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

Karl Benedict

2022-Apr-20

GEO Working Group: Data Working Group (Data-WG)
Subgroup: *Data Sharing and Data Management Principles* (Data-WG/DSDMP)

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		

Visualize some relationships

GEO Data Management Principles - DataONE

```
geo_x_single(
  "GEO Data Management Principles Mapped to DataONE Lifecycle Elements",
  "DataONE",
  "dataone.png")
```

```
## Warning in eattrs[[nam[i]]][idx] <- attrs[[nam[i]]]: number of items to replace
## is not a multiple of replacement length
```

```
## Using `sugiyama` as default layout
```

GEO Data Management Principles - NSTC

```
geo_x_single(
  "GEO Data Management Principles Mapped to NSTC Lifecycle Elements",
  "NSTC",
  "nstc.png")
```

```
## Warning in eattrs[[nam[i]]][idx] <- attrs[[nam[i]]]: number of items to replace
## is not a multiple of replacement length
```

```
## Multiple parents. Unfolding graph
```

```
## Multiple roots in graph. Choosing the first
```

```
## Using `tree` as default layout
```

GEO Data Management Principles - EEA

```
geo_x_single(
  "GEO Data Management Principles Mapped to EEA Lifecycle Elements",
  "EEA",
  "eea.png")
```

```
## Warning in eattrs[[nam[i]]][idx] <- attrs[[nam[i]]]: number of items to replace
## is not a multiple of replacement length
```

```
## Multiple parents. Unfolding graph
```

```
## Multiple roots in graph. Choosing the first
```

```
## Multiple parents. Unfolding graph
```

```
## Multiple roots in graph. Choosing the first
```

```
## Using `tree` as default layout
```

GEO Data Management Principles - NIST

```
geo_x_single(
  "GEO Data Management Principles Mapped to NIST Lifecycle Elements",
  "NIST",
  "nist.png")
```

```
## Warning in eattrs[[nam[i]]][idx] <- attrs[[nam[i]]]: number of items to replace
## is not a multiple of replacement length
```

```
## Using `sugiyama` as default layout
```

GEO Data Management Principles - FAIR

```
geo_x_single(
  "GEO Data Management Principles Mapped to FAIR Principles Elements",
  "FAIR",
  "fair.png")
```

```
## Warning in eattrs[[nam[i]]][idx] <- attrs[[nam[i]]]: number of items to replace
## is not a multiple of replacement length
```

```
## Using `sugiyama` as default layout
```

GEO Data Management Principles - TRUST

```
geo_x_single(
  "GEO Data Management Principles Mapped to TRUST Principles Elements",
  "TRUST",
  "trust.png")
```

```
## Warning in eattrs[[nam[i]]][idx] <- attrs[[nam[i]]]: number of items to replace
## is not a multiple of replacement length
```

```
## Using `sugiyama` as default layout
```

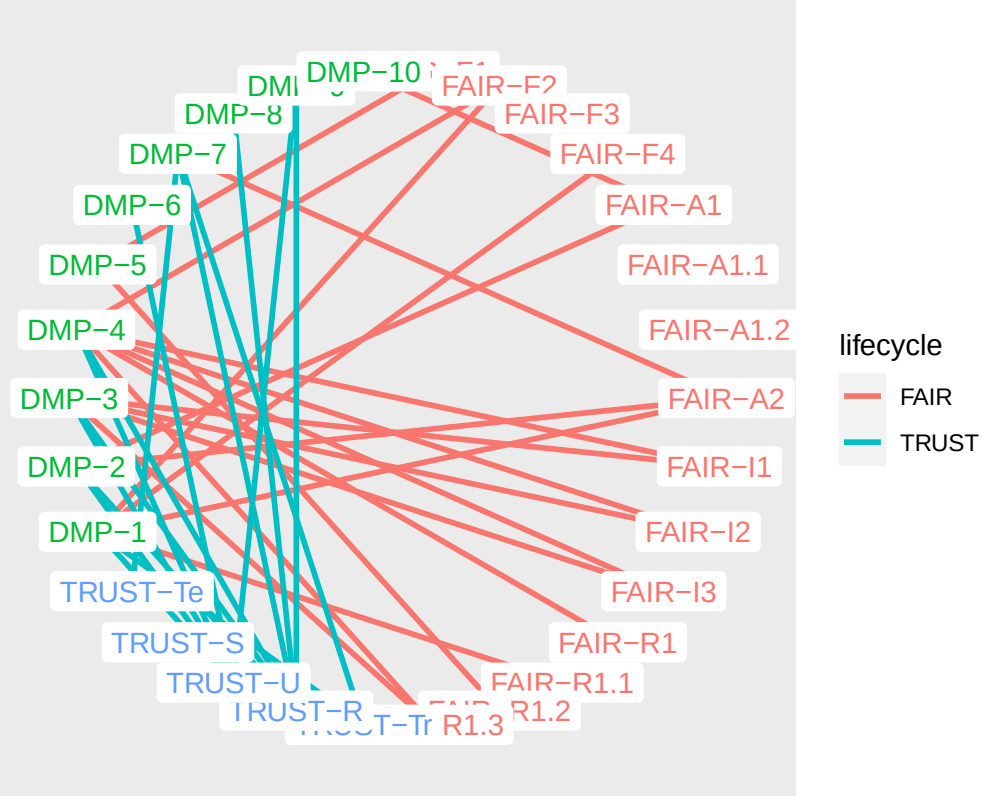
GEO Data Management Principles - Reference Frameworks

```
# combined graph for all principles
graph <- graph_from_data_frame(
    d = filter(e_dmp_lifecycle, (lifecycle == "FAIR" | lifecycle == "TRUST") ),
    vertices = filter(n_all, (source == "FAIR" | source == "TRUST" | source == "GEO") ),
    directed = FALSE)
```

```
## Warning in eattrs[[nam[i]]][idx] <- attrs[[nam[i]]]: number of items to replace
## is not a multiple of replacement length
```

```
  xy <- layout_(graph,in_circle())
  ggraph(graph, layout = "linear", circular = TRUE) +
    coord_fixed() +
    geom_edge_link0(aes(color = lifecycle), width = 1) +
    geom_node_point(aes(color = as.factor(source)), size = 1, show.legend = FALSE) +
    geom_node_label(aes(label = short_label,
                        color = as.factor(source)),
                    label.size = 0,
                    show.legend = FALSE) +
    xlim(-1.1, 1.1) +
    ylim(-1.1, 1.1) +
    ggtitle("Crosswalk of GEO, FAIR, and TRUST principles") +
    theme(plot.title = element_text(size = 7, face = "bold"))
```

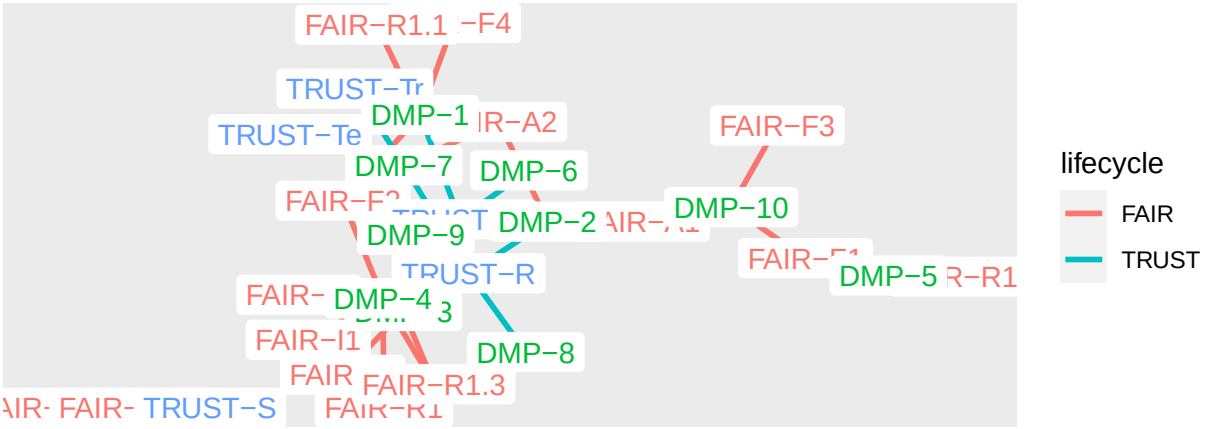
Crosswalk of GEO, FAIR, and TRUST principles




```
ggsave(paste("images/circular_","principles.png", sep = ""), width = 12, height = 12, dpi = 600)
ggraph(graph) +
  coord_fixed() +
  geom_edge_link0(aes(color = lifecycle), width = 1) +
  geom_node_point(aes(color = as.factor(source)), size = 1, show.legend = FALSE) +
  geom_node_label(aes(label = short_label,
                      color = as.factor(source)),
                  label.size = 0,
                  show.legend = FALSE) +
  ggtitle("Crosswalk of GEO, FAIR, and TRUST principles") +
  theme(plot.title = element_text(size = 7, face = "bold"))
```

```
## Using `stress` as default layout
```

Crosswalk of GEO, FAIR, and TRUST principles




```
  ggsave(paste("images/auto_","principles.png", sep = ""), width = 18, height = 18, dpi = 600)
```

```
graph <- graph_from_data_frame(
  d = e_dmp_lifecycle,
  vertices = n_all,
  directed = FALSE)
```

```
## Warning in eattrs[[nam[i]]][idx] <- attrs[[nam[i]]]: number of items to replace
## is not a multiple of replacement length
```

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

Under development