GEOSS Data Management Principles:
Importance and Implementation

Alex de Sherbinin / Associate Director / CIESIN, Columbia University
Gregory Giuliani / Lecturer / University of Geneva
Joan Maso / Researcher and GIS developer / CREAF
“Ensure that data and information of different origin and type are comparable and compatible, facilitating their integration into models and the development of applications to derive decision support tools.”

GEO Strategic Plan 2016-2025

- Interoperability
- Quality (Fitness for use)
- Trustworthiness
GEOSS Data Management Principles (DMPs)

- Build on GEOSS Data Sharing Principles
- Set standard for good data (management) practices

GEOSS DMPs Approved in April 2015:
- Long version
- Condensed version

GEOSS DMPs Implementation Guidelines
Endorsed in November 2015
GEOSS DMPs

10 Principles in 5 categories

1. Discoverability
2. Accessibility
3. Usability
4. Preservation
5. Curation
GEOSS DMPs condensed

**Earth observations** will be **catalogued** or otherwise advertised on the internet so that they can be **discovered**, and will be **accessible** online using **open-standard encodings and services**. Data and services will be comprehensively **documented** using international or community-approved standards, and to the extent possible, peer-reviewed publications, so that users can understand and make use of the data. Metadata will include **access and use conditions**, the **results of quality control procedures**, and **provenance statements** indicating the origin and processing history of the dataset or product. Data and associated metadata will be **protected from loss** and periodically **verified to ensure integrity, authenticity and readability**. **Corrections and updates** to data and metadata records will be performed as required. Finally, **persistent identifiers** will be assigned to data so that they can be tracked and cited and data providers can be acknowledged.
Implementation of DMPs

“A priority mission for GEO is to encourage the implementation of the Principles (DMPs) by organizations contributing to GEOSS”

GEO Strategic Plan 2016-2025
GEOSS DMPs Implementation Guidelines

For each DMP:
• Terms used to describe the principle and its implementation
• Explanation of the DMP
• Guidance on implementation with examples

November 2015
Implementation guidelines uptake?

**Monitoring** DMPs implementation

1. **GEOSS Data Providers**
   a) Data Repositories Certification
   b) Status checker...

2. **Dataset/Collection**
   a) Data fitness for use: Certification, DMP Labels...
   b) User feedback...
From Principles to Implementation

WDS Data Sharing Principles

Purpose
The International Council for Science - World Data System (ICSU-WDS) aims to promote equitable access to quality-assured scientific data, data services, products and information towards long-term data stewardship. Furthermore, ICSU-WDS is committed to foster agreed-upon data standards and conventions, and providing mechanisms to facilitate data management. As the leading international, multidisciplinary organization in the provision of ICSU-WDS has adopted Data Sharing Principles to advance its goals.

The Principles express core ethical commitments that are operationalized in WDS Core and Network Members. It is anticipated that existing organizational policies align with the Principles. Partners and Associate Members are not subject to certification and therefore are not bound by them, but they are encouraged to do so. The Principles embody the spirit of open science and diverse communities of data producers and data users, and thus could be adopted science for the public good.

2009-15

DOI: 10.5281/zenodo.34354

Data Principles

For further data principles are recommended for adoption by the Belmont Forum e-Infrastructure and Data Management projects.

- Durable: Through verifiable and digital means, with data access and user conditions, including licensing, usage, and data stewardship requirements. Data should be permanent, unique, and recognizable identifiers should be maintained.
- Identifiable: through verifiable and digital means, with data access and user conditions, including licensing, usage, and data stewardship requirements. Data should be permanent, unique, and recognizable identifiers should be maintained.
- Shareable: Data should be sharable through verifiable and digital means, with data access and user conditions, including licensing, usage, and data stewardship requirements. Data should be permanent, unique, and recognizable identifiers should be maintained.
- Interoperable: Data should be interoperable through verifiable and digital means, with data access and user conditions, including licensing, usage, and data stewardship requirements. Data should be permanent, unique, and recognizable identifiers should be maintained.
- Accessible: Data should be accessible through verifiable and digital means, with data access and user conditions, including licensing, usage, and data stewardship requirements. Data should be permanent, unique, and recognizable identifiers should be maintained.
- Reusable: Data should be reusable through verifiable and digital means, with data access and user conditions, including licensing, usage, and data stewardship requirements. Data should be permanent, unique, and recognizable identifiers should be maintained.

DOI: 10.1038/sdata.2016.18

Comment: The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson et al.

There is no single model to ensure the long term sustainability of scientific data. A diverse set of methodologies—ranging from academia, industry, funding agencies, and scholarly publishers to name a few—exist to manage and ensure a sustainable and responsible use of principles that are: 

- Open (Open access to data and associated documentation)
- Findable (Findability and discoverability)
- Accessible (Availability and accessibility)
- Interoperable (Interoperability and interoperability)
- Reusable (Reusability and reusability)

DOI: 10.1038/sdata.2016.18

www.nature.com/sdata

2016
Practical Implementation

Trustworthy Data Repositories (TDRs)

- Formal Certification ISO 16363
  - (0 Repositories, 100+ requirements)

- Extended Certification
  - (2 Repositories, 34 requirements)

- CoreTrustSeal Certification
  - (150+ Repositories, 16 requirements)
Core Trustworthy Data Repository (TDR) certification

Catalogue of requirements (16):

Context
1. Organizational infrastructure (6)
2. Digital object management (8)
3. Technology (2)

Applicant feedback

Certification procedures:
1. Self-assessment: documented public evidence + compliance level
2. Peer-review (2 reviewers)
3. Renewal (3 years)

2009-15
CORE TDR Requirements

Context
Repository type, designated community, Level of curation performed, Outsource partners, Impact

Organizational infrastructure
• **R1.** DR has an explicit mission to provide access to and preserve data in its domain.
• **R2.** DR maintains all applicable licenses covering data access & use & monitors compliance
• **R3.** DR has a continuity plan to ensure ongoing access to and preservation of its holdings.
• **R4.** DR ensures that data are created, curated, accessed, and used in compliance with disciplinary and ethical norms.
• **R5.** DR has adequate funding and sufficient numbers of qualified staff managed through a clear system of governance
• **R6.** repository adopts mechanism(s) to secure ongoing expert guidance and feedback (either in-house, or external, scientific guidance).
CORE TDR Requirements

Digital object management

• **R7.** DR guarantees the integrity and authenticity of the data.
• **R8.** DR accepts data & metadata based on defined criteria to ensure relevance and understandability for users.
• **R9.** DR applies documented processes and procedures in managing archival storage of the data.
• **R10.** DR assumes responsibility for long-term preservation and manages this function in a planned and documented way.
• **R11.** DR has appropriate expertise to address technical data and metadata quality sufficient to make quality evaluations.
• **R12.** Archiving takes place according to defined workflows from ingest to dissemination.
• **R13.** DR enables users to discover the data and refer to them in a persistent way through proper citation.
• **R14.** DR enables reuse of the data over time, ensuring that appropriate metadata support the understanding and use of the data.
CORE TDR Requirements

Technical infrastructure

• **R15.** DR functions on well-supported operating systems and other core infrastructural software and is using hardware and software technologies appropriate to the services it provides to its Designated Community.

• **R16.** The technical infrastructure of the repository provides for protection of the facility and its data, products, services, and users.
### Mapping CORE TDR to GEOSS DMPs

<table>
<thead>
<tr>
<th>DMP label</th>
<th>Label</th>
<th>Usability</th>
<th>Preservation</th>
<th>Curation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discoverable</td>
<td>1</td>
<td>D</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accessible</td>
<td>2</td>
<td>A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard encoding using</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Well documented metadata</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Traceable</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quality documented</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Preserved</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Periodically verified</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reviewed and refreshed</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tagged with permanent ID</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**R1.**

**R2.**

**R3.**

**R4.**

**R5.**

**R6.**

**R7.**

**R8.**

**R9.**

**R10.**

**R11.**

**R12.**

**R13.**

**R14.**

**R15.**

**R16.**
Myths about Core TDR Certification

• Costly!
• Pass or fail!
• Difficult for small data providers...
• Time consuming!
• Not yet recognized?
Thank you!
GEOSS DMPs in detail

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

Accessibility
• DMP-2: Data will be accessible via online services, including, at minimum, direct download but preferably user-customizable services for visualization and computation.
GEOSS DMPs in detail

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.

• 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.
Usability (ctd.)

- 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.
- 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.
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.

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

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.

- DMP-10: Data will be assigned appropriate persistent, resolvable identifiers to enable documents to cite the data on which they are based and to enable data providers to receive acknowledgement of use of their data.
Alex de Sherbinin  adesherbinin@ciesin.columbia.edu
Gregory Giuliani  gregory.giuliani@unige.ch
Joan Maso  joan.maso@uab.cat
making data work for our planet

GEOSS/HACK 2018

classified for our planet

powered by GEOSS

#GEOSSHACK18
bit.ly/GEOSSHACK18