Use Case Analysis to Support Information Quality

H. K. Ramapriyan (SSAI/NASA GSFC ESDIS), David Moroni (JPL/CalTech), Robert Downs (SEDAC/Columbia University), Ross Bagwell (SSAI/NASA GSFC ESDIS)

There has been much discussion in the ESIP Information Quality Cluster (IQC) recently about use cases, the need for developing them and analyzing them. To help clarify the purpose of use cases and support use case development and analysis to be conducted by ESIP IQC, the process and lessons learned by the NASA Earth Science Data System Working Groups (ESDSWG) Data Quality Working Group (DQWG) from use case analyses conducted during 2014-2015 are described and summarized below.

During 2014-2015, the DQWG initially looked at the various standards and practices from the point of view of assessing and identifying solutions that can be adopted in the Earth Observing System Data and Information System (EOSDIS) Distributed Active Archive Centers (DAAC) environment. However, it became clear quite early that it would be useful to capture the issues related to conveying data quality information to users through a set of use cases, given how the data quality is currently being represented in various datasets held at the DAACs. The use cases were needed to address datasets offered (even if only represented by a small subset) by the DAACs and had to cover a broad class of users. A total of 16 use cases were defined, and information about each use case captured using a template. (Note: This template has been slightly modified for the capture of use cases in the IQC). At a fundamental level, a use case provides a number of vital pieces of information that enable the decision makers to address the needs of a data user through a derivation of one or more recommendations. The following elements of a use case are considered vital to these ends: Narrative, Scope (Primary and/or Secondary), Chronology, and Success Criteria (see examples below).

With further discussion, it became evident that it would be beneficial to analyze the use cases using four of the following focus areas, each of which is considered by the DQWG as relevant thematic areas which are vital to the capture and dissemination of data quality information: 1) Accuracy, Precision and Uncertainty, 2) Distinguishability, 3) Applicability, and 4) Usability. Four subgroups were formed to address each of these thematic areas separately in order to get a broad coverage and obtain distinct sets of recommendations that applied to each of these areas. The mission and vision of each of these subgroups are given in Appendix A. Even though these subgroups were disparate in their formulation and function, it was later discovered and agreed that each subgroup has needs and concerns that overlap. An example of the analysis of one of the use cases, with the recommendations from each of the subgroups, is given below:

Use case: Aquarius Salinity Data Quality Issue Noted in Coastal Region

Author: David Moroni, JPL

Narrative: A user of Aquarius Salinity data has done comparisons between buoy data and sea surface salinity from Aquarius in coastal areas. Differences in some cases are quite large prompting the user to ask questions about data quality. The user has made suggestions on how to better implement data quality information in the file structure.

Primary Scope: Quantitative - Product

> *Rationale*: User, specifically in a level 3 gridded product, would like information on a quality flag implemented that allows for easy interpretation of the quality of sea surface salinity at a given pixel.

Secondary Scope: Quantitative – Science

> *Rationale*: Improved quality flags, as indicated by the Primary Scope, would result in substantial impacts to the quality of scientific results, specifically when comparing sea surface salinity between Aquarius and buoys.

Chronology:

- User directly contacts the DAAC regarding the data quality concern.
- DAAC representative contacts the Aquarius Science Team regarding the issue.
- The Aquarius Science Team asks the DAAC to provide more details such as graphics or diagrams describing the issue.
- The DAAC provides the Aquarius Science Team with the appropriate info.
- The Aquarius Science Team vets the data quality issue and makes a recommendation on how to move forward.

Success Criteria:

- Successful relay of user-reported data quality concerns from the DAAC to the data producer.
- Proper vetting by the science team of the data quality issue.
- A solution is reached by the science team on how to account for this issue, such as proper flagging for data at the coast or improved documentation warning users not to use this data for coastal applications.
- The user is made aware of the outcome.

Use Case Analysis:

The purpose of the use case analysis is to identify: 1. Why the issues identified in the narrative exist, 2. Whether there are possible solutions to address the issues; and 3. What needs to be done by the data system and/or the data producers in order to meet the success criteria. The success criteria collectively indicate the idealized workflow (which graciously assumes that all of the requisite solutions exist to satisfy the criteria of each use case) of how the issue would be

resolved. In the following sections, we cover the comments and recommendations from each of the subgroups named above. The recommendations are grouped into data systems and science recommendations. The former are actions to be taken by NASA's Earth Science Data and Information System (ESDIS) Project and/or DAACs to address the issues identified by the use case, and the latter are actions to be taken by science groups (Principal Investigators, data producers and/or NASA HQ Program Scientists).

Accuracy, Precision and Uncertainty Subgroup:

Comments:

- User seems less concerned with quantifying accuracy/precision (user seems capable of doing this when comparing to buoy data) and more concerned with acquiring a level of confidence in the uncertainty of the data in coastal vs. non-coastal regions.
- User would like to see coastal quality information provided within the data file structure.
- Documentation of uncertainty and/or per pixel uncertainty would help.
- Quality flags codified in a manner to provide varying degrees of confidence levels to the uncertainty of the data in coastal regions would help.
- It would be good to provide guidance on how ISO 19157 accommodates including uncertainty, A&P in metadata fields.
- Although this use case is specifically in response to Level 3 data, it could potentially be applicable to Level 2 data as well.
- Nonetheless, it points out a common deficiency with gridded (i.e., Level 3 and Level 4) datasets in which data is very often "pre-screened" (i.e. quality flags are applied pre-emptively to the data) and consequently leaves little to no evidence of transcendent quality flag artifacts present in the higher level data products.
- Science team was required to "vet" the user inquiry (i.e., verify the data discrepancies).
- DAAC is left "unarmed" with the appropriate documentation and knowledgebase of resources to directly address the user's inquiry. Therefore, the DAAC is left with no other recourse than to contact the Aquarius Science Team.
- Success is strongly dependent on reliable imitation and responsiveness of both the DAAC and the data producer (i.e., the Aquarius Science Team).
- The above dependency could be significantly attenuated with improved foresight to capture such information (which is already well-known by the science team) in the dataset metadata and/or documentation.

Recommendations (Data Systems):

- Incorporate a checklist for oceanographic data that includes the capturing of "known issues" specifically for coastal regions.
- Consider using the following ECHO/ISO Quality Attributes: QAFRACTIONGOODQUALITY (#30), QAPERCENTGOODQUALITY (#263), QAPERCENTOTHERQUALITY (#263), QAPercentOutOfBoundsData, AutomaticQualityFlag, OperationalQualityFlag, ScienceQualityFlag, QAPERCENTOTHERQUALITY

- Host a web page that captures known quality issues.
- Provide a webified data quality screening service to filter-out data that is of a user defined quality specifications based upon data quality flags (prototype of this exists at JPL).

Recommendations (Science):

- Complete a checklist for oceanographic data that includes the capturing of "known issues" specifically for coastal regions.
- Ensure all known issues discovered by the science teams are reported to the DAACs in a timely manner.
- Work with DAACs to provide data quality information through appropriate data formatting and metadata specifications (i.e., CF, ISO, ACDD, ECHO, etc.).
- Work with DAACs to provide data quality information through a standardized quality flagging schema (the GHRSST model for quality confidence levels is a great example).

Distinguishability Subgroup:

Comments:

- Flight Project and Science Data Systems teams need to do a better job of conveying the limitations of specific datasets, which need to be included in documentation and dataset descriptions.
- Determination of cost/benefit for user specified flag levels is an interesting idea, but we don't believe this would take us into the direction of Distinguishability.
- User forums should be established where data users can interact directly with the Flight Projects to avoid intermediary delay and potential misinterpretation of dataset limitations.
- User forums can also function as a live crowd-source "vetting" field in which consensus may be reached on the prevalence and likely causes of specific dataset limitations.
- A more public "peer review", enabled by user forums, may also help promote more pro-active dissemination of known issues before they are haphazardly discovered by novice data users.

Recommendations (Data Systems):

- DAAC to establish user forums enabling users to interact directly with the data producers to avoid intermediary delay and potential misinterpretation of dataset limitations.
- ESDIS to enable user forums as a live crowd-source "vetting" field in which consensus may be reached on the prevalence and likely causes of specific dataset limitations.

Recommendations (Science):

- Do a better job of conveying the limitations of specific datasets, which need to be included in documentation and dataset descriptions.
- HQ to support open and public "peer review" to help promote increased discovery, reduced latency, and dissemination of known issues.

Applicability Subgroup:

Comments:

- This use case calls for data applications in coastal regions of the ocean, particularly in areas where co-located ocean buoy data is available.
- Quality issues are noted and suggestions for improvement have been made.
- If recommendations from the other subgroups are followed, there should be enough information for determining suitability for specific applications.
- User determines applicability of data based on quality criteria.
- Data provider can assist user with expert models of applicability, especially for novice users.
- Use case is applicable for various data processing levels.
- DAAC expertise allows addressing the user with appropriate documentation and expertise for user inquiry.
- Success solely depends upon how applicable the use case is to the user needs, along with response of data provider.

Recommendations (Data Systems):

- HQ should provide a standard set of documents to be provided to investigators and potential proposers; documents should describe what quality information should be provided and how.
- Develop capabilities for users to comment on various categorized aspects of the applicability of the data quality characteristics that could be vetted and included in the publicly available information about each data set or collection.
- Develop capabilities for determining and recording the suitability of the data quality characteristics as they apply to the context of each data set within the EOSDIS data holdings.
- Develop capabilities for investigators to annotate and describe the “fitness for use” of the data as it applies to the data quality characteristics.
- The goal should be to have enough publicly available information about each data set or collection for the user to ascertain the applicability and not be required to contact the DAAC; the information should be included within the data set.
- Data Systems should provide capabilities to allow data quality indicators for applicability.
- Develop capabilities for data sets to be searched by selecting among values for a particular quality indicator and usage application (the latter would be based on a priori assessments of “fitness for use”).

Recommendations (Science):

- DAACs should provide investigators with guidelines that describe categories of data quality and request that the investigators provide information and evidence about the applicability of the data set for each category.

- DAACs should provide capabilities for users to refine search results based on quantifiable data quality criteria, including confidence levels and the values of quality flags.
- DAACs should establish lists of variables that can be selected by a user to search for all data sets that contain the selected variable.
- Quality flags should be related to a quantifiable metric that directly relates to the usefulness, validity, and suitability of the data.
- All of the information above should be made public and easy to locate.
- Science teams need to define and/or create data quality indicators that can best describe the quality characteristics of a data product.
- Science teams should provide definitions for each quality indicator and a description of how each quality indicator can be used (documentation, user guide, and in search system).

Usability Subgroup:

Comments:

- Implement a quality flag that allows for easy interpretation and application of the quality of sea surface salinity at a given pixel. Distinctions in the data quality flag should be apparent between coastal and open-ocean regions.
- Large differences between buoy and Aquarius data at some places have been observed by the user
- User would like to see accuracy indication via quality flag.
- Documentation of uncertainty and/or per pixel uncertainty would help.
- It would be good to provide guidance on how ISO 19157 accommodates including uncertainty, accuracy and precision in metadata fields.
- Users discover, assess, and use Earth science data by accessing, understanding, and evaluating quality aspects of the data.
- Improving the availability and accessibility of information about the quality of the data offers affordances for using the data.
- Ease-of-use of the data is dependent upon obtaining detailed information about the quality of the data.
- Providing objective quality information about the data improves the potential usability of the data.
- Facilitating multiple quality indicators about the data increases the potential usability of the data for diverse uses
- The availability of tools for conducting quality reviews and for using quality information about the data increases opportunities for improving the usability of data.

Recommendations (Data Systems):

- The quality flag should be described in the data documentation and in the list of FAQs about the dataset.

- For ease-of-use, the quality flag could follow the confidence level flag as used by GHRSST.
- The goal should be to have a enough publicly available information whereby the user is not required to contact the DAAC.
- Provide users with information on the distribution of errors for each data set, including the results of an outlier analysis for each variable.
- Provide easy-to-use quality flags (like the confidence level flag used by GHRSST).
- Develop capabilities for users to search for data products that contain the same variable or variables as a particular data product of interest.
- Use OPeNDAP or THREDDS to enable users to remotely interrogate data for the purposes of quality assessment, subsetting, aggregation, co-location, and visualization.

Recommendations (Science):

- Quality flags should directly correspond to a quantifiable metric, such as the related uncertainty, confidence intervals, and confidence levels.
- The information above should be made public and easy to locate.
- Develop a data quality plan for each data product.
- Quality flags should be publicly accessible and directly correspond to a quantifiable metric, such as the related uncertainty, confidence intervals, and confidence levels.
- Create data quality indicators to describe quality characteristics of a data product (Data Quality Screening Service in GESDISC, Default Quality Flags and Advanced Quality Control from MODIS subset tool, webification to extract quality indicators on the fly and also subset on the fly using quality indicators at PODAAC).

Merged Recommendations:

Below are the recommendations that were derived by merging the recommendations from all the subgroups and eliminating duplicates.

Data Systems:

- Provide a standard set of documents to be provided to investigators and potential proposers; documents should describe what quality information should be provided and how.
- Incorporate a checklist for data that includes the capturing of "known issues" for particular regions or time intervals.
- Consider using the following ECHO/ISO Quality Attributes: QAFRACTIONGOODQUALITY (#30), QAPERCENTGOODQUALITY (#263), QAPERCENTOTHERQUALITY (#263), QAPercentOutOfBoundsData, AutomaticQualityFlag, OperationalQualityFlag, ScienceQualityFlag
- Host a prominent web page that captures known quality issues. Provide enough publicly available information with self-describing documentation so the need for users to contact the DAACs is minimized.
- Provide easy-to-use quality flags. Provide a webified data quality screening service to filter-out data that meet user defined quality specifications based on data quality flags

- Describe quality flags in the data documentation and in the list of FAQs about the dataset.
- Set up user forums enabling users to interact directly with the data producers to avoid intermediary delay and potential misinterpretation of dataset limitations.
- Employ user forums that may function as a live crowd-source "vetting" field in which consensus may be reached on the prevalence and likely causes of specific dataset limitations.

Science:

- Enable open and public peer review to help promote increased discovery, reduced latency, and dissemination of known issues.
- Provide a checklist for data that includes the capturing of known issues for particular regions or time intervals.
- Convey fully the limitations of specific datasets, for inclusion in documentation and dataset descriptions.
- Work with DAACs to provide data quality information through a standardized quality flagging schema (e.g., GHRSST model for quality confidence levels).
- Make quality flags publicly accessible and directly corresponding to a quantifiable metric, such as the related uncertainty, confidence intervals, and confidence levels.
- Provide data quality information through appropriate data formatting and metadata specifications (i.e., CF, ISO, ACDD, ECHO, etc.).
- Ensure all known issues discovered by the science teams are reported to the DAACs in a timely manner.
- Describe any restrictions on the use of the data and clearly display the rights enabling the use and adaption of the data and of the data quality information.

Merging across use cases:

As the DQWG analyzed each of the 16 use cases and arrived at merged recommendations as exemplified above, the full set of recommendations from the 16 groups was further analyzed and merged to eliminate duplicates. Merging the recommendations across the subgroups, across use cases and accounting for similarity and complementarity of data system and science recommendations, resulted in a total of 93 recommendations. The DQWG members then prioritized these 93 recommendations individually. Based on the consensus derived through this process, the 12 recommendations with the highest priority were identified.

Next, the DQWG assessed and identified a subset of 4 out of the 12 recommendations, which would be considered readily achievable by DAACs using existing solutions that are known to exist in an open-source, operational environment. These 4 recommendations were referred to as “Low Hanging Fruit” (LHF) recommendations. A total of 25 potential solutions were identified and prioritized according to operational readiness and ease of implementation/integration. Each of these 25 solutions was then mapped to one or more of the 4 LHF recommendations. The final result was the construction of 8 implementation recommendations, each of which contained a relevant subset of the 25 solutions.

Lessons Learned (for applying to IQC's use case analysis):

- Subgroups were useful to ensure breadth of coverage in the analyses.
- Moving forward, the IQC, does not need to be "officially" divided into subgroups, especially given the progress made with DQWG and the relatively small number of new use cases; however, similar thematic subgroups for time-constrained use case evaluation (such as at ESIP breakout sessions) may provide more valuable insights.
- It was useful to identify issues, document comments and develop recommendations in a free-thinking "brainstorming" mode, even though a large number of recommendations resulted.
- Consolidation and prioritization took some effort, but was an important step in removing duplicate recommendations, aggregating recommendations that were related and synergistic, and focusing on recommendations that were generalized enough to be carried out across multiple Earth science disciplines and DAACs.
- Much work remains in mapping recommendations to existing standards and solutions as well as getting buy-in from the community of data producers, distributors and users.
- More work is also needed in specifically identifying standards and solutions that are operationally mature, open-source, and relatively easy to integrate and implement across multiple DAACs, projects, and missions. We have since coined a new term, "Re-use Readiness", to help us focus on these types of standards and solutions.

Acknowledgements

The majority of the work in this brief note is the result of the authors' participation in the Data Quality Working Group, one of several NASA Earth Science Data System Working Groups. The authors would like express their appreciation for comments by Ge Peng (CICS/NCEI) and Chung-Yi (Sophie) Hou (ESIP Student Fellow).

Appendix A

The mission and vision of each of the four thematic subgroups that the DQWG formed to facilitate analysis of use cases are provided here.

1. ***Accuracy, Precision and Uncertainty Subgroup***

Mission:

Recommend the type of information about data accuracy, precision and uncertainty estimates/ characterizations users need, and how such information should be conveyed and disseminated to users.

Vision:

The necessary information about data accuracy, precision and uncertainty in NASA-provided Earth science datasets are conveyed to all types of users (machine and human) such that it is easy to ingest and interpret.

2. ***Distinguishability Subgroup***

Mission:

Provide recommendations using proven methods and solutions in communicating data quality information for Earth Science remote sensing datasets in a way that empowers greater liberty and discernment to the data user to make informed distinctions between datasets.

Vision:

Improve communication of data quality information to empower Earth Science data users across all disciplines and experience levels to confidently and more accurately discern similarities and differences between datasets.

3. ***Applicability Subgroup***

Mission:

Provide recommendations based on existing solutions being utilized by ESDIS, DAACs, Flight Projects, and Making Earth System Data Records (ESDRs) for Use in Research Environments (MEaSUREs) investigators to communicate how data quality characteristics apply to datasets to allow the user to determine their relevance for his/her purpose.

Vision:

Users are able to understand the data quality information available and can assess whether or not the data are useful for their application or purpose (i.e., "fitness for use").

4. ***Usability Subgroup***

Mission:

Improve capabilities of Earth science data users to discover, assess, and use Earth science data by accessing, understanding, and evaluating quality aspects of the data.

Vision:

Users are able to easily and effectively discover, assess, and use Earth science data based on information about the quality of the data.