ESDS Resource Center for Data Producers (RCDP) WG “Year 1” Report

2024 March 3

Peter Leonard (ADNET Systems, Inc.)
Hampapuram Ramapriyan (SSAI)

Part 1 - DPDG

An overview of the Resource Center for Data Producers (RCDP) first requires an overview of the Data Product Development Guide (DPDG) for Data Producers.

The purpose of the Data Product Development Guide (DPDG) for Data Producers is stated in its Abstract:

This Data Product Development Guide (DPDG) for Data Producers was prepared for the Earth Observing System Data and Information System (EOSDIS) by the DPDG Working Group, one of the Earth Science Data System Working Groups (ESDSWGs), to aid in the development of NASA Earth Science data products.

The DPDG is intended for those who develop Earth Science data products and are collectively referred to as “data producers.” This guide is primarily for producers of Earth Science data products derived from remote sensing, in situ and model data that are to be archived at an EOSDIS DAAC (Distributed Active Archive Center). However, producers of other Earth Science data products will also find useful guidance.

Versions 1.0 and 1.1 of the DPDG were published by ESCO in July 2020 and October 2021, respectively.

Versions 1.0 and 1.1 of the DPDG Quick Start Guide (QSG), a condensed version of the DPDG, were published by ESCO in June 2021 and April 2022, respectively.

Version 1.1 of the DPDG is 66 pages in length, while Version 1.1 of the QSG is only 19 pages in length.

All published versions of the DPDG and QSG are available from https://www.earthdata.nasa.gov/esdis/esco/standards-and-practices/data-product-development-guide-for-data-producers

The differences between Version 2.0 of the DPDG and Version 1.1 are summarized in the Change Explanation of the Version 2.0 document:

Many changes have been implemented in this revision to Version 1.1. New material, including Figure 3, has been added regarding data products from airborne and field campaign investigations. Section 3.3 on cloud-optimized formats and services is new. Section 8 has been expanded to include Earthdata Pub. Appendices D and E have been revised to include mappings of the attributes to the Unified Metadata Model profiles, and expanded to include the Findability, Accessibility, Interoperability, and Reusability (FAIR) sub principles.

Version 2.0 of the DPDG was originally submitted to ESCO on 2023-05-06, the document was sent out for community review by ESCO on 2023-05-26, and the ESCO review closed on 2023-06-23, with roughly 430 comments having been received from the reviewers.

Twenty three V2.0 DPDG Editing Team Telecons were held between the 2023 and 2024 ESDSWG Annual Meetings. Four of these took place prior to the submission of the original Version 2.0 DPDG to ESCO: 2023-03-28, 2023-04-11, 2023-05-03, 2023-05-04. The remaining nineteen took place after the ESCO review of the original Version 2.0 concluded: 2023-08-09, 2023-08-17, 2023-08-24, 2023-08-30, 2023-09-06, 2023-09-13, 2023-09-19, 2023-10-05, 2023-10-10, 2023-10-16, 2023-10-23, 2023-11-29, 2023-12-04, 2023-12-06, 2023-12-21, 2023-12-27, 2024-01-03, 2024-01-04 and 2024-02-07.

The notes from these Telecons are available at https://docs.google.com/document/d/1STx3XlQHqj-WxsebajJAxiYMEVMkgYLjgeELriCpQZo/edit

The revised Version 2.0 of the DPDG was submitted to ESCO on 2024-02-15.

A key question - do we really need a Quick Start Guide (QSG) for V2.0 ?

Part 2 - RCDP

The ESDS Resource Center for Data Producers (RCDP) WG was formed at the 2023 ESDSWG Annual Meeting with the following mission:

1. Consider the creation of an RCDP Web area that presents Web links for many resources for Earth Science data product development.
2. Assess existing ESDIS Web areas related to data product development.
3. Make recommendations to ESDIS regarding the handling of the DPDG after Version 2.0.

4. Consider the implementation of a Universal Help Desk for Earth Science data product development.

Seven RCDP WG Telecons were held between the 2023 and 2024 ESDSWG Annual Meetings: 2023-04-04, 2023-05-02, 2023-06-06, 2023-08-01, 2023-09-05, 2023-11-07 and 2024-02-06.

The notes from these Telecons are available at https://docs.google.com/document/d/1STx3XlQHqj-WxsebajJAxiYMEVMkgYLjgeELriCpQZo/edit

Progress was made on all four of these items during the 2023/24 term, but, ultimately, most of our energy was focused on items 3) and 4).

Regarding the Governance of the DPDG After Version 2.0

The governance of the DPDG after Version 2.0 will have the final say regarding the following:

1. How comments regarding the DPDG will be collected.
2. When a revision of the DPDG will be made.
3. What revisions will be made.
4. Who will make these revisions (i.e., the selection of the editing team).
5. What format will be used for the master copy of the DPDG.
6. Where the master copy of the DPDG will reside.

Regarding item 1), the proposed RCDP Web site could play a role in the collection of comments regarding revisions of the DPDG.

Regarding items 3 and 4), Bob Downs suggested that a set of revisions to the DPDG could be proposed by anyone via the ESDS WG process.

Regarding the Format of Master Copy of the DPDG After Version 2.0

An informal poll of the RCDP WG regarding the best format for the master copy of the DPDG after Version 2.0 was conducted in the second half of February 2024, and what follows is a condensed summary of the results. A large group of poll respondents did not see any problem with continuing with MS Word for the format of the master copy of the DPDG. The antithetical opinion, also favored by a large group of the poll respondents, is that the master copy of the DPDG should reside in GitHub, which essentially rules out MS Word format, because MS Word is not GitHub friendly. This pro-GitHub group of respondents recommends that either Markdown or AsciiDoc or even HTML be used for the master copy of the DPDG. One respondent to the poll pointed out that the CF Conventions documentation is now being archived in GitHub in AsciiDoc format (see the many files with a .adoc suffix at https://github.com/cf-convention/cf-conventions).

Regarding a Universal Help Desk for Earth Science Data Product Development

Earthdata Forum

Peter Leonard quickly assessed that the Earthdata Forum could serve nicely as a Universal Help Desk for Earth Science data product development. The question Peter asked was regarding how to handle missing data in an Earth Science data product, and an accurate and comprehensive answer was received in a reasonable amount of time.

It was noticed that searching the Earthdata Forum for a string containing an underscore (e.g., missing_value) turns up nothing at all, which is a significant limitation, because many CF and ACDD attribute names include one or more underscores, and many CMR ShortNames also include one or more underscores.

DPDG ChatBot

We also considered the idea of developing a DPDG ChatBot that could answer questions regarding Earth Science data product development.

We note that it only makes sense to make use of a DPDG ChatBot, if the volume of questions regarding Earth Science data product development is too high for the Earthdata Forum to handle (i.e., greater than perhaps a few questions per day).

Ge Peng (UAH) produced an experimental DPDG ChatBot using Langchain in early 2024 that had ingested the original Version 2.0 DPDG.

A list of test questions regarding Earth Science data product development was developed to ask the ChatBot:

1. What are the preferred data formats for a data product that is to be archived at a DAAC?
2. What sort of testing of my data product should I carry out?
3. What is the difference between the data format, the file format, and the product format of a data product?
4. What is the difference between ShortName and LongName?

5. What is the difference between long_name and LongName?
6. How should I document the provenance of my data product?
7. How can I make my data product smaller without losing any information?
8. How do I go about obtaining a DOI for my data product?
9. Are there any special considerations that I should take into account if my data are destined to be stored in and served from a cloud environment?
10. How do I find out where to archive my data?
11. What kind of quality information do I need to include with my data product?
12. Why do I need so much metadata in my data product?
13. When do I get a permanent identifier for my data product?
14. What are the best metadata compliance checkers?

How well the experimental DPDG ChatBot performed depends upon who you ask.

Here are some of Peter Leonard's concerns regarding the experimental DPDG ChatBot:

- The experimental ChatBot could not read and ingest the information in tables, which is a serious shortcoming, because the DPDG includes an enormous amount of information in tables.
- The experimental ChatBot seems to ignore the spelling out of abbreviations. For example, the tested ChatBot could not deduce from the spelling out of HDF5 (i.e., Hierarchical Data Format Version 5) that HDF5 is a data format.
- A human expert can interpret and make word extrapolations, such as going from "Form" to "Format" in the abbreviation for netCDF-4 (i.e., network Common Data Form Version 4), which the experimental ChatBot did not seem to be able to do.
- The experimental ChatBot could not handle technical terms containing one or more underscores (it effectively replaces the underscore with a space character). For example, the CF long_name attribute, a variable-level attribute, was interpreted by the experimental ChatBot as being "long name" (i.e., with no underscore), and subsequently the experimental ChatBot conflated the long_name attribute with CMR LongName, the latter being a global attribute. Consequently, two entirely different attributes were treated by the experimental ChatBot as being one-and-the-same.
- The questions asked by beginners frequently include minor flaws, and, in some cases, major flaws - a human expert can identify what exactly is being asked by posing counter questions - can the experimental ChatBot do this?

Ge Peng pointed to the hybrid version of https://www.chatclimate.ai/ as a successful science ChatBot prototype (see Vaghefi, S.A., Stammbach, D., Muccione, V. et al. "ChatClimate: Grounding conversational AI in climate science." Nature, Communications Earth and Environment, Volume 4, Article 480 (2023). https://doi.org/10.1038/s43247-023-01084-x).

Regarding a RCDP Web Site

One of the goals of the RCDP WG is to develop a Web site that includes a comprehensive set of Web links related to Earth Science data product development. This list of Web links is to start from the Web links listed in V2.0 DPDG, and then expand from there. A list of Web links for tools useful for Earth Science data product development would be emphasizedl.

This was not completed during “Year 1” of the RCDP WG, partially because the references in the revised V2.0 DPDG document did not stabilize until the revised document was submitted to ESCO in mid February 2024.

Reviews of ESDIS Web Resources Related to Earth Science Data Product Development

Three examples of ESDIS Web resources related to Earth Science data product development are the Earthdata Forum (https://forum.earthdata.nasa.gov), Earthdata Pub (https://pub.earthdata.nasa.gov), and the Algorithm Publication Tool (APT, https://www.earthdata.nasa.gov/apt).

As described above, the Earthdata Forum was evaluated and was found to be a useful resource, though it would be more useful if successful searches of the posts on the Earthdata Forum could be done for a string containing one or more underscores.

We did not evaluate Earthdata Pub.

Our colleagues who tried to use the APT were not very impressed with this Tool.

Part 3 - Acronyms and Abbreviations

APT
AsciiDoc
CF
DAAC
DPDG
Earthdata Pub
EOSDIS

ESCO
ESDIS
ESDSWG
GitHub
HDF5
Markdown
MS Word
NASA
QSG
RCDP
WG