



Data Stewardship Maturity Matrix – Use Case Study Session

ESIP Summer Meeting
July 2015



Table of Content

- Two different examples of using the Data Stewardship Maturity Matrix:
 1. National Center for Atmospheric Research (NCAR) Global Climate Four-Dimensional Data Assimilation (CFDDA) Hourly 40 km Reanalysis
 2. Santa Barbara Coastal (SBC) Long Term Ecological Research (LTER): pH time series: Water-sample pH and CO₂ system chemistry, ongoing since 2011
- Evaluation was performed also to provide feedback regarding Rev.1 v3.1 02/26/2015 of the Data Maturity Matrix template.



NCAR Global Climate Four-Dimensional Data Assimilation (CFDDA) Hourly 40 km Reanalysis



CFDDA Dataset

- Dataset Description:

NCAR Global Climate Four-Dimensional Data Assimilation (CFDDA) Hourly 40 km Reanalysis dataset is a dynamically-downscaled dataset with that was created using NCAR's CFDDA system. The dataset contains three-dimensional hourly analyses in netCDF format for the global atmospheric state from 1985 to 2005 on a 40 km horizontal grid (0.4 degree grid increment) with 28 vertical levels.

- Dataset Landing Page:

<http://rda.ucar.edu/datasets/ds604.0/#!description>

CFDDA Dataset - Continued

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 • 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. 	<ul style="list-style-type: none"> • It would be helpful if the references to OAIS and ISO19115 as community standards are included in the evaluation criteria.
<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: 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. 	<ul style="list-style-type: none"> • Similar to preservability, it would be helpful if the examples of the community-standard data service provided in the paper are also referenced here.
<i>Usability</i>	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	<ul style="list-style-type: none"> • 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 	<ul style="list-style-type: none"> •



CFDDA Dataset - Continued

- Dataset Identifier:

<http://dx.doi.org/10.5065/D6M32STK>

- Dataset's Data Maturity Matrix Evaluation Result:

http://wiki.esipfed.org/images/a/a1/CFDDA_DataMaturityMatrix_Evaluation_Rev1.pdf



SBC LTER: pH time series: Water-sample pH and CO₂ system chemistry, ongoing since 2011



SBC LTER Dataset

- Dataset Description:

Data are a time-series of pH and carbonate chemistry in manually-collected sea water samples from nine near-shore locations along the Santa Barbara Channel, California, USA, and intended to benchmark data from moored pH instruments. Data collection began in June 2011 and is ongoing.

- Dataset Landing Page

<http://sbc.lternet.edu/cgi-bin/showDataset.cgi?docid=knb-lter-sbc.75>



SBC LTER Dataset - Continued

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	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	<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?
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	<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?
Usability	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”?



SBC LTER Dataset - Continued

- Dataset Identifier:

<http://dx.doi.org/10.6073/pasta/e06713c1ee0e48b41162767b4610b451>

- Dataset's Data Maturity Matrix Evaluation Result:

http://wiki.esipfed.org/images/f/fc/SBC75_DataMaturityMatrix_Evaluation_Rev1.pdf



Lessons Learned

- The associated publication is a valuable resource before start working on matrix.
 - Peng, G., Privette, J. L., Kearns, E. J., Ritchey, N. A., & Ansari, S. (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](https://doi.org/10.2481/dsj.14-049)
- The focus of the matrix is on evaluating the data instead of the information organization.
 - When an information organization applies its processes consistently to all of its data, the maturity level of the organization and its data could be correlated.
 - However, there is often a distribution of different maturity levels within an information organization.



Lessons Learned - Continued

- The maturity evaluation process should involve members with different roles/responsibilities from the dataset/project team.
- The process of formalizing the evidences in order to justify the rating for each of the maturity categories can provide the following benefits:
 - Allow the information organization to reflect on the current status of its data stewardship process.
 - Identify existing data stewardship practices that can be enhanced further.
 - Prioritize gaps or needs in stewardship areas that require additional attention.
 - Create roadmap to improve maturity level for the individual datasets as well as the overall dataset collection.



Additional Resources

- Data Maturity Matrix Wiki –
 - http://wiki.esipfed.org/index.php/Data_Maturity_Matrix
 - Meeting notes from the Maturity Matrix Evaluation Review Session:
 - http://wiki.esipfed.org/index.php/Data_Maturity_Matrix/meeting_notes_20150406
 - Discussions and clarifications for each matrix category were provided.
 - Frequently Asked Questions