IMPLEMENTING THE GROUP ON EARTH OBSERVATIONS DATA MANAGEMENT PRINCIPLES: OBSERVATIONS OF A SCIENTIFIC DATA CENTER

R. R. Downs [a] *

[a] Center for International Earth Science Information Network (CIESIN), The Earth Institute, Columbia University, 61 Route 9W, Palisades, New York, United States of America - rdowns@ciesin.columbia.edu

KEY WORDS: Scientific Data Management, Data Stewardship, Data Preservation, Data Curation, Data Usability, Data Discovery

ABSTRACT:

The Group on Earth Observations (GEO) Data Management Principles (DMP) provide direction for managing geospatial data and related information products and services. Offering opportunities for enabling discovery, accessibility, usability, preservation, and curation, the GEO DMP challenge repositories, such as scientific archives and data centers, to improve practices that foster the use of Earth science data today and in the future. In addition, the Data Management Principles Implementation Guidelines (IG) offer many practical suggestions for implementing the DMP with examples that can inform the consideration of options for improving geospatial data management practices. Implementing such improvements offers value to the users of geospatial data by enabling data providers to support the use of the data products and services that they disseminate. Adopting these improvements also can assist repositories in their efforts to meet the requirements for attaining data repository certification, which offers value for repositories and their stakeholders. This article shows how repositories can improve data management practices for geospatial data by adopting the GEO DMP, with examples drawn from a scientific data center. Current and future opportunities for improving data management practices to attain data repository certification also are described along with practical approaches that repositories can adopt in the short term.

1. INTRODUCTION

1.1 Enabling Data Use by Managing Geospatial Data

The management of geospatial data is necessary to facilitate its use, today and in the future. Like users of other kinds of scientific data, users of geospatial data need to be able to understand the data so that the data may be used to answer the scientific questions and test the hypotheses that are being investigated for a particular study. Similarly, geospatial data users also need to be able to find the data products and services that they will use. Upon identifying candidate data products and services for their study, users also need to be able to explore candidate data to determine whether a particular data product or service has the potential to meet their information needs. Upon identifying data that has the potential to meet their information needs, users need to be able access the data using tools or services for which they either have familiarity or have the potential to gain familiarity efficiently, so that they can conduct their intended analyses.

In the future, use of scientific data, including geospatial data, will depend on similar capabilities to those just described. However, we also can expect that much of the infrastructure will have changed, including the hardware, operating systems, and software components upon which use of the data is dependent. Recognizing the challenges for these current and future scenarios for using geospatial data, as well as other kinds of scientific data, data creators, stewards, application developers, distributors, and other data stakeholders have a responsibility to ensure that the data are properly managed so that current and future users will be able to discover, explore, and use the data that are being created today. In light of this responsibility, stakeholders who have faced challenges with scientific data management and have attained experience in overcoming such challenges, should offer guidance for others to manage scientific data, including geospatial data, so that the data can be properly prepared for use by current and future users.

1.2 The Group on Earth Observations Data Management Principles

The Group on Earth Observations (GEO) has developed Data Management Principles (DMP) to provide direction for managing geospatial data and related information products and services. Offering opportunities for enabling discovery, accessibility, usability, preservation, and curation, the GEO DMP challenge repositories, such as scientific archives and data centers, to improve practices that foster the use of Earth science data today and in the future. The GEO DMP contain ten principles that are recommended for improving the management of Earth science data.

In addition, the Data Management Principles Implementation Guidelines (IG) offer many practical suggestions for implementing the DMP with examples that can inform the consideration of options for improving geospatial data management practices (Group on Earth Observations, 2015). The GEO DMP and the GEO DMP IG are organized by topics and, within each topic, provide a descriptive title of the principle that is recommended. Each of the topics and principle titles of the GEO DMP are listed in Table 1 (Group on Earth Observations, 2015).

* Corresponding author

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLII-3-W2-51-2017 | © Authors 2017. CC BY 4.0 License.

Category	GEO DMP Principle Title
Discovery	DMP-1: Metadata for Discovery
Accessibility	DMP-2: Online Access
Usability	DMP-3: Data Encoding DMP-4: Data Documentation DMP-5: Data Traceability DMP-6: Data Quality-Control
Preservation	DMP-7: Data Preservation DMP-8: Data and Metadata Verification
Curation	DMP-9: Data Review and Reprocessing DMP-10: Persistent and Resolvable Identifiers

Table 1. Topics and Titles of GEO Data Management Principles

1.3 Data Management at the NASA Socioeconomic Data and Applications Center

The NASA Socioeconomic Data and Applications Center (SEDAC) has been managing Earth science data and related products and services for almost twenty-five years. During this time, SEDAC has been acquiring, integrating, and developing data to prepare and produce data products and services for use by scientists, decision-makers, and educators and their students, representing many disciplinary fields of inquiry. Many of the policies, procedures, and practices for managing data at SEDAC can serve as examples of ways in which the GEO DMP can be adopted by others. In the following sections, observations of SEDAC data management are briefly described as examples to demonstrate how implementation of the GEO DMP can be conducted by members of the Earth science data community. In many examples, references also are provided to offer additional information and explanation of the examples that are described.

2. EXAMPLES OF DATA MANAGEMENT

2.1 Enabling Data Discovery by Implementing DMP-1: Metadata for Discovery

SEDAC creates metadata to describe each dataset that it disseminates. The metadata records are completed and reviewed to ensure that correct metadata is available to enable discovery. In addition, the abstract and the recommended citation for each dataset are displayed online on the data landing page as well as in the associated metadata to enable users to gain a general understanding of the data upon viewing the data landing page. Furthermore, the metadata is accessible from each dataset landing page and is available online in HTML, XML, and Text formats.

Simplified language is used for the values of metadata elements so that diverse audiences, including students within various grade levels, can understand the data. For example, the language that is used for the rights declaration has been standardized so that the rights for using the data can be understood by diverse audiences. Likewise, the names of locations and keyword terms are included to enable understanding by potential users.

In addition to disseminating the metadata on the SEDAC website, metadata records are distributed to other data cataloging services, such as the NASA Common Metadata Repository and the DataCite catalog, to reach potential users who may not visit the SEDAC website initially as part of their search to discover data products and services.

2.2 Facilitating Data Access by Implementing DMP-2: Online Access

SEDAC offers 24/7 online Access to its data products and services from the SEDAC Website. Permissions are negotiated with data producers to enable open access to data, whenever possible. As a result of such negotiations, attribution is the only condition for using many datasets distributed by SEDAC. The Creative Commons Attribution (CC By) license is an easy to understand license that can be applied to enable the use of open access information resources under the condition that attribution is provided to the source of the information resource (Creative Commons, 2013). The CC By license has been applied to maps disseminated by SEDAC for over a decade and, more recently, to the GPWv4 collection of datasets that have been produced by CIESIN.

Each data collection is described within a collection overview webpage. In addition to providing a description of the collection, the overview webpage contains links to a description of methods employed to develop the collection and links to the datasets within the collection, the Map Gallery for the collection, and any Map Services that are available for the collection. Furthermore, the collection overview webpage contains links to citations of data products and services within the collection and to Frequently Asked Questions (FAQs) about the data within the collection.

For each dataset within a collection, users can access a landing page for the dataset, which provides a general description of the dataset, including the abstract title, purpose, recommended citation, and a list of available formats for the data. In addition to the information displayed on the dataset landing page, tabs enable access for downloading the data and for obtaining maps, map services, documentation, and metadata.

2.3 Fostering Data Usability by Implementing DMP-3: Data Encoding

Diverse community needs inform the development of the data products and services that SEDAC disseminates. SEDAC data products and services are used across multiple disciplines. An analysis of the journals in which SEDAC data were cited identified several major fields of inquiry that are using the cited data, including environment and ecology, social sciences, geosciences, biology and chemistry, economics and business, engineering, plant and animal science, multidisciplinary studies, and clinical medicine (Downs and Chen, 2015).

Requirements for usability of the data are determined from an identification of community needs, which are obtained from the current literature, recent events, the SEDAC User Working Group, NASA, and other working groups and committees (Downs and Chen, 2003). Such determinations include the identification, selection, and review of data products and services, the selection and conversion to multiple data formats for enabling access to the data, and the development of tools and services for using the data, which evolve over time to reflect changes in community practice.

2.4 Fostering Data Usability by Implementing DMP-4: Data Documentation

Data documentation is needed to facilitate understanding of the data and their potential for use. In addition to the descriptions of data offered within metadata records, data documentation enables users and potential users to learn about the data, how they were developed, and how they have been used. As one of its recent initiatives, the SEDAC Configuration Management Board (CMB) has developed a Data Documentation Template to provide guidance for the documentation of data disseminated by SEDAC. The template provides elements for data producers and developers to describe the data for use by diverse audiences. The outline for the NASA SEDAC Data Documentation Template appears in Table 2.

Documentation for
<Dataset Title>
<Documentation Publication Date>
<Authors>

Abstract
Data set citation
Suggested citation for documentation
Contact to provide feedback on documentation
Table of Contents
I. Introduction
II. Data and Methodology
III. Data Set Description(s)
IV. How to Use the Data
V. Potential Use Cases
VI. Limitations
VII. Acknowledgments
VIII. Disclaimer
IX. Use Constraints
X. Recommended Citation(s)
XI. Source Code
XII. References
XIII. Documentation Copyright and License
Appendix 1. Contributing Authors & Documentation Revision History
Appendix 2. Data Revision History

Table 2. NASA SEDAC Data Documentation Template

2.5 Fostering Data Usability by Implementing DMP-5: Data Traceability

At SEDAC, ensuring data traceability involves capturing the provenance of each dataset to record the history of the creation, development, and production of the dataset. Capturing provenance also includes preserving any permissions received for the data, if applicable, along with contact information for the data producer, as well as any relevant communications with the data producer. Ensuring data traceability also includes capturing and describing the methodology using the new documentation template, including the research study design, and a reference to the original article that describes the data collection and any derivation and processing techniques involved, if applicable. Likewise, the instruments and any software utilized are described in the documentation, as well as the data revision history and any changes that were made in each version, along with any errors that were corrected, and any format conversions that occurred.

2.6 Fostering Data Usability by Implementing DMP-6: Data Quality-Control

Data quality-control ensures that any errors can be identified and corrected. SEDAC conducts internal and external reviews of its data products and services by employing a data lifecycle review process, where the results of each review stage provide feedback for the previous stage. The review process begins early in the data lifecycle, prior to data acquisition, by assessing community needs, as previously described. Next, a data nomination is completed for review by NASA and the SEDAC User Working Group (UWG), which also represents the user community and provides guidance on data development and dissemination plans. Based on the guidance offered, data development and dissemination plans are developed and also are reviewed by NASA and the SEDAC UWG. An internal alpha review is conducted during data development and the SEDAC CMB conducts a review prior to the beta release, which involves external reviewers, including NASA and the SEDAC UWG. Prior to production release of the data, another review is conducted by the SEDAC CMB to ensure that the data product has met the conditions for release. After release, comments and suggestions that are offered through SEDAC reference services also are gathered to inform the FAQs and future development.

2.7 Fostering Data Preservation by Implementing DMP-7: Data Preservation

SEDAC conducts data preservation and stewardship activities by archiving data and related information and by conducting periodic audits to assess its capabilities for maintaining the security and sustainability of its operations. SEDAC follows procedures for routinely accessioning each dataset, which includes appraisal, generating message digests for each dataset, establishing file and format inventories, and maintaining redundant storage facilities. Developing a portfolio approach to sustainability facilitates planning to ensure continuing access to SEDAC data (Downs and Chen, 2016). SEDAC also selects superseded data for acquisition into the SEDAC Long-Term Archive, which has been established collaboratively with the Columbia University Libraries and the Columbia University Earth Institute (Downs et al., 2007).

Regular audits of SEDAC, including routine NASA security and risk audits, ensure that its capabilities are developed to meet evolving security requirements. SEDAC also has conducted self-assessments and external test audit for ISO 16363 compliance (Downs and Chen, 2010). More recently, SEDAC also has attained World Data System Certification

2.8 Fostering Data Preservation by Implementing DMP-8: Data and Metadata Verification

Data and metadata verification is completed by conducting regular reviews of data and metadata content, media, platforms, systems, and hardware. SEDAC follow procedures for verifying media integrity, conducts periodic inspections of access to content, performs media refreshment and testing, and tests the integrity of its Archival Information Packages by completing integrity verification with SHA-1 message digests. Any changes to a dataset require creation of a new version of the dataset and associated message digests.

Periodic review of hardware and system software initiates the replacement of components prone to failure. Market studies on media, servers, and platforms identify new hardware and software for replacing legacy resources. Preventative measures and onsite security audits identify potential vulnerabilities to be addressed for improving security. And two factor authentication helps to reduce additional potential vulnerabilities.

2.9 Fostering Data Curation by Implementing DMP-9: Data Review and Reprocessing

SEDAC review procedures are initiated from pre-acquisition to post-production and include inspections by data producers, tests conducted by project scientists, and previously described reviews conducted by NASA, the SEDAC UWG, alpha and beta reviewers, the SEDAC CMB, and users who voluntarily identify errors and offer suggestions for improving data products and services. Questions and feedback received by SEDAC reference services populate the FAQs with answers to common questions, enabling users to find answers to questions that might be encountered when using a particular data product or service. Also, as previously mentioned, recommendations that are received are reviewed to identify possible improvements that can inform the development of new versions of data and to identify needed corrections or address any other issues or concerns with the data that have been reported.

2.10 Fostering Data Curation by Implementing DMP-10: Persistent and Resolvable Identifiers

Persistent and resolvable identifiers are created at SEDAC by assigning a Digital Object Identifier (DOI) to each dataset that is released. DOIs are recognized by many publishers as acceptable persistent identifiers that can be included in the citations of data that appear within the bibliographies and references sections of journal publications (Duerr, et al., 2011). SEDAC has established policy guidelines and procedures for routinely assigning a DOI to each dataset and for maintaining the DOIs to enable persistent identification of the data over time. SEDAC also records the DOIs of related resources within the DataCite metadata schema that is populated for a particular dataset, linking the data to documentation, related data products, and publications. In addition, the DOI that is assigned to each resource is included within the recommended citation for that resource.

SEDAC also assigns a DOI to new documentation that is developed using the NASA SEDAC Data Documentation Template. In addition, SEDAC is exploring additional opportunities for leveraging persistent identifiers to improve the curation and usability of the data that it disseminates.

3. PRACTICAL APPROACHES AND CONCLUSIONS

In this section, practical approaches that are recommended for adoption by data providers in the short term are described along with conclusions.

3.1 Practical Approaches for Repositories to Improve Data Management

A few approaches that are recommended for adoption in the short term to improve data management practices are described to offer practical ways in which data producers, developers, and distributors can begin implementing the GEO DMP.

Developing a template to guide the documentation of data products can enable stakeholders to begin documenting data products and services early in the data lifecycle. Creating such documentation early in the data lifecycle can assist in the planning of activities and the organization of information and can be revised as work progresses.

Involving data users and other stakeholders in reviews of data products at each stage throughout the data lifecycle can improve the quality of data being produced, developed, and disseminated. Beginning early in the data lifecycle with reviews can identify issues that can be corrected before they become major concerns.

Assigning a persistent identifier to each data product and including the persistent identifier in the recommended data citation that displays on the data landing page can assist data users in citing the data that they use for a publication. In addition, adopting persistent identifiers can enable the inclusion of data in external catalogs where new users can discover the data.

Conducting self-assessments and external audits using instruments that measure capabilities for managing data can improve data stewardship by identifying areas where improvement is needed (Downs and Chen, 2013). In addition, attaining certifications attained by completing such audits can demonstrate the achievements in data stewardship to data stakeholders.

3.2 Conclusions

For each of the principles recommended by the GEO DMP, several examples from the NASA SEDAC have been offered to describe how data management can be improved at scientific archives, data centers, and other distributors of Earth science data products and services. In addition, some practical approaches are recommended for implementation in the short term so that data distributors can begin considering such opportunities and implementing improvements to their infrastructure to start adopting the principles of the GEO DMP in an incremental manner.

ACKNOWLEDGEMENTS

The article is based on the presentation, “Implementing the Group on Earth Observations Data Management Principles: Lessons from a Scientific Data Center”, which the author prepared for presentation on 10 May 2017, during the 37th International Symposium on Remote Sensing of Environment, which was held in Tshwane, South Africa on 8-12 May 2017. The author appreciates the support received from the National Aeronautics and Space Administration (NASA) under contract NNG13HQ04C for the Socioeconomic Data and Applications Distributed Active Archive Center (DAAC).

REFERENCES

Creative Commons. 2013. Attribution 4.0 International License. https://creativecommons.org/licenses/by/4.0/legalcode

Data Management Principles (DMP) Implementation Guidelines: Life-Cycle Data Management Principles. 2015. Group on Earth Observations. https://www.earthobservations.org/documents/geo_xii/GEO-XII_10_Data%20Management%20Principles%20Implementation%20Guidelines.pdf

Duerr. RE, Downs, R.R., Tilmes, C., Barkstrom, B., Lenhardt, W.C., Glassy, J., Bermudez, L.E., Slaughter. P. 2011. On the Utility of Identification Schemes for Digital Earth Science Data:

An Assessment and Recommendations. *Earth Science Informatics*, 4(3): 139-160. http://dx.doi.org/10.1007/s12145-011-0083-6

Downs, R.R., Chen, R.S. 2003. Cooperative Design, Development, and Management of Interdisciplinary Data to Support the Global Environmental Change Research Community. *Science & Technology Libraries*. 2003; 23(4): 5-19. http://dx.doi.org/10.1300/J122v23n04_02

Downs, R.R., Chen, R.S., Lenhardt WC, Bourne W, Millman D. 2007. Cooperative Management of a Long-Term Archive of Heterogeneous Scientific Data. *Proceedings, Ensuring the Long-Term Preservation and Value Adding to Scientific and Technical Data (PV 2007)*. Oberpfaffenhofen/Munich, Germany. 9–11 Oct 2007. http://www.pv2007.dlr.de/Papers/Downs_CooperativeManagementOfALongTermArchive.pdf

Downs, R.R., Chen, R.S. 2010. Self-Assessment of a Long-Term Archive for Interdisciplinary Scientific Data as a Trustworthy Digital Repository. *Journal of Digital Information*, 11(1). http://journals.tdl.org/jodi/article/view/753

Downs, R.R., Chen, R.S. 2013. Independent Evaluation of a Scientific Data Center for Compliance with the ISO 16363 Requirements for Audit and Certification of Trustworthy Digital Repositories. *Research Data Alliance Plenary 2 Meeting*, Washington, DC, 16 Sep 2013. http://hdl.handle.net/10022/AC:P:21793

Downs, R.R., Chen, R.S. 2015. Bridging Disciplines: Assessing the Interdisciplinary Impact of Open Data. 41st International Association for Social Science Information Services and Technology (IASSIST) Annual Conference, Minneapolis, MN, Jun 2-5 2015. http://dx.doi.org/10.7916/D8J38SDZ

Downs, R.R., Chen, R.S. 2016. A Portfolio Approach to a Sustainable Business Model for Scientific Data Stewardship. *SciDataCon 2016*, Denver, CO 11-13 September 2016. http://www.scidatacon.org/2016/sessions/45/paper/273/