Data Stewardship Maturity Matrix (DSMM) – Introduction and Application

Ge Peng, Ph.D.

Cooperative Institute for Climate and Satellites, North Carolina (CICS-NC)
NC State University and NOAA’s National Centers for Environmental Information (NCEI)

Library of Congress Annual Digital Preservation – DSA Meeting,
18 – 19 September 2017, Washington, DC, USA

NOAA Satellite and Information Service | National Centers for Environmental Information

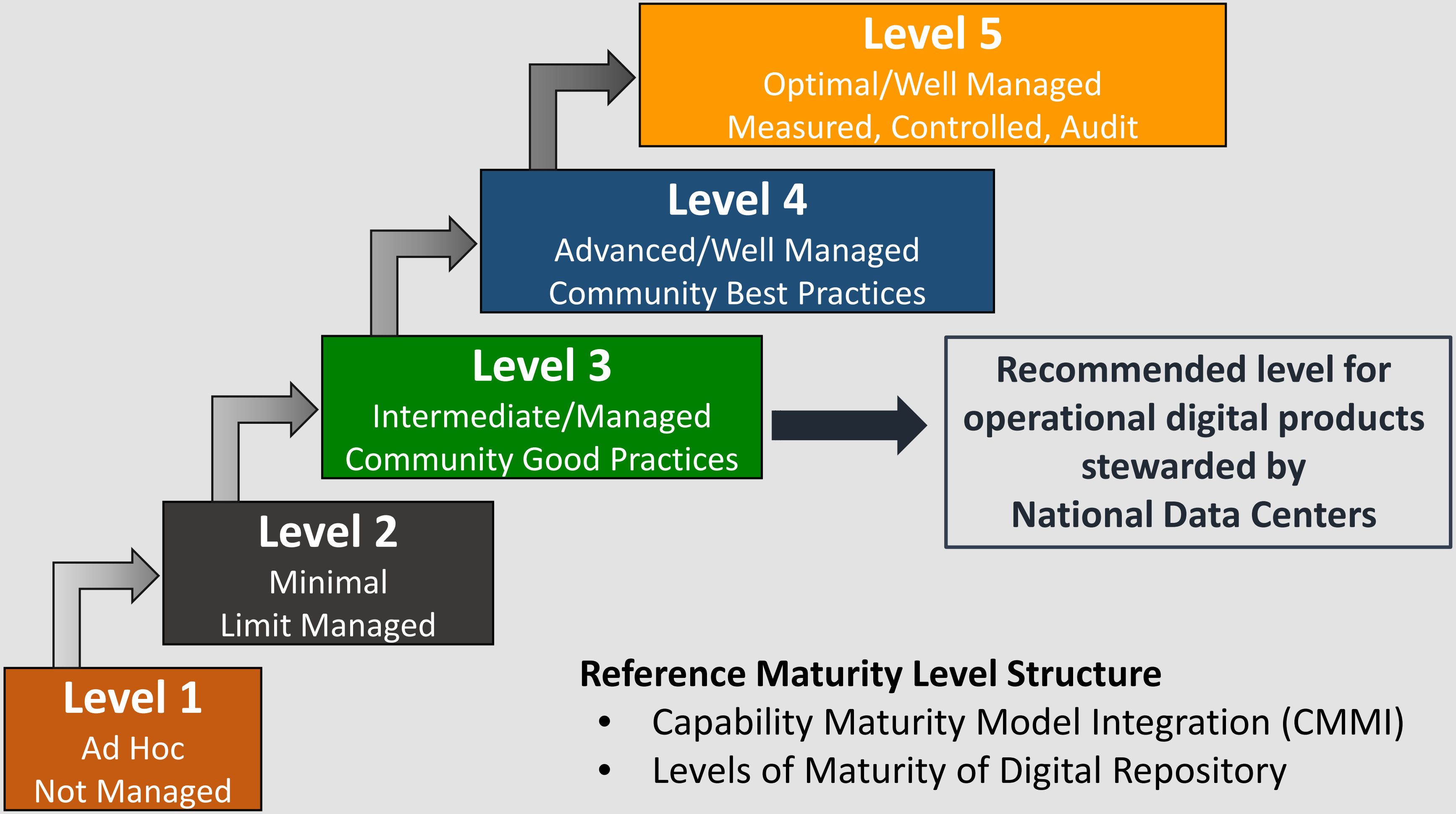
What Is the DSMM?

A Unified Framework for Measuring Stewardship Practices Applied to Individual Data Products

Developed by CICS-NC/NCEI & By Domain Subject Matter Experts, Leveraging

- **Institutional Knowledge**
- **Community Best Practices and Standards**

DSMM Follows CMMI Level Structure



Level 5
Optimal/Well Managed
Measured, Controlled, Audit

Level 4
Advanced/Well Managed
Community Best Practices

Level 3
Intermediate/Managed
Community Good Practices

Recommended level for operational digital products stewarded by National Data Centers

Level 2
Minimal
Limit Managed

Level 1
Ad Hoc
Not Managed

Reference Maturity Level Structure

- Capability Maturity Model Integration (CMMI)
- Levels of Maturity of Digital Repository

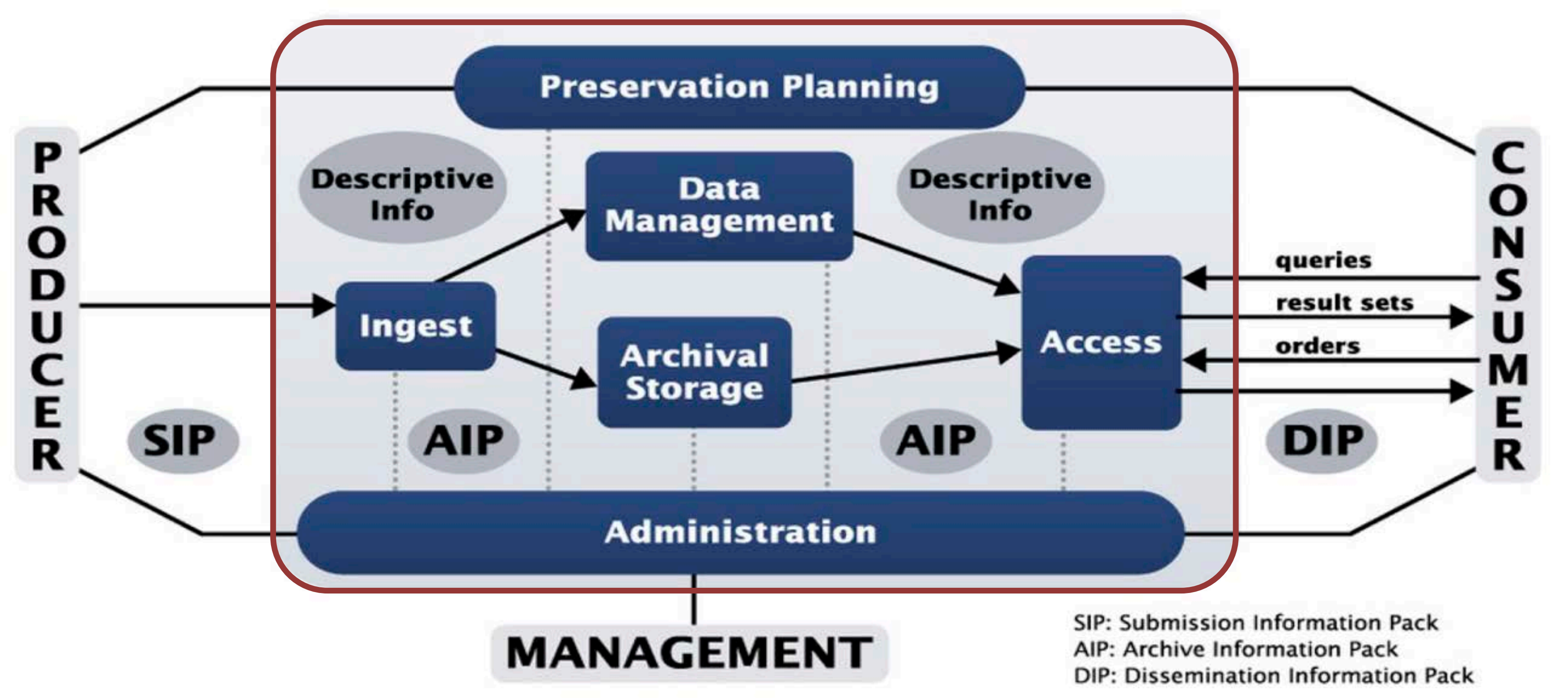
DSMM – Key Components

Practices in *Nine* Quasi-Independent Key Components

- *Preservability*
- *Accessibility*
- *Usability*
- *Production Sustainability*
- *Data Quality Assurance*
- *Data Quality Control/Monitoring*
- *Data Quality Assessment*
- *Transparency/Traceability*
- *Data Integrity*

Scope of DSMM

Functional Entities of the Open Archival Information System (OAIS)



SIP: Submission Information Pack
AIP: Archive Information Pack
DIP: Dissemination Information Pack

(CCSDS Version: 650x0m2-2012; Image Source: Lotar International)

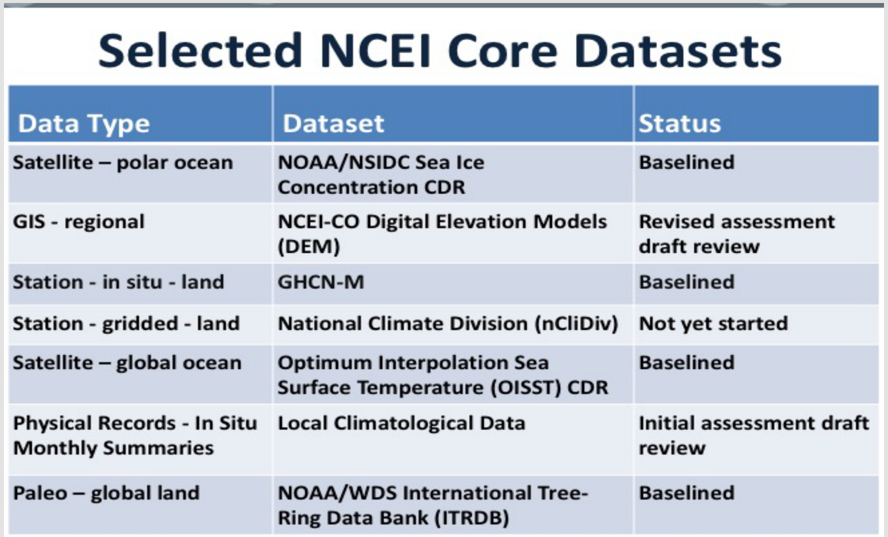
DSMM Vetting Process

- **Community Engagement: Feedback and Collaboration**
  - Internal (Domain SMEs from NOAA Data Centers: NCDC, NGDC, and NODC –> NCEI)
  - External (SMEs from ESIP Data Stewardship Committee; ESIP, AMS and AGU meetings)

DSMM Vetting Process

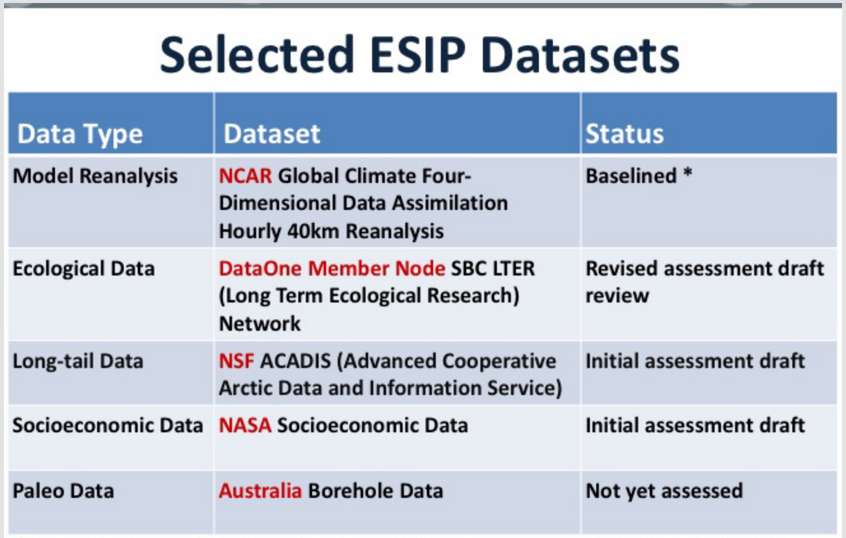
- **Community Engagement: Feedback and Collaboration**
  - Internal (Domain SMEs from NOAA Data Centers: NCDC, NGDC, and NODC –> NCEI)
  - External (SMEs from ESIP Data Stewardship Committee; ESIP, AMS and AGU meetings)

- **Use Case Studies**
  - NCEI Core Datasets – Different data types managed by same organization
  - ESIP DSC Datasets – Different disciplines managed by different organizations



Selected NCEI Core Datasets

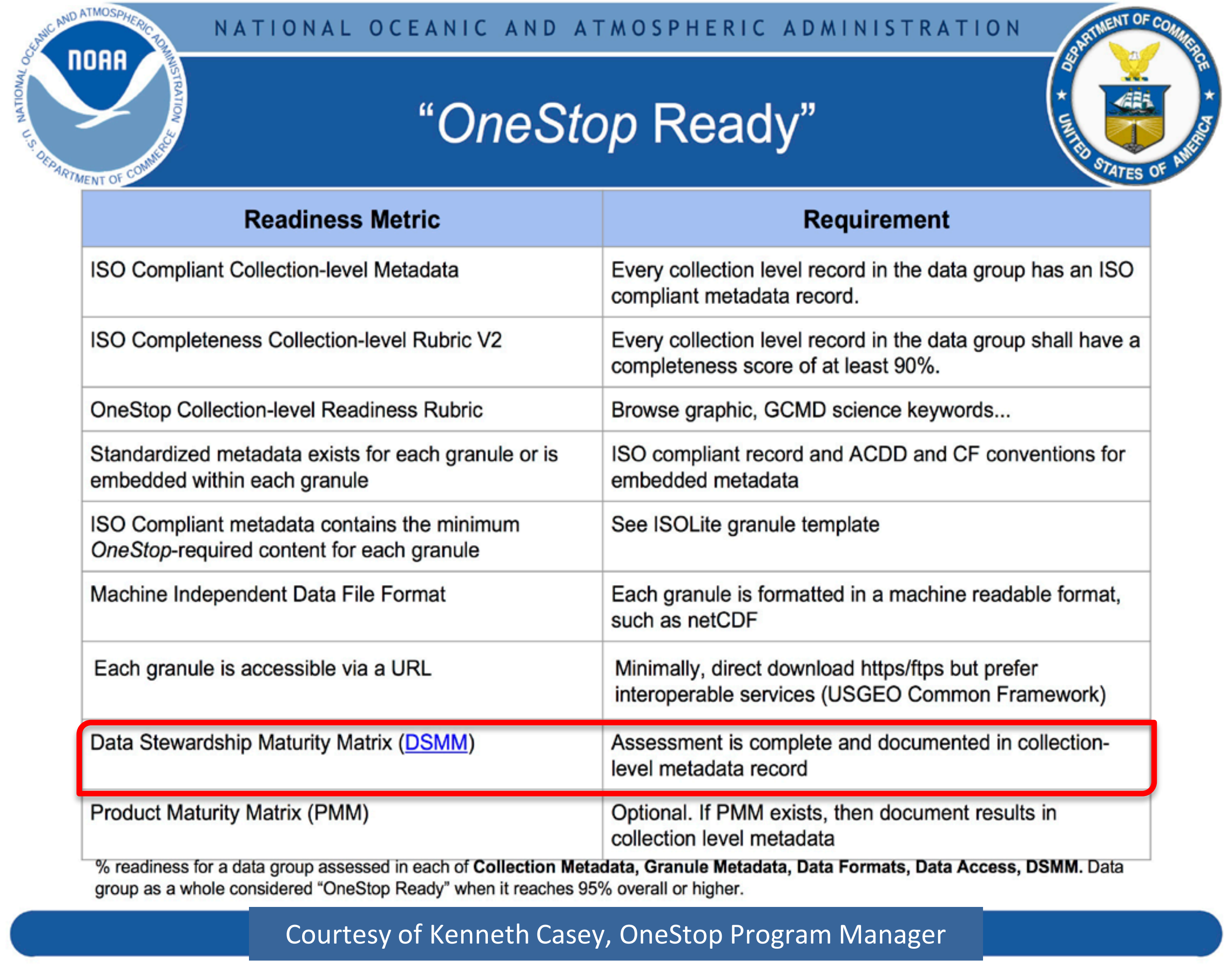
Data Type	Dataset	Status
Satellite – polar ocean	NOAA/NSIDC Sea Ice Concentration CDR	Baselined
GIS - regional	NCEI-CO Digital Elevation Models (DEM)	Revised assessment draft review
Station - in situ - land	GHCN-M	Baselined
Station - gridded - land	National Climate Division (nCliDiv)	Not yet started
Satellite – global ocean	Optimum Interpolation Sea Surface Temperature (OISST) CDR	Baselined
Physical Records - In Situ Monthly Summaries	Local Climatological Data	Initial assessment draft review
Paleo – global land	NOAA/WDS International Tree-Ring Data Bank (ITRDB)	Baselined



Selected ESIP Datasets

Data Type	Dataset	Status
Model Reanalysis	NCAR Global Climate Four-Dimensional Data Assimilation Hourly 40km Reanalysis	Baselined *
Ecological Data	DataOne Member Node SBC LTER (Long Term Ecological Research) Network	Revised assessment draft review
Long-tail Data	NSF ACADIS (Advanced Cooperative Arctic Data and Information Service)	Initial assessment draft
Socioeconomic Data	NASA Socioeconomic Data	Initial assessment draft
Paleo Data	Australia Borehole Data	Not yet assessed

DSMM Applications & Implementation



NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION

"OneStop Ready"

Readiness Metric	Requirement
ISO Compliant Collection-level Metadata	Every collection level record in the data group has an ISO compliant metadata record.
ISO Completeness Collection-level Rubric V2	Every collection level record in the data group shall have a completeness score of at least 90%.
OneStop Collection-level Readiness Rubric	Browse graphic, GCMD science keywords...
Standardized metadata exists for each granule or is embedded within each granule	ISO compliant record and ACDD and CF conventions for embedded metadata
ISO Compliant metadata contains the minimum *OneStop*-required content for each granule	See ISOLite granule template
Machine Independent Data File Format	Each granule is formatted in a machine readable format, such as netCDF
Each granule is accessible via a URL	Minimally, direct download https/ftps but prefer interoperable services (USGEO Common Framework)
Data Stewardship Maturity Matrix (DSMM)	Assessment is complete and documented in collection-level metadata record
Product Maturity Matrix (PMM)	Optional. If PMM exists, then document results in collection level metadata

% readiness for a data group assessed in each of **Collection Metadata, Granule Metadata, Data Formats, Data Access, DSMM.** Data group as a whole considered "OneStop Ready" when it reaches 95% overall or higher.

Courtesy of Kenneth Casey, OneStop Program Manager

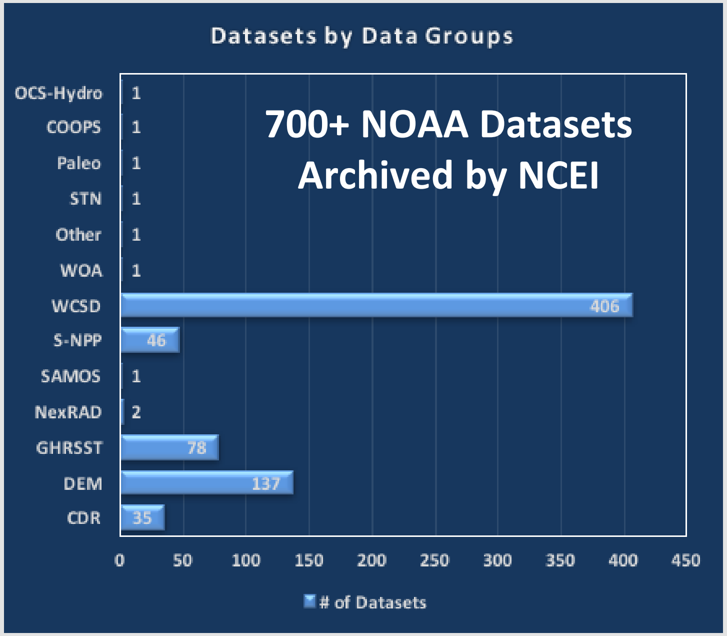
- **OneStop Ready**
- **OneStop DSMM Implementation**
  - ➢ Best practices,
  - ➢ Workflows,
  - ➢ Tools

DataOne User Group Meeting Poster:

tinyurl.com/DSMM-OneStop-Poster

DSMM Applications & Implementation

Data Quality Descriptive Information

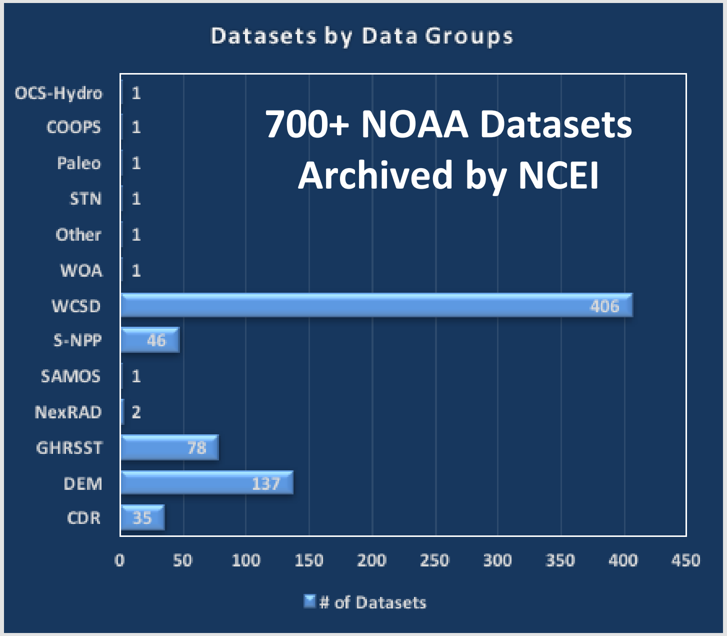


700+ NOAA Datasets Archived by NCEI

Evaluating data product stewardship maturity (as of 1/31/2017), mostly done by OneStop Metadata Team

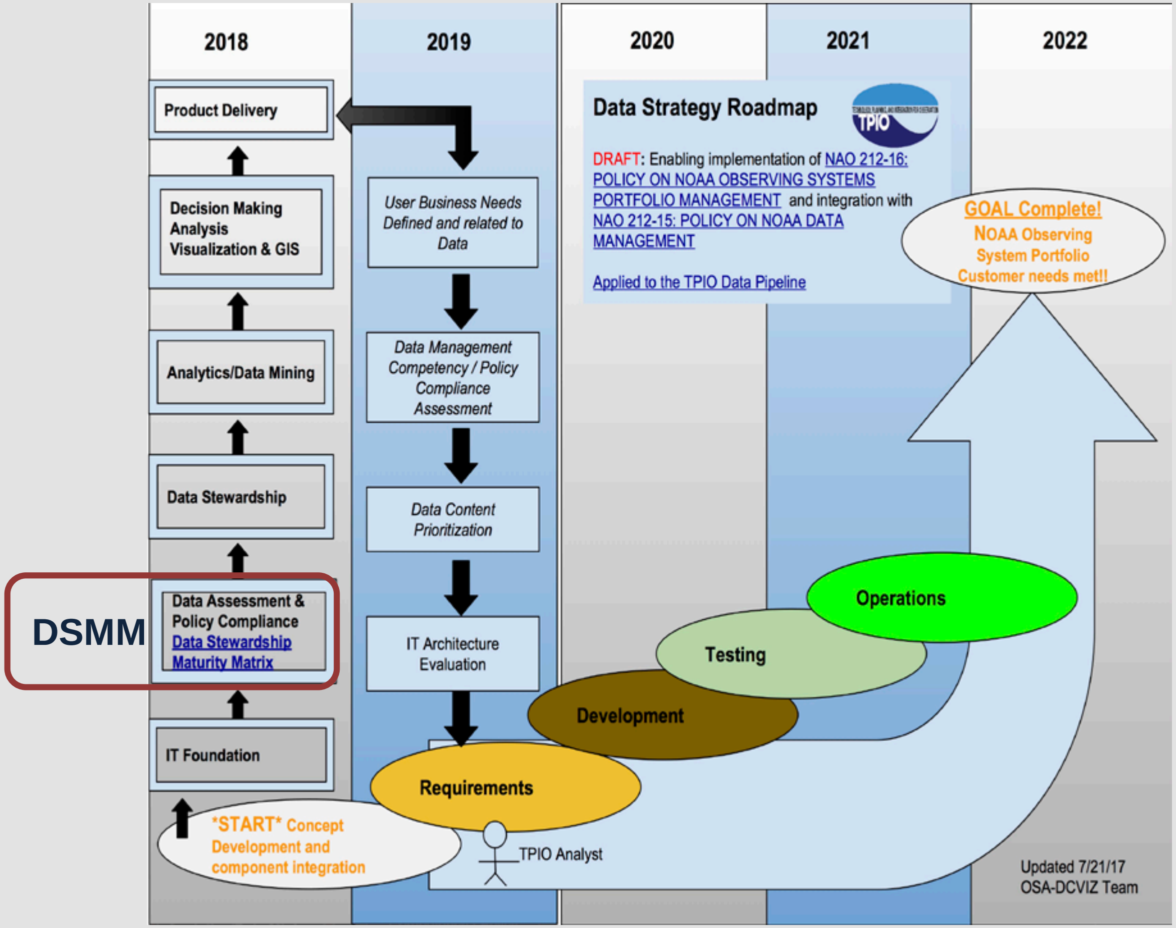
DSMM Applications & Implementation

Data Quality Descriptive Information



700+ NOAA Datasets Archived by NCEI

Evaluating data product stewardship maturity (as of 1/31/2017), mostly done by OneStop Metadata Team



Product Delivery

Decision Making Analysis Visualization & GIS

Analytics/Data Mining

Data Stewardship

Data Assessment & Policy Compliance Data Stewardship Maturity Matrix

IT Foundation

START Concept Development and component integration

User Business Needs Defined and related to Data

Data Management Competency / Policy Compliance Assessment

Data Content Prioritization

IT Architecture Evaluation

TPIO Analyst

Requirements

Development

Testing

Operations

Data Strategy Roadmap

DRAFT: Enabling implementation of NAO 212-16: POLICY ON NOAA OBSERVING SYSTEMS PORTFOLIO MANAGEMENT and integration with NAO 212-15: POLICY ON NOAA DATA MANAGEMENT

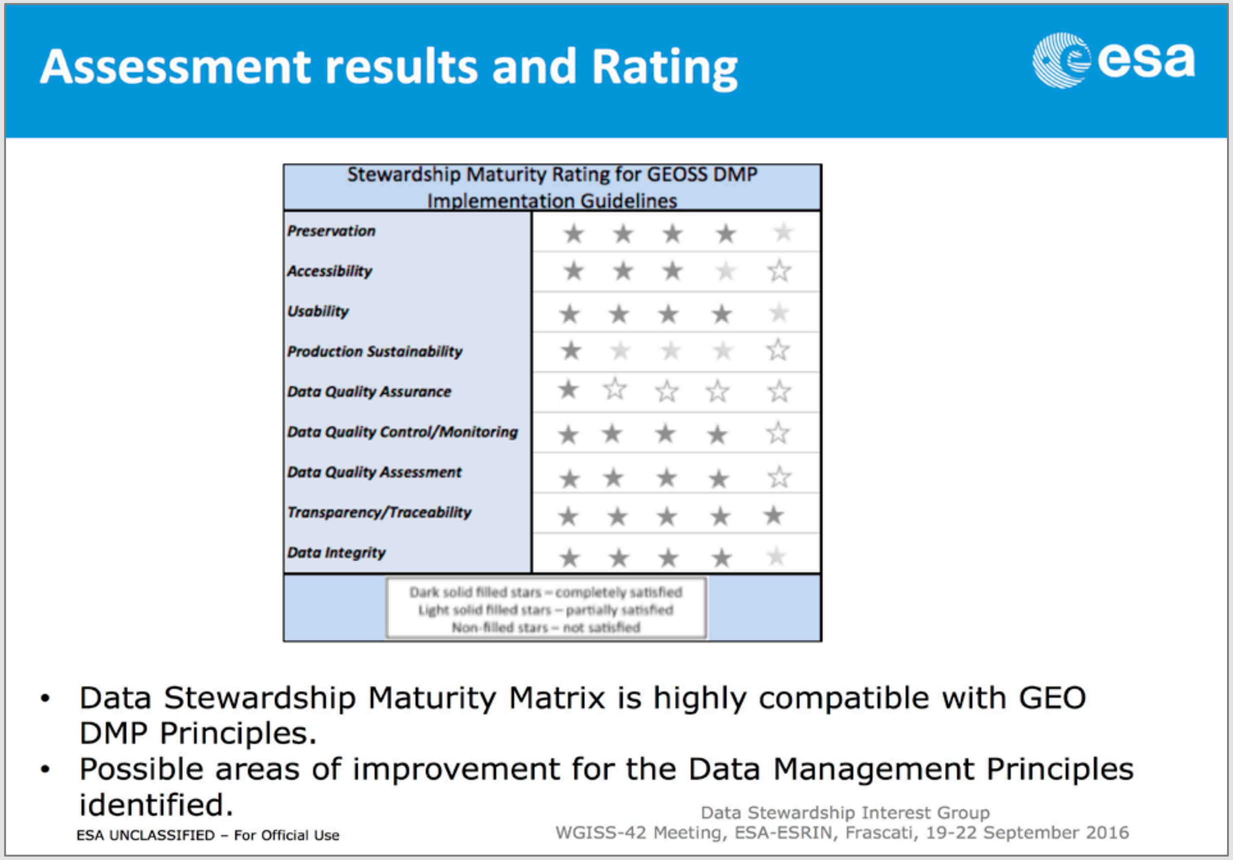
Applied to the TPIO Data Pipeline

GOAL Complete! NOAA Observing System Portfolio Customer needs met!!

Updated 7/21/17
OSA-DCVIZ Team

NOAA TPIO Data Strategy Roadmap (Courtesy of Matthew Austin, Team Lead)

DSMM Applications & Implementation



Assessment results and Rating

Stewardship Maturity Rating for GEOSS DMP Implementation Guidelines					
Preservation	★	★	★	★	✫
Accessibility	★	★	★	✫	☆
Usability	★	★	★	★	✫
Production Sustainability	★	✫	✫	✫	☆
Data Quality Assurance	★	☆	☆	☆	☆
Data Quality Control/Monitoring	★	★	★	★	☆
Data Quality Assessment	★	★	★	★	☆
Transparency/Traceability	★	★	★	★	★
Data Integrity	★	★	★	★	✫

Dark solid filled stars – completely satisfied
Light solid filled stars – partially satisfied
Non-filled stars – not satisfied

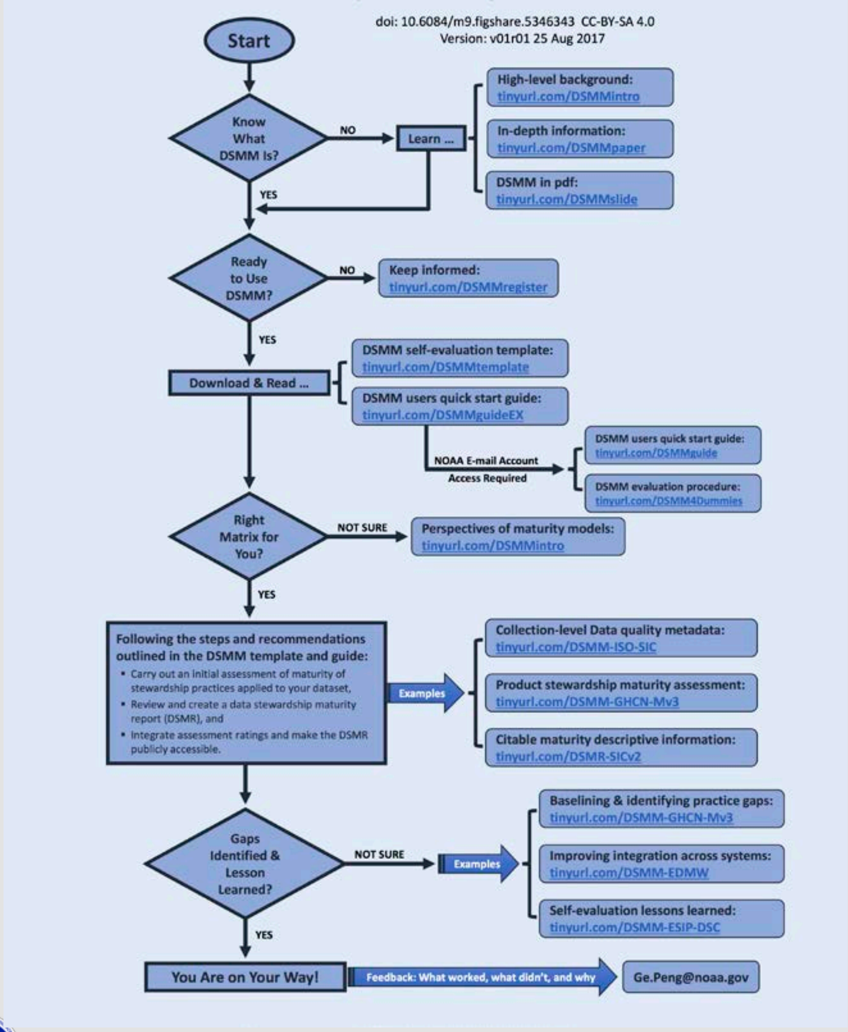
- Data Stewardship Maturity Matrix is highly compatible with GEO DMP Principles.
- Possible areas of improvement for the Data Management Principles identified.

ESA UNCLASSIFIED – For Official Use

Data Stewardship Interest Group
WGISS-42 Meeting, ESA-ESRIN, Frascati, 19-22 September 2016

Evaluating GEO Data Management Principles by European Space Agency (ESA) Data Stewardship Interest Group (Albani 2016)

Getting to Know & to Use DSMM



doi: 10.6084/m9.figshare.5346343 CC-BY-SA 4.0
Version: v01r01 25 Aug 2017

Start

Know What DSMM Is?

NO

Learn ...

High-level background:
tinyurl.com/DSMMintro

In-depth information:
tinyurl.com/DSMMpaper

DSMM in pdf:
tinyurl.com/DSMMslide

YES

Ready to Use DSMM?

NO

Keep informed:
tinyurl.com/DSMMregister

YES

Download & Read ...

DSMM self-evaluation template:
tinyurl.com/DSMMtemplate

DSMM users quick start guide:
tinyurl.com/DSMMguideEX

NOAA E-mail Account
Access Required

DSMM users quick start guide:
tinyurl.com/DSMMguide

DSMM evaluation procedure:
tinyurl.com/DSMM4Dummies

Right Matrix for You?

NOT SURE

Perspectives of maturity models:
tinyurl.com/DSMMintro

YES

Following the steps and recommendations outlined in the DSMM template and guide:

- Carry out an initial assessment of maturity of stewardship practices applied to your dataset,
- Review and create a data stewardship maturity report (DSMR), and
- Integrate assessment ratings and make the DSMR publicly accessible.

Examples

Collection-level Data quality metadata:
tinyurl.com/DSMM-ISO-SIC

Product stewardship maturity assessment:
tinyurl.com/DSMM-GHCN-Mv3

Citable maturity descriptive information:
tinyurl.com/DSMR-SICv2

Gaps Identified & Lesson Learned?

NOT SURE

Examples

Baselining & identifying practice gaps:
tinyurl.com/DSMM-GHCN-Mv3

Improving integration across systems:
tinyurl.com/DSMM-EDMW

Self-evaluation lessons learned:
tinyurl.com/DSMM-ESIP-DSC

YES

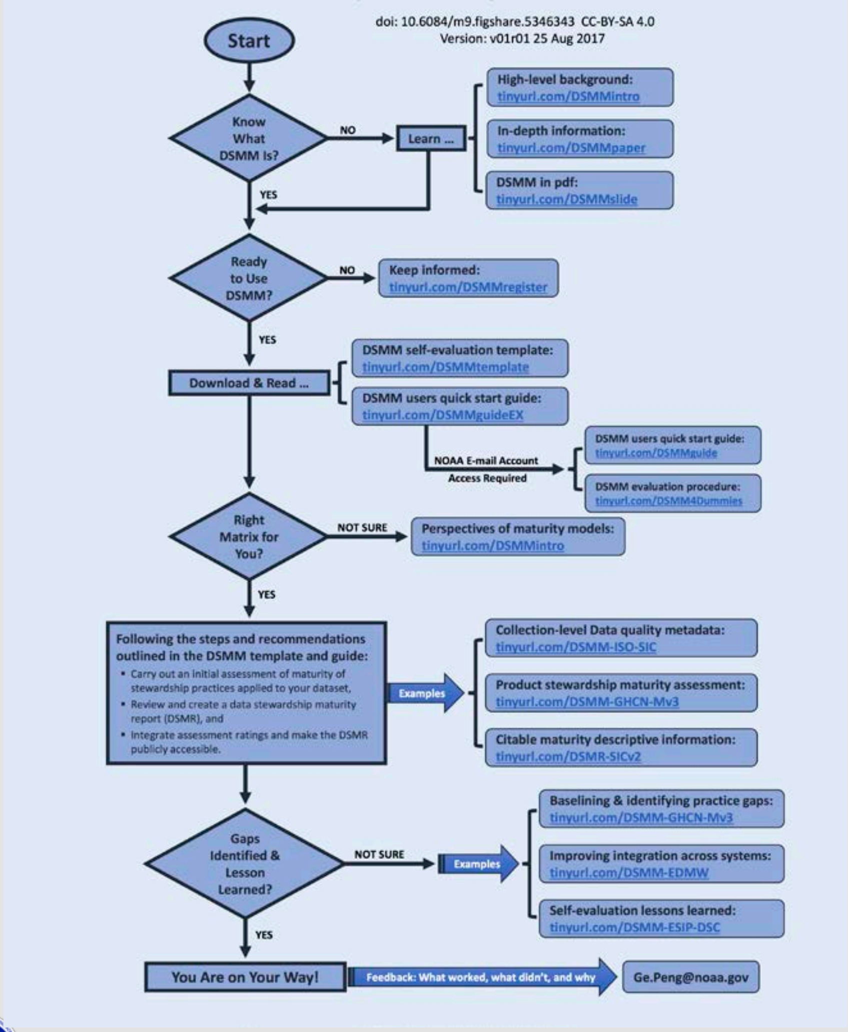
You Are on Your Way!

Feedback: What worked, what didn't, and why

Ge.Peng@noaa.gov

- **Published on figshare – A gradual way to get relevant information with clickable links**
- **Download: tinyurl.com/DSMM-FlowChart**

Getting to Know & to Use DSMM



doi: 10.6084/m9.figshare.5346343 CC-BY-SA 4.0
Version: v01r01 25 Aug 2017

Start

Know What DSMM Is?

NO

Learn ...

High-level background: tinyurl.com/DSMMintro

In-depth information: tinyurl.com/DSMMpaper

DSMM in pdf: tinyurl.com/DSMMslide

YES

Ready to Use DSMM?

NO

Keep informed: tinyurl.com/DSMMregister

YES

Download & Read ...

DSMM self-evaluation template: tinyurl.com/DSMMtemplate

DSMM users quick start guide: tinyurl.com/DSMMguideEX

NOAA E-mail Account Access Required

DSMM users quick start guide: tinyurl.com/DSMMguide

DSMM evaluation procedure: tinyurl.com/DSMM4Dummies

Right Matrix for You?

NOT SURE

Perspectives of maturity models: tinyurl.com/DSMMintro

YES

Following the steps and recommendations outlined in the DSMM template and guide:

- Carry out an initial assessment of maturity of stewardship practices applied to your dataset,
- Review and create a data stewardship maturity report (DSMR), and
- Integrate assessment ratings and make the DSMR publicly accessible.

Examples

Collection-level Data quality metadata: tinyurl.com/DSMM-ISO-SIC

Product stewardship maturity assessment: tinyurl.com/DSMM-GHCN-Mv3

Citable maturity descriptive information: tinyurl.com/DSMR-SICv2

Gaps Identified & Lesson Learned?

NOT SURE

Examples

Baselining & identifying practice gaps: tinyurl.com/DSMM-GHCN-Mv3

Improving integration across systems: tinyurl.com/DSMM-EDMW

Self-evaluation lessons learned: tinyurl.com/DSMM-ESIP-DSC

YES

You Are on Your Way!

Feedback: What worked, what didn't, and why

Ge.Peng@noaa.gov

- Published on figshare – A gradual way to get relevant information with clickable links
- Download: tinyurl.com/DSMM-FlowChart

Contact me

Ge.Peng@noaa.gov

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

THANK YOU

NCEI

NOAA NATIONAL CENTERS FOR ENVIRONMENTAL INFORMATION

Backup Slides

www.ncei.noaa.gov
www.climate.gov

NCEI Climate Facebook: http://www.facebook.com/NOAANCEIclimate
NCEI Ocean & Geophysics Facebook: http://www.facebook.com/NOAANCEIoceangeo
NCEI Climate Twitter (@NOAANCEIclimate): http://www.twitter.com/NOAANCEIclimate
NCEI Ocean & Geophysics Twitter (@NOAANCEIocngeo): http://www.twitter.com/NOAANCEIocngeo

Why Do We Need a DSMM?

A more formal approach to stewardship that supports rigorous compliance verification

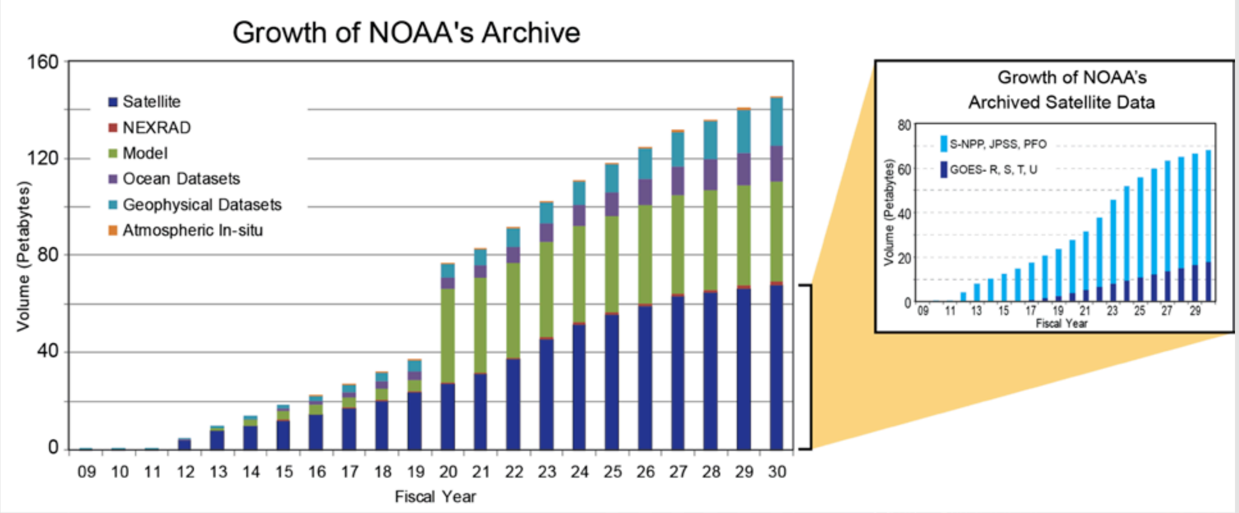
- *U.S. Information Quality Act (2001);*
- *Guidelines for Ensuring and Maximizing the Quality, Objectivity, Utility, and Integrity of Information (OMB 2002);*
- *Open Data and Data Sharing Policy (OMB, 2013; OSTP, 2013);*

Ensure the federally funded data are

- **preserved and secure**
- **available, discoverable, and accessible**
- **credible and understandable**
- **usable and useful**
- **sustainable and extendable**
- **citable, traceable, reproducible**

Why Do We Need a DSMM?

NOAA: 2000+ parameters
NCEI: 800+ collections

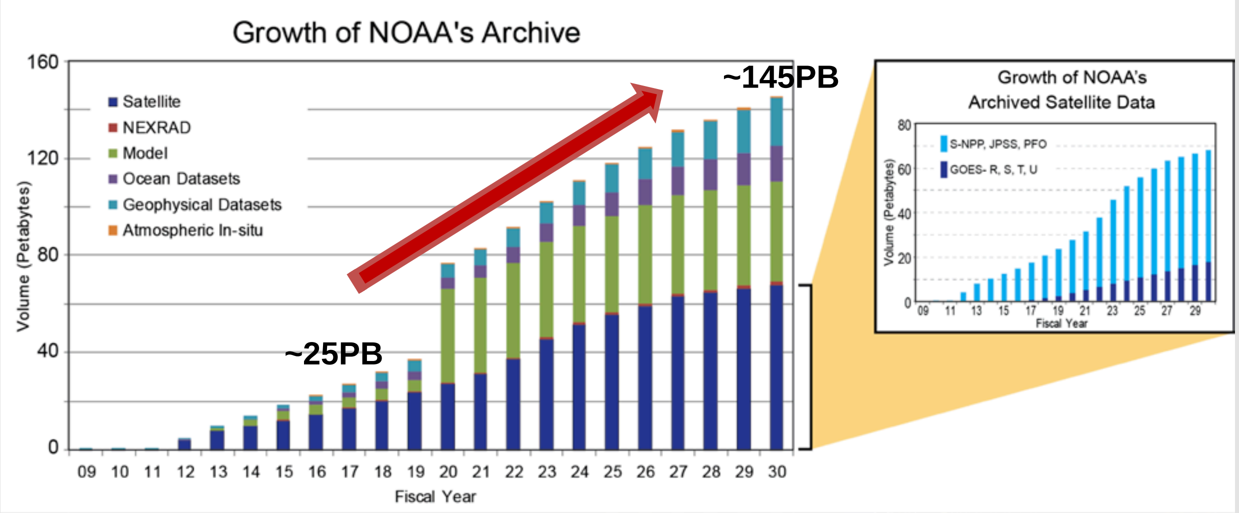


Ensure the federally funded data are

- ➢ preserved and secure
- ➢ available, discoverable, and accessible
- ➢ credible and understandable
- ➢ usable and useful
- ➢ sustainable and extendable
- ➢ citable, traceable, reproducible

Why Do We Need a DSMM?

NOAA: 2000+ parameters
NCEI: 800+ collections



Ensure the federally funded data are

- ➢ **preserved and secure**
- ➢ **available, discoverable, and accessible**
- ➢ **credible and understandable**
- ➢ **usable and useful**
- ➢ **sustainable and extendable**
- ➢ **citable, traceable, reproducible**

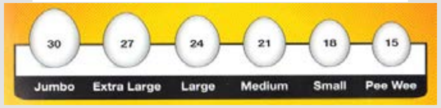
Data Stewardship

- ▪ Scalable
- ▪ Transparent
- ▪ Content-rich
- ▪ Interoperable
- ▪ Timely

Why Do We Need a Consistent Framework?

Statement: This is a good, big apple.

- What does “good” mean?
- What does “big” represent?



WE KNOW

USDA Prime is better quality than USDA Select!
(http://meat.tamu.edu/beefgrading/)

Extra Large is indeed larger than Large!

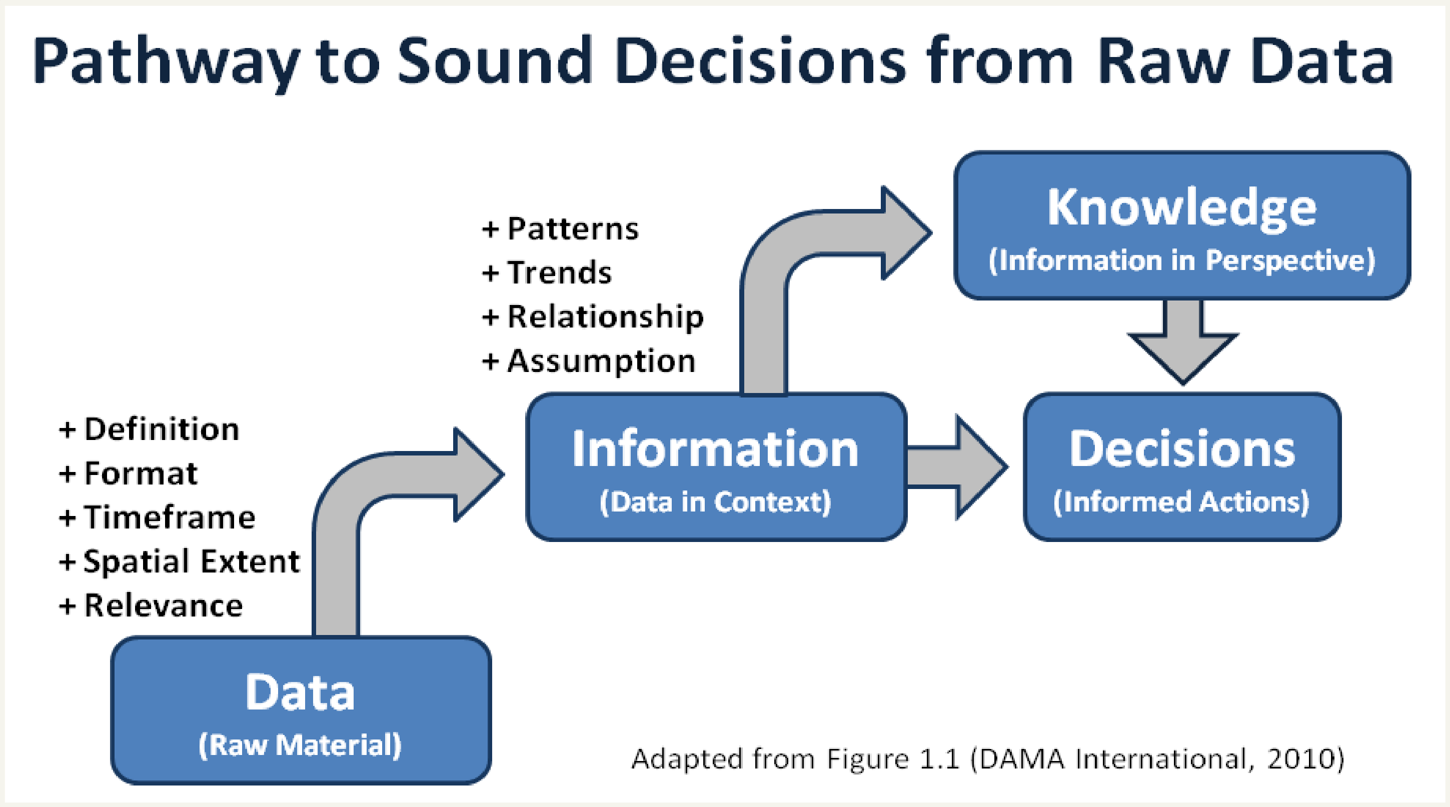
- ➢ **well-defined, implemented & audited**

The Same Goes for Individual Datasets!

DSMM Defines Measureable, Five-Level Progressive Practices *in* ***Nine*** *Quasi-Independent Key Components*

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 being preservable*				
Accessibility	*The state of 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 being transparent, trackable, and traceable*				
Data Integrity	*The state of data integrity being verifiable*				

Why Should We Care?



Pathway to Sound Decisions from Raw Data

Data
(Raw Material)

+ Definition
+ Format
+ Timeframe
+ Spatial Extent
+ Relevance

Information
(Data in Context)

+ Patterns
+ Trends
+ Relationship
+ Assumption

Knowledge
(Information in Perspective)

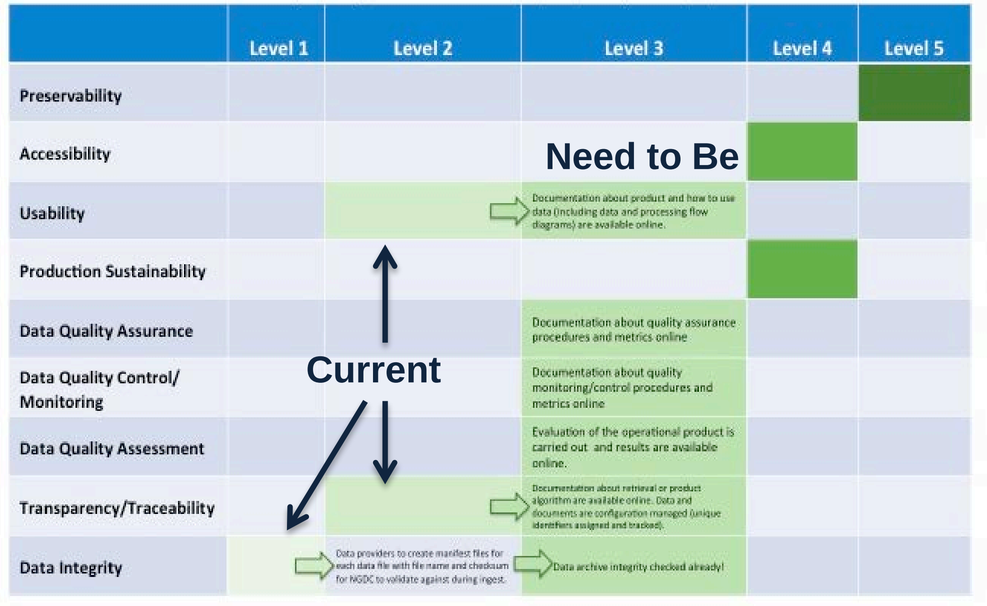
Decisions
(Informed Actions)

Adapted from Figure 1.1 (DAMA International, 2010)

Sound decisions reply on sound data and information!

Ways to Utilize DSMM & Results

- To know the current state of your dataset(s) – maturity scoreboard
- To know where you want or need to be – stewardship requirements
- To know how to get there – roadmap forward (informed, actionable steps)



	Level 1	Level 2	Level 3	Level 4	Level 5
Preservability					🟩
Accessibility				🟩	
Usability		🟩	Documentation about product and how to use data (including data and processing flow diagrams) are available online.		
Production Sustainability				🟩	
Data Quality Assurance			Documentation about quality assurance procedures and metrics online		
Data Quality Control/ Monitoring			Documentation about quality monitoring/control procedures and metrics online		
Data Quality Assessment			Evaluation of the operational product is carried out and results are available online.		
Transparency/Traceability		🟩	Documentation about retrieval or product algorithm are available online. Data and documents are configuration managed (unique identifiers assigned and tracked).		
Data Integrity	🟩	Data providers to create manifest files for each data file with file name and checksum for NGDC to validate against during ingest.	Data archive integrity checked already!		

Stewardship Maturity Scoreboard and Roadmap Forward

- A reference model for stewardship planning and resource allocation – informed decision-making support
- A consolidate source and transparency for information about stewardship practices – assessment with detailed justifications
- Content-rich quality metadata – enhanced discoverability and usability

NCEI/CICS-NC Data Stewardship Maturity Matrix

Document ID: NCDC-CICS-SMM_0001
Version: Rev. 1. 12/09/2014

Dataset Name

Maturity Level as of mm/dd/yyyy

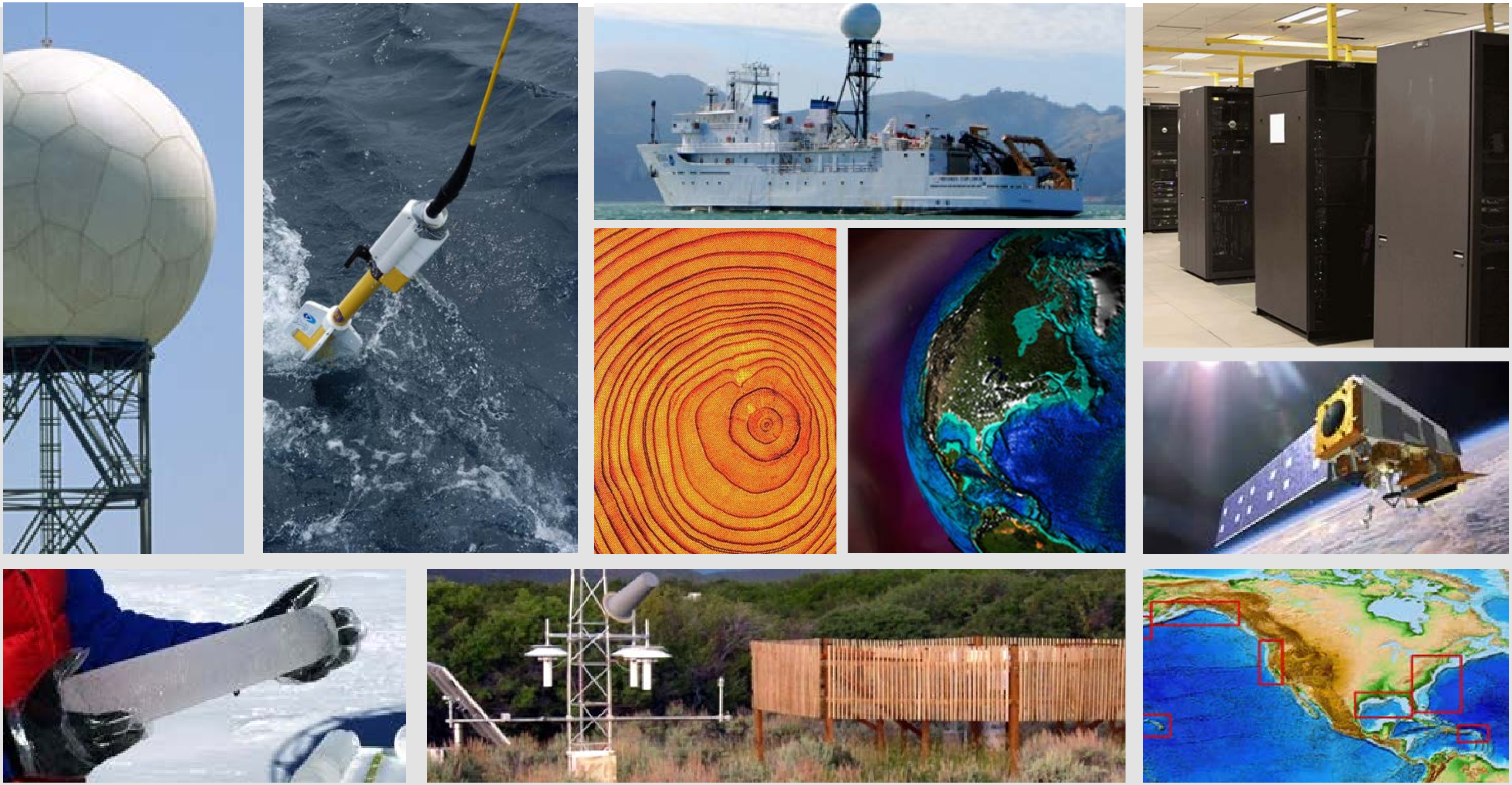
Stewardship Maturity Matrix for Digital Environmental Data Products

Maturity Scale	Preservability	Accessibility	Usability	Production Sustainability	Data Quality Assurance	Data Quality Control/Monitoring	Data Quality Assessment	Transparency /Traceability	Data Integrity
Level 1 – Ad Hoc **Not Managed**	Any storage location Data only	Not publicly available Person-to-person	Extensive product-specific knowledge required No documentation online	Ad Hoc or Not applicable No obligation or deliverable requirement	Data quality assurance (DQA) procedure unknown or none	None or Sampling unknown or spotty Analysis unknown or random in time	Algorithm/method/model theoretical basis assessed (method and results online)	Limited product information available Person-to-person	Unknown or no data ingest integrity check
Level 2 - Minimal **Managed Limited**	Non-designated repository Redundancy Limited archiving metadata	Publicly available Direct file download (e.g., via anonymous FTP server) Collection/dataset level searchable	Non-standard data format Limited documentation (e.g., user's guide) online	Short-term Individual PI's commitment (grant obligations)	Ad Hoc and random DQA procedure not defined and documented	Sampling and analysis are regular in time and space Limited product-specific metrics defined & implemented	Level 1 + Research product assessed (method and results online)	Product information available in literature	Data ingest integrity verifiable (e.g., checksum technology)
Level 3 - Intermediate **Managed Defined, Partially Implemented**	Designated archive Redundancy Community-standard archiving metadata Conforming to limited archiving process standards	Level 2 + Non-standard data service Limited data server performance Granule/file level searchable Limited search metrics	Community Standard-based interoperable format & metadata Documentation (e.g., source code, product algorithm document, processing or/and data flow diagram) online	Medium-term Institutional commitment (contractual deliverables with specs and schedule defined)	DQA procedure defined and documented and partially implemented	Level 2 + Sampling and analysis are frequent and systematic but not automatic Community metrics defined and partially implemented Procedure documented and available online	Level 2 + Operational product assessed (method and results online)	Algorithm/method/model Theoretical Basis Document (ATBD) & source code online Dataset configuration managed (CM) Unique Object Identifier (OID) assigned (dataset, documentation, source code) Data citation tracked (e.g., utilizing Digital Object Identifier (DOI) system)	Level 2 + Data archive integrity verifiable
Level 4 - Advanced **Managed Well-Defined, Fully Implemented**	Level 3 + Conforming to community archiving standards	Level 3 + Community-standard data services Enhanced data server performance Conforming to community search metrics Dissemination report metrics defined and implemented internally	Level 3 + Basic capability (e.g., subsetting, aggregating) & data characterization (overall/global, e.g., climatology, error estimates) available online	Long-term Institutional commitment Product improvement process in place	DQA procedure well documented, fully implemented and available online with master reference data Limited data quality assurance metadata	Level 3 + Anomaly detection procedure well-documented and fully implemented using community metrics, automatic, tracked and reported Limited quality monitoring metadata	Level 3 + Quality metadata assessed (method and results online) Limited quality assessment metadata	Level 3 + Operational Algorithm Description (OAD) online, OID assigned, and under CM	Level 3 + Data access integrity verifiable Conforming to community data integrity technology standard
Level 5 - Optimal **Level 4 + Measured , Controlled , Audit**	Level 4 + Archiving process performance controlled, measured, and audited Future archiving standard changes planned	Level 4 + Dissemination reports available online Future technology and standard changes planned	Level 4 + Enhanced online capability (e.g., visualization, multiple data formats) Community metrics of data characterization (regional/cell) online External ranking	Level 4 + National or international commitment Changes for technology planned	Level 4 + DQA procedure monitored and reported Conforming to community quality metadata & standards External review	Level 4 + Cross-validation of temporal & spatial characteristics Physical consistency check Conforming to community quality metadata & standards Dynamic providers/users feedback in place	Level 4 + Assessment performed on a recurring basis Conforming to community quality metadata & standards External ranking	Level 4 + System information online Complete data provenance available online	Level 4 + Data authenticity verifiable (e.g., data signature technology) Performance of data integrity check monitored and reported

Dataset Information: URL Goes Here
Dataset POC: Name & E-mail Here

SMM POC: Ge.Peng@noaa.gov

NCEI Ingests and Archives Environmental Data from U.S. and International Sources



Data spans stone-age to space-age … from the depths of the ocean to the sun … and across the globe

NCEI Products Span From Local to Global and Weekly to Decadal Scales

	Daily/Weekly	Monthly	Seasonal – Annual	Annual to Decadal
Local	Snowfall Impact Index – FEMA, disaster response	Heating & Cooling Degree Days – Energy Sector	Temperature & Precipitation Outlooks – Agriculture	Coastal Digital Elevation Models (DEM) – Hazard Mitigation
Regional	Hurricane Tracks – Emergency Planners	Solar Activity/Sun Spots – Power Distribution	Billion $ Disasters, Climate Extremes Index – Insurance	Climate Normals – Construction, Infrastructure, Agriculture
National & Global	Tsunami Warning – Emergency Managers	Global and U.S. Climate Summaries – Numerous Sectors	Annual State of the Climate Reports – Scientists	IPCC & National Climate Assessments – Gov't Policymakers

NCEI Data & NOAA BigData Initiative

- NOAA has a lot of data – often under-utilized
- Five major data alliances
- Weather/Climate/Model data and products
- 27 October 2015 - Amazon Web Service provides full access, for the first time, to the entire Level II data from the NOAA’s Next Generation Weather Radar (NEXRAD) network – over 300 terabytes – growing at about 50 terabytes per year
- NOAA GOES-16 Provisional data – Amazon Web Services & Open Cloud Consortium.

Google Cloud Platform

Microsoft

amazon web services

OCC OPEN COMMONS CONSORTIUM

IBM