Stewardship Maturity Matrix
- A Unified Framework for Assessing Data Quality and Usability Practices
Applied to Individual Digital Environmental Data Products

Ge Peng
NOAA’s Cooperative Institute for Climate and Satellite – NC State University (CICS-NC)
and NOAA’s National Climatic Data Center (NCDC)

Jeff Privette
NCDC/NOAA

In Collaboration with
Ed Kearns and Nancy Ritchey
NCDC/NOAA

10 July 2014
ESIP Summer Meeting
Why Do We Need a Stewardship Maturity Matrix?

- The value and quality of a data set depends – in part – on the stewardship practices applied after its production.

Hypothetic questions to a data center:

1. Congress: Are your datasets compliant with the U.S. Data Quality Act? If not, then what?
3. Modelers: Is the quality of a routinely updated product being assessed?

Problem: Most of data centers currently cannot readily convey - or even assess – the level of stewardship practices for its stakeholders or customers. No community scorecard exists.

- This is a vulnerability – and an opportunity!

Solution: Define a Stewardship Maturity Matrix to assess stewardship practices applied to individual data products
What Is Scientific Data Stewardship?

- Activities to preserve or improve the information content, accessibility, and usability of environmental data and metadata (NRC, 2007)

Activities to ensure or improve the **quality and usability** of environmental data and metadata

**Ensuring Data Are Always Meaningful**

**Scientific Data Stewards**

- Data Quality Screening/Assurance/Control/Evaluation/Assessment/Monitoring
  - Common data format
  - Spatial/temporal characteristics
  - Uncertainty estimates

*Environmental data* are recorded observations and measurements of the physical, chemical, biological, geological, or geophysical properties or conditions of the oceans, atmosphere, space environment, sun, and solid earth, as well as correlative data and related documentation or metadata (NOAA, 2008).
How Do We Identify Key Components and Define Levels of Stewardship Maturity Matrix

- Policies
  - U.S. Laws
  - Agencies’ recommendation
  - Experts’ recommendation

- Processes
  - Data Archive
  - Data/Metadata Management
  - Data Quality Management

- Procedures Standards
  - Data Preservation
  - Data Governance
  - Data Provenance

- Tasks
  - Evaluate product
  - Verify data integrity
  - Monitor data quality

- Non-Functional Requirements

- Functional Key Areas

- Key Matrix Components

- Community Practices
  - Relevant
  - Measurable
  - Progressive
  - Independent (?)
Maturity Levels 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
- Climate Data Record Maturity Matrix (CDRMM)
Goals?

- General
- Simple
- Concise

Assess & Convey & Path Forward

Not to Reinvent Wheels

Leveraging

- Community accepted good and best practices and standards
- Subject Matter Experts

The ESIP Community
Key Components of Stewardship Maturity Matrix

Scientific Data Stewardship

- Data Quality Screen/Assurance/Monitor/Control
- Data Usability
- Data Integrity
- Production Sustainability
- Transparency/Traceability
- Discovery and Access
- Preservation
- Best Practice
- Best Practice

Data Quality Assessment/Validation
Who Could Use the Matrix?

- **Data providers and scientific stewards**
  - to evaluate and improve the quality and usability of their products against community best practices

- **Modelers, decision-support system users, and scientists**
  - to improve their products and uncertainty estimates
  - to make investment and use decision

- **Data managers/stewards of data centers and repositories**
  - to validate their compliance or lack of to community accepted stewardship practice or standards
  - to assess the current state
  - to create a roadmap forward to improve or enhance its stewardship maturity of practices applied to a certain product or all its holdings

- **General data users**
  - to make an educated choice on selecting or utilizing a dataset
Your Comment & Suggestion Are Appreciated!

• Should community best practices in each key component be given intermediate (level 3) or advanced (level 4) maturity level, i.e., benchmark or upper-bound?

What is a Best Practice?

- **Wikipedia**: A best practice is a method or technique that has consistently shown results superior to those achieved with other means, and that is used as a **benchmark**.
- **Cambridge Dictionary**: a working method or set of working methods that is officially accepted as being **the best** to use in a particular business or industry, usually described formally and in detail.

*Dilbert is a Best Practice*
Your Comment & Suggestion Are Appreciated!

• Should community best practices in each key component be given intermediate (level 3) or advanced (level 4) maturity level, i.e., benchmark or upper-bound?

• Need for multiple data quality key components?
  
  **DQ Assurance/Screen** procedures are to ensure that the products meet the requirements – building it right.

  **DQ Assessment/Validation** is to ensure that the products are scientifically sound – building the right thing.

  **DQ Monitor/Control** procedures are to detect anomalous in the data input and/or outputs
Your Comment & Suggestion Are Appreciated!

• Should community best practices in each key component be given intermediate (level 3) or advanced (level 4) maturity level, i.e., benchmark or upper-bound?

• Need for multiple data quality key components?

• How can we Integrate our effort with the ESIP community?

Ge.Peng@noaa.gov

http://www.slideshare.net/gepeng86/sds-mm-20140527betav10
http://www.slideshare.net/gepeng86/long-term-scientificdatastewardship20140226
Talking Points for Discussion

The Data Quality Management Cycle

Data Quality Key Function Areas

- Plan: Requirements/Specifications
- Do: Measure/Assess/Describe
- Check: Screen/Assurance/Monitor/Control
- Act: Resolve/Report/Notify

- Roles and responsibilities?
- Consistent guidelines of good & best practices in each function area?
- An efficient way to report, resolve, and notify timely and effectively?
Backup Slides
# Dataset Name

## Stewardship Maturity Matrix for Digital Environmental Data Products

<table>
<thead>
<tr>
<th>Maturity Scale</th>
<th>Preservation</th>
<th>Accessibility</th>
<th>Data Integrity</th>
<th>Usability</th>
<th>Production</th>
<th>Quality</th>
<th>Data Quality</th>
<th>Data Quality</th>
<th>Data Quality</th>
<th>Data Quality</th>
<th>Data Quality</th>
<th>Transparency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 1 – Ad Hoc Not Managed</td>
<td>Any storage location Data only</td>
<td>Not publicly available Person-to-person</td>
<td>Unknown or no data integrity check</td>
<td>Extensive product-specific knowledge required No document online</td>
<td>Ad Hoc No obligation or deliverable requirement</td>
<td>Unknown or none</td>
<td>Data quality assurance (DQA) procedure unknown or none</td>
<td>Sampling: unknown or spotty Analysis: unknown or random in time</td>
<td>Algorithm theoretical basis assessed</td>
<td>Limited product information available Person-to-person</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 2 - Minimal Managed Limited</td>
<td>Non-designated repository archive redundancy Limited archival metadata</td>
<td>Publicly available Not searchable online</td>
<td>Data ingest integrity verifiable (e.g., checksum technology)</td>
<td>Non-standard data format Limited document (e.g., user's guide) online</td>
<td>Short-term Individual PI's commitment (grant obligations)</td>
<td>Completeness and redundancy check</td>
<td>Ad Hoc and random Procedure not defined and documented</td>
<td>Sampling and analysis are regular in time and space Limited product-specific metric defined &amp; implemented</td>
<td>Level 1 + Research product assessed</td>
<td>Product information available in literature</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 3 - Intermediate Managed Defined, Partially Implemented</td>
<td>Designated repository/archive redundancy Conforming to community archiving process and metadata Conforming to limited archival standards</td>
<td>Available online Limited data server performance Collection/dataset searchable</td>
<td>Level 2 + Data archive integrity verifiable</td>
<td>Community Standard-based Interoperable format &amp; metadata Documentation (e.g., source code, product algorithm document, processing or data flow diagram) online</td>
<td>Medium-term Institutional commitment (e.g., deliverables with specs and schedule defined)</td>
<td>Level 2 + Consistency check (i.e., temporal, logical)</td>
<td>Procedure defined and documented and partial implemented</td>
<td>Sampling and analysis are regular, frequent, systematic but not automatic Community metric defined and partially implemented Procedure documented and available online</td>
<td>Level 2 + Operational product assessed</td>
<td>Algorithm Theoretical Basis Document (ATBD) &amp; source code online Dataset configuration managed (CM) Data citation tracked (e.g., utilizing DOI system) Unique object identifier (DOI) assigned (dataset, documentation, source code)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 4 - Advanced Managed Well-Defined, Fully Implemented</td>
<td>Level 3 + Conforming to community archiving standards</td>
<td>Level 3 + Enhanced data server performance Granule/file searchable</td>
<td>Level 3 + Data access integrity verifiable</td>
<td>Conforming to community data integrity technology standard</td>
<td>Level 3 + Basic capability (e.g., subsetting, aggregating) &amp; data characterization (overall) global, e.g., climatology, error estimates available online</td>
<td>Level 3 + Long-term Institutional commitment Product improvement process in place</td>
<td>Level 3 + Accuracy check Procedures well documented and available online Limited data quality assurance metadata</td>
<td>Procedure well documented and available online with master reference data Limited data quality assurance metadata</td>
<td>Level 3 + Quality metadata assessed</td>
<td>Level 3 + OAD (Operational Algorithm Description) online (CM + DOI)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 5 - Optimal</td>
<td>Level 4 + Archiving process performance controlled, measured, and audited Future archiving process and standard changes planned</td>
<td>Level 4 + Dissemination reports available online Future technology changes planned</td>
<td>Level 4 + Data authenticity verifiable (e.g., data signature technology) Performance of data integrity check monitored and reported</td>
<td>Level 4 + Enhanced online capability (e.g., visualization, multiple data formats) Community metric set of data characterization (regional) online External ranking</td>
<td>Level 4 + National or International commitment Changes for technology planned</td>
<td>Level 4 + Conforming to community quality metadata &amp; standards</td>
<td>Level 4 + Conforming to community quality metadata &amp; standards</td>
<td>Level 4 + External review</td>
<td>Level 4 + Assessment performed on a recurring basis Conforming to community quality metadata &amp; standards External ranking</td>
<td>Level 4 + System information online Complete data provenance</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Version: SDS_MM_20140527_beta_v1.0**
POC: Ge.Feng@noaa.gov, Jeff.Privette@noaa.gov
Data Stewardship

All activities that preserve and improve the information content, accessibility, and usability of data and metadata (NRC, 2007) and that ensure or improve the quality and usability of environmental data.
**PRINCIPLE #6: Data and metadata require expert stewardship**

Data stewardship encompasses all activities that preserve and improve the information content, accessibility, and usability of data and metadata.

---

**Long-term CDR stewardship is important**

- Stewardship of the scientific data
- Scientific vigilance: Scientists provide oversight of data stewardship

---

**Eight Non-Functional Requirements on CDR**

(No Particular Order)

- Preserved
- Sustainable
- Extensible
- Accessible
- Transparent
- Reproducible
- Scientifically Defensible
- Assessed/Improved

J. Privette, CDRP

---

**NOAA Administrative Order 212-15 (2008)**

Management of environmental and geospatial data and information

- Include end-to-end data stewardship in data management planning
- Take appropriate steps to ensure acceptable accuracy, precision, representativeness, documentation, and long-term continuity of NOAA's quality data sets for the user community

---

**Data Quality Act (Publ. Law 106-554, 2001)**

- OMB to provide policy and guideline for ensuring and maximizing the quality, objectivity, utility, and integrity of information

---


- Data quality, objectivity, and integrity (security of information)
- Transparency — high degree for influential information
- Reproducibility
The Reference Model for an Open Archival Information System (OAIS)

The OAIS Reference Model attempts to comprehensively identify the responsibilities and components of an archival system, including:

- the roles of people and institutions that interact in an archive: Producer, Management, and Consumer
- the digital objects: information package
- the major functions: six higher-level functions – Ingest, Data Management, Archival Storage, Access, Preservation Management, and Administration, representing thirty-three lower-level functions

Source: OAIS Overview 22Arp2008.ppt


[http://wiki.esipfed.org/index.php/NOAA_Sessions_on_NIDIS_CLASS_the_Data_Centers_and_much_much_more](http://wiki.esipfed.org/index.php/NOAA_Sessions_on_NIDIS_CLASS_the_Data_Centers_and_much_much_more)

Source: www.dlib.org