

Example Data Management Plan: NSF General

Product of a DataONE Best Practices Workshop.

Background Information (not part of a DMP)

Title:

Improving the long-term preservability of HDF-formatted data by creating maps to file contents.

Purpose of project:

This proposed project will create ancillary metadata that will enable future users to read all HDF (Hierarchical Data Format) formatted data. Currently, users must have an HDF software library to read the data, or employ a tool that uses that library. It is uncertain how long the library, which is required to read the files, will be in existence. Loss of this library would mean the loss of all data in HDF format. Currently there are about 800 different types of HDF-formatted files for remote sensing data at NASA alone. This stewardship project will protect against loss of the HDF software libraries currently required to read the HDF files.

The HDF Group (www.hdfgroup.org), a non-profit company whose mission is to sustain the HDF technologies and support HDF user communities worldwide, is developing the map generation software and the specification or schema for maps for HDF data files.

Example DMP

Data Management Plan

Improving the long-term preservability of HDF-formatted data by creating maps to file contents.

1. Types of Data Produced

The products of this research are software to create maps, which are machine- and human-readable, containing information about the format and content of an HDF file. These maps can be used in the future to develop new software that enables reading of HDF files.

Map generation software will be developed for all platforms currently supported. Map generation software will be tested on the entire suite of HDF file types currently held by NASA. For quality control, individual archives that use the software to map their HDF data are responsible for verifying the quality of the maps created by performing a bit-by-bit comparison of data read using the maps and data read using existing HDF reading tools, presumably on a randomly chosen statistically significant subset of each type of HDF file in their holdings. To ensure the self-describing nature of the map files, non-HDF literate, non-native English speaking programming personnel will be used to develop the demonstration software described below. This will verify that the instruction and content of the map files are broadly understandable. The map files created as a result of this project will have the same name as the original HDF file from which they are derived, with the additional extension *.map*.

One ancillary data product produced by this project will be demonstration software (a reader) that uses the maps to read the HDF files, along with example input/output files and contextual information. While the existence of this demonstration software is not necessary to allow users in the future to read HDF data files, having a demonstration reader available in the near term will help the user community learn how to use the map files. The reader and associated files will be available as open source software at SourceForge.org, along with a copy held at the National Snow and Ice Data Center (NSIDC). Subversion will be used for source code and associated file version control.

Another ancillary product of this proposed work will be a collection of one of each kind of HDF file contained in the NASA archives. This collection of HDF files will be compiled and made available from the NSIDC archive.

The data manager during and after the project for the HDF file collection and the reader software will be the PI. The HDF Group is responsible for management of the software during the project.

2. Data and Metadata Standards

The maps that will be produced are self-describing and require no documentation. Once created, the maps will become part of the data product, and will be included as one component of the metadata for the HDF product.

3. Policies for Access and Sharing

The reader software is open source (licensed under GPLv3) and is freely available. The HDF data at NSIDC is freely available with no restrictions. The collection of HDF files has a suggested citation, which is described in the metadata.

The user community for the map writer is expected to be any repository with HDF data. The user community for the map reader is expected to be any user of HDF data in the future who wishes to re-use or adapt map-reading software to their own purposes. The user community for the collection of

HDF files is expected be the community of tool developers who need a wide range of data types with which to test their product.

4. Policies for Re-use, Distribution

This proposed project will create ancillary metadata that will enable future users to read all HDF (Hierarchical Data Format) formatted data. The reader software is open source (licensed under GPLv3) and is freely available. The HDF data at NSIDC is freely available with no restrictions. The collection of HDF files has a suggested citation, which is described in the metadata.

5. Plans for Archiving and Preservation

The writer software will be preserved by the HDF Group for the life of the HDF libraries. The HDF Group uses industry-standard best practices to ensure the integrity of their software and systems. Once the map writer has been used to generate maps for every HDF file in existence, the continued existence of the writer software is not required. The reader software will be preserved at SourceForge.org for as long as there is community interest. The collection of HDF files will be preserved at NSIDC as long as utility is deemed high.

Example DMP