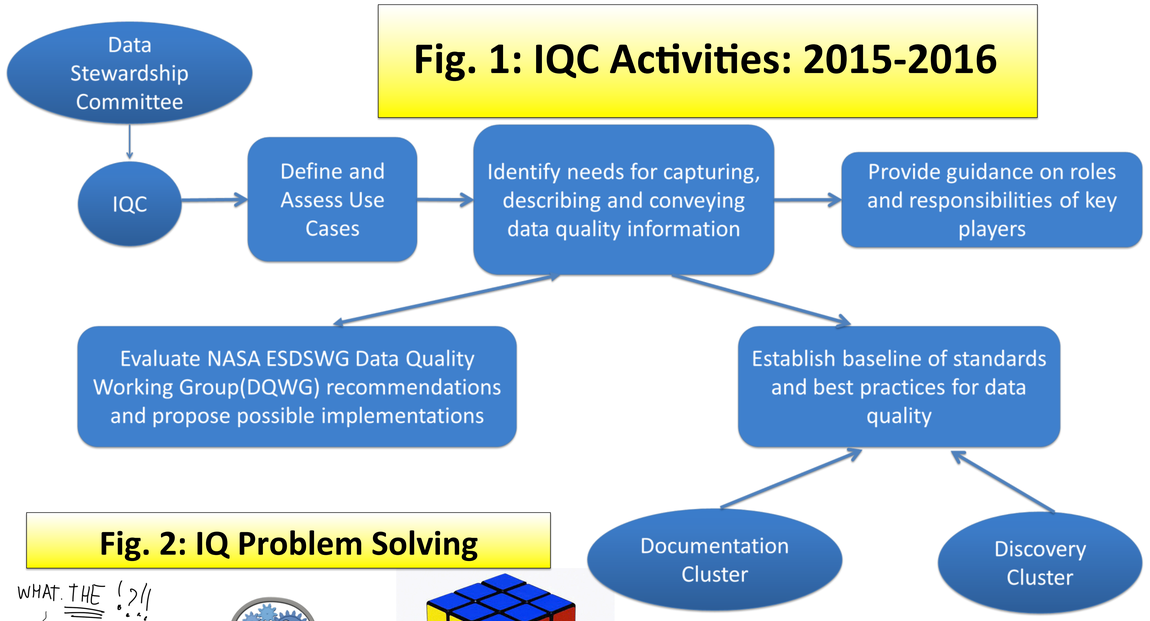
Ensuring and Improving Information Quality for Earth Science Data and Products: Role of the ESIP Information Quality Cluster

David Moroni[1] (David.F.Moroni@jpl.nasa.gov)
Hampapuram Ramapriyan[2,3] (Hampapuram.Ramapriyan@ssaihq.com), Ge Peng[4,5] (Ge.Peng@noaa.gov)
[1]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA,
[2]NASA Goddard Space Flight Center, [3] Science Systems and Applications, Inc.,
[4]NOAA's Cooperative Institute for Climate and Satellites - North Carolina (CICS-NC), [5]NOAA's National Centers for Environmental Information (NCEI),

Abstract: Quality of Earth science data products is always of concern to users regardless of the type of products. The following represent four unique aspects, the collection of which constitutes information quality: science, product, stewardship and service. With increasing requirements on ensuring and improving information quality coming from multiple government agencies and throughout industry, there have been considerable efforts toward improving information quality during the last decade, much of which has not been well vetted in a collective sense until recently. Given this rich background of prior work, the Information Quality Cluster (IQC), established within the Federation of Earth Science Information Partners (ESIP) in 2011, and reactivated in the summer of 2014, has been active with membership from multiple government agencies, institutions, and organizations. The vision of IQC is "**to become internationally recognized as an authoritative and responsive resource of information and guidance to data providers on how best to implement data quality standards and best practices for their science data systems, datasets, and data/metadata dissemination services.**" IQC's objectives and activities, aimed at ensuring and improving information quality for Earth science data and products, are discussed briefly, including recent development and evaluation of use cases. During 2016, several members of the IQC have led the development and assessment of four use cases. The purpose of IQC's use cases is to identify issues related to collecting and conveying quality information to users, and recommending improvements for implementation by data producers and data distributors. An accompanying poster (Peng, Ramapriyan, and Moroni: http://commons.esipfed.org/node/9625) presents in more detail how various maturity matrices address and support the four aspects of information quality mentioned above.

Fig. 1: IQC Activities: 2015-2016



Information Quality Domains:

1. ***Science***: accuracy, precision, uncertainty, validity, and suitability for use (fitness for purpose).
2. ***Product***: completeness of science quality assessments, documentation, metadata, provenance, context, etc
3. ***Stewardship***: integration, management, and preservation of data and metadata.
4. ***Service***: accessibility, searchability, trustworthiness, usability, subject matter expertise, user assistance.

Fig. 2: IQ Problem Solving

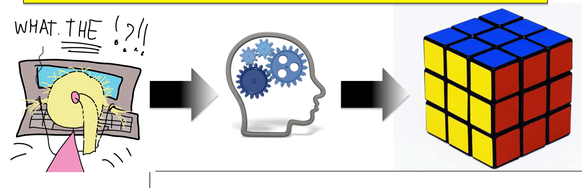
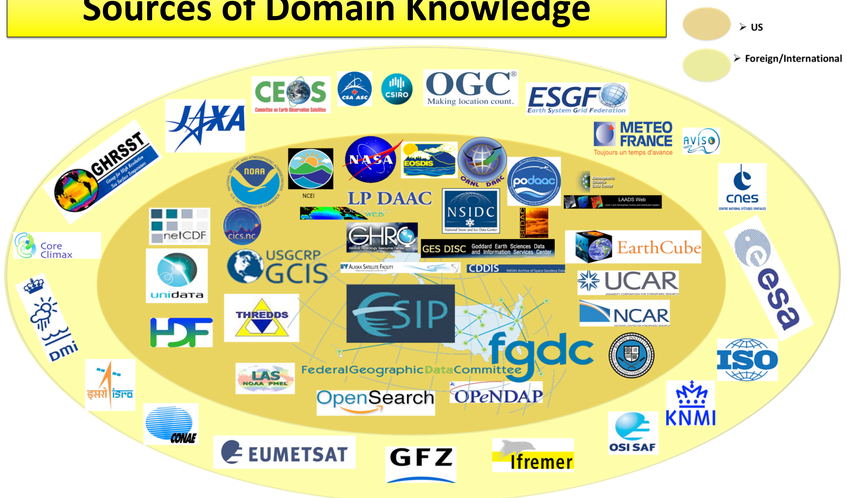


Fig. 3: Scope of Mutual Influence and Sources of Domain Knowledge



Acknowledgements: Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsement by the United States Government or the Jet Propulsion Laboratory, California Institute of Technology. These activities were carried out across multiple United States government-funded institutions (noted above) under contracts with the National Aeronautics and Space Administration (NASA) and the National Oceanic and Atmospheric Administration (NOAA). Government sponsorship acknowledged.

2016 IQ Use Case Evaluation and Recommendations Summary:

1. "**Dataset Rice Cooker Theory**" - Bob Downs, David Moroni, and Joseph George
   - **Problem**: Data products assimilated using heterogeneous data sources are plagued with unknown and/or incomplete quality characterization, thus leading to improper handling of input data used in assimilation.
   - **Solutions**: a) Full disclosure of errors, uncertainties, and all available quality information of input data; b) Understanding and properly conveying how above quality issues impact the final data product; c) analysis as a function of time and at the pixel level; d) ensure that the integrity and intended use of data is upheld through proper integration of heterogeneous data sources.
2. "**Appropriate Amount/Extent of Documentation for Data Use**" - Ge Peng, Steve Olding, and Lindsey Harriman
   - **Problem:** Reducing the amount of time required for the typical data user to determine what information they need and increasing the level of satisfaction via user feedback on the proper use of information provided to the user.
   - **Solutions:** a) Provide tiered access to user-relevant information – DOI landing pages, User guides, ATBDs; b) effective product filtering algorithms and tools; c) user feedback mechanism.
3. "**Improving Use of SBC LTER Data Portal**" - Margaret O'Brien, Sophie Hou, Ross Bagwell, and Hampapuram Ramapriyan.
   - **Problem**: The Santa Barbara Coastal (SBC) Long Term Ecological Research (LTER) data portal is seeking how to improve the user's ability to discover the SBC LTER data collections and help the user to better understand the data characteristics as a means of determining if a dataset of interest is suitable for a user's intended usage.
   - **Solutions**: a) Data providers should supply sufficient information on data quality to help the SBC LTER data portal improve categorization of various datasets; b) refine the keyword search algorithm; c) improve the data discovery interface using usability evaluations; d) advise data producers to use proper terms in "tag" datasets; e) review datasets for completeness and quality of information.
4. "**Citizen Science**" – Ruth Duerr, Han Qin, Chung-Lin Shie, and Bhaskar Ramachandran
   - **Problem**: Crowd-sourced data often lacks support from subject-matter experts, thus translating to unknown or poorly understood data characteristics.
   - **Solutions**: a) Training data producers on how to better sample and document each measurement/observation; b) improved communication between data distributors and data producers; c) encourage data users to read associated documentation and provide feedback on use and issues found.