

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 SBC LTER: pH time series: Water-sample pH and CO2 system chemistry, ongoing since 2011 as of 03/26/2015

Dataset Title	SBC LTER: pH time series: Water-sample pH and CO2 system chemistry, ongoing since 2011
Dataset Information URL	http://sbc.lternet.edu/cgi-bin/showDataset.cgi?docid=knb-lter-sbc.75
Data Provider POC (Name; E-mail; Affiliation)	Gretchen Hofmann; hofmann@lifesci.ucsb.edu ; University of California at Santa Barbara, Department of Ecology, Evolution and Marine Biology
Dataset POC (Name; E-mail; Affiliation)	Information Manager; sbelter@msi.ucsb.edu ; Santa Barbara Coastal LTER
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; Affiliation)	Sophie Hou; houl@illinois.edu ; University of Illinois at Urbana-Champaign
SMM Original Assessment Date (MM/DD/YYYY)	03/26/2015
SMM Original Assessment POC (Name; E-mail; Affiliation)	Sophie Hou; houl@illinois.edu ; University of Illinois at Urbana-Champaign
SMM Last Modified Date (MM/DD/YYYY)	03/26/2015
SMM Last Modification POC (Name; E-mail; Affiliation)	Sophie Hou; houl@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	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		
Key Component						Stewardship Maturity Rating /Justification or Evidence	Comments/Recommendation
<i>Preservability</i>	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	<ul style="list-style-type: none"> • Level: 3.5 • Dataset is archived with LTER dedicated local data repository. • Dataset is regularly backed up as part of LTER stewardship practices. • Although the dataset does not use ISO19115 metadata format, its Ecological Metadata Language (EML) format is widely recognized and adopted within the ecological discipline. 	<ul style="list-style-type: none"> • Does the “standard” really mean an ISO level or just in the context of community spec?
<i>Accessibility</i>	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	<ul style="list-style-type: none"> • Level: 2.5 • Although the dataset’s data file is available for public access, registration is required before the data file can be downloaded directly. • In addition, although the dataset is not searchable on the file level, its metadata is exposed to the users for search. • Further, while the LTER SBC’s local repository does not offer additional data services or data server performance, LTER SBC is one of DataONE’s member nodes. As a result, LTER SBC datasets are also available through DataONE Mercury. LTER also implements PASTA (Provenance Aware Synthesis Tracking Architecture) to ensure PASTA will automatically harvest data from LTER sites into a central warehouse and making the data available through a standard and well defined software interface. 	<ul style="list-style-type: none"> • How could “search metrics” be used to reflect the accessibility of a data’s maturity level? • Do data service and data server applicable for all data types? What if the data is meant to be used without needing data server capability such as visualization? • There might be a mismatch between the addressing the data’s accessibility versus the types of applications that help make the data accessible? • Perhaps a better category would be to judge data’s readiness for machine use, and examine both data and metadata?
<i>Usability</i>	Extensive product-specific knowledge required	Non-standard data format	Community standard-based interoperable format & metadata	Level 3 + Basic capability (e.g., subsetting, aggregating) & data	Level 4 + Enhanced online capability (e.g., visualization, multiple data formats)	<ul style="list-style-type: none"> • Level: 3 • The format for the dataset’s data file is csv. 	<ul style="list-style-type: none"> • This criteria might be better titled as “understandability”?

	No documentation online	Limited documentation (e.g., user's guide) online	Documentation (e.g., source code, product algorithm document, processing or/and data flow diagram) online	characterization (overall/global, e.g., climatology, error estimates) available online	Community metrics of data characterization (regional/cell) online External ranking	<ul style="list-style-type: none"> All the relevant information, such as description, parameters, methods of the dataset, is publicly viewable on the dataset's landing page. 	<ul style="list-style-type: none"> If this criteria is called "understandability," the previous criteria can be "usability" instead of "accessibility"?
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	<ul style="list-style-type: none"> Level: 3 The dataset was produced with institutional commitment. 	<ul style="list-style-type: none"> This criteria seems to be more about the resources that support the data? If yes, doesn't this refer back to the organization instead of the data?
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	<ul style="list-style-type: none"> Level: 2 Currently, this is no standardized DQA procedure defined, documented, and implemented. DQA is performed for the dataset based on the information manager's experience. 	<ul style="list-style-type: none"> Perhaps the different maturity levels should reflect the evidence of the different types of data quality assurance performed?
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	<ul style="list-style-type: none"> Level: 1 Sampling and analysis as well as product-specific metrics evaluation are performed selectively. 	<ul style="list-style-type: none"> Similar to Data Quality Assurance, perhaps the wording for the levels could reflect whether the referenced practices have been applied instead of the availability of the processes?
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	<ul style="list-style-type: none"> Level: 4 Dataset's data quality is assessed during the ingest phase. 	<ul style="list-style-type: none"> Should the task of ensuring "that the products are scientifically sound" be part of data stewardship activities? Perhaps "Data Quality" is more about data relevance to the proposed science domain?
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	<ul style="list-style-type: none"> Level: 3 The dataset does have DOI assigned and the citation is tracked. In addition, even though the information about production creation is not presented in the ATBD/OAD format, the information is available both under the dataset's public landing page and as part of the metadata. 	<ul style="list-style-type: none"> I think "traceability" is used to mean provenance in this context? If yes OID is more about identification?

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	<ul style="list-style-type: none"> • Level: 1 • 	•
-----------------------	---	--	--	---	---	--	---

Creative Commons License – Attribution (BY)-NC (Non-Commercial)

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 criteria 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.

Disclaimer: This template is provided “as is” without any representations or warranties, express or implied. NCDC or CICS-NC makes no representations or warranties in relation to this template or the information and materials provided on this template. Use for the template is intended for use as a preliminary stewardship maturity assessment of a dataset, utilizing the latest NCDC/CICS-NC scientific data stewardship maturity matrix.






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 template 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 1: 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	

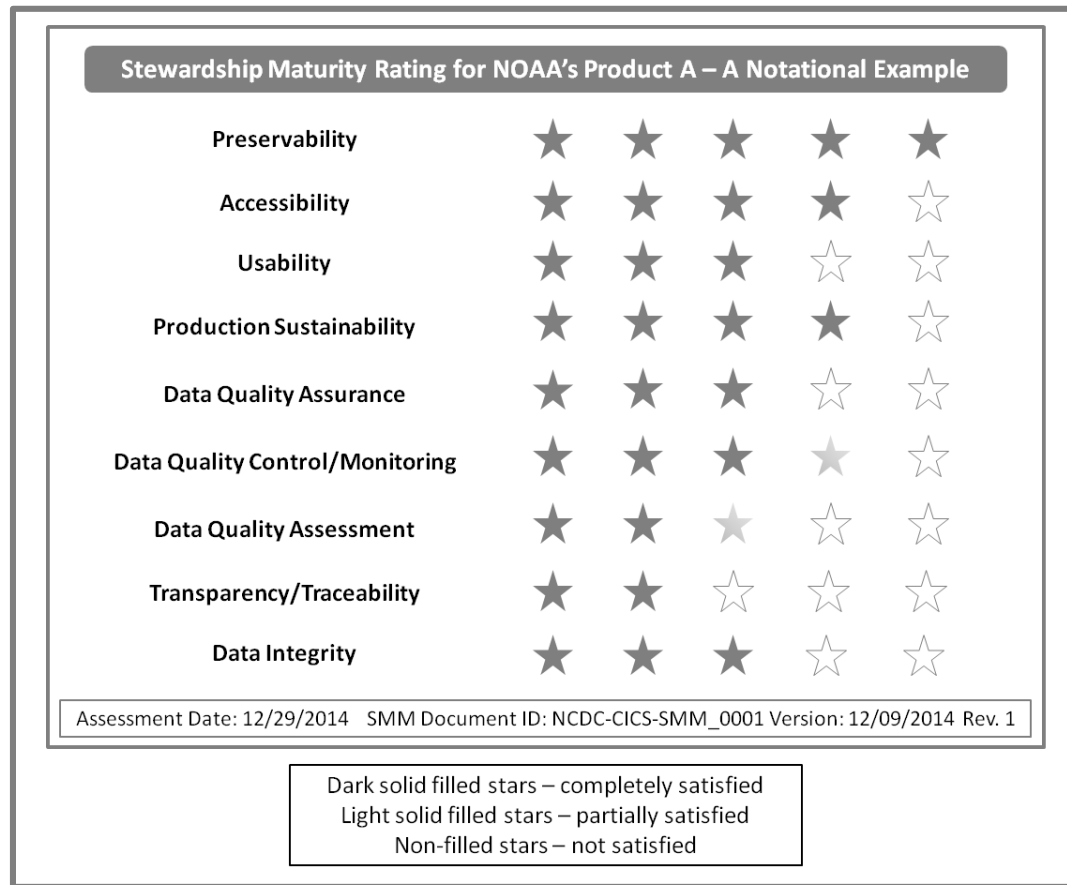


Figure 1: An example of displaying the stewardship maturity rating of your dataset*

*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.