Linking the GEO Data Sharing and Data Management Principles to other Reference Lifecycles and Principles

Karl Benedict - University of New Mexico / Earth Science Information Partners

2022-May-25

GEO Working Group: Data Working Group (Data-WG)
Subgroup: *Data Sharing and Data Management Principles* (Data-WG/DSDMP). In particular, the following subgroup members provided invaluable input into the development of the approach used in the development of this analysis: Bente Lija Bye, Eugenio Trumpy, Chris Jarvis, Jose Miguel Rubio Iglesias, Ethan McMahon, Robert R Downs, Chris Shubert, Sebastian Claus, Paula De Salvo

This document summmarizes the process and outputs of an analytic approach taken to increase the Data-WG and broader GEO community understanding of the relationship between the GEO Data Sharing (pg 11) and Data Management Principles (pg. 10) (referred to as *DSDMP* hereafter) and other data lifecycle models and reference principles (referred to as *reference frameworks* hereafter) that have been developed since the development of the GEO principles as part of the 2016-2025 GEO Strategic Plan. This document presents both a narrative description of the process followed in developing the initial connections between the DSDMP and reference frameworks, and summary visualizations of the preliminary data.

The work presented herein was initiated in July 2021 within the Data-WG/DSDMP through a discussion within the subgroup to identify an initial set of reference frameworks to focus on in the identification of connections between the DSDMP and those reference frameworks. The identification of these connections was intended to serve three purposes:

- Identify gaps in the coverage by DSDMP concepts of elements of the reference frameworks
- Inform discussions for further development of the DSDMP with specific insights gained from the process of gap identification
- Enable enhanced communication of the DSDMP to audiences familiar with the reference frameworks through communication of the identified connections between the frameworks with which they are familiar and the DSDMP.

Reference lifecycles and principles were included in the analysis to both address questions of how the DSDMP relate to the process steps emphasized in lifecycle models, and the more conceptual elements of the reference principles. Through addressing both reference lifecycles and principles we can gain a more holistic assessment of the current relationship between the DSDMP and the community's broader practice informed by values. The initial set of lifecycles and principles identified by the Data-WG/DSDMP included:

- NOAA Environmental Data Management Framework (EDMF - not yet completed)
- US National Science and Technology Council Common Framework for EO Data (NSTC - preliminary connections defined)
- European Environment Agency Data/Information Management Framework (EEA - preliminary connections defined)
- (US) National Institute for Standards and Technology Research Data Framework (NIST - preliminary connections defined)
- DataONE Data Lifecycle (DataONE - preliminary connections defined)
- FAIR Principles (FAIR - preliminary connections defined)
- TRUST Principles (TRUST - preliminary connections defined)
- CARE Principles (CARE - not yet completed)

Data Collection

Following the identifcation of the reference frameworks to be used in the analysis a shared Google spreadsheet was developed in which the preliminary mappings between the DSDMP and each of the reference frameworks. The use of the spreadsheet allowed for rapid prototyping of the data model for capturing and organizing the developed mappings. The spreadsheet includes an `Instructions` worksheet that provides background information about the content and structure of the spreadsheet, a `Lifecycles` worksheet that provides reference information and labels for each of the selected reference frameworks, a `Crosswalk-DataSharingPrinciples` worksheet that provides reference information about the individual GEO data sharing principles and the mapping between those principles and the reference frameworks, and a `Crosswalk-DataManagementPrinciples` worksheet that provides reference information about the GEO data management principles and the mapping between those principles and the reference frameworks.

The tabular structure within the prototype spreadsheet enables streamlined extraction of content of descriptive information about the individual DSDMT and reference frameworks and the identified connections between them.

Analysis

The extraction of data managed in the prototype spreadsheet is accomplished through R code (in the form of the R markdown document used to create this document and other analytic products) that:

- Reads the content of the individual data containing worksheets
  - `Lifecycles`
  - `Crosswalk-DataSharingPrinciples`
  - `Crosswalk-DataManagementPrinciples`
- Extracts and formats the data from each worksheet
- Presents the extracted connection information in tabular form
- Visualizes the connection information for graphic interpretation
- Presents the reference information about the DSDMP and reference frameworks

Developed Crosswalk Information

The following table summarizes the connections defined thus far between the GEO DSDMP and the reference frameworks.

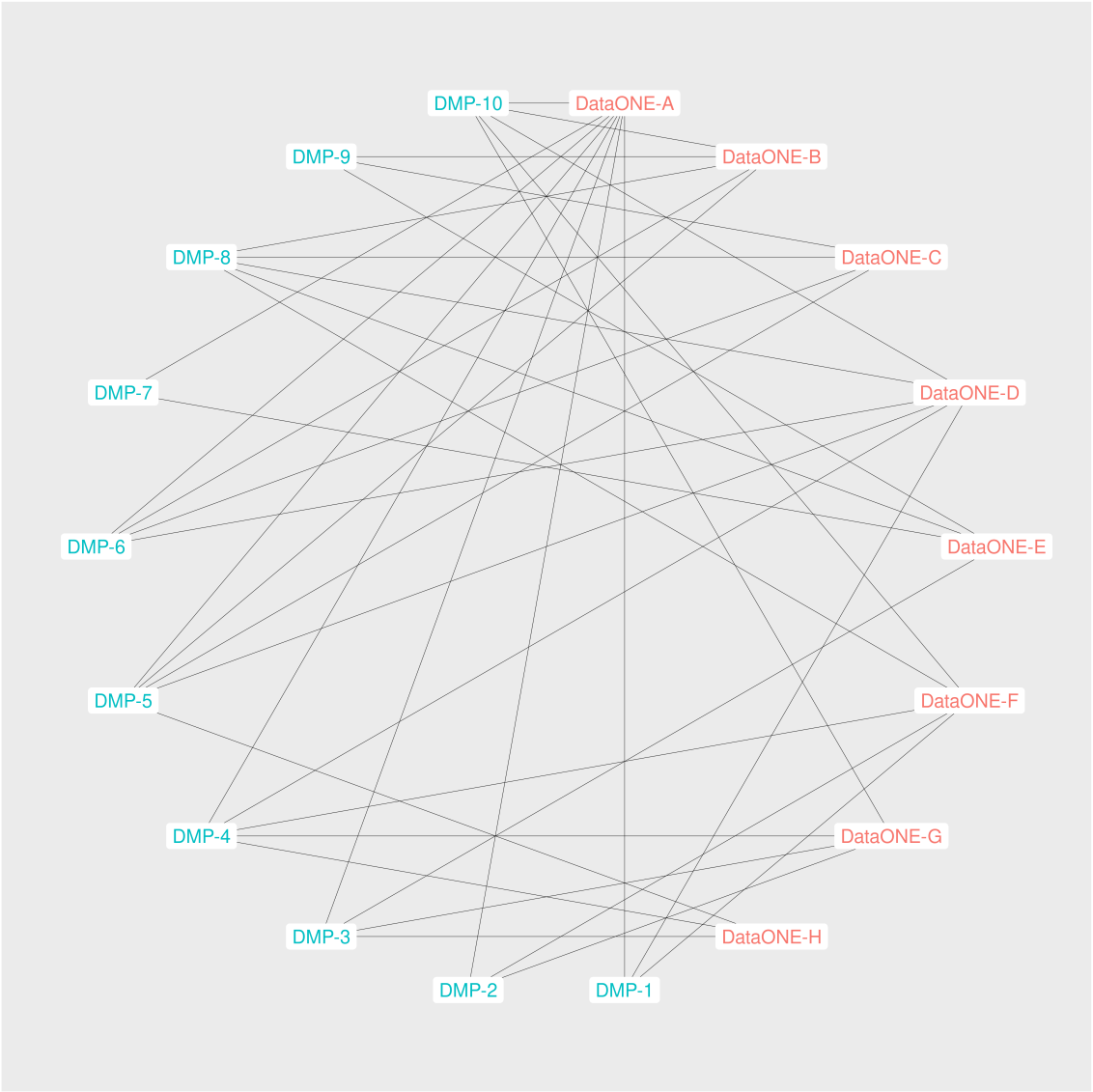
Data Management Principle	NSTC	EEA	NIST	DataONE	FAIR	TRUST
DMP-1: Metadata for Discovery	NSTC-A	EEA-J	NIST-B	DataONE-A	FAIR-F2	TRUST-U
DMP-1: Metadata for Discovery				DataONE-D	FAIR-F4	TRUST-Tr
DMP-1: Metadata for Discovery				DataONE-F	FAIR-A2	
DMP-1: Metadata for Discovery					FAIR-R1.1	
DMP-2: Online Access	NSTC-B	EEA-K	NIST-B	DataONE-A	FAIR-A1	TRUST-R
DMP-2: Online Access				DataONE-F	FAIR-A2	TRUST-U
DMP-2: Online Access				DataONE-G		
DMP-3: Data Encoding	NSTC-D	EEA-H	NIST-A	DataONE-A	FAIR-I1	TRUST-R
DMP-3: Data Encoding			NIST-B	DataONE-E	FAIR-I2	TRUST-U
DMP-3: Data Encoding			NIST-D	DataONE-G	FAIR-I3	

Data Management Principle	NSTC	EEA	NIST	DataONE	FAIR	TRUST
DMP-3: Data Encoding			NIST-E	DataONE-H	FAIR-R1.3	
DMP-4: Data Documentation	NSTC-C	EEA-J	NIST-B	DataONE-A	FAIR-F2	TRUST-R
DMP-4: Data Documentation			NIST-C	DataONE-D	FAIR-I1	TRUST-U
DMP-4: Data Documentation			NIST-D	DataONE-F	FAIR-I2	
DMP-4: Data Documentation			NIST-E	DataONE-G	FAIR-I3	
DMP-4: Data Documentation				DataONE-H	FAIR-R1	
DMP-4: Data Documentation					FAIR-R1.3	
DMP-5: Data Traceability		EEA-J	NIST-D	DataONE-A	FAIR-F1	
DMP-5: Data Traceability			NIST-E	DataONE-B	FAIR-R1.2	
DMP-5: Data Traceability				DataONE-C		
DMP-5: Data Traceability				DataONE-D		
DMP-5: Data Traceability				DataONE-H		
DMP-6: Data Quality-Control		EEA-I	NIST-C	DataONE-A		TRUST-U
DMP-6: Data Quality-Control			NIST-D	DataONE-B		
DMP-6: Data Quality-Control				DataONE-C		
DMP-6: Data Quality-Control				DataONE-D		
DMP-7: Data Preservation		EEA-L	NIST-D	DataONE-A	FAIR-A2	TRUST-Tr
DMP-7: Data Preservation			NIST-F	DataONE-E		TRUST-R
DMP-7: Data Preservation						TRUST-Te
DMP-8: Data and Metadata Verification		EEA-L	NIST-D	DataONE-B		TRUST-R
DMP-8: Data and Metadata Verification			NIST-F	DataONE-C		
DMP-8: Data and Metadata Verification				DataONE-D		
DMP-8: Data and Metadata Verification				DataONE-E		
DMP-8: Data and Metadata Verification				DataONE-F		
DMP-9: Data Review and Reprocessing		EEA-L	NIST-F	DataONE-B		TRUST-R
DMP-9: Data Review and Reprocessing				DataONE-C		TRUST-U
DMP-9: Data Review and Reprocessing				DataONE-E		
DMP-10: Persistent and Resolvable Identifiers	NSTC-A		NIST-E	DataONE-A	FAIR-F1	

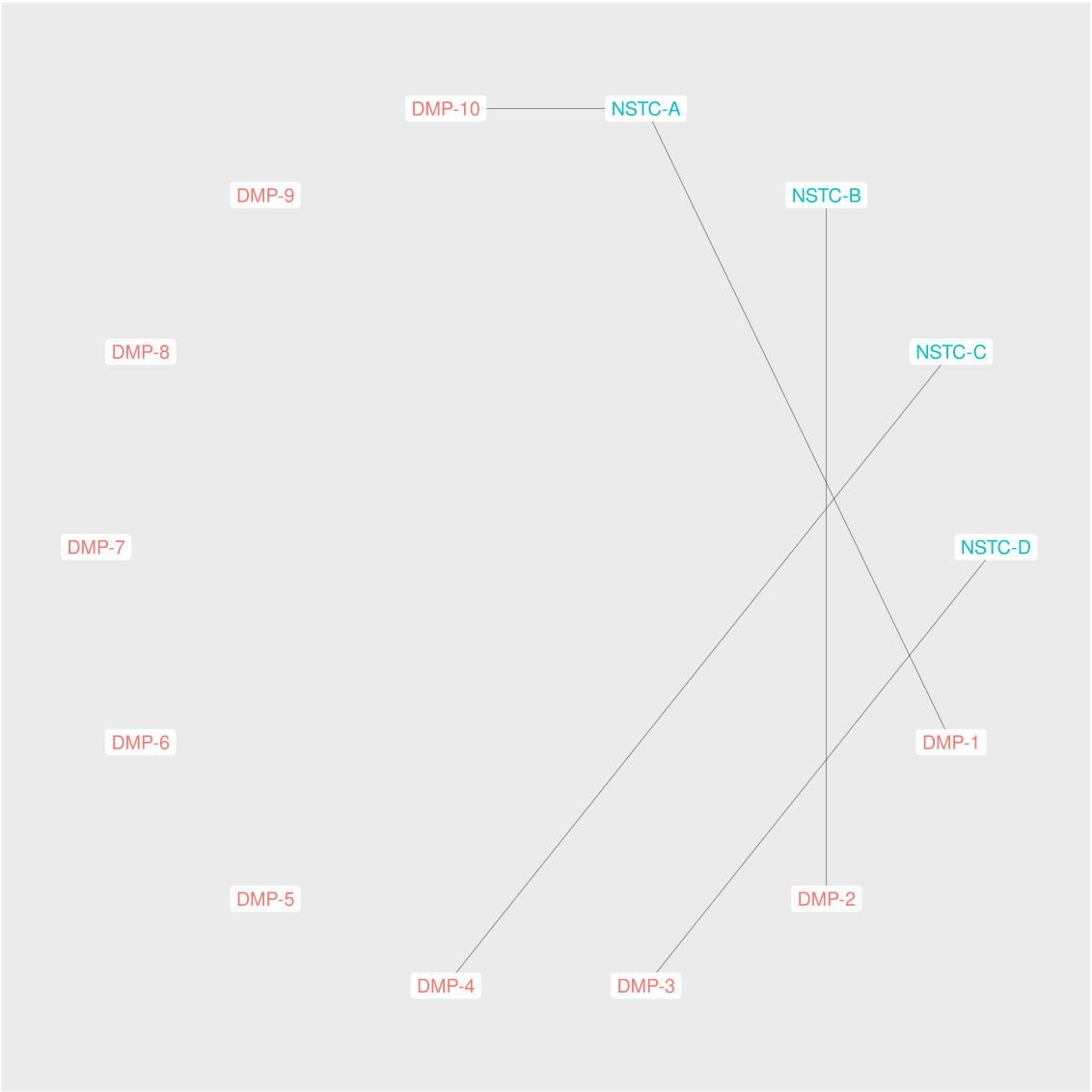
Data Management Principle	NSTC	EEA	NIST	DataONE	FAIR	TRUST
DMP-10: Persistent and Resolvable Identifiers				DataONE-B	FAIR-F3	
DMP-10: Persistent and Resolvable Identifiers				DataONE-D	FAIR-A1	
DMP-10: Persistent and Resolvable Identifiers				DataONE-F		
DMP-10: Persistent and Resolvable Identifiers				DataONE-G		

Visualization of DSDMP Relationships with Reference Lifecycles and Principles

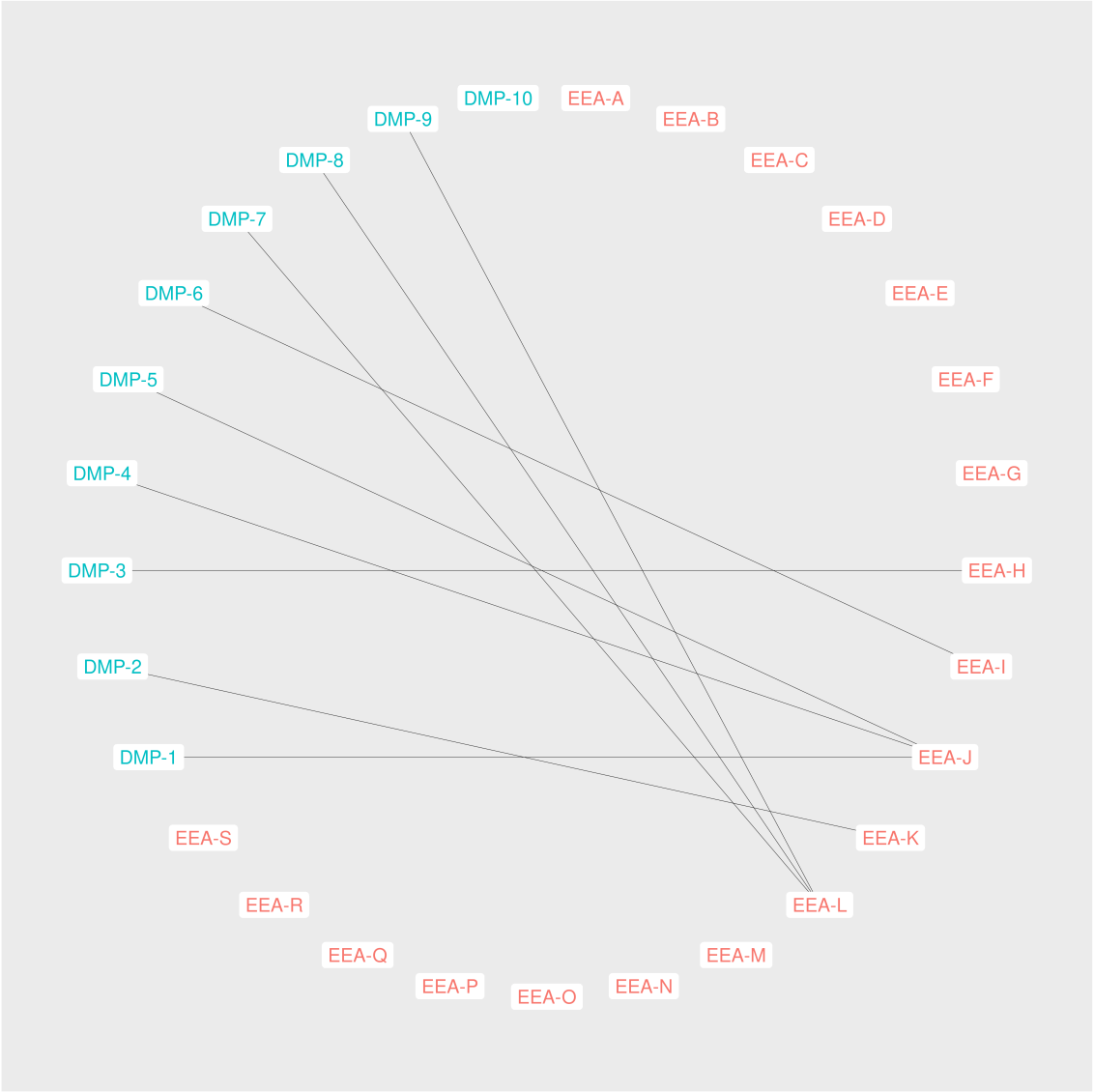
GEO Data Management Principles Mapped to DataONE Lifecycle Elements



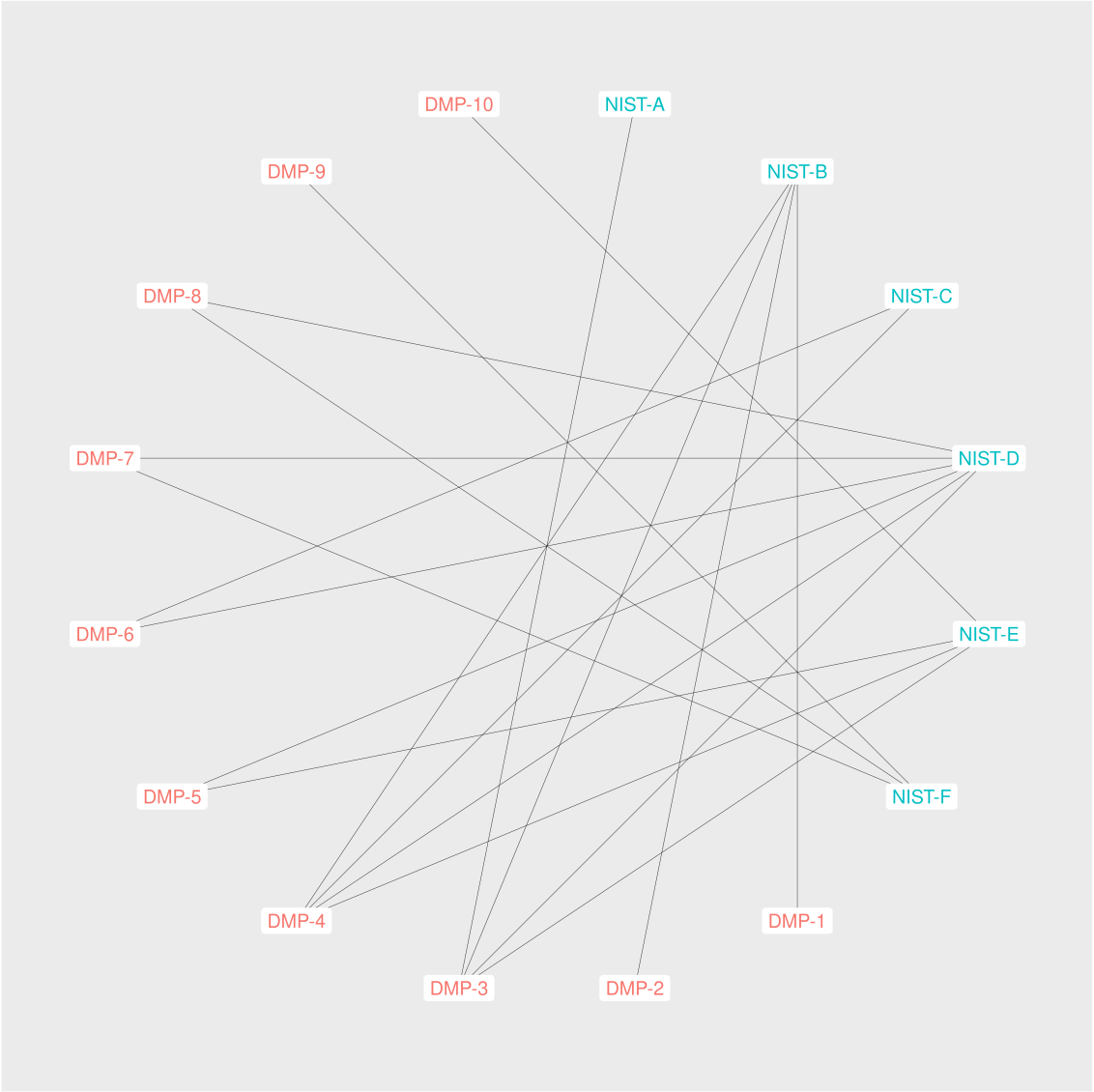
GEO Data Management Principles Mapped to NSTC Lifecycle Elements



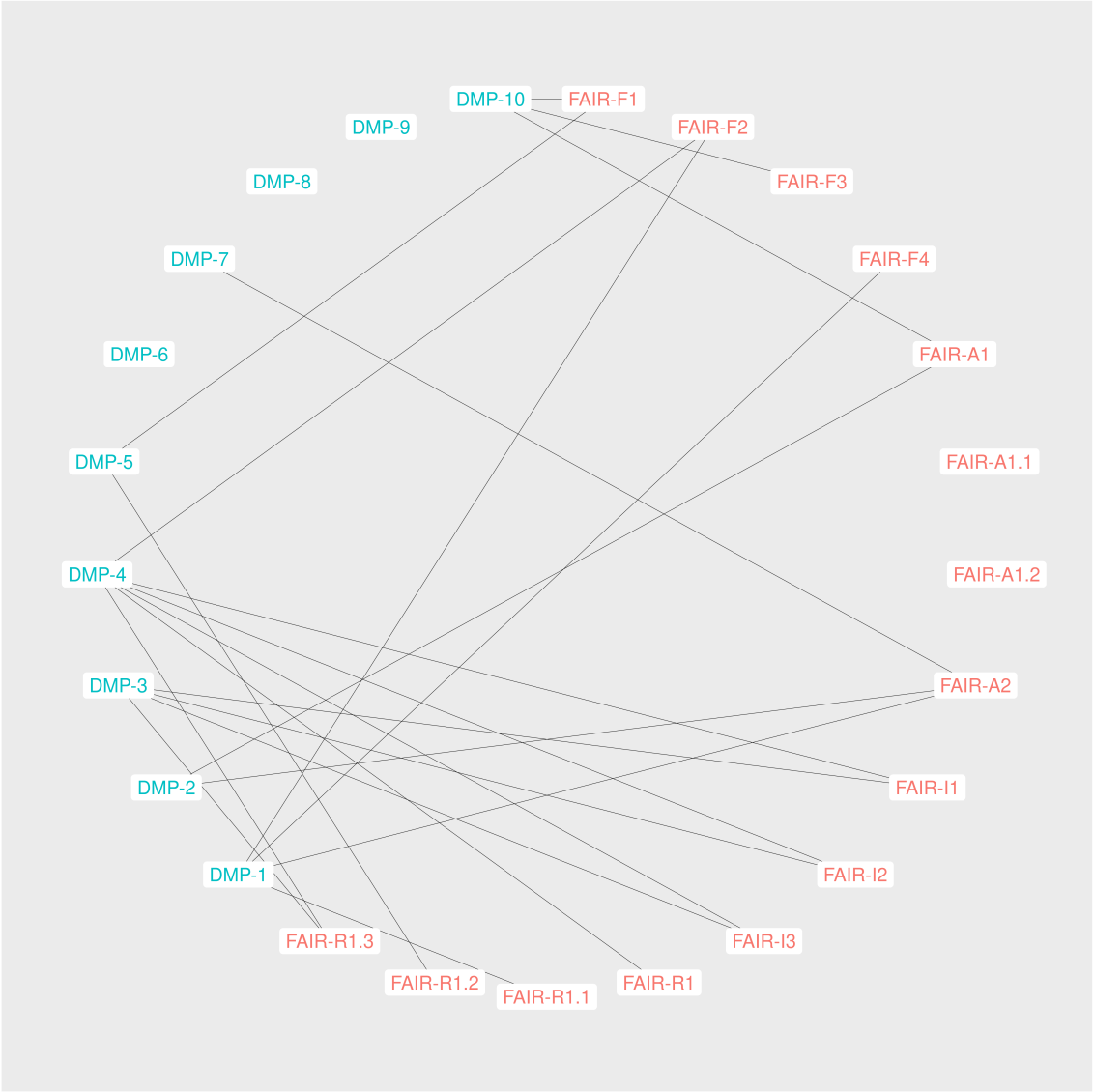
GEO Data Management Principles Mapped to EEA Lifecycle Elements



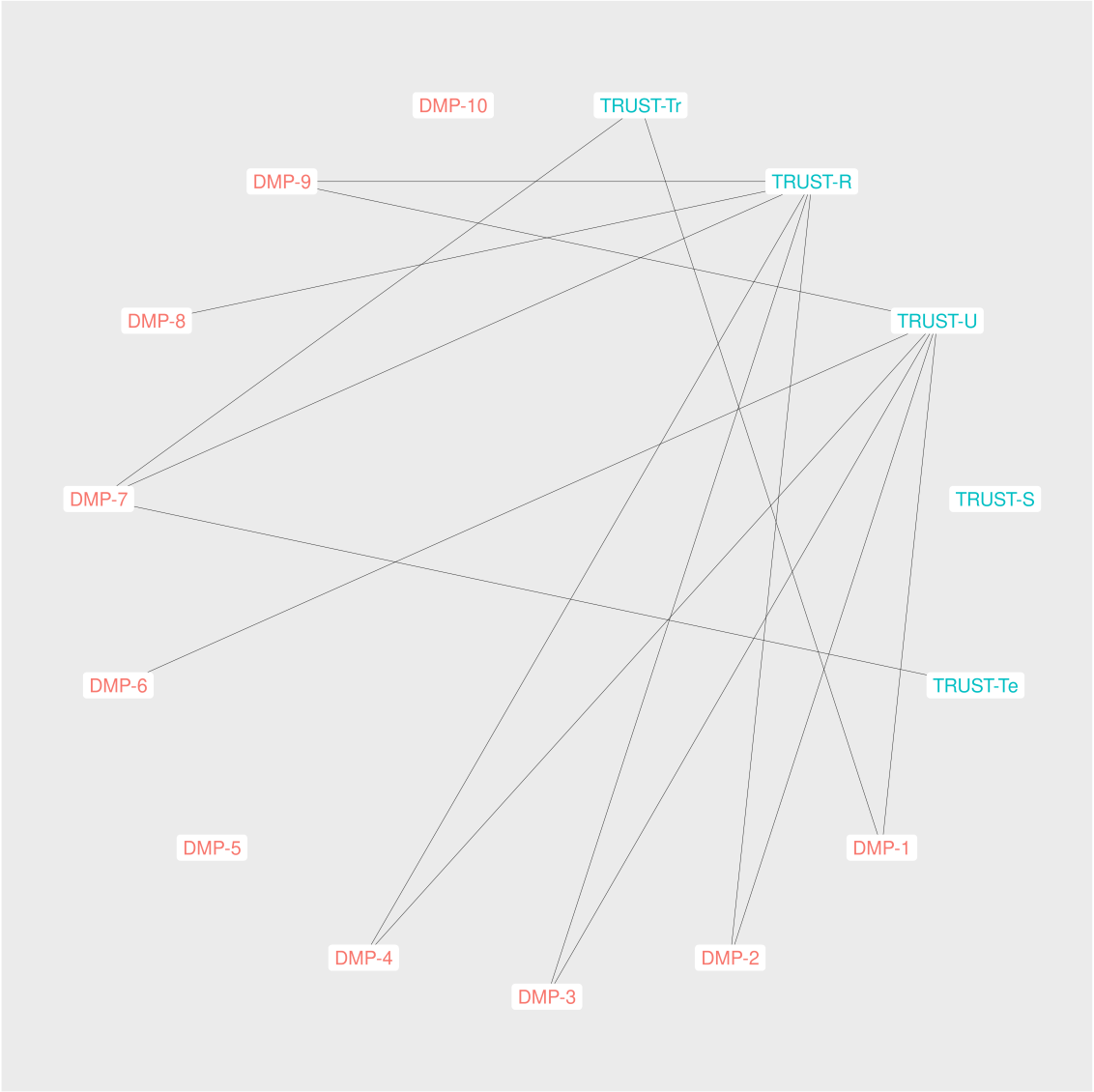
GEO Data Management Principles Mapped to NIST Lifecycle Elements



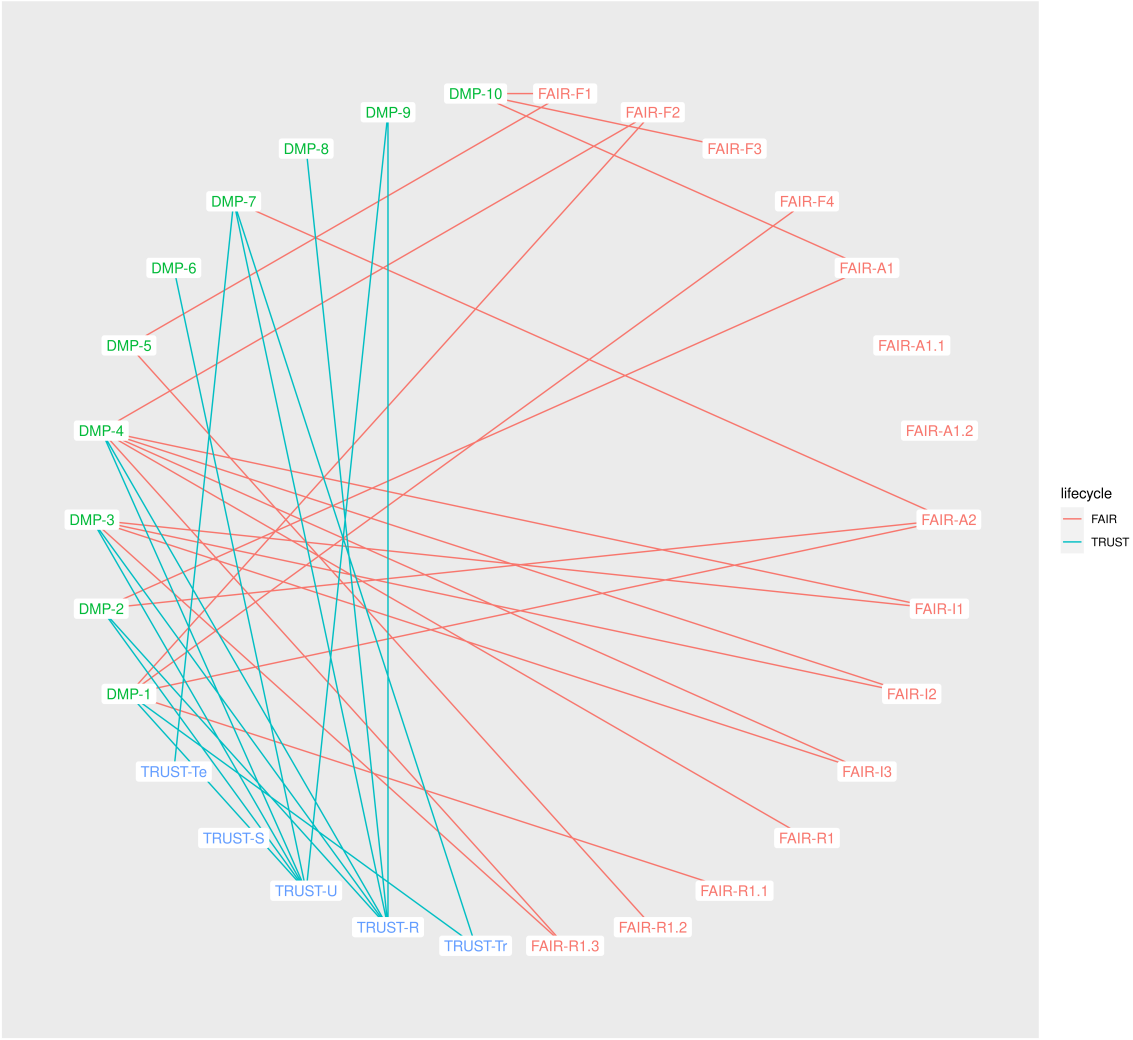
GEO Data Management Principles Mapped to FAIR Principles Elements



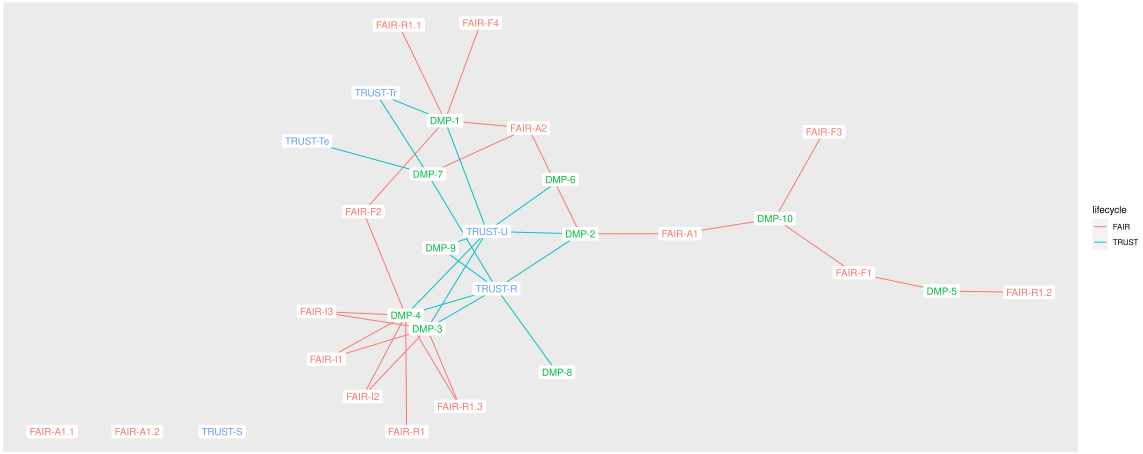
GEO Data Management Principles Mapped to TRUST Principles Elements



Crosswalk of GEO, FAIR, and TRUST principles



Crosswalk of GEO, FAIR, and TRUST principles



Crosswalk of GEO, FAIR, and TRUST principles in network form

Next Steps

While the current spreadsheet provides a useful initial platform for capturing and sharing the initial mappings, it does not provide a scalable data structure that will enable streamlined collection of data from multiple contributors, allowing for cross-validation of identified connections. Next steps for work on this project include the following:

- Transitioning to a data model that will enable capture and management of connection information from multiple contributors - enabling cross validation of identified connections.
- Expansion of the data model to capture information about the nature of the connections
- Develop an online dashboard that provides current connection information based upon community contributed data
- Publish the results of the analysis in one or more Earth Science data publication venues

Appendix A - Reference Information for the Data Sharing and Data Management Principles

Lifecycle/Principle label and Description
EDMF-A: Requirements Definition
EDMF-B: Planning
EDMF-C: Development
EDMF-D: Operations
EDMF-E: Collection
EDMF-F: Processing
EDMF-G: Quality Control
EDMF-H: Documentation
EDMF-I: Cataloging
EDMF-J: Dissemination
EDMF-K: Preservation
EDMF-L: Stewardship
EDMF-M: Usage Tracking
EDMF-N: Final Disposition
EDMF-O: Discovery
EDMF-P: Reception
EDMF-Q: Understanding
EDMF-R: Analysis
EDMF-S: Value-Added Products
EDMF-T: User Feedback
EDMF-U: Citation
EDMF-V: Tagging
EDMF-W: Gap Assessment
NSTC-A: Data Search and Discovery

Lifecycle/Principle label and Description
NSTC-B: Data-Access
NSTC-C: Data Documentation
NSTC-D: Compatible Formats and Vocabularies
EEA-A: Requirements Analysis
EEA-B: Component Identification
EEA-C: Content Specification
EEA-D: Data Availability Audit
EEA-E: Data Source Identification
EEA-F: Development
EEA-G: Acquisition (input)
EEA-H: Processing
EEA-I: Quality Control
EEA-J: Documentation& Cataloguing
EEA-K: Publication (output)
EEA-L: Preservation / Archiving
EEA-M: Analysis
EEA-N: Value Added Product Creation
EEA-O: Dissemination
EEA-P: Usage Tracking
EEA-Q: User Feedback
EEA-R: Citation
EEA-S: Gap Analysis
NIST-A: Envision
NIST-B: Plan
NIST-C: Generate / Acquire
NIST-D: Process / Analyze
NIST-E: Use / Reuse
NIST-F: Preserve / Discard
DataONE-A: Plan
DataONE-B: Collect
DataONE-C: Assure
DataONE-D: Describe
DataONE-E: Preserve
DataONE-F: Discover
DataONE-G: Integrate

Lifecycle/Principle label and Description
DataONE-H: Analyze
FAIR-F1: (meta)data are assigned a globally unique and persistent identifier
FAIR-F2: data are described with rich metadata
FAIR-F3: metadata clearly and explicitly include the identifier of the data it describes
FAIR-F4: (meta)data are registered or indexed in a searchable resource
FAIR-A1: (meta)data are retrievable by their identifier using a standardized communications protocol
FAIR-A1.1: the protocol is open, free, and universally implementable
FAIR-A1.2: the protocol allows for an authentication and authorization procedure, where necessary
FAIR-A2: metadata are accessible, even when the data are no longer available
FAIR-I1: (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
FAIR-I2: (meta)data use vocabularies that follow FAIR principles
FAIR-I3: (meta)data include qualified references to other (meta)data
FAIR-R1: meta(data) are richly described with a plurality of accurate and relevant attributes
FAIR-R1.1: (meta)data are released with a clear and accessible data usage license
FAIR-R1.2: (meta)data are associated with detailed provenance
FAIR-R1.3: (meta)data meet domain-relevant community standards
TRUST-Tr: To be transparent about specific repository services and data holdings that are verifiable by publicly accessible evidence
TRUST-R: To be responsible for ensuring the authenticity and integrity of data holdings and for the reliability and persistence of its service
TRUST-U: To ensure that the data management norms and expectations of target user communities are met
TRUST-S: To sustain services and preserve data holdings for the long-term
TRUST-Te: To provide infrastructure and capabilities to support secure, persistent, and reliable services
DMP-1: Metadata for Discovery
DMP-2: Online Access
DMP-3: Data Encoding
DMP-4: Data Documentation
DMP-5: Data Traceability
DMP-6: Data Quality-Control
DMP-7: Data Preservation
DMP-8: Data and Metadata Verification
DMP-9: Data Review and Reprocessing
DMP-10: Persistent and Resolvable Identifiers