

# Data Preservation as a Legacy

Denise Bleakly, retired Sandia National Laboratories, GISP

April 2025

# 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.

# Data is Growing, Growing....GONE!

- Data preservation is not a GIVEN! Even well curated data can be corrupted, lost or erased.
- One legacy you can leave your projects is well managed, well curated and well archived data.

## 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**

# Myths about digital data

**Myth: Digital data lasts forever**

- **Reality:** Digital data is fragile and susceptible to decay

**Myth: Copying data to a new storage format preserves the data**

- **Reality:** Copying data to a new format is part of data preservation, but the contextual metadata and associated files are also needed

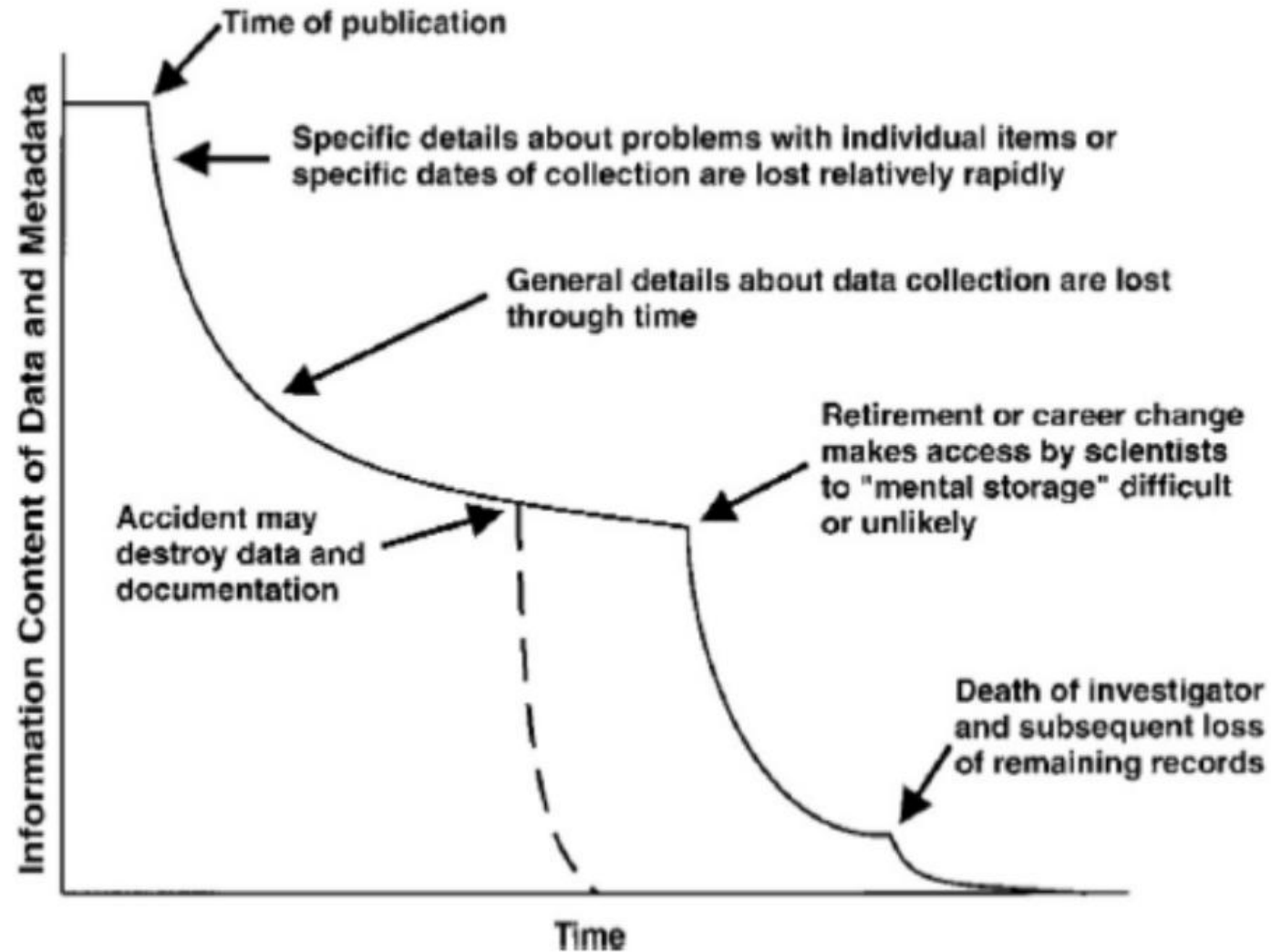
**Myth: All digital data can be kept**

- **Reality:** Not everything can and should be saved.

**Myth: Paper is no longer needed to preserve information**

- **Reality:** Printed copies on acid free paper with ph-neutral inks can last 100's of years. Some organizations are creating paper or microfiche copies of their most valuable records or data

# Loss of Data Over Time



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

# Can we actually preserve data for more than 50+ years?

**The answer is NO!**

- “If our goal is to preserve the information for a century or longer, it is evident that any system, no matter how well designed or well-supported or preservation-supporting, is destined to become obsolete and unsupportable long before the century mark.”
- However, There are data curation activities and preservation steps that make it more likely that geospatial data will survive over the long term

# Historical 100 Year Context

“If we consider our 100-year reference timespan, archives from the middle to the end of that span will be faced with curating information for which all links to the original creators and context have been severed.

To see this, one only has to consider the challenges, in the year 2009 of curating materials created in 1959 or 1909.”

OR

***Consider digital files created in 2024...What data from 1924 or 1904 can you find or use?***



# How can we preserve spatial data for the long term?

## *What is realistic?*

- A series of data package hand-offs that occur at many levels:
- Between different types of media and storage subsystems
- Different object frameworks and organizational schemes
- Different repository systems
- Different institutions and policy regimes
- Different application communities with diverse assumptions and interests.

# 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

# 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.

# 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

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

*Geospatial data may require extensive, product-specific context to interpret and render making preservation efforts difficult*

*Just think...*

- How many pieces of software do you use? Is it documented?
- How many models or modifications do you have to your scripts? Are they documented, noted and connected to the data?
- Or, after finishing a project, did you document what software(s) you used, any models, any testing, any QA/QC activities, and connect it with the project data? Is it in a format that people can access and is it connected to the data?

## Digital Data Preservation is a multi step process that involves many different elements

- Digital preservation will be achieved through a digital preservation infrastructure that ensures data integrity, format and media sustainability, and information security

# Why Does Geospatial Data Preservation Cost So Much?

- **Knowing what to migrate** requires knowledgeable subject matter experts and information professionals
- **Migration of the data** itself to new versions of software and hardware
- **Building and maintaining indexes** to preserved/archived data
- **Cost of hardware/software** to build spatial data repositories; License agreements for archive software
- **Costs of archive maintenance** for archive software Maintaining functionality and context of archived data
- **Creating and maintaining FGDC compliant metadata** that will contain information about the geospatial data being managed and will be key for decisions about the data into the future

# 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).



# Elements of a Data Preservation Strategy

## 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

- A data preservation strategy will identify what the records requirements are
- Will allow for planning for periodic records to be created as to where data resides, where the metadata, data dictionary, and where the contextual information is stored

## Legal

- A data preservation strategy can address regulatory requirements (such as permit requirements, other Federal Government Requirements, state requirements, etc.)
- A data preservation strategy can identify and address any data ownership and data Sharing requirements will allow for planning for periodic “archives” (every 10 + years)

## Financial

- 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)

# What Digital Preservation IS NOT

- **Digital preservation is NOT the same as digitization projects**

The confusion comes from the mistaken belief that anything digital will last forever.

Anything digitally born and created needs preservation measures just as much as anything digitized at a later stage

- **Archiving is NOT the same as preservation**

Archiving is a first step in preservation not the final step

- **Long-term Storage is NOT the same as preservation**

Digital preservation is not just a “storage issue”

There is no guarantee that data stored is preserved in such a way that the data can still be read and understood in the future.

- **When systems are replaced**, tacit knowledge, appropriate software and documentation are often lost.

- **Preservation is NOT Permanent Access**

A file can be accessible but not understandable if the metadata and context information is lost.

# Key Data Preservation Practices

1. Storing data in well supported, open formats
2. Use widely adopted standards
3. Bundling data, metadata and context information together using something like “Bagit”
4. Store a graphical representation of the data ( e.g., as a pdf) with the data and metadata bundle
5. Data should be free from external dependencies
6. Use persistent identifiers/Digital Object Identifiers (DOI’s)
7. Ensure all information objects are self-contained and independently understandable.
8. Preserve geographic data in a way that non geospecialists can understand
9. Plan for technological obsolescence – media migration every 3-5 years, data format migration every 10-25 years
10. 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.

## Develop Data File “Breadcrumbs”

Even with a good data life cycle plan and a good record keeping system, we discovered that data can be lost...

- **A File Plan** - is one way of periodically leaving a "bread crumb" in the records file system that describes the following:
  - What the geospatial data is
  - Where the geospatial data is stored
  - Currency of the data
  - Status of the data

# What Have We Learned About Data Preservation?

“Preserving digital information for a century will require a series of handoffs, occurring repeatedly at many levels:

- between different types of media and storage subsystems,
- different object frameworks and organizational schemes,
- different repository systems,
- different institutions and policy regimes, and
- different, diverse application communities.”

# How Can I Start with Data Preservation?

Gradually a set of best practices is emerging for  
ensuring digital data continuity

- **Develop a “7<sup>th</sup> Generation” Mindset**
- **Begin with the End in Mind**
  - For every new project or data set, create a data management plan
- **Start Early**
  - Waiting for a task or report to be completed is too late to preserve critical information
- **Data Preservation is a Team Effort**
  - Each team member has a role to play – documenting the data set, regularly backing the data up, working with records and data managers, working with IT staff for data storage


# 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](https://library.wheaton.edu/sites/default/files/Digital_Preservation_Plan.pdf) 

## **Oak Ridge National Laboratory, Distributed Active Archive Center (DAAC) – Best Practices for data management; Preserve: Protecting Data for Long-Term Use**

<https://daac.ornl.gov/datamanagement/>

## **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> 4/4/2025

# 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***