State of Idaho. The current approach that is in place and being monitored for effectiveness is described here. Its implementation relies on continuing commitment to EPSCoR DM/CI goals by many participants. Although specific responsibilities are assigned to particular individuals and groups mentioned previously, all can play an active role in monitoring the effectiveness of this policy and recommend changes as necessary. Ongoing education for faculty will also be needed to foster their willingness to advance the EPSCoR DM/CI agenda. The implementation of this plan comprises:

1. Data management is a standing agenda item on the ELT monthly meetings. The data flow within the RII will be reviewed, and any anticipated or current difficulties will be identified. The cause of the difficulty and the RII researchers likely to be affected will be documented in this discussion and a mitigating action item recorded and distributed to the team. The CI-WG has primary responsibilities for reporting the status of CI activities and may be the first to identify issues related to the policy and bring them to the ELT.

2. The institutional leads from the ELT are responsible for monitoring and following up with individual researchers who need to provide the data. Any barriers will be clarified by the institutional lead.

3. Early communication of potential or actual problems is essential to ensure that members of the research team are not disadvantaged through missing or late information. The CI-WG and CI staff at respective universities will be responsible for anticipating and coordinating these difficulties in partnership with the ELT.

4. Planned approaches for when *Type I-IV* barriers are encountered:

    a. *Type I or Type II Barriers*: a brief request will be prepared for the Idaho EPSCoR Office to identify appropriate additional training or needed resources. Decisions on the scope of legacy data to be included within the time and budget will be made within the scientific theme teams.

    b. *Type III Barriers*: the difficulties will be communicated to all participants with a plan for modifications to the data collection program in the following year. These modifications will include contingency plans should sensor problems be encountered again.

    c. *Type IV Barriers*: these issues are the most difficult to resolve and the most difficult to generalize. If the institutional lead from the ELT cannot resolve this issue, it will be passed to the State EPSCoR Director whose responsibility will be to assess the implications to the goals of the RII and the broader scientific community. S/he will consider extenuating circumstances and develop a course of resolution. Measures such as relief from teaching or other institutional duties to provide the time to address the issue, removal of individuals from the project, elimination of future funding or disciplinary action shall be conducted by the State Director in close consultation with the VPR of the relevant University.

5. The ELT will keep track of the results of action items, and will review them at the next ELT – CI-WG meeting. Additional measures will be taken as needed until a timely resolution is reached.