CEOS

Working Group on Information Systems and Services
Data Stewardship Interest Group

WGISS Data Management and Stewardship Maturity Matrix Version 1.0

Doc. Ref.:   CEOS.WGISS.DSIG.DSMM
Date: 26 September 2017
Issue: 1.0

INTENDED AUDIENCE

This document is intended to assist data managers in Earth Observation (EO) data centres in the task of ensuring Earth Observation space data sets long-term preservation, curation, accessibility, discoverability and usability.

The intended audience comprises:
1. Data providers
   - To evaluate and improve the quality and usability of their products
2. Modellers, decision-makers, and scientists
   - To improve their products
   - To make investments and take decision
3. Data managers/stewards of data centres and repositories
   - To validate their compliance or lack of stewardship practice or standards
   - To assess the current state of their data holdings and repositories
   - To create a roadmap to improve or enhance the stewardship maturity of practices applied to all data holdings

BACKGROUND

This white paper presents a WGISS Data Management and Stewardship maturity assessment model in the form of a matrix for Earth Observation datasets.

In the extended environment of Maturity Matrices and Models, the Maturity Matrix for "Long-Term Scientific Data Stewardship", of Peng, and Jeffrey L. Privette, & Others (2015) [9], represents a systematic assessment model for measuring the status of individual datasets. In general, it provides information on all aspects of the data records, including all activities needed to preserve and improve the information content, quality, accessibility, and usability of data and metadata. This was used as a starting point of the WGISS Data Management and Stewardship Maturity Matrix, presented in this document.

In parallel, the GEO Data Management Principles Task Force was tasked with defining a common set of GEOSS Data Management Principles (DMP-IG) [8]. These principles address the need for discovery, accessibility, usability, preservation, and curation of the resources made available through GEOSS.

The content of the WGISS Data Management and Stewardship Maturity Matrix represents the result of a combined analysis performed on the DMP-IG and a consultation at European level, with the Long Term Data Preservation Working Group.

SCOPE

The scope of the WGISS Data Management and Stewardship Maturity Matrix is in line with the following formal definitions:

- Data stewardship "encompasses all activities that preserve and improve the information content, accessibility, and usability of data and metadata" (National Research Council 2007);
- Data management "includes all activities for planning, execution and oversight of policies, practices and projects that acquire, control, protect, deliver and enhance the value of data and information assets." (Mosely et al. 2009).

In addition to data preservation, accessibility, and usability, this Maturity Matrix also covers discoverability, use and services capabilities, program budget, data policy, etc.

The main objective is to measure and verify the overall implemented or to be implemented preservation lifecycle. It can be used to create a stewardship maturity scoreboard of dataset(s), identify potential gaps, and define a roadmap forward for scientific data stewardship improvement; or to provide data quality and usability information to users, stakeholders, and decision makers.

It is flexible and adaptable through a tailoring, with respect to the requirements and objectives of the data owners, highlighted after the initial process foreseen in the Preservation Workflow [3].

RATIONALE

The rationales for applying the WGISS Data Management and Stewardship Maturity Matrix are:

- ✓ Providing data quality, usability information to users, stakeholders, and decision makers;
- ✓ Providing a reference model for stewardship planning and resource allocation;
- ✓ Allowing the creation of a roadmap for scientific data stewardship improvement;
- ✓ Providing detailed guidelines and recommendations for preservation;
- ✓ Evaluating if the preservation follows best practices;
- ✓ Giving a technical evaluation of the level of preservation and helping with self-assessment of preservation;
- ✓ Providing a status of the preservation, but doesn't offer information on numbers or averages related to preservation;

✓ Helping to break down problems related to preservation, and to understand the costs associated with each preservation level;
✓ Funding agencies can define certain goal levels that they would like to reach.

It is a self-assessment and it is applied at dataset level.

EARTH OBSERVATION DOMAIN APPLICABILITY

This Maturity Matrix may be adopted to facilitate and improve CEOS WGISS Data Management and Stewardship Maturity Matrix activities and achievements. It needs to be adapted to take into account specific Earth Observation requirements and already existing Best Practices. A dataset appraisal activity should define the desired level to be reached for each maturity matrix component, for a specific mission/dataset, for example based on:

✓ Mapping versus final user exploitation capabilities;
✓ Mapping w.r.t. data preservation commitments, budgets, responsibilities and preservation requirements.

Different missions' datasets can have different targets and varying maturity level ratings.

COMPONENTS AND SCALING

In this document, the WGISS Data Management and Stewardship Maturity Matrix is related with the Data Management Principles-Implementation Guidelines [8]. The relevant DMP components are:

DMP Implementation Guidelines	Stewardship Maturity Matrix Key Components
DMP-1: Discoverability	Accessibility
DMP-2: Accessibility	Accessibility
DMP-3: Encoding	Usability
DMP-4: Documentation	Usability
DMP-5: Traceability	Transparency /Traceability
DMP-6: Quality	Data Quality Control/ Monitoring Data Quality Assessment Data Quality Assurance
DMP-7: Preservation	Preservability
DMP-8: Verification	Data Integrity
DMP-9: Review and reprocessing	Production Sustainability
DMP-10: Persistent and resolvable identifiers	Transparency /Traceability

The DMP Implementation Guidelines were analysed in order to align the principles to the maturity matrix format. A weight was given to each component of the principles, in order to create an incremental order for the implementation of the preservation principles.

The maturity of data management and stewardship for each component, applied to individual datasets, can be assessed on a four-level maturity scale. The four levels are associated with measurable data management practices.

- Level-0:   Not Managed
- Level-1:   Limit Managed
- Level-2:   Managed
- Level-3:   Well Managed

COMPONENTS AND ASSOCOIATED MATURITY LEVELS

DISCOVERABILITY

DMP-1: METADATA FOR DISCOVERY

Data and all associated metadata will be discoverable, through catalogues, search engines. Data access and use conditions, including licenses, will be clearly indicated.

Level-0	1) No catalogue available 2) No advertising available
Level-1	1) Advertising available. 2) Catalogue search available at product level with minimum set of metadata.
Level-2	1) Detailed catalogue search available at product level. 2) Product metadata oriented towards an international standard (e.g. ISO, OGC, INSPIRE, etc.) 3) Data Records Collection and Associated Knowledge searchable [3], [4]. 4) Collection metadata oriented towards an international standard (e.g. ISO, OGC, INSPIRE, etc.)
Level-3	1) Product metadata fully compliant with an international standard (e.g. ISO, OGC, INSPIRE, etc.) 2) Collection metadata fully compliant with an international standard (e.g. ISO, OGC, INSPIRE, etc.) 3) Catalogue accessible via an accepted international or community agreed upon standards protocol. 4) Data policy on the use conditions/restrictions and legal constraints of the data, available in metadata. 5) Periodic updates of metadata in the catalogue (e.g. contact point). 6) Quality indicator metadata available and discoverable. 7) Search results ordered by relevancy. 8) Seamless transition from discovery to access.

ACCESSIBILITY

DMP-2: ONLINE ACCESS

Data will be accessible via online services, including, at a minimum, direct download, but preferably user customizable services for access, visualization and analysis.

Level-0	Data are not accessible online.
Level-1	Basic online services available for data access (e.g. FTP/HTTP direct download).
Level-2	1) Simple Access Architecture through metadata - e.g. Data Access through a catalogue service. 2) Data access system oriented towards an international standard (e.g. OpenSearch, ISO).

Level-3	1) Data access system fully compliant with an international standard (e.g. OpenSearch, ISO). 2) Data policy regarding use conditions and restrictions of the data, available in the metadata. 3) Visualisation services allowing a user to view images of data (e.g. Web Map Services for geospatial data, browse image services). 4) Reporting system available (e.g. user statistics, data access reports, system availability reports, etc.). 5) Hosted processing (e.g. on the fly processing). 6) Quick adoption to new technologies and standards evolution.

USABILITY

DMP-3: DATA ENCODING

Data should be structured using encodings that are widely accepted in the target user community and aligned with organizational needs and observing methods, with preference given to non-proprietary international standards.

Level-0	Data Not Structured.
Level-1	Basic schema for automated data use.
Level-2	1) Use of non-proprietary international standards encodings for syntactic interoperability. If a proprietary format is used, it has to be formally and semantically described. 2) Periodically repackaging/ reformatting of archived data.
Level-3	Accepted and Available semantic encoding standards for complete interoperability.

DMP-4: DATA DOCUMENTATION

Data will be comprehensively documented, including all elements necessary to access, use, understand, and process, preferably via formal structured metadata, based on international or community approved standards. To the extent possible, data will also be described in peer-reviewed publications and referenced in the metadata record.

Level-0	Partial and incomplete mission documentation (satellite, data and product documentation).
Level-1	1) Already existent mission documentation available and preserved for the long term. 2) No link between mission documentation and data records.
Level-2	1) Documentation produced, published and well described (covering the format, metadata, and methods used in creating and validating the data). 2) Link between mission documentation and data records created and managed (internal use only).
Level-3	1) Standards based metadata for documentation (e.g. to support the reproducibility of science). 2) Link between mission documentation and data records

	published.

DMP-5: DATA TRACEABILITY

Data will include provenance metadata, indicating the origin and processing history of raw observations and derived products, to ensure full traceability of the product chain.

Level-0	Limited product information available (not online).
Level-1	Product information available (not online).
Level-2	Dataset tested for presence of correct provenance metadata (presence, completeness and correctness). Well described product information available online.
Level-3	1) Automatic metadata generation for provenance documentation. 2) Complete and updated data provenance available online.

DMP-6: DATA QUALITY-CONTROL

Data will be quality-controlled and the results of quality control shall be indicated in the metadata; data made available in advance of quality control will be flagged in metadata as unchecked.

Level-0	1) No control and monitoring check. 2) No quality indicator in metadata. 3) No procedures documentation.
Level-1	1) Basic data quality control and monitoring check. 2) Minimal set of quality control procedures documented and available.
Level-2	1) Quality indicator post-processing available. 2) Quality control procedures documented and available online.
Level-3	1) Data quality control fully compliant with an international standard 2) Quality indicator pre and post processing available in the metadata [6]. 3) Quality metadata assessed.

PRESERVATION

DMP-7: DATA PRESERVATION

Data will be protected from loss and preserved for future use; preservation planning will be for the long term, and include guidelines for loss prevention, retention schedules, and disposal or transfer procedures.

Level-0	1) Uncontrolled storage location. 2) Only data are stored. 3) Data Records archiving not managed.
Level-1	1) Basic archiving for original data records preservation: - The entity in charge of data long term preservation is identified and designated;

	- Minimal redundancy and metadata preservation; 2) Assessment of SW preservation.
Level-2	1) Preservation repository certified internally: - Documented storage procedures (planning of periodic media refreshment) - Redundancy managed (e.g. back-up, different media technology) - Basic archiving processes measured and controlled 2) Community-standard for archiving metadata (e.g. AIP).
Level-3	1) Preservation repository officially certified (e.g. ISO 19363, CoreTrustSeal) 2) Periodic technology refreshment 3) Identify and manage the basic preservation of relevant mission SW, ensuring that preserved data can be recreated. 4) Continuity of service availability (Business Continuity, Disaster and Recovery, etc.).

DMP-8: DATA AND METADATA VERIFICATION

Data and associated metadata held in data management systems will be periodically verified to ensure integrity, authenticity and readability.

Level-0	No Data/Associated Knowledge integrity, authenticity and readability check.
Level-1	Data Records/Associated Knowledge integrity basic check (e.g. checksum).
Level-2	1) Data Records/Associated Knowledge content integrity check and verification. 2) Media readability and accessibility testing.
Level-3	1) Automatic Data Records/Associated Knowledge content integrity check and verification. 2) Data authenticity verifiable internally and by the final user. 3) Automatic verification process, including monitoring and reporting.

CURATION

DMP-9: DATA REVIEW AND REPROCESSING

Data will be managed to perform corrections and updates in accordance with reviews, and to enable reprocessing as appropriate; where applicable, this shall follow established and agreed procedures.

Level-0	No reprocessing activities planned.
Level-1	1) Minor updates and bugs corrections of data records implemented. 2) Data Records repackaging and/or reformatting.
Level-2	Reprocessing for calibration and/or algorithm improvement.

Level-3	1) Reprocessing for time-series creation (e.g. FDR for ECV). 2) Roadmap for technology evolution.

DMP-10: PERSISTENT AND RESOLVABLE IDENTIFIERS

Data will be assigned appropriate persistent, unique and resolvable identifiers to enable documents to cite the data used and to enable data providers to receive acknowledgement for the use of their data.

Level-0	No persistent and resolvable identifiers available
Level-1	1) Persistent identifier assignment only for particular Data Records Collections. 2) Basic landing pages management (e.g. manual generation and updates, no common template).
Level-2	1) Persistent identifier assignment to all disseminated Data Records Collections. 2) Automatic landing page generation and extensive management of landing pages.
Level-3	Persistent identifier [7] created for all accessible data records.

WGISS Data Management and Stewardship Maturity Matrix

The complete WGISS Data Management and Stewardship Maturity Matrix is presented in the picture below:

	Discoverability	Accessibillity	Usability				Preservation		Curation	
	DMP-1 **Metadata for Discovery**	**DMP-2** **Online Access**	**DMP-3** **Data Encoding**	**DMP-4** **Data Documentation**	**DMP-5** **Traceability**	**DMP-6** **Quality**	**DMP-7** **Preservation**	**DMP-8** **Verification**	**DMP-9** **Reprocessing**	**DMP-10** **Persistent Identifier**
Level-0 Not Managed	1) No catalogue available 2) No advertising available	Data are not accessible online	Data Not Structured	Partial and incomplete mission documentation (mission, data and product documentation)	Limited product information available (not online)	1) No control and monitoring check 2) No quality indicator in metadata 3) No procedures documentation	1) Uncontrolled storage location. 2) Only data are stored 3) Data Records archiving not managed	No Data/Associated Knowledge integrity, authenticity and readability check	No reprocessing activities planned	No persistent and resolvable identifiers available
Level-1 Limit Managed	1) Advertising available 2) Catalogue search available at product level with minimum set of metadata	Basic online services available for data access (e.g. FTP/HTTP direct download)	Basic schema for automated data use	1) Already existent mission documentation available and preserved for the long term 2) No link between mission documentation and data records	Product information available (not online)	1) Basic data quality control and monitoring check 2) Minimal set of quality control procedures documented and available	1) Basic archiving for original data records preservation: - The entity in charge of data long term preservation is identified and designated - Minimal redundancy and metadata preservation 2) Assessment of SW preservation	Data Records/Associated Knowledge integrity basic check (e.g. checksum)	1) Minor updates and bugs corrections of data records implemented 2) Data Records repackaging and/or reformatting	1) Persistent identifier assignment only for particular Data Records Collections 2) Basic landing pages management (e.g. manual generation and updates, no common template)
Level-2 Managed	1) Detailed catalogue search available at product level 2) Product metadata oriented towards an international standard (e.g. ISO, OGC, INSPIRE, etc.) 3) Data Records Collection and Associated Knowledge searchable. 4) Collection metadata oriented towards an international standard (e.g. ISO, OGC, INSIPIRE, etc.)	1) Simple Access Architecture through metadata - e.g. Data Access through a catalogue service 2) Data access system oriented towards an international standard (e.g. OpenSearch, ISO)	1) Use of non-proprietary international standards encodings for syntactic interoperability. If a proprietary format is used, it has to be formally and semantically described. 2) Periodically repackaging/ reformatting of archived data.	1) Documentation produced, published and well described (covering the format, metadata, and methods used in creating and validating the data) 2) Link between mission documentation and data records created and managed (internal use only)	Dataset tested for presence of correct provenance metadata (presence, completeness and correctness). Well described product information available online	1) Quality indicator post-processing available 2) Quality control procedures documented and available online	1) Preservation repository certified internally: - Documented storage procedures (planning of periodic media refreshment) - Redundancy managed (e.g. back-up, different media technology) - Basic archiving processes measured and controlled 2) Community-standard for archiving metadata (e.g. AIP)	1) Data Records/Associated Knowledge content integrity check and verification 2) Media readability and accessibility testing	Reprocessing for calibration and/or algorithm improvement	1) Persistent identifier assignment to all disseminated Data Records Collections 2) Automatic landing page generation and extensive management of landing pages
Level-3 Well Managed	1) Product metadata fully compliant with an international standard (e.g. ISO, OGC, INSPIRE, etc.) 2) Collection metadata fully compliant with an international standard (e.g. ISO, OGC, INSPIRE, etc.) 3) Catalogue accessible via an accepted international or community agreed upon standards protocol 4) Data policy on the use conditions/restrictions and legal constraints of the data, available in metadata 5) Periodic updates of metadata in the catalogue (e.g. contact point) 6) Quality indicator metadata available and discoverable 7) Search results ordered by relevancy. 8) Seamless transition from discovery to access	1) Data access system fully compliant with an international standard (e.g. OpenSearch, ISO) 2) Data policy regarding use conditions and restrictions of the data, available in the metadata. 3) Visualisation services allowing a user to view images of data (e.g. Web Map Services for geospatial data, browse image services) 4) Reporting system available (e.g. user statistics, data access reports, system availability reports, etc.) 5) Hosted processing (e.g. on the fly processing) 6) Quick adoption to new technologies and standards evolution	Accepted and Available semantic encoding standards for complete interoperability	1) Standards based metadata for documentation (e.g. to support the reproducibility of science) 2) Link between mission documentation and data records published	1) Automatic metadata generation for provenance documentation 2) Complete and updated data provenance available online	1) Data quality control fully compliant with an international standard 2) Quality indicator pre and post processing available in the metadata 3) Quality metadata assessed	1) Preservation repository officially certified (e.g. ISO 19363, CoreTrustSeal) 2) Periodic technology refreshment 3) Identify and manage the basic preservation of relevant mission SW, ensuring that preserved data can be recreated. 4) Continuity of service availability (Business Continuity, Disaster and Recovery, etc.)	1) Automatic Data Records/Associated Knowledge content integrity check and verification 2) Data authenticity verifiable internally and by the final user 3) Automatic verification process, including monitoring and reporting	1) Reprocessing for time-series creation (e.g FDR for ECV) 2) Roadmap for technology evolution	Persistent identifier created for all accessible data records

Reference Document

[1] CEOS, "**EO Data Preservation Guidelines Best Practices**",
http://ceos.org/document_management/Working_Groups/WGISS/Interest_Groups/Data_Stewardship/Recommendations/EO%2Data%20Preservation%20Guidelines_v1.0.pdf

[2] CEOS, "**EO Preserved Data Set Content**",
http://ceos.org/document_management/Working_Groups/WGISS/Interest_Groups/Data_Stewardship/Recommendations/EO%20Preserved%20Data%20Set%20Content_v1.0.pdf

[3] CEOS, "**Long Term Preservation of Earth Observation Space Data: Preservation Workflow**",
http://ceos.org/document_management/Working_Groups/WGISS/Interest_Groups/Data_Stewardship/Best_Practices/Preservation%20Workflow_v1.0.pdf

[4] CEOS, "**Associated Knowledge Preservation Best Practices**",
http://ceos.org/document_management/Working_Groups/WGISS/Documents/WGISS%20Best%20Practices/CEOS%20Associated%20Knowledge%20Preservation%20Best%20Practices_v1.0.pdf

[5] CEOS, "**Generic Earth Observation Data Set Consolidation Process**",
http://ceos.org/document_management/Working_Groups/WGISS/Interest_Groups/Data_Stewardship/Best_Practices/GenericEarthObservationDataSetConsolidationProcess_v1.0.pdf

[6] CEOS, "**Long-Term Preservation of Earth Observation Space Data: Glossary of Acronyms and Terms**",
http://ceos.org/document_management/Working_Groups/WGISS/Interest_Groups/Data_Stewardship/White_Papers/EO-DataStewardshipGlossary_v1.2.pdf

[7] CEOS, "**CEOS Persistent Identifier Best Practices**",
http://ceos.org/document_management/Working_Groups/WGISS/Documents/WGISS%20Best%20Practices/CEOS%20Persistent%20Identifier%20Best%20Practices_v1.2.pdf

[8] GEOSS, "**Data Management Principles**",
https://www.earthobservations.org/documents/dswg/201504_data_management_principles_long_final.pdf

[9] Peng, Privette, & Others, **Scientific Data Stewardship Maturity Matrix**

1. https://datascience.codata.org/articles/abstract/10.2481/dsj.14-049/
2. http://tinyurl.com/DSMMtemplate
3. http://www.slideshare.net/gepeng86/scientific-data-stewardship-maturity-matrix
4. http://tinyurl.com/DSMM-FlowChart