Managing Idaho’s Landscapes for Ecosystem Services (MILES)

Idaho EPSCoR Research Data Management Plan

Version 5.0 (October 12, 2015)

The MILES Data Management Plan provides procedural detail related to the policies outlined in the MILES Idaho EPSCoR Research Data Policy (IE-RDP).

1. Types of Data

This project leverages existing national, regional, and institutional data collections while also acquiring substantial new ecological, biophysical, and socio-economic data. Broadly, we anticipate using the following types of *existing* data products:

- Ecological/biophysical data at local and regional scales including state and federal databases on hydrological flow, river floodplain management, in-stream flow policies, and water quality.
- Collections of vegetation management data (fire history, exotic/invasive species increases).
- Secondary data sources including demographics (US Census Bureau), agricultural prices and costs of production (USDA surveys), land values (local tax assessor), water rights (Idaho Department of Water Resources), and parcel level land use data.
- Hydrological and landscape databases developed at University of Idaho, Boise State, and Idaho State via NSF funding (*e.g.* Dry Creek experimental watershed through CUAHSI HIS).
- Climate data via observations and modeling across multiple scales.
- Digital elevation and other remote sensing data and data products.
- Soil Survey Geographic Database (SSURGO) distributed by the Natural Resources Conservation Services (NRCS)
- Bathymetry data for lakes and ponds
- Historic and current local weather data
- Historic and current snow water equivalent data (depth, precipitation, etc.)

We will *acquire* a diverse set of new and biophysical, ecological, and socio-economic data:

- Remote sensing imagery of land use and land cover including LiDAR.
- Quantitative/qualitative socio-economic data from structured surveys, interviews, narratives, focus groups, and values mapping, social, cultural, economic, market values and costs, demographic analysis of migration flows, and migration preferences.
- Surface/ground water sample data for water resource risk determination.
- Geophysical data including river discharge, local runoff, groundwater, and geological layers including soils and vegetation.
- Plant/vegetation characteristics (invasive/introduced plants), seed dispersal, pollinators, noise disturbance, soil changes associated with exurban development and conversion of open land.
- Responses of plant-animal-soil-water interactions and ecosystem nutrient cycling to change in post-fire communities associated with land use scenarios.
- Geospatially explicit social data derived from engagement with individuals in communities combined with cross-cultural traditional knowledge.

The project will *generate* wide range of data, data derivatives, tools, model outputs, and visualization products:

- Remote sensing and geospatial data products such as classified maps, improved elevation models, etc.
- Model simulation results (*e.g.* population growth model, hydrological model, crop choice model, etc.)
- Socioeconomic and policy in reports and/or tabular forms
- Tools and software (*e.g.* data visualization)

2. Data Formats and Metadata Standards

Project participants will use well-defined data and metadata formats aligned with existing data sharing networks (*e.g.* DataONE, Northwest Knowledge Network (NKN), ISU's GIS TReC) and the data sharing policy developed under the previous EPSCoR RII. Researchers will document their research data and products throughout the course of the project using appropriate metadata standards as described below.

2.1 Data Formats

This project requires use and management of a heterogeneous set of domain-driven data formats. Researchers will use standard, published data formats wherever possible, including:

- *LiDAR* – LiDAR data will be stored using the Common LiDAR Data Exchange Format (.LAS). Derived products (e.g. DEMs) will use standard formats (*e.g.* GeoTIFF) or be exposed via web services.
- *Geospatial data* – Leverage existing support for ESRI GIS server software by supporting common ESRI file formats for raster/gridded data (NetCDF, HDF, image formats, etc.) and vector formats (*e.g.* ESRI shapefile, file geodatabases , or simple ASCII files). For geospatial web services, researchers are encouraged to use open and standard-compliant formats, such as Open Geospatial Consortium (OGC) standards for Web Feature Services (WFS) and Web Map Services (WMS).
- *Time Series Point Observation Data* – Researchers will use existing institutional installations of the CUAHSI Hydrologic Information System (HIS) for hydrologic observation data. Non-hydrologic types of fixed-point observation data may use a combination of tabular formats (e.g. .CSV) or custom databases as needed but will be accompanied with appropriate descriptive metadata to facilitate re-use and interoperability.
- *Tabular* – Esp. SES data (*e.g.* surveys, interviews, etc.). We support appropriate common, standard tabular file formats such as CSV, XLS, or databases accompanied with sufficient metadata. Many social science survey collection and analysis tools (e.g., SPSS, Survey Monkey) have the capability to export results as tabular data in these forms. Many tabular social science data are accompanied by codebooks (a version of metadata) and we expect that these would be included (likely in PDF format) along with the tabular data that would be published.

2.2 Metadata Standards

Project participants will document research data and products using a core set of metadata standards as described in the Idaho EPSCoR *data sharing policy*. Where applicable and possible, researchers will use the ISO 19115-2 or ISO 19115 2003 geospatial metadata standard, however we support the following core set of metadata standards:

- *ISO 19115-2*: General purpose; most applicable to geospatial, gridded, and imagery data and services.
- *Content Standard for Digital Geospatial Metadata (CSDM) metadata standard from FGDC (Federal Geographic Data Committee)*: A federally developed legacy geospatial metadata standard. Soon deprecated by ISO 19115; still widely used for geospatial data, esp. by agencies.
- *CUAHSI WaterML/ODM*: Site-specific hydrologic data (*e.g.* streamflow, water quality, etc.) to be stored in CUAHSI HIS which tracks site and observation specific metadata.
- *Ecological Metadata Language (EML)*: Specifically for documenting ecology data products.

- *Geographic Markup Language (GML/ ISO 19136)*: an XML encoding for the storage and transactions of geographic information
- *Dublin Core/XML*: Generic resource descriptions enabling discovery of resources (documentation, images, multimedia); applies to SES data.
- *Darwin Core/XML*: Used for documenting geographic distribution of species and specimens in collections.
- *NetCDF Climate and Forecast (CF) Metadata:* A set of metadata conventions for array-oriented scientific data (NetCDF, network Common Data Form).
- *Data Document Initiative (DDI):* A metadata standard for describing socioeconomic and behavioral data, particularly survey-based data.

Metadata standards are not available for models, visualization products, or code as they are for data as described above. However, it is important to document and make accessible the entire body of MILES work, which includes a great deal of these types of products. One of the above broad formats, such as Dublin Core, may be recommended for use with these products. Documentation for models and visualization products should also include information about the computing environment: the version number of the software used, the operating system used, and, where appropriate, input files specifying parameter values and other settings should be included with the documentation. The CI Working Group will consult and work closely with the scientists generating these products to make sure they are well documented and made discoverable through the same means as other data products.

3. Research Product and Data Management Repositories

Idaho EPSCoR identifies and promotes the use of the following existing data management repositories:

- The Northwest Knowledge Network (NKN) – a regional data management system that provides storage, retrieval and protection services across the life cycle of data.
- Interactive Numeric & Spatial Information Data Engine (INSIDE) Idaho – Idaho's statewide geospatial clearinghouse for archiving and distributing geospatial data.
- Idaho State University GIS Training and Research Center (GISTrec) Data Services – geospatial data repository and data services, with regional focus on eastern Idaho.
- Boise State University Data Repository – data and metadata management infrastructure to enable discovery and sharing of research data.
- The Consortium of Universities for the Advancement of Hydrologic Science (CUAHSI) Hydrologic Information System (HIS) network in Idaho – provides web services, tools, standards and procedures that enhance access to more and better data for hydrologic analysis.

Idaho EPSCoR is also actively involved with other existing and future data management efforts and repositories. Research faculty will be notified as they become available. Additional repositories and data management efforts include the following:

- The Long Term Ecological Research (LTER) Network – a collaborative effort of 26 sites representing diverse ecosystems and research emphases for investigating ecological processes over long temporal and broad spatial scales.
- Data Observation Network for Earth (DataONE) – poised to be the foundation of new innovative environmental science through a distributed framework and sustainable cyberinfrastructure that meets the needs of science and society for open, persistent, robust, and secure access to well-described and easily discovered Earth observational data. Idaho's NKN repository is a member node of DataONE.

- Critical Zone Observatory (CZO) Data – publication, sharing and integration of heterogeneous data collected at CZO sites
- Interuniversity Consortium for Political and Social Research (ICPSR) – an international consortium of over 700 academic institutions and research organizations; it maintains a catalog of more than 500,000 files of research in the social sciences. It hosts 16 specialized collections of data in education, aging, criminal justice, substance abuse, terrorism, and other fields.

For use as a limited access repository for sharing preliminary data and research products among MILES collaborators, Idaho EPSCoR and the ELT approves use of NKN's OwnCloud service.

Idaho EPSCoR identifies and promotes use of the following guidelines for sharing models, code, visualization products, etc.

All models should be registered with an online model repository appropriate to the discipline and type of model. Models should be registered by individuals using their own accounts rather than using an institutional account. Links to these records will also be provided in the NKN portal. The following repositories are approved for use. The additional of other repositories can be added as necessary, if requested by a MILES-affiliated researcher and approved by the CI Working Group and Executive Leadership Team. All source code, tools, and applications must include adequate documentation and/or within code comments to understand their functionality. The following repositories are approved for use:

- OpenABM – a node in the Computational Modeling in Socio-Ecological Sciences (CoMSES) Network for agent-based models (https://www.openabm.org/)
- Community Surface Dynamics Modeling System (CSDMS) – a repository for spatially explicit models (http://csdms.colorado.edu/wiki/Main_Page)
- Systems Dynamics Case Repository – a repository for system-dynamics models (http://cases.systemdynamics.org/)
- GitHub – a repository of open source code, models, and tools (http://github.com)

4. Data Sharing, Privacy and Intellectual Property

In recognition of the NSF's commitment to the principle that the various forms of data collected with public funds belong in the public domain, we have adopted a data sharing policy (IE-RDP) that facilitates the process of making data that has been collected with NSF support available to other researchers and to the public. Idaho EPSCoR adopts the following specific policies and practices related to the timeliness of sharing data and research products, and to appropriate citation.

For sharing with other MILES researchers:

- **Processed data:** All processed data products will be made immediately available to other Idaho EPSCoR researchers via the limited access repository approved by the ELT these data may should be considered provisional and the researcher is responsible for including appropriate documentation (e.g., ReadMe.txt file) with these data to ensure that collaborators are aware of the provisional status of the data and the appropriate contact person.
- **Raw data**: Researchers may choose to include raw data with the processed data they provide in the limited access repository. However, each researcher should be prepared to share and discuss raw data along with processed data at the request of other MILES participants.
- **Derived data products:** All derived data products will be made immediately available to other Idaho EPSCoR researchers via a limited access repository approved by the ELT.
- **Research Products:** All research products should be made immediately available to other

Idaho EPSCoR researchers via a limited access repository approved by the ELT. Appropriate documentation should be included. Researchers who use and document their products on GitHub or similar sites may elect to provide links to these in a text document for other researchers to access, as an intermediate step before documenting each tool with complete metadata.

For sharing with the public:

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 below 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.

- **Processed data:** At a minimum, researchers will catalog and deposit processed data and corresponding metadata in a public access repository approved by the ELT (see Section 3) no later than **one year** from the time of the raw data acquisition or collection (e.g., date of field survey); when the data is deposited a data embargo period may be requested. As an alternative to providing separate files for raw and processed data, the researcher may provide raw data along with code/scripts that generate the processed version of the data.
- **Raw data**: At a minimum, researchers will catalog and deposit raw data and corresponding metadata in a public access repository approved by the ELT (see Section 3) no later than **one year** from the time of the raw data acquisition or collection (e.g., date of field survey); when the data is deposited a data embargo period may be requested. Researchers are encouraged to include the raw data with the processed data product and metadata upon submission to a public access repository. In cases where the raw data files are large (>3TB) and the researcher has concerns about providing raw data files of this size, the researcher will consult NKN and the MILES Data Manager for a workable solution. As an alternative to providing separate files for raw and processed data, the researcher may provide raw data along with code/scripts that generate the processed version of the data.
- **Derived data products:** All derived data products will be made publicly accessible via an approved public access repository (see Section 3) no later than **two years** from the time of their creation.
- **Research Products:** All research products will be made publicly accessible via an approved public access repository (see Section 3) no later than **two years** from the time of their creation.

Considerations for graduate students and post-doctoral scholars:

Graduate student and post-doctoral scholars who received funding from MILES to conduct research that results in the creation of research data and research products must ensure that these data and products are fully documented and uploaded to an approved public access repository (see IE-MDMP) **prior to graduation or end of the post-doctoral period.** These researchers, by nature of their positions, will only be affiliated with the participating institutions for a temporary length of time and, therefore, must build the timeline for sharing their data and products based on both the timeframes outlined above and the length of their own tenure at their institution. Graduate students in particular are encouraged to include this timeline in their data management plan that they develop with their advisor. The timeline and responsibilities for data documentation and sharing will differ depending on the type of research conducted (e.g., field collection v. modeling). Graduate students/post-docs and advisors are encouraged to include data management are a topic in their scheduled check-in meetings, proposal defense, and at then end of each semester, etc.

Use of the **data embargo** process will ensure that data is stored in an approved public access repository (see IE-MDMP) so that MILES can meet its data sharing and management goals and so that these researchers will still have time to submit manuscripts for publication after their graduate and post-doc period is complete.

A hypothetical timeline for a graduate student pursuing a Masters degree and collecting data in the field is as follows:

- Year 1
  - 1st semester
    - Plan research timeline
    - Learn MILES data requirements
    - Write data management plan
  - 2nd semester & summer
    - Research – collection and processing of data (documentation with Intermediate Metadata.docx or similar)
- Year 2
  - 1st semester
    - Share – Make sure processed and raw data are provided to MILES on OwnCloud
    - Research – analysis of data (documentation with Intermediate Metadata.docx or similar)
    - Revisit and if necessary revise research timeline and data management plan
    - Proposal defense
  - 2nd semester
    - Research – analysis and finalization of data and related products
    - Share – Make sure derived data and products are provided to MILES on OwnCloud
    - Documentation – Write final metadata for data and products (use NKN's Metadata Editor or similar) (can be concurrent with Share step below)
    - Share – Upload data and products along with final metadata to approved repository (NKN or other) and request data embargo (can be concurrent with Documentation step above)
    - Write and defend thesis
    - Ensure that all data and products are discoverable through NKN's portal and/or MILES website

5. Documentation, Privacy and Intellectual Property

Metadata: All data will be described and catalogued with appropriate, complete metadata using an explicit common core set of metadata standards, as specified in Section 2.2.
Data Citation: After data have been deposited and cataloged in an approved repository (see Section 3), users of the data should appropriately acknowledge the National Science Foundation, Idaho EPSCoR and the individual investigators responsible for the data. Any use of data provided by 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.
Privacy and Confidentiality: We are explicitly compliant with federal and state laws surrounding data privacy including the protection of personal financial information through the Gramm-Leach-Bliley Act,

personal medical information through HIPAA, HITECH, and other regulations.
Human Subject Data: All such data (*e.g.,* surveys) will be collected and managed only by personnel with adequate human subject protection certification.

6. Data Management as a Training Component

Students engaged in RII will be required to design an individual data management plan as part of their research project. Faculty members involved in the modeling and cyberinfrastructure components will advise students in the development of effective, project specific data management plans. This requirement will contribute to the professional development of both the students and their faculty mentors. The CI Working Group and the MILES data manager will provide guidance and training to faculty and students to facilitate the creation of these data management plans, as well as the writing of metadata and other data management practices. It is the hope that these guidance and training materials will empower all MILES affiliated researchers, faculty, and leadership to fulfill their roles and responsibilities in terms of data documentation, sharing, and publishing as described in the MILES Research Data Policy and Data Management Plan.

7. Plans for Long Term Archival and Curation of Data

All raw and derived data used and referenced in publications produced through this RII proposal will be archived, curated, and made accessible into the future via established data archival networks. We will extensively leverage previous EPSCoR funding in the Northwest Knowledge Network (NKN) data repository (UI/INL) to host, archive, and curate data/metadata within Idaho. As NKN is a regional member node to NSF DataONE, RII data may be archived and accessible at national and international levels via DataONE. Other data archives may also be used as indicated in Section 3. Models, visualization products, and other tools developed with be documented and made available as described in Section 2.2 and 3. These products should all be open source where possible. In keeping with NSF policies, all source code should be open source.

8. Revisions

This data management plan will be reviewed annually and revised as needed by the RII project data manager.