Elements of a Data Preservation Strategy

Denise Bleakly, Retired Sandia National Laboratories, GISP

National States Geographic Information Council (NSGIC)

4 December 2025

Contact information: *d3bleakly@gmail.com* *505.239.3306*

Acknowledgements

- I would like to thank the New Mexico Geographic Information Council (NMGIC) for funding this work and encouraging the development of this presentation for a wider audience
- Mr. Leland Pierce for encouraging me in developing this topic area and for supporting the effort for getting this information out to a larger audience
- National States Geographic Information Council (NSGIC) for assisting in the development of this presentation and providing support for hosting it on their education portal.

Geospatial Data Preservation Strategy Elements

In this talk we will cover the following:

- Why geospatial data preservation is difficult
- The basic elements of a data preservation strategy
- Robust metadata development and maintenance
- The use of FAIR data principles as infrastructure for data preservation
- Best practices for long term geospatial data preservation

- *NOTE: This presentation contains more information than I will be presenting. It was developed for future reference and to serve as a resource*

Have You Ever…

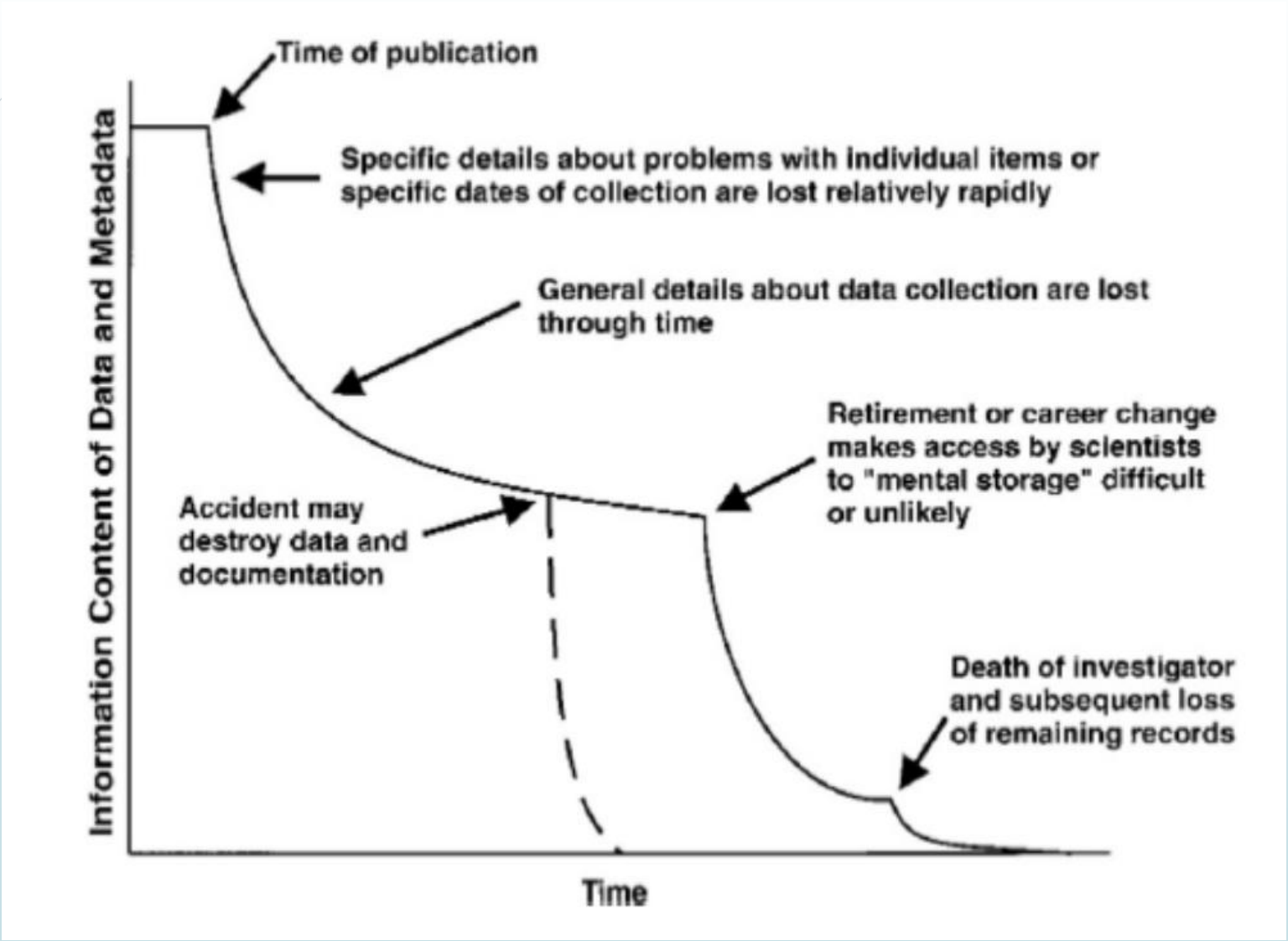
- **Lost track of where you put your files?**
- **Couldn't find an important file?**
- **Couldn't retrieve data you swore you archived?**
- **Found data that you archived, but it was unreadable or corrupt?**
- **Archived data and tried to retrieve the data years later only to discover that it wasn't what you thought it was?**

These are all examples of what can happen to digital data during your working career…

Now, imagine what would happen if your future colleagues

needed to access your data 25+ years from now

Loss of Data Over Time



1997, Mitchener, et al., "Non Geospatial Metadata for the Ecological Sciences. Ecological Applications, 7(1), 1997, pp 330-342.

Threats to Digital Data Preservation

- **Dramatic events** such as floods, earthquakes and political upheaval
- **Understanding the data:** Future users may be unable to understand or use the data (because of the semantics, format, processes or algorithms involved).
- **The chain of evidence** may be lost and there may be lack of certainty of provenance or authenticity.
- **Non-maintainability** of essential hardware, software, or support environment may make the information inaccessible.
- **Access and use restrictions** may not be respected in the future, jeopardizing proper reuse
- **Loss of ability to identify the location of data**
- **Institutions may cease to exist:** The current custodian of the data, whether an organization or project, may cease to exist at some point in the future.
- **Institutions may let us down:** The ones we may trust to look after the digital holdings may let us down (budget cuts, data priorities, change in management).

Why is it Hard to Preserve Geospatial Data over the Long Term?

- **No Uniform Data Model** – geospatial data are represented in a wide variety of data types: vector and raster; topological and non-topological; discrete and continuous domains
- **Proprietary Formats** – formats are closely tied to specific software systems, which are not always backward compatible (e.g., ESRI Geodatabases)
- **Multiple granule sizes** - data range from individual features to thematic layers of features to heterogeneous spatial databases
- **Relational Data systems** – store complex datasets

NMGIC NEW MEXICO GEOGRAPHIC INFORMATION COUNCIL 1984-2024 40 YEARS

Why is it Hard to Preserve Geospatial Data over the Long Term?

- **Large Size** – gigabyte sizes growing by terabytes are common
- **Long-lived Programs** – Geospatial data sets can be long lived, years or decades of data collection is common
- **Extensive context** - capturing enough contextual information around a geographic data set can be challenging
- **Dynamic Data** – some datasets change daily and are ever-growing, capturing contextual data and processing methods for preservation is a challenge. Geospatial data may require extensive, product-specific context.

Why Do I need a Geospatial Data Preservation Strategy?

- **Data preservation is part of a data lifecycle plan**. If you don't have a data management plan, you probably don't have a geospatial data preservation strategy
- **In practice, it has been shown that geospatial data are hard to preserve over the long term, and specific steps need to be taken to increase the likelihood that the data will survive longer than the project lifespan**
- **A geospatial data preservation strategy**, is a document which lays out distinct actions to be taken that will greatly increase the chances that the geospatial data you manage will be preserved into the future.
- **Digital preservation infrastructure:** Digital preservation will more likely be achieved through a designated digital preservation infrastructure that ensures data integrity, format and media sustainability, and information security

What to Preserve?

Preservation is so much more than making sure you can read a digital file: for geospatial data, it's all about context

- The original raw and reprocessed data
- All the relevant information including metadata, documentation related to intermediate processing steps or algorithms, unique analysis information, scripts developed
- Important calibration or model input data
- Instrument design prints
- PDF's of maps or other output
- Other data as required by law, policy, regulation, or agreement

Basic Elements of a Geospatial Data Preservation Strategy

A Data Preservation Strategy addresses these topics:

Data Life Cycle

- Planning for data preservation should occur with the creation of the data (Cradle-to- grave data planning)
- Periodic reviews of the data during the data life cycle (1, 10, 25, 100 year philosophy)

Records Requirements

- Identifies what the records requirements are from a project, program and organizational perspective
- Plans for periodic records to be created as to where data resides, where the metadata, data dictionary, and where the contextual information is stored

Basic Elements of a Geospatial Data Preservation Strategy Cont'd

Legal

• Identifies regulatory requirements (such as permit requirements, other Federal Government Requirements, state requirements, etc.)

• Identifies and addresses any data ownership and data Sharing requirements which will allow for planning for periodic "archives" (every 10 + years)

Documentation and Metadata

- Documents what metadata and documentation standards are used to document the data and provides the dataset informational context

Basic Elements of a Geospatial Data Preservation Strategy cont'd

File Formats

- Documents what file formats the data are in and what data format standards are followed
- Identifies the needs for data format conversions for future long-term archiving

Data Prioritization

- Documents and describes how the project/program decides what data is prioritized for preservation activities

Storage and geographic location

- Documents storage systems, locations and multiple copies of data & metadata to prevent loss of data

Basic Elements of a Geospatial Data Preservation Strategy cont'd

Data Integrity

- Describes procedures to prevent, detect, and recover from unexpected or deliberate changes to data
- Describes procedures to check the fixity upon ingest of the data to ensure the data ingested is the same as the data provided

Information Security

- Describes procedures to prevent human-caused corruption of data, deletion and unauthorized access to the data / metadata

Financial

- Identifies and describes costs for data preservation that can be planned for as part of the data lifecycle
- Planning for periodic data refreshing and associated costs (every 5-10 years)

Wow! A Data Preservation Strategy is overwhelming!

All of the elements of a structured data preservation strategy can be overwhelming, but there are a few tools and techniques you can implement on a daily basis that will increase the likelihood that your data will be preserved over the long haul:

- *Develop robust metadata and maintain the metadata*
- *Follow the FAIR data principles*
- *Follow best geospatial data practices*

Geospatial Metadata is KEY to long-term geospatial data preservation

Geospatial Metadata:

From the USGS:

- *Metadata describe information about data, including who, what, where, when, why, and how, so that it can be understood, re-used, and integrated with other data. Metadata records follow a standard format to enable interoperability.*
- *Metadata are crucial for any use or reuse of data; no one can responsibly re-use or interpret data without metadata that explains how the data were created, why, where it is geographically located, and details about the structure of the data.*
- *Metadata are used for enabling data discovery, understanding data, analysis and synthesis, maintaining longevity of data, tracking the progress of a research project, and demonstrating the return on investment for research at an institution.*

From the USGS Metadata Tab on the USGS Data Management Website, : https://www.usgs.gov/data-management/metadata-creation, accessed 8/29/25

Metadata Standards Requirements

- ***Federal agencies are mandated*** by Executive Order 12906 to use metadata standards endorsed by the Federal Geographic Data Committee (FGDC) below:
- Content Standard for Digital Geospatial Metadata (CSDGM)
- International Organization for Standardization (ISO) series of standards (19115, 19115-2, 19139, etc.). There is an ongoing effort to move towards adopting the ISO metadata standard.
- Both FGDC-CSDGM and ISO require metadata to be formatted in Extensible Markup Language (.xml)

From the USGS Metadata Tab on the USGS Data Management Website, : https://www.usgs.gov/data-management/metadata-creation, accessed 8/29/25

Metadata Resources

Executive Order 12906, directed that Federal agencies document geospatial resources using the **Content Standard for Digital Geospatial Metadata (CSDGM)**

Content Standard for Digital Geospatial Metadata (CSDGM), Vers. 2 (FGDC-STD-001-1998) is the current version of this FGDC authored and endorsed standard.

- https://www.fgdc.gov/standards/projects/metadata/base-metadata/index_html

Since the publication of EO 12906, the FGDC has endorsed several ISO Geospatial Metadata Standards that are now encouraged for use.

The Federal Geographic Data Committee (FGDC) maintains a website to download current versions of metadata standards:

https://www.fgdc.gov/resources/download-geospatial-standards

FAIR Data Practices as a Data Preservation Tool

The FAIR data practices were developed in the mid-2000's for improving the sharing of scientific data, which included data discovery, accessibility, interoperability and reusability.

Although the FAIR data principles were not developed as a data preservation tool, they provide a great starting point to develop the data management infrastructure that will help in long term data preservation

Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.*
The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016).
https://doi.org/10.1038/sdata.2016.18

FAIR Data Principles

The FAIR data principles are:

- **F– Findable** – is the data discoverable
- **A – Accessible** – can the data be accessed
- **I – Interoperable** – does the data contain enough information to determine interoperability
- **R- Reusable** – is the data reusable

Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.*
The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016).
https://doi.org/10.1038/sdata.2016.18

FAIR Data Guiding Principles

To be FINDABLE

- F.1 (meta)data are assigned a globally unique and persistent identifier
- F.2 data are described with rich metadata (defined by R1 below)
- F.3 metadata are clearly and explicitly includes the identifier of the data it describes
- F.4. (meta)data are registered or indexed in a searchable resource

Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.*
The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016).
https://doi.org/10.1038/sdata.2016.18

FAIR Data Guiding Principles

To be FINDABLE

- F.1 (meta)data are assigned a globally unique and persistent identifier
- F.2 data are described with rich metadata (defined by R1 below)
- F.3 metadata are clearly and explicitly includes the identifier of the data it describes
- F.4. (meta)data are registered or indexed in a searchable resource

Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.*
The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016).
https://doi.org/10.1038/sdata.2016.18

FAIR Data Guiding Principles

To be ACCESSIBLE

- A.1 (meta)data are retrievable by their identifier using a standardized communication protocol
- A.1.1 the protocol is open, free and universally implementable
- A.1.2 the protocol allows for an authentication and authorization procedure, where necessary
- A.2 metadata are accessible, even when the data are no longer available

Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.*
The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016).
https://doi.org/10.1038/sdata.2016.18

FAIR Data Guiding Principles

To be INTEROPERABLE

- I.1 (meta)data use a formal, accessible, shared and broadly applicable language for knowledge representation
- I.2 (meta)data use vocabularies that follow FAIR principles
- I.3 (meta)data include qualified references to other (meta)data

Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.*
The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016).
https://doi.org/10.1038/sdata.2016.18

FAIR Data Guiding Principles

To be REUSEABLE

- R.1 meta(data) are richly describe with a plurality of accurate and relevant attributes
- R.1.1 (meta)data are released with a clear and accessible data usage license
- R.1.2 (meta)data are associated with detailed provenance
- R.1.3 (meta)data meet domain-relevant community standards

Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.*
The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016).
https://doi.org/10.1038/sdata.2016.18

Key Data Preservation Practices

1. ***Metadata!*** Develop a strong metadata development and maintenance program
2. Use widely adopted standards
3. Storing data in well supported, open formats
4. Data should be free from external dependencies
5. Adopt the use persistent identifiers/Digital Object Identifiers (DOI's) for datasets
6. Bundling data, metadata and context information together using something like "Bagit"

Key Data Preservation Practices

7. Store a graphical representation of the data (e.g., as a pdf) with the data and metadata bundle
8. Ensure all information objects are self-contained and independently understandable.
9. Preserve geographic data in a way that non-geospecialists can understand
10. Plan for technological obsolescence – media migration every 3-5 years, data format migration every 10-25 years
11. The 3-2-1 rule should be applied: three data copies in at least two formats, with at least one copy stored in a separate secure location.

A Note About File Formats

Geospatial data projects use a wide variety of proprietary data formats as part of the data development and analysis process.

Proprietary file formats are one of the main reasons geospatial data does not survive in the long term

To prevent geospatial data from being obsolete and not re-useable, long-term data preservation requires the data to be converted into currently adopted, well documented file formats.

File Formats

The National Archives and Records Administration (NARA) maintains a current list of acceptable file formats for all data required to go into a federal data repository.

The most current list of acceptable geospatial, imagery, documentary, textual and database file formats is located at:

https://www.archives.gov/records-mgmt/policy/transfer-guidance-tables.html

The Library of Congress (LOC) *also maintains a list of currently acceptable file formats if your content will be subject to storage at the Library of Congress*

The link to the geospatial data formats table is:

https://www.loc.gov/preservation/resources/rfs/geo-carto.html

Note: all links tested and verified on 8/30/2025

How does this work in *the Real World*?

- **Data preservation is a *TEAM EFFORT***
- ***Focus on metadata development and maintenance*** *– start the metadata early in data set development, don't use metadata as a "punishment" , if team members are having issues with metadata development and maintenance, work to make the metadata easier to develop and maintain.* ***Talk about metadata issues as a TEAM***
- ***Implement the FAIR data principles.*** *The FAIR data principles contain the infrastructure needed for data preservation. This does not need to be a new program, work within your existing team structure to start practicing the FAIR data principles*
- ***Review your data management plans and strategies.*** *Think of your data management plan as a cradle to grave guideline. If you don't have any preservation strategies in your data management plan, add some!*
- ***Follow best geospatial data practices***

Keep Up The Good Work!

- ***You all work on very important work for your projects and agencies***
- ***The data you collect, manage, curate and preserve will be used by future scientists and engineers*** –
  - Think of the Early LandSat imagery From the 1970's has played a critical role in our current studies on Climate Change
  - Ask your self – "If I were to move on to another job, would my colleagues know where to find and use the data I have created?"
  - Create file "Breadcrumbs" You can start data preservation practices by documenting the data you create, maintain metadata, and work with others of your team to back up the data and archive it.

Questions?

By working together as a Geospatial Community, we can preserve the vital geographic data we have created and manage it for future generations

Resources for Developing a Digital Data Preservation Strategy

Over the last several years, some important resources have emerged that can assist in developing a ***Digital Data Preservation Strategy***

National Archives – Digital Data Preservation Strategy
https://www.archives.gov/preservation/electronic-records/digital-preservation-strategy#

Wheaton College Library and Archives – Data Preservation Plan
https://library.wheaton.edu/sites/default/files/Digital_Preservation_Plan.pdf

National Digital Stewardship Alliance (NDSA)– Levels of Digital Preservation
https://ndsa.org/publications/levels-of-digital-preservation/

Ecological Informatics, 3rd edition – Chapter 3: scientific data management, documentation, metadata and preservation https://link.springer.com/book/10.1007/978-3-319-59928-1

Note: all links tested and verified on 8/30/2025

Additional Resources

FAIR Data Principles

- Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016). https://doi.org/10.1038/sdata.2016.18
- Thompson, P.T., Ojha, S., Powell, C.D. *et al.* A proposed FAIR approach for disseminating geospatial information system maps. *Sci Data* **10**, 389 (2023). https://doi.org/10.1038/s41597-023-02281-1

National Digital Stewardship Alliance (NDSA) Levels of Digital Preservation

- Bailey, Jefferson, et al. "The NDSA Levels of Digital Preservation: An Explanation and Uses." (2013). https://digitalpreservation.gov/documents/NDSA_Levels_Archiving_2013.pdf
- **NDSA website**

https://ndsa.org/publications/levels-of-digital-preservation/

Digital Preservation Strategies

Shimray, Somipam R., and C. Kodanda Ramaiah. "Digital preservation strategies: an overview." *11th National Conference on Recent Advances in Information Technology-2018*. 2018.

https://www.researchgate.net/profile/Somipam-Shimray/publication/327221006_Digital_Preservation_Strategies_An_Overview/links/5b80da10a6fdcc5f8b6592f4/Digital-Preservation-Strategies-An-Overview.pdf

Data Preservation Strategy Worksheet

The following slides contain the contents of a data preservation strategy worksheet. It provides an outline of the content of a Data Preservation Strategy and contains a series of questions for consideration as you work through the elements of the data preservation strategy.

Geospatial Data Preservation Plan Worksheet

This worksheet was developed based on templates from Data Management Plans, data preservation strategies and FAIR data principles. It is meant to assist in capturing the information needed to think about data preservation and how it can be implemented in your work group.

Administrative	*Basic information about the project, data collected and the context for which the data was developed*
Basic information	Project title, contact name, contact details
Summary	Summary of data set, explain the purpose of the dataset
Guidance Documents	• Identify Organizational requirements for data preservation, • list any organizational guidance documents • list federal government guidance documents (if applicable)
Legal	*Identify any legal requirements for the preservation of your data*
	Are there any regulatory requirements for your data? (such as permit requirements, other Federal Government Requirements, state requirements, etc.) Sharing requirements will allow for planning for periodic “archives” (every 10 + years)
	Identify and address any data ownership and data distribution restrictions
	Is privacy protection in place? Have you removed personal or sensitive information from your data to ensure privacy protection?
	Have you established who owns the copyright of your data?
	Have you documented how the data should be institutionally credited or cited?

Documentation & Metadata	*Documentation of the data to enable contextual understanding and long-term usability*
	List standards and guidance used • FGDC Content Metadata standard • ISO standards • File format standards
	Inventory content and location of descriptive metadata of data sets
	Do you have an indexing system for your metadata?
	Geospatial content – have you described the map projection, datum, and other geospatial parameters for your project in general? Are the details in the dataset specific metadata?
	Have you documented, software, models or codes used in the development of the data?
File Formats	*This describes the file types, data structures and naming conventions to aid long-term preservation and reuse*
	Inventory the file formats in use
	Identify repository specific file formats required (NARA/LOC/Project)
	Plan for data set conversion to required file formats
	Have you documented any naming conventions used within your files?
Data Prioritization	*This describes how the project decides what data is prioritized for preservation activities*
	Identify program/project data priorities
	Identify organizational records retention requirements
	Identify data needing to be preserved to meet legal requirements

Storage and Geographic Location	*Storage systems, locations, and multiple copies of data/metadata to prevent loss of data*
	Have you identified where your data and metadata will be stored for the long term?
	Has the data and metadata been moved from hard drives, workstations and other external media to the organization data storage system?
	Have you identified the long term repository for your data?
	Do you have at least two complete copies of your data and metadata? Is one of them stored in a different location?
Data Integrity	*Procedures to prevent, detect and recover from unexpected or deliberate changes to data*
	Describe any processes you have in place to check that the data has been unchanged.
	Describe any QA/QC processes you have for data integrity.
	Describe any processes you have in place for identifying and correcting any changes to data.
Information Security	*Procedures to prevent human-caused corruption of data, deletion and unauthorized access*
	Identify who has read, write, move and delete authorization to remove individual files and data
	Document access restrictions for content
	Maintain logs of who performed what actions on files, including deletions and preservation actions
Financial	*Identify long term costs for data preservation, data refreshing and data archiving*
	Costs for data preservation can be planned for as part of the data lifecycle
	Planning for periodic data refreshing and associated costs (every 5-10 years)

FAIR Data Principles	
Findable	
	Are the data assigned a globally unique and persistent identifier
	Are the data described with complete metadata
	Does the metadata clearly and explicitly includes the identifier of the data it describes
	Are the data and metadata registered or indexed in a searchable resource?
Accessible	
	Are the data and metadata retrievable by their identifier using a standardized communications protocol?
	Is the communications protocol, open, free and universally implementable?
	Does the protocol allow for an authentication and authorization procedure, when and where necessary
	Is the metadata accessible, even when the data are no longer available?
Interoperable	
	Does the data and metadata use a formal, accessible, shared and broadly applicable language for knowledge representation?
	Does the data and metadata use vocabularies that follow FAIR principles
	Does the data and the metadata include qualified references to other applicable metadata/data
Reuseable	
	Does the metadata richly describe the data with complete, accurate and relevant attributes
	Are the data and metadata released with a clear and accessible data usage license
	Are the data and metadata associated with detailed provenance
	Does the data and metadata meet domain-relevant community standards?

Resources for More Information

National Archives and Records Administration (NARA)

- Digital Preservation Strategy 2022-2026 *https://www.archives.gov/preservation/digital-preservation/strategy*
- The National Archives and Records Administration (NARA) maintains a current list of acceptable file formats for all data required to go into a federal data repository. *https://www.archives.gov/records-mgmt/policy/transfer-guidance-tables.html*

The Library of Congress (LOC) *also maintains a list of currently acceptable file formats if your content will be subject to storage at the Library of Congress*. *https://www.loc.gov/preservation/resources/rfs/geo-carto.html*

FAIR Data Principles

- Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18
- Thompson, P.T., Ojha, S., Powell, C.D. *et al.* A proposed FAIR approach for disseminating geospatial information system maps. *Sci Data* 10, 389 (2023). https://doi.org/10.1038/s41597-023-02281-1

National Digital Stewardship Alliance (NDSA) Levels of Digital Preservation

- Bailey, Jefferson, et al. "The NDSA Levels of Digital Preservation: An Explanation and Uses." (2013). https://digitalpreservation.gov/documents/NDSA_Levels_Archiving_2013.pdf
- **NDSA website**: https://ndsa.org/publications/levels-of-digital-preservation/

Digital Preservation Strategies

Shimray, Somipam R., and C. Kodanda Ramaiah. "Digital preservation strategies: an overview." *11th National Conference on Recent Advances in Information Technology-2018*. 2018.

https://www.researchgate.net/profile/Somipam-Shimray/publication/327221006_Digital_Preservation_Strategies_An_Overview/links/5b80da10a6fdcc5f8b6592f4/Digital-Preservation-Strategies-An-Overview.pdf

Data Management Plan Templates

- **United States Geological Survey (USGS) Data Management Plan** website: https://www.usgs.gov/data-management/data-management-plans
- **Data Management Plan (DMP) Tool Templates** (A public collection of a wide variety of data management plan templates including Department of Defense (DOD) and Department of Energy (DOE): https://dmptool.org/public_templates