NASA Data Quality Working Group 2015-2016 Findings

ESIP IQC Telecon - 14 April 2016

Chair: David Moroni[1] **Co-Chair**: H. K. "Rama" Ramapriyan[2]

[1]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA

[2]Science Systems and Applications, Inc., Goddard Space Flight Center, Earth Science Data Information Systems, Greenbelt, MD

16 Active Participants:

1. Ed Armstrong (JPL/PO.DAAC)

2. Ross Bagwell (ESDIS)

3. Stacie Doman Bennett (LPDAAC)

4. Charlene DiMiceli (UMD)

5. Robert Downs (SEDAC)

6. Pierre Guillevic (UMD)

7. Peter Hall (SSAI/GSFC)

8. Molly Hardman (NSIDC)

9. George Huffman (GSFC)

10. *Siri Jodha Khalsa (NSIDC)*

11. Tiffany Matthews (ASDC)

12. Steve Olding (ESDIS)

13. *Donna Scott (NSIDC)*

14. Marc Simard (JPL)

15. Chung-Lin Shie (GSFC)

16. *Yaxing Wei (ORNL)*

2015-2016 Action Plan

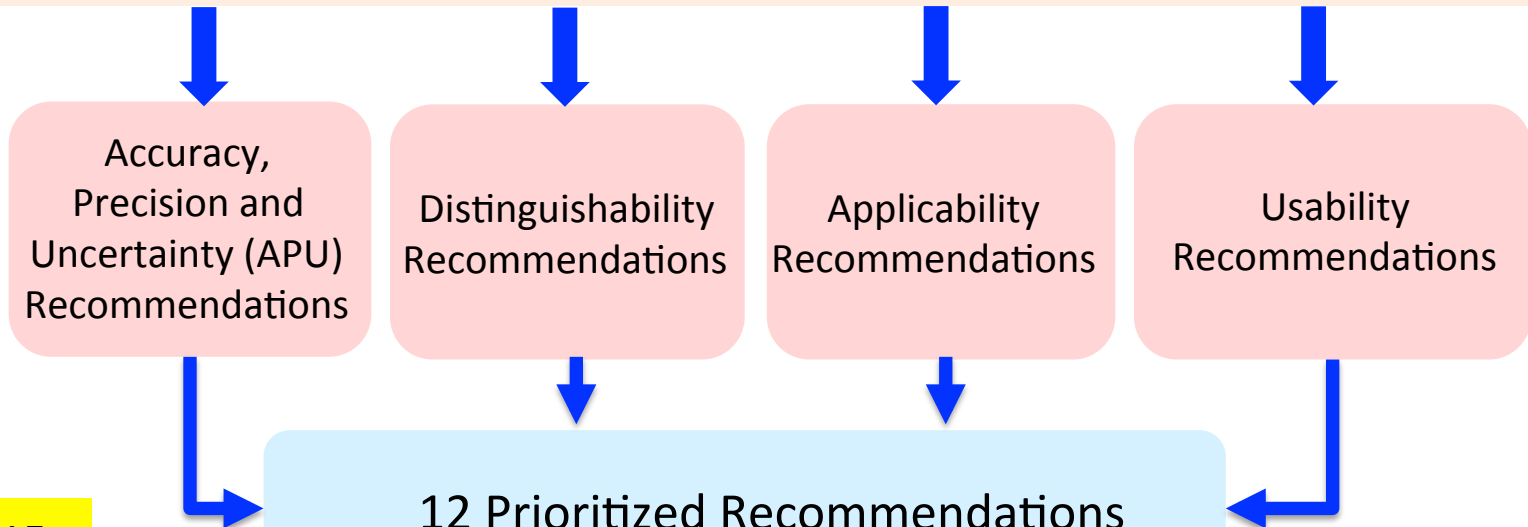| **Mission Statement** *"Discover and assess data quality standards and practices in the inter-agency and international arena to improve upon existing recommendations relevant to ESDIS, DAAC's, and NASA Data Providers."*	**Stakeholders** • NASA HQ • ESDIS • DAACs • SIPSs • MEaSUREs PI's • MEaSUREs Program • NASA Earth science instrument teams.
Approach • Complete recommendations report to ESDIS • Determine the costs, benefits, and feasibility of implementing or borrowing key elements from several existing data quality standards and practices • Continue reaching out to international communities to discover additional standards and practices not yet included in the above list.	**Outcomes, Deliverables, Milestones** ✓ Final report of recommendations from 2014 (REC2014). ✓ Mapping REC2014 to current capabilities and standards ✓ Assessment of cost, benefits and feasibility of implementation of REC2014 ✓ Vetting of REC2014 at Summer/Winter ESIP and Fall AGU. ✓ Acceptance and evaluation of new data quality use cases received from inter-agency and international data users and data providers ✓ Discovery of new solutions. ✓ Drafting of recommendations from 2015-2016 activity.

Trajectories and Outcomes

16 Use Cases Relevant to the NASA Earth Science Data and Information Systems

Accuracy, Precision and Uncertainty (APU) Recommendations

Distinguishability Recommendations

Applicability Recommendations

Usability Recommendations

Year 1: 2014-2015

12 Prioritized Recommendations

Year 2: 2015-2016

4 Low Hanging Fruit Recommendations

Data Systems Integration Recommendations

Science and Applications Recommendations

Implementation Strategies & Solutions

Assessment Report

Data Quality Management Phases

- Phase 1: ***Capturing***
  - deriving, collecting and organizing the information
- Phase 2: ***Describing***
  - documenting and procuring the information for public consumption
- Phase 3: ***Facilitating Discovery***
  - publishing and providing access to the information
- Phase 4: ***Enabling Use***
  - enhancing the utility of the information

Low-Hanging Fruit Recommendations

1. **Capturing:**

   DAACs: Maintain continuous and effective communication with data producers throughout the duration of their projects. ***Data Producers***: Develop a data quality plan for each data product and submit it along with the data for dissemination.

2. **Describing:**

   DAACs: Describe quality flags in the data documentation and in the list of Frequently Asked Questions (FAQs) about the dataset. ***Data Producers***: Provide users with a list of quality flags for questionable values along with descriptions for each quality flag (e.g., as provided by MODIS land products).

3. **Facilitating Discovery:**

   DAACs: Host a prominent web page that captures known quality issues. ***Data Producers***: Convey fully the limitations of specific datasets, for inclusion in documentation and dataset descriptions.

4. **Enabling Use:**

   DAACs: Provide enough publicly available information with self-describing metadata and documentation such that the need for users to contact the DAACs is minimized.

   Data Producers: produce and deliver the quality info needed to enable use in this context.

Remaining 8 DQ Recommendations

1. ***DAACs***: **Provide standard documents** for data producers which describe categories of quality information for documentation and metadata capture; ***Program Office***: Provide these documents to proposers of new datasets.

2. ***DAACs***: **Capture version id, processing history, and lineage** for any dataset in which multiple dataset versions exist.

3. ***DAACs***: Request information about the contribution of the various input data that are used to process a higher level product; ***Data Producers***: include **information about accuracy/uncertainty of input** datasets used along with products.

4. ***DAACs***: Provide **easy-to-use quality flags** using standardized metadata and documenting the lineage and derivations of each quality flag; ***Data Producers***: Provide quality flags corresponding to a quantifiable metric, such as the uncertainty, confidence intervals, and confidence levels.

5. ***DAACs***: Employ standards-based **metadata consistency checking tool** that meets usability needs and generates reports using a metadata scoring framework; ***Program Office***: Provide guidance on how data quality related attributes will be evaluated in the metadata scoring framework.

6. ***DAACs***: Include **documentation** on how **accuracy and uncertainty of products** were determined; ***Data Producers***: Provide all data with added quality and/or uncertainty flags for the areas that have potential limitations.

7. ***DAACs***: **Inform users** as soon as possible **when data are compromised** and provide status updates promptly; ***Data Producers***: provide information to data providers promptly regarding any compromised datasets.

8. ***DAACs***: Provide standing recommendations quickly to **alternative datasets** when a dataset has been retired or quarantined.

Low-Hanging Fruit Committees

- Formed out of a need to **quickly identify and evaluate existing solutions** for feasibility of integration and implementation across ESDIS.

- Provided a **mapping** of these solutions **to LHF recommendations** and a **consolidated** set of **implementation recommendations**.

- Science and Applications:
  - Chairs: **Siri Jodha Khalsa** and **Donna Scott**
  - Scope: solutions for **capturing** and **describing** DQ information and **facilitating discovery of science data**.

- Data Systems Integration:
  - Chairs: **Yaxing Wei** and **Bob Downs**
  - Scope: solutions for **DAACs** and **science data systems** to capture and describe DQ info, facilitate discovery, and **enable use** of science data.

Implementation Recommendations

__Preface__: each of these map to 1 or more of the 4 LHF Recommendations (slide 5), and contain 1 or more "implementation ready" solutions. These have also been prioritized according to maturity and difficulty of integration/implementation. Prioritization scheme described in afternoon poster.

1. Guidance on ISO Metadata Standards
2. Improve Access and Understanding of DQ Info
3. Metadata Authoring and Validating Tools
4. Develop Tools to Leverage DQ Info
5. Recommendations for File-level Metadata
6. Review Board/Team for DQ and Usability
7. Facilitate Communication Between Data Producers and DAACs
8. Guidance and Best Practices on Representing DQ Info

Top Tier Solutions (out of 25)

1. ***NASA Schema for ISO Metadata***
   http://standards.iso.org/iso/19115/19115.zip

2. ***EUFAR Metadata Creator***
   http://176.31.165.18:8080/emc-eufar/

3. ***Metadata Compliance Checker***
   http://podaac-uat.jpl.nasa.gov/mcc/

4. ***Ocean CO2 Metadata Collection Form***
   http://mercury.ornl.gov/OceanOME/

5. ***Kayako*** https://support.earthdata.nasa.gov/

6. ***NCO Utilities for Granule Metadata Authorship, Editing, and Standardization***
   http://nco.sourceforge.net/

Ongoing and Future Work

- Provide additional use cases to our ESIP partner: Information Quality Cluster (IQC).
    - http://wiki.esipfed.org/index.php/Information_Quality
- Provide a final comment period to NASA stakeholders.
    - Will help determine priority and feasibility of proposed recommendations and solutions.
- Engage inter-agency and international experts on Earth science data quality needs and solutions.
- Provide Archive Centers and Data Producers with more direct data quality guidance using an endorsed set of implementation strategies and solutions.
- Build a lasting and authoritative knowledgebase of data quality best practices that can also be extended beyond NASA (i.e., ESIP IQC).

2016-2017 Action Plan

Mission Statement "*Discover and assess data quality recommendations and **solutions** in the inter-agency and international arena to improve upon existing **technologies**, practices, and standards in support of **end-to-end data lifecycle** stewardship in the NASA Earth science domain.*"	**Stakeholders** • NASA HQ • ESDIS • DAACs • SIPSs • MEaSUREs PI's • MEaSUREs Program • NASA Earth science instrument teams • ESIP Information Quality Cluster (IQC)
Approach • Re-evaluate "low-hanging fruit" recommendations for feasible solutions. • Determine "Re-use Readiness" for recommended DQ solutions. • Deliver/evaluate new use cases for ESIP IQC. • Continue engaging interagency and international partners. • Strategize new development concepts that can be leveraged to facilitate data quality recommendations.	**Outcomes, Deliverables, Milestones** • Acceptance and evaluation of new DQ use cases. • Establish an DQ Solutions Master List with up-to-date and well-vetted solutions. • DQ Solutions "Re-use Readiness" Framework. • Create a Data Call template for ESDIS to evaluate DQ compliance across all DAACs. • Discover new solutions as they become known. • Develop strategic guidelines for AIST and ACCESS PI's to foster technology and standards development to satisfy unresolved recommendations of DQWG.

Conclusions

- Full 2015-2016 Assessment report, submitted to ESDIS POCs on 31 March 2016, available on DQWG wiki: **http://bit.ly/dqwg**

- Members of the **two DQWG committees have identified solutions** addressing LHF recommendations to the extent possible based on the knowledge constrained to the participating members.

- **Not all DAACs and MEaSUREs projects were represented on the committees or in the DQWG**. Therefore, the list of solutions may be incomplete and corresponding levels of maturity and difficulty are open for future discussion.

- We recommend that a "**data call"** be employed by the ESDIS Project **to identify a more complete list** that applies to the LHF recommendations and better captures levels of maturity and difficulty for each solution.

- We propose that the **consolidated solutions list be a "living document"** that is periodically updated and extended to cover the DQWG recommendations broader than the current set of LHF recommendations.