

Common Metadata Repository (CMR) Metadata Update (2017)

Table of Contents

- [Table of Contents](#)
- [Introduction](#)
- [Metadata Sources](#)
- [UMM-Collection Completeness](#)
- [Commonly Used Documentation Objects](#)
 - [Notes](#)
- [NASA DAAC Metadata Evolution](#)
- [CMR Metadata Groups](#)
 - [NASA vs. IDN Comparison](#)
 - [NASA vs. SciOps Comparison](#)

Introduction

The focus of this Case Study is to identify and compare the usage of metadata elements and attributes in CMR metadata collections as well as to identify and compare the completeness of UMM-Profile concepts in CMR metadata collections. The metadata usage studies include a comparison of NASA metadata with IDN and SciOps metadata as well as an evaluation of Commonly Used Documentation Objects (CUDOs). This work updates our prior analysis of CMR metadata in several important ways:

- 1) We retrieved new metadata records for all collections in the CMR during March 2017. This increased the size of our sample from ~4000 records from the NASA DAACs to over 32,000 records from the DAACs, SciOps, and the International Directory Network (IDN).
- 2) We added a new metric to our calculations that reports the percent of records in a metadata group (e.g. DAAC) that include a concept or item. This provides important information for the collection managers as well as providing information on the usage of various metadata elements. For example, we can distinguish items that occur once in every record from those that occur multiple times in some records.
- 3) We developed new visualizations for comparing metadata collections and used these visualizations to compare:
 1. DAAC records in 2016 to DAAC records in 2017
 2. DAAC records to SciOps records
 3. DAAC records to IDN records

Metadata Sources

The CMR includes metadata from many sources inside and outside of NASA. These are the sources collections we analyzed, the collection acronyms, and the number of records in the samples:

NASA Distributed Active Archive Centers	Acronym	Count
Alaska Satellite Facility	ASF	161
Crustal Dynamics Data Information System	CDDIS	38
Global Hydrology Resource Center	GHRC	1044
Goddard Earth Sciences Data and Information Center	GES_DISC	361
Level 1 and Atmosphere Archive and Distribution System	LAADS	130
Land, Atmosphere Near real-time Capability for EOS	LANCERMODIS	6
Land, Atmosphere Near real-time Capability for EOS	LANCEAMSR2	154
Langley Research Center	LARC	406
Langley Research Center	LARC_ASDC Atmospheric Science Data Center	606
Land Process DAAC - EOS Core System	LPDAAC_ECS	285
National Snow and Ice Data Center Version 0	NSIDCV0	223
National Snow and Ice Data Center EOS Core System	NSIDC_ECS	784

Ocean Biology Processing Group	OBPG	132
Oak Ridge National Laboratory	ORNL	12 16
Ozone Monitoring Instrument Near Real Time	OMINRT	5
Physical Oceanography DAAC	PODAAC	603
Socioeconomic Data and Applications Center	SEDAC	202
U.S. Geological Survey Earth Resources Observation Systems	USGS_EROS	11
International Directory Network	IDN	
Australian Antarctic Data Centre	AU_AADC	25 59
European Space Agency	ESA	103
European Organisation for the Exploitation of Meteorological Satellites	EUMETSAT	23
Indian Space Research Organisation	ISRO	19
Japan Aerospace Exploration Agency	JAXA	340
Fire Information for Resource Management System	LM_FIRMS	1
NOAA's National Centers for Environmental Information	NCEI	54 48
U.S. Geological Survey Long Term Archive	USGS_LTA	130
SciOps Collections	SciOps	
Advanced Cooperative Arctic Data and Information Service	ACADIS	393
Centro de Datos Antarticos, Argentina	AR	142
Biological and Chemical Oceanography Data Management Office	BCO-DMO	136
National Antarctic and Arctic Data Center, China	CN	134
Columbia University	COLUMBIA	214
Carbon Dioxide Information Analysis Center, Environmental Sciences Division, Oak Ridge National Laboratory, U. S. Department of Energy	DOE	202
Geologic Division, U.S. Geological Survey, U.S. Department of the Interior	DOIUSGSGD	128
Open File Services Section, Publications Warehouse, Eastern Region, Publications, U.S. Geological Survey, U. S. Department of the Interior	DOIUSGSPUBS	105
SOUTHEAST ECOLOGICAL SCIENCE CENTER, U.S. GEOLOGICAL SURVEY, U.S. DEPARTMENT OF THE INTERIOR	DOIUSGSSESC	207
Inter-American Institute for Global Change Research, Data and Information System	IAI-DIS	116
Marine Biodiversity Information Network, Scientific Committee on Antarctic Research, International Council for Science	ICSU	112
International Ocean Biogeographic Information System	IOBIS	295
National Institute of Polar Research, Ministry of Education, Science, Sports and Culture, Japan	JP	112
Korea Polar Research Institute, Republic of Korea	KR	329
Georgia Coastal Ecosystems, Long-Term Ecological Research Network Office	LTER	177
National Snow and Ice Data Center	NSIDC	187
Antarctica New Zealand, New Zealand Antarctic Institute, New Zealand	NZ	857
Research Data Archive, Computational and Information Systems Laboratory, National Center for Atmospheric Research, University Corporation for Atmospheric Research	UCAR	437
Surface Ocean Lower Atmosphere Study, School of Environmental Sciences, University of East Anglia, UK	UEA	104
British Oceanographic Data Centre, Natural Environment Research Council, United Kingdom	UK	33

Global Land Cover Facility, University of Maryland	UMD	169
Global Resource Information Database - Geneva, Division of Early Warning and Assessment, United Nations Environment Programme	UNEPDEWA	373
UNEP Regional Office for Asia Pacific, United Nations Environment Programme	UNEPROAP	162
United States Antarctic Program Data Center	USAP	190
North Inlet-Winyah Bay Reserve, Baruch Marine Field Laboratory, Belle W. Baruch Institute for Marine and Coastal Sciences, University of South Carolina	USC	151

UMM-Collection Completeness

[Complete results and links.](#)

We examined completeness of the NASA and IDN metadata groups with respect to the UMM-Collection recommendation. Nine of the fifteen required elements are complete in all these metadata collections (see Table 1).

Table 1 - UMM Concept Percent Completeness in NASA Collections

Required Concept	% Complete	Required Concept	% Complete	Required Concept	% Complete	Required Concept	% Complete
Metadata Dates	100%	Abstract	100%	Keyword	100%	Platform Short Name	97%
Resource Identifier	100%	Data Dates	100%	Related URL	94%	Instrument Short Name	93%
Resource Title	100%	Processing Level	99%	Temporal Extent	100%	Project Name	73%
Resource Version	100%	Responsibility	100%	Spatial Extent	95%		

[Summary Tables](#) include concept names (with links to information describing the concepts in the [ISO Explorer](#)), ISO paths used to search for the concepts, summary guidance relevant to the specific concepts, histograms that show the number of records in each collection that are missing the concept as well as links to table that shows the specific records that are missing various elements.

Commonly Used Documentation Objects

[Complete results and links.](#)

All scientific documentation includes contact information for people and organizations, identifiers, references to external resources (online and offline), spatial and temporal extents, keywords, and other items that occur multiple times. ISO metadata includes standard representations for these objects (and others) and it is helpful to use these standard representations as templates throughout a metadata collection.

We examined usage of these [Commonly Used Documentation Objects](#) (CUDOs) across NASA and IDN Collections and identified a number of differences across collections. We also identify collections with more complete information that can be used as examples for guiding improvement of others.

Notes

Contact Information: Most contact information in the CMR is limited to organization names and roles and contact information as part of the resource citation is rare. The [email](#) element of the contact information is important across all contact information but it is absent from many collections and contact sections.

Identifiers: Identifiers are complete across NASA and IDN for metadata records and for resource citations but are not consistently used for other items, e. g. platforms, instruments, missions.

Citations: Resource citations are complete in all collections. The ISO standard includes mechanisms for over thirty types of external documentation sources, e.g. algorithm descriptions, quality reports, scientific papers, etc. These capabilities are generally unused in CMR metadata because they generally do not exist in the primary source dialects (DIF, ECHO).

Online Resources: Most collections contain online resources for data distribution and many of those URL have associated names. Fewer have descriptions that might help users understand the function of the URL.

Spatial Extents: Minimum bounding rectangles are the most commonly used spatial extent and they are complete in 50% of the NASA and IDN collections.

Temporal Extents: Temporal extents are generally more common than spatial extents in NASA and IDN collections.

NASA DAAC Metadata Evolution

[Complete results and links.](#)

This report updates the metadata evaluation that we did during 2016 and provides an opportunity to identify how the CMR metadata have evolved over the year. The total number of records increased by over 50% during this time. We introduced a new visualization to [summarize](#) this comparison. Table 2 summarizes the results and provides links to Tables that show the elements that changed:

Table 2. Counts of completeness changes in
NASA DAAC Collections - 2016-2017

		2017		
		None	Some	All
2016	All		4	22
	Some	21		5
	None		48	

The largest change identified is forty-eight elements that were introduced to the metadata during 2017. These forty-eight elements existed in Some 2017 collections, and did not exist in any (None) 2016 collections. The second largest change is the deletion of twenty-one elements that existed in some 2016 collections and in no 2017 collections (None). This change was primarily due to an improvement in the translation from the CMR into ISO.

CMR Metadata Groups

The CMR includes three groups of metadata records with separate and distinct histories and processing paths, see Table 1. The first, referred to as the *NA SA Collection*, is made up of metadata records originally created at DAACS using the ECHO dialect. The second, referred to as the *IDN Collection*, includes records from major International data providers that are ingested into the CMR by SciOps. The third collection, referred to as *SciOps*, includes metadata records more than 1500 sources that originated in the Global Change Master Directory (GCMD) and the DIF dialect. Each of these collections includes sources that are analyzed separately with the expectation that they may have homogeneous characteristics. Of course, the validity of this assumption may vary with collection and source.

Table 1. Metadata Groups in the Common Metadata Repository (CMR)

Group Title	# Records	Group History	Major components - # Records
NASA	6367	Traditional DAAC Metadata – ECHO Dialect	GES-DISC – 1044 ORNL – 1216 18 DAAC Collections
IDN	8702	Non-NASA Collections – Managed by SciOps – Typically, DIF dialect	NOAA_NCEI – 5488 AU_AADC – 2559 8 Miscellaneous Collections Collections
SciOps (formerly GCMD)	5465	Miscellaneous, mostly non-NASA – DIF Dialect	NZ – 857 UCAR – 437 ACADIS – 393 Korea Polar - 329

Comparisons between these metadata groups are influenced by the fact that the collections that originate in ECHO contain much more content (406 items) than the collections that originate in DIF (175 items). Much of this content is related to additional attribute information and detailed contact information that exists in ECHO but not DIF.

A clear pattern that emerges from these comparisons is that items tend to exist or be complete in all or none of the collections that originate in DIF (IDN and SciOps). This reflects the homogeneity of content in these collections that may result from management by one group (SciOps) and marked differences between the content of these collections and those that originate in ECHO from various NASA DAACs.

NASA vs. IDN Comparison

[Complete results and links.](#)

The IDN group includes metadata collections from many large international data producers and providers. We had anticipated that these collections might provide insight into metadata practices and priorities of these organizations. In fact, these metadata are collected and shepherded into the CMR by SciOps and it appears that they reflect SciOps metadata management practices more than they reflect the metadata practices of the originating organizations. See [NASA vs. IDN](#) for the comparison.

NASA vs. SciOps Comparison

[Complete results and links](#)

The SciOps group includes more than 13,000 metadata records that originated in the GCMD and were provided by nearly 2000 data providers, all non-IDN members. These providers are diverse and more than 1700 of them each have fewer than ten records in CMR. We selected twenty-five providers with more than 100 records for the comparison of [NASA vs. SciOps](#).