Ensuring and Improving Information Quality for Earth Science Data and Products – Role of the ESIP Information Quality Cluster


Ge Peng, North Carolina State University & NOAA National Centers for Environmental Information

David Moroni, Jet Propulsion Laboratory, California Institute of Technology

Chung-Lin Shie, University of Maryland, Baltimore County & NASA Goddard Space Flight Center

September 12, 2016

H. K. Ramapriyan’s work was supported by a NASA contract with SSAI. Ge Peng is supported by NOAA under a Cooperative Agreement with NCSU. David Moroni’s work is supported by a NASA contract with the JPL, CalTech, Pasadena, CA, Chung-Lin Shie’s work was supported by NASA under a Cooperative Agreement with UMBC. Members of IQC are from many different organizations, funded under various contracts. Government sponsorship acknowledged.

Presented at SciDataCon 2016, Denver, CO
Topics

• ESIP Federation
• Information Quality Cluster (IQC)
  — Definition of Information Quality
  — IQC Objectives
• “Many Global Players”
• Related Activities
• ESIP IQC Activities
• Conclusion
Earth Science Information Partners (ESIP)

- ESIP Federation – established in 1998; currently >180 members
- From [http://esipfed.org/](http://esipfed.org)
  - **Mission:** To support the networking and data dissemination needs of our members and the global Earth science data community by linking the functional sectors of observation, research, application, education and use of Earth science.
  - **Vision:** To be a leader in promoting the collection, stewardship and use of Earth science data, information and knowledge that is responsive to societal needs.
- **Collaboration Areas** – 5 Committees, 3 Working Groups, 16 Clusters
Information Quality Cluster

• Vision
  – Become *internationally recognized* as an *authoritative and responsive information resource* for guiding the implementation of *data quality standards and best practices* of the science data systems, datasets, and data/metadata dissemination services.

• *Closely connected to Data Stewardship Committee*

• Open membership (as with all Collaboration Areas in ESIP)
Information Quality

• **Scientific quality**
  – Accuracy, precision, uncertainty, validity and suitability for use (fitness for purpose) in various applications

• **Product quality**
  – how well the scientific quality is assessed and documented
  – Completeness of metadata and documentation, provenance and context, etc.

• **Stewardship quality**
  – how well data are being managed, preserved, and cared for by an archive or repository

• **Service Quality**
  – how easy it is for users to find, get, understand, trust, and use data
  – whether archive has people who understand the data available to help users.

Information Quality is a combination of all of the above
ESIP Information Quality Cluster (IQC) - Objectives

• Share Experiences
• Actively evaluate best practices and standards for data quality from the Earth science community.
• Improve collection, description, discovery, and usability of information about data quality in Earth science data products.
• Support:
  – Data product producers with information about standards and best practices for conveying data quality; provide mentoring as needed
  – Data providers/distributors/intermediaries establish, improve, and evolve mechanisms to assist users in discovering, understanding, and applying data quality information properly.
• Consistently provide guidance to data managers and stewards on the implementation of data quality best practices and standards as well as for enhancing and improving maturity of their datasets.
Many Global Players
Some Related Activities

- NASA ESDSWG Data Quality WG (Recommendations) – 2014–present
- NOAA Dataset Lifecycle Stage based Maturity Matrices – 2009–present
- Quality Assurance framework for Earth Observation (QA4EO)
- EUMETSAT CORE-CLIMAX Data System Maturity Matrices
- NASA Earth Science Data System Working Groups (ESDSWG) – Metrics Planning and Reporting WG (Product Quality Checklists) – 2010-2012
- GEOSS Data Quality Guidelines and GEO DMP Implementation Guidelines
- CEOS Essential Climate Variables (ECV) Inventory Questions
- NCAR Community Contribution Pages

Covered in subsequent charts
NASA Earth Science Data System Working Groups (ESDSWG) – Data Quality Working Group DQWG

- **Mission:** “Assess existing data quality standards and practices in the inter-agency and international arena to determine a working solution relevant to Earth Science Data and Information System Project (ESDIS), Distributed Active Archive Centers (DAACs), and NASA-funded Data Producers.”

- **Initiated in March 2014**

- **2014-2015:**
  - 16 use cases analyzed, issues identified from users’ points of view and ~100 recommendations made for improvement
  - Consolidated into 12 high-priority recommendations

- **2015-2016:**
  - Extracted 4 “Low Hanging Fruit” (LHF) recommendations from previous 12
  - 25 solutions to address these recommendations have been identified and assessed for operational maturity and readiness for implementation, with an initial focus on four “low-hanging fruit” recommendations; solutions that exist as open-source and in an operational environment were ranked as highest priority for implementation.

- **Details in poster by Yaxing Wei et al**
Data Stewardship Maturity Matrix

- NOAA NCEI/CICS-NC Scientific Data Stewardship Maturity Matrix (DSMM) provides a unified framework for assessing the maturity of measurable stewardship practices applied to individual digital Earth Science datasets that are publicly available.

- Assesses maturity in 9 categories (e.g., preservability, accessibility, data quality assessment, data integrity) at 5 levels (1 = Not Managed; 5 = Optimally Managed).

- Provides understandable data quality information to users including scientists and actionable information to decision-makers.


(Self-assessment template: tinyurl.com/DSMMtemplate)
QA4EO

- Established and endorsed by the Committee on Earth Observation Satellites (CEOS) in response to a Group on Earth Observations (GEO) Task DA-06-02 (now Task DA-09-01a)


- **Key Principles** (from [http://qa4eo.org/docs/QA4EO_guide.pdf](http://qa4eo.org/docs/QA4EO_guide.pdf))
  - “In order to achieve the vision of GEOSS, Quality Indicators (QIs) should be ascribed to data and products, at each stage of the data processing chain - from collection and processing to delivery
  - A QI should provide sufficient information to allow all users to readily evaluate a product’s suitability for their particular application, i.e. its “fitness for purpose”.
  - To ensure that this process is internationally harmonized and consistent, the QI needs to be based on a documented and quantifiable assessment of evidence demonstrating the level of traceability to internationally agreed (where possible SI) reference standards.”

- **Framework and 10 Key Guidelines** established (e.g., establish Quality Indicator, establish measurement equivalence, expression of uncertainty)

- A few cases studies are available that illustrate QA4EO-compliant methodologies [e.g., NOAA Maturity Matrix for CDRs, WELD: Web - Enabled Landsat Data (NASA-funded MEaSUREs Project), ESA Sentinel-2 Radiometric Uncertainty Tool]
ISO 19157:2013 - Geographic information -- Data quality*

• Establishes principles for describing the quality of geographic data
  – Defines components for describing data quality
  – Specifies components and content structure of a register for data quality measures
  – Describes general procedures for evaluating the quality of geographic data
  – Establishes principles for reporting data quality

• Defines a set of data quality measures for use in evaluating and reporting data quality

• Applicable to data producers providing quality information to describe and assess how well a data set conforms to its product specification

• Applicable to data users attempting to determine whether or not specific geographic data are of sufficient quality for their particular application

• Examples of DQ Elements: Completeness, Thematic Accuracy, Logical Consistency, Temporal Quality, Positional Accuracy

ESIP IQC Activities

Data Stewardship Committee

IQC

Define and Assess Use Cases

Identify needs for capturing, describing and conveying data quality information

Provide guidance on roles and responsibilities of key players

Evaluate NASA ESDSWG Data Quality Working Group (DQWG) recommendations and propose possible implementations

Establish baseline of standards and best practices for data quality

Documentation Cluster

Discovery Cluster
Conclusion

• Capture, description, discovery, and usability of information about data quality in Earth science data products is critical for proper use of data.

• Many groups are involved in developing and documenting best practices.

• ESIP Information Quality Cluster, as a multilateral group is well placed to promoting standards and best practices.

• Membership in IQC is open – you are invited to participate!
Thank you for your attention!

Hampapuram.ramapriyan@ssaihq.com
David.F.Moroni@jpl.nasa.gov
Ge.Peng@noaa.gov
chung-lin.shie-1@nasa.gov
ESIP Information Quality Cluster Activities

- Coordinate use case studies with broad and diverse applications, collaborating with the ESIP Data Stewardship Committee and various national and international programs
- Identify additional needs for consistently capturing, describing, and conveying quality information
- Establish and provide community-wide guidance on roles and responsibilities of key players and stakeholders including users and management
- Prototype innovative ways of conveying quality information to users
- Evaluate NASA ESDSWG DQWG recommendations and propose possible implementations.
- Establish a baseline of standards and best practices for data quality, collaborating with the ESIP Documentation Cluster and Earth Science agencies.
- Engage data provider, data managers, and data user communities as resources to improve our standards and best practices.
NASA MEaSUREs - Product Quality Checklists

- Making Earth System Data Records for Use in Research Environments (MEaSUREs)
- NASA-funded, typically 5-year projects generating long-term consistent time series
- Product Quality Checklists (PQC) indicate completeness of Quality Assessment, metadata, documentation, etc.
- PQC templates - developed in 2011 and adopted in 2012
- Questions asked address science quality, documentation quality, usage and user satisfaction
NCAR Climate Data Guide*

- Community contributed datasets, reviews
- Focuses on “limited selection of data sets that are most useful for large-scale climate research and model evaluation”
- Contributed reviews answer 10 key questions; Examples of topics addressed
  - strengths, limitations, and typical applications of datasets
  - Comparable datasets
  - Methods of uncertainty characterization
  - utility for climate research and model evaluation.

CDR Maturity Matrix

- NOAA NCEI Climate Data Record (CDR) Maturity Matrix assesses readiness of a product as a NOAA satellite CDR
- Assesses maturity in 6 categories (software readiness, metadata, documentation, product validation, public access, utility) at 6 levels
- Provides consistent guidance to data producers for improved data quality and long-term preservation
- EUMETSAT’s CORE-CLIMAX Matrix – based on CDR Maturity Matrix; contains guidance on uncertainty measures
- [http://www1.ncdc.noaa.gov/pub/data/sds/cdr/Guidelines/Maturity_Matrix_Template.xlsx](http://www1.ncdc.noaa.gov/pub/data/sds/cdr/Guidelines/Maturity_Matrix_Template.xlsx)
## NOAA CDR Maturity Matrix

<table>
<thead>
<tr>
<th>Maturity</th>
<th>Software Readiness</th>
<th>Metadata</th>
<th>Documentation</th>
<th>Product Validation</th>
<th>Public Access</th>
<th>Utility</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Conceptual development</td>
<td>Little or none</td>
<td>Draft Climate Algorithm Theoretical Basis Document (C-ATBD); paper on algorithm submitted</td>
<td>Little or None</td>
<td>Restricted to a select few</td>
<td>Little or none</td>
</tr>
<tr>
<td>2</td>
<td>Significant code changes expected</td>
<td>Research grade</td>
<td>C-ATBD Version 1+ ; paper on algorithm reviewed</td>
<td>Minimal</td>
<td>Limited data availability to develop familiarity</td>
<td>Limited or ongoing</td>
</tr>
<tr>
<td>3</td>
<td>Moderate code changes expected</td>
<td>Research grade; Meets int'l standards: ISO or FGDC for collection; netCDF for file</td>
<td>Public C-ATBD; Peer-reviewed publication on algorithm</td>
<td>Uncertainty estimated for select locations/times</td>
<td>Data and source code archived and available; caveats required for use.</td>
<td>Assessments have demonstrated positive value.</td>
</tr>
<tr>
<td>4</td>
<td>Some code changes expected</td>
<td>Exists at file and collection level. Stable. Allows provenance tracking and reproducibility of dataset. Meets international standards for dataset</td>
<td>Public C-ATBD; Draft Operational Algorithm Description (OAD); Peer-reviewed publication on algorithm; paper on product submitted</td>
<td>Uncertainty estimated over widely distributed times/location by multiple investigators; Differences understood.</td>
<td>Data and source code archived and publicly available; uncertainty estimates provided; Known issues public</td>
<td>May be used in applications; assessments demonstrating positive value.</td>
</tr>
<tr>
<td>5</td>
<td>Minimal code changes expected; Stable, portable and reproducible</td>
<td>Complete at file and collection level. Stable. Allows provenance tracking and reproducibility of dataset. Meets international standards for dataset</td>
<td>Public C-ATBD, Review version of OAD, Peer-reviewed publications on algorithm and product</td>
<td>Consistent uncertainties estimated over most environmental conditions by multiple investigators</td>
<td>Record is archived and publicly available with associated uncertainty estimate; Known issues public. Periodically updated</td>
<td>May be used in applications by other investigators; assessments demonstrating positive value</td>
</tr>
<tr>
<td>6</td>
<td>No code changes expected; Stable and reproducible; portable and operationally efficient</td>
<td>Updated and complete at file and collection level. Stable. Allows provenance tracking and reproducibility of dataset. Meets current international standards for dataset</td>
<td>Public C-ATBD and OAD; Multiple peer-reviewed publications on algorithm and product</td>
<td>Observation strategy designed to reveal systematic errors through independent cross-checks, open inspection, and continuous interrogation; quantified errors</td>
<td>Record is publicly available from Long-Term archive; Regularly updated</td>
<td>Used in published applications; may be used by industry; assessments demonstrating positive value</td>
</tr>
</tbody>
</table>
# Dataset Name

## Stewardship Maturity Matrix for Digital Environmental Data Products

<table>
<thead>
<tr>
<th>Maturity Level</th>
<th>Preservability</th>
<th>Accessibility</th>
<th>Usability</th>
<th>Production Sustainability</th>
<th>Data Quality Assurance</th>
<th>Data Quality Control/Monitoring</th>
<th>Data Quality Assessment</th>
<th>Transparency / Traceability</th>
<th>Data Integrity</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Level 1</strong> - Ad Hoc Not Managed</td>
<td>Any storage location Data only</td>
<td>Not publicly available Person-to-person</td>
<td>Extensive product-specific knowledge required No documentation online</td>
<td>Ac Hoc or Not applicable No obligation or deliverable requirement</td>
<td>Data quality assurance (DQA) procedure unknown or none</td>
<td>None or Sampling Analysis unknown or unreliable in time</td>
<td>Algorithm/method/model theoretical basis assessed (method and results online)</td>
<td>Limited product information available Person-to-person</td>
<td>Unknown or no data ingest integrity check</td>
</tr>
<tr>
<td><strong>Level 2</strong> - Minimal Managed Limited</td>
<td>Non-designated repository Redundancy Limited archival metadata</td>
<td>Publicly available Direct file download (e.g., via anonymous FTP server) Collection/dataset level searchable</td>
<td>Non-standard data format Limited documentation (e.g., user’s guide) online</td>
<td>Short-term individual PI’s commitment (grant obligations)</td>
<td>Ad Hoc and random DQA procedure not defined and documented</td>
<td>Sampling and analysis are regular in time and space Limited product-specific metrics defined &amp; implemented</td>
<td>Level 1 + Research product assessed (method and results online)</td>
<td>Product information available in literature</td>
<td>Data ingest integrity verifiable (e.g., checksum technology)</td>
</tr>
<tr>
<td><strong>Level 3</strong> - Intermediate Managed Defined, Partially Implemented</td>
<td>Designated archive Redundancy Community-standard archival metadata Conforming to limited archiving process standards</td>
<td>Level 2 + Non-standard data service Limited data server performance Granule/file level searchable Limited search metrics</td>
<td>Community Standard-based interoperable format &amp; metadata Documentation (e.g., source code, product algorithm documentation, processing o/sample data flow diagram) online</td>
<td>Medium-term institutional commitment (contractual deliverables with scope and schedule defined)</td>
<td>DQA procedure defined and documented and partially implemented</td>
<td>Level 2 + Sampling and analysis are frequent and systematic with automatic community metrics defined and partially implemented Procedure documented and available online</td>
<td>Level 2 + Operational product assessed (method and results online)</td>
<td>Algorithm Theoretical Basis Document (ATBD) &amp; source code online Dataset configuration managed (CM) Unique Object Identifier (DOI) assigned (dataset, documentation, source code) Data citation tracked (e.g., utilizing Digital Object Identifier (DOI) system)</td>
<td>Level 3 + Data archive integrity verifiable</td>
</tr>
<tr>
<td><strong>Level 4</strong> - Advanced Managed Well-Defined, Fully Implemented</td>
<td>Level 3 + Conforming to community archiving standards</td>
<td>Level 3 + Community-standard data services Enhanced data server performance Conforming to community search metrics Dissemination report metrics defined and implemented Internally</td>
<td>Level 3 + Basic capability (e.g., subsetting, aggregation) &amp; data characterization (e.g., climatology, error estimates) available online</td>
<td>Long-term institutional commitment Product improvement process in place</td>
<td>DQA procedure well documented, fully implemented and available online with master reference data Limited data quality assurance metadata</td>
<td>Level 3 + Anomaly detection process well documented and fully implemented using community metrics, automatic, tracked and reported Limited quality monitoring metadata</td>
<td>Level 3 + Quality metadata assessed (method and results online)</td>
<td>Operational Algorithm Description (OAD) online, OID assigned, and under CM</td>
<td>Level 3 + Data access integrity verifiable Conforming to community data integrity technology standard</td>
</tr>
<tr>
<td><strong>Level 5</strong> - Optimal</td>
<td>Level 4 + Archiving process performance controlled, measured, and audited Future archiving standard changes planned</td>
<td>Level 4 + Dissemination reports available online Future technology and standard changes planned</td>
<td>Level 4 + Enhanced online capability (e.g., visualization, multiple data formats) Community metrics of data characterization (regional/area) online External ranking</td>
<td>Level 4 + National or international commitment Changes for technology planned</td>
<td>Level 4 + DQA procedure monitored and reported Conforming to community quality metadata &amp; standards External review</td>
<td>Level 4 + Assessment performed on a recurring basis Conforming to community quality metadata &amp; standards External ranking</td>
<td>Level 4 + System information online Complete data provenance online</td>
<td>Level 4 + Data authenticity verifiable (e.g., signature technology) Performance of data integrity check monitored and reported</td>
<td></td>
</tr>
</tbody>
</table>