SEAD | Sustainable Environment Actionable Data

Data Management Plan

This document is intended as a guide on the five main areas required by the National Science Foundation (NSF) for a Data Management Plan (DMP). The plan utilizes online data services provided by SEAD (Sustainable Environment — Actionable Data) as a supporting mechanism for managing, sharing, curating, preserving, and publishing data. SEAD provides secure access-controlled Project Spaces in which teams and individual researchers can incrementally develop datasets, and then submit their work for publication and long-term preservation. SEAD's technology guides project teams through a publication process where collections from the team's secure Project Space are curated, packaged, given a Digital Object Identifier (DOI), matched with a trusted long-term repository, and registered with the DataOne catalog. The SEAD Data Management Standards and Practices page on the SEAD website can be referenced for additional information.

The five main areas that must be covered in an NSF data management plan are:

1. Types of data
2. Data and metadata standards
3. Policies for Access and Sharing and Provisions for Appropriate Protection/Privacy
4. Plans for archiving and preservation of access
5. Policies and Provisions for Re-Use and Re-Distribution

I. Types of Data

This section of your DMP should include information on the sources of data in your project, including instruments and models/analysis tools. For each source, your DMP should describe the types, formats and prospective sizes of data that your project will generate. Additional materials required to make your data useful, such as experimental procedures, software, calibration information, test suites, and other forms of formal and informal documentation should also be described.

Your plan should include the decisions you make about what data and additional materials to preserve as records of your project. SEAD can support any decision you make regarding what to keep during the project, how and when to upload and annotate your data, and what and when to publish and preserve for the longer term. SEAD encourages and enables project teams to publish and preserve more of their data than just the final results and to consider including raw data, control experiments, calibration information, negative results, reports, notes, and other information. One of the advantages to using SEAD is that its hosted, access-controlled Project Spaces allow your team to upload and share any and all materials *as you create them* and to then decide, at milestones within your project, which subsets of this information to publish. Many repositories working with SEAD support publication of new versions of data and derived datasets. SEAD highly encourages researchers to take advantage of these capabilities to incrementally

enrich and expand their project's published data products over time.

II. Data and Metadata Standards

Your data needs to be adequately described for it to be discoverable and reusable. Your plan should include information about the file formats you will use and the additional metadata and provenance information you will record. It is usually considered to be best practice to publish data in well-documented, open formats (such as CSV, PDF, and HDF) and to use standard vocabularies for metadata (such as Dublin Core and W3C Provenance). However, considerations such as the availability of widely used applications that can produce and read the formats you use, or the need to record metadata that is not yet standardized, may make it useful to store data in proprietary or custom formats and to use multiple metadata vocabularies and/or custom terms. Your DMP should justify your choices.

SEAD supports the publication of data in any format and for project teams to use any metadata vocabulary(ies). By default, SEAD uses terms from the Dublin Core (DC) vocabulary and utilizes the W3C Provenance specification. Project Space administrators can add new terms, from external community vocabularies, or project-specific terms, as desired. SEAD software also allows teams to describe data and document relationships between data files as necessary to make it useful to others.

To support your data publication process, SEAD provides a Staging Area where you can review and dynamically compare your data and metadata against the requirements of specific repositories to identify any issues you may have with unsupported formats and/or missing required metadata. This process enables you to quickly update your submission to assure that it meets the policies of the repository you choose.

III. Policies for Access and Sharing and Provisions for Appropriate Protection/Privacy

Your DMP should describe who will have access to your data and when that access will be provided. You should also address the security and privacy concerns related to your data and how they will be addressed.

SEAD provides mechanisms for you to implement your access choices. The SEAD team has also worked to assure that your data are secure within SEAD's services. You may wish to include the following information about SEAD in your DMP.

SEAD's Project Spaces are access-controlled and use encrypted (https) communications. Project Spaces are created by project teams within SEAD's hosted services. Project Space administrators control access to a specific project space, and the privileges granted to users within that space (e.g. to upload and annotate data, or to have view-only access). You have full control over who in your team has administrative rights within SEAD (who can therefore control who sees the data and who is able to edit/add/publish data). Teams also have the option to mark some or all of their data publicly visible, or to allow an extended team to have view-only access to the data, providing a "preprint" or "preview" mechanism for data before they are officially published. You should also

describe whether any of your project's data will be made available this way.

SEAD's operations have undergone review by the Center for Trustworthy Scientific Cyberinfrastructure (http://trustedci.org/) and its software has been internally reviewed for cybersecurity vulnerabilities including the top ten categories as identified by the Open Web Application Security Project (https://www.owasp.org/index.php/Top_10_2013-Top_10). Machines housing SEAD services are monitored using the standard software/procedures implemented by SEAD's parent organizations – the National Center for Supercomputing Applications (NCSA) at the University of Illinois, the University of Indiana, and the University of Michigan.

IV. Plans for Archiving and Preservation of Access

NSF's decision to subject DMPs to peer-review as part of your proposal means that what is considered adequate and/or best practice for data management, publication, and preservation may vary by discipline and program. SEAD's services are designed to support you in creating a DMP that will meet expectations in your community. Information you may wish to include in this section about SEAD is given below.

SEAD's design supports incremental upload and annotation of data making it possible for project teams to manage more of their data, with more metadata, and earlier in their process. By allowing this active use of data during the project, SEAD technology enables researchers to add an additional layer of quality assurance that does not exist when data is submitted for publication and annotated only at the end of the project's life. SEAD's RESTful API can be used to avoid manual transcription errors, and many features of SEAD user interface (including support for controlled vocabularies, use of type-ahead capabilities to encourage re-use of previously entered terms, automatic association of names with researcher IDs, previews and metadata extraction) are designed to reduce input errors. Data within SEAD's Project Spaces are stored on redundant disk arrays and are periodically backed-up.

Publication in SEAD involves automated data and metadata review, assignment of a persistent global identifier (by default a Digital Object Identifier (DOI)) and submission to one of several long-term repositories working with SEAD that have documented policies and practices for preserving data. By default, publication through SEAD results in the creation and preservation of Open Archives Initiative Object Reuse and Exchange (OAI-ORE) documentation of the contents, structure, and metadata as published along with the submitted data. SEAD also registers all published data with the DataOne federated data catalog, which provides faceted search over data from a broad range of projects.

All repositories that accept data published through SEAD work to assure the preservation of the published datasets and provide continuing access to them. Individual repositories working with SEAD may also be able to provide further review of data as it is published, longer retention periods, migration of data to new formats over time, and additional policies that may be important for your data publications. SEAD encourages researchers to discuss such specific needs with SEAD, its existing partners, and/or institutional or domain repositories that accept SEAD data packages.

SEAD will work to preserve data in perpetuity and seek a minimum planned data retention period of 5 years from partner repositories. Should SEAD and its partner institutions become unable to continue maintaining the data, efforts will be made to contact the project team that published the data and/or their institutions to transfer the publications to another provider. To date, SEAD has been used to publish collections including more than 2.2 M files representing more than 1.4 TB of information, with the oldest collections having been managed within SEAD without data loss for more than 4 years. SEAD is directly involved in national and international efforts, including the Research Data Alliance and National Data Service Consortium, to provide universal enduring storage for research data and will work to make data published through SEAD compatible with such services as they emerge.

V. Policies and Provisions for Re-Use and Re-Distribution

Your DMP should note the licenses you plan to use. You may also wish to describe the value your data is expected to have for other researchers/other projects and explain how your plan addresses their needs (i.e. that relevant data is provided with sufficient description for it to be reused).

SEAD encourages the use of open data licenses such as the Creative Commons (CC) license for data publications. Teams can request a different license if it is supported by the repository chosen for publishing and archiving of data. Unless arranged beforehand, SEAD will make published data available with its open data license in an unrestricted manner and will not enforce access control or other license provisions. SEAD also works with repositories that support data embargoes, and/or enforce access controls for sensitive, restricted-use data (e.g. data involving privacy issues). SEAD encourages researchers to discuss such specific capabilities with SEAD in advance.