Data Management Principles Implementation Guidelines

For information.

Table of Contents

1 INTRODUCTION

The GEO Data Management Principles Task Force was created by GEO and tasked with defining a common set GEOSS Data Management Principles. These principles address the need for discovery, accessibility, usability, preservation, and curation of the resources made available through GEOSS. To support the implementation of the principles the Task Force has drafted the following guidelines that data providers and others can use as they seek to implement the principles, and to provide a basis for assessing how well the principles are being adhered to in practice.

Each Data Management Principle (DMP) has an associated implementation guideline for data providers to follow. The following topics are covered in the guidelines for each DMP:

- Terms used to describe the principle and its implementation;
- Explanation of the DMP;
- Guidance on implementation with examples;
- Metrics to measure the level of adherence to the DMP; and
- Resource implications of implementing the DMP.

A compilation of terms appears in Appendix A. A list of references appears in Appendix B.

This paper is intended to be a living document that will evolve over time as more is learned about implementing the GEOSS DMPs.

2 DMP-1: METADATA FOR DISCOVERY

DMP Category: Discovery.

DMP-1: Data and all associated metadata will be discoverable, through catalogues and search engines, and data access and use conditions, including licenses, will be clearly indicated.

2.1 Terms

The following terms, as they relate to this guideline, are defined in Appendix A:

Broker, Catalogue, Clearinghouse, Core Elements, Discovery, Discovery Services, License, Metadata, Metadata Element, Network Services, Queryable, Search Engine, Use Conditions

2.2 Explanation of the principle

A visitor to a library should be able to find a desired book without having to look at every book in the library. The library's catalogue allows the visitor to search information about the books (e.g. author, ISBN number, genre), to discover where to find the book and under what conditions or restrictions the book might be read or borrowed. This "information about the book" is its metadata. Likewise, a user looking for Earth Observation resources (data, web services, models, etc.) should be able to find what they want by searching the metadata associated with that resource, including how the resource can be accessed and whether there are restrictions or conditions placed on its use. GEOSS maintains a catalogue of resources and like a library catalogue it does not keep copies of resources but instead manages the metadata that allows users to locate and access the resources.

Not all users begin a search for resources by going to a catalogue. Instead they may use general purpose search engines. For this reason catalogues may have portals, such as the geoportal, for use by humans, as well as programmatic interfaces (APIs) meant for access by search engines, metadata harvesters and the portals of other communities.

2.3 Guidance on Implementation, with Examples

The following types of metadata are particularly important for discoverability and reuse:

- a descriptive title and abstract;
- identity and contact information (e.g. ORCIDs) for the individuals responsible for the creation of the resource;
- identity and contact information for the individuals responsible for the management of the resource;
- geographic location or boundaries;
- temporal coverage;
- keywords describing the resource and the scientific or practical domain to which it applies;
- information on conditions and restrictions on use, in particular license information; and
- web links to the resource and to further information about the resource.

The following guidelines will help ensure that data and services are discoverable and usable. Adherence to these guidelines is checked and assessed through certification of the data repository or service using baseline certifications such as Data Seal of Approval or World Data System certifications.

- catalogue entries should be in accordance with an accepted international or community agreed upon standards (e.g. DataCite, Dublin Core, ISO19115, etc.), and all core elements of the standard should be completed;
- The catalogue should be accessible via an accepted international or community agreed upon standard protocol (e.g. OAI-PMH, OpenSearch, OGC CSW, etc.);
- The metadata kept in the catalogue should be periodically checked for validity and links to the resources are still valid and responding. If metadata are maintained in the catalogue for resources that no longer exist, a mechanism should be provided to point to updated versions, if any, or suitable explanations be provided for why resources no longer exist;
- The catalogue should provide search capabilities and the search results should display in a relevance-ranked order to reflect the user's query;
- GEOSS Data/Resource Providers are encouraged to register Catalogues over individual resources, where multiple resources are to be made discoverable;
- As an alternative to creating a catalogue with a search interface, a data provider may post metadata, with links to the associated data, in a web-accessible location, which can then be harvested by search engines or metadata aggregators; and
- Since some resources may have restrictions or other conditions of use, these should be clearly indicated in the metadata. Examples include limits on distribution, intended use, as well as licenses.

See also Data Management Principle 4 'Data Documentation', which gives additional guidelines regarding documentation for that allows data to be used, understood and processed.

2.4 Metrics to measure level of adherence to the principle

Appropriate metrics relate to: 1) whether the metadata provides appropriate information for discovery and about reuse conditions; 2) whether the system providing the catalogue information follows good practice in terms of standards and performance; and, 3) whether the repository is certified as a trusted digital repository.

There are many projects and components that contribute to the implementation and measuring of metrics. Some examples of these include:

- Projects:
  - The GeoViQua project utilizes metadata quality indicators.
- Service checkers:
  - FGDC Service status checker;
  - JRC Service checker.
- Performance indicators and availability:
  - A catalogue is a system or service with high availability (e.g. > 99%), and should be engineered so that:
    - no single points of failure;
    - reliable cross over;
    - detection of failures as they occur.
  - Communities indicate the need for tools for validation (metadata, service, data – resources):
    - Some tools interpret standards differently;
    - Compliant resources should have undergone the certification process;
    - Need for reference implementations and consistent, widely publicized and well-known community-accepted implementation guidelines.
  - There should be a mechanism for data users to supply feedback as to the level of metadata adoption. In many cases, this is best known by the data user, and can serve as a qualitative metric.

2.5 Resource Implications of Implementation

The following activities are Key resource consumers include the activities of metadata authoring and maintenance, and standing up and maintaining a catalogue service. Examples of cost estimates to cover these activities have been made by many data management organizations, such as the Italian National research Council (CNR) and various EC member states. In particular, CNR have made cost estimates for operating the GEO Discovery and Access Broker (DAB), as well as other EC member states having cost estimates to operate a Spatial Data Infrastructure (SDI).

3 DMP-2: ONLINE ACCESS

DMP Category: Accessibility

DMP-2: Data will be accessible via online services, including, at a minimum, direct download but preferably user-customizable services for access, visualization and analysis.

3.1 Terms

The following terms, as they relate to this guideline, are defined in Appendix A:

Authentication/Authorization, Online, SSO (Single Sign-On)

3.2 Explanation of the principle

The storage and distribution of data has evolved dramatically in recent decades. These developments include the vast increases in the availability of data online and the speed of transfer, as well as the ability to run queries over numerous datasets using Application Program Interfaces (APIs). Users now expect data to be available on demand, via online services, i.e. a web address. Currently, this means a URL responding to HTTP, HTTPS, or FTP based protocols.

To meet a wide variety of use cases, particularly analysis at scale, users expect data to be usable by a human via a user interface (providing at least download but also tools for visualisation and analysis) and to be 'machine-usable' via an API.

There are several types of online services. A few of these are:

- Direct access service, allowing the user to download data to their computer;
- Direct Web service, allowing a machine to download a large number of files;
- Browse services, which allow users to inspect representations of candidate files before ordering;
- Visualisation services allowing a user to view images of data and possibly to superpose it on other data. For geospatial data this would typically be via a Web Map Service (e.g. OGC WMS / WMTS);
- In place processing of the data:
    - Since the volume of data is increasing dramatically, it is desirable to perform processing and analysis of the data in place, i.e. before downloading the source data;
    - The OGC WPS provides a standardized way to remotely execute processing.;
    - In order to ease the transfer of the processors, some techniques can be used: e.g. virtualization, or docker techniques.

3.3 Guidance on Implementation, with Examples

1. **Simple architecture**: The data access architecture should be simple to implement
2. **Use of standards**: The data access system should rely on standards. Examples of standards are :
    - HTML
    - OGC standards,
    - OPeNDAP,
    - CEOS OpenSearch,

3. **Archived data repackaging/reformatting:** Data should be provided in the standard formats that are needed by the designated communities and in exchange formats to facilitate interchange between archives.

In order to ease the work of the user, the URL for accessing the data should be present within the metadata provided by the catalogue service. The use of a standardized interface (like OPeNDAP OpenSearch, OGC, etc.) is preferred. This allows the use of existing tools and also helps resources to be more widely used.

Many data repositories require knowledge of the identity of those requesting data. For this reason it is desirable to enable automatic user authentication and authorization. SSO is recommended to open data more widely and ease use. As several SSO protocols exist, a common protocol or a federation of interoperable protocols is recommended.

3.4 Metrics to measure level of adherence to the principle:

Online data accessibility using a standard browser or web service indicates adherence.

3.5 Resource Implications of Implementation

Simple data accessibility can be accomplished with minimal cost using freely available resources. Costs increase when providing and maintaining access to additional tools, services, and related information.

4 DMP-3: DATA ENCODING

DMP Category: Usability

DMP-3: 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.

4.1 Terms

The following terms, as they relate to this guideline, are defined in Appendix A:

4.2 Explanation of the Principle

Usability of data, and especially automated use, depends strongly on the extent to which end users (both human and machine) can rely on standardized encoding as tools, applications, and algorithms are typically designed to work with such. Use of standardised encodings brings benefits to the end user and limits the amount of time spent on transforming data, and therefore is a key to ***interoperability***.

Complete interoperability needs three conditions to be met (Hugo 2009):

Schematic Interoperability defines the structure (schema) in which the data will be offered by a service. For many applications, this schema is critical for correct binding, but schema are likely to vary within a common framework depending on specific applications.

Syntactic Interoperability: this defines the way in which data services will be invoked (Hugo 2008). In many cases, such standards make provision for query parameters and sub-setting of data sets. OPeNDAP has started working on an additional refinement, in that requests for derived data ("offerings"), for example based on statistical analysis, can also be included into the service syntax. Such concepts, which allow requests for processing to be sent to data, instead of the other way round, is a major requirement in the field of Big Data applications (Fulker and Gallagher 2013). Definition of the parameters depend to some extent on semantic interoperability and conventions.

Semantic Interoperability ensures that the content of the schema (the data itself) can be understood by humans or machines (Heffin and Hendler 2000). It is the most complex of the interoperability requirements, and attempts to establish common ontologies, vocabularies, and frameworks such as "essential variables" (OOPC 2015), are all designed to address semantic interoperability. A subset or refinement of semantic interoperability concerns the protocols or methodologies used to gather the data – sometimes critical for valid collations or combinations. Some frameworks for essential variables in Earth and environmental observation science attempt to provide such protocols and methodologies.

In practice, true semantic interoperability is difficult to achieve, often requiring brokering and mediation to align with a standard. A future consideration is the extent to which it will be possible to persist such mediations for re-use. Agreement on a workable set of syntactic (service), schematic, and semantic standards for the typical data families in use by the community can help in some cases.

4.3 Guidance on Implementation, with Examples

The availability and acceptance of ***syntactic encoding standards*** are at a high level of maturity, and that these standards cover the majority of data families that the GEO community uses routinely. Examples include the map and sensor services defined by OGC SWE (2011), OPeNDAP and NetCDF services (OGC Network Common Data Form 2015; Common Data Model 2015), and the work done by WMO in respect of globally available meteorological data (WMO Information System 2015). The extent to which the community has implemented these standards is, however, highly variable, with implementation of Sensor Observation Services lagging seriously behind Web Mapping Services and the use of OPeNDAP and NetCDF. Practitioners should select the standards and open-source implementations of these appropriate to their data family, internal information technology platforms, and capabilities, as a preferred means of providing access to publicly available data sets.

Communities have also developed a portfolio of content standards in support of ***schematic interoperability***. Examples include the provision of KML (OGC KML 2008), GML (OpenGIS® Geography Markup Language 2007), GeoJSON (GeoJSON Format Specification 2015), and other similar standards for the encoding of spatial data, and the SensorML (OGC® SensorML 2014) suite for encoding of time series and sensor observations. Interoperability in the field of especially spatial data sets, whether these are vector data or raster data sets, is highly mature, and it is common for applications and web components to support a wide variety of data schema. Best practice and guidance should stress the application of these widely adopted standards whenever possible.

The most diverse landscape is found in respect of ***semantic interoperability*** and content standard encoding to support it. Some communities have access to mature content standards (for example the Biodiversity community through TDWG (TDWG Standards 2015), the Climate Modelling community through essential climate variables (GCOS Essential Climate Variable(s) 2015), and WaterML(OGC® WaterML 2015)), and there are significant efforts to establish ontology, vocabularies, and name services for a wide variety of disciplines. A major concern is centered on this diversity, and it is often difficult for implementers and end users to select from the large number of options available. GEO is in a position to address this problem – firstly through creation of definitive registries of resources that are available, and by working towards community consensus on the most appropriate resources to use. In general, best practice in the absence of such guidance will be to use any published vocabularies, ontologies, and name services appropriate to the field of study rather than none at all.

4.4 Metrics to measure level of adherence to the principle

Measuring adherence to a schema offered by a data service depends on the data format (MIME type): in the case of XML encodings, the structure and vocabulary (in other words, both schematic and to some extent semantic interoperability) can be tested against the XSD (XML Schema Document). Other encodings (GeoJSON, text, or binary encodings) do not support such automated validation and have to be explicitly tested.

It will often only be possible to evaluate or test the compliance of a data set and/ or service by submitting such a data set or service to a validation service, but to our knowledge only a few such services exist or are in practical use. OGC makes several test services and suites available (OGC Validator 2007).

4.5 Resource Implications of Implementation

Implementation in the Earth and Environmental observation domain can be aided by the availability of free and open source software and can reduce the cost of deploying standardised data services. Offerings range from spatial databases (PostGres), through data servers (GeoServer, Sensor Observation Services, OPeNDAP) to visualisation tools (Global Imagery Browse Services, OpenLayers).

Implementation requires human resources with experience and knowledge in the domain of interest, spatial data, and computing. There is a growing need for this combination of skills as seen in the emergence of careers in data science. Their contributions range from systems development, configuration, and maintenance to content publication and standardisation. They may also provide assistance with development of vocabularies, name services, and content standards.

In practice, none of these ideal aspects of interoperability are likely to be realised, requiring brokering and mediation. The target of such brokering or mediation can be any of the three types of interoperability. A major consideration is the extent to which it will be possible to persist such mediations for future re-use.

From this, we deduce that a truly interoperable environment can only be realised if communities of practice converge towards a workable set of syntactic (service), schematic, and semantic standards for the typical data families that the community uses, and that brokering and mediation services and definitions are visible and available to practitioners.

5 DMP-4: DATA DOCUMENTATION

DMP Category: Usability

DMP-4: 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 referenced in the metadata record.

5.1 Terms

The following terms, as they relate to this guideline, are defined in Appendix A:

Community-Approved Standards, Documented, International Standards

5.2 Explanation of the Principle

The proper use of metadata for the purpose of data documentation helps ensure that data users can access, use, understand, and process data. Usability of data is maximized when all appropriate elements of metadata are utilized. Partial documentation of data negatively impacts its usability in two main ways. First, one or more aspects of documentation can be handled partially, while others are handled completely and can happen when not all appropriate metadata elements have been populated for a given aspect of documentation. Second, one or more aspects of documentation can be ignored completely, meaning none of the metadata elements have been populated for that aspect of documentation.

The purpose of using formal standards-based metadata for data documentation is to maximize the use and reuse of the metadata across community and disciplinary boundaries. Standards facilitate the sharing of metadata between data providers and data users, either directly or via mediation technology.

When applicable, data producers should publish, in the peer-reviewed literature, the methods used in creating and validating the data. These and other descriptions can assist users in understanding various aspects of the data in ways not easily captured by formal metadata and should reference the data. However, publications are not a substitute for formal metadata, which should reference such works to enable discovery of additional documentation contained in referenced publications.

5.3 Guidance on Implementation, with Examples

Implementation requires populating metadata elements with appropriate content. Formal metadata standards for comprehensive data documentation include, among others, ISO 19115-1 (Standards ISO 2014), ISO 19115-2 (Standards ISO 2009), ISO 19139 (Standards ISO 2007), ISO 19157 (Standards ISO 2013), Dublin Core (Standards ISO-2 2009), Darwin Core, Directory Interchange Format (DIF), and Climate and Forecast (CF) metadata conventions.

Each metadata standard contains a set of suggested elements, or fields, which should be populated to cover three categories of metadata, including Descriptive, Structural, and Administrative metadata. It is the responsibility of the data providers to create and populate the metadata according to the standard used. Data users should have an expectation that, if the standard is followed, the dataset metadata can be read and utilized appropriately.

5.4 Metrics to measure level of adherence to the principle

Measuring consistent adherence to metadata creation and population guidelines can be very problematic. It is relatively easy to determine if the suggested metadata fields have been left empty or populated, but it is much more difficult to determine if populated metadata fields have been populated properly, or in a meaningful way. For example, a metadata field used to point to where the associated data can be found may be populated incorrectly or populated with a link that resolves to a location where access or use of the data may not be possible. The question then becomes whether the link was wrong or the metadata expressing the manner in which the data can be accessed and used is incomplete or wrong. Finally, following the example just mentioned, even if a link to data, and the associated metadata fields that explain how to access it, are populated correctly, it is still possible for the data to be misunderstood if appropriate semantic metadata is not available.

Four levels of metrics should be used to determine adherence to DMP-4:

- Measure the completeness of the suggested metadata fields for the standard used, reporting the percentage of fields meaningfully populated;
- Count the number of metadata references to other sources of documentation that describe the associated data;
- Measure whether links work correctly, reflecting dependencies between metadata fields and information on the accessibility of other documentation; and
- Measure the semantic success of the metadata, indicating the level at which the associated data can be understood and used in a meaningful manner.

5.5 Resource Implications of Implementation

Organizational, administrative, financial, technical, and operational resources are needed to implement the guidelines and the metrics necessary for measuring adherence to DMP-4. Organizational resources include policy formulation to reflect adherence and the value of adherence to the organization. Administrative resources include workflow definitions and review to validate adherence. Financial

resources include budgets for people, software, and hardware for implementation. The hardware costs may be minimal compared to resources for professional development on metadata generation, software creation and maintenance, process improvement, and evaluation. Technical resources include tools and documents to implement the metadata generation, its testing, and adherence metrics. Operational resources include the time and people needed to integrate the metadata generation and adherence metrics into routine processes of the data provider. Tools for capturing metadata are available, both commercially and in open source.

6 DMP-5: DATA TRACEABILITY

DMP Category: Usability

DMP-5: 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.

6.1 Terms

The following terms, as they relate to this guideline, are defined in Appendix A: Provenance, Traceability

6.2 Explanation of the principle

Provenance information is given as part of the metadata to the data. Some provenance information can be captured automatically by the process step tools involved and accumulated in the metadata of the resulting dataset. Ideally, a process step will inherit the provenance of data sources and add information about the current process step. Other elements of provenance can be captured manually, including names of parties that created, updated or maintained the dataset.

Provenance is considered a complement to data quality information. In the absence of quantitative information about the uncertainties of the data, expert users can infer data quality estimations from the uncertainties of the sources and from the confidence in the process steps applied. In addition, the reputation of the responsible party of the dataset sources and the result can be used to increase confidence in, as well as an indication of the uncertainties on, the dataset.

The accessibility of the original data source's metadata and processing algorithm descriptions is also a metric of the usability of the provenance information. If provenance is describing sources and processing tools that are not available (or at least have some available documentation), such information cannot be effective in the end.

Provenance can help users identify a problem in a basic dataset or improve it. Provenance information about other products can help to identify which products were derived from the affected dataset. Provenance can help to recreate (or reproduce) the dataset when the problem in the basic dataset is fixed or an improved version is available. Provenance information can also be used to assess the homogeneity of a dataset series where some members of the series originated from sources with different time extents or different versions of the processing algorithms. Provenance can be provided at different levels such as dataset series, dataset, feature, attribute type, attribute etc. For example, this is useful to determine the source of features of even attribute values in the case that a dataset is the result of merging elements features from different sources. Provenance at the dataset level is usually stored in the dataset metadata (that, in the case of GEOSS, it is accessible by the Discovery and Access Broker) while provenance at the feature and attribute level is usually stored in the dataset itself as additional properties of the feature, requiring data access to get them.

6.3 Guidance on Implementation, with Examples

1. **Automatic metadata creation**: Tools that create and manipulate the data also should produce provenance documentation automatically to avoid losing steps or incorrectly documenting metadata. Tools need to inherit the provenance from previous sources. References to algorithms and versions need to be added.
2. **Provenance metadata presence and completeness**: Datasets should be tested for the presence of metadata about provenance information, which should include a clear sequential description of all sources, processing steps, and responsible parties.
3. **Provenance metadata correctness**: Ensure that data sources are documented using universal identifiers (many times local file names are documented) and ideally pointing to accessible sources, that processing algorithms are well maintained and accessible, and that responsible party information is current and points to an accessible party.
4. **Provenance Visualization**: Provenance information can sometimes be very complex. Tools for interpreting provenance and generating graphs can enhance understanding.

6.4 Metrics to measure level of adherence to the principle

1. Presence of information about data sources, process steps, and responsible parties in the metadata distributed with the data. This can be done by verifying the sources and process steps documented in the lineage model of ISO 19115 and ISO 19115-2" Geographic Information – Metadata" that the Discover and Access Broker provide for each GEOSS resource.
2. The accessibility of the original data source's metadata and processing algorithm descriptions is a metric of the usability of the provenance. For sources, this can be obtained by checking the source URI and finding out if they are available for downloading.

6.5 Resource Implications of Implementation

This is part of the metadata process and the costs can be absorbed in this concept. There are two associated costs:

1. Implementing automatic metadata procedures in the processing tools and processing chain;
2. Complementing the automatic tools with a manual edition and review.

7 DMP-6: DATA QUALITY-CONTROL

DMP Category: Usability

DMP-6: Data will be quality-controlled and the results of quality control shall be indicated in metadata; data made available in advance of quality control will be flagged in metadata as unchecked.

7.1 Terms

The following terms, as they relate to this guideline, are defined in Appendix A: Data Quality Indicator, Quality Control

7.2 Explanation of Principle

The quality-control of data is necessary to enable use of the data, especially by individuals who were not involved in the creation of the data. A data quality review should verify consistency, accuracy, and precision of values, fitness for use, completeness and correctness of documentation, and validity and fullness of metadata (Peer et al., 2014), as well as other aspects of the data. Ideally, the data quality review should be conducted prior to dissemination so that prospective user communities can determine the potential for using the data by consulting the results of the data quality review. Prospective users should be able to easily determine the potential for use for their own purposes by assessing data quality review results recorded in data quality indicators of the metadata that describe the data. The absence of values for data quality indicators in metadata is an indication that a data quality review has not been conducted.

7.3 Guidance on Implementation, with Examples

One or more tiers of data quality review should be completed, either independently or in succession. The review also can be conducted as an internal review, an open review, a blind review, or a double-blind review, depending on community practices. An internal quality review may be officiated by the data producer, either manually or automatically. External open reviews offer opportunities for the research community to review and comment on data quality. Blind or double-blind data quality reviews also may be conducted externally by members of the research community. Ideally, an external party, such as a data center, archive, repository, or publisher will officiate an external review to ensure that it is conducted independently of the data producer. The officiator facilitates the review by providing access to the data, any dependent tools, services, related information, and documentation. They specify the review criteria, recruit reviewers, ensure the integrity of the process, receive commentary, and report the results.

Officiators should enable reviewers to determine the extent to which the data meet each criterion. Besides providing context by describing the profile, purpose, scope, collection period, phenomenon studied, and lineage or provenance, documentation should describe collection methods, processes, each variable measured, instrumentation, meaning of each variable value, any input data, previous versions, reasons for missing values, descriptions of uncertainties, and post-collection processing. Sources of support for data collection should be described as well as any considerations for interpretation or restrictions for collection, storage, transmission, access, or use, including any approvals or licenses received with regard to such conditions or restrictions. Names and affiliations of data producers and contributors should be documented for the review process, except for double-blind reviews.

The data quality review should evaluate the data, in terms of relevant criteria that are applicable to a variety of uses of the potential user community. Data quality indicators should distinguish between the dataset level and the individual file level. In consultation with the community, the data quality review officiator should define each criterion to be used for the review. Archives, data centers, and publishers may consult with their respective community representatives to define the criteria for data quality reviews to be conducted on data acquired for their collections.

7.4 Metrics to measure level of adherence to the principle

The officiator of the data quality review provides capabilities to ensure that the results of and justification for each reviewer's decisions, including area of expertise, are documented to complete the data quality report and determine the score for each data quality indicator. The results should record each reviewer's decisions, the criteria used for the data quality review, a definition for each criterion and the meaning of each value, and the extent to which the data met each criterion within data quality indicators to clearly communicate the results determined for each criterion. The officiator should resolve discrepancies between decisions of individual reviewers for a particular criterion to provide a decisive determination about the quality of the reviewed data. For example, the officiator may request

clarification from individual reviewers or request a review by an additional reviewer to break a tie vote for any particular criterion.

The value of the data quality indicator should be included in the metadata that describe the data along with the definition of the indicator or a reference to the definition. If a data quality review was not conducted prior to metadata creation, the metadata should state that the data quality review was not completed. If a particular criterion was not included in the data quality review, the indicator for that criterion should state that the data quality review was not completed.

7.5 Resource Implications of Implementation

Except for automated reviews, at least two reviewers should be recruited to conduct independent data quality reviews. Each data quality reviewer should possess expertise relevant to the use of the data and their type of use should be recorded. Candidate data quality reviewers must report to the officiator, any potential conflicts of interest prior to accepting a review assignment and recuse themselves from the review process when conflicts exist. Determinations of conflicts of interest should be completed prior to conducting the review.

Each reviewer should be provided with access to the review criteria, the data, documentation, metadata, and any tools or services needed to access or use the data (Callahan, 2015). Associated products, tools, or services should be accessible by the reviewers and described to enable inspection and use. Each reviewer should be provided with capabilities for rendering and inspecting these resources and with instructions to enable unimpeded use of the data and related resources.

8 DMP-7: DATA PRESERVATION

DMP Category: Preservation

DMP-7: 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.

8.1 Terms

The following terms, as they relate to this guideline, are defined in Appendix A: Archive, Digital Migration, Long Term, Long Term Preservation, Succession Plan

8.2 Explanation of the principle

Data are a valuable asset for reuse and underpin the scholarly record. The preservation of data in digital format requires certain actions to be performed: this includes such things as preservation planning, scheduled transformation of file-type to avoid obsolescence and plans for asset transfer in the eventuality that the repository is obliged to close. These actions are detailed in the Reference Model for an Open Archival Information System (OAIS). Repositories which through their mission, organisational setup and business processes are able to fulfill these actions in a sustainable way qualify as Trusted Digital Repositories (TDRs).

A TDR:

- Has an explicit mission to provide access to and preserve data in its domain or in accordance with a stated collection policy;
- Has a continuity plan ensuring ongoing access and preservation of holdings;
- Assumes responsibility for long-term preservation and manages this function in a planned and documented way; and

- The repository enables reuse of the data over time, ensuring that appropriate metadata are available to support the understanding and use of the data.

8.3 Guidance on Implementation, with Examples

Data contributed to GEOSS should be preserved for the long term and protected from loss for future use in trusted digital repositories (TDRs). Each requirement above is accompanied by guidance text—as part of the certification criteria for the Data Seal of Approval, the ICSU World Data System[1] or the joint DSA-WDS Criteria[2] currently under development—to assist GEOSS data contributors to conduct a self-assessment.

The guidance below indicates the types of evidence required to certify the trustworthiness of a data repository.

- TDRs are responsible for stewardship of digital objects, ensuring that they are stored in an appropriate environment for required durations and that the holdings are accessible and available, both currently and in the future. Depositors and users must understand that preservation of, and continued access to, the data is an explicit role of the repository;
- The repository, data depositors, and Designated Community need to understand the level of responsibility required for each deposited item in the repository. The repository must have the legal authority to complete their responsibilities and must document procedures to assure their completion; and
- Repositories must ensure that data can be understood and used effectively into the future despite changes in technology. This Requirement evaluates the measures taken to ensure that data are reusable.

8.4 Metrics to measure level of adherence to the principle

Recommended compliance levels for each of the requirements in the section above:

0 -- Not applicable;

1 -- The repository has not considered this yet;

2 -- The repository has a theoretical concept;

3 -- The repository is in the implementation phase; and

4 -- The guideline has been fully implemented in the repository.

Recommended metrics for the evaluation of a trustworthy data repository:

8.4.1 Mission/Scope

- Explicit statements of the long-term preservation role within the organization's mission, with approval by the governing authority.

8.4.2 Continuity of access

- The level of responsibility undertaken for data holdings, including any guaranteed preservation periods.

[1] WDS Certification criteria and guidance: https://www.icsu-wds.org/services/certification

[2] DSA–WDS Partnership WG Catalogue of Common Requirements: https://goo.gl/WnAau0

- Medium-term (3-5-years) and long-term (> 5 years) plans ensure continued availability and accessibility of the data. Descriptions of contingency plans and responses to rapid changes of circumstance and long-term planning indicate options for relocation or transition of activities to another body or return of data holdings to their owners (i.e., data producers). For example, what will happen in the case of cessation or withdrawal of funding, a planned ending of funding for a time-limited project repository, or a shift of host institution interests?

8.4.3 Organizational infrastructure

- The repository is hosted by a recognized institution (ensuring long-term stability and sustainability) appropriate to its Designated Community; and
- The repository has sufficient funding, including staff resources, IT resources, and a budget for attending meetings when necessary. Ideally this should be for a three- to five-year period.

8.4.4 Appraisal

- What is the repository's approach if the metadata provided are insufficient for long-term preservation?

8.4.5 Documented storage procedures

- How is data storage addressed by the preservation policy?
- Does the repository have a strategy for redundant copies? If so, what is it?
- Are data recovery provisions in place? What are they?
- Are risk management techniques used to inform the strategy?
- What checks are in place to ensure consistency across archival copies?
- How is deterioration of storage media handled and monitored?

8.4.6 Preservation plan

- Is the 'preservation level' for each item understood? How is this defined?
- Does the contract between depositor and repository provide for all actions necessary to meet the responsibilities?
- Is the transfer of custody and responsibility handover clear to the depositor and repository?
- Does the repository have the rights to copy, transform, and store the items, as well as provide access to them?
- Is a preservation plan in place?
- Are actions relevant to preservation specified in documentation, including custody transfer, submission information standards, and archival information standards?
- Are there measures to ensure these actions are taken?

8.4.7 Data reuse

- Are plans related to future migrations in place?
- How does the repository ensure understandability of the data?

8.5 Resource Implications of Implementation

The Common Requirements described above reflect the basic characteristics of trustworthy repositories based on the Catalogue of Common Requirements developed by the DSA-WDS Partnership Working Group on Repository Audit and Certification, a Working Group (WG) of the Research Data Alliance. Their goal is to create a set of harmonized common criteria for certification of repositories at the **basic level**, drawing from the requirements already put in place by the Data Seal of Approval (DSA) and the ICSU World Data System (WDS). The ultimate aim is to build a global framework for repository certification that moves from the basic level to the extended level (nestor-SEAL DIN 31644) to the formal (ISO 16363) level.

As should be expected of a comprehensive accreditation process, providing sufficient evidence is somewhat involved and the amount of time and effort needed for the self-assessment depends on the level of maturity of the repository. Entities with existing business process and records management procedures or experience with audits or certifications should spend less time preparing the self-assessment. In general, while very well-prepared repositories may only need a few person-days to complete the assessment, the process usually takes two weeks to three months.

Several individuals may need to contribute to the assessment, which can require discussion with other data management and technical experts in the organization. Thus, it is difficult to estimate resource requirements for the self-assessment phase.

9 DMP-8: DATA AND METADATA VERIFICATION

DMP Category: Preservation

DMP-8: Data and associated metadata held in data management systems will be periodically verified to ensure integrity, authenticity and readability.

9.1 Terms

The following terms, as they relate to this guideline, are defined in Appendix A:

Authenticity, Integrity, Readability

9.2 Explanation of the principle

Important among the actions performed by TDRs described above in DMP-7, is periodic checking and transformation (file migration) of data to ensure that they do not become obsolete. Constant and careful maintenance of the preserved data sets (data and associated knowledge) is necessary to ensure data integrity, authenticity, readability and thus usability over the long term. Archive and Data Management Systems' curation and maintenance consist of all the activities aimed at guaranteeing the integrity, authenticity and readability of the archived and preserved data. This covers the storage of equipment, media and hard disk arrays in secured and environmentally controlled rooms, and a set of defined activities to be performed on routine basis, such as migration to new systems and media, in accordance with the technology and consumer market evolution, data compacting and data format/packaging conversion. Data holders and archive owners need to design a maintenance scheme for their Archives and Data Management System to guarantee the integrity of the archived and collected data.

9.3 Guidance on Implementation, with Examples

1. **Archived data refreshment**: Periodically perform a migration of the archived data ("media refreshment") to the most adequate proven technology for data storage, to ensure data access preservation. Technology selection should not only be based on

technical and cost aspects, but should also aim at the minimization of environmental impact (e.g. in terms of power consumption, thermal dissipation, etc.);

2. **Archived data formats description**: Provide formal description of old archiving formats to allow the conversion to new standard formats, which will increase technical compatibility and reduce diversity of formats and interfaces between archives;
3. **Archived data duplication**: Maintain identical copies of all archived data applying one of the security levels defined below:
    a. Dual copy in the same geographical location (but different buildings) to avoid data loss due to media degradation or obsolescence, or
    b. Dual copy in the same geographical location (but different buildings) based on different technology to avoid technology based principle failures, or
    c. Dual copy in two different geographical locations to safeguard the archive from external hazards (e.g. floods, other natural and technological hazards, etc.), or
    d. Dual copy in two different geographical locations, based on different technologies to avoid technology based principle failures.
4. **Archive system components migration (hardware):** Perform periodical migration of archive system components to new hardware platforms.
5. **Media readability and accessibility tests:** Perform periodical test for media readability and accessibility on a representative set of the archived data.
6. **Archive content integrity:** Periodically verify the integrity of the archive collection/content through integrity check on a representative set of the archived data.
7. **Data content integrity:** Ensure that archived content and associated information remains unchanged and, if changes are made, that these are documented, and that this documentation is preserved and made available as well (provenance information).

9.4 Metrics to measure level of adherence to the principle:

Measures for the level of adherence include the Data Preservation Guidelines in point C above or to **ISO 16363:2012 - Space data and information transfer systems - Audit and certification of trustworthy digital repositories** (CCSDS 652.0-M-1), the standard used to assess the trustworthiness of a generic digital repository.

9.5 Resource Implications of Implementation

Estimating the cost in terms of resources for long-term digital preservation has received much attention from many organisations (e.g. companies, digital libraries, research data centres) interested in preserving their data and depends on the organization and on the data to be preserved (e.g. volume, format, etc.) and can therefore only be modelled here. Cost modelling techniques are used to estimate the costs involved in digital asset preservation and their economic impact on the organisation. Generic Cost models follow two main steps:

1. *Identifying resource costs and activities;*
2. *Assigning resource costs to activities and Assigning activity costs to cost objects*.

1. *Identifying resource costs and activities*

Activities identified for the Archiving process include managing storage, refreshment, migration, reporting, back-up, reformatting/repackaging, test and integrity verification, and reporting on archived data formats. Resources needed to complete the cost analysis include human resources and equipment, office/work space, IT services and technology, and other utilities. Usability and integrity are core parameters for quantifying impact.

Activities	Parameters	Impact
Manage Storage	• Usability (Readability, Authenticity); • Integrity.	This activity is very important in order to ensure the physical preservation of digital data and consequently the physical access to it, that is to maintain data and technologies (HW, SW) used for accessing the data. If this activity is incorrectly performed, the risk of losing the data, as well as the ability to access the data, is very high.
Manage Refreshment Manage Migration Manage Reporting Manage Back-up Manage Reformatting/ Repackaging Manage Test and Integrity Verification	• Usability (Readability; Authenticity) • Integrity	It is very important in order to ensure the physical preservation of digital data and consequently the physical access to it, and its availability over time. Without such activities, the data can be lost in the long term, without the possibility to recover it or, if not correctly managed, the access to data could be lost.
Report on archived data format	• Integrity	These activities are relevant in order to ensure the traceability of each action on the data. This can support the integrity and completeness of data and information provided to the data users.

2. *Assigning resource costs to activities and Assigning activity costs to cost objects*.

The aforesaid step should be done with simulation and estimation value.

10 DMP-9: DATA REVIEW AND REPROCESSING

DMP Category: Curation

DMP-9: 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.

10.1 Terms

The following terms, as they relate to this guideline, are defined in Appendix A: Data Curation, Data Reprocessing, File Format, Format Conversion

10.2 Explanations of the principle

Curation, normally [4] implies most, if not all the activities of DMPs 1 to 10. Thus, its meaning as one of the 5 foundational elements of the DMPs is narrower than its usual meaning, focusing exclusively in activities beyond appraisal/selection of data and data preservation (DMPs 7 & 8) and other activities intended to ensure discoverability (DMPs 1 & 4), accessibility (DMP 2), and usability (DMPs 3 to 6). In particular it focuses on correction, updating and reprocessing of data records (DMP 9) and the use of unique and persistent identifiers (DMP 10).

Most data management planning ends with its ingestion and the processing and interpretation of raw data. But, since data processors, who preserve the integrity and authenticity of the data, are well versed with software developments, advancements in computing technology, and processing algorithms, it has produced, as a natural development, the practice of extracting more and more information from the available data. This coincides with the key "social" and "scientific" goals of providing data to distinctive communities: long-term data sets and their usability by multiple stakeholders and communities. Combining such technological processes with scientific knowledge has led to the addition of new essential elements, adding value to data records, such as a) review [leading to corrections and updating] and b) reanalysis [with or without reprocessing i) when new technologies, including new formats for presentation, emerge, or ii) when data are reviewed by other communities using different processing tools]

Updates and Corrections have increasingly become a major purpose of databases in order to facilitate comparisons between different sets of data (e.g. between in situ observations -regionally, temporally, by technique, by investigator, etc.-, as well as between in situ and remotely sensed observations). Updating and correcting processed data can be time consuming, resource intensive, and constrained by time and interpreter choices to meet user needs.

Reprocessing can produce higher quality data (in particular fidelity images of multiple datasets of different categories of earth observations) than those created during initial processing. Data reprocessing is often necessary and can include, e.g., updating of the instrument calibration, taking account of current knowledge about sensor degradation and radiometric performance; or applying new knowledge in terms of data correction and/or derived products algorithms. Reprocessing also can change the output file format. F**ormat conversion** or **reformatting** might be an additional and usual consequence not necessary linked to reprocessing.

10.3 Guidance on Implementation, with Examples

Updates and corrections to submitted data sets is encouraged. Records of updates and corrections should be maintained; summaries of updates should be posted in the database, and users should be notified. Whether it should be the provider's or the data curator's responsibility to ensure that the current data in the archive is identical to the data used in the most recent publications or current research is open to debate. But such responsibilities should be stated in data provision arrangements and transparent to users. Corrections might initiate debates (e.g. the July 2015 NOAA corrections of the dataset questioning the hiatus and slowdown of 21st century global temperature rise) but should not

prevent implementation of correction policies and methodologies and results from being open to the designated communities.

Reprocessing should be strongly considered when 1) the quality of the end product from processing does not meet the objectives of the designated community and there is technology (whether new or from another community) available to improve it; 2) the data were processed with different objectives or with objectives appropriate only at the time of its processing; 3) when acquisitions of more data in adjoining areas or in the same area (with new parameters or type), necessitates reanalysis; 4) when new techniques and processing steps are more suitable to tackle the problem in the issue-area; 5) when new software is more suitable for processing the data; and/or 6) when new processing skills, experience and knowledge offer improvements.

Reprocessing has limitations. It can strain resources, including time, personnel, and expertise, requiring more quality control, interpretation, data handling and additional computer resources. Dataset or collection-specific limitations include software or hardware (e.g. processing systems and algorithm differences in various data sets limit or enhance ultimate quality), geographic-bound or time-bound data sets with bad data quality that are not suitable for reprocessing with the new technologies, and when new reprocessing techniques cannot overcome errors made during acquisition etc. Ideally, reprocessing should deliver new data products that are part of a very long time series. At times, data reprocessing needs a previous phase as a proof of concept before it becomes a broader initiative or a consolidated policy. Communication of strengths, limitations and uncertainties of reprocessed observations and reanalysis data to the developer community and the extended research community, including the new generations of researchers and the decision-makers, is crucial for further advancement of observational data records.

10.4 Metrics to measure level of adherence to the principle

10.4.1 Substantive metrics:

Since usability is the main purpose of curation, metrics have traditionally been linked to citation metrics. Other metrics also are being considered (e.g. US NAS analysis of indicators of STI activities in the US and abroad that NCSES should produce; metrics on socio-economic benefits of interdisciplinary data curation from the Use of Earth Observations [5]). Concerning GEO, CEOS has made an unprecedented effort to develop a roadmap with specificity, actionability, responsibility, and desired outcomes in terms of quantitative metrics of ECVs, and there are ongoing exercises to provide metrics for the EBVs by the GEOBPN Leipzig Center. Qualitative descriptions also are valuable and should not be abandoned. See, e.g., Conway *et al*, describing impact of curation of data on disasters, health, energy, climate, water, ecosystems and agriculture [6]. Agreement on universal metrics may be difficult.

10.4.2 Process-based metrics:

Does it make sense to "create" a metrics system (or scoreboard) based on whether institutional processes of updating, correction and reprocessing policies are under study, development or already in place, similar to those in DMP7? Or similar to the DCC data appraisal metrics?

10.5 Resource Implications of Implementation

Both updating and corrections, as well as reprocessing, are detailed, labor intensive, time-consuming, and prone to errors. Each reusable data set or collection requires specific reprocessing steps or techniques appropriate for the specific data set or group. Many variables impact the effectiveness of reprocessing, such as reprocessing challenges at individual facilities (time, expertise, computer equipment, quality and completeness of reprocessing instructions) and change due to technological evolution, since reprocessing requires precision, as well as periodic retraining to assure staff competence.

Reprocessing is still not considered strictly necessary in many areas. Climate change related observations are the paradigmatic data sets that need reprocessing since a major difficulty in understanding past climate change is that most systems used to make the observations that climate scientists now rely on were not designed with their needs in mind. Current observation system requirements for climate monitoring and model validation such as those specified by GCOS are rarely aligned with the capabilities of historical observing systems, emphasizing continuity and stability. It is no surprise that the GEO 2009-2011 Work Plan has only one task specifically addressing reprocessing: CL-06-01a on Sustained Reprocessing and Reanalysis of Climate Data. But even in this area, e.g. in the CEOS 2014-2016 Work Plan considers that only the data from the TOPEX/Poseidon mission ended in 2006 -VC-13-, although it admits -CMRS-3: Action plan (first version)- that it is necessary to create the conditions for delivering further climate data records from existing observational data by targeting processing gaps/shortfalls/opportunities (e.g., cross-calibration, reprocessing).

Alternatives to reprocessing such as OTFR (on-the-fly reprocessing) that generate real-time new data products or other dynamic data processing techniques (as well as migration to intermediate XML for file format conversions or e-streaming technologies) are still in their initial research or development phases.

11 DMP-10: PERSISTENT AND RESOLVABLE IDENTIFIERS

DMP Category: Curation

DMP-10: Data will be assigned appropriate persistent, unique and resolvable identifiers to enable documents to cite the data on which they are based and to enable data providers to receive acknowledgement for use of their data.

11.1 Terms

The following terms, as they relate to this guideline, are defined in Appendix A: A persistent, unique and resolvable identifier, Persistence, Resolution to Location, Unique Identity

11.2 Explanation of the principle

Assigning a persistent, unique and resolvable digital identifier to data allows researchers and other users to communicate unambiguously the data that were used in the published research and contributes to the transparency and reproducibility of research. Persistent, unique and resolvable identifiers are an important component in the mechanism and practice of citation. They remove ambiguity about which work or data has been cited and easily allow citations to be counted and used as a metric for research contributions.

Data citations allow the user to locate the evidence underpinning a research statement, which is critical for scientific practice and the process of verification, and they provide acknowledgment of a source, which has become culturally important in the practice of attributing intellectual debt and as one of the metrics for assessing research contributions.

Improving data citation practice is an important step to ensure that contributions of data creators and data curators are acknowledged. In turn, such recognition should lead to proper financial support for data sharing and data stewardship, which are essential research lifecycle activities.

Thus, the Joint Declaration of Data Citation Principles [https://www.force11.org/group/joint-declaration-data-citation-principles-final] states:

Sound, reproducible scholarship rests upon a foundation of robust, accessible data. For this to be so in practice as well as theory, data must be accorded due importance in the practice of scholarship and in the enduring scholarly record. In other words, data should be considered legitimate, citable products

of research. Data citation, like the citation of other evidence and sources, is good research practice and is part of the scholarly ecosystem supporting data reuse.

All the Data Citation Principles are relevant to this Data Management Principle.

Relatedly, the San Francisco Declaration on Research Assessment (DORA) [http://www.ascb.org/dora/] calls for metrics relating to the value and impact of all research outputs, including datasets and software, to be included in the assessment of research contributions.

11.3 Guidance on Implementation, with Examples

The persistence, resolvability and uniqueness of an identifier depend on responsibility being taken to enact and maintain a series of key functions.

- **Persistence and uniqueness of the identifier:** a registration authority must ensure that the identifier is unique and that information is maintained that unambiguously associates the identifier with the resource. The identifier itself (the string of numbers or letters in whatever format) must be maintained and must not change;
- **Persistence of resolution of identifier to location:** a mechanism must be provided that enables the resource to be found at a specific location on a network. As noted above, this will generally be to a freely accessible 'landing page' providing detailed metadata relating to the data resource. If the data resource is moved, steps must be taken to ensure that the identifier resolves to the new location;
- **Persistence of landing page:** If for whatever reason the data holder needs to remove (de-accession or destroy the data itself) the landing page must be maintained and must provide information that this step has been taken. The identifier and metadata must persist even if the data resource has been destroyed;
- **Persistence checking:** to maintain these functions regular checking of link resolution, resource persistence and location should be undertaken.

Organizations that maintain and provide access to data resources should ensure that these functions are carried out, whether by the organization itself or by a third party.

The key words here are persistence and responsibility. The authors of Clark et al. 2015, recommend that all organizations endorsing the Joint Declaration of Data Citation Principles adopt a "Persistence Guarantee':

[Organization/Institution Name] is committed to maintaining persistent identifiers in [Repository Name] so that they will continue to resolve to a landing page providing metadata describing the data, including elements of stewardship, provenance, and availability.

[Organization/Institution Name] has made the following plan for organizational persistence and succession: [plan].

The capacity to deliver such a guarantee corresponds to some of the criteria for being a Trusted Digital Repository (TDR) [see above, DMP-7 and reference DSA/WDS]

11.3.1 Persistent Identifier Schemes

A number of persistent identifier schemes exist. The principal ones, summarized in Clark et al. 2015, include PURLs (Permanent Uniform Resource Locators), the Handle System, ARKs (Archival Resource Keys), CrossRef and DataCite DOIs (Digital Object Identifiers). Some databases and data archives use their own identifier system and maintain the resolution between these identifiers and a location themselves.

DOIs are built on the Handle System. CrossRef and DataCite are Registration Agencies that provide services for registering and resolving DOIs and ensure persistence by requiring specific commitments from registering organizations and by actively monitor compliance.

The following table is adapted from Clark et al. 2015 and summarises the approach of the most important identifier schemes used for identifying data to maintain persistence.

Scheme	Authority	Resolution URI	Achieving Persistence	Enforcing Persistence	Action on Removal of Data Resource
PURL	Online Computer Library Centre (OCLC)	https://purl.org	Registration	None	Domain owner responsibility
ARK	Various Name Assigning or Mapping Authorities	http://n2t.net; Name Mapping Authorities	User-defined policies	Hosting server	Host-dependent; metadata should persist
Handle	Corporation for National Research Initiatives (CNRI)	http://handle.net	Registration	None	Identifier should persist
DataCite DOI	DataCite	http://dx.doi.org	Registration with contract	Link checking	DataCite contacts owners; metadata should persist

Data contributed to GEOSS should be assigned appropriate persistent**, unique** and resolvable identifiers. Both the organisation holding the data and GEOSS should indicate clearly how the data should be cited by those using the data in published work.

11.4 Metrics to measure level of adherence to the principle

Measures of adherence are as follows:

1. Assigning appropriate, persistent, unique and resolvable identifiers to data sets contributed to GEOSS;
2. Resolution of the identifier to the data landing page;
3. Clear statement on the landing page and in the GEOSS entry of how to cite the data; and
4. Good practice data citation in the GEO community.

11.5 Resource Implications of Implementation

Data archives should subscribe to a service that generates unique persistent identifiers for data and should assign an identifier to each data product that is released to the public. The data identifier assignments may be initiated automatically or manually by the archive. The recommended citation for each data product should include the data product identifier.

APPENDIX A

DEFINITIONS OF TERMS

Access Rights Information: The information that identifies the access restrictions pertaining to the Content Information, including the legal framework, licensing terms, and access control. It contains the access and distribution conditions stated within the Submission Agreement, related to both preservation (by the repository) and final usage (by the Consumer). It also includes the specifications for the application of rights enforcement measures. [From DMP-7]

Archive: An organization that intends to preserve information for access and use by a Designated Community. [From DMP-7]

Authentication: Authentication is the process of giving users access to systems based on their identity. Authentication merely ensures that the user is who he or she claims to be, but says nothing about the access rights of the user. Usually it is based on a username and password. [From DMP-2]

Authenticity: The degree to which a person (or system) regards an object as what it is purported to be. Authenticity is judged on the basis of evidence. [From DMP-7]

Authenticity: *the property of authentic data and associated metadata as being what they purport to be — reliable assets that over time have not been altered, changed or otherwise corrupted.*

Assuring continued authenticity is an essential but intransigent preservation consideration for digital data and records. Authenticity verification requires the use of metadata. The critical change for IT practices is that metadata is now very important and must be safeguarded with the same priorities as the data. Authenticity must involve the entire process from submission of information to a repository, creation of the data record containing the necessary metadata, and security and reliability of the stored information record. Validation of the information at the time of submission is crucial. This includes secure transmission and authentication but may also extend into requirements on the processes producing the information, such as ensuring who is the author or owner of the information (Context and Provenance information). [From DMP-8]

Authorization: Authorization is the process of granting or denying access to a resource that can be a web service or a dataset. [From DMP-2]

Broker: Transforms a dataset from one standard into another. A broker can read and mediate among the many standards and specifications used by different communities of practice.[3] [From DMP-1]

Catalogue: A data catalogue is a collection of metadata about datasets. [From DMP-1]

Clearinghouse: In general a clearinghouse provides a central access point for value-added topical guides that identify, describe, and evaluate Internet-based information resources. A clearinghouse is a system of servers located on the Internet that contain field-level descriptions of available digital data. This descriptive information, known as metadata, are collected in a standard format to facilitate query and consistent presentation across multiple participating sites. A clearinghouse uses readily available Web technology for the client side and uses standards for the query, search, and presentation of search

[3] http://www.eurogeoss.eu/broker/Pages/TheEuroGEOSSBrokeringPlatform.aspx

results to the Web client. A clearinghouse provides information about who is providing which authorized geoinformation for which application (GETIS).[4] [From DMP-1]

Community-Approved Standards: Standards that are typically narrowly focused, and published and maintained by scientific or disciplinary communities, such as official Communities of Practice, or more informal groups that represent a certain discipline or area of interest. [From DMP-4]

Consumer: The role played by those persons, or client systems, who interact with repository services to find preserved information of interest and to access that information in detail. This can include other repositories, as well as internal repository persons or systems. [From DMP-7]

Core Elements: the minimum subset of metadata fields that need to be maintained for a dataset. [From DMP-1]

Curation[5]: Activities required to make deposited data preservable or usable now and in the future. Depending on technological changes, curation may be required at certain points in time throughout the data lifecycle. [From DMP-7]

Data: A reinterpretable representation of information in a formalized manner suitable for communication, interpretation, or processing. Examples of data include a sequence of bits, a table of numbers, the characters on a page, the recording of sounds made by a person speaking, or a moon rock specimen. [From DMP-7]

Data Curation: is the active and on-going management of data through its lifecycle of interests and usefulness to scholarship, science, and education. These activities should "enable data discovery and retrieval, maintain its quality, add value, and provide re-use over time" and include "authentication, archiving, management, preservation, retrieval, and representation" [1] [From DMP-9]

Data Quality Indicator: Values specifying the level of quality determined for each data quality criterion. [From DMP-6]

Data Reprocessing: Treatment of data reclaimed from existing data sets to obtain new data products. [From DMP-9]

Designated Community: An identified group of potential Consumers who should be able to understand a particular set of information. The Designated Community may be composed of multiple user communities. A Designated Community is defined by the Archive and this definition may change over time. [From DMP-7]

Digital Migration: The transfer of digital information, while intending to preserve it, within the repository. It is distinguished from transfers in general by three attributes: a focus on the preservation of the full information content that needs preservation; a perspective that the new archival implementation of the information is a replacement for the old; and an understanding that full control and responsibility over all aspects of the transfer resides with the repository. [From DMP-7]

Digital Object: An object composed of a set of bit sequences. [From DMP-7]

Discovery: the act of finding or learning something for the first time (Merriam-webster). [From DMP-1]

Discovery Services: making it possible to search for data sets and services on the basis of the content of the corresponding metadata and to display the content of the metadata.[6] [From DMP-1]

[4] http://cordis.europa.eu/project/rcn/60644_en.html

[5] Term not present in the OAIS glossary

Documented: Data that has associated metadata, where the metadata elements contain information necessary to assist data users in accessing the data, using the data, understanding the data, and processing the data. [From DMP-4]

Encoding: TBD. [From DMP-3]

Essential Variables: TBD. [From DMP-3]

File Format: The internal structure and encoding of a digital object, which allows it to be processed, or to be rendered in human-accessible form [2]. [From DMP-9]

Format Conversion: copying the digital content from one type of storage medium to another, in a permanent attempt to outrun the obsolescence of one generation after another of data carriers and their associated hardware; frequently, format conversions rather involve translation from one data format (or file format) to another, to outrun the obsolescence of the format and its associated software [3] [From DMP-9]

GeoJSON: TBD. [From DMP-3]

Identifier: a Persistent, Unique and Resolvable Identifier: A maintainable digital identifier that allows a digital object (a file or set of files) to be referenced. [From DMP-10]

Ingest[2]: The process of entering data and associated metadata into a data repository. [From DMP-7]

Integrity[2]: Internal consistency or lack of corruption of digital objects. Integrity can be compromised by hardware errors even when digital objects are not touched, or by software or human errors when they are transferred or processed. [From DMP-7]

Integrity: *the property of safeguarding data and associated metadata accuracy and completeness. Integrity refers to the assurance that data and associated metadata are not lost or damaged as a result of malicious or inadvertent activity.*

The most important measure to ensure integrity of stored digital information is access control. Additional protection is provided by checksums that may be applied to individual records, files or disk structures. The best protection is to store however several copies of each data record in separate systems under separate administration and possibly also in separate locations. At least three copies should exist in order to enable a majority vote to determine the correct version and to grant the Data Preservation. [From DMP-8]

International Standards: Standards that are published and maintained by recognized international Standards Development Organizations, such as IEEE, ISO, OGC, etc. [From DMP-4]

License: A permission or a set of permissions regarding whatever is licensed. When I give someone a license to do something, I give them the permission to do it. I have rights that I license. [From DMP-1]

Long Term: A period of time long enough for there to be concern about the impacts of changing technologies, including support for new media and data formats, and of a changing Designated Community, on the information being held in a repository. This period extends into the indefinite future. [From DMP-7]

Long Term Preservation: The act of maintaining information, Independently Understandable by a Designated Community, and with evidence supporting its Authenticity, over the Long Term. [From DMP-7]

6

http://inspire.ec.europa.eu/documents/Network_Services/TechnicalGuidance_DiscoveryServices_v3.0.pdf

Metadata: information describing data sets and data services and making it possible to discover, inventory and use them - alternative: data about data. [From DMP-1]

Metadata Element: a discrete unit of metadata, in accordance with ISO 19115 and 19139. [From DMP-1]

Network Services: computing services that make it possible to discover, transform, view and download data and to invoke data and e-commerce services. [From DMP-1]

Open Archival Information System (OAIS): An Archive, consisting of an organization, which may be part of a larger organization, of people and systems, that has accepted the responsibility to preserve information and make it available for a Designated Community. It meets a set of responsibilities, as defined in section 4, that allows an OAIS Archive to be distinguished from other uses of the term 'Archive'. The term 'Open' in OAIS is used to imply that this Recommendation and future related Recommendations and standards are developed in open forums, and it does not imply that access to the Archive is unrestricted. [From DMP-7]

Persistence: the identifier and the resolution are persistent in that some entity, depending on the system involved, takes responsibility for ensuring that the information 1) defining the identifier's relationship to a specific resource and 2) the resolution to a given location are maintained. [From DMP-10]

Preferred Formats: Formats that a repository can reasonably assure will remain readable and usable. Typically, these are the de facto standards employed by a particular discipline. [From DMP-7]

Producer: The role played by those persons or client systems that provide the information to be preserved. This can include other repositories or internal repository persons or systems. [From DMP-7]

Provenance: *Part of the metadata that documents the history of the content information. This information tells the origin or source of the content information, any processes and changes that may have taken place since it was originated, and who has had custody of it since it was originated. It is sometimes referred to as lineage.*

Complete provenance information is part of the information required for assessing the validity and fitness for purpose of a dataset or product. It is composed of references and descriptions of the data sources, data processes and algorithms used. It also includes a description of responsible parties involved in all the steps of the process chain. [From DMP-5]

Provenance Information: The information that documents the history of the Content Information. This information tells the origin or source of the Content Information, any changes that may have taken place since it was originated, and who has had custody of it since it was originated. The Archive is responsible for creating and preserving Provenance Information from the point of Ingest; however, earlier Provenance Information should be provided by the Producer. Provenance Information adds to the evidence to support Authenticity. [From DMP-7]

Quality-Control: Data quality-control is conducted by reviewing data to assess their potential for use. [From DMP-6]

Queryable: a metadata element that can be queried upon or that is part of a query. [From DMP-1]

Readability: *the property of assuring data and associated metadata usage over the long term.*

All activities such as reformatting, data refreshment and duplication, etc., aim to grant that the data and the associated metadata are accessible and readable for the entire retention period, and that they are viewed and understood. [From DMP-8]

Reference Model: A framework for understanding significant relationships among the entities of some environment, and for the development of consistent standards or specifications supporting that

environment. A reference model is based on a small number of unifying concepts and may be used as a basis for education and explaining standards to a non-specialist. [From DMP-7]

Resolvable: The identifier contains information that enables access (e.g. via browser click) to a specific location on a network, even if the location of the metadata or data has changed. Most often this will be to metadata presented on a landing page that acts as a proxy for the data resource. [From DMP-10]

Search Engine: Computer program that can search indexed topics. [From DMP-1]

SSO – Single Sign-On: Single sign-on (SSO) is an authentication process that allows a user to access multiple servers with one set of login credentials. With SSO, a user logs in once and gains access to different servers, without the need to re-enter log-in credentials each time. [From DMP-2]

Succession Plan: The plan of how and when the management, ownership and/or control of the repository holdings will be transferred to a subsequent repository in order to ensure the continued effective preservation of those holdings. [From DMP-7]

Traceability: Property of a measurement result whereby the result can be related to a reference through a documented unbroken chain of process steps and source each contributing to the measurement of the uncertainty.

In other words, it is the capability to trace back the data to its origins. Traceability implies that provenance information is complete enough for the user to assess the uncertainty of the data. This is one of the aims of documenting provenance. Another term that is relevant to provenance is *reproducibility*, which would require the provenance information to be complete enough, and in a clear sequence, to enable recreation of the data from its sources by applying the process steps. [From DMP-5]

Unique Identity: The identifier communicates unique information confirming that it refers to, and only to, a given object - in other words, the identifier itself is unique, while the thing it is identifying may not be. The uniqueness is maintained by the registration authority. [From DMP-10]

Use Conditions: something that limits or restricts the use or reuse of a resource; a qualification. [From DMP-1]

Web-Service: The W3C defines a Web service generally as: “a software system designed to support interoperable machine-to-machine interaction over a network.” [From DMP-2]

APPENDIX B

REFERENCES

From DMP-3

Fulker, D. and Gallagher, J. (2013) Extending OPeNDAP to Offer Remapping Services. *Geophysical Research Abstracts*. [Online] EGU2013-10236, 2013 EGU General Assembly 2013. 15. Available from: http://meetingorganizer.copernicus.org/EGU2013/EGU2013-10236.pdf.

GCOS. (2015) *GCOS Essential Climate Variable (ECV) Data Access Matrix*. [Online] Available from: http://www.gosic.org/ios/MATRICES/ECV/ECV-matrix.htm

GeoJSON. (2015) *The GeoJSON Format Specification*. [Online] Available from: http://geojson.org/geojson-spec.html

Hefflin, J. and Hendler, J. (2000) Semantic Interoperability on the Web. *Extreme Markup Languages 2000, August 15-18, Montreal, Canada*. [Online] Available from:

http://www.cs.umd.edu/projects/plus/SHOE/pubs/extreme2000.pdf.

Hugo, W. (2009) Meta-Data Implementation For The Environmental Sciences: Options, Benefits And Issues - Saeon Case Study. *African Digital Scholarship and Curation Conference, 12-14 May 2009*. [Online]. Available from: http://www.ais.up.ac.za/digi/docs/hugo_present.pdf.

Hugo, W. (2008) Beyond Spatial Data Infrastructure: Knowledge And Process Extensions. *Ecological Circuits*. [Online] 2 (2008). p.25. Available from: http://web.archive.org/web/20120512061432/http://eepublishers.co.za/images/upload/beyond%20spatial.pdf.

OGC. (2007) *OpenGIS® Geography Markup Language (GML) Encoding Standard*. [Online] Available from: http://www.opengeospatial.org/standards/gml

OGC. (2007) *OGC Validator*. [Online] Available from: http://cite.opengeospatial.org/teamengine/

OGC. (2008) *OGC® KML*. [Online] Available from: http://www.opengeospatial.org/standards/kml

OGC. (2011) *OGC® SWE Common Data Model Encoding Standard*. [Online] Available from: http://www.opengeospatial.org/standards/swecommon

OGC. (2014) *OGC® SensorML: Model and XML Encoding Standard*. [Online] Available from: http://www.opengeospatial.org/standards/sensorml

OGC. (2015) *OGC Network Common Data Form (netCDF) Standards Suite*. [Online] Available from: http://www.opengeospatial.org/standards/netcdf

OGC. (2015) *OGC® WaterML*. [Online] Available from: http://www.opengeospatial.org/standards/waterml

OOPC. (2015) *Essential Variables*. [Online] Available from: http://ioc-goos-oopc.org/obs/ecv.php

TDWG. (2015) *TDWG Standards*. [Online] Available from: http://www.tdwg.org/standards/

WMO. (2015) *WMO Information System*. [Online] Available from: http://www.wmo.int/pages/prog/www/WIS/

From DMP-4

1. ISO 19115:
    a. ISO 19115-1: 2014 Geographic information -- Metadata -- Part 1: Fundamentals:

       http://www.iso.org/iso/home/store/catalogue_ics/catalogue_detail_ics.htm?csnumber=53798http://www.iso.org/iso/home/store/catalogue_ics/catalogue_detail_ics.htm?csnumber=53798 ;

    b. ISO 19115-2: 2009 Geographic information -- Metadata -- Part 2: Extensions for imagery and gridded data:

       http://www.iso.org/iso/catalogue_detail.htm?csnumber=39229http://www.iso.org/iso/catalogue_detail.htm?csnumber=39229 .
2. ISO 19139:2007 Geographic information -- Metadata -- XML schema implementation:

   http://www.iso.org/iso/catalogue_detail.htm?csnumber=32557 ;
3. ISO 19157:2013 Geographic information -- Data quality http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=32575 ;
4. https://earthdata.nasa.gov/standards/preservation-content-spechttp://www.iso.org/iso/catalogue_detail.htm?csnumber=32557 ;
5. Dublin Core http://dublincore.org/http://dublincore.org/ ;
6. Darwin Core http://rs.tdwg.org/dwc/http://rs.tdwg.org/dwc/ ;
7. DIF :

   https://earthdata.nasa.gov/standards/directory-interchange-format-dif-standardhttps://earthdata.nasa.gov/standards/directory-interchange-format-dif-standard ;
8. CF: http://cfconventions.org/http://cfconventions.org/ ;
9. GeoViQua:
    a. http://www.geoviqua.org/http://www.geoviqua.org/ ;
    b. GEOLabel http://www.geoviqua.org/GeoLabel.htm .

From DMP-5

Isuru Suriarachchi (2015) Survey of Provenance Practices in Data Preservation Repositories. Research Data Alliance: https://rd-alliance.org/filedepot_download/767/632 ;

GeoNetwork4MetaViz – Enabling an open source metadata catalogue to visualize lineage information (2015): http://geoportal-glues.ufz.de/documents/Fact_Sheet_MetaViz.pdf ;

Demir, I.; Mills, H.; Oh, J.; et. al. (2015), Geoscience Papers of the Future: The Importance of Software Sharing for Science Reproducibility. Presented at EarthCube All Hands Meeting, Washington, DC, 27-29 May 2015: http://earthcube.org/document/2015/geoscience-papers-future .

From DMP-6

Callahan S. 2015. Data without Peer: Examples of Data Peer Review in the Earth Sciences. D-Lib Magazine, 21(1/2). http://dx.doi.org/10.1045/january2015-callaghan ;

Peer L, Green A, Stephenson E. 2014. Committing to Data Quality Review. International Journal of Digital Curation, 9(1), 263-291 http://dx.doi.org/10.2218/ijdc.v9i1.317 .

From DMP-8

Reference: CEOS EO Data Preservation Guidelines.

- Keeping Research Data Safe (KRDS) Project: http://www.beagrie.com/krds.php
- Activity Based Costing (ABC) methodology;
- Angus Whyte and Andrew Wilson. Appraise & Select Research Data for Curation. Digital Curation Centre and Australian National Data Service "working level" guide, 25 October 2010: http://www.dcc.ac.uk/resources/how-guides/appraise-select-data .

From DMP-9

F.1 Note references

[1] Plato L. Smith II *Exploring Data Curation and Management Programs, Projects, and Services through Metatriangulation* (2012).

[2]Brown, A., 2006a. Automatic Format Identification Using PRONOM and DROID, The National Archives].

[3] C.M. Sperberg-McQueen, *What Constitutes Successful Format Conversion? Towards a Formalization of `Intellectual Content´*, in The International Journal of Digital Curation Issue 1, Volume 6 (2011).

[4] As defined in the Oxford Glossary or Common Definitions.

[5] See M. T. Borzacchiello & M. Craglia, Socio-Economic Benefits from the Use of Earth Observation (Report from the International Workshop held at Joint Research Center, Ispra, 11-13 July 2011), available at:

http://ies.jrc.ec.europa.eu/uploads/SDI/publications/EOBenefitWS_JRCTechReport_final.pdf;

or, concerning format conversions, -see C.M. Sperberg-McQueen, 2011, What Constitutes Successful Format Conversion?). See also C.M. Sperberg-McQueen, What Constitutes Successful Format Conversion? Towards a Formalization of 'Intellectual Content, The International Journal of Digital Curation Issue 1, Volume 6 (2011).

[6] Conway, Esther, Sam Pepler, Wendy Garland, David Hooper, Fulvio Marelli, Luca Liberti, Emanuela Piervitali, Katrin Molch, Helen Glaves, and Lucio Badiali. Ensuring the Long Term Impact of Earth Science Data through Data Curation and Preservation. Information Standards Quarterly, Fall 2013, 25(3): 28-36. http://www.niso.org/publications/isq/2013/v25no3/conway/ .

F.2 Curation and reprocessing examples

See the process of reanalyzing undertaken by the Analysis Centers (ACs) of the IGS of the full history of GPS data collected by the IGS global network since 1994 in a fully consistent way using the latest models and methodology. [IGS: the International GNSS Service GNSS: Global Navigation Satellite System] http://acc.igs.org/reprocess.html .

On the reprocessing of atmospheric chemistry Ether data website: http://www.pole-ether.fr/etherTypo/index.php?id=1678&L=1 .

On seismic data reprocessing: http://www.searcherseismic.com/our-services/2d--3d-seismic-data-reprocessing.htm .

And, in particular, on the reprocessing of Multi-channel Seismic-Reflection Data Collected in the Beaufort Sea http://pubs.usgs.gov/of/2000/ofr-00-460/ .

Most of NASA satellite mission data are reprocessed multiple times with improvements made each time. This results in multiple versions (variously referred to as Collection, Version or Edition) of data products. For example, see:

1.- http://oceancolor.gsfc.nasa.gov/cms/REPROCESSING;

2.- https://daac.ornl.gov/MODIS/MODIS-menu/reprocessing.html ;

3.- http://disc.sci.gsfc.nasa.gov/Aura/data-holdings/MLS/index.shtml ;

4.- https://earthdata.nasa.gov/modis-terra-collection-6-aerosol-cloud-and-other-atmospheric-level-2-and-level-3-products-released ;

On sea ice data reprocessing:

http://nsidc.org/data/nsidc-0508

https://climatedataguide.ucar.edu/climate-data/sea-ice-concentration-data-reprocessed-ssmr-ssmi-eumetsat .

On the 2010-2013 ESA reprocessing campaign of all the SMOS (Soil Moisture and Ocean Salinity) data: http://cp34-bec.cmima.csic.es/ocean-reprocessed-dataset/

On the Landsat 8 data held in the USGS archives reprocessing, introducing corrections affecting both the Operational Land Imager (OLI) and the Thermal Infrared Sensor (TIRS): http://landsat.gsfc.nasa.gov/?p=7435 .

On the ESA Landsat 5 Reprocessing, see Alessandra Paciucci et al**,** *Landsat 5 Reprocessing: Case Study Into Reprocessing and Data Configuration 4-11-2013.*

On the reprocessing of global vegetation images: http://proba-v.vgt.vito.be/content/reprocessing-proba-v-data-finalized .

On NOAA´s project on AVHRR (Advanced Very High Resolution Radiometer) reprocessing in TIMELINE (TIMe Series Processing of Medium Resolution Earth Observation Data assessing Long - Term Dynamics In our Natural Environment): see Katrin Molch et al, NOAA AVHRR Data Curation and Reprocessing - TIMELINE (2013), and for Europe and North Africa, C.M. Frey C. et al, (2015) *AVHRR re-processing over Europe and North Africa.* 36th International Symposium on Remote Sensing of Environment, 11-15 May, Berlin, Deutschland.

From DMP-10

Ball, A. & Duke, M. (2012). 'How to Cite Datasets and Link to Publications'. DCC How-to Guides. Edinburgh: Digital Curation Centre. Available online: http://www.dcc.ac.uk/resources/how-guides .

Joint Declaration of Data Citation Principles https://www.force11.org/group/joint-declaration-data-citation-principles-final .

San Francisco Declaration on Research Assessment (DORA) [http://www.ascb.org/dora/]

Starr J, Castro E, Crosas M, Dumontier M, Downs RR, Duerr R, Haak LL, Haendel M, Herman I, Hodson S, Hourclé J, Kratz JE, Lin J, Nielsen LH, Nurnberger A, Proell S, Rauber A, Sacchi S, Smith

A, Taylor M, Clark T. (2015) Achieving human and machine accessibility of cited data in scholarly publications. PeerJ Computer Science 1:e1 https://dx.doi.org/10.7717/peerj-cs.1

Tonkin, E. (2008) 'Persistent Identifiers: Considering the Options' Ariadne Issue 56 http://www.ariadne.ac.uk/issue56/tonkin/

Current GEO Recommendations, GEOSS Data Citation Guidelines: Version 2 http://www.gstss.org/library/GEOSS_Data_Citation_Guidelines_V2.0.pdf

DCC Guide, 'How to Cite Datasets', provides summary of minimal information for data citations, explanations of DOIs, discusses current issues (including granularity) and implementation issues (including versioning in the context of time series data): http://www.dcc.ac.uk/resources/how-guides/cite-datasets

DataCite Metadata Schema v 3.1: https://schema.datacite.org/meta/kernel-3/index.html

Re: Assigning identifiers to different versions of a dataset, particularly for time series data.

Dryad DOI Usage: http://wiki.datadryad.org/DOI_Usage

UKDA Approach to Persistent Identifiers and Versioning: http://www.bl.uk/aboutus/stratpolprog/digi/datasets/workshoparchive/LousieCortin_IdentifiersForTheUKDA_May2012.pdf

Re: Assigning identifiers and citing data (subsets) created dynamically by database queries.

Recommendations from Research Data Alliance Working Group on Data Citation: https://rd-alliance.org/filedepot/folder/262?fid=667

Re: machine accessibility of cited data by mean of a persistent and unique identifier.

Clark et al. 2015 'Achieving human and machine accessibility of cited data in scholarly publications' https://dx.doi.org/10.7717/peerj-cs.1

APPENDIX C

ACRONYMS AND ABBREVIATIONS

API – Application Programming Interface

ARK – Archival Resource Key

CCSDS – Consultative Committee for Space Data Systems

CEOS – Committee on Earth Observation Satellites

CF – Climate and Forecast

CNR – National Research Council

CNRI – Corporation for National Research Intiatives

CSW – Catalogue Service for the Web

DAB – Discovery and Access Broker

DCC – Digital Curation Centre

DIF – Directory Interchange Format

DMP – Data Management Principle

DOI – Digital Object Identifier

DORA – Declaration on Research Assessment

DSA – Data Seal of Approval

EBV – Essential Biodiversity Variables

ECV – Essential Climate Variables

FGDC – Federal Geographic Data Committee

GCOS – Global Climate Observing System

GEO –Group on Earth Observations

GeoJSON – Geographic JavaScript Object Notation

GEOSS – Global Earth Observation System of Systems

GeoViQua – Quality Aware Visualization for the Global Earth Observing System of Systems

GETIS – Geo-Processing Networks in a European Territorial Interoperability Study

GML – Geography Markup Language

HW – Hardware

ICSU – International Council for Science

IEEE – Institute of Electrical and Electronics Engineers

ISO – International Organization for Standardization

JRC – Joint Research Centre

JSON – JavaScript Object Notation

KML – Keyhole Markup Language

MIME – Multipurpose Internet Mail Extensions

NAS – National Academy of Sciences

NCSES – National Center for Science and Engineering Statistics

NetCDF – Network Common Data Form

NOAA – National Oceanic and Atmospheric Administration

OAI-PMH – Open Archives Initiative Protocol for Metadata Harvesting

OAIS – Open Archival Information System

OCLC – Online Computer Library Centre

OGC – Open Geospatial Consortium

OPeNDAP – Open-source Project for a Network data Access Protocol

OTFR – On-the-Fly Reprocessing

PURL – Permanent Uniform Resource Locator

SDI – Spatial Data Infrastructure

SensorML – Sensor Model Language

SSO – Single Sign-On

STI – Science, Technology, and Innovation

SW – Software

SWE – Sensor Web Enablement

TDR – Trusted Digital Repository

TDWG – Taxonomic Databases Working Group (also known as Biodiversity Information Standards)

TOPEX – Ocean Topography Experiment

URI – Uniform Resource Identifier

WaterML – Water Markup Language

WDS – World Data System

WMO – World Meteorological Organization

WMS – Web Map Service

WMTS – Web Map Tile Service

WPS – Web Processing Service

XML – Extensible Markup Language

XSD – XML Schema Definition

APPENDIX D

LIST OF CONTRIBUTORS

The following list of contributors is in alphabetic order, implying no order of priority or responsibility.

Name	Organization	Role
▪ Albani, Mirko	▪ ESA	▪ Author, DMP-8 Lead
▪ Alonso, Enrique	▪ RDA	▪ Author, DMP-9 Lead
▪ Baker, Garry	▪ UK	▪ Author
▪ Browdy, Steven	▪ OMS Tech / IEEE	▪ Author, DMP-4 Lead, Editor
▪ Chen, Bob	▪ ICSU	▪ Author
▪ De Lathouwer, Bart	▪ OGC	▪ Author, DMP-1 Lead
▪ Downs, Robert	▪ ICSU	▪ Author, DMP-6 Lead, Editor
▪ Duerr, Ruth	▪ ESIP	▪ Author
▪ Hodson, Simon	▪ CODATA	▪ Author, DMP-10 Lead, Editor
▪ Hugo, Wim	▪ WDS	▪ Author, DMP-3 Lead
▪ Khalsa, Siri Jodha Singh	▪ IEEE	▪ Editor
▪ Kishor, Puneet	▪ CC	▪ Author
▪ Maso, Joan	▪ Spain	▪ Author, DMP-5 Lead
▪ Mokrane, Mustapha	▪ WDS	▪ Author, DMP-7 Lead
▪ Moreno, Richard	▪ CEOS	▪ Author, DMP-2 Lead