NASA ESDSWG Data Quality Working Group Overview

David Moroni\textsubscript{1} (david.f.moroni@jpl.nasa.gov)
H. K. “Rama” Ramapriyan\textsubscript{2} (Rama.Ramapriyan@nasa.gov)

\textsuperscript{1}Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA
\textsuperscript{2}Science Systems and Applications, Inc., Goddard Space Flight Center, Earth Science Data Information Systems, Greenbelt, MD

Summer ESIP 2015
Information Quality Cluster
Outline

• Mission and Scope
• Milestones and Trajectories
• Use Cases to Recommendations
• Data Quality Management Phases
• Prioritization and Consolidation of Recommendations
• Recommendation Categories
• Summary of Recommendations
• Key Findings
• Conclusions
**Mission and Scope**

**Our Mission**: Discover and assess the existing 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.
Use Cases to Recommendations

• Google Doc Form is point of entry.
  – All use cases are transferred to a Google spreadsheet.
• 16 total use cases were submitted and evaluated by four subgroups:
  – Accuracy, Precision and Uncertainty
  – Applicability
  – Distinguishability
  – Usability
• Subgroups developed topic-relevant recommendations from each use case, separated by Data System and Science.
• Four unique data quality management phases were identified.
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
Prioritization and Consolidation of Recommendations

• 94 recommendations extracted from 16 use cases.
• Consolidated down to 12 sets of “high priority” recommendations, grouped by 7 high-level categories.
  – Matched up with one or more of the 4 relevant Data Quality Management Phases corresponding to each category.
• Recommendations split into one of two domains:
  – **Data System**: Intended for data distribution and archive systems
    • May require interaction with providers/producers to ensure requisite information is made available to data centers and users.
  – **Science**: Intended for scientific data producers
    • May require coordination with project/program management and science teams.
Recommendation Categories

1. **General (1)**: broad scope with generic means for implementation.
2. **Standard Documents & Processes (2)**: aimed at describing data quality through standardized documentation and metadata.
3. **Quality of Input Datasets used in Generating Products (1)**: focuses on the extraction, derivation, and characterization of quality of datasets used as input toward a final data product.
4. **Quality Flags & Indicators (2)**: how quality flags are defined in documentation and metadata; how quality flags are disseminated.
5. **Metadata Consistency Checking (1)**: considers the ways in which various metadata attributes are checked for consistency and compliance to existing standards.
6. **Publicizing Quality Issues (4)**: in which data quality information can be packaged and disseminated for public consumption and utilization.
7. **Dataset Recommendations (1)**: recommending alternative datasets when the primary dataset is compromised or obsolete relative to current calibrations and processing methods.
# Summary of Recommendations – Pt. 1

<table>
<thead>
<tr>
<th>Management Phase ID#</th>
<th>Rec. Category</th>
<th>Data Systems Rec’s</th>
<th>Science Rec’s</th>
</tr>
</thead>
<tbody>
<tr>
<td>1, 2</td>
<td>General</td>
<td>Maintain <em>continuous and effective communication with data producers.</em></td>
<td><em>Develop</em> and submit a <em>data quality plan</em> for each dataset intended for dissemination.</td>
</tr>
<tr>
<td>1, 2</td>
<td>Standard Documents &amp; Processes</td>
<td>Provide a <em>standard set of documents</em> to be provided to potential data producers which <em>describe categories of quality information</em> for documentation and metadata capture.</td>
<td><em>Provide</em> these <em>documents</em> to <em>proposers of new datasets</em>, with the goal of <em>containing requisite information</em> for each data quality category.</td>
</tr>
<tr>
<td>1</td>
<td>Standard Documents &amp; Processes</td>
<td><em>Capture version id, processing history, and lineage</em> for any dataset in which multiple dataset versions exist.</td>
<td></td>
</tr>
</tbody>
</table>
## Summary of Recommendations – Pt. 2

<table>
<thead>
<tr>
<th>Management Phase ID#</th>
<th>Rec. Category</th>
<th>Data Systems Rec’s</th>
<th>Science Rec’s</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Quality of Input Datasets used in Generating Products</td>
<td>Request <em>information about the contribution of the various input data</em> that are used to process a higher level product.</td>
<td>Producers to include <em>information about accuracy/uncertainty of input datasets</em> used along with products.</td>
</tr>
<tr>
<td>2, 4</td>
<td>Quality Flags &amp; Indicators</td>
<td>Describe <em>quality flags</em> in the data documentation and in list of FAQs about the dataset.</td>
<td>Provide users with a <em>listing and characterization of each quality flag</em>.</td>
</tr>
<tr>
<td>1, 2, 3, 4</td>
<td>Quality Flags &amp; Indicators</td>
<td>Provide <em>easy-to-use quality flags</em> using standardized metadata and documenting the <em>lineage and derivations of each quality flag</em>.</td>
<td>Provide <em>quality flags corresponding to a quantifiable metric</em>, such as the <em>uncertainty, confidence intervals, and confidence levels</em>.</td>
</tr>
</tbody>
</table>
## Summary of Recommendations – Pt. 3

<table>
<thead>
<tr>
<th>Management Phase ID#</th>
<th>Rec. Category</th>
<th>Data Systems Rec’s</th>
<th>Science Rec’s</th>
</tr>
</thead>
<tbody>
<tr>
<td>1, 2, 4</td>
<td>Metadata Consistency Checking</td>
<td>Employ standards-based metadata consistency checking tool that meets usability needs and generates reports using a metadata scoring framework.</td>
<td>Provide guidance on how data quality related attributes will be evaluated in the metadata scoring framework.</td>
</tr>
<tr>
<td>2, 3, 4</td>
<td>Publicizing Quality Issues</td>
<td>Host a prominent web page that captures known quality issues.</td>
<td>Convey fully the limitations of specific datasets in publicized documentation.</td>
</tr>
<tr>
<td>2, 3, 4</td>
<td>Publicizing Quality Issues</td>
<td>Provide enough publicly available information with self-describing metadata and documentation such that the need for users to contact the data providers is minimized.</td>
<td></td>
</tr>
</tbody>
</table>
## Summary of Recommendations – Pt. 4

<table>
<thead>
<tr>
<th>Management Phase ID#</th>
<th>Rec. Category</th>
<th>Data Systems Rec’s</th>
<th>Science Rec’s</th>
</tr>
</thead>
<tbody>
<tr>
<td>1, 2, 4</td>
<td>Publicizing Quality Issues</td>
<td>Include <em>documentation</em> on how accuracy and uncertainty of products were determined.</td>
<td>Provide all data with <em>added quality</em> and/or <em>uncertainty flags</em> for the areas that have <em>potential limitations</em>.</td>
</tr>
<tr>
<td>2, 3</td>
<td>Publicizing Quality Issues</td>
<td><em>Inform users as soon as possible when data are compromised</em> and provide status updates promptly.</td>
<td>Data producers to <em>provide information to data providers promptly</em> regarding any compromised datasets.</td>
</tr>
<tr>
<td>3, 4</td>
<td>Dataset Rec’s</td>
<td>Provide <em>standing recommendations</em> quickly to alternative datasets when a dataset has been retired or quarantined.</td>
<td></td>
</tr>
</tbody>
</table>
Key Findings

- **Users commonly lack the means to distinguish between datasets.**
- **Users lack information about uncertainties at the pixel level.**
- **Data producers lack a standard set of instructions for generating and representing the requisite data quality information.**
- **Tools/services for assessing data quality are not consistently available for every dataset.**
- NASA MEaSUREs Projects provide a path for developing standard instructions for presenting data quality information.
- Data providers can leverage “Amazon-esque” faceted search mechanisms to facilitate more expedient distinguishability.
- User forums can be used to extract and evaluate data quality information among users, science teams, data providers, and data producers.
- Evaluation tools and scoring frameworks have been already developed for verifying aspects of data quality and metadata consistency, which could be leveraged or improved upon.
Conclusions

• DQWG final report was submitted to ESDIS at end of June. Publication of report is pending due to ESDIS review.

• There are many challenges, but also just as many if not more opportunities to address these challenges.

• DQWG is committed to working closely with the ESIP Information Quality Cluster on finding new solutions, formulating new recommendations, and improving upon existing recommendations.

• DQWG is also open to exploring other ESIP groups for cross-cutting technologies and implementations that may significantly address the existing data quality challenges.
Acknowledgements

- Ed Armstrong
- Ross Bagwell
- Stacie Doman Bennett
- Charlene DiMiceli
- Feng Ding
- Ted Habermann
- Pierre Guillevic
- Steve Olding
- Bill Rossow
- Donna Scott
- Suhung Shen
- Chung-Lin Shie
- Marc Simard
- Gilberto Vicente
- Yaxing Wei