Revised GEO Data Sharing and Data Management Principles

This document is submitted by the Secretariat to the Programme Board for decision.

1　INTRODUCTION

In 2015, the GEO Data Management Principles Task Force was tasked with defining a common set GEO Data Management Principles[1]. These principles address the need for discovery, accessibility, usability, preservation, and curation of data and related resources that are shared. Such resources also should be shared as open data in accordance with the GEO Data Sharing Principles[2]. The GEO Data Management Principles complement the FAIR Principles and TRUST Principles, which also are being adopted across research communities. The GEO Data Management Principles can be applied to the entire data management lifecycle, while the FAIR principles (Wilkinson M.D. et al.) focus primarily on aspects of metadata, including persistent identifiers. The TRUST principles (Lin D. et al.) primarily focus on the curation and preservation of data and related resources. To support the implementation of the principles, the GEO Data Management Principles Guidelines have been developed so that data providers and other stakeholders can use them as a reference as they seek to implement the principles. These guidelines can also be used when assessing how well the principles are being followed in practice. This current version of the guidelines is the result of a revision process conducted by the GEO Data Working Group in 2022.

These guidelines are applicable to the management of all Earth Observation data products, both remote sensing and in-situ data, as well as other types of data products and services. These guidelines are intended to cover raw data and higher level products, including Analysis Ready Data (ARD) and data products that are produced on-demand by services. While the principles can be applied to services that generate data on-demand; this guidance focus primarily on data and related products. The Data Management Principles can be applied to Decision Ready Information (DRI) and other Knowledge assets, but these guidelines do not explicitly offer recommendations for them.

The GEO Data Branding website[3] offers a self-assessment opportunity for data providers to generate a GEO label that reflects the level of implementation of the DMPs. The GEOSS Yellow Pages[4] (https://www.geoportal.org/yellow-pages) provide information about the level of implementation of the DMPs as have been declared once by data providers and covering all their data. At this stage, the GEO has not defined any process for independent certification of a data provider against the DMPs.

Guidance for implementing the GEO Data Management Principles (DMP) is offered within each section of the guidelines. The following topics are covered in the guidelines for each of the DMP:

---

[1] https://www.earthobservations.org/documents/dswg/201504_data_management_principles_long_final.pdf
[2] https://www.earthobservations.org/dswg.php
[3] https://geolabel.info/
[4] https://www.geoportal.org/yellow-pages

- Title and Category of the DMP,
- Explanation of the DMP,
- Guidance on implementation, standards applicable and examples;
- Metrics to measure level of adherence to the principle;
- Resource Implications of Implementation

The Appendix contains a compilation of terms, an explanation about overarching interoperability levels, references, a summary of changes with respect to the original version and an acknowledgement of contributions.

Most of the guidelines reference community standards and practices that foster implementation of a principle in an interoperable manner by enabling one or more of the following levels of interoperability, which are further described in Appendix B: syntactic, semantic, schematic, and legal interoperability.

2   DMP-1: METADATA FOR DISCOVERY

DMP Category: Discovery.

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

2.1   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 bookshelf. The library's catalogue allows the visitor to search information about the books (e.g. author, ISBN number, genre, keywords), 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 s/he wants by searching and retrieving the metadata associated with that digital resource, including information on how the resource can be accessed and whether there are any restrictions or conditions placed on its use. Consistent with the FAIR Data Principles, data should be Findable, Accessible, Interoperable, and Reusable (Wilkinson, et al., 2016). Currently, metadata is concentrated in metadata catalogues that are web services that can be interrogated to find the relevant datasets for an application. GEOSS maintains a catalogue of resource descriptions (metadata) and, like a library catalogue, does not keep copies of resources, but manages the metadata that facilitates discoverability, allowing users to locate and access the resources. Such resource descriptions may point to other catalogues - analogous to a library directing a visitor to other libraries' catalogues.

Not all users begin a search for resources by going to a specific catalogue. Some start by a study or a publication that cite artifacts like data and data fragments, or even software and methods to create new data. They use full citations and PIDs (see DMP-10) to find that dataset directly. Other users might be more inclined to use search engines instead of catalogues. For this reason catalogues may link to portals, such as a geoportal, for use by humans, as well as programmatic interfaces (Application Programming Interfaces - APIs) meant for access by search engines, metadata harvesters and the portals of other communities. The discovery of data is enabled by the syntactic interoperability provided by the combination of catalogue services and API and the metadata standard (commonly expressed in XML) that make use of XML schemas to improve schematic interoperability. However, in practice, true semantic interoperability is difficult to achieve, often requiring brokering and mediation to align with the preferred metadata standard

used by the catalogue. A future consideration is the extent to which it will be possible to persist such mediations for re-use.

2.2   Guidance on Implementation, with examples

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

- a descriptive title and a description (present in. DataCite, Dublin Core, ISO 19115, DCAT, etc. standards);
- identity and contact information (e.g. ORCID, RoR, WikiDATA) for the individuals responsible for the creation of the resource;
- identity and contact information for the individuals responsible for the management of the resource (present in DataCite, ISO 19115, etc. standards);
- geographic location or boundaries (present in DataCite, ISO 19115, etc. standards);
- temporal coverage (present in DataCite, ISO 19115, etc. standards);
- keywords describing the resource and the scientific or practical domain to which it applies (present in DataCite, Dublin Core, ISO 19115, DCAT, etc. standards);
- conditions and restrictions on use, in particular license information (present in ISO 19115, etc. standards); and
- web links to the resource and to further information about the resource (present in, DataCite, ISO 19115, etc. standards).

Data discovery is facilitated by following guidelines and standards that offer recommendations to help ensure that data and services are discoverable. Adherence to these guidelines is checked and assessed through certification of the data repository or service using baseline certifications, such as the [Core Trustworthy Data Repositories Requirements (CoreTrustSeal, 2016)](), or ISO 16363 certification.

- Catalogue entries (metadata documents) should be in accordance with accepted international or community (governance) agreed upon standards (e.g. DataCite, Dublin Core, ISO 19115, etc.), and all mandatory 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, STAC, OGC API Records, DCAT-APetc.), including search capabilities where results of user queries display in relevance-ranked order;
- The metadata kept in the catalogue should be periodically checked for validity and to ensure that accessibility is maintained through valid links from persistent and resolvable identifiers to the resources (Rauber et al., 2015), as described in DMP-10, Persistent Resolvable Identifiers. If metadata are maintained in the catalogue for resources that no longer exist, a mechanism (checksum) should be provided to point to updated versions, if any, or suitable explanations should be provided for why resources no longer exist;
- GEOSS Data/Resource Providers are encouraged to register their catalogues that describe individual resources, where multiple resources are to be made discoverable in the GEOSS Yellow Pages;
- The catalogue should provide access to resources in HTML in a way that common search engines can index their resources (e.g. using schema.org tags).
- As an alternative to create a catalogue with a search interface, a data provider may post metadata, with unique and persistent identifiers that link to the associated data, in a web-accessible location (sometimes referred to Web Accessible Folder) which can then be harvested by search engines or metadata aggregators (also see DMP-10);

- Even if the GEOSS Data Sharing Principles recommend otherwise, 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;
- All data should specify licenses, including the data that conforms to open data licenses. Metadata should also indicate if the data policy is conformant to the GEOSS Data-CORE (see the GEOSS Data Sharing Principles)

See also GEO Data Management Principles 4 'Data Documentation', which offers additional guidelines regarding documentation that allows data to be used, understood and processed.

2.3   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 established practices in terms of standards and performance; and, 3) whether the repository is certified as a trustworthy digital repository in accordance with ISO 16363, the World Data System, the Data Seal of Approval, or other certification efforts.

There are many components contributing to the implementation and measuring of metrics. Some examples are:

- Service checkers, either existing or to be developed:
  - FGDC Service status checker https://statuschecker.fgdc.gov/dashboard/geossd_114 ;
  - EC-JRC INSPIRE Reference Validator https://inspire.ec.europa.eu/validator/home/index.html
  - Committee on Earth Observation Satellites (CEOS) Working Group on Information Systems and Services (WGISS) Data Management and Stewardship Maturity Matrix (2017).
  - The GEOSS DAB also performs some testing on the metadata catalogue registered in the GEOSS Yellow Pages and provides some recommendations before accepting the catalogue as part of the metadata provided by the GEOSS Platform.
  - FAIRness https://fairassist.org/, e.g. web demo using F-UJI is available at https://www.f-uji.net

- Performance indicators and availability:
  - A catalogue data discovery service should be engineered so that it:
    - Contains no single points of failure;
    - Implements reliable crossover;
    - Detects failures as they occur.
  - Communities indicate the need for tools for validation (metadata, service, data – resources):
    - Some tools interpret standards differently (publish mapping rules);
    - 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.

2.4　Resource Implications of Implementation

Efforts to enable discoverability can be key resources for consumers and include activities such as metadata authoring and maintenance, and standing up and maintaining a catalogue service. Creating metadata can be a labor-intensive activity, which should be completed correctly, initially, since it can be more costly if corrections must be completed (Palaiologk, et al., 2012). Metadata is commonly distributed in XML format; however, metadata editors exist that can make the elaboration of metadata less error-prone and tedious and will generate XML automatically from forms content.　Metadata reviews and automated checking can also reduce such costs.　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 has 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　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 online data and the speed of transfer, as well as the ability to run queries over numerous datasets via Web APIs for cloud computing, datacubes, virtual research environments (VRE), etc.　Users now expect data to be available on demand, via online services. This is particularly relevant for products that can be updated frequently, as it allows running the production process only when the product is needed and to access the latest update. Currently, this mainly means a URL responding to HTTP(S), or secured variants of FTP based protocols but also direct processing on the cloud

To meet a wide variety of use cases, particularly analysis at large scale, users expect data to be usable to a human via a user interface (providing at least download but also tools for visualization and analysis) and to be 'machine-usable' via a Web API. The approach or algorithm for accessing the data should be described as formalized provenance metadata to foster use (see DMP-5). Access services and tools must be comprehensively described and be based on recognized standards in order to be useful.

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

- Direct access service, allowing the user to download data to their computer in the original format (e.g. a web accessible folder or Cloud Optimized GeoTIFF)
- Direct Web service, providing capabilities for interoperability among services (e.g. OGC API - Environmental Data Retrieval, OGC API - Features, OGC API - Coverages, Sensor Things API);
- Browse services, which allow users to inspect representations of candidate datasets before ordering;
- Visualization 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, OGC API - Maps, OGC API - Tiles);

- In place processing of the data:

  o Since the volume of data is increasing dramatically, it is desirable to perform processing and analysis of the data in place (e.g. in the cloud), i.e. before or instead of downloading the source data;
  o The OGC Web Processing Service (WPS) standard provides a standardized way to remotely execute processing;
  o The OGC APIs– Processes standard provides a standardized Web API for remotely executing processing through interfaces described according to the OpenAPI Specification.
  o Jupyter notebook to in place processing data using Python scripts.
  o Virtual Research Environment or Virtual laboratories to in-place processing data
  o In order to ease the transfer of the processors, some techniques can be used: e.g. virtualization, or containers techniques such as Docker

3.2   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 recognized standards. Examples of standards are :

    - Traditional geospatial (OGC WxS),
    - Web API standards (e.g. OGC API-Features, OGC API-Processes)
    - Near-continuous raster and/or Multidimensional (NetCDF/HDF5/ OPeNDAP)
    - Formats optimized for the Cloud (Cloud optimized GeoTIFF)
    - Time Series (OGC Sensor Observation Services, TimeSeriesML), OGC SensorThings API, and
    - Tabular data in database tables or as spreadsheet tables (CSV).

3. **Archived data repackaging/reformatting:** Data should be provided in the open standard formats that are needed by the designated communities and in exchange formats to facilitate interchange between archives. Formats recognized by the GDAL/OGC libraries will facilitate exchange and transformation.

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, STAC, OGC, etc.) is preferred. This allows the use of existing tools and also helps resources to be more widely used.

Many data repositories require some knowledge of the identity of who is requesting data. This will require a user management system. However, this imposes a technical barrier that may discourage some users. If after considering the advantages and disadvantages, it is deemed necessary, it is desirable to enable automatic user authentication and authorization. Single Sign On (SSO) is a way to ease the burden of the authentication process by reusing common credentials. As several SSO protocols exist, a common protocol (e.g. OpenID connect) or a federation of interoperable protocols is recommended. Protocols that only rely on cookies should be avoided, as cookies do not commonly allow service to service communication. Instead, protocols based on token exchanges are preferable.

3.3 Metrics to measure level of adherence to the principle

Online data accessibility using a standard browser or web service indicates adherence. Service availability and quality should be measured for throughput (reply or query speed, processing latency), and causes of failure (processing error / request error / no data / …), etc.

3.4 Resource Implications of Implementation

Providing simple accessibility for small numbers of users to access low volumes of data that have been prepared in advance can be accomplished with minimal cost using freely available resources and can be aided by the availability of free and open source software that can reduce the cost of deploying standardized data services. Offerings range from spatial databases (PostgreSQL), through data servers (GeoServer, OGC Sensor Observation Services, OPeNDAP) to visualization tools (Global Imagery Browse Services, OpenLayers).

Costs increase when preparing and curating data and when providing and maintaining access to additional tools, services, and related information. Large scale data storage and access services are offered commercially (e.g. Amazon Cloud) and from public and public-private service providers (e.g. EUDAT, Helix Nebula, Copenicus DIAS). In practice, few of these ideal aspects of interoperability are likely to be realised without brokering or mediation. The target of such brokering or mediation can be any of the four types of interoperability, as described under DMP-1. A major consideration is the extent to which it will be possible to persist such mediations for future re-use.

The cost of interconnecting geospatial solutions involving several Web Services and Web APIs from different vendors can be reduced by the use of recognized standard interfaces. The compatibility of such interfaces can be guaranteed by adopting implementations that passed a Compliance Testing and Certification. Compliance Testing tools, such as those available from the OGC, enable a vendor to test whether their products correctly implement a particular standard. Information about OGC Compliance Testing can be found on the OGC website (ogc.org/compliance). For the EU-INSPIRE directive, web-service validation tools are offered based on ISO and OGC standards; the INSPIRE Validator assesses distributed data against their conformance classes (FeatureTypes). This INSPIRE Validator can be found on the INSPIRE Website (https://inspire.ec.europa.eu/validator/home/index.html) An INSPIRE Monitoring System measures indicators to compare the characteristics and quality of different data providers and member states. NASA Earthdata also offers an equivalent validation tool.

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 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 standardized encodings brings benefits to the end user and limits the amount of time spent on transforming data, and therefore is a key to *interoperability*, as described under DMP-1.

4.2   Guidance on Implementation, with examples

The availability and acceptance of **syntactic encoding standards** are at a high level of maturity, and these standards cover the majority of data families that the GEO community uses routinely. Examples are sensor services defined by OGC SWE (2011), OPeNDAP service and NetCDF format (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 OGC Sensor Observation Services lagging seriously behind Web Mapping Services and the use of OPeNDAP and NetCDF (with or without the Climate and Forecast (CF) Metadata Conventions). The use of Python xarray and their related file formats (e.g, Zarr) is gaining traction in the image analysis community, stimulated by open code snippets that can be easily and remotely executed by Jupyter notebooks. Practitioners should select the open standards and implementations of these appropriate to their community, internal information technology platforms, and capabilities, as a preferred means of providing access to publicly available datasets.

Communities have also developed a portfolio of content standards in support of **schematic interoperability** (see Appendix B). For example, GML (OpenGIS® Geography Markup Language 2007) requires a GML application schema that ensures schematic interoperability, while GeoJSON (GeoJSON Format Specification 2016) is schemaless. Standards for raster data such as the GeoTIFF, NetCDF, HDF5, JPEG2000 rely on metadata to specify the schema and semantics. TimeSeriesML and O&M are examples of encodings of time series and sensor observations. Interoperability practices, for spatial datasets, whether vector datasets or raster datasets, are highly mature, and it is common for applications and web components to support a wide variety of data schema. Formats recognized by the GDAL/OGC libraries will facilitate exchange of data among many GIS products. 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**, as described under the Appendix B, Metadata for Discovery, 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 ontologies, vocabulary services, federated vocabulary services, FAIR vocabularies, and naming services for a wide variety of disciplines [Cox, 2021]. 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 available resources, and by fostering 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 vocabulary, ontology, linked data capability, and name service appropriate to the field of study rather than none at all.

Many observational data are distributed as tabular data (e.g. CSV). In this case, particular attention should be placed in describing the columns in a way that the variable meaning and the units of measurement are clearly stated. The O&M standard is an alternative to encode in-situ data that provides this information. This is mentioned again in DMP-4

4.3 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 dataset and/or service by submitting such a dataset 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). Metrics also might relate to the amount of data available in a well-defined and documented format (% of the whole data holding) and to the availability of data format specifications offered by a particular service.

4.4 Resource Implications of Implementation

Implementation requires the labor of experts to assess and identify encoding schemes that are appropriate for the communities served and the data products and services that are being disseminated. In cases where selected coding schemes are not based on standards that are openly accessible, there may be costs to acquire the standards that will be used or to acquire manuals or training on the use of the encoding schemes that have been selected for adoption. In addition to resources that are needed to encode in accordance with a selected schema, resources for verification of the encoding also will be needed. If software or services for verification are not available, they may need to be developed. The acquisition of services, software, and tools also may be necessary, when these are not provided as open source or as free of charge community services. For example, registry services, software procurement, and consulting services may be needed. In addition to the resources needed to adopt and implement encoding, resources, including labor, may be needed to maintain the encoding schemes that have been adopted as encoding schemas and standards evolve and practices change for how they are used. From this, we deduce that a truly interoperable environment can only be realized 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. Formats recognized by the GDAL/OGC libraries can be easier to produce due to the range of tools available using them.

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 Explanation of the Principle

Data documentation (metadata) should enable users and potential users to determine whether the data will meet their needs and help them to access, use, understand, and process the data. Usability of data is maximized when data documentation is complete to enable

understandability of the data and all appropriate elements of metadata are utilized. Partial documentation of data can happen in two intensities: 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 and to support the reproducibility of science. Standards facilitate the understanding of metadata between data providers and data users, either directly or via mediation technology.

When applicable, data producers should publish, in the peer-reviewed open 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.

5.2   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 19110, ISO 19157 (Standards ISO 2013), ISO 19158 (Standards ISO 2012), Dublin Core (Standards ISO-2 2009), Darwin Core, Directory Interchange Format (DIF), DataCite, STAC, 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, including data producers and distributors, 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. In the absence of a usable community standard, the documentation should describe the data, their quality, how they were produced, the instruments and variables employed, and how the data can be accessed and used.

One aspect that is often forgotten in metadata is the description of the data model (the schema) and the meaning of each element of the data (semantic interoperability). For example, if data is provided in a tabular format, the variable meaning and units of measure for each column should be provided in the documentation. In XML, an XML schema can provide validation of the format but does not provide descriptions on the meaning of the elements and attributes of the data. The JSON schema is more suitable for describing the elements and attributes in text but lacks the capability to specify units of measure.

When data is produced by scientists or other users without GIS skills, they can be overwhelmed by the number of details they are asked to provide in metadata. Separation of the concepts used in metadata is difficult for a non-expert. Tools to support the creation of metadata (such as GeoNetwork forms) with clear documentation and examples similar to the ones they are being produced can help a lot.

5.3 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, a.k.a. in a meaningful way. For example, a title cannot be expressive enough, a keyword could contain incorrect topics, the CRS could be wrong, a metadata field used to point to where the associated data can be found may point to the wrong file 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 outdated, or whether 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 level at which the associated data can be understood and used in a meaningful manner. Also, measure the extent to which metrics on Linked Data references and keywords have been selected from formal vocabularies.

5.4 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. Costs can be reduced dramatically if the metadata is produced at the same time the data is produced and by the same people that have the knowledge about the data and its generation process. Production of metadata can be included in the data processing chain and partially automated. If metadata is produced at a later stage or by different people, the effort in finding out every detail to be documented multiplies.

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   Explanation of the principle

Provenance information (a.k.a. lineage) is provided by data producers and stewards, as part of the metadata, to record sources of the data, any changes to the data, and the chain of custody of the data. Some provenance information can be captured automatically by the processing tools involved in a processing chain 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 (e.g. the role of each data source involved, other input parameters and the purpose of 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 a fundamental part of the reproducibility of a dataset result and contributes to an open knowledge approach. It also contributes to the trust and is often used as one of the first selection criteria.

Provenance is a necessary 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, evidence, describing the processing steps or the results of any tests that were conducted on the data, can provide an indication of the uncertainties of the data, as further described in DMP-6.

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 should include persistent links to such sources.

In case of inconsistencies with other data, provenance can help users identify a problem in a basic dataset or to improve it. Provenance information about other products can help to identify which products were derived from the affected dataset or which control, transformation or curation has been applied to the dataset. If complete enough, 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 also help to understand the efforts done to make a dataset time series homogenous. Provenance can be provided at different levels such as collection, dataset series, dataset, feature, attribute type, attribute etc. For example, this is useful to determine the source of a single feature of even a single attribute value in the case that a dataset is the result of merging features from different sources. Provenance at the dataset level is usually stored in the dataset metadata (that, in the case of GEOSS, is accessible by the Discovery and Access Broker) (DMP-1 and DMP-4) 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 (DMP-2). Assigning version numbers to each release and providing access to earlier versions of the dataset is also recommended, particularly if the earlier version has been published with a persistent identifier and could have been cited previously. Some metadata schemas, such as ISO and DataCite, offer capabilities for linking metadata entries on the basis of versioning and progressions in a series.

6.2   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. Descriptions of methodology and protocols are sometimes published separately and referred to via URL or persistent identifier, especially when they apply to multiple metadata records.
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 and provenance data 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.

6.3   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 19115-1 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, processing codes, 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.

6.4   Resource Implications of Implementation

This is part of the metadata process and the costs can be absorbed in this concept. There are three 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.
3. Implementing automated assessment of metrics.

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   Explanation of Principle

This principle is not favoring data that have better accuracy or less uncertainties. Instead, it focuses on the need for conducting a quality-control process to the data and sharing the results to enable use of the data, especially by individuals who were not involved in the creation or processing of the data. Data Quality is multidimensional and a data quality review should verify several components of the data including consistency, accuracy, and precision of values, completeness and correctness of documentation, and validity and fullness of metadata (Peer et al., 2014), as well as other aspects of the data, including uncertainty and any limitations on use. There are two moments where the data quality reviews and appreciation happens. The first is under the responsibility of the producer that should conduct it prior to data sharing and dissemination so that prospective user communities can determine the potential for using the data by consulting the results of the data quality review in a timely manner. 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. Results of data assessments also can inform decisions on whether to invest resources in the data. Secondly, the data users can report their experiences with the data and complement the producer quality with geospatial user feedback. User feedback can range from informal rating and comments to comprehensive quality metrics and usage reports. The absence of values for data quality indicators in metadata is an indication that a data quality review has not been conducted.

7.2   Guidance on Implementation, with examples

One or more tiers of data quality assessments 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 using terminology that is understandable by the community so that the results of the data quality assessment are available and usable. An assessment of the data also may be conducted as part of the peer-review of the article that describes the data. The result of this assessment may result in a human readable quality report as well as a set of formal quality indicators referring to one component of data quality and using previously documented methodologies. The ISO 19157 data quality standard provides a set of quality measures that can be done to the data as well as a way to document numerical and conformance results in the metadata. The QualityML is an initiative initiated by the EC FP7 GeoViQua project to create an extendable registry of quality measures in interoperable format and associated to permanent identifiers. This idea is now taken on board by the ISO 19157-3 standard.

The data quality concept is becoming broader and broader, as demonstrated in the DAM-NL work (http://www.dama-nl.org/wp-content/uploads/2020/11/3DQ-Dictionary-of-Dimensions-of-Data-Quality-version-1.2-d.d.-14-Nov-2020.pdf) where 60 dimensions of data quality were formalized. 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 and the location of processing codes, if applicable. 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. Officiators also should report needed corrections to ensure that they are addressed in subsequent data releases.

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 series level and the individual dataset file level. In consultation with the community, established practices, or standards, the data quality review officiator should define each criterion to be used for the review. In in-situ and in citizen science data could be necessary to report data quality at the observation label. The Sensor Thing API data model (based on O&M) includes this possibility. 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. The TRUST Principles for the data repositories include a Responsibility principle emphasizing that *"TRUSTworthy repositories take responsibility for the stewardship of their data holdings and for serving their user community"*. Responsibility is demonstrated by adhering to the designated community's metadata and curation standards, along with providing stewardship of the data holdings e.g. technical validation, documentation, quality control, authenticity protection, and long-term persistence."

Geospatial User Feedback is an OGC standard data model to encode quality related metadata directly provided by the users. Geospatial User Feedback is based on the idea of individual feedback elements that can include ratings, comments, publications, additional quality reports, usage reports and significant events conditioning the interpretation of the data. Geospatial User Feedback can affect the whole dataset or only a portion of it, and it is related to the dataset by its unique identifier (see DMP-10). Recently it has been proved that the same system could be also used to report new knowledge extracted from the data by users iteratively

7.3   Metrics to measure level of adherence to the principle

When assessing valuable scientific data products, determine whether 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 and measure result. Assess whether the results 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. Determine whether the officiator has resolved discrepancies between decisions of individual reviewers for a particular criterion to provide a decisive determination about the quality of the reviewed data. For example, determine whether the officiator has requested clarification from individual reviewers or requested a review by an additional reviewer to break a tie vote for any particular criterion.

Assess whether the value of the data quality indicator has been 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, determine whether the metadata states that the data quality review was not completed. If a particular criterion was not included in the data quality review, determine whether the indicator for that criterion states that the data quality review was not completed. Determine whether other indicators describe the quality assessment framework or the kind of review that was completed, such as internal review, metadata completeness, formal review for a peer reviewed article, and end-user feedback.

Determine if the data catalogue is connected to a Geospatial User Feedback implementation and users are providing user feedback, and this feedback is presented as a supplement of the official producer metadata.

7.4   Resource Implications of Implementation

Ideally, except for automated reviews and processes that produce multiple datasets, 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.

A Geospatial User Feedback implementation should be made available in the metadata portal. It could be a system that accompanies the data catalogue (such as the GeoNetwork user feedback plug-in) or it could be an external system that is connected directly by the client side with an open Web API (e.g. the NiMMbus system developed in the NextGEOSS project).

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   Explanation of the principle

Data are valuable assets for reuse and underpin the scholarly record. The preservation of data in digital format requires certain actions to be performed: this includes preservation planning, scheduled transformation of file-type to avoid obsolescence, system backup and recovery plans for the possibility of system failures and corruption, 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) (CCSDS, 2012). Repositories which

through their mission, organizational setup and business processes are able to fulfil these actions in a sustainable way, may qualify as Trustworthy 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, and that obsolete data formats and services are updated as required.

The ISO 19165 (www.iso19165.org) is also based on the ideas defined in the OAIS and provides a data model that enriches the ISO19115 metadata model with metadata particularly designed to serve geospatial digital libraries in their preservation tasks. It also provides a solution for the information package based on ISO 29500-2 Open Packaging Convention (being the MMZx format and implementation of it).

8.2   Guidance on Implementation, with examples

- Archived data refreshment: Periodically perform or automate 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.). Additional copies can also be created and stored in a simple human readable format;
- Archived data formats description: Provide formal description of old archiving formats to automate or allow the conversion to new standard formats, which will increase technical compatibility and reduce diversity of formats and interfaces between archives;
- Archived data duplication: Maintain identical copies of all archived data applying one of the security levels defined below:

  1. Dual copy in the same geographical location (but different buildings) to avoid data loss due to media degradation or obsolescence, or
  2. Dual copy in the same geographical location (but different buildings) based on different technology to avoid technology based principle failures, or
  3. Dual copy in two different geographical locations to safeguard the archive from external hazards (e.g. floods, other natural and technological hazards, etc.), or
  4. Dual copy in two different geographical locations, based on different technologies to avoid technology based principle failures.

- Archive system components migration (hardware): Perform periodical migration of archive system components to new hardware platforms.

Data contributed to GEOSS should be preserved for the long term and protected from loss for future use in trustworthy digital repositories (TDRs). The Core Trustworthy Data Repositories Requirements provides guidance for meeting such requirements as part of its certification criteria (CoreTrustSeal, 2016), which also can be used by 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.

8.3 Metrics to measure level of adherence to the principle

Recommended compliance levels for each of the types of evidence described 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:

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

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

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, these resources should be budgeted for in three- to five-year periods.

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

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?

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?

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

8.4 Resource Implications of Implementation

In the preparation and design of a dataset, some resources should be dedicated to elaborate a preservation plan that defines how and when the dataset will enter the operation phase and enter in the preservation phase conducted by a Core Trustworthy Data Repository. Some resources should be set aside to prepare for the preservation phase. The Core Trustworthy Data Repositories Requirements, described above, reflect the basic characteristics of trustworthy data repositories (CoreTrustSeal, 2016). The requirements represent a set of harmonized common criteria for certification of data repositories 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   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 datasets (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. While also pertinent to DMP- 7 *Data Preservation*, this principle 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. Verification should include routine tests for resolvability of persistent identifiers, readability, fixity, and provenance. Please note the provenance in preservation includes the aspects described in DMP-7 as well as all actions done to preserve, verify and upgrade the data in the preservation phase.

9.2   Guidance on Implementation, with examples

- **<u>Media readability and accessibility tests:</u>** Perform periodical test for media readability and accessibility on a representative set of the archived data.
- **<u>Archive content integrity:</u>** Periodically verify the integrity of the archive collection/content through integrity check on a representative set of the archived data.
- **<u>Data content integrity:</u>** 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).

9.3   Metrics to measure level of adherence to the principle:

Assess whether verification of readability and accessibility tests, integrity checking, and provenance capture, as described above, has been completed and documented. Determine whether certification has been obtained for compliance with the CoreTrustSeal (2016) requirements or with 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.

9.4   Resource Implications of Implementation

Estimating the cost in terms of resources for data and metadata verification has received much attention from many organizations (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 management and their economic impact on the organization.

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 Explanation of the principle

Curation, normally 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 the ingestion and the processing and interpretation of raw data. However, data processing organizations that preserve the integrity and authenticity of the data are well versed in software development, advancements in computing technology and processing algorithms. As a result, the practice of extracting more and more information from the available data has evolved naturally. This coincides with the key "social" and "scientific" goals of providing data to distinctive communities: long-term datasets 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 [identifying errors and problems] 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 activity in 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. When possible, automated techniques should be created to reduce the amount of manual review required for data and metadata.

Reprocessing can produce higher quality data 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 (e.g. in the creation of remote sensing high level products). Reprocessing also can change the output file format. **Format conversion** or **reformatting** might be an additional consequence not necessarily linked to reprocessing.

10.2 Guidance on Implementation, with examples

Updates and corrections to submitted datasets are encouraged. Records of updates and corrections should be maintained as well as the original data; summaries of updates should be

posted in the database, and users should be notified as part of the provenance information (DMP-5). 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 openly visible 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). However, such debates should not prevent implementation of correction policies and methodologies and openness of results to the designated communities along with continued access to the previous versions of the data, especially those that were provided a persistent identifier and possibly cited.

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 (this can be particularly true with the aggregation of more in-situ datasets or the inclusion of citizen science data); 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 datasets limit or enhance ultimate quality), geographic-bound or time-bound datasets 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 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.

10.3   Metrics to measure level of adherence to the principle

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 (T. Borzacchiello & M. Craglia (2011), Conway, Esther (2011)). 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 GEOBON 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.

Process-based metrics:

Metrics for institutional processes that guide updating, correction, and reprocessing should include:

- existence of a process (logic) for update and reprocessing of holdings
- existence of appropriate metadata structures to capture update and reprocessing information

10.4   Resource Implications of Implementation

Both updating and corrections, as well as reprocessing, are detailed, labor intensive, time-consuming, and prone to errors. Each reusable dataset or collection requires specific reprocessing steps or techniques appropriate for the specific dataset 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 considered necessary in many areas. Climate change related observations are the paradigmatic datasets that need reprocessing since a major difficulty in understanding past climate change is that most systems used to make the in-situ observations on which climate scientists now rely 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 ways of 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.

If the data is used frequently by many users (such as the long time series of remote sensing data) a re-processing of a collection can have an impact on users and the tools they use. The impact and cost could be reduced by careful planning and advertisement of the reprocessing plans in advance.

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   Explanation of the principle

Assigning a persistent, unique and resolvable digital identifier to data allows researchers and other users to communicate unambiguously about 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 contributions of data to published research.

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.

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 DMP.

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.

Persistent and resolvable identifiers are the basis to relate data, metadata and user feedback together as well as to set up relations among datasets in a distributed system of systems. This is well understood by the Linked Data community but still not completely implemented in geospatial data.

11.2   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 metadata on 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;
- **Maintaining arbitrary views of data**: For arbitrary views of data, such as those generated via a query, provide continuing access by versioning and timestamping the data and by storing executable queries to access the timestamped data and assigning persistent identifiers to those queries.

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 keywords 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]

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), and 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 monitoring 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. Additional information about these schemes and others is available from the DataCite Metadata Working Group (2016).

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
DOI	DataCite, Crossref, and others	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 organization holding the data and GEOSS should indicate clearly how the data should be cited by those using the data in published works.

11.3   Metrics to measure level of adherence to the principle

Measures of adherence are as follows:

1. Assigning appropriate, persistent, unique and resolvable identifiers to datasets 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.

11.4   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. Service providers often charge fees for assigning persistent identifiers. The data identifier assignments may be initiated automatically or manually by the archive and must be maintained so that the persistent identifier resolves to the current location of the data landing page, which should describe the data sufficiently to enable potential users to decide whether the data have the potential to meet their needs. The recommended citation for each data product should include the data product identifier. An alternative procedure that can help reduce cost and can also help solving the DMP-7 is to send the datasets to a well funded open data repository that generates a persistent identifier such as Zenodo or Pangaea.

Annex A

Terms and Definitions

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

Broker: A piece of software that 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.[5] [From DMP-1]

Catalogue: A collection of metadata about datasets. [From DMP-1]

Clearinghouse: 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, is collected in a standard format to facilitate query and consistent presentation across multiple participating sites. A clearinghouse uses readily

---

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

available Web technology for the client side and uses standards for the query, search, and presentation of search results to the Web client. A clearinghouse provides information about who is providing which authorized geoinformation for which application (GETIS).[6] [From DMP-1]

Cloud computing: On-demand availability of computer system resources, especially data storage (cloud storage) and computing power, without direct active management by the user. [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]

Curation[7]: 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: Interpretable 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 cube: a multi-dimensional ("n-D") array of values. Typically, the term datacube is applied in contexts where these arrays are massively larger than the hosting computer clouds. In the geospatial domain, an example could be a time series of harmonized Analysis ready data image data. [From DMP-1]

Data Curation: 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"[8] [From DMP-9]

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

Data Reprocessing: Data that resulted from the recreation of a given data product again, for example, with improved algorithms, to generate a newer version of the data product. [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

---

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

[7] Term not present in the OAIS glossary

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

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

Discovery Services: Services that make it possible to search for datasets and services on the basis of the content of the corresponding metadata and to display the content of the metadata.[9] [From DMP-1]

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

Encoding: The process of putting a sequence of characters (letters, numbers, punctuation, and certain symbols) into a specialized format for efficient transmission or storage. [From DMP-3]

Essential Variables: Variables known to be critical for observing and monitoring a given facet of the Earth system. [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[10]. [From DMP-9]

Format Conversion: Translation of digital content from one format to another to avoid obsolescence of the format and its associated software. Format conversions may also be performed to satisfy needs of users in different user communities[11]. [From DMP-9]

GeoJSON: Format for encoding a variety of geographic data structures in a JSON encoding. [From DMP-3]

Geospatial user feedback: Information about a resource directly provided by users, that can usually be used as a complement of the producer data quality as well as to generate cumulative knowledge on a dataset [From DMP-6]

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

---

[9]

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

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

[11] 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).

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, however, is to store 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 ensure 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]

Linked Data: Data that follows a set of design principles for sharing machine-readable interlinked data on the Web. When combined with Open Data (data that can be freely used and distributed), it is called Linked Open Data (LOD). The collection of Semantic Web technologies (RDF, OWL, SKOS, SPARQL, etc.) provides an environment where applications can query that data, draw inferences using vocabularies, etc.

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, the unavailability of the original authors or team that produced the data to respond to users' questions, 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]

Metadata: information describing datasets 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, e.g. 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 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: An identifier and its resolution are persistent in that some entity, when 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 conducted by reviewing data to assess their potential for use. [From DMP-6]

Queryable: an 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]

Reproducibility: The ability to replicate the data or results of a study. This 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-4]

Resolvable identifier: 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: 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,* as defined previously. [From DMP-5]

Unique Identity: The identifier that communicates unique information confirming that it refers to, and only to, a given object. [From DMP-10]

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

User Management System - a system that allows users to authenticate themselves and get authorized to access various resources such as applications, or data that an administration has granted access in advance to particular user.

Virtual Research Environment: an online system helping researchers to collaborate. Features usually include collaboration support, data analysis, visualization, or simulation environment. VREs have become important in fields where research is primarily carried out in teams which span institutions and even countries. [From DMP-1]

Web Application Programming Interface: An application programming interface that allows a web browser or a web app to communicate with a web server to manage web resources. It enables data transmission between one software product and another in the web. [From DMP-2]

Web-Service: A software system designed to support interoperable machine-to-machine interaction over a network. [From DMP-2]

Annex B

Levels of Interoperability

Achieving interoperability is a fundamental task for a machine to machine communication and the interaction with humans to get the exact meaning of the content and use of data. Interoperability has to apply on different levels. The most important levels referenced in this document are:

Syntactic Interoperability 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 is an example of additional refinement, in that requests for derived data ("offerings"), e.g. 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) and a clear ongoing trend (OGC Earth Observation Cloud Platform Concept Development Study Report 2021). Definition of the parameters depend to some extent on semantic interoperability and conventions.

Schematic Interoperability defines the structure (application schema, data model) in which the data will be offered by a service. For many applications, this schema is critical for correct binding, but schemas are likely to vary within a common framework depending on specific applications (Hugo, 2008).

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. Semantic Interoperability attempts to establish agreed ontologies, common vocabularies (Cox et al 2021), and terminology frameworks such as "essential variables" (OOPC 2015), are all designed to address a common understanding of meaning.

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.

Legal Interoperability allows discovery, integration, and use of data by communicating rights and restrictions that are associated with the data so that they are understood by people and can be interpreted by machines. For example, if data have been encoded to indicate that there are no restrictions on use, the data are eligible to appear in the results retrieved when searching for data that are not restricted due to intellectual property rights or restrictions that have been assigned to the data.

Agreement on a workable set of syntactic (service), schematic, and semantic standards for the typical data domains in use by the community can be an appropriate scalable approach.

Annex C

References

For the introduction:

GEO Data Branding [Online] Available from: https://geolabel.info

GEO Data Sharing Principles [Online] Available from: https://www.earthobservations.org/dswg.php

GEOSS Yellow pages service as registration for new Data Providers. [Online] Available from: https://www.geoportal.org/yellow-pages

Lin D, Crabtree J, Dillo I, Downs RR, Edmunds R, Giaretta D, De Giusiti M, L'Hours H, Hugo W, Jenkyns R, Khodiyar V, Martone M, Mokrane M, Navale V, Petters J Sierman B, Sokolova DV, Stockhause M, Westbrook J. 2020. The TRUST Principles for Digital Repositories. Scientific Data 7, 144. https://doi.org/10.1038/s41597-020-0486-7

Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, Bouwman J. The FAIR Guiding Principles for scientific data management and stewardship. Scientific data. 2016 Mar 15;3(1):1-9. https://doi.org/10.1038/sdata.2016.18

For DMP-1

CEOS WGISS Data Stewardship Interest Group, 2017: WGISS Data Management and Stewardship Maturity Matrix. Doc. Ref: CEOS.WGISS.DSIG.DSMM. Version: v1.0 26 September 2017.

Core Trustworthy Data Repositories Requirements. (2016). CoreTrustSeal. [Online] Available from: https://www.coretrustseal.org/

FGDC Service status checker: [Online] Available from: https://statuschecker.fgdc.gov/dashboard/geossd_114

GEOSS Yellow Pages [Online] Available from: https://www.geoportal.org/yellow-pages

GEOSS Data-Core [Online] Available from: https://www.earthobservations.org/documents/dswg/Annex%20IX%20-%20Instructions%20for%20GEOSS%20Data%20Providers%20How%20to%20place%20tags%20in%20the%20Metadata%20for%20GEOSS%20Data_CORE.pdf

OGC Earth Observation Cloud Platform Concept Development Study Report (2021) [Online] Available from: https://docs.ogc.org/per/21-023.html

Palaiologk, A. S., Economides, A. A., Tjalsma, H. D., & Sesink, L. B. (2012). An activity-based costing model for long-term preservation and dissemination of digital research data: the case of DANS. International journal on digital libraries, 12(4), 195-214. [Online] Available from: http://dx.doi.org/10.1007/s00799-012-0092-1

Rauber, A., Asmi, A., van Uytvanck, D., and Pröl, S. 2015. Data Citation of Evolving Data: Recommendations of the Working Group on Data Citation (WGDC). Research Data Alliance. [Online] Available from: https://rd-alliance.org/system/files/documents/RDA-DC-Recommendations_151020.pdf

White Paper: Mechanisms to Share Data as Part of GEOSS Data-CORE1White Paper: Mechanisms to Share Data as Part of GEOSS Data-CORE, [Online] Available from: https://www.earthobservations.org/documents/dswg/Annex%20VI%20-%20%20Mechanisms%20to%20share%20data%20as%20part%20of%20GEOSS%20Data_CORE.pdf

Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data, 3. [Online] Available from: https://dx.doi.org/10.1038/sdata.2016.18

For DMP-2

CEOS WGISS Data Stewardship Interest Group, 2017: WGISS Data Management and Stewardship Maturity Matrix. Doc. Ref: CEOS.WGISS.DSIG.DSMM. Version: v1.0 26 September 2017. [Online] Available from: http://ceos.org/document_management/Working_Groups/WGISS/Interest_Groups/Data_Stewardship/White_Papers/WGISS%20Data%20Management%20and%20Stewardship%20Maturity%20Matrix_Final.xlsx

Geoserver [Online] Available from: https://geoserver.org/

OGC. (2021) *OGC®* API - Environmental Data Retrieval [Online] Available from: http://www.ogc.org/standards/ogcapi-edr

OGC. (2019) *OGC® API - Features - Part 1: Core* [Online] Available from: http://www.ogc.org/standards/ogcapi-features

OGC. (2019) *OGC® API - Maps - Part 1: Core* [Online] Available from: http://www.ogc.org/standards/ogcapi-maps

OGC. (2019) *OGC® API - Tiles - Part 1: Core* [Online] Available from: http://www.ogc.org/standards/ogcapi-tiles

OGC. (2021) *OGC®* SensorThings API Part 1: Sensing Version 1.1 [Online] Available from: https://www.ogc.org/standards/sensorthings

OGC (2012) *OGC®* Sensor Observation Service Interface Standard *Version 2.0* [Online] Available from: https://www.ogc.org/standards/sos

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

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

OPeNDAP: [Online] Available from: https://www.opendap.org/

OpenLayers: [Online] Available from: https://openlayers.org/

For DMP-3

Fulker, D. and Gallagher, J. (2013) Extending OPeNDAP to Offer Remapping Services. *Geophysical Research Abstracts*. EGU2013-10236, 2013 EGU General Assembly 2013. 15. [Online] 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*. 2 (2008). p.25. [Online] 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. (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. (2010) *OGC® Observations and Measurements 2.0 (O&M 2.0)* [Online] Available from: https://www.ogc.org/standards/om

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/

For DMP-4

ISO 19115-1: 2014 Geographic information -- Metadata -- Part 1: Fundamentals. [Online] Available from:
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

ISO 19115-2: 2009 Geographic information -- Metadata -- Part 2: Extensions for imagery and gridded data. [Online] Available from: http://www.iso.org/iso/catalogue_detail.htm?csnumber=39229http://www.iso.org/iso/catalogue_detail.htm?csnumber=39229

ISO 19139:2007 Geographic information -- Metadata -- XML schema implementation [Online] Available from: http://www.iso.org/iso/catalogue_detail.htm?csnumber=32557

ISO 19157:2013 Geographic information -- Data quality [Online] Available from: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=32575 ; https://earthdata.nasa.gov/standards/preservation-content-spechttp://www.iso.org/iso/catalogue_detail.htm?csnumber=32557

Dublin Core [Online] Available from: http://dublincore.org/http://dublincore.org/

Darwin Core [Online] Available from: http://rs.tdwg.org/dwc/http://rs.tdwg.org/dwc/

Directory Interchange Format: (DIF) [Online] Available from: https://earthdata.nasa.gov/standards/directory-interchange-format-dif-standardhttps://earthdata.nasa.gov/standards/directory-interchange-format-dif-standard

Linked data: [Online] Available from: https://www.w3.org/standards/semanticweb/data

CF Metadata Conventions: [Online] Available from: http://cfconventions.org/

GeoViQua [Online] Available from: http://www.creaf.uab.cat/projectes/geoviqua/

GEOLabel [Online] Available from: www.geolabel.info

For DMP-5

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

GeoNetwork4MetaViz – Enabling an open source metadata catalogue to visualize lineage information (2015). [Online] Available from: 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. [Online] Available from: http://earthcube.org/document/2015/geoscience-papers-future

For DMP-6

Callahan S. 2015. Data without Peer: Examples of Data Peer Review in the Earth Sciences. D-Lib Magazine, 21(1/2). [Online] Available from: 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. [Online] Available from: http://dx.doi.org/10.2218/ijdc.v9i1.317 .

QualityML: Quality Indicators Dictionary and Markup Language. [Online] Available from: https://www.qualityml.org/

Zabala A., Maso J., Bastin L., Giuliani G., Pons X. (2021) Geospatial User Feedback: How to Raise Users' Voices and Collectively Build Knowledge at the Same Time. ISPRS International Journal of Geo-Information. Vol.. DOI: 10.3390/ijgi10030141. [Online] Available from: https://www.mdpi.com/2220-9964/10/3/141

OGC. (2021) *OGC® SensorThings API Part 1: Sensing Version 1.1* [Online] Available from: https://www.ogc.org/standards/sensorthings

OGC. (2016) *OGC® Geospatial User Feedback Standard: Conceptual Model* [Online] Available from: http://www.opengeospatial.org/standards/guf

OGC. (2010) *OGC® Observations and Measurements 2.0 (O&M 2.0)*: [Online] Available from: https://www.ogc.org/standards/om

For DMP-7

Consultative Committee for Space Data Systems (CCSDS). 2012. Reference Model for an Open Archival Information System. Magenta Book CCSDS 650.0-M-2. [Online] Available from: https://public.ccsds.org/Pubs/650x0m2.pdf (Also available as ISO 14721:2012)

ISO 19165-1 Geographic information - Preservation of digital data and metadata - Part 1: Fundamentals information website. [Online] Available from: https://www.iso19165.org

For DMP-8

CEOS EO Data Preservation Guidelines.

Keeping Research Data Safe (KRDS) Project. [Online] Available from: http://www.beagrie.com/krds.php

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. [Online] Available from: http://www.dcc.ac.uk/resources/how-guides/appraise-select-data .

For DMP-9

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

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

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).

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), [Online] Available from: https://publications.jrc.ec.europa.eu/repository/handle/JRC66935?mode=full

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).

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. [Online] Available from: http://www.niso.org/publications/isq/2013/v25n03/conway/

IGS: the International GNSS Service GNSS: Global Navigation Satellite System [Online] Available from: http://acc.igs.org/reprocess.html

Reprocessing of Multi-channel Seismic-Reflection Data Collected in the Beaufort Sea. [Online] Available from: http://pubs.usgs.gov/of/2000/ofr-00-460/

Ocean Color and SST Processing History. [Online] Available from: https://oceancolor.gsfc.nasa.gov/reprocessing/

Reprocessing Complete for MODIS Aqua and Terra LST&E V6 Data Products. [Online] Available from: https://lpdaac.usgs.gov/news/reprocessing-complete-for-modis-aqua-and-terra-lste-v6-data-products/

Sea ice data reprocessing. [Online] Available from: http://nsidc.org/data/nsidc-0508

Climate Data, Sea Ice Concentration Data Reprocessed From SSMR & SSMI, by Eumetsat. [Online] Available from: https://climatedataguide.ucar.edu/climate-data/sea-ice-concentration-data-reprocessed-ssmr-ssmi-eumetsat

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

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

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

Reprocessing of global vegetation images. [Online] Available from: http://proba-v.vgt.vito.be/content/reprocessing-proba-v-data-finalized .

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.

For DMP-10

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

ESIP (2012). "Data Citation Guidelines for Data Providers and Archives". doi:10.7269/P34F1NNJ, [Online] Available from: http://commons.esipfed.org/node/308.

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

San Francisco Declaration on Research Assessment (DORA) [Online] Available from: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. [Online] Available from: https://dx.doi.org/10.7717/peerj-cs.1

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

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

DCC Guide. How to Cite Datasets. [Online] Available from: http://www.dcc.ac.uk/resources/how-guides/cite-datasets

DataCite Metadata Working Group. (2016) DataCite Metadata Schema for the Publication and Citation of Research Data v 4.0. Datacite. [Online] Available from: https://schema.datacite.org/

Dryad DOI Usage. [Online] Available from: http://wiki.datadryad.org/DOI_Usage

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

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

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 [Online] Available from: https://dx.doi.org/10.7717/peerj-cs.1

Annex D

Acronyms and Abbreviations

ARD – Analysis Ready Data

ARK – Archival Resource Key

ATS – Abstract Test Suite

CCSDS – Consultative Committee for Space Data Systems

CEOS – Committee on Earth Observation Satellites

CF – Climate and Forecast (cf convention)

CNR – National Research Council

CNRI – Corporation for National Research Initiatives

CSV – Comma-Separated Values

CSW – Catalogue Service for the Web - OGC Standard

DAB – Discovery and Access Broker - Component of the GEOSS platform

DCC – Digital Curation Centre

DIF – Directory Interchange Format

DMP – Data Management Principle

DOI – Digital Object Identifier

DORA – Declaration On Research Assessment

DRI – Decision Ready Information

DSA – Data Seal of Approval

EBV – Essential Biodiversity Variables

ECV – Essential Climate Variables

FAIR – Findable, Accessible, Interoperable, Reusable

FGDC – Federal Geographic Data Committee

FTP – File Transfer Protocol

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 - OGC Standard

GUF – Geospatial User Feedback - OGC Standard

HDF5 – Hierarchical Data Format Version 5 - OGC Standard

HTML – Hyper Text Markup Language

HTTP – HyperText Transfer Protocol

HTTPS – HyperText Transfer Protocol Secure

HW – Hardware

ICSU – International Council for Science

IEEE – Institute of Electrical and Electronics Engineers

ISO – International Organization for Standardization

INSPIRE - Infrastructure for Spatial Information in Europe

JRC – Joint Research Centre

JSON – JavaScript Object Notation

MIME – Multipurpose Internet Mail Extensions - Internet IETF Standard

NAS – National Academy of Sciences

NCSES – National Center for Science and Engineering Statistics

NetCDF – Network Common Data Form - OGC Standard

NOAA – National Oceanic and Atmospheric Administration

O&M – Observations and Measurements

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

QualityML – Quality Indicators Dictionary and Markup Language

SDI – Spatial Data Infrastructure

SensorML – Sensor Model Language - OGC Standard

SSO – Single Sign-On

STAC – SpatioTemporal Asset Catalog

STI – Science, Technology, and Innovation

SW – Software

SWE – Sensor Web Enablement - OGC Standard

TDR – Trusted Digital Repository

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

TOPEX – Ocean Topography Experiment

TRUST – Transparency, Responsibility, User focus, Sustainability, and Technology

URI – Uniform Resource Identifier - Internet IETF Standard

URL: Uniform Resource Locator - Internet IETF Standard

WaterML – Water Markup Language - OGC Standard, part 1 and 2 endorsed by WMO

Web API – Web Application Programming Interface

WDS – World Data System

WMO – World Meteorological Organization

WMS – Web Map Service - OGC Standard

WMTS – Web Map Tile Service - OGC Standard

WPS – Web Processing Service - OGC Standard

XML – Extensible Markup Language - W3C Standard

XSD – XML Schema Definition - W3C Standard

Annex E

Summary of Changes

This document is a revision of the original document "Data Management Implementation Guidelines" presented in the GEO-XII plenary in November 2015 (https://www.earthobservations.org/documents/dswg/201504_data_management_principles_long_final.pdf). This appendix describes significant changes that this version introduces.

Changes that apply to the whole document:

- The terms/keywords were found difficult to maintain and the subsection including them in the description of each DMP principle was removed. Instead, a common list of terms and definitions as an Appendix A is provided.
- For each DMP, the corresponding label icon is included.
- For each DMP, cross-references and links to other principles are added.

Changes by section:

Introduction

- A new paragraph contextualizing the DMP with similar initiatives such as the FAIR and TRUST principles are included.
- A new paragraph clarifying that the GEO DMP are applicable to raw, data ARD and DRI is added.
- A new paragraph to highlight the existence of the GEO Data Branding website for self-assessment and the GEOSS Yellow Pages initiatives for data providers is included.

DMP-1: Metadata for Discovery

Explanation of the principle

- Relation between DMP-1 and FAIR principles is added.
- Levels of interoperability are now considered.
- Discussion on XML schemas is included.

Guidance on Implementation

- For each mentioned type of metadata element, an example is provided.
- The Core Trustworthy Data Repositories Requirements (CoreTrustSeal, 2016), or ISO 16363 certifications for data discovery replaces the World Data System and the Data Seal of Approval.
- It is suggested that the catalogue should provide access to resources in HTML in a way that common search engines can index their resources (e.g. using schema.org tags).
- The fact that metadata should indicate conformity to the GEOSS Data-CORE (from GEO DSP) is added.

Metrics to measure

- The World Data System and the Data Seal of Approval is included as a possible metric to measure the level of adherence to the principle.

- "Projects" subsection was removed as it became obsolete and difficult to maintain.
- Some service checkers are included such as the CEOS WGISS 2017 and the GEOSS Discovery and Access Broker (GEOS DAB) one.

Resources

- The knowledge that XML generators and validators of metadata schemas can reduce costs and time is added
- Examples of data management organizations that have estimated the costs related to the metadata discovery are provided

DMP-2: Online Access

Explanation of the principle

- The datacubes and Virtual Research Environments (VRE) are included as a complement to the Web APIs
- Examples for some types of online services or formats were included (e.g. Could Optimized GeoTIFF (COG))
- New types of in place processing data services that provide access to data too are included: e.g. OGC APIs, Jupiter Notebooks and the previously mentioned VRE.

Guidance on Implementation

- New standards for online data access are included: OGC APIs, NetCDF, HDF5, OPeNDAP and COG.

Resources

- A new paragraph includes examples of Compliance Testing tools available for online data access, such as the OGC compliance service, the INSPIRE validator, or NASA Earthdata validator.

DMP-3: Data Encoding

Explanation of the principle

- Since the levels of interoperability are cited by some DMPs, the description of the different levels of interoperability was moved to Appendix B.

Guidance on Implementation

- A paragraph specifying the importance for encoding observational data (in situ measurements) that are commonly distributed as tabular data (e.g. CSV) is included
- The use of formats recognized by OGC/GDAL libraries is suggested as a guarantee for open source solutions to recognize and read the data encoding.

DMP-4: Data Documentation

Guidance on Implementation

- A new paragraph highlighting the importance of data models and schemas for describing and documenting data, such as XML and JSON schemas is added.

- A new paragraph suggesting that the convenience to use metadata editors (such as the use of GeoNetwork forms) and templates for data documentation can be useful.

DMP-5: Data Traceability

Explanation of the principle

- A new sentence is included recommending the assignment of a different version numbers for each dataset release.

Resources

- A new associated cost was added as "Implementing automated assessment of metrics".

DMP-6: Data Quality Control

Explanation of the principle

- A sentence that highlights the differences between data quality and data uncertainty is included.

Guidance on Implementation

- The possibility of data users reporting their experiences with the data to complement the producer quality with geospatial user feedback is added.
- It is suggested that in some cases (such as in-situ and in citizen science data) it could be necessary to report data quality at the observation label, in the same way that it is included in the OGC Sensor Things API data model.
- DMP-6 is aligned with TRUST principles.
- A new paragraph describing the Geospatial User Feedback OGC Standard as another quality control mechanism is added.

Metrics to measure

- The need to measure feedback provided from end-users (from Geospatial User Feedback implementation) as part of the official producer metadata is added.

Resources

- Suggestions on how to implement a Geospatial User Feedback system in metadata portals are added.

DMP-7: Data Preservation

Explanation of the principle

- A new paragraph that refers to the ISO 19615-1 as a solution for geospatial digital preservation is added.

Guidance on Implementation

- Archived data refreshment, archived data formats description, archived data duplication and archive system components migration (hardware) as ways to implement data preservation are moved from DMP-8

DMP-8: Data and Metadata Verification

Guidance on Implementation

- The content regarding the archived data refreshment, archived data formats description, archived data duplication and archive system components migration (hardware) was moved to DMP-8.

Metrics to measure

- The core CoreTrustSeal (2016) is added as new standard to assess the level of adherence of DMP-8.

Resources

- The generic cost models workflow (1-Identifying resource costs and activities; 2-Identifying resource costs and activities) was removed as it provided much more detail than the other DMPs

DMP-9: Data Review and Reprocessing

Metrics to measure

- The process-based metrics section is converted into a paragraph and focus on institutional processes that guide updating, correction, and reprocessing.

Resources

- A new paragraph highlighting the impact and costs of data reprocessing, and how it can be reduced if the users are advertised in advance is added.

DMP-10: Persistent and Resolvable Identifiers

Guidance on Implementation

- Maintaining arbitrary views of data is added as a key function for persistent identifiers.

Resources

- The section is extended with recommendations and resources to maintain and/or generate persistent data identifiers.

Others

- Appendix B is included describing the levels of interoperability previously in DMP-3. This appendix is referenced from several DMPs.
- Appendix E (this appendix) is produced to detail the most important changes done from the original version.

Annex F

Acknowledgement of Contributions

We are grateful for the contributions to these guidelines from the following contributors, who are listed in alphabetical order, implying no order of priority or responsibility.

Under the GEO Task Force 2015/18

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, Robert S.	ICSU	Author
de La Beaujardière, Jeff	US	Reviewer
De Lathouwer, Bart	OGC	Author, DMP-1 Lead
Downs, Robert R.	ICSU	Author, DMP-6 Lead, Editor
Duerr, Ruth	ESIP	Author
Goldstein, Justin	ESIP, USGCRP	Reviewer
Haslinger, Florian	EPOS	Reviewer
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 and reviewer
Mayernick, Matthew	ESIP	Reviewer
Mokrane, Mustapha	WDS	Author, DMP-7 Lead
Moreno, Richard	CEOS	Author, DMP-2 Lead
Peng, Ge	ESIP	Reviewer
Ramapriyan, Hampapuram K.	ESIP	Reviewer
Robinson, Erin	ESIP	Author
Stein, Alfred	The Netherlands, ITC	Reviewer
Wolfe, Robert	ESIP, USGCRP	Author

Under the GEO Data Working Group (TOR), Subgroup on Data Sharing & Data Management Principles 2021/22

Name	Organization	Role
Brobia, Alba	Spain, CREAF	Reviewer, Editor
Bye, Bente Lilja	Consultant	Reviewer
de La Beaujardière, Jeff	US	Reviewer
Downs, Robert R.	ICSU	Reviewer, Author, Editor
Goldstein, Justin	ESIP, USGCRP	Reviewer
Name	Organization	Role
Haslinger, Florian	EPOS	Reviewer

Iglesias, Jose Miguel Rubio	EEA	Reviewer
Maso, Joan 🆔	Spain, CREAF	Reviewer, Author
McMahon, Ethan	WRI	Reviewer, Coordinator
Robinson, Erin	ESIP	Author
Schubert, Chris 🆔	TU Wien	Coordinator, Reviewer, Editor
Stein, Alfred	The Netherlands, ITC	Reviewer
Trumpy, Eugenio	CNR	Reviewer
Voidrot, Marie-Francoise	Open Geospatial Consortium	Reviewer
Wolfe, Robert	ESIP, USGCRP	Author