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
What Is the DSMM?


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

• Institutional Knowledge
• Community Best Practices and Standards
DSMM Follows CMMI Level Structure

Level 1
Ad Hoc
Not Managed

Level 2
Minimal
Limit Managed

Level 3
Intermediate/Managed
Community Good Practices

Level 4
Advanced/Well Managed
Community Best Practices

Level 5
Optimal/Well Managed
Measured, Controlled, Audit

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

Recommended level for operational digital products stewarded by National Data Centers
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)
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                                    |
| Model Reanalysis                           | NCAR Global Climate Four-Dimensional Data Assimilation Hourly 40km Reanalysis |
| Ecological Data                            | DataOne Member Node SBC LTER (Long Term Ecological Research) Network |
| Long-tail Data                             | NSF ACADIS (Advanced Cooperative Arctic Data and Information Service) |
| Socioeconomic Data                         | NASA Socioeconomic Data                    |
| Paleo Data                                 | Australia Borehole Data                    |
## DSMM Applications & Implementation

### OneStop Ready

<table>
<thead>
<tr>
<th>Readiness Metric</th>
<th>Requirement</th>
</tr>
</thead>
<tbody>
<tr>
<td>ISO Compliant Collection-level Metadata</td>
<td>Every collection level record in the data group has an ISO compliant metadata record.</td>
</tr>
<tr>
<td>ISO Completeness Collection-level Rubric V2</td>
<td>Every collection level record in the data group shall have a completeness score of at least 90%.</td>
</tr>
<tr>
<td>OneStop Collection-level Readiness Rubric</td>
<td>Browse graphic, GCMD science keywords...</td>
</tr>
<tr>
<td>Standardized metadata exists for each granule or is embedded within each granule</td>
<td>ISO compliant record and ACDD and CF conventions for embedded metadata</td>
</tr>
<tr>
<td>ISO Compliant metadata contains the minimum OneStop-required content for each granule</td>
<td>See ISOLite granule template</td>
</tr>
<tr>
<td>Machine Independent Data File Format</td>
<td>Each granule is formatted in a machine readable format, such as netCDF</td>
</tr>
<tr>
<td>Each granule is accessible via a URL</td>
<td>Minimally, direct download https/ftp but prefer interoperable services (USGEO Common Framework)</td>
</tr>
</tbody>
</table>

**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

---

*Courtesy of Kenneth Casey, OneStop Program Manager*

### OneStop DSMM Implementation

- Best practices,
- Workflows,
- Tools

DataOne User Group Meeting Poster: [tinyurl.com/DSMM-OneStop-Poster](http://tinyurl.com/DSMM-OneStop-Poster)
DSMM Applications & Implementation

Data Quality Descriptive Information

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

Data Quality Descriptive Information

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

700+ NOAA Datasets Archived by NCEI

NOAA TPIO Data Strategy Roadmap (Courtesy of Matthew Austin, Team Lead)
DSMM Applications & Implementation

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

Evaluating GEO Data Management Principles by European Space Agency (ESA) Data Stewardship Interest Group (Albani 2016)
Getting to Know & to Use DSMM

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

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

Backup Slides

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 (@NOAANCEIoceangeo): http://www.twitter.com/NOAANCEIoceangeo
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

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

Sound decisions reply on sound data and information!

Adapted from Figure 1.1 (DAMA International, 2010)
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)
- 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

## Dataset Name

**Stewardship Maturity Matrix for Digital Environmental Data Products**

<table>
<thead>
<tr>
<th>Maturity Scale</th>
<th>Preservability</th>
<th>Accessibility</th>
<th>Usability</th>
<th>Production Sustainability</th>
<th>Data Quality Assurance</th>
<th>Data Quality Control/Monitoring</th>
<th>Data Quality Assessment</th>
<th>Transparency/Traceability</th>
<th>Data Integrity</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Level 1 – Ad Hoc Not Managed</strong></td>
<td>Any storage location (Data only)</td>
<td>Not publicly available (person-to-person)</td>
<td>Extensive product-specific knowledge required</td>
<td>Ad Hoc or not applicable (no obligation or deliverable requirement)</td>
<td>Data quality assurance (DQA) procedure unknown or none</td>
<td>None or sampling unknown or spotty</td>
<td>Algorithm/method/model theoretical basis assessed (method and results online)</td>
<td>Limited product information available (person-to-person)</td>
<td>Unknown or no data ingest integrity check</td>
</tr>
<tr>
<td><strong>Level 2 - Minimal Managed Limited</strong></td>
<td>Non-designated repository (Redundancy Limited), archiving metadata (Non-standard data format, limited documentation)</td>
<td>Publicly available (direct file download via anonymous FTP server, collection/dataset level searchable)</td>
<td>Non-standard data format (limited documentation, e.g., user's guide online)</td>
<td>Short-term individual PI's commitment (grant obligations)</td>
<td>Ad Hoc and random DQA procedure not defined and documented</td>
<td>Sampling and analysis are regular in time and space (limited product-specific metrics defined &amp; implemented)</td>
<td>Level 1 + Research product assessed (method and results online)</td>
<td>Product information available in literature</td>
<td>Data ingest integrity verifiable (e.g., checksum technology)</td>
</tr>
<tr>
<td><strong>Level 3 - Intermediate Managed Defined, Partially Implemented</strong></td>
<td>Designated archive (Redundancy Community-standard archiving metadata Conforming to archiving process standards)</td>
<td>Level 2 + Non-standard data service (Limited data server performance, granule/file level searchable, limited search metrics)</td>
<td>Community standard-based interoperable format &amp; metadata documentation, source code, product algorithm, document processing and/or data flow diagram online</td>
<td>Medium-term institutional commitment (contractual, deliverables with schedule defined)</td>
<td>DQA procedure defined and documented and partially implemented</td>
<td>Level 2 + Sampling and analysis are frequent and systematic but not automatic (community metrics defined and partially implemented; Procedure documented and partially implemented)</td>
<td>Level 2 + Operational product assessed (method and results online)</td>
<td>Algorithm/method/model theoretical basis &amp; source code online</td>
<td>Data access integrity verifiable</td>
</tr>
<tr>
<td><strong>Level 4 - Advanced Managed Well-Defined, Fully Implemented</strong></td>
<td>Level 2 + Community-standard data services (Enhanced data server performance, conforming to community search metrics)</td>
<td>Level 3 + Basic capability (e.g., subsetting, aggregating) &amp; data characterization (global, regional, e.g., climatology, error estimates) available online</td>
<td>Long-term institutional commitment (Product improvement process in place)</td>
<td>DQA procedure well-documented, fully implemented and available online with master reference data</td>
<td>Level 3 + Anomaly detection procedure well-documented and fully implemented using community metrics, automatic, tracked and reported</td>
<td>Level 3 + Quality metadata assessed (method and results online)</td>
<td>Level 3 + Operational Algorithm Description (OAD) online, OID assigned, and under CM</td>
<td>Level 3 + Data access integrity verifiable</td>
<td>Conforming to community data integrity technology standard</td>
</tr>
<tr>
<td><strong>Level 5 - Optimal</strong></td>
<td>Level 4 + Archiving process performance controlled, measured, and audited</td>
<td>Level 4 + Dissemination reports available online</td>
<td>Level 4 + Enhanced online visibility: multiple data formats</td>
<td>Level 4 + National or international coordination (changes for technology planned)</td>
<td>Level 4 + DQA procedure monitored and reported</td>
<td>Level 4 + Assessment performed on a recurring basis</td>
<td>Level 4 + System information online (Complete data provenance available online)</td>
<td>Level 4 + Data authenticity verified (e.g., data signature technology)</td>
<td>Performance of data integrity check monitored and reported</td>
</tr>
</tbody>
</table>

**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**
- Snowfall Impact Index
  - FEMA, disaster response

**Monthly**
- Heating & Cooling Degree Days
  - Energy Sector
- Solar Activity/Sun Spots
  - Power Distribution

**Seasonal – Annual**
- Temperature & Precipitation Outlooks
  - Agriculture
- Billion $ Disasters, Climate Extremes Index
  - Insurance
- Global and U.S. Climate Summaries
  - Numerous Sectors
- State of the Climate in 2013
  - Annual State of the Climate Reports
  - Scientists

**Annual to Decadal**
- Coastal Digital Elevation Models (DEM)
  - Hazard Mitigation
- Climate Normals
  - Construction, Infrastructure, Agriculture
- 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.