The Scientific Data Stewardship Maturity Assessment Model Template

Template Version: NCDC-CICS-SMM-0001-Rev.1 v3.1 02/26/2015

Stewardship Maturity Matrix (SMM) for NCAR Global Climate Four-Dimensional Data Assimilation (CFDDA) Hourly 40 km Reanalysis as of 03/25/2015

Dataset Title	**NCAR Global Climate Four-Dimensional Data Assimilation (CFDDA) Hourly 40 km Reanalysis**
Dataset Information URL	**http://rda.ucar.edu/datasets/ds604.0/#!description**
Data Provider POC (Name; E-mail; Affiliation)	Terri Betancourt; terrib@ucar.edu; National Center for Atmospheric Research (NCAR) Research Applications Laboratory (RAL)
Dataset POC (Name; E-mail; Affiliation)	Grace Peng; grace@ucar.edu; National Center for Atmospheric Research (NCAR) Computational & Information Systems Laboratory (CISL), Data Support Section (DSS)
SMM Version (Document ID and Version Numbers)	NCDC-CICS-SMM_0001_Rev.1 12/09/2014
SMM POC (Name; E-mail; Affiliation)	Ge Peng; Ge.Peng@noaa.gov; Cooperative Institute for Climate and Satellites, North Carolina (CICS-NC), North Carolina State University & NOAA's National Climatic Data Center (NCDC)
SMM Assessment Version (v<nn>r<mm>, e.g., v01r00)	**V01r00**
SMM Assessment POC (Name; E-mail; Affilication)	Sophie Hou; hou@illinois.edu; University of Illinois at Urbana-Champaign
SMM Original Assessment Date (MM/DD/YYYY)	**03/25/2015**
SMM Original Assessment POC (Name; E-mail; Affiliation)	Sophie Hou; hou@illinois.edu; University of Illinois at Urbana-Champaign
SMM Last Modified Date (MM/DD/YYYY)	**03/25/2015**
SMM Last Modification POC (Name; E-mail; Affiliation)	Sophie Hou; hou@illinois.edu; University of Illinois at Urbana-Champaign
SMM Modified Date (MM/DD/YYYY)	
SMM Modification POC (Name; e-mail; Affiliation)	**(*** Repeat these last two lines to capture the SMM modification history ***)**

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	Stewardship Maturity Rating /Justification or Evidence	Comments/Recommendation
Preservability	Any storage location Data only	Non-designated repository Redundancy Limited archiving metadata	Designated archive Redundancy Community-standard archiving metadata Conforming to limited archiving standards	Level 3 + Conforming to community archiving standards	Level 4 + Archiving process performance controlled, measured, and audited Future archiving standard changes planned	• **Level: 3** • The designated archive is NCAR's Research Data Archive (RDA). • Data is regularly backed up as part of RDA's stewardship practices. • Although RDA currently only uses customized metadata format, RDA uses community-standard controlled vocabularies (GCMD) to represent its data parameters. • RDA has plans to crosswalk between its current metadata format and the ISO19115 in order to review and determine the applicability of the result for implementation. • Additional standardized processes and documentations have been planned for the ingest process.	• It would be helpful if the references to OAIS and ISO19115 as community standards are included in the evaluation criteria.
Accessibility	Not publicly available Person-to-person	Publicly available Direct file download (e.g., via anonymous FTP server) Collection/dataset level searchable online	Level 2 + Non-standard data service Limited data server performance Granule/file level searchable Limited search metrics	Level 3 + Community-standard data service Enhanced data server performance Conforming to community search metrics Dissemination report metrics defined and implemented internally	Level 4 + Dissemination reports available online Future technology and standard changes planned	• **Level: 1.5** • Although CFDDA's data are available for public access, registration and/or log in is required before data files can be downloaded directly. • In addition, although CFDDA's data are separated into sub-collections (type 1: grouped by individual year and then by the months of the year; type 2: grouped by data parameter), this level of granularity is not searchable online.	• Similar to preservability, it would be helpful if the examples of the community-standard data service provided in the paper are also referenced here.
Usability	Extensive product-specific knowledge required No documentation online	Non-standard data format Limited documentation (e.g., user's guide) online	Community standard-based interoperable format & metadata Documentation (e.g., source code, product algorithm	Level 3 + Basic capability (e.g., subsetting, aggregating) & data characterization (overall/global, e.g.,	Level 4 + Enhanced online capability (e.g., visualization, multiple data formats) Community metrics of data characterization (regional/cell) online	• **Level: 3** • The file format for CFDDA's data is netCDF. • The documentations regarding CFDDA are included as part of the data's public landing page, and the	•

			document, processing or/and data flow diagram) online	climatology, error estimates) available online	External ranking	information can be accessed and downloaded directly without log in. • Recommendations regarding the best tools to view and visualize CFDDA data have also been included on the landing page. • However, the data cannot be readily visualized, manipulated, or analyzed as-is in the online environment.	
Production Sustainability	Ad Hoc or Not applicable No obligation or deliverable requirement	Short-term Individual PI's commitment (grant obligations)	Medium-term Institutional commitment (contractual deliverables with specs and schedule defined)	Long-term Institutional commitment Product improvement process in place	Level 4 + National or international commitment Changes for technology planned	• **Level: 3** • The dataset was produced with institutional commitment.	• What would considered to be the definition of short, medium, and long term in terms of time scale? • The evaluation criteria might also need to highlight the availability of committed human resources (skillsets/expertise/knowledge)?
Data Quality Assurance	Data quality assurance (DQA) procedure unknown or none	Ad Hoc and random DQA procedure not defined and documented	DQA procedure defined and documented and partially implemented	DQA procedure well documented, fully implemented and available online with master reference data Limited data quality assurance metadata	Level 4 + DQA procedure monitored and reported Conforming to community quality metadata & standards External review	• **Level: 2.5** • Currently, this is no standardized DQA procedure defined, documented, and implemented. • Prior to ingest, data files for CFDDA were inspected, and necessary modifications were made to the files to ensure data accuracy. However, the review process was not a standardized process.	• It might be helpful to provide short definitions to clarify the differences between the 3 different quality elements.
Data Quality Control/Monitoring	None or Sampling unknown or spotty Analysis unknown or random in time	Sampling and analysis are regular in time and space Limited product-specific metrics defined & 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 3 + Anomaly detection procedure well-documented and fully implemented using community metrics, automatic, tracked and reported Limited quality monitoring metadata	Level 4 + Cross-validation of temporal & spatial characteristics Physical consistency check Conforming to community quality metadata & standards Dynamic providers/users feedback in place	• **Level: 1** • Extensive Data Quality Monitoring was performed by the data provider, but not independently verified by the data archive. • Data Quality reports/concerns will be investigated, documented by the archive, and made available online.	• To confirm, the Data Quality Control/Monitoring should be applied to the final dataset? In other words, do Data Quality Control/Monitoring measures implemented during the data generation stage count?
Data Quality Assessment	Algorithm/method /model theoretical basis assessed (methods and results online)	Level 1 + Research product assessed (methods and results online)	Level 2 + Operational product assessed (methods and results online)	Level 3 + Quality metadata assessed Limited quality assessment metadata	Level 4 + Assessment performed on a recurring basis Conforming to community quality metadata & standards External ranking	• **Level: 4** • CFDDA's data quality is assessed based on the reviews of the data products that have been produced from CFDDA.	• Would it be possible to clarify the term "quality metadata assessed" a bit further? After reading the paper, I am still not sure if it is the quality of the metadata that I should be evaluating or is it the quality of the *process* for assessing metadata quality that I should be evaluating?

Transparency /Traceability	Limited product information available Person-to-person	Product information available in literature	Algorithm 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 3 + Operational Algorithm Description (OAD) online, OID assigned, and under CM	Level 4 + System information online Complete data provenance online	• **Level: 4.5** • I would consider CFDDA to have transparency/traceability level to be 4.5 because even though the provenance information is not structured in the ATBD/OAD format, the analogous information is available as part of the data's land page and the information can be accessed and downloaded directly without log in. • RDA has plans to obtain DOI for CFDDA.	• Does ATBD and OAD apply to all data types? Based on this website (http://eospso.nasa.gov/content/algorithm-theoretical-basis-documents) , it seems to apply to only instrument based data? If ATBD and OAD do not apply to all data types, should it be a requirement to achieve Level 3 and 4? In other words, if ATBD and OAD do not apply to a particular data type and this data type has all of its other provenance available online, how should the level be assigned?
Data Integrity	Unknown or no data ingest integrity check	Data ingest integrity verifiable (e.g., checksum technology)	Level 2 + Data archive integrity verifiable	Level 3 + Data access integrity verifiable Conforming to community data integrity technology standard	Level 4 + Data authenticity verifiable (e.g., data signature technology) Performance of data integrity check monitored and reported	• **Level: 1** •	•

Citation for the paper: Peng, G., J.L. Privette, E.J. Kearns, N.A. Ritchey, and S. Ansari, 2015: A unified framework for measuring stewardship practices applied to digital environmental datasets. *Data Science Journal*, **13**, 231 - 253. Doi: 10.2481/dsj.14-049.

Citation for this template: Peng, G., 2015: The scientific data stewardship maturity assessment model template. Version: NCDC-CICS-SMM-0001-Rev.1 v3.1 02/26/2015. figshare. DOI: http://dx.doi.org/10.6084/m9.figshare.1211954. Date Accessed: mm/dd/yyyy.

Note: 1) All criterions need to be completely satisfied at the lower maturity level(s) before moving on to a higher maturity level, even if some practices were satisfied at the higher maturity level.

2) Use brown color-coded text to indicate that more information is needed or the evidence may not be true for all data files in a data collection or it may require additional assessment.

3) The color scheme for the maturity levels is provided in Table I.

4) A recommended way of displaying the stewardship maturity assessment result is shown in Figure 1.

Register: Users are encouraged to register to receive e-mail notifications of future updates. To do so, please send an e-mail with your name and affiliation to Maturity.Matrix@gmail.com with a subject line of SDS_MM_Register or register at http://goo.gl/kUW5Qq. Constructive comments and suggestions are encouraged.

NCDC or CICS-NC does not warrant that the template will be constantly available or available at all or the information within the template is complete, true, accurate, adequate or non-misleading. NCDC or CICS-NC will not be liable to you (whether under the law of contract, the law of torts or otherwise) in relation to the contents of, or use of, or otherwise in connection with, this temple for any direct loss, for any indirect, special or consequential loss; or for any business losses, loss of revenue, income, profits or anticipated savings, loss of contracts or business relationships, loss of reputation or goodwill, or loss or corruption of information or data. By using this template, you agree that the limitations of liability set out in this template disclaimer are reasonable. If you do not think they are reasonable, you must not use this template.

The layout or/and content of the matrix and template are subject to change any time without notification.

Stewards who carried out their self-evaluations of the stewardship maturity of their datasets are encouraged to document justifications in detail (with URL links if applicable) and make them available to data users at the dataset web sites to allow transparency and feedback from the users.

Any opinions or recommendations expressed here are those of the people who have carried out the assessment and do not necessarily reflect the views of NCDC or CICS-NC.

Table I: Scientific Data Stewardship Maturity Matrix Scale Definition and RGB Color Scheme

Maturity Scale	Definition	Color Code	R	G	B	Color
Level 1	Ad Hoc/Unknown: Not Managed	Lighter Green	229	244	224	
Level 2	Minimal: Managed Limited	Light Green	203	234	192	
Level 3	Intermediate: Managed, defined, partially implemented	Green	176	223	161	
Level 4	Advanced: Managed, Well-defined, fully implemented	Dark Green	85	168	57	
Level 5	Optimal: Level 4 +, measured, controlled, audit	Darker Green	56	112	38	

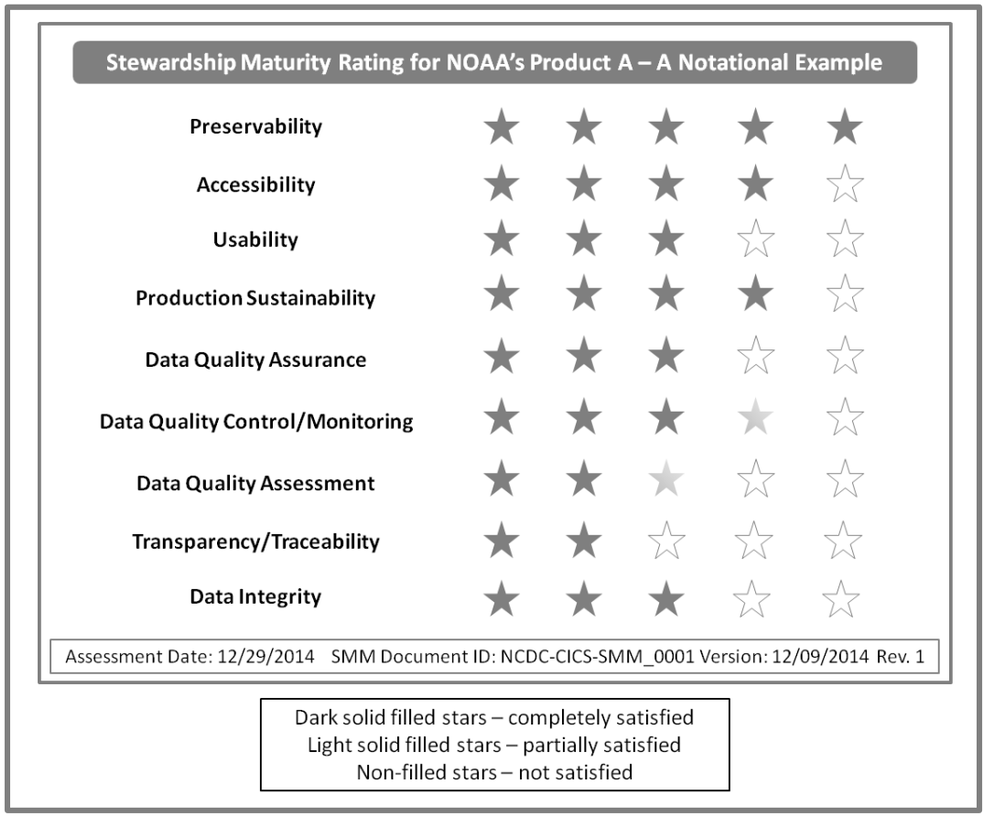

Figure 1: An example of displaying the stewardship maturity rating of your dataset*

*****To request an external review of the stewardship maturity assessment for your dataset, send your detailed assessment result utilizing this template to Maturity.Matrix@gmail.com with a subject line of SDS_MM_Scoreboard. As there is no operational support for this review service at this time, no guarantee will be made on the turn-around time.