Data Sharing Plan

Version 1.1 (February 14, 2019)

INTRODUCTION

Data sharing promotes many goals of the Idaho NSF EPSCoR Genes to Environment: Modeling, Mechanisms, and Mapping (GEM3) award OIA-1757324. Data sharing allows scientists to expedite the translation of research results into knowledge, products, and procedures to improve human-environment systems.

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. We do this through 1) active communication between project leaders, research teams, and the Data Management (DM) Working Group[1]; 2) through collaboration, to more efficiently allocate and share resources (*e.g.* personnel, equipment, and software); and 3) recognition that 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.

The new paradigm for research data management is for rapid and open access, and this document outlines a trajectory towards this end for Idaho NSF EPSCoR GEM3. However, one or several institution's 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 support from researchers like you.

[1] The DM Working Group consists of one Data Manager from each of the Idaho universities (BSU, ISU, and UI) and any other GEM3 participates who is interested in attending.

TABLE OF CONTENTS

DEFINITIONS

Research Data

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.*[2]

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, 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 to generate research conclusions.

We define ***processed data*** as the version of the data that is used in subsequent analyses to generate research conclusions. These may also be referred to as 'cleaned' data – those that have undergone the initial process of data quality assurance checks. Some examples of the way raw data are 'cleaned' include 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 trailheads classified for recreationists' behavior based on the aggregation of interview results.

Any of the above types of research data can be classified as sensitive. ***Sensitive data*** is data from human subject, student records, health outcomes, finances, personally identifiable information (PII), the location of endangered species, or the location of historical artifacts. If you work with these types of data, additional local, state, and federal laws may apply. Please contact your GEM3 data manager for assistance with secure storage and de-identification.

[2] National Academy of Sciences, National Academy of Engineering, and Institute of Medicine 2009, page 22

Non-data Research Products

We define ***research products*** as those unique intellectual contributions to the accumulation of knowledge under the GEM3 project that cannot be defined as research data (as defined above). These include models, tools, scripts, code, and intellectual products. Only those products that are new contributions should be shared as GEM3 research products. An existing tool that is used by a GEM3 researcher to generate data would not meet this definition. However, a collection of newly written Python scripts packaged as a new ArcToolbox would.

While research data does not qualify for a patent, research products may be ***patent worthy*** if the product is novel and has commercial uses. Contact your GEM3 data manager or institutional patent office for more information.

Documentation and Sharing

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

> ***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
>
> ***data dictionary***: also known as a data definition matrix, provides detailed information about the data, such as definitions of data elements, their meanings, allowable values, and relationships between elements.
>
> ***DOI or digital object identifier:*** a stabile identifier that allows for proper attribution and linking
>
> ***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 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[5])

Timely Data Sharing

We define ***timely*** data sharing as making data and research products publicly available as quickly as necessary to preserve the value of the data. Frequency of release should account for key audiences and downstream needs. For purposes of GEM3, we define this as ***6 months*** for raw and processed data and ***two years*** for derived data and non-data research products (as detailed in our NSF-approved GEM3

---

[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/ES14E00402.1

[4] Chan, L., et al. 2002. Budapest Open Access Initiative. http: / /www.opensocietyfoundations .org/ openaccess/read

[5] Gacek, C., and B. Arief. 2004. The many meanings of open source. I- Software 21:34–40.

Proposal). All data products should be made publicly available before a student researcher is to graduate.

We define ***data embargo*** as the process by which access to data download is restricted for a certain amount of time (generally 1 to 2 years) before it is made available as open access. 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. At the end of the embargo period, data files will be available for download by the public. This process will allow GEM3 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 SHARING PLAN

Researcher Accountability and Responsibilities

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 **sensitive data** that require additional protections[6] are responsible for communicating and requesting any additional assistance needed from a GEM3 Data Manager or the DM Working Group in proper curation of these data. Data that is deemed sensitive may be exempt from sharing on a public repository, but sensitive data must still be documented, archived, and secured in accordance with this plan and the NSF-approved GEM3 Proposal.

Graduate student and post-doctoral scholars who received funding from GEM3 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 repository **prior to graduation or end of the post-doctoral period.**

It is the responsibility of individual researchers to contribute their raw and derived data products and non-data research products, along with valid standards-based metadata, in a timely manner (defined above). The institutional leads from the Joint Leadership Team (JLT) are responsible for monitoring and following up with individual researchers who need to provide data. The State EPSCoR Director has the responsibility for enforcement. 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. 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.

[6] Additional protections include but are not limited to IRB review, Animal Care and Use review, Gramm-Leach-Bliley Act (GLBA), Health Insurance Portability and Accountability Act (HIPAA), Health Information Technology for Economic and Clinical Health Act (HITECH), Family Educational Rights and Privacy Act of 1974 (FERPA), Federal OMB Circular A-110 guidelines, and any federal and state wildlife laws and policies concerning release of sensitive information such as the *precise location* species have been observed.

Acquiring Data via a Data Sharing Agreement

Acquiring existing research data from an outside agency or partner often requires the data recipient to enter into a **Data Sharing Agreement** with the data holder. Such agreements are not just contracts between individual researchers and data holder, but also between the researcher's Institution and the data holder. As such, it is important that researchers engage their institution's legal representative early in the discussions to ensure that appropriate Data Sharing Agreements are drafted. These legal experts will have template data sharing agreements available and will advocate for the rights of both the institution and you as a researcher. Contact a GEM3 data manager if you are being asked to sign a Data Sharing Agreement to help guide you through the process.

Audiences for Data Sharing

Research data and research products created by GEM3 participants will be shared with two audiences: 1) other GEM3 participants and 2) the 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.

For **sharing with GEM3 participants**, researchers should be prepared to share and discuss any raw, processed, and derived data as well as other research products as soon as it is available. Researchers should make metadata available to other GEM3 researchers via the data submission process of the GEM3 website (www.idahogem3.org) as soon as data is ready for analysis. Data shared in this manner will have public searchable metadata and private, password-protected file download.

For **sharing with the public**, researchers should make every effort to document these data with appropriate metadata in approved repositories early in the process and no later than two years after research completion. Researchers may request a **data embargo** when they submit data to an approved public repository and request a **DOI**. Using the data embargo process and DOIs are the best ways for researchers to meet the GEM3 policy requirements for sharing data and to protect the first right to publish manuscripts using that data.

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. For example, Microsoft 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). 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.

All research data and research data products must be documented and cataloged with appropriate, complete metadata. Metadata should be provided in two forms. First, discovery level metadata (i.e. title, abstract, authors, keywords, use restrictions) helps people search for, find, and appropriately reuse your data. Second, a detailed metadata file should accompany your dataset (as a PDF, TXT, XML) and provide more details on how the data was collected, cleaned, manipulated, analyzed, and otherwise created. This file should answer the questions of how, what, why, when, and where, and allow others to fully understand your dataset and could include a data dictionary.

Shared data without adequate, corresponding metadata has limited intrinsic value and is incomplete. If you are not familiar with metadata or metadata standards, please consult a GEM3 Data Manager or the DM Working Group.

Data Repositories

To share data with other GEM3 participants, researchers will upload preliminary data and non-data products into the GEM3 limited-access repository located at www.idahogem3.org. There, researchers can find contact information for their institution's GEM3 Data Manager, and they can find a form for easy data upload. Data shared in this manner will have public searchable metadata to allow data discovery and private, password-protected file download.

To share data with the public, Idaho EPSCoR identifies and promotes the use of the existing data repositories and repositories for non-data research products (Table 1). Data should be registered at these locations by individuals using their own accounts rather than using an institutional account.

Table 1. A list of Idaho EPSCoR GEM3 approved repositories for data and non-data research products, along with information on the types of information that is suitable for upload in each.

Name and Location	**Types of information accepted**
Northwest Knowledge Network (NKN) https://www.northwestknowledge.net	All types of data from any GEM3 researcher
Boise State University Scholarworks https://scholarworks.boisestate.edu/	All types of data from any GEM3 researcher
Idaho State University GIS Training and Research Center (GISTrec) Data Services http://giscenter.isu.edu/	Geospatial data from ISU GEM3 researchers
Interactive Numeric Spatial Information Data Engine (INSIDE) Idaho http://insideidaho.org/	Geospatial data from Idaho
GenBank https://www.ncbi.nlm.nih.gov/genbank/	Genomic sequence data greater than 200 bp long
The Consortium of Universities for the Advancement of Hydrologic Science (CUAHSI) Hydrologic Information System (HIS) network in Idaho http://wrp.nkn.uidaho.edu/	Hydrologic data from Idaho
OpenABM https://www.openabm.org/	Agent-based models
Community Surface Dynamics Modeling Systems (CSDMS) http://csdms.colorado.edu/wiki/Main_Page	Spatially explicit models
Systems Dynamics Case Repository http://cases.systemdynamics.org/	System-dynamics models
Github http://github.com/	Open source code, models, and tools

Licensing

ALL GEM3 products should be licensed under the **Creative Commons Attribution Non-Commercial Share-Alike (CC BY-NC-SA)** or the **Open Data Commons Attribution License**. Attribution means that others must credit the authors of the dataset when using it. Noncommercial means that others may only use the data for non-commercial purposes. Share-alike means that if others create something new with the data and they want to share the new product, it must be distributed under the same license as the original data. For more licensing options see Creative Commons or Open Data Commons.

Research Data Acknowledgements

After data have been deposited and cataloged in an approved public repository, 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. 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. 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 OIA-1757324."

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.

Intellectual Property and Ownership

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 DOI for their data. The Northwest Knowledge Network at the University of Idaho, the Library at Boise State University, and potentially other approved repositories, can issue DOIs for stored datasets. Researchers are encouraged to request DOIs for their 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.

Researchers who believe their research products may have commercial value should contact their institutional patent office.

In the event of a Data Breach

The security of the data we collect is of upmost importance to maintain the confidence and trust of our research subjects, stakeholders, and collaborators. If you suspect that your research data has been compromised in any way, contact your institution's Information Technology Services (ITS) office immediately. Next notify the GEM3 JLT, your research team lead, and if applicable your institutional IRB.

They will work with you to notify all appropriate parties and ensure the future security of your research data.

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.

These products should all be open source where possible. In keeping with NSF policies, all source code should be open source.