Summary of Evaluations of Use Cases – Information Quality Cluster – October 2016

A concise description of four use cases, their evaluations and recommendations are presented below. These resulted from discussions within groups formed at the summer ESIP meeting in July 2016, followed by presentation to the Information Quality Cluster, and subsequent editing by the group participants. The names of each of the group participants are shown below.

Use Case 1

Group Participants: Bob Downs, David Moroni, Joseph George

Title: Dataset Rice Cooker Theory

Author: Bob Downs (rdowns@ciesin.columbia.edu)

Champion: Bob Downs

Data Quality Information Management Phase: Capture, Describe, Facilitate Discovery, Enable Use

Recommendation Category: Relevance to Application, Fitness for Intended Usage, Fitness for Alternative Usage, Discoverability, Data Usage, Data Recommendations, Documentation, Quality Assurance Procedures, Quality Flags and Indicators, Quality Information Dissemination, Quality of Input Datasets used in Generating Products, Procedures, User Interaction

Relevant Success Criteria:
1. All parties involved are responding to communication.
2. All parties involved have the requisite knowledge, skillsets, and funding to resolve these issues.
3. The user makes adjustments to account for this data quality issues.
4. Established process to document and track status.
5. The distributor should have all the information needed to update or create documentation to be made known to all other data users.
6. Fundamentally, enough distinctions can be made between the quality of various input datasets in relation to the integrated dataset.

Recommendations for Data Producers:
Exercise due diligence in exercising the discovery and disclosure of the errors and uncertainties of the input. Furthermore, we would like them to be able to convey how this relates cumulatively to the quality of the final data product. Finally, we would like to have this expressed as a function of time and at the pixel level.

Recommendations for Data Distributor:
Ensure close interaction with all level 4 data producers as well as proactive data management planning prior to the delivery of the data. Ensure that the information is conveyed to the user in a human-readable manner that is complete as possible. Finally, ensure discovery of this information is optimal to the extent that the user can easily discover this information with minimal contact with the data distributor.

Recommendations for User:
Ensure that the integration of multiple data sources, with differing quality information, does not compromise their intended use.

Current Degree of Compliance: 4 – Implementations are operationally mature at one or more agencies/institutions

Complying Agency/Institution: NASA/ESDIS, NOAA/NCEI, JAXA, UK Met Office, Australian Bureau of Meteorology

Existing Solution(s):
http://sedac.ciesin.columbia.edu/data/collection/gpw-v4
https://www.ghrsst.org/products-and-services/product-specification/l4-gridded-sst/

Justification:
These data collections and processing methods demonstrate ways in which data can be processed from multiple sources of input data and likewise account for the errors and uncertainties of the input data. Users and potential user communities need to know how the quality of the data products will impact their use and the quality of the science being produced.

Use Case 2
Group Participants: Ge Peng, Steve Olding, Lindsey Harriman

Title: Appropriate Amount/Extent of Documentations for Data Use

Author: Ge Peng (ge.peng@noaa.gov)

Champion: Ge Peng

Data Quality Information Management Phase: Capture, Describe, Facilitate Discovery, Enable Use

Recommendation Category: Relevance to Application, Fitness for Intended Usage, Discoverability

Relevant Success Criteria:
Quantitative: Amount of time the user spends finding/determining what they need
Qualitative: Positive feedback from the user that they are satisfied with the use and information provided.

Recommendations for Data Producers: Data producer should/be required to provide an ATBD (Algorithm Theoretical Basis Document) that is publicly accessible

Recommendations for Data Distributor:

- Provide publicly accessible data documentation in tiers: DOI landing page, user guide, ATBD.
- Provide a filtering method for users to decide how to find product/necessary documentation (a la Amazon shopping) based on product/data center-relevant criteria
- Provide a feedback mechanism for users
- Work with producers to create consistent User Guides.

Recommendations for User: Provide feedback via a set mechanism

Current Degree of Compliance: 3 – Prototype implementations have been deployed by one or more agencies/institutions

Complying Agency/Institution: NASA/ESDIS, NOAA/NCEI

Existing Solution(s): Some tiered documentation is available; faceted search implemented in some cases to get the users started.

Justification: Defining and communicating levels of recommended documentations will

- Help guide end-users on what document to look for to get appropriate information based on the level of their data use needs,
- Help data centers establish document templates that they need to curate,
- Help data providers supply information needed for data stewardship and use/service.

Use Case 3

Group Participants: Ross Bagwell, Hampapuram Ramapriyan (Rama), Sophie Hou, and Margaret O'Brien

Title: Understanding and Identifying Datasets using SBC LTER Data Portal

Authors: Margaret O'Brien and Sophie Hou (Sophie Hou hou@ucar.edu; margaret.obrien@ucsb.edu)

Champion: Sophie Hou

Data Quality Information Management Phase: Capture, Facilitate Discovery

Recommendation Category: Discoverability, Quality Assurance Procedures, Quality Flags and Indicators, User Interaction

Relevant Success Criteria:

- User is able to discover the SBC LTER data collections.
- User is able to understand the characteristics of the datasets and determine if the dataset is of interest by reviewing the metadata provided by SBC LTER.

Recommendations for Data Producers:

- Provide information on data quality.
- Provide sufficient information to help with categorization of the datasets.

Recommendations for Data Distributor:

- Refine the algorithm for the "one box" (Google) search.
- Improve interface for the discovery (e.g. perform usability evaluations).
- Provide advice to data producers regarding the terms used to tag datasets.
- Review datasets for quality/completeness of information.

Recommendations for User: N/A

Current Degree of Compliance: 1 – Considered by one or more agencies/institutions but not yet scoped

Examples of Complying Agency/Institution that data distributor might reference: NASA/ESDIS, NCAR

Existing Solution(s):

- Follow community standards for keywords/controlled vocabularies (e.g. GCMD).
- There are several existing usability evaluation techniques that can be applied.

Justification:

- GCMD provides a well-known set of keywords/controlled vocabularies that researchers should have common understanding of the definitions used.
- The usability techniques will help with the design of the webpages to improve discoverability.

Use Case 4

Group Participants: Han Qin, Chung-Lin Shie, Bhaskar Ramachandran, and Ruth Duerr

Title: Citizen Science

Author: Ruth Duerr (ruth.duerr@ronininstitute.org)

Champion: Ruth Duerr

Data Quality Information Management Phase: Capture, Describe, Facilitate Discovery, Enable Use

Recommendation Category: Relevance to Application

Relevant Success Criteria:

- If your project involves small scale areas you may be able to find subject-matter experts (e.g., local sea ice experts),
- Large scale popular projects with high public interest can use multiple-eyes to crowd source the results (i.e., figure out where the outliers are by looking at the distribution of answers).
- Focus on specific measurement, try to avoid involving multiple measurements so that training is reasonable

Recommendations for Data Producers: training

Recommendations for Data Distributor: communications with producers, anonymized producer information in the data (need to be careful if populations are small - for example if you need to ensure the anonymity of an 85 yr old women in a very small village of 100 people, simply removing names from the data is not going to be sufficient)

Recommendations for User: read the documentation, provide feedback on use and issues found

Current Degree of Compliance: 4 – Implementations are operationally mature at one or more agencies/institutions

Complying Agency/Institution: NASA/ESDIS, NOAA/NCEI, NSF, USGS, ESIP

Existing Solution(s): Goodchild, Michael F., and Linna Li. "Assuring the quality of volunteered geographic information." Spatial statistics 1 (2012): 110-120.

Justification: N/A