Data Call User Guide

Preface: This document, produced by the ESDSWG Data Quality Working Group (DQWG) sub-group on the "Data Call Pilot Study", is designed as a supplementary guidance document for the Data Call Template which may be accessed here: https://bit.ly/dqdatacall

Introduction: The purpose of the Data Call Template is to provide a tool for DAACs to evaluate, on a qualitative basis, the overall quality of all varieties of individual Earth science datasets already publicly available which are based on observational data (i.e., this evaluation template is not intended for model data that makes prognostic predictions or excludes direct observational data), and are to be described within two distinct categories: Science and Product. The idea of a Data Call stems from a recommendation that originated from the conclusion statement of the 2015-2016 Executive Summary of the ESDSWG Data Quality Working Group. Contextually, the 2015-2016 Executive Summary requested a Data Call to assist ESDIS with conducting an inventory of data quality attributes that could be readily addressed by one or more of the actionable solutions as listed in the Solutions Master List. For reference, the 2015-2016 DQWG Executive Summary and Solutions Master List are available at the following links:

- https://wiki.earthdata.nasa.gov/display/ESDSWG/2015-2016+Final+Report
- https://wiki.earthdata.nasa.gov/x/2pASBg

Quality Distinctions:

As mentioned above, the quality of Earth science datasets can be broken down into high-level categories, hereby referred to as Science and Product. These distinctions were formally defined by the ESIP Information Quality Cluster (IQC), in close collaboration with the DQWG, and are well described in the following publication: http://dlib.org/dlib/july17/ramapriyan/07ramapriyan.html

More succinctly, these are defined as follows:

- ***Science Quality*** - Pertains to the data quality lifecycle/management phase that involves artifacts related to the defining, development, and validation of a particular version of dataset.
- ***Product Quality*** - Pertains to the data quality lifecycle/management phase that involves artifacts related to the production, assessment, and delivery of a particular version of a dataset. For the purpose of the DQWG, we have also included the ESIP IQC's additional information quality categories (i.e., Service and Stewardship) within the "Product" category. For our specific purposes, Product Quality is therefore extended to also include artifacts originating from data quality lifecycle/management phases relating to maintenance, preservation, dissemination, enabling use, and providing users with support/services.

Another way to think about the differences between the Science vs. Product quality is with respect to who is responsible for generating the particular artifacts in each category. In general, the data producers (typically involving the PI, science data systems team, and science teams involved with cal/val) are responsible for the majority of artifacts that provide the Science Quality. Likewise, in general, the DAACs are typically responsible for the majority of the artifacts that provide us with Product Quality.

Quality Attributes:

- ***Science Quality***
    a. Name of Technical Lead - Name of the person at the DAAC or affiliated with the DAAC who functions as the subject matter expert for the given dataset. This person may called upon later by ESDIS to help clarify any information that has been submitted for this dataset. Note: the technical lead for the "Product Quality" may be a different person than what is stated for "Science Quality".
    b. DAAC Name - This attribute allows for proper identification of the DAAC in which the given dataset officially resides in public distribution and in a manner that is consistent with its reporting to EMS/CMR/ECHO/REVERB/GCMD.
    c. Dataset DOI - Contains the Digital Object Identifier (DOI) for the dataset that is being evaluated by the template. All datasets funded by NASA and being exported to CMR/GCMD/ECHO/REVERB should contain a DOI.
    d. Product Maturity - Drop down validation is used to direct input to the proper designation, as defined by this document: https://science.nasa.gov/earth-science/earth-science-data/data-maturity-levels.
    e. Cal/Val Documentation Maturity - Drop down validation is used to designate the current state that the cal/val documentation is in, whether it is: Preliminary, Technical Report, or Refereed Journal. Preliminary documents include presentations given at science team meetings or white papers that have not been internally vetted by a science team. Technical Reports are published documents that have been vetted by a science team, but have not undergone external peer review. A Refereed Journal covers documents that have gone through external peer review in which the reviewers are selected by the publishing company.
    f. Quality Flags/Indicators Exist at Each Grid Point - A Boolean (i.e., Yes or No) expression of whether or not these types of quality flags or indicators are provided with the dataset.
    g. Quality Flags/Indicators Exist across Multiple Grid Points (e.g. Global Data File Level or Across a Specified Distribution of Grid Points) - A Boolean (i.e., Yes or No) expression of whether or not these types of quality flags or indicators are provided with the dataset.
    h. Uncertainties are Documented - A Boolean (i.e., Yes or No) expression of whether or not the uncertainties of dataset have been documented. Insertion of uncertainty information as a data variable and/or the metadata of a dataset constitutes itself as a form of documentation.
    i. Biases are Documented - A Boolean (i.e., Yes or No) expression of whether or not the biases of dataset have been documented. Insertion of bias information as a data variable and/or the metadata of a dataset constitutes itself as a form of documentation.
    j. Error Contribution of Input Data - This applies only to Level 3 and 4 datasets. Purpose is to provide a high level expression of whether or not the error contribution of the input data is captured somewhere. Selections include "known for all sources", "known for some sources", and "unknown".
    k. Sources of Input Data - This applies only to Level 3 and 4 datasets. Please cite the source/sensor combination, processing level, version number, and name of the data provider.
    l. Time Series Stability - Only applies to datasets in which time series is in excess of 3 years. The purpose is to assess whether the time series is capturing a trend; the physical correctness and statistical significance of the trend may be addressed in cal/val documentation. Options include: "upward trend", "neutral", "downward trend", "unknown", "N/A".
    m. Time Series Reliability - Applies to all datasets. The purpose is to assess whether the time series is reliable for a particular scientific application. Options include: "climate", "synoptic", "unknown", "N/A". Climate applications are generally reserved for datasets in which the calibration is temporally consistent, data gaps are statistically insignificant, and the time series length is long enough (generally 10 years or more for most climate measurement parameters); for some datasets, the time series may be too short, but if the dataset is well-calibrated with other datasets, which when combined together form a sufficiently long time-series to formulate a Climate Data Record (CDR), then this could still be applicable to Climate. Synoptic usually indicates time scales that are sub-seasonal.

n. Applicable ECV - Insert the corresponding name of the Essential Climate Variable (ECV), as defined by CEOS, that this dataset corresponds to. Note: this is an optional field and only applies to datasets intended for climate assessment and/or climate model intercomparison.

- ***Product Quality***
  a. Name of Technical Lead - Name of the person at the DAAC or affiliated with the DAAC who functions as the subject matter expert for the given dataset. This person may called upon later by ESDIS to help clarify any information that has been submitted for this dataset. Note: the technical lead for the “Product Quality” may be a different person than what is stated for “Science Quality”.
  b. DAAC Name - This attribute allows for proper identification of the DAAC in which the given dataset officially resides in public distribution and in a manner that is consistent with its reporting to EMS/CMR/ECHO/REVERB/GCMD.
  c. Dataset DOI - Contains the Digital Object Identifier (DOI) for the dataset that is being evaluated by the template. All datasets funded by NASA and being exported to CMR/GCMD/ECHO/REVERB should contain a DOI.
  d. Climate and Forecasting (CF) Compliant - CF Conventions are endorsed by NASA and ESDIS as an extension of the recommendations put forth by the ESDSWG Dataset Interoperability Working Group (DIWG). In addition to providing a very technical NASA-specific CF specifications document (https://cdn.earthdata.nasa.gov/conduit/upload/506/ESDS-RFC-021-v0.01.pdf), the DIWG also published an official ESDIS Standards Office (ESO) document providing conveying reasons for advocating CF across NASA EOSDIS: https://earthdata.nasa.gov/standards/climate-and-forecast-cf-metadata-conventions. The primary purpose for DQWG adopting this as a quality-specific attribute is to support usability through interoperability and also to promote standardization with respect to CF representation of quality flags and quality indicators. This field represents a simple Yes/No or N/A Boolean expression to indicate whether compliance is adhered to. N/A is for datasets that are not in netCDF, HDF, or ASCII format. Compliance can be verified by using the PO.DAAC Metadata Compliance Checker (MCC) Tool: https://podaac-uat.jpl.nasa.gov/mcc/. The latest CF standard specifications can be found here: http://cfconventions.org/.
  e. Attribute Conventions for Dataset Discovery (ACDD) Compliant - ACDD Conventions are endorsed by NASA and ESDIS as an extension of the recommendations put forth by the ESDSWG Dataset Interoperability Working Group and the ESIP Documentation Cluster. The primary purpose for DQWG adopting this as a quality-specific attribute is to support usability through interoperability and also to promote standardization with respect to ACDD representation of geospatial metadata. This field represents a simple Yes/No or N/A Boolean expression to indicate whether compliance is adhered to. N/A is for datasets that are not in netCDF, HDF, or ASCII format. Compliance can be verified by using the PO.DAAC MCC Tool: https://podaac-uat.jpl.nasa.gov/mcc/. More info can be found here: http://wiki.esipfed.org/index.php/Category:Attribute_Conventions_Dataset_Discovery.
  f. ISO 8601 Compliant - Represents the date/time standards for metadata in all types of data records and is required for compliance to CF and ISO 19115 (see below). The primary purpose for DQWG adopting this as a quality-specific attribute is to support usability through interoperability and also to promote standardization with respect to ISO representation of date/time metadata This field represents a simple Yes/No or N/A Boolean expression to indicate whether compliance is adhered to. N/A is for datasets that are not in netCDF, HDF, or ASCII format. The PO.DAAC MCC tool automatically checks for this: https://podaac-uat.jpl.nasa.gov/mcc/. More info can be found here: https://www.iso.org/iso-8601-date-and-time-format.html.
  g. ISO 19115 Compliant - An XML container file and/or Directory Interchange Format (DIF) metadata record (as would be provided to ESDIS for CMR/GMCD/ECHO/REVERB) is extractable and/or provided to ESDIS in this specific geospatial metadata format for the given dataset being evaluated. The primary purpose for DQWG adopting this as a quality-specific attribute is to support usability through interoperability and also to promote standardization with respect to ISO representation of geospatial metadata. This field represents a simple Yes/No Boolean expression to indicate whether compliance is adhered to. More info can be found here: https://earthdata.nasa.gov/standards/iso-19115.
  h. Calibration/Validation Visibility - This is a complementary field to what is being asked for under the Science Quality attribute referred to above as “Cal/Val Documentation Maturity”. This field is asking for whether this documentation exists either exclusively at the DAAC, at both the DAAC and some other 3rd party, or whether it is entirely non-existent.
  i. User Guide Exists - User guides are considered a paramount fixture of ensuring usability and basic understanding of a dataset, including its fitness for use and listing of known issues and limitations of the data. User guides also function as a container of citations and references for publications that are relevant to a variety of data quality topics, including but not limited to calibration/validation and uncertainty quantification/characterization. This field requests a simple Boolean Yes/No expression of whether a user guide exists for the given dataset.
  j. Data Management Plan (DMP) Exists - The DMP is considered one of the foundational documents for a dataset prior to its integration into a DAAC, which is intended to lay out the complete lifecycle management plan for a dataset, including a variety of data quality topics. This field requests a simple Boolean Yes/No expression of whether a DMP exists for the given dataset.
  k. Interface Control Document (ICD) Exists - The ICD is also considered one of the foundational documents for a dataset prior to its integration into a DAAC, which is intended to lay out the end-to-end production/delivery interface architecture and defines the access and distribution components and dependencies for a given dataset. This field requests a simple Boolean Yes/No expression of whether an ICD exists for the given dataset.
  l. Algorithm Theoretical Basis Document (ATBD) Exists - The ATBD is considered the foundational document of a dataset prior to the creation of the dataset, which provides technical details about the algorithm(s), both at the engineering level and the science level, and their workflow in producing geophysical measurement parameters from remote sensing observational platforms and sensors. This field requests a simple Boolean Yes/No expression of whether an ATBD exists for the given dataset.
  m. ATBD Link and/or Reference - Not all ATBDs are publicly available, but if they are, it is good to know where they reside. This field requests a URL or citable reference (such as a DOI) for the persistent, publicly accessible location of this ATBD.
  n. Alternative Usages - The intended, primary usage of a dataset is usually well known. What is not so well known are its alternative usages, which are often discovered long after the dataset becomes available to the general public. This field addresses the data quality concern of applicability and is therefore provided as an optional, free-text field to denote any alternative usages or applications of a dataset that go beyond its intended purpose for use.
  o. Temporal Application - This field intends to capture the temporally appropriate science applications of a given dataset. Selectable options include: synoptic, seasonal, and climate. Synoptic typically includes time scales of hours to weeks. Seasonal typically includes 1 to 3 months. Climate scales are typically on the order of 3 or more years, depending on the scale of the specific climate oscillation.
  p. Geographical Application - This field intends to capture the geographical coverage of a given dataset. Selectable options include: regional or global. Datasets that cover all 360 degrees of longitude and come very close to covering a minimum of 50% of latitudinal extent are generally considered global in coverage.
  q. Data Integrity Validation - This field intends to capture whether a per-dataset data file integrity check is performed by the DAAC when the file is first brought into the DAAC. These are typically done by compiling a cryptographic hash, such as an MD5 checksum or some sort of cryptographic digital signature which proves that the all of the data files for a given dataset have not mutated upon being delivered to the DAAC. Selectable options include: checksum, digital signature, none.

r. Known Issues Captured and Updated - This field intends to capture the location of any documentation containing data issues (i.e., spacecraft, instrument, outages, cal/val, algorithm updates, etc…) that may provide a data user to determine data suitability and quality for their own specific scientific application using a given dataset. Sometimes this can be collocated with the user guide. This is an optional free-text field, but it is desirable to insert a URL to the documentation source if it exists; if it is self-contained within the user guide, one may simply state: Self-Contained in User Guide.
s. Metadata is Queryable and Extractible - This field is intended to support usability through interoperability and also to promote ease of access of dataset metadata which may be used for data reduction and/or data fusion applications. This field is only intended to be used by DAACs who already provide some sort of public-facing tool or service (e.g., REST API) that allows any outside data user to query a given dataset's metadata and extract that metadata for further analysis. This also let's ESDIS know which datasets are being fully supported by these types of tools/services. Selectable options include: Yes, Only Queryable, Only Extractable, No.
t. Self-Describing Data Format - This field is intended to support usability through interoperability and also to promote transparency of information that is pertinent to assessing the quality and suitability of a dataset at the granule-level. Selectable options include: netCDF-3, netCDF-4, HDF-4, HDF-5, HDF-EOS, BUFR, GRIB, ASCII, N/A.
u. Uses NASA-Approved Format - For more information on this, please refer to: https://earthdata.nasa.gov/user-resources/standards-and-references
v. ESDIS Metrics System (EMS) Metrics Reporting Enabled - This field is intended to inform ESDIS that they are capable of viewing and analyzing usage metrics for a given dataset, which also allows ESDIS to get a better sense of a how a particular dataset is being used by the community. Selectable options include: Yes or No.
w. EMS Metrics Reporting Start Date - As a compliment to the above EMS Reporting Enabled attribute, this informs ESDIS as to how far back they can begin searching and analyzing the usage metrics for a given dataset. This is a free-text attribute that is requesting the approximate day, month and year for when the EMS reporting began.

Authors and Contributors

David Moroni (lead-author)

Co-authors: Hampapruam Ramapriyan, Yaxing Wei, Donna Scott, Bob Downs, Deborah Smith, Chung-Lin Shie, Zhong Liu, Beth Huffer, and James Tilton