NOAA
National Satellite and Information Service

April 20-22, 2021

Data Stewardship Quality Assessment at an Enterprise-scale

CoMET and the DSMQ

Nancy Ritchey, NOAA, NESDIS, NCEI, Archive Branch Chief
Paul Lemieux III, NOAA, NESDIS, NCEI, Physical Scientist

Presentation Outline

- DSMM Overview
  - Background
  - Key Components
  - Matrix Template
- DSMM Incorporation into Collection Metadata Editing Tool (CoMET)
  - DSMM to DSMQ Translation
  - Scoring and Mapping (how we did it)
  - DSMQ to DSMR Conversion (in active development)
    - NOAA Library's Institutional Repository (including DOI minting)
    - Cross-linking Data and Assessment

NOAA OneStop

To improve the discovery of, access to, and usability of NOAA's vast and diverse collection of big data.

- Modern, user-friendly discovery interface
- Open infrastructure
- Standards and best practices

Data Stewardship Maturity Matrix (DSMM)

A unified framework for measuring *Stewardship Practices* applied to individual digital earth sciences data products

http://tinyurl.com/DSMMintro

DSMM Evaluates Stewardship Maturity in Nine Key Components

- Consistent framework for assessing and reporting quantifiable stewardship practices.
- Allows for greater stewardship quality transparency, builds trust among user communities, and contributes to the reproducibility of NOAA's data products
- Developed jointly by domain experts leveraging institutional knowledge and community best practices and standards
- Vetted through use case studies with diverse datasets managed by different organizations, in collaboration with NC State University, NCEI Data Stewardship Division, and U.S. and int'l data stewardship groups

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
Preservability	*The state of dataset being preservable*				
Accessibility	*The state of dataset being publicly searchable and accessible*				
Usability	*The state of data product being easy to understand and use*				
Production Sustainability	*The state of data production being sustainable and extendable*				
Data Quality Assurance	*The state of data product quality being assured/screened*				
Data Quality Control /Monitoring	*The state of data product quality being controlled and monitored*				
Data Quality Assessment	*The state of data product quality being assessed*				
Transparency /Traceability	*The state of data product being transparent, trackable, and traceable*				
Data Integrity	*The state of data integrity being verifiable*				

DSMM Template

Dataset Title and URL

Assessment Version

Assessment Ratings

Assessor Info and Dates

Stewardship Maturity Matrix (SMM) as of <04/24/2019>

For <EPS_VIIRS_AF_EDR>

Dataset Title	**NOAA JPSS Visible Infrared Imaging Radiometer Suite (VIIRS) Active Fires Environmental Data Record (EDR) from NDE**
Dataset Information URL	https://doi.org/10.7289/V5GH9G0T
Data Provider POC (Name; E-mail; Affiliation)	Customer Engagement Branch, ncei.sat.info@noaa.gov, DOC/NOAA/NESDIS/NCEI > National Centers for Environmental Information, NESDIS, NOAA, U.S. Department of Commerce
Dataset POC (Name; E-mail; Affiliation)	Ivan Csiszar, ivan.csiszar@noaa.gov, DOC/NOAA/NESDIS/STAR > Center for Satellite Applications and Research, NESDIS, NOAA, U.S. Department of Commerce
SMM Version (Document ID and Version Number)	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 (NCSU)
SMM Template Version (Document ID and Version Numbers)	NCDC-CICS-SMM_0001_Rev.1 v4.0 06/23/2015
SMM Template 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 (NCSU)
SMM Assessment Version (v<nn>r<mm>, e.g., v01r00)	**v02r02**
SMM Assessment Date (MM/DD/YYYY)	**05/19/2017**
SMM Assessment POC (Name; E-mail; Affiliation)	**Paul Lemieux III, paul.lemieux@noaa.gov, Earth Resources Technology, Inc.**
Stewardship Maturity Ratings (each key component) (kc1/kc2/kc3/kc4/kc5/kc6/kc7/kc8/kc9)	**4.5 / 5 / 4.5 / 5 / 4.5 / 4.75 / 3.5 / 3.5 / 5**
SMM Original Assessment Date (MM/DD/YYYY)	**09/14/2016**
SMM Original Assessment POC (Name; E-mail; Affiliation)	**Paul Lemieux III, paul.lemieux@noaa.gov, Earth Resources Technology, Inc.**
SMM Last Modified Date (MM/DD/YYYY)	**04/24/2019**
SMM Last Modification POC (Name; E-mail; Affiliation)	**Paul Lemieux III, paul.lemieux@noaa.gov, Riverside Technology, Inc.**
SMM Modified Date (MM/DD/YYYY) [2]	**04/24/2019**
SMM Modification POC (Name; E-mail; Affiliation)	**Paul Lemieux III, paul.lemieux@noaa.gov, Riverside Technology, Inc.**
[1] NCEI includes the organizations previously referred to as National Climatic Data Center (NCDC), National Geophysical Data Center (NGDC), and National Oceanographic Data Center (NODC).	
[2] Repeat these last two lines to capture the SMM modification history	

DSMM Template cont.

Justification Statements

Key Components

Rating Criteria

Additional Comments

Maturity Scale (across) / Key Component (below)	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 And Justification or Evidence	Comments
Preservability (The state of being preservable)	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 4.5 - Archived by NCEI which is a NOAA [illegible]gnated archive compliant to NARA stand[illegible]. - Metadata following ISO 19115 [illegible]. - Compliant to OIAS RM. - Plans to update metadata to ISO 19115-1 at a later date and may be a pilot dataset for the OneStop initiative. - Using CLASS.	CLASS is CMMI-Level 3. No known audits on the archiving processes.
Accessibility (The state of being searchable and accessibl[illegible])	Not publicly availabl[illegible] [illegible]n-to-person	Publicly available Dir[illegible]e [illegible]load (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 5 - Collection level searchable online. - CLASS ordering: http://www.class.noaa.gov/saa/products/search?sub_id=0&datatype_family=JPSS_GRAN - Granules searchable and orderable via CLASS. - Direct file download available via CLASS FTP: ftp://ftp-npp.bou.class.noaa.gov/ - CLASS has dissemination reports available internally and externally. Users have to e-mail the CLASS Help Desk to request access to the metrics tools. - New technology for OneStop search and discovery planned (i.e. ElasticSearch, Hyrax Servers, etc.) This is part of the JPSS data group that will be OneStop ready.	A CLASS improvement (GEARS) is in the works.
Usability (The state of being easy to use)	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 document, processing or/and data flow diagram) online	Level 3 + Basic capability (e.g., subsetting, aggregating) & data characterization (overall/global, e.g., climatology, error estimates) available online	Level 4 + Enhanced online capability (e.g., visualization, multiple data formats) Community metrics of data characterization (regional/cell) online External ranking	❖ Level 4.5 - Satellite community standard netCDF data format compliant to ACDD and CF conventions. - Data paper {{Schroeder, Oliva, et al, 2014} Schroeder, W., Oliva, P., et al, (2014), The new VIIRS 375 m active fire detection data product: algorithm description and initial assessment, _Rem. Sens. Env._, 143, 85—96, doi:10.1016/j.rse.2013.12.008.} published in peer-reviewed journal available online here: https://doi.org/10.1016/j.rse.2013.12.008 - Data visualization capability and data download in multiple formats available online at the following two pages: http://viirsfire.geog.umd.edu/pages/mapsData.php and	Multiple visualization options through JPSS LTM and UMD project page. Multiple formats available for CONUS through the UMD Active Fires link. No known external rankings.

DSMM Template cont.

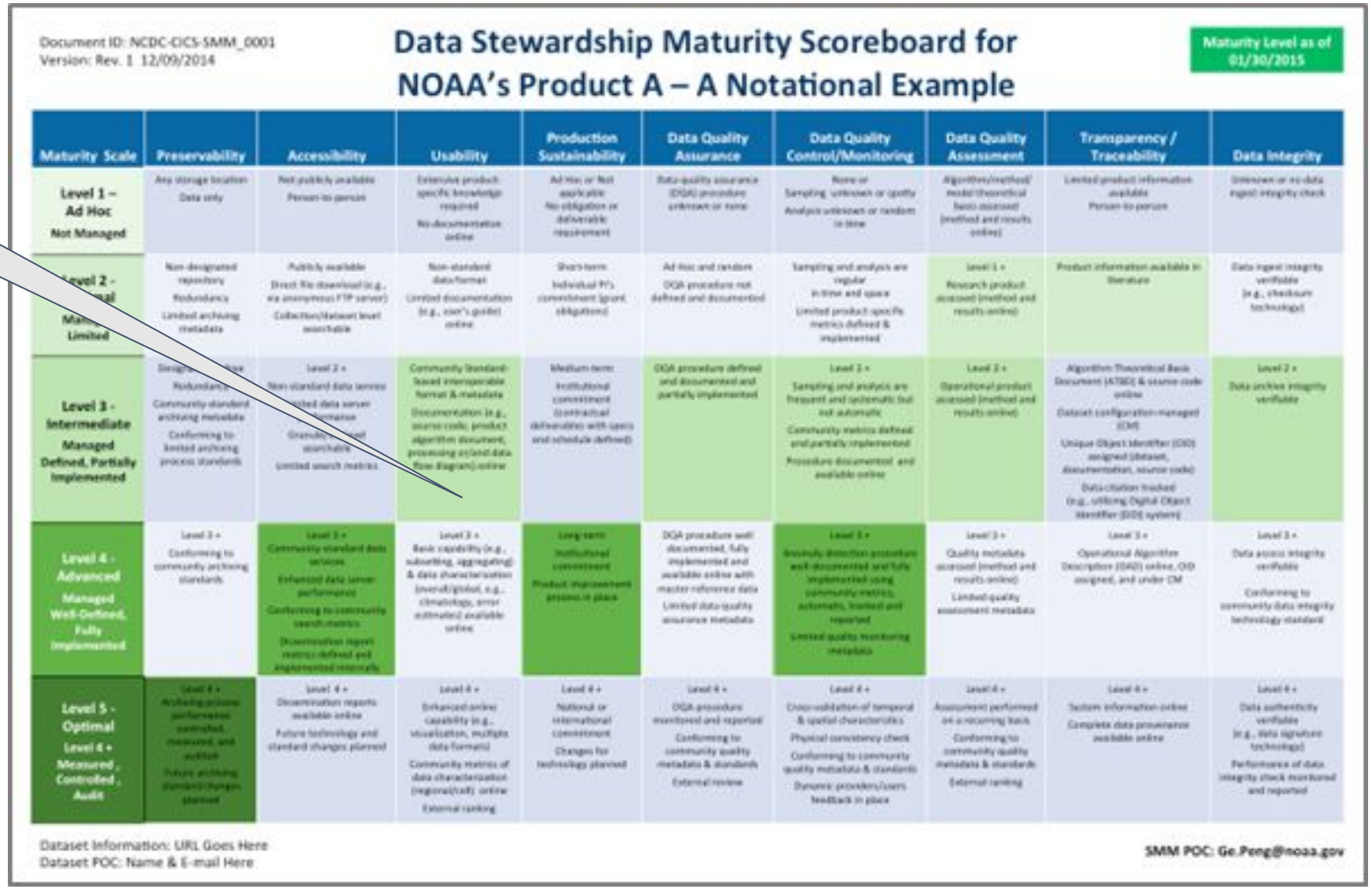
Assessment Version

Color coded for better visibility

Dataset Information

Assessment Date

Assessor POC



Document ID: NCDC-CICS-SMM_0001
Version: Rev. 1 12/09/2014

Data Stewardship Maturity Scoreboard for NOAA's Product A – A Notational Example

Maturity Level as of 01/30/2015

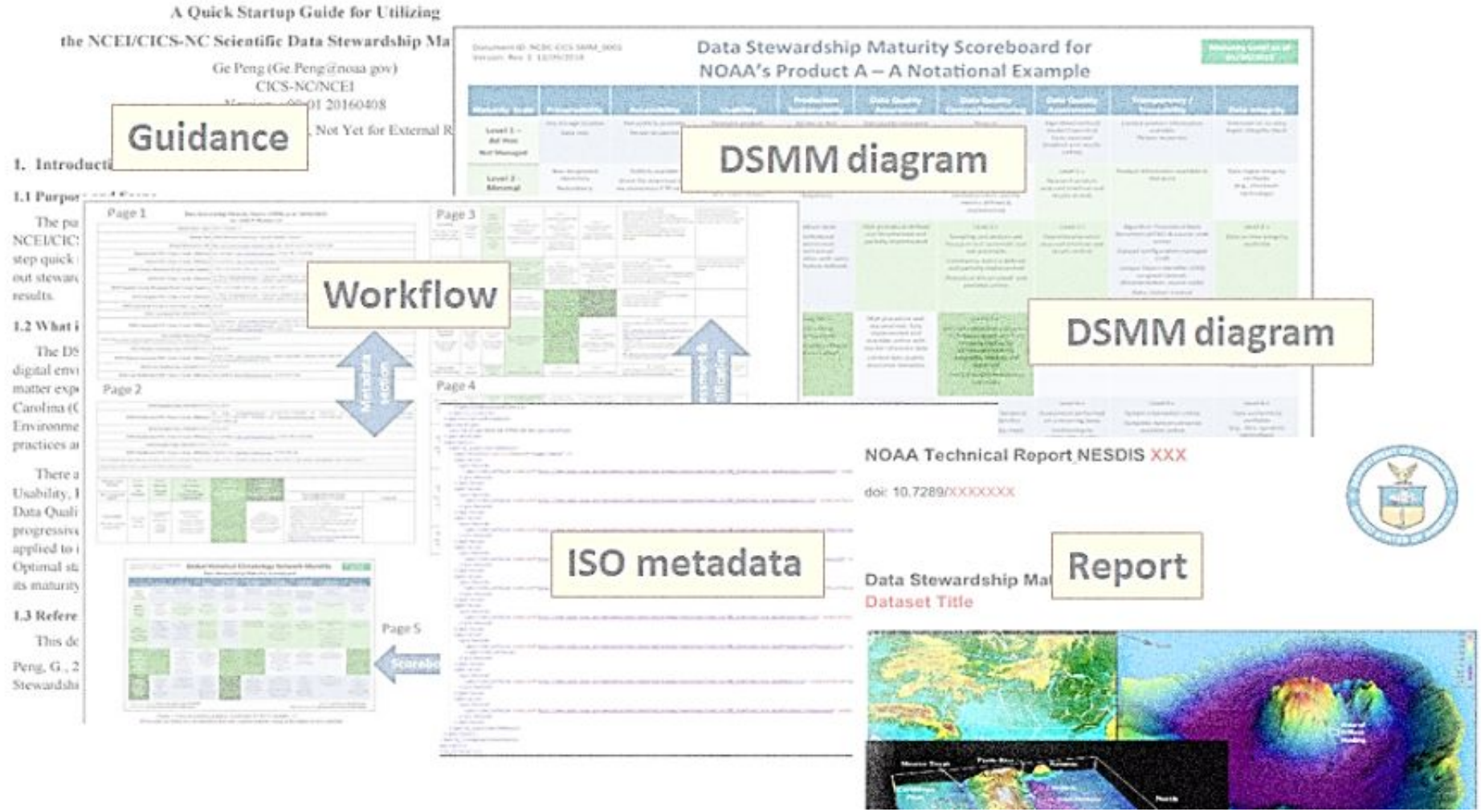
Maturity Scale	Preservability	Accessibility	Usability	Production Sustainability	Data Quality Assurance	Data Quality Control/Monitoring	Data Quality Assessment	Transparency / Traceability	Data Integrity
Level 1 – Ad Hoc Not Managed	Any storage location Data only	Not publicly available Person-to-person	Extensive product-specific knowledge required No documentation online	Ad Hoc or Not applicable No obligation or deliverable requirement	Data quality assurance (DQA) procedure unknown or none	None or Sampling unknown or spotty Analysis unknown or random in time	Algorithm/method/model theoretical basis assessed (method and results online)	Limited product information available Person-to-person	Unknown or no data ingest integrity check
Level 2 – [illegible] Limited	Non-designated repository Redundancy Limited archiving metadata	Publicly available Direct file download (e.g., via anonymous FTP server) Collection/dataset level searchable	Non-standard data format Limited documentation (e.g., user's guide) online	Short-term Individual PI's commitment (grant obligations)	Ad Hoc and random DQA procedure not defined and documented	Sampling and analysis are regular in time and space Limited product specific metrics defined & implemented	Level 1 + Research product assessed (method and results online)	Product information available in literature	Data ingest integrity verifiable (e.g., checksum technique)
Level 3 - Intermediate Managed Defined, Partially Implemented	Designated [illegible] Redundancy Community standard archiving metadata Conforming to limited archiving process standards	Level 2 + Non-standard data service [illegible] data server performance Granule-level searchable Limited search metrics	Community Standard-based interoperable format & metadata Documentation (e.g., source code, product algorithm document, processing or/and data flow diagram) online	Medium-term Institutional commitment (contractual deliverables with specs and schedule defined)	DQA procedure defined and documented and partially 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 2 + Operational product assessed (method and results online)	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 2 + Data archive integrity verifiable
Level 4 - Advanced Managed Well-Defined, Fully Implemented	Level 3 + Conforming to community archiving standards	Level 3 + Community standard data services Enhanced data server performance Conforming to community search metrics Dissemination report metrics defined and implemented externally	Level 3 + Basic capability (e.g., subsetting, aggregating) & data characterization (overall/global, e.g., climatology, error estimates) available online	Long-term Institutional commitment Product improvement process in place	DQA procedure well documented, fully implemented and available online with master reference data Limited data quality assurance metadata	Level 3 + Anomaly detection procedure well documented and fully implemented using community metrics, automatic, tracked and reported Limited quality monitoring metadata	Level 3 + Quality metadata assessed (method and results online) Limited quality assessment metadata	Level 3 + Operational Algorithm Description (OAD) online, OID assigned, and under CM	Level 3 + Data access integrity verifiable Conforming to community data integrity technology standard
Level 5 - Optimal Level 4 + Measured, Controlled, Audit	Level 4 + Archiving process performance controlled, measured, and audited Future archiving standards changes planned	Level 4 + Dissemination reports available online Future technology and standard changes planned	Level 4 + Enhanced online capability (e.g., visualization, multiple data formats) Community metrics of data characterization (regional/cell) online External ranking	Level 4 + National or international commitment Changes for technology planned	Level 4 + DQA procedure monitored and reported Conforming to community quality metadata & standards External review	Level 4 + Cross-validation of temporal & spatial characteristics Physical consistency check Conforming to community quality metadata & standards Dynamic providers/users feedback in place	Level 4 + Assessment performed on a recurring basis Conforming to community quality metadata & standards External ranking	Level 4 + System information online Complete data provenance available online	Level 4 + Data authenticity verifiable (e.g., data signature technique) Performance of data integrity check monitored and reported

Dataset Information: URL Goes Here
Dataset POC: Name & E-mail Here

SMM POC: Ge.Peng@noaa.gov

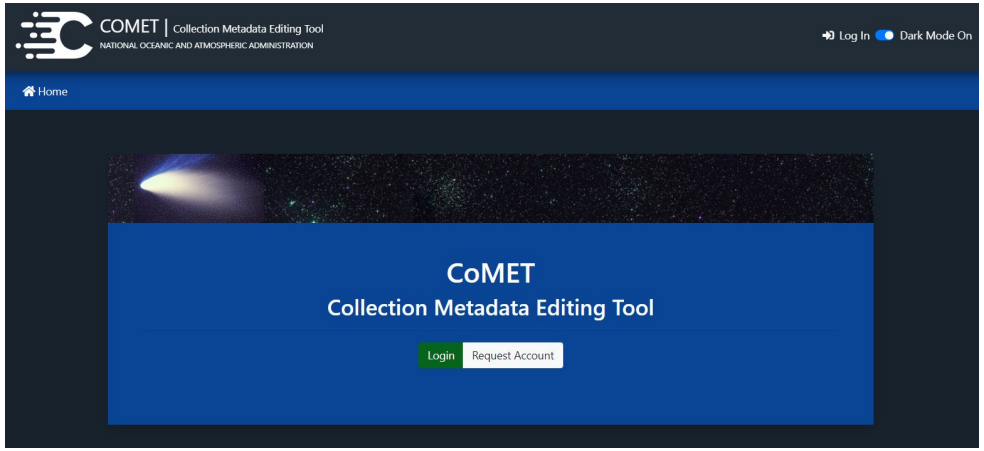
DSMM Guidance, Template, Tools

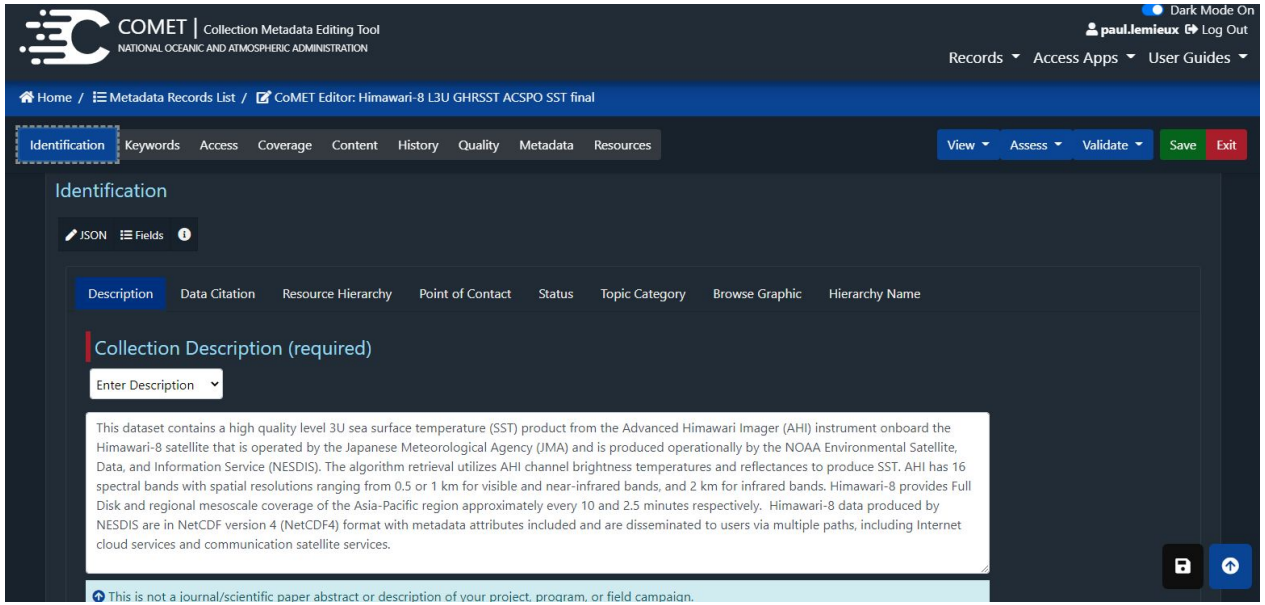
Unable to scale up as a manual process, too labor intensive and too complex!



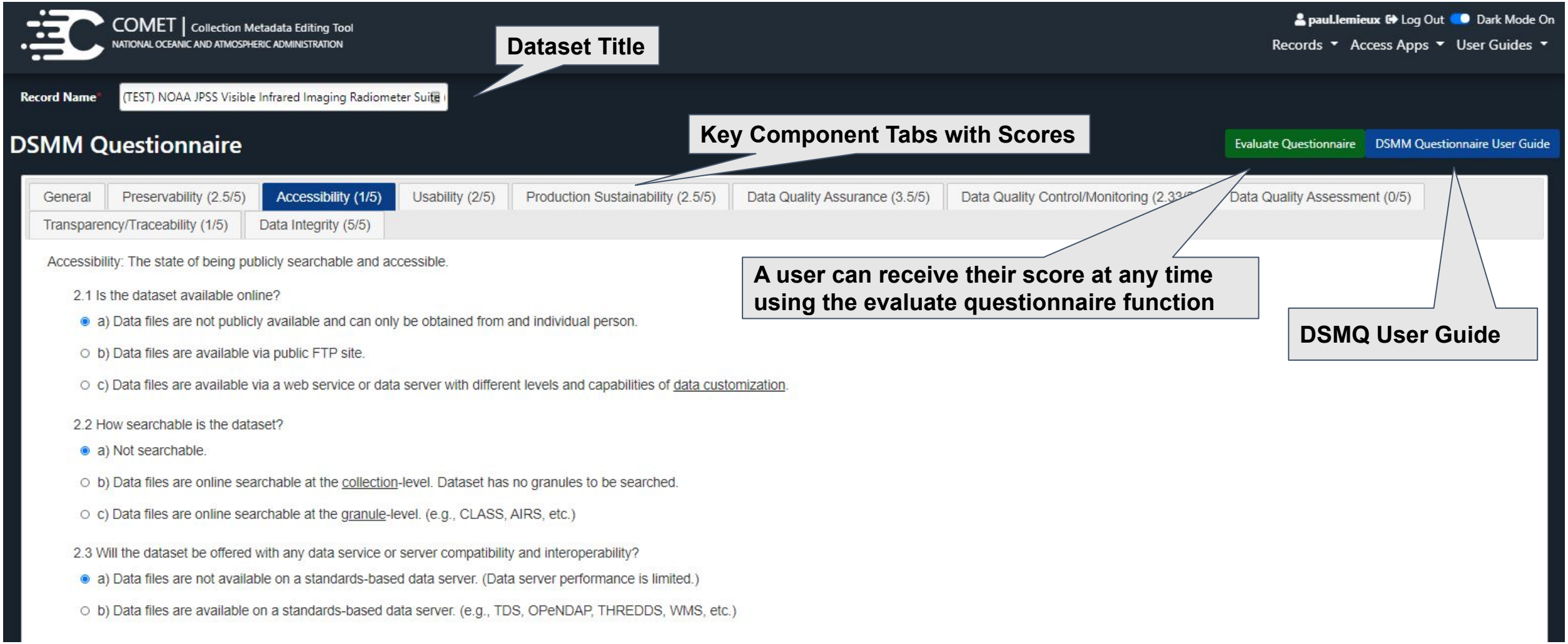
DSMM Template Evolution with CoMET

- CoMET provides the enterprise architecture needed to evolve the DSMM into a fully functional and easy to use enterprise tool for assessing stewardship quality of environmental datasets.
- The Collection Manager tool suite is an integral component of the NCEI archive enterprise infrastructure.
- CoMET supports the documentation of data sets using the NCEI Collection Metadata template, and assists users in meeting all NCEI and NOAA standards and best practices to support FAIR data principles.





DSMM Template Evolution with CoMET



DSMQ Scoring and Mapping

- Scores were determined by identifying every possible permutation of answer combinations and assigning a score to EACH one
- CoMET developers translated those answer combinations and scores into their code

Key Components

Scores

Answer Combinations

A1 | *fx* | Section ID

	A	B	C	D
28	2	Accessibility	2.00	2.1b and 2.2b and 2.3a and 2.4a and 2.5a-c and 2.6a-c
29	2	Accessibility	2.25	2.1b-c and 2.2c and 2.3a and 2.4a and 2.5a-c and 2.6a-c
30	2	Accessibility	2.25	2.1b-c and 2.2b and 2.3a and 2.4b and 2.5a-c and 2.6a-c
31	2	Accessibility	2.50	2.1b-c and 2.2b and 2.3b and 2.4a and 2.5a-c and 2.6a-c
32	2	Accessibility	2.50	2.1b-c and 2.2c and 2.3a and 2.4b-c and 2.5a-c and 2.6a-c
33	2	Accessibility	2.75	2.1b-c and 2.2c and 2.3b and 2.4a and 2.5a-c and 2.6a-c
34	2	Accessibility	2.75	2.1b-c and 2.2b and 2.3b and 2.4b-c and 2.5a-c and 2.6a-c
35	2	Accessibility	3.00	2.1b and 2.2c and 2.3b and 2.4b and 2.5a and 2.6a-c
36	2	Accessibility	3.25	2.1b and 2.2c and 2.3b and 2.4b and 2.5a and 2.6a-c
37	2	Accessibility	3.50	2.1c and 2.2c and 2.3b and 2.4b and 2.5a and 2.6a-c
38	2	Accessibility	3.50	2.1b and 2.2c and 2.3b and 2.4c and 2.5a and 2.6a-c
39	2	Accessibility	3.50	2.1b and 2.2c and 2.3b and 2.4b and 2.5b and 2.6a-c
40	2	Accessibility	3.75	2.1c and 2.2c and 2.3b and 2.4c and 2.5a and 2.6a-c
41	2	Accessibility	3.75	2.1c and 2.2c and 2.3b and 2.4b and 2.5b and 2.6a-c
42	2	Accessibility	3.75	2.1b and 2.2c and 2.3b and 2.4c and 2.5b and 2.6a-c
43	2	Accessibility	4.00	2.1c and 2.2c and 2.3b and 2.4c and 2.5b and 2.6a-c
44	2	Accessibility	4.50	2.1c and 2.2c and 2.3b and 2.4c and 2.5b and 2.6b-c
45	2	Accessibility	4.50	2.1c and 2.2c and 2.3b and 2.4c and 2.5c and 2.6a
46	2	Accessibility	5.00	2.1c and 2.2c and 2.3b and 2.4c and 2.5c and 2.6b-c
47	3	Usability	1.00	3.1a and 3.2a and 3.3a-c and 3.4a-c and 3.5a-b
48	3	Usability	1.50	3.1a and 3.2a and 3.3a-c and 3.4a-c and 3.5a-b
49	3	Usability	2.00	3.1b and 3.2a and 3.3a-c and 3.4a-c and 3.5a-b
50	3	Usability	2.50	3.1c and 3.2b and 3.3a-c and 3.4a-c and 3.5a-b
51	3	Usability	2.50	3.1b and 3.2c and 3.3a-c and 3.4a-c and 3.5a-b
52	3	Usability	3.00	3.1c and 3.2c and 3.3a and 3.4a and 3.5a-b

DSMQ Overall Scoring

- Each section was evaluated, producing over 1,600 possible answer combinations
- CoMET provides real-time evaluation of each key component and a final score upon completion of the assessment

A1 | fx | Section

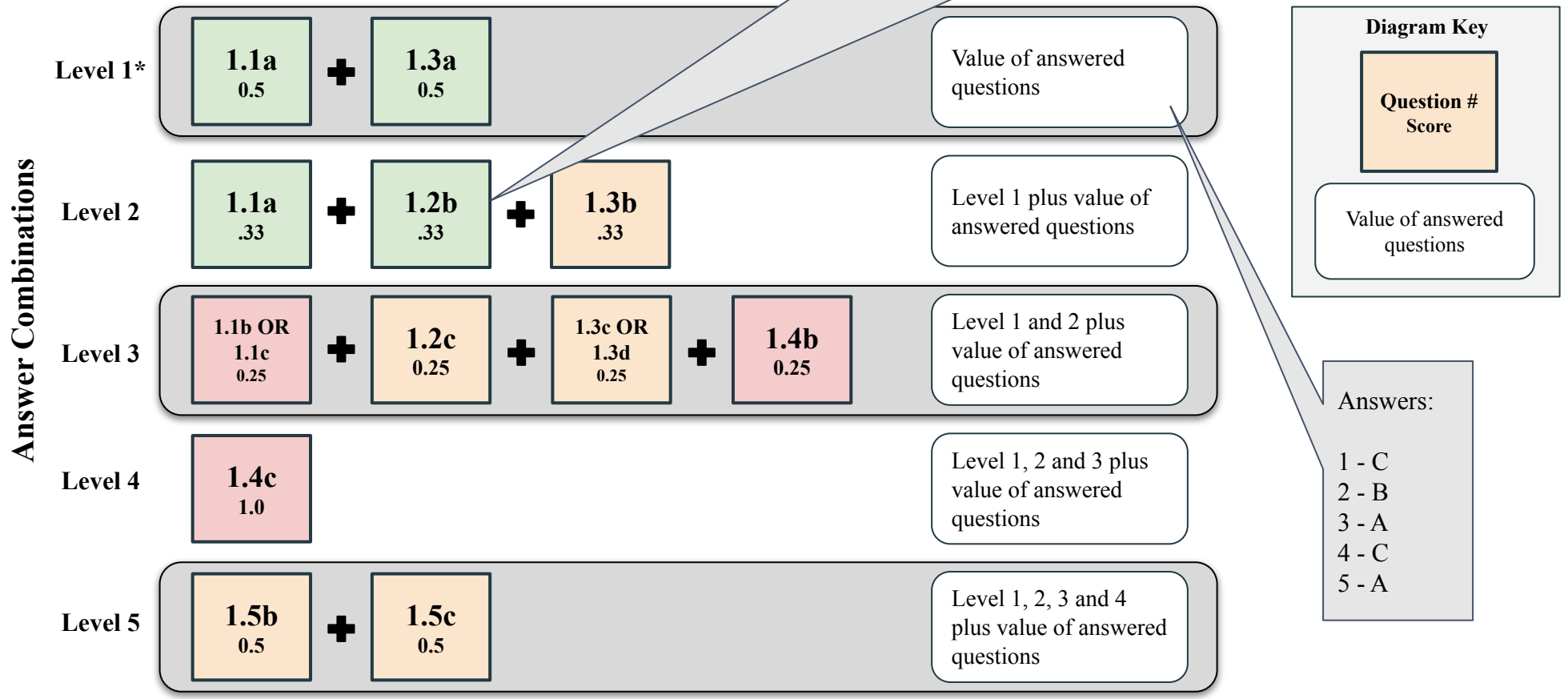
	A	B
1	**Section**	**Possible Combinations**
2	Preservability	324
3	Accessibility	486
4	Usability	162
5	Production Sustainability	27
6	Data Quality Assurance	72
7	Data Quality Control/Monitoring	360
8	Data Quality Assessment	64
9	Transparency/Traceability	90
10	Data Integrity	45
11		

Total = 1,630 !!!

Implementing DSMQ Scoring Logic

Section 1 - Preservability

In this example, respective answers satisfy a score of 1.66. This is displayed in green. The red cells represent answers that would have been satisfied but were out of range of the potential score due to lack of previous satisfactions.

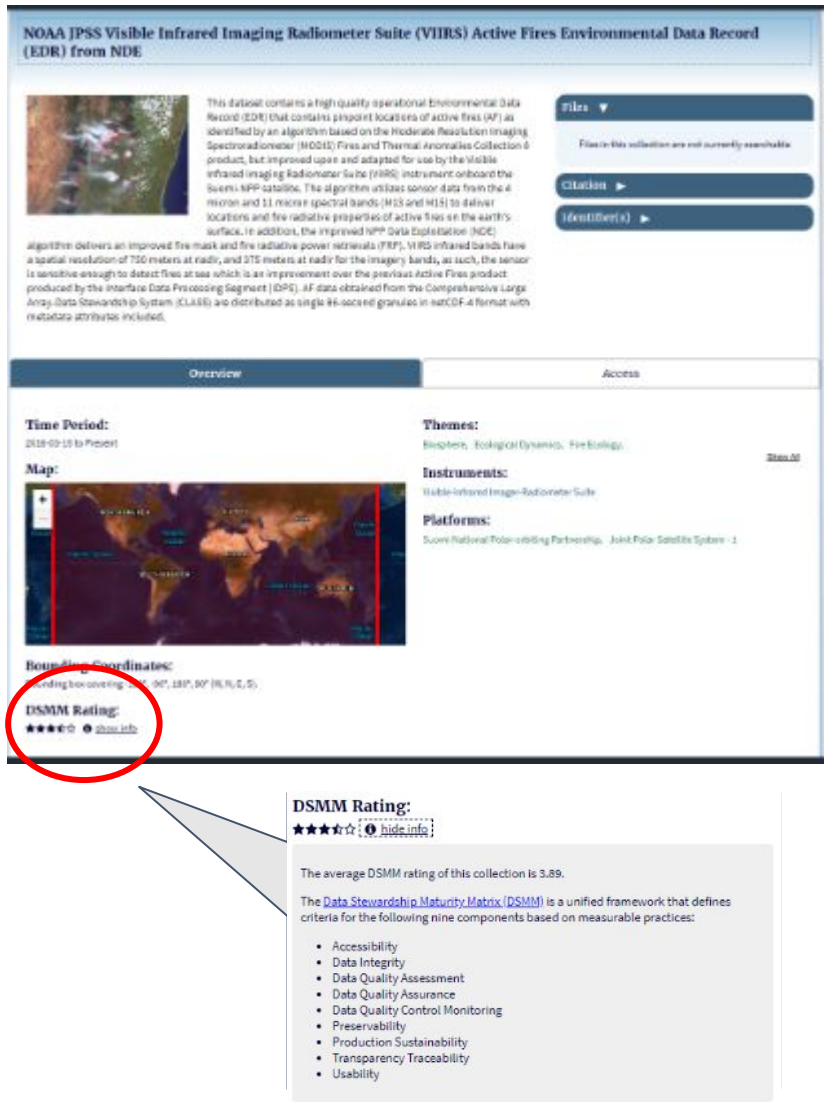


*All criteria of a level needs to be satisfied before advancing to the next level

DSMQ Results -> ISO Metadata

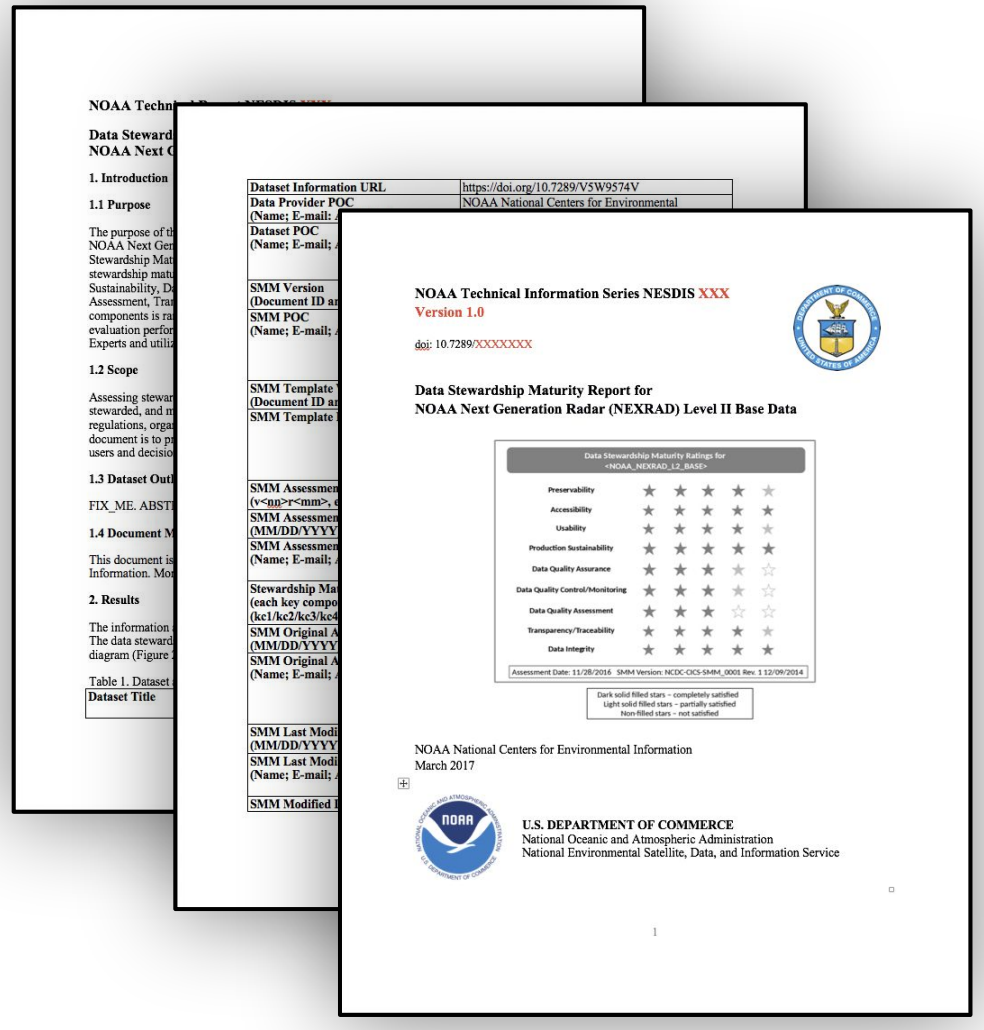
- Each key component and the overall score are automatically incorporated into the ISO Data Quality section by CoMET including a reference to the DSMM journal article, the DSMM template and to the DSMR
- Used by NOAA OneStop for search relevancy ranking. DSMM score can break ties between similar search results

```
            <gmd:value>
              <gco:Record>
                <gco:CodeListValue
codeList="https://data.noaa.gov/resources/iso19139/schema/resources/Codelist/MM_StewCodelists.xml#DataQualityAssessment"

codeListValue="minimal">minimal</gco:CodeListValue>
              </gco:Record>
            </gmd:value>
            <gmd:value>
              <gco:Record>
                <gco:CodeListValue
codeList="https://data.noaa.gov/resources/iso19139/schema/resources/Codelist/MM_StewCodelists.xml#Accessibility"

codeListValue="optimal">optimal</gco:CodeListValue>
              </gco:Record>
            </gmd:value>
            <gmd:value>
              <gco:Record>
                <gco:CodeListValue
codeList="https://data.noaa.gov/resources/iso19139/schema/resources/Codelist/MM_StewCodelists.xml#DataQualityControlMonitoring"

codeListValue="intermediate">intermediate</gco:CodeListValue>
              </gco:Record>
            </gmd:value>
```



Data Stewardship Maturity Report (DSMR)

- CoMET will be able to export a NOAA technical report from the DSMQ results including graphics, references, and assessment details
- Collaboration with NOAA Central Library for preservation of reports. DOIs assigned for citability and tracing
- Report cross-linked via DOIs in associated collection metadata



Additional Information and Links

OneStop
https://data.noaa.gov/onestop/#

CoMET/DSMQ Application & User Guide
https://data.noaa.gov/cedit

DSMM Resources
https://tinyurl.com/DSMMintro
https://tinyurl.com/DSMMpaper
https://tinyurl.com/DSMMtemplate

THANKS!

Questions?

Feel free to reach out…

nancy.ritchey@noaa.gov

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

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