CISESS
Cooperative Institute for
Satellite Earth System Studies

CODATA 2019 Beijing
19–20 September 2019
Beijing, China

A Holistic Framework for Supporting Evidence-Based Institutional Research Data Management

Ge Peng[1,2], PhD

In Collaboration With

Jeffrey L. Privette[2], Edward Kearns[3], Nancy Ritchey[2], Otis Brown[1], Curt Tilmes[4], Sky Bristol[5], Hampapuram Ramapriyan[4,6], and Thomas Maycock[1]

[1] North Carolina Institute for Climate Studies (NCICS), North Carolina State University, Asheville, NC 28801 USA;
[2] NOAA National Centers for Environmental Information (NCEI), USA;
[3] NOAA Office of the Chief Information Officer (OCIO), USA;
[4] NASA Goddard Space Flight Center (GSFC), USA;
[5] United States Geological Survey (USGS), USA;
[6] Science Systems and Applications, Inc. (SSAI), USA

CODATA 2019 Meeting, Beijing, China, September 19, 2019

NC STATE UNIVERSITY NCICS NOAA NASA USGS science for a changing world

Main Challenges for Institutional RDM

- Increasing Quantity and Variety of Digital Research Data,
- Evolving Users Requirements,
- Increased Federal Requirements,
- Multi-Perspectives of Data Management and Stewardship,
- Multi-Dimensions of Data and Information Quality.

Increased Federal Requirements → Quality Attributes

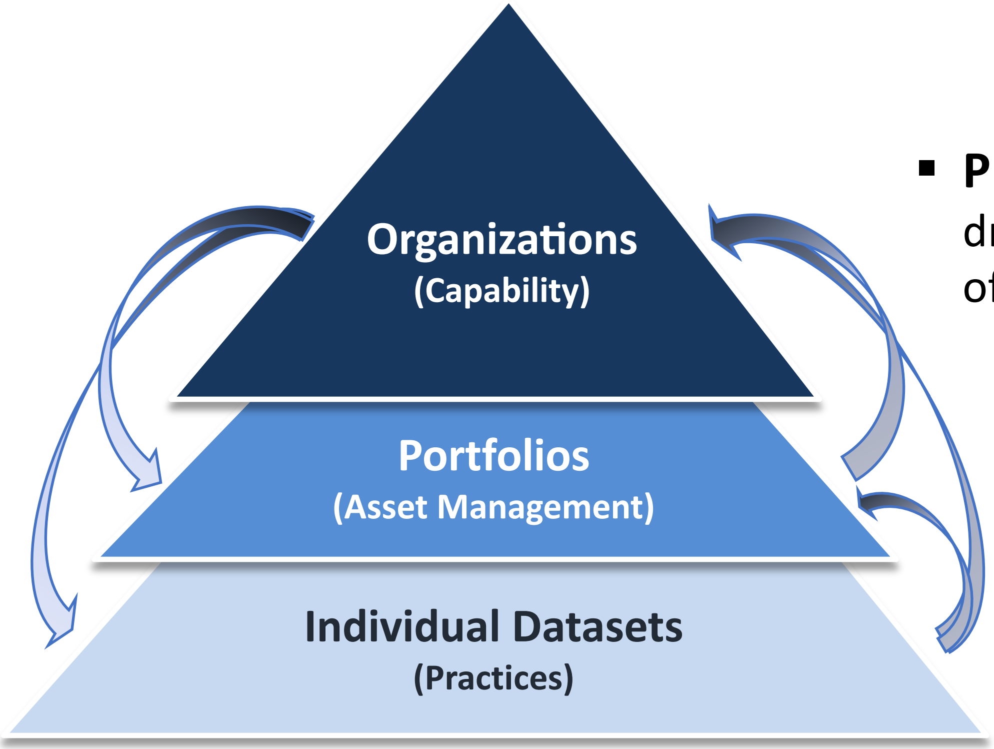
- US Public Laws
  - Information Quality Act (106-554 2000),
  - DATA Act (113-101 2014),
  - OPEN Government Data Act (115-435 2019, Title II).
- US Federal Policies
  - Information Quality Act Guidelines (OMB 2002), revised in 2019,
  - Open Data Policy – Managing Information as an Asset (OMB 2013),
  - Increasing access (OSTP 2013)

Important Quality Attributes for U.S. Federally Funded Digital Research Data Include:

- Accuracy, Integrity, Utility, Transparency, Traceability, Preservability, Accessibility, Interoperability, Usability.

- **Compliance reporting with support evidences**

Multi-Perspectives of Institutional RDM



- **Process** (过程)
  driven by achievement
  of a desired outcome
- **Procedure** (流程)
  driven by completion
  of the task
- **Practices** (实践)
  actual use of something

Data Production: Processes ensure a data product is produced in a right way while practices ensure the produced product is a right one.

There Are Many Data Quality Attributes!

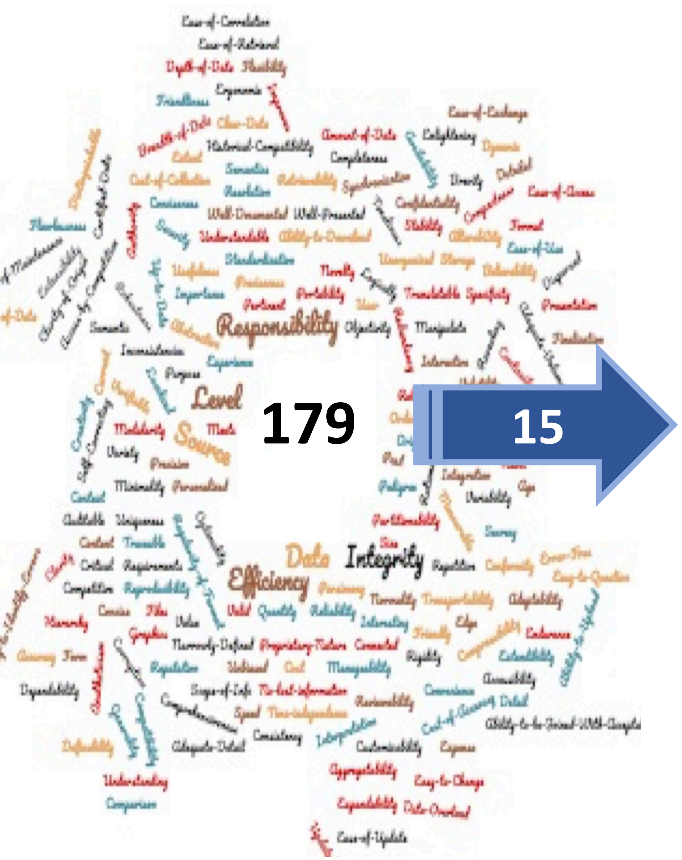
Wang and Strong 1996, *J. Management Info. Sys.*

Many are overlapping

- Accuracy
- Correctness
- Free from bias
- Validated
- …

Data quality is not just about accuracy any more!

Multi-Dimensions of Data and Information Quality



Quality Attributes	Dimensions
accuracy, objectivity, believability, reputation,	➢ Intrinsic
relevance, timeliness, completeness, value-added, appropriate amount of data,	➢ Contextual
ease of understanding, concise representation and representational consistency, interpretability,	➢ Representational
accessibility, access security.	➢ Accessibility

(Wang and Strong 1996, *J. Management Info. Sys.)*

Multi-Dimensions of Data and Information Quality

Perspective

Based on open data and data sharing principles

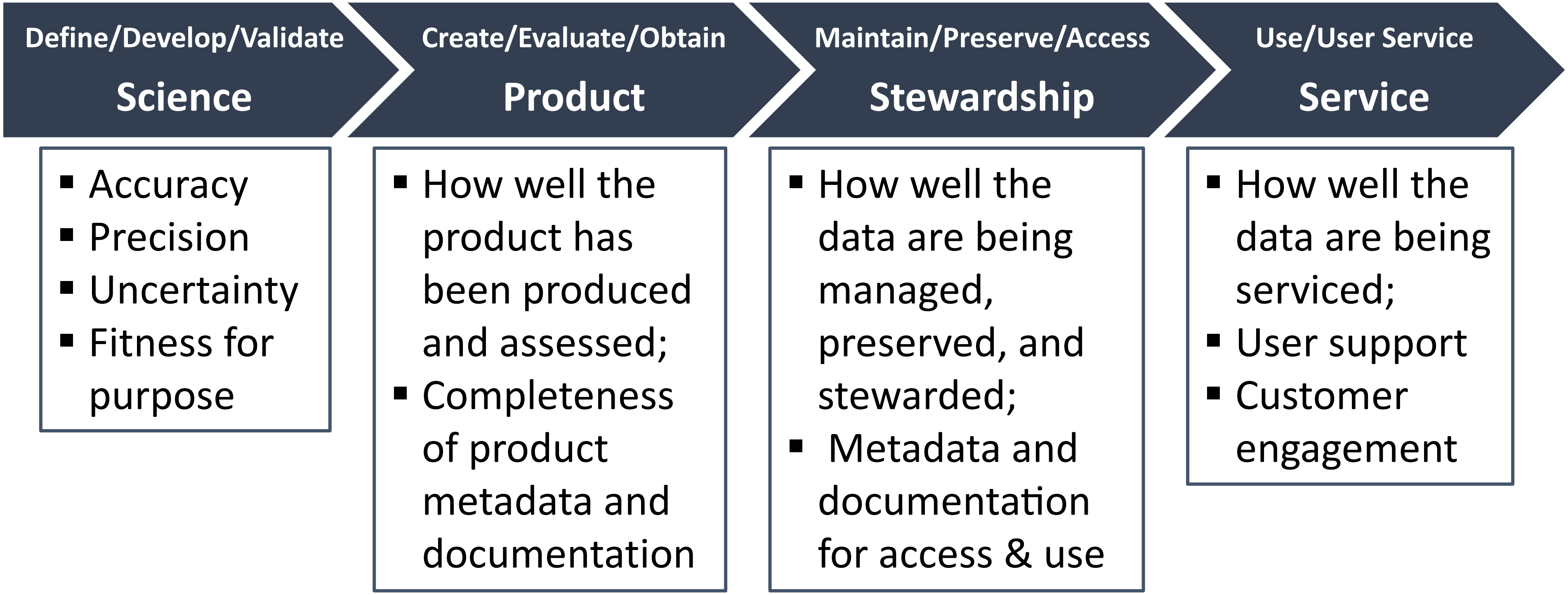
(Wilkinson et al. 2016, *Scientific Data*)

Quality Attributes

- **Findability,**
- **Accessibility,**
- **Interoperability,**
- **Reusability.**

Multi-Dimensions of Data and Information Quality

Stages of Data Product Lifecycle



Define/Develop/Validate

Science

- Accuracy
- Precision
- Uncertainty
- Fitness for purpose

Create/Evaluate/Obtain

Product

- How well the product has been produced and assessed;
- Completeness of product metadata and documentation

Maintain/Preserve/Access

Stewardship

- How well the data are being managed, preserved, and stewarded;
- Metadata and documentation for access & use

Use/User Service

Service

- How well the data are being serviced;
- User support
- Customer engagement

(Ramapriyan et al. 2017, *D.-Lib Magazine*)

Needs to be Holistic and Integrated

Institutional Research Data Management

- A lot of Moving Parts, many may have already been in place;
- Cross-Department;
- Cross-Discipline.

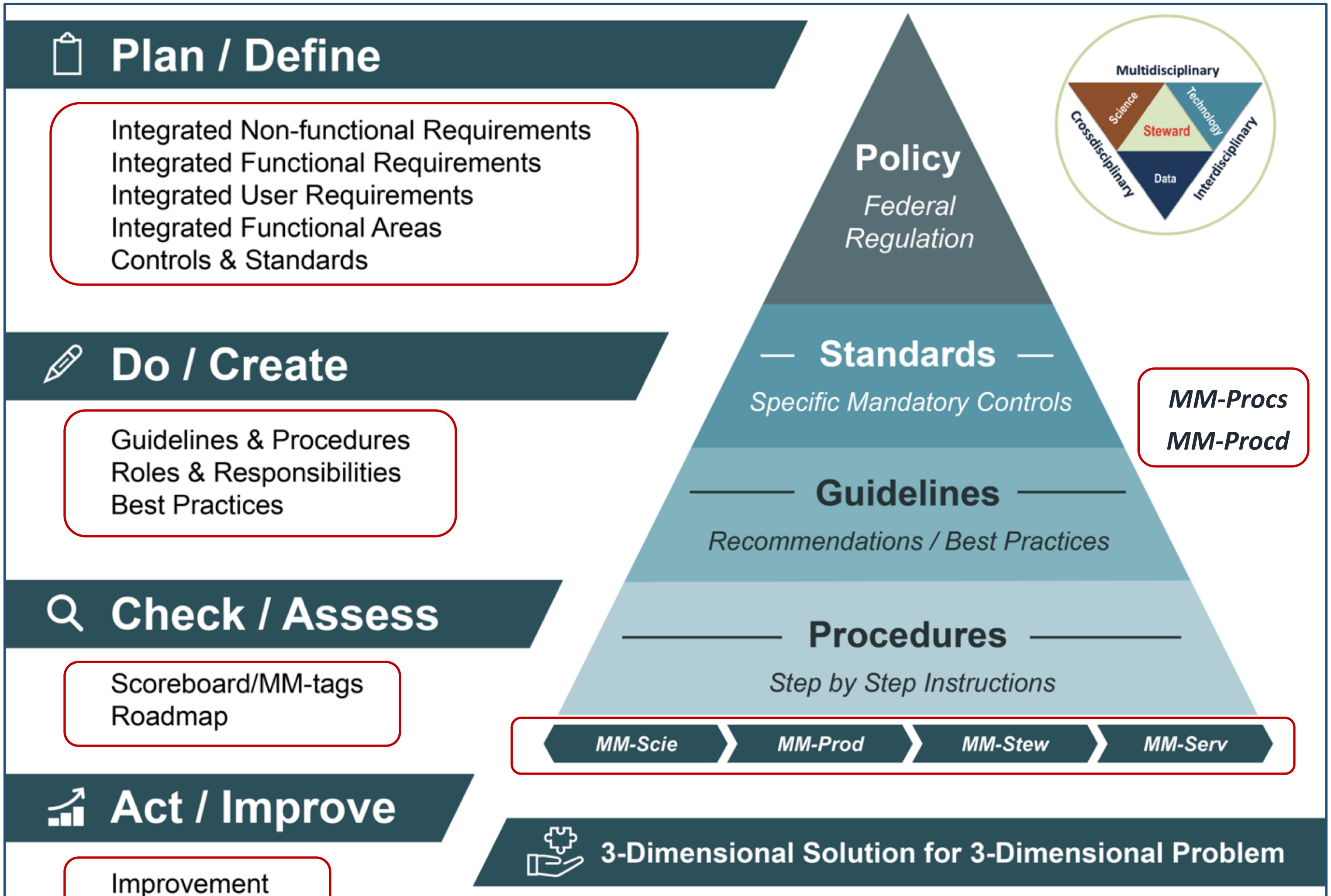
Institutions need to demonstrate the compliance by reporting with support evidences!

But How? Where to Start?

Needs to have a Holistic and Integrated Approach:

- To be utilized without much upfront cost,
- Enterprise-wide,
- Evidence-based,
- Support continuous improvement.

High-Level, Holistic Framework For Institutional RDM



Plan / Define

Integrated Non-functional Requirements
Integrated Functional Requirements
Integrated User Requirements
Integrated Functional Areas
Controls & Standards

Do / Create

Guidelines & Procedures
Roles & Responsibilities
Best Practices

Check / Assess

Scoreboard/MM-tags
Roadmap

Act / Improve

Improvement
Standardization

Policy
Federal Regulation

Standards
Specific Mandatory Controls

Guidelines
Recommendations / Best Practices

Procedures
Step by Step Instructions

MM-Procs
MM-Procd

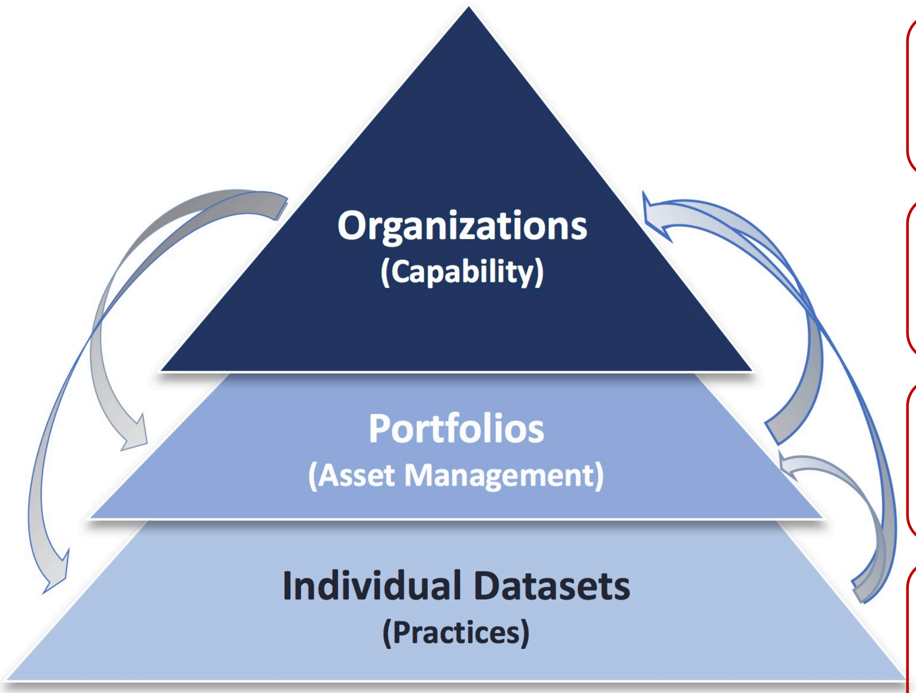
MM-Scie *MM-Prod* *MM-Stew* *MM-Serv*

3-Dimensional Solution for 3-Dimensional Problem

(Adapted from Peng et al. 2018, *Data Science Journal*)

Examples of Maturity Assessment Models

Tiers of Maturity Assessment within Context of Scientific Data Stewardship

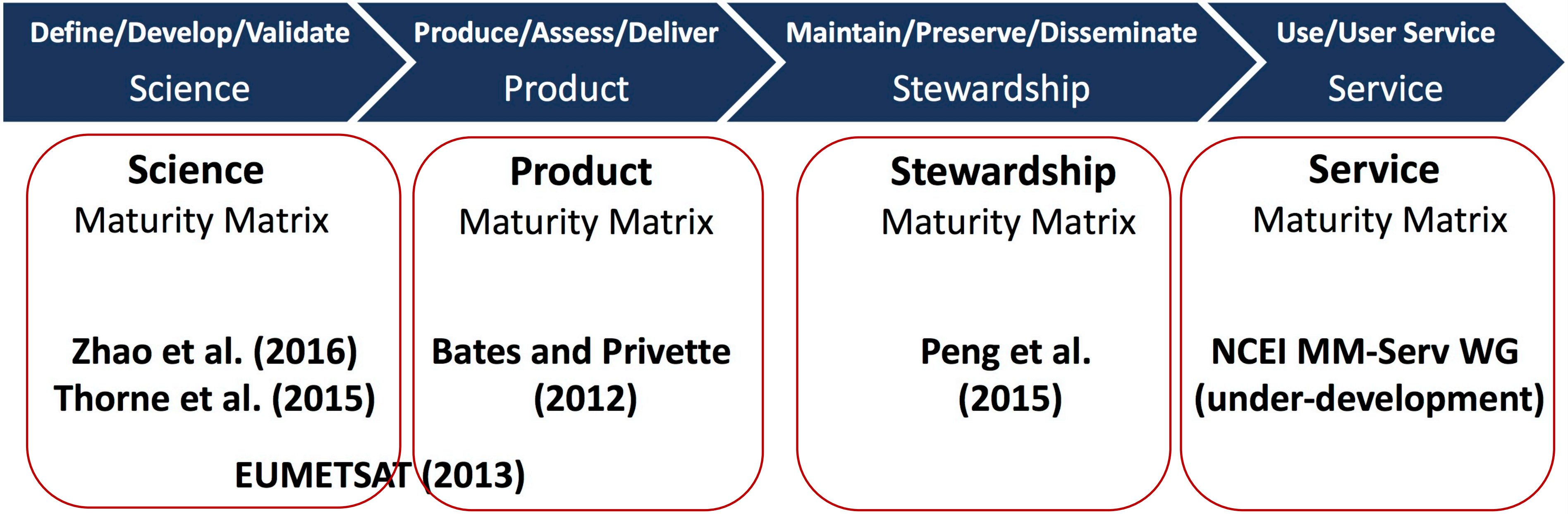


(Peng 2018, *Data Science Journal*)

- **Repository Processes Maturity**
  e.g., CMMI Data Management Maturity Model (CMMI 2014)
- **Repository Procedures Maturity**
  e.g., ISO Repository Trustworthiness Certification (ISO 16363 2012)
- **Asset Management Maturity**
  e.g., NGDA Lifecycle Maturity Assessment Model (FGDC 2015)
- **Stewardship Practices Maturity**
  e.g., NCEI/CICS-NC Data Stewardship Maturity Matrix (Peng et al. 2015)

Examples of Dataset Maturity Assessment Models

Data Product Lifecyle-Stage-Based Maturity Assessment Models



Define/Develop/Validate

Science

Produce/Assess/Deliver

Product

Maintain/Preserve/Disseminate

Stewardship

Use/User Service

Service

Science Maturity Matrix

Zhao et al. (2016)
Thorne et al. (2015)

Product Maturity Matrix

Bates and Privette (2012)

EUMETSAT (2013)

Stewardship Maturity Matrix

Peng et al. (2015)

Service Maturity Matrix

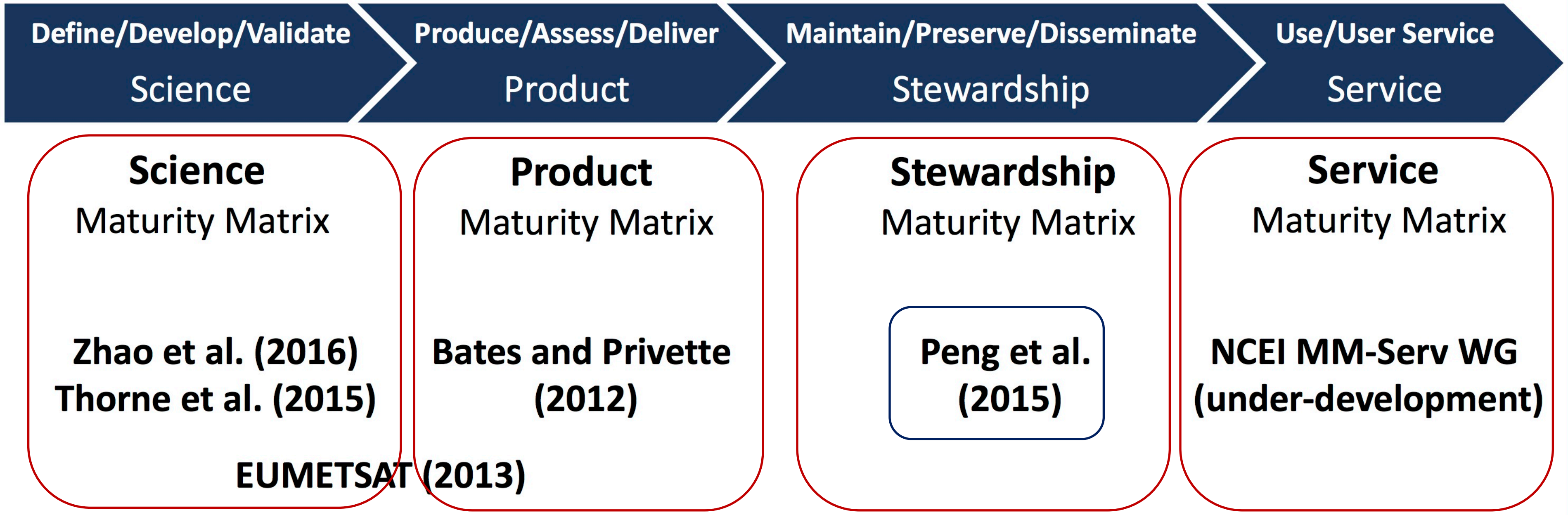
NCEI MM-Serv WG (under-development)

(Peng 2018, *Data Science Journal*)

- **WMO Stewardship Maturity Matrix for Climate Data,**
- **CEOS Data Management and Stewardship Maturity Matrix.**

Examples of Dataset Maturity Assessment Models

Data Product Lifecyle-Stage-Based Maturity Assessment Models



Define/Develop/Validate
Science

Produce/Assess/Deliver
Product

Maintain/Preserve/Disseminate
Stewardship

Use/User Service
Service

Science
Maturity Matrix

Zhao et al. (2016)
Thorne et al. (2015)

Product
Maturity Matrix

Bates and Privette (2012)

EUMETSAT (2013)

Stewardship
Maturity Matrix

Peng et al. (2015)

Service
Maturity Matrix

NCEI MM-Serv WG (under-development)

Scientifically sound and utilized

Fully documented and transparent

Well-preserved and integrated

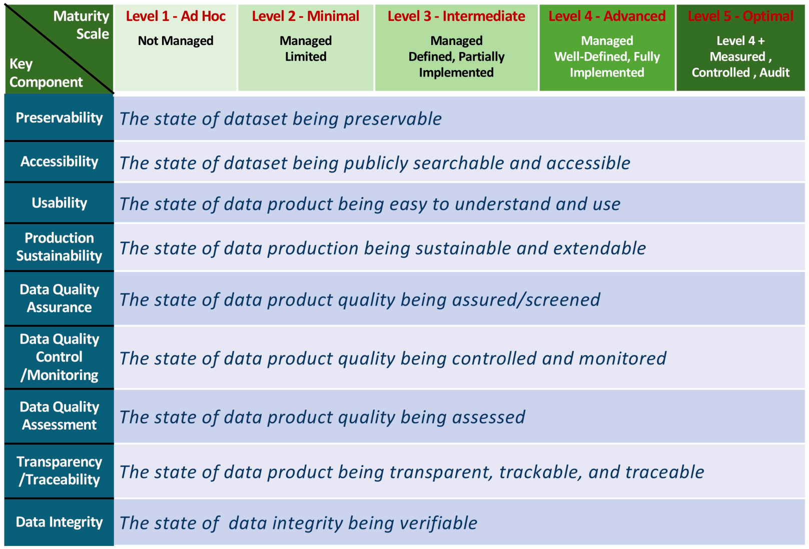
Readily obtainable and usable

Possible to Assess the Maturity of Individual Datasets?

(NOAA *OneStop* Application of a Data Stewardship Maturity Matrix)

What Is the DSMM?

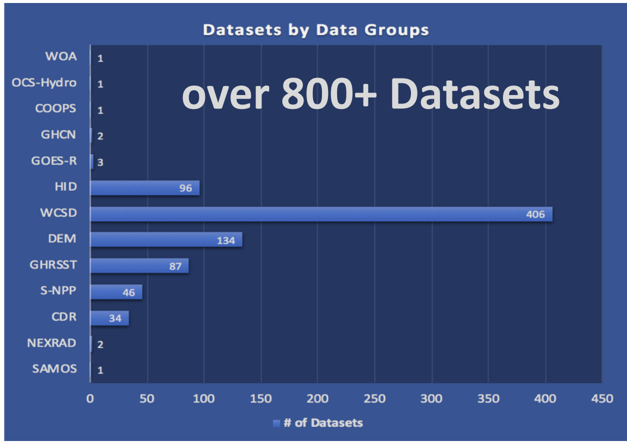
- **A Unified Framework** for measuring stewardship practices applied to individual data products,
- **Developed Jointly** by domain Subject Matter Experts (i.e., data management, science, and technology),
- **Leveraged** institutional knowledge and community best practices and standards,



Maturity Scale / Key Component	Level 1 - Ad Hoc Not Managed	Level 2 - Minimal Managed Limited	Level 3 - Intermediate Managed Defined, Partially Implemented	Level 4 - Advanced Managed Well-Defined, Fully Implemented	Level 5 - Optimal Level 4 + Measured , Controlled , Audit
Preservability	*The state of dataset being preservable*				
Accessibility	*The state of dataset being publicly searchable and accessible*				
Usability	*The state of data product being easy to understand and use*				
Production Sustainability	*The state of data production being sustainable and extendable*				
Data Quality Assurance	*The state of data product quality being assured/screened*				
Data Quality Control /Monitoring	*The state of data product quality being controlled and monitored*				
Data Quality Assessment	*The state of data product quality being assessed*				
Transparency /Traceability	*The state of data product being transparent, trackable, and traceable*				
Data Integrity	*The state of data integrity being verifiable*				

- Used and reused by various data management and stewardship organizations,
- Used to curate structured, rich, machine and human readable quality information metadata and documents.

(Peng et al. 2019, *Data Science Journal; ncics.org/dsmm*)



Key Takeaways

Institutional Research Data Management:

- is a multi-perspective and multi-dimensional problem,
- requires an integrated data-centric framework.

Our framework

- follows the Plan-Do-Check-Act (PDCA) cycle,
- provides a tool to address RDM activities as a consistent, integrated, dataset-centric system,
- includes the application of maturity assessment models,
- supports informed decision-making process.

References

Peng, G., 2018: The state of assessing data stewardship maturity – An overview. Data Science Journal. 17, doi: 10.5334/dsj-2018-007.

Peng, G., J.L. Privette, C. Tilmes, S. Bristol, T. Maycock, J.J. Bates, S. Hausman, O. Brown, and E. J. Kearns, 2018: A Conceptual Enterprise Framework for Managing Scientific Data Stewardship. Data Science Journal, 17. doi:10.5334/dsj-2018-015.

Peng, G., A. Milan, N. Ritchey, R. P. Partee II, S. Zinn, PE. McQuinn, Lemieux III, R. Ionin, D. Collins, P. Jones, A. Jakositz, and K.S. Casey, 2019: Practical Application of a Stewardship Maturity Matrix for the NOAA OneStop Program. Data Science Journal, 18. doi:10.5334/dsj-2019-041.

Ramapriyan, H K, Peng, G, Moroni, D, and Shie, C L, 2017: Ensuring and Improving Information Quality for Earth Science Data and Products. D.-Lib Magazine, 23, DOI:10.1045/july2017-ramapriyan.

Questions?

Contact Me:

gpeng@ncsu.edu

ORCID: http://orcid.org/0000-0002-1986-9115

Twitter: @DrPengAtAVL