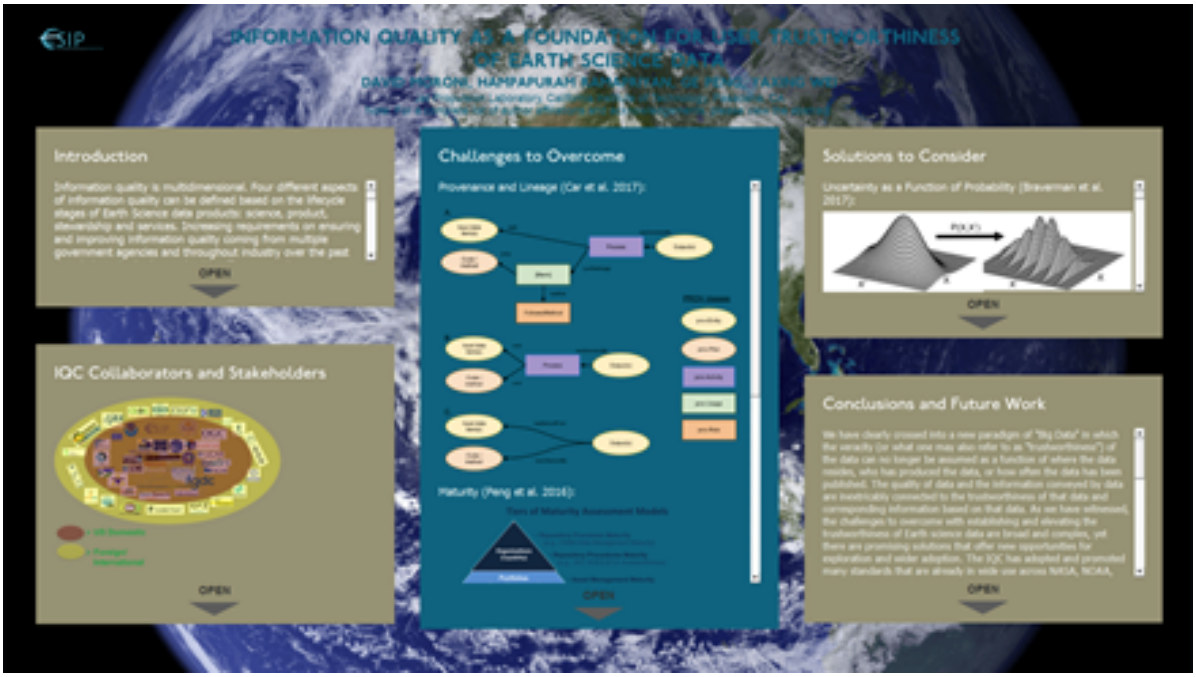
Information Quality as a Foundation for User Trustworthiness of Earth Science Data



David Moroni, Hampapuram Ramapriyan, Ge Peng, Yaxing Wei

Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA
Note: For a complete list of author affiliations and acknowledgements, please view the abstract.

PRESENTED AT:

AGU FALL MEETING
New Orleans
11-15 Dec. 2017
What will *you* discover?

267915

INTRODUCTION

Information quality is multidimensional. Four different aspects of information quality can be defined based on the lifecycle stages of Earth Science data products: science, product, stewardship and services. Increasing requirements on ensuring and improving information quality coming from multiple government agencies and throughout industry over the past decade have birthed considerable efforts toward improving information quality, 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 organizations. The IQC's objectives and activities, aimed at ensuring and improving information quality for Earth science data and products, are intended to be a vital and ongoing contribution toward improving the trustworthiness of Earth science data to a vast and interdisciplinary community of data users. Fast forwarding to 2017, the IQC has led multiple panel sessions to survey and cross-examine the challenges posed in the various aspects of information quality. What was discovered to be most lacking is the transparency of data lineage (i.e., provenance and maturity), uniform methods for uncertainty characterization, and uniform quality assurance data and metadata. While solutions to these types of issues exist, it was also revealed that most data producers have little time to investigate and collaborate toward a consistent approach across all of the Earth science disciplines. The IQC has positioned itself as a community platform to bring together all relevant stakeholders from data producers, repositories, program managers, and the end users. A combination of both well-vetted and "trailblazing" solutions are presented to address how data trustworthiness can be elevated and maintained through optimized extraction, curation, and dissemination of information quality artifacts.

IQC COLLABORATORS AND STAKEHOLDERS

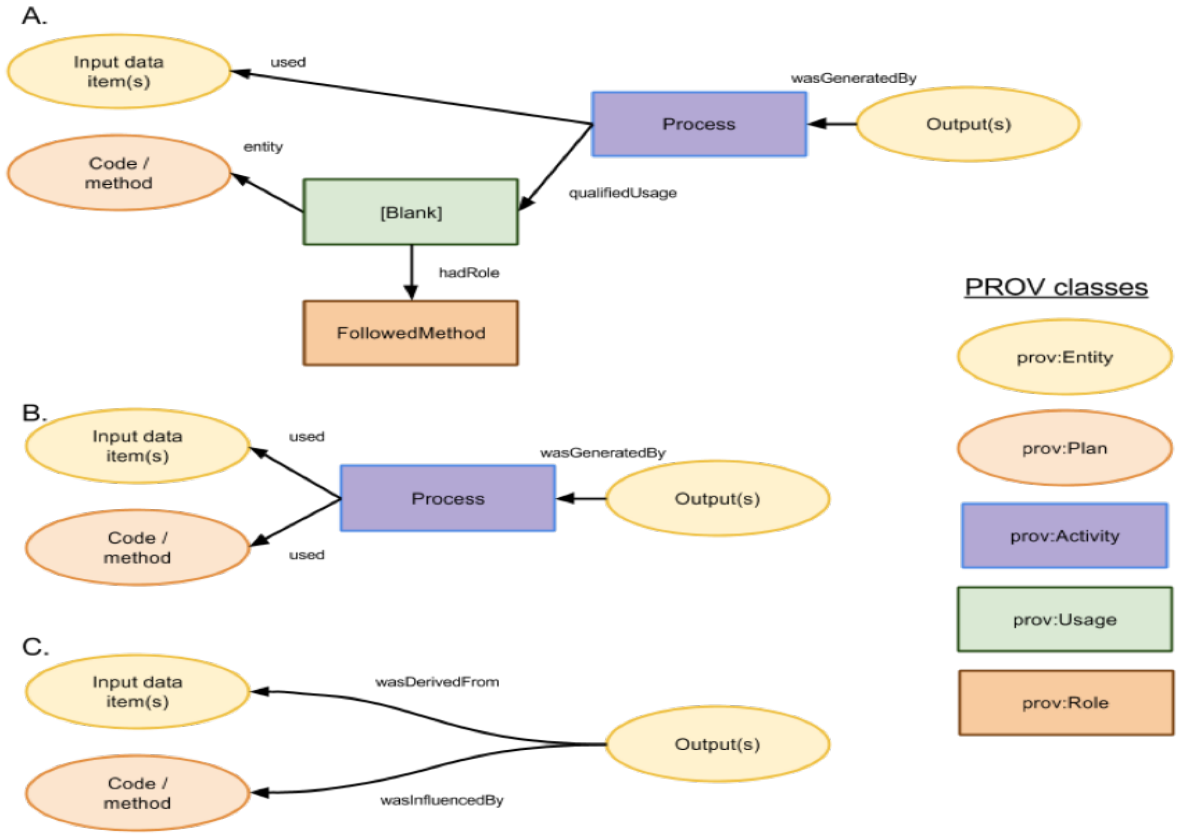
➢ US Domestic

➢ Foreign/ International

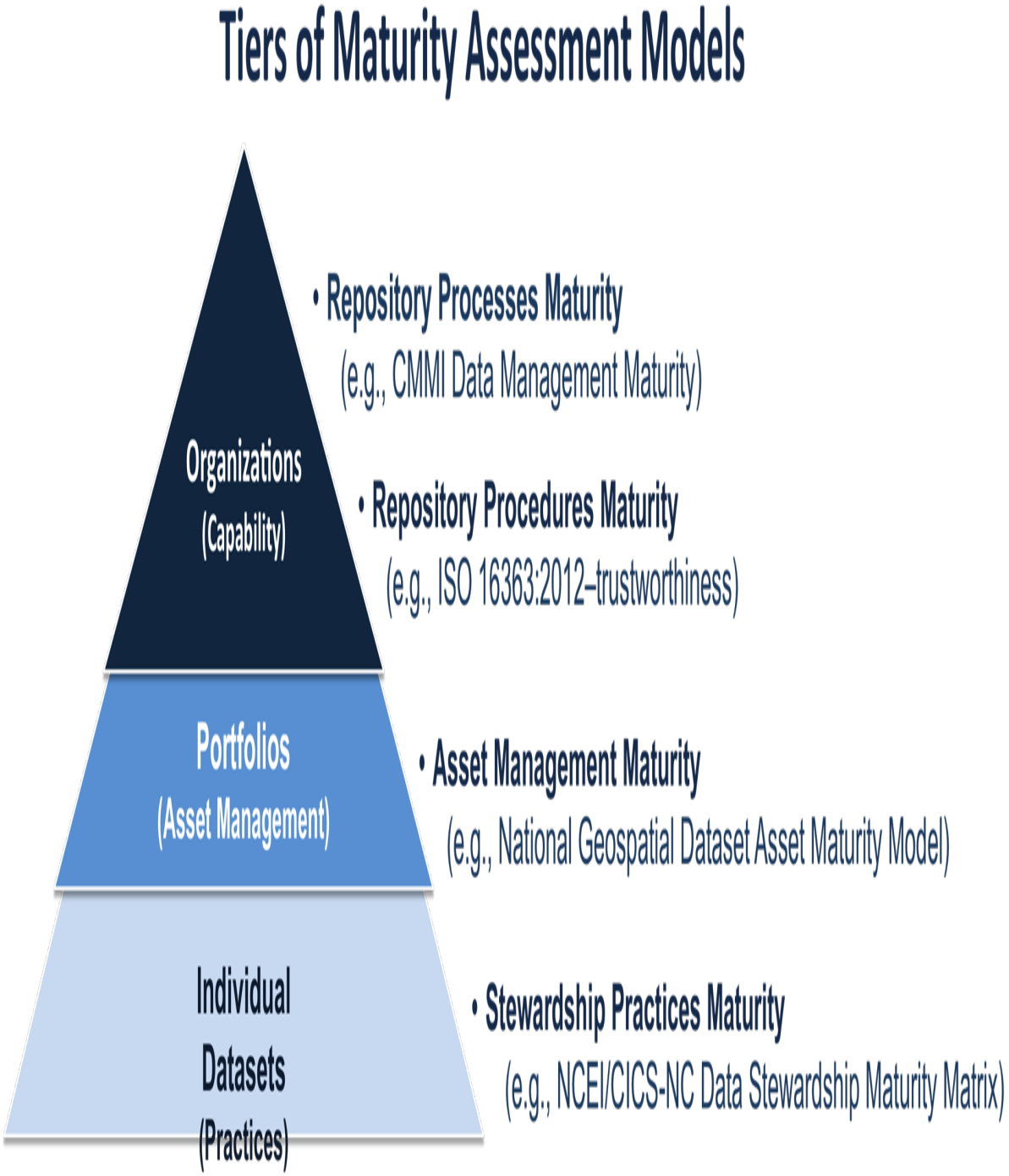
Disclaimer: Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsement by the United States Government or the Jet Propulsion Laboratory, California Institute of Technology.

CHALLENGES TO OVERCOME

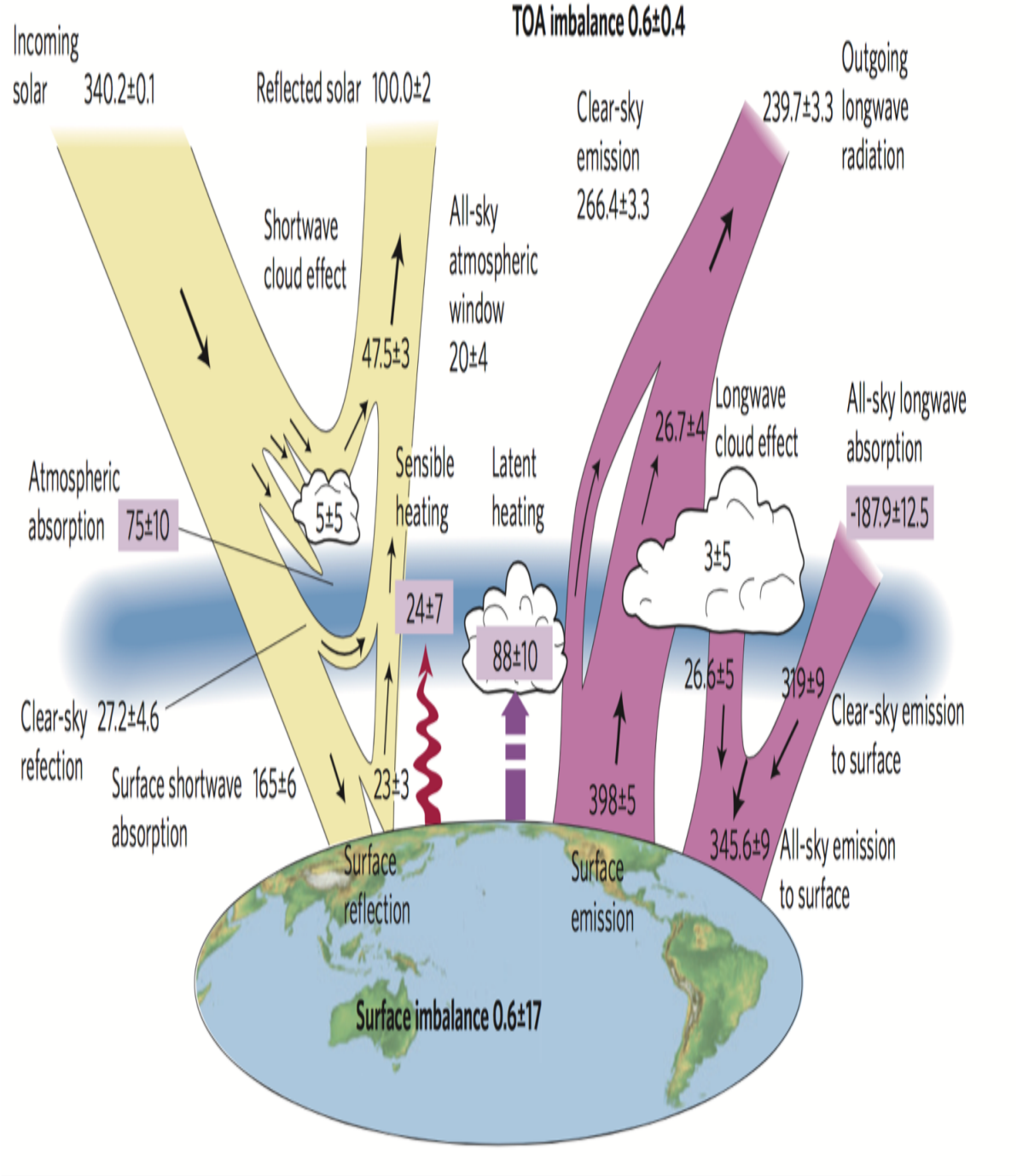
Provenance and Lineage (Car et al. 2017):



Maturity (Peng et al. 2016):



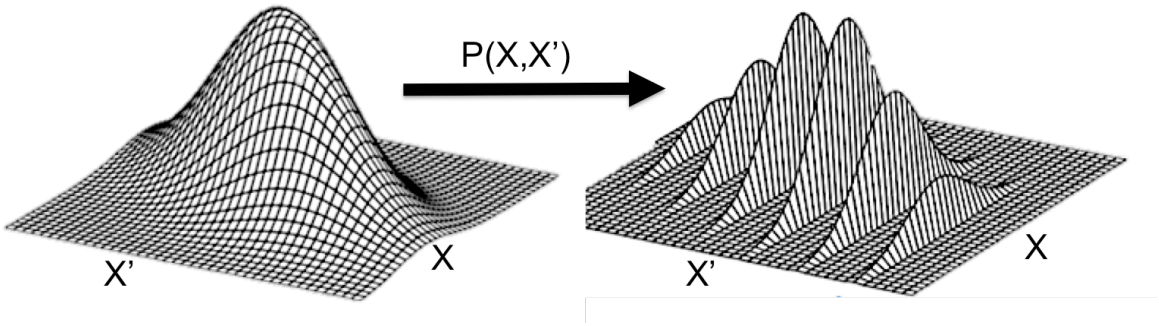
Uncertainty (example of Earth's energy balance, Stephens et al. 2012):



Earth's energy balance is an example of why the impact of measurement uncertainty becomes more pronounced when in combination with other measurements that are required to compute a single geophysical estimate.

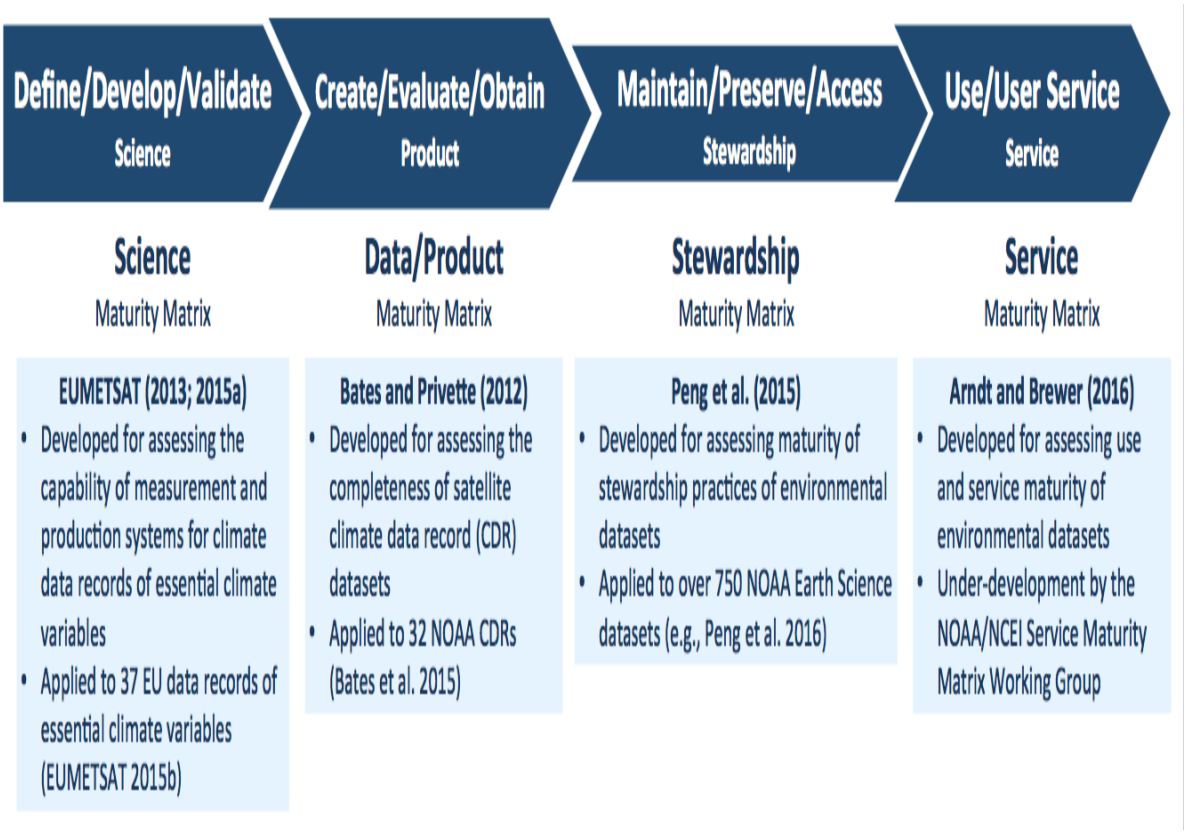
SOLUTIONS TO CONSIDER

Uncertainty as a Function of Probability (Braverman et al. 2017):

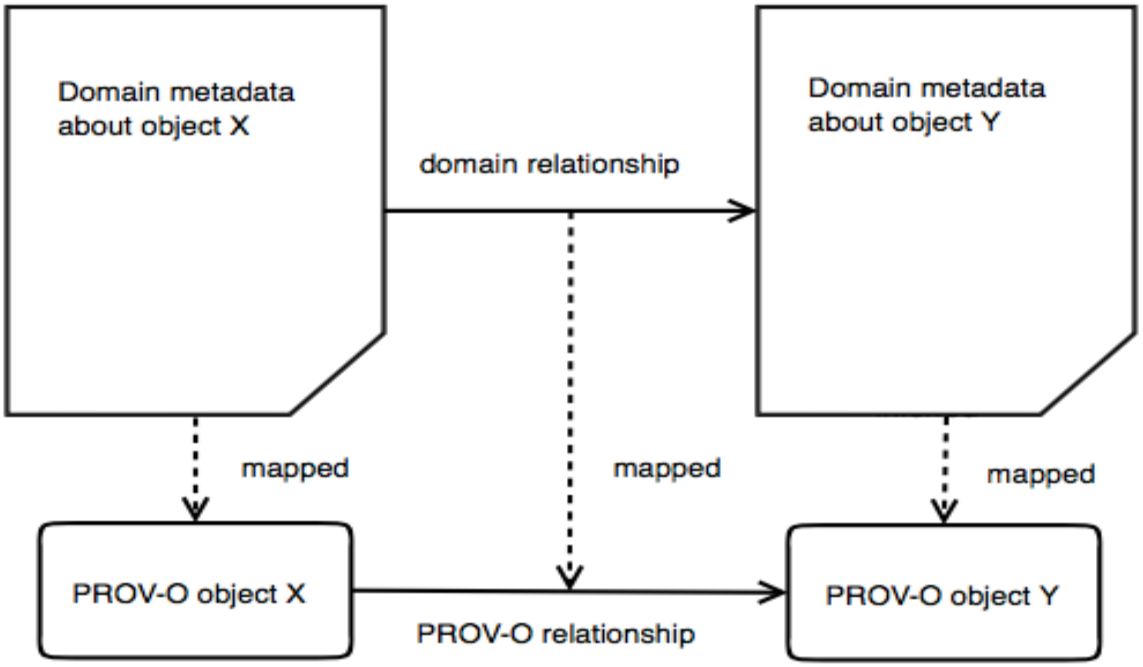


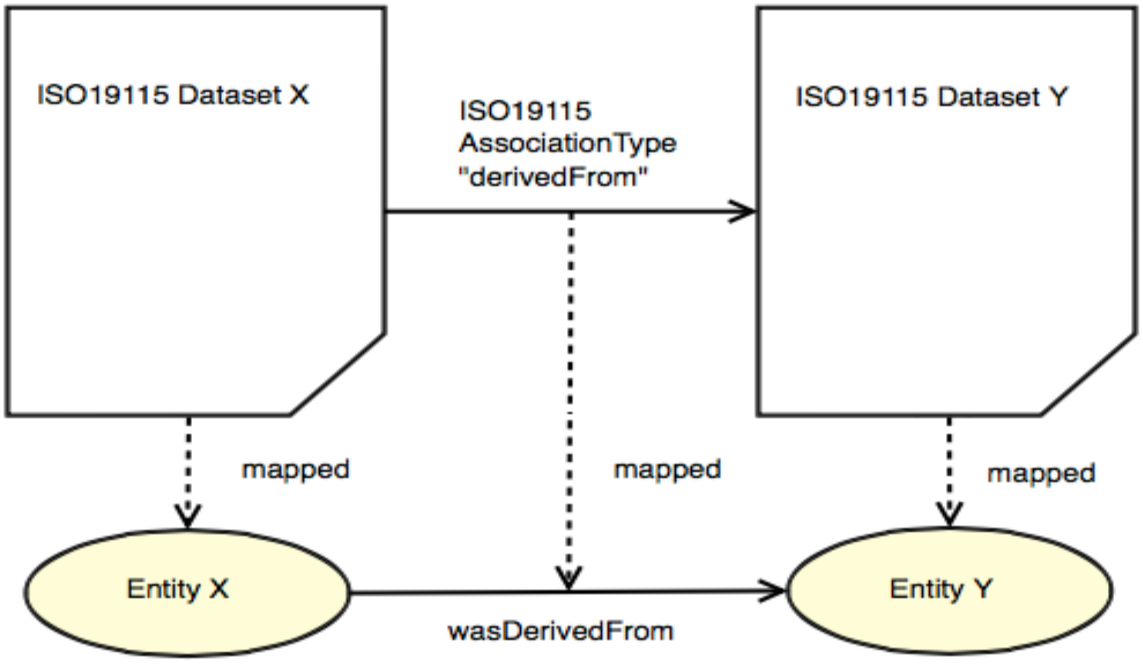
P(X,X') represents an arbitrary multi-variate probability distribution function in which X represents the measured quantity while X' represents the state estimate (or similation) of the measured quantity. Probability-based uncertainty provides constrained levels of confidence in the uncertainty estimates compared to mean square error (MSE) and provides more physically meaningful uncertainty information that conforms with the shape of any data distribution, accounting for both outliers and inliers.

Maturity Assessment as a Function of Dataset Lifecycle Stages (Peng et al. 2016):



Provenance Patterns can be used to clarify Dataset Relationships and also facilitate Data Reuse Fitness Assessments (Car et al. 2017).





Use Case Collection and Evaluation:

	Capture	Describe	Facilitate Discovery	Enable Use
Science	9	16	9	5
Product	11	18	10	5
Stewardship	7	11	6	6
Service	5	10	6	5

Table depicts the number of use cases (out of a total of 20) that were evaluated by the IQC and determined to correspond to specific Data Quality Management Phases (columns) and Information Quality aspects (rows). For more details, see Rama et al. 2017.

CONCLUSIONS AND FUTURE WORK

We have clearly crossed into a new paradigm of "Big Data" in which the veracity (or what one may also refer to as "trustworthiness") of the data can no longer be assumed as a function of where the data resides, who has produced the data, or how often the data has been published. The quality of data and the information conveyed by data are inextricably connected to the trustworthiness of that data and corresponding information based on that data. As we have witnessed, the challenges to overcome with establishing and elevating the trustworthiness of Earth science data are broad and complex, yet there are promising solutions that offer new opportunities for exploration and wider adoption. The IQC has adopted and promoted many standards that are already in wide use across NASA, NOAA, and across a variety of international agencies and organizations who are responsible for the production and stewardship of Earth science data. Such standards and ongoing work include: NOAA/NCEI Data Stewardship Maturity Model, GEOSS Data Management Principles, GEOSS Data Management Plan Implementation Guidelines (DMP-6 on Quality Control), WDS/RDA Assessment of Data Fitness for Use Working Group, NASA Data Quality Working Group, WMO Guidelines on Climate Data Management, QA4EO'11, ISO/TS 19157-2013, Climate and Forecasting (CF) Quality Flagging Schema, and ISO 19158 (providing a "quality assurance framework"). The IQC's network of collaboration is still growing and open to new membership. A white paper on the topic of data uncertainty in an interdiscplinary Earth science remote sensing context is being developed with the intention of providing a statistical foundation for proper uncertainty estimation, interpretation and application; feedback on this whitepaper is welcome from all who are willing to contribute. In addition to hosting sessions and invited speakers on special topics at the bi-annual ESIP conferences, the IQC features monthly teleconferences promoting the breadth of information and data quality issues and solutions from experts in their respsective fields. More can be learned about past and ongoing IQC activities by visiting http://wiki.esipfed.org/index.php/Information_Quality.

DISCLOSURES

> CLICK TO ENTER CONTENT

ABSTRACT

Information Quality as a Foundation for User Trustworthiness of Earth Science Data

David Moroni[1], Hampapuram Ramapriyan[2], Ge Peng[3], Yaxing Wei[4]

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

[2]Science Systems and Applications, Inc. / NASA Goddard Space Flight Center

[3]Cooperative Institute for Climate and Satellites - North Carolina (CICS-NC) / NOAA National Centers for Environmental Information

[4]Oak Ridge National Laboratory, Oak Ridge, TN

Abstract

Information quality is multidimensional. Four different aspects of information quality can be defined based on the lifecycle stages of Earth Science data products: science, product, stewardship and services. 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 organizations. The IQC's objectives and activities, aimed at ensuring and improving information quality for Earth science data and products, are also considered vital toward improving the trustworthiness of Earth science data to a vast and interdisciplinary community of data users. During 2016, several members of the IQC have led the development and assessment of four use cases. This was followed up in 2017 with multiple panel sessions at the 2017 Winter and Summer ESIP Meetings to survey the challenges posed in the various aspects of information quality. What was discovered to be most lacking is the transparency of data lineage (i.e., provenance and maturity), uniform methods for uncertainty characterization, and uniform quality assurance data and metadata. While solutions to these types of issues exist, most data producers have little time to investigate and collaborate to arrive at and conform to a consensus approach. The IQC has positioned itself as a community platform to bring together all relevant stakeholders from data producers, repositories, program managers, and the end users. A combination of both well-vetted and "trailblazing" solutions are presented to address how data trustworthiness can be elevated and maintained through optimized extraction, curation, and dissemination of information quality artifacts.

Acknowledgements: 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). 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. Government sponsorship acknowledged.

REFERENCES

Peng, G., H. Ramapriyan, D. F. Moroni, 2016: The State of Building a Consistent Framework for Curation and Presentation of Earth Science Data Quality. Poster: IN41C.1666, AGU 2016 Fall Meeting, 12 – 16 December 2016, San Francisco, CA, USA.

Ramapriyan, H., G. Peng, D. Moroni, C-L Shie**, "**Ensuring and Improving Information Quality for Earth Science Data and Products", *D-Lib Magazine,* July 2017, DOI: https://doi.org/10.1045/july2017-ramapriyan

Stevens, G. L., J. Li, M. Wild, C.-A. Clayson, N. Loeb, S. Kato, T. L'Ecuyer, P. W. Stackhouse Jr., M. Lebsock, T. Andrews, "An update on Earth's energy balance in light of the latest global observations", *Nature Geoscience*, September 2012, 5, 691-696, DOI: https://doi.org/10.1038 /ngeo1580

CHAT SETTINGS

Info

Choose Date and Time: - Choose Date and Time -

Edit message:

```
The chatroom will open 10 minutes before start.

If you missed the chat, but would like to contact me, please use
the Contact Author button at the bottom of my presentation.
```

SAVE SETTINGS START CHAT

CHAT INFORMATION

THE CHATROOM WILL OPEN 10 MINUTES BEFORE START. IF YOU MISSED THE CHAT, BUT WOULD LIKE TO CONTACT ME, PLEASE USE THE CONTACT AUTHOR BUTTON AT THE BOTTOM OF MY PRESENTATION.

JOIN CHAT

IN ORDER TO JOIN CHAT, PLEASE ENTER YOUR NAME AND EMAIL ADDRESS

Enter your name here:

Enter your email address he

JOIN CHAT