SAND2021-14023C

A Geospatial Data Preservation and Curation Strategy at the DOE Office of Legacy Management – 22411

Denise R. Bleakly*, Joshua I. Linard **, and Matthew Cuneo***
*Sandia National Laboratories
**DOE Office of Legacy Management
***RSI Entech

ABSTRACT

Identifying and planning preservation and curation activities associated with geospatial data will improve the ability of the U.S. Department of Energy Office of Legacy Management (DOE/LM; hereafter, LM) to support its core mission of protecting human health and the environment. Legacy Management is expected to manage multiformat digital information content, including text documents, geographic information system (GIS) databases, geospatial data sets, survey information aerial photographs, satellite imagery, ground photography (documenting surface conditions), spreadsheets, and relational databases. Due to the fragile nature of digital materials and continually evolving hardware, software, standards, and file formats, challenges exist in implementing an effective geospatial data preservation plan. This paper documents the LM strategy for preserving and curating geospatial data within the context of its data lifecycle-management framework.

The strategy consists of different preservation and curation elements, specific activities, and key enabling factors to ensure LM geospatial data maintains its visibility, accessibility, understandability, linkability and interoperability, trustworthiness, and security (VAULTS). Preservation elements enable the effective preservation of LM geospatial data and recognize the need for flexible strategies to adapt to ongoing changes in scale, technology, and standards. Digital preservation activities will leverage a digital preservation infrastructure to ensure data integrity, format and media sustainability, and information security. To meet its preservation and curation objectives, LM will address key enabling factors, which include responsibilities traditionally associated with program sponsors, intended to highlight critical data management responsibilities.

Preservation strategies involve establishing the fundamental logic necessary to evaluate ramifications on geospatial data encountered through its life. Documented standards and procedures will provide guidance on minimum metadata and preferred file formats. A risk-based approach to setting preservation priorities is recommended to establish scheduled assessments to identify at-risk formats. Geospatial data will be managed in a repository designed to provide reliable, long-term access to its owners. Data authenticity will be determined and maintained to ensure the record accurately represents the original. Metadata and ancillary data documentation will be created to provide essential contextual, administrative, descriptive, and technical information. To identify emerging risks, practices, and standards to continually improve its geospatial data preservation program, LM will actively engage with the DOE and local, national, and international digital preservation communities to share information and experiences, seek guidance, and collaborate to address digital preservation challenges.

Digital preservation activities comprise the specific operations and maintenance necessary to fully leverage geospatial data preservation strategies. Infrastructure comprising hardware, software, networks, storage, related equipment, and facilities used to develop, test, operate, monitor, manage and/or support information technology services need regular evaluation to ensure they meet preservation requirements. Detailed processes defining data integrity and quality checks must be regularly conducted and assessed for continued relevance. Data format and preservation media must be regularly assessed to ensure geospatial data sustainability. Information security processes require definition for controlled access to preserved data and change documentation.

Long-term geospatial data preservation requires commensurate budget resources to ensure achievable goals, a human resources capital plan to ensure knowledge management, and advocacy to clarify to senior managers the organizational value of sustained data governance and infrastructure.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

INTRODUCTION

The objective of geospatial data preservation is the long-term retrievability and use of data within its original context, not just the data content. Geospatial data is key to the U.S. Department of Energy Office of Legacy Management (DOE/LM; hereafter, LM) mission, and its preservation will be the foundation of current and future data and information preservation efforts. [1] The importance of data preservation was recognized by the US government in that the U.S. National Archives and Records Administration (NARA) was established and preserves records to protect citizens' rights, ensure government accountability, and document the national experience. [2] Inclusive to that responsibility, the NARA provides access to the digital record and digital surrogate content determined to have sufficient historical or other value to warrant continued preservation by the Federal Government per 44 U.S.C. §§ 2107 and 2203(g). [3] [4] The determination of data value is often coordinated with federal agencies to create data and with specific regard to geospatial data, agencies must ensure data are included on agency records schedules approved by NARA. [5] Ultimately, some time in the future (more than 75+ years) LM will transfer permanent records to NARA, according to both LM and NARA records retention schedules.

Legacy Management creates numerous data types to support its core mission of protecting human health and the environment and has been charged with the responsibility for long-term surveillance and maintenance, workforce restructuring and benefits, property management, land use planning, and community assistance for 101 legacy sites in the US and the territory of Puerto Rico (Fig. 1). As part of that mission, the LM Records, Information, and Knowledge Management team is responsible for policies, protocols, and procedures for preserving and dispositioning unclassified sensitive and unclassified program materials and administrative records and information from LM sites. Associated record schedules indicate the duration for which they must be managed (e.g., 75 years). Historically, these records and information mostly comprised of hard-copy documents, although the receipt of digital records and information is increasingly common and provides new preservation challenges, especially when considering large or complex digital data.

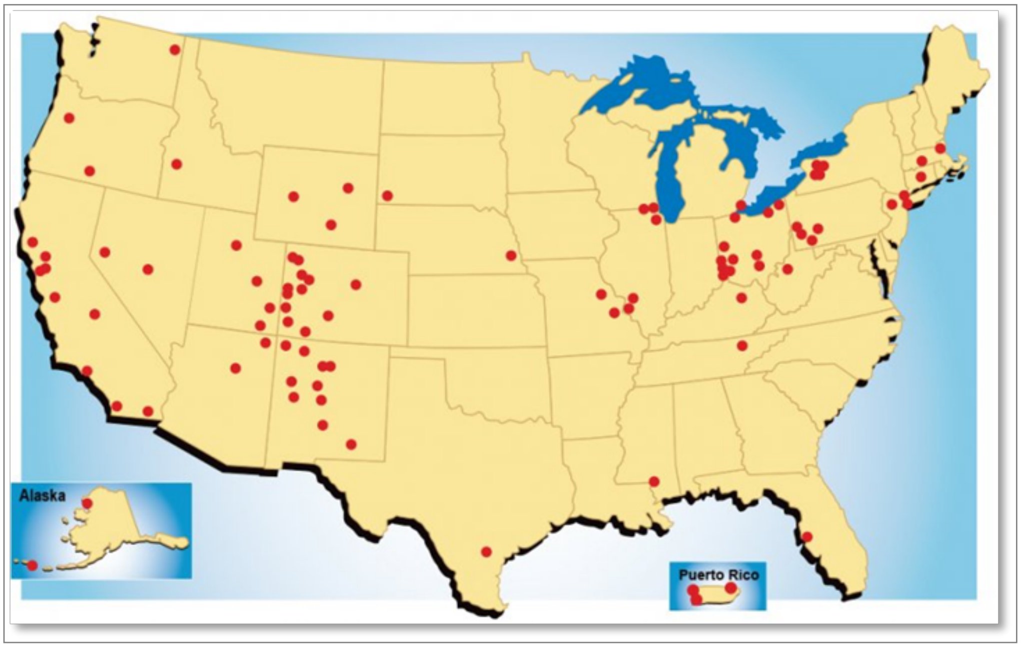

Fig. 1 DOE Legacy Management Legacy Site Locations.

Geospatial data, one example of complex data, consists of numerous related data arrays, data layers, and data formats that require sophisticated applications and infrastructure for use, and requires subject matter experts (SMEs) to understand the content and context. With LM, the Environmental and Spatial Data

Management (ESDM) team is expected to manage all types of geospatial data comprising temporally variable two- and three-dimensional survey coordinates, boundaries, and areas (i.e., Global Positioning System [GPS] locations, fence lines, property boundaries, remotely sensed imagery, digital elevation models, analytical samples, or groundwater contaminant plume extents). To ensure this kind of complex data has the context and content necessary to meet preservation expectations, the data must be curated throughout its lifecycle (Fig. 2). The LM Data Lifecycle includes steps relating to the ingestion of data to the geospatial database system and the provisioning of the data into the master database after initial quality control and quality assurance and documentation have occurred and the data has been entered into the process of data curation for its lifespan.

Data curation activities are those associated with active data maintenance, in contrast to data preservation, which is associated with the long-term retrievability of data in its original context. The objectives of curation are to ensure the data remains visible, accessible, understandable, linked, trusted, and secure (VAULTS). [6] [7] [8] Legacy Management is using VAULTS as a data curation tool instead of the more recognizable FAIR (Findable, Accessible, Interoperable, Reusable) data practices to remain consistent with internal LM data initiatives and to include authoritative data and data security in its data curation workflow. [9] Data curation begins during planning when data is identified for achieving organizational goals and initial preservation requirements established (e.g., retain the data for 75 years). As the data moves through its lifecycle, information is accumulated so, when the data is ready for preservation, there is confidence that a future user has enough information to understand its provenance. At some duration threshold (e.g., 10 years), the preserved data must be assessed to reduce threats to its long-term value and digital obsolescence. If risks are identified, a mitigation planning effort initiates a new lifecycle at the end of which data is preserved again.

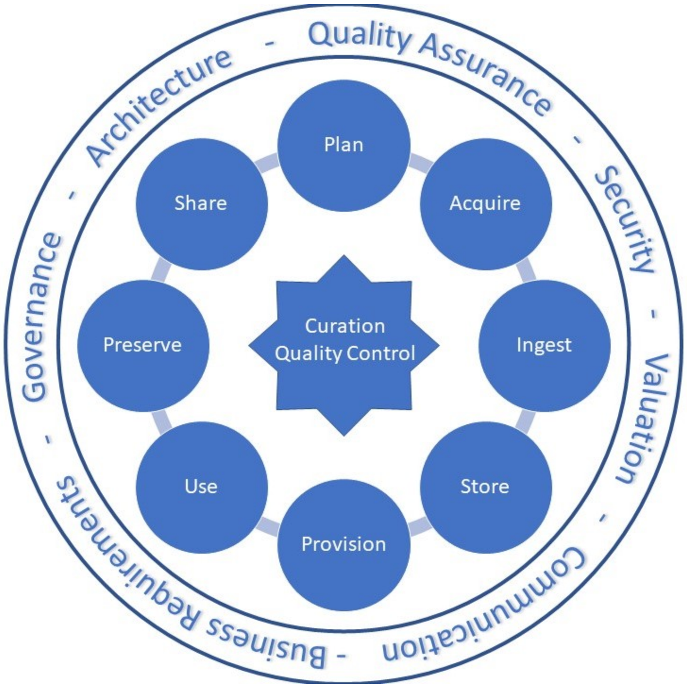

Fig. 2. DOE Legacy Management Data Lifecycle.

Risks to the long-term retrievability and use of data within its original context increase with longer retention schedules and is complicated by workforce succession, evolving technology, and organizational changes. Preservation strategies must be adopted to capture as much information as possible, so successive personnel understand as much context as possible. Understanding the context of data often relies on capturing metadata by personnel engaged in the initial data lifecycle, from planning through use.

Technology continues to evolve and can impact data, supporting applications, and underlying hardware. Documentation of the current data management framework and the activities associated with maintaining that framework can facilitate the transition to emerging technological advances. Organizations can also change over time in terms of mission, scope, and structure and can impact the ability to sustain resources needed to preserve data through their retention schedules. Addressing key enabling factors, traditionally associated with data sponsors, is necessary to mitigate those risks.

PRESERVATION STRATEGY DEVELOPMENT PROCESS

The LM team has initiated a process to develop a geospatial data preservation strategy to preserve the geospatial data entrusted to LM. The team used the LM data lifecycle management framework (Fig. 2) as a roadmap to creating a preservation strategy that would leverage the data lifecycle framework and assist in identifying concrete actions that could be employed as part of data curation and preservation activities.

Geospatial data is a subset of digital data, which have unique characteristics and challenges to curation activities and data preservation. [10] [11] In many instances, geospatial data may require extensive, product-specific proprietary formats and context information to interpret and render, making preservation efforts difficult. [11] [12] Some of the unique and challenging aspects of geospatial data are [11] [13] [14] [15] [16]:

- No Uniform Data Model – geospatial data are represented in a wide variety of data types: vector and raster; topological and non-topological; and 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.
- 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.

Given the unique issues and challenges with geospatial data, the team was initially unsure where to start, what to include, or how to move forward in thinking about a digital data preservation strategy. As such, the team wanted to learn from others by understanding other government agencies' approach to digital data preservation work; to identify key process of the data lifecycle that may need to be explored more, (in which the team discovered that data curation was one of those processes it needed to develop more fully); and to identify the best data practices that will help with geospatial data preservation. These sources helped the team to frame discussions, discover important considerations, and to realize the topic was much more extensive than anticipated.

The key resources that helped frame the LM data preservation strategy include:

- The NARA Digital Data Preservation Strategy, which provided LM with the format and content it needed to address. [3] The team thought mirroring the NARA strategy would provide confidence in identifying the issues that would need to be addressed in a preservation strategy.

- The Wheaton College Library and Archives Digital Preservation Plan, [17] which provided an outline of data preservation activities to consider from a library and archives perspective and provided an understanding of a risk-based approach to prioritizing data sets for preservation, as well as some of the challenges libraries and archival programs face with data preservation activities.

- The Oak Ridge National Laboratory, Distributed Active Archive Center [DAAC) [18] [19] from which the team identified data management best practices, the need to assign a persistent digital object identifier, and a data preservation perspective from the data archive perspective.

- The National Digital Stewardship Alliance (NDSA) – The Levels of Digital Preservation [20] from its "Levels of Digital Preservation" chart, the team identified the five functional areas (Storage, Integrity, Control, Metadata, and Content) and the four levels of preservation activities (Level 1 – Know your content, Level 2 – Protect your content, Level 3 – Monitor your content, and Level 4 – Sustain your Content) of data preservation. This provided the team with direct actions to help increase the likelihood of digital geodata being preserved over the long term.

- The first six chapters of "*Ecological Informatics, Data Management and Knowledge Discovery"* [21] provide a grounding in the topics of data management and planning, scientific databases, quality assurance and quality control of scientific data, the creation and managing of metadata, and the preservation of data for long-term use.

- The United States Geological Survey (USGS), Data Management Website. [22] The USGS has invested time and resources to provide a data management website addressing the USGS requirements in a direct and helpful way. The site discusses and fully explores each aspect of the science data lifecycle, explains requirements under public law and USGS policies and procedures, and provides additional resources for training and further information. From this site, the team learned of a systematic way to explore the science data lifecycle of another federal government agency other than the DOE.

As LM collects and manages vast amounts of environmental and scientific data each year, these sources helped the team to better understand the nature of managing scientific data and tying it to a scientific data lifecycle. It also made the team aware of the need to curate data so when long-term data preservation actions take place, strategies have been addressed and implemented to increase the likelihood the data has been preserved.

The NARA, "*Strategy for Preserving Digital Archival Materials,"* provided the LM team the format and content needed to address preservation strategy elements, digital preservation activities, and key enabling factors [3] From this framework, the team developed how to address each element.

The resulting geospatial data preservation strategy consists of curation and preservation actions, key enabling factors, and best practices for digital data preservation. The preservation strategy presented here has been generalized for this article by removing specific LM actions and more generically discussing the actions and activities to enable adoption by other organizations. It is intended to address geospatial data determined by LM to have sufficient historical, regulatory, or other value to warrant ongoing and continued preservation.

PRESERVATION STRATEGY ELEMENTS

Legacy Management will employ the following strategies to enable the effective preservation of its geospatial content while recognizing the need for flexibility to adapt to ongoing changes in scale, technology, and standards. The goal is to reduce risk and achieve best practices to preserve and maintain access to its digital geographic content.

Documentation of Standards and Procedures

This element documents existing LM internal standards and procedures related to records and data management and are, wherever possible, tiered to DOE and NARA standards and procedures and documents how LM standards and procedures relate to guidance documents of each agency.

From this, LM develops and provides guidance on minimum metadata and preferred file formats for digital geospatial data and records transferred to LM, which promotes the use of open standards-based formats and accepted voluntary, community-based standards to help facilitate future access and preservation. LM uses NARA's preferred file formats for file acceptance. [23] Currently NARA's geospatial data preservation formats include ESRI Shapefiles, GeoTIFF, Geographic Markup Language (GML), Topologically Integrated Geographic Encoding and Referencing Files (TIGER), Keyhole Mark Up Language (KML), Vector Product Format (MIL-STD-2407), ESRI Arc/Info Interchange File Format TerraGo Geospatial PDF (GeoPDF) and the Spatial Data Transfer Standard (SDTS). [23]

Data Curation

Legacy Management practices data curation as part of the data lifecycle and as a preservation action. Data curation is the active management of a collection of data that reduces threats to data long-term value, it enables data discovery and retrieval, maintains data quality, and aids in reuse over time through data authentication, data archiving, metadata review and creation, digital preservation, data transformation, and mitigating digital obsolescence. In practice, data curation also involves the traditional archival activities of assessing and selecting data products and processing details for long-term preservation. [24] [25]

Legacy Management will use two sets of workflow tools to assist in the data curation activities –the Data Curation Network (DCN) C-U-R-A-T-E-D workflow, and the Department of Defense (DoD) V-A-U-L-T-S workflow. [26] [6] [8] The team chose these workflows because they were concrete examples of the types of activities needed for curation and these can apply to individual geospatial data sets or collections of geospatial data, and they help with the planning and budgeting for specific curation activities for datasets under LM responsibility.

The C-U-R-A-T-E-D workflow is a process for data curators to use to establish the condition of the data and associated metadata and documentation [24] (TABLE 1).

Curation Activities	
C	Check files/code and read documentation (risk mitigation, file inventory, appraisal/selection)
U	Understand the data (or try to) (revies files/environment, QA/QC issues, read me files)
R	Request missing information or changes (tracking provenance of changes and their reasons)

A	Augment metadata for findability (digital object identifier, metadata, discoverability, etc.)
T	Transform file formats for reuse (transformations to simpler file formats, conversion tools, data visualization)
E	Evaluate data for accessibility, interoperability, trustworthiness, and security, which also applies to software licenses, responsibility, standards, and metrics for tracking use
D	Document all curation activities throughout the process (for example, curation logs, correspondence, configuration management, and QA/QC checks over time.)

TABLE I Data Curation Network C-U-R-A-T-E-D Workflow

Legacy Management will also use the VAULTS workflow (TABLE II), a DoD workflow for data management activities [7] [8] which has been adopted by LM within its geospatial data management group [6]. Although the two workflows are similar, the VAULTS workflow addresses data integrity, trustworthiness and addresses data security issues, and C-U-R-A-T-E-D does not.

Curation Activities	
V	Visible – can the data be discovered?
A	Accessible – can the data be accessed?
U	Request missing information or changes (tracking provenance of any changes and why)
L	Linked and Interoperable – Is there enough documentation to determine whether the data is reusable? What formats are the data available in? What standards were followed?
T	Trustworthy – Can we assess the quality of the data?
S	Secure – Is the data secure? How is the data protected?

TABLE II. Department of Defense V-A-U-L-T-S Workflow

Additionally, LM will reappraise data periodically to determine whether a particular data set needs further curation through data migration, data refreshing, data archiving, placement in a data repository or whether it can be disposed. The reappraisal process also determines whether the dataset still meets:

- Business needs
- Legal requirements
- Regulatory requirements
- Historical documentation requirement(s)
- Other unique needs.

Prioritization

This element describes how LM will take a risk-based approach to setting digital geographic data preservation priorities and performs curation and preservation activities on a schedule derived by reviewing DOE, and other Federal government regulatory compliance activities required of LM. Performing scheduled assessments of the file formats in LM holdings will alert LM to at-risk formats for which practical preservation strategies are not yet determined or where the necessary actions are technically complex.

File Management

This element describes how LM will store geospatial data content in its Geospatial Enterprise system [6] to provide ongoing management and access to the content throughout its lifecycle. Legacy Management will manage and maintain trusted/authoritative copies of the geographic and geospatial data in its planned master digital repository. The mission of the digital repository is to provide reliable, long-term access to managed digital resources to its designated community, now and in the future. At some point, LM may federate with the Data.gov, or Geoplatform.gov platforms. [27] Legacy Management will minimize the number of file formats that must be actively managed by normalizing files into selected formats that retain the significant characteristics of the original format in the operational database, as well as in the longer-term low-access storage. The original data files will be retained as part of the record of the data in low-access storage.

Authenticity

Authenticity refers to the trustworthiness of the record as an accurate representation of the original. Legacy Management will ensure authenticity by documenting all digital preservation actions and file access logs, and complete conversion checks to determine the data conversion is successful and accurate as per LM data governance procedures.

Metadata and Data Set Documentation

This element describes how LM will provide geospatial metadata, contextual information, and additional documentation, as needed, for each dataset. Legacy Management is in the process of determining how to assign persistent digital identifiers (such as DOIs) [28] and record preservation metadata about each digital object, data stored as computer files, and requiring applications software for viewing, to aid in the preservation of digital holdings over time through manual and automated preservation processes. The metadata required includes the geospatial metadata pursuant to ISO 19115X/19139x for each geospatial digital object [29] [30], data package metadata/documentation that describes any software/hardware needed to interpret the data and any contextual information for the data object. Finally, preservation metadata, which provides essential contextual, administrative, descriptive, and technical information, are preserved along with the digital object.

Organizational Relationships

This preservation element describes how LM will actively engage with the DOE, local, national, and international digital preservation communities to share information and experiences, seek guidance, and collaborate to address digital preservation challenges. This engagement will help LM identify emerging risks, practices, and standards to continually improve its geospatial data preservation practices.

LEGACY MANAGEMENT DIGITAL PRESERVATION ACTIVITIES

Legacy Management digital preservation activities will undergo ongoing assessment using appropriate voluntary, community-based assessment instruments, such as the National Digital Stewardship Alliance (NDSA) Preservation Levels [20], which measure program capabilities and maturity. Digital preservation will be achieved through a digital preservation infrastructure that ensures data integrity, format and media sustainability, and information security.

Infrastructure

The LM digital preservation infrastructure (hardware, software, networks, storage, related equipment, and facilities used to develop, test, operate, monitor, manage and/or support information technology services) includes such topics such as:

- Storage, network capacity, systems, and tools for the ingest or creation, processing, active file management, and preservation of LM acquired and managed geospatial datafiles.
- Processes to regularly review and update systems and tools that may be developed or procured by LM to meet business needs.
- Affordable, managed, replicated content storage infrastructure for geospatial data managed by LM. Replication includes one preservation copy in a different storage environment, preferably in a remote geographic region.
- Tools to inventory all born-digital files and digital surrogates upon ingest.
- Tools for file format transformations to perform file migrations over time as formats become obsolete and at-risk.
- Standardized workflow processes for associating native digital files and digital copies (digital surrogates) with record identifiers and metadata and ensuring files are in appropriate preservation storage and access server locations (on-premises or in the cloud).

Data Integrity

Data integrity actions are those activities that review, ensure, and document a dataset is appropriately documented and has not been changed or modified from its original intended form. Examples include:

- Inventory all incoming files and log the results of all ingest events, as well as all later lifecycle events, such as format transformations, file movement, and audits.
- Ingest files, a process that includes malware scanning and the checking of file fixity, which refers to the validation that a file has not been altered from a previous state.
- Copy content off physical media, incorporating the use of write-blockers, devices that prevent accidental damage to the content on the physical media, as appropriate.
- Performs periodic audits of all born-digital electronic record files and digital surrogates stored in the preservation repository, including fixity checks.
- Repair and/or replace files with fixity issues.
- Perform yearly audits of logs to validate whether files in the preservation repository have remained unchanged and uncorrupted over time.
- Perform an annual sample audit of media containing permanent records that are retained by LM in accordance with the DOE Records retention schedule.
- Before media containing permanent records are 10 years old, recopy onto tested and verified new electronic media.

Format and Media Sustainability

Assessing data formats and media sustainability, is a process to determine the data file format, and evaluates whether the data format is sustainable by activities, such as:

- The characterization and validation of file formats at the point of ingest. Characterization refers to the identification and description of a file's technical characteristics, such as its production environment. It is usually captured by technical metadata. Validation refers to confirming that the file in hand conforms to the expected characteristics of its type.
- Create a process which identifies file formats that are no longer sustainable (e.g., are no longer created by or accessible through current software).
- Create normalized versions of at-risk format files. Normalization refers to converting all files of a particular type (e.g., maps, color images, etc.) to a chosen file sustainable format.
- Perform automated and manual format migrations or other preservation activities based upon

identified unsustainable file formats.

- Monitor the larger preservation community and technological environment for signs of unsustainability and obsolescence in formats, media, and equipment.

Information Security

Information security, is the preservation action that protects the data over the short and long term, includes:

- Identifying who has access to the physical media items; access to ingest and processing systems and services; and read, write, and execute authorization to folders and files on inhouse servers and in cloud storage systems
- Performing a scheduled review of individuals and groups who have read, write, and execute authorization to folders and files on servers
- Ensuring no one person has write access to all files
- Maintaining a system of record logs of actions on files, including deletions and preservation actions and who has done these actions.

KEY ENABLING FACTORS FOR SUCCESSFUL LONG-TERM GEOSPATIAL DATA PRESERVATION EFFORTS

Many factors will contribute to the ultimate success of the LM Geospatial Data Preservation Strategy. This section is intended to highlight the critical factors that must be addressed by LM to meet the objectives. The NARA "*Data Preservation Strategy"* [3] discussed what NARA considered to be the key enabling factors for enabling the data preservation, including budget resources, staffing resources, information technology infrastructure, guidance on standards, and guidance and policy on data preservation, which are addressed within the context of the LM data preservation strategy.

Budget Resources

With this preservation strategy, LM acknowledges digital geospatial data preservation is a significant additional responsibility for LM. As such, it will develop an analysis of the long-term budget resources necessary to perform the tasks of its Geospatial Data Digital Preservation Strategy. Resources for geospatial SMEs for data curation and preservation activities, interactions with LM recorded information management teams, and work with the trusted digital repository are items that will be addressed in the budget resource needs. By identifying future budget resource needs and planning for them in a longer-term budget documentation, LM expects that budget resources will be made available.

Staffing Resources

With this preservation strategy, LM acknowledges digital geospatial data preservation is a significant business process that crosses multiple business units. Legacy Management will develop a separate human resource plan to support this function, and lobby for the budget resources to obtain the staff needed for curation and preservation activities.

Information Technology Infrastructure

Legacy Management will require a planning process that identifies infrastructure needs to support digital preservation, including systems and tools, storage, network capacity, data integrity, and information system security. This should document relevant operational and governance processes, including those for forecasting for storage and network capacity and planning for and implementing additional capacity and technology refreshes.

Guidance on Standards to Records Creators for Geospatial Data

Legacy Management will continue to develop and promulgate guidance to LM transition sites (the long-term stewardship sites that have been remediated and transferred to LM) for technical, format, and metadata standards to ensure the sustainability of geographic data files and digital surrogates (files that have been digitized) of geographic products, such as maps. Over the long term, as LM implements the Geospatial Data Preservation Strategy, requiring the LM transition sites follow recommended data documentation and metadata documentation, file format requirements, and contextual information, should assist in the long-term sustainability of these types of geospatial data.

Guidance and Policy for Digital Geographic Data Preservation

As part of its mission, LM produces and receives geospatial data and hosts information from sites around the country. As part of this responsibility, LM will develop and promulgate further internal guidance and policy as it implements this geospatial data preservation strategy.

DIGITAL DATA PRESERVATION BEST PRACTICES

Data preservation best practices have been developed from many different perspectives and applications – from the scientific and environmental data community; [18] from the university data management and library community [12] [11] [31] [32]; from the large-scale data repository community, which has published guidelines for data users for submission of data [33] [19]; and from a consortium of European mapping agencies and state archives. [34] All these different sources for data preservation best practices had general and specific recommendations ranging from discussions on format types, standards, the inclusion of a graphic representation of the data in a data package, the use of persistent identifiers, metadata, contextual information, and suggestions of what users in the future might want or need.
A common core of 10 best practices for digital data preservation emerged from these sources:

1. Store data in well-supported, open formats
2. Use widely adopted standards
3. Bundle data, metadata, and context information together using a file packaging format, such as “Bagit”
4. Store a graphical representation of the data (e.g., as a pdf) with the data and metadata bundle
5. Ensure data is free from external dependencies
6. Use persistent identifiers/Digital Object Identifiers (DOIs
7. Ensure all information objects are self-contained and independently understandable.
8. Preserve geographic data in a way that non-geo-specialists can understand
9. Plan for technological obsolescence – media migration every 3 to 5 years, data format migration every 10 to 25 years
10. Apply the 3-2-1 rule: three data copies in at least two formats, with at least one copy stored in a separate secure location.

CONCLUSIONS

There is no grand strategy, action, or series of actions that will guarantee preservation of geospatial data through time. Legacy Management has the stewardship responsibility for managing its legacy site data for more than 75 years), which has resulted in developing the data preservation strategy discussed here. Due to file content issues, file format, metadata and contextual information, and the need for preservation, whether the preservation method will be either a single geospatial data set or a collection of geospatial data will be individually determined. By sharing its data preservation strategy process, LM intends to assist other agencies in understanding their internal data curation and preservation processes, and to identify areas for implementing best practices thereby increasing the likelihood geospatial data will be preserved well into the future.

REFERENCES

1. DOE Office of Legacy Management. 2016. *2016-2025 Strategic Plan, DOE/LM-1477.* Strategic Plan, Department of Energy. http://energy.gov/lm.
2. National Archives and Records administration. 2017. “Strategy for Preserving Digital Archival Materials.” Strategy. Accessed March 12, 2021. https://www.archives.gov/files/preservation/electronic-records/digital-pres-strategy-2017.pdf.
3. House of Representatives, Congress. 2008. “44 U.S.C. 2107 - Acceptance of Records for Historical Preservation.” Washington, D.C>: U.S. Government Publishing office, December 30. Accessed October 2024, 2021. https://www.govinfo.gov/app/details/USCODE-2008-title44/USCODE-2008-title44-chap21-sec2107.
4. Congressional Research Service. 2018. *The Geospatial Data Act of 2018.* CRS-R45348, Washington DC: Congressional Research Service. Accessed May 31, 2019. https//crsreports.congress.gov.
5. Federal Data Strategy. 2020. *Federal Data Strategy.* Accessed June 17, 2020. https://strategy.data.gov/overview/.
6. DOE Office of Legacy Management, Woolpert Engineering. 2019. *Enterprise Geospatial Stgrategy Implementation Plan (November 2019 Draft).* Implementation Plan, Department of Energy.
7. Department of Defense. 2020. “DOD Data Strategy: Unleashing Data to Advance the National Defense Strategy.” Washington D.C. , September 30 . Accessed July 2021. https://media.defense.gov/2020/Oct/08/2002514180/-1/-1/0/DOD-DATA-STRATEGY.PDF.
8. —. 2021. “DOD Memorandum: Creating Data Advantage.” May 05. Accessed July 2021. https://media.defense.gov/2021/May/10/2002638551/-1/-1/0/DEPUTY-SECRETARY-OF-DEFENSE-MEMORANDUM.PDF.
9. Wilkinson, M, M Dumontier, IJsbrand Jan Aalbersberg, and Et al. 2016. “The FAIR Guiding Principles for Scientific Data Management and Stewardship.” *SCI Data* 3 (160018). Accessed October 12, 2020. doi:https://doi.org/10.1038/sdata.2016.18.
10. Bleakly, Denise R. 2002. *Long-Term Spatial Data Preservation and Archiving: What are the Issues?* Research Report, Albuquerque: Sandia National Laboratories. Accessed April 13, 2020. doi:https://doi.org/10.2172/793225.
11. Janée, Greg. 2009. “Preserving Geospatial Data: The National Geospatial Digital Archive Approach.” Accessed June 30, 2020. http://www.ngda.org/docs/Pub_Janee_Arch09_09.pdf.
12. Janée, Greg, James Frew, and Terry Moore. 2009. “Relay-supporting Archives: Requirements and Progress.” *International Journal of Digital Curation* 4 (1): 57-70. Accessed February 21, 2021. doi: https://doi.org/10.2218/ijdc.v4i1.78.

13. Sweetkind-Singer, Julie, Mary L Larsgaard, and Tracy Erwin. 2006. “Digital Peservation of Geospatial Data.” Edited by Jamie Stoltenberg and Abraham Parrish. *Library Trencds* (Board of Trustees, University of Illinois) 55 Geographic Information Systems and Libraries (2): 304-314. Accessed October 2020.

14. McGarva, Guy, Steve Morris, and Greg Janée. 2009. *Technology Watch Report: Preserving Geospatial Data.* Digital Preservation Coalition. https://www.dpconline.org/docs/technology-watch-reports/363-preserving-geospatial-data-by-guy-mcgarva-steve-morris-and-gred-greg-janee/file.

15. Downs, Robert R. n.d. “Management, Curation, and Preservation of Geospatial Data: Introductory Perspectives.” *Journal of Map & Geography Liberaries* 11 (2): 117-122. Accessed October 2020. doi:10.1080/15420353.2015.1064848.

16. Lauriault, Tracey P, Peter L Pulsifer, and D.R. Fraser Taylor. 2011. “Chapter 2: The Preservation and Archiving of Geospatial Digital Data: Challenges and Opportunites for Cartographers.” In *Preservation in Digital Cartography, Lecture Notes in Geoinformation and Cartographys*, edited by M Jobst, 25-55. Springer-Verlag Berlin Heidelberg. Accessed April 20, 2020. doi:DOI 10.1007/978-3-642-12733-5_2.

17. Wheaton College Digital Preservation Team. 2019. *Digital Preservation Plan.* June 14. Accessed July 2021. https://library.wheaton.edu/sites/default/files/Digital_Preservation_Plan.pdf.

18. Cook, Robert B, Yaxing Wei, Leslie A Hook, Suresh K.S. Vannan, and John J McNelis. 2018. “Chapter 6: Protecting Data For Long-Term Use.” Chap. 6 in *Ecological Informatics*, edited by Friedrich Recknagel and William K Michener, 89-113. Springer. doi:DOI 10.10007/978-3-319-59928.1.

19. Oak Ridge National Laboratory. 2021. *ORNL Distribute Active Archive Center (DAAC) Data Management. [Online] 2021. [Cited: February 15, 2021.] https://daac.ornl.gov/datamanagement/.*

20. Levels of Preservation Revisions Working Group. 2019. “Levels of Digital Preservation Matrix V2.0.” National Digital Stewardship Alliance, October. Accessed April 16, 2021. doi:https://osf.io/2mkwx/.

21. Recknagel, Friedrich, and William K Michener, . 2017. *Ecological Informatics: Data management and knowledge discovery.* Springer.

22. United States Geological Survey (USGS). 2020. *USGS Data Management Data Lifecycle Model.* Accessed April 4/13/2020, 2020. https://www.usgs.gov/products/data-and-tools/data-management/data-lifecycle.

23. National Archives and Records Administration (NARA). 2020. *Appendix A: Tables of File Formats.* Accessed March 30, 2020. https://www.archives.gov/records-mgmt/policy/transfer-guidance-tables.html.

24. Data Curation Network (DCN). 2020. *The DCN Curation Workflow, the CURATED Steps.* Accessed December 3, 2020. https://datacurationnetwork.org/outputs/workflows/.

25. Bose, Rajendra, and Femke Reitsma. 2005. *Advancing Geospatial Data Curation.* Edinburgh: Digital Curation Centre. Accessed March 19, 2021. http://www.ukoln.ac.uk/events/pv-2005/pv-2005-final-papers/030.pdf.

26. Data Curation Network (DCN). 2018. *CURATE Workflow.* Accessed October 13, 2020. https://datacurationnetwork.org/outputs/workflows/.

27. Federal Geographic Data Committee (FGDC). 2020. “Geoplatform 2020.” *Geoplatform.gov.* July. Accessed October 22, 2020. https://www.geoplatform.gov/about/.

28. Cook, Robert B., Suresh K.S. Vannan, Benjamin F. McMurry, Daine M. Wright, Y. Wei, Alison G. Boyer, and J. H. Kidder. 2016. “Implementation of data citataions and persistent Identifiers at the ORNL DAAC.” *Ecological Informatics* (Elsevier) 33: 10-16. Accessed February 4, 2021. doi:https://doi.org/10.1016/j.ecoinf.2016.03.003.

29. Federal Geographic Data Committee. 2021. *ISO 191** Suite of Geospatial Metadata Standards - Federal Geographic Data Committee.* Accessed October 22, 2021. https://www.fgdc.gov/metadata/iso-suite-of-geospatial-metadata-standards.

30. Federal Geographic Data Committee (FGDC). 2019. “FGDC Technical Guidance: Data.gov and the Geoplatform Metadata Recommendations.” Technical Guidance, 32. Accessed April 7, 2020. https://www.geoplatform.gov/wp-content/uploads/2020/01/FGDC_Technical_Guidance_Datagov_GeoPlatform_NGDA_Metadata_Recommendations_20191209.pdf.

31. Smithsonian Libraries. 2018. “Smithsonian Data Management Best Practices Storage, Archiving, and Preservation Preparation.” Washington D.C.: Smithsonian Institution. Accessed April 16, 2020. https://library.si.edu/sites/default/files/tutorial/pdf/storagearchivingpreservation20180307.pdf.

32. Geospatial Multistate Archive and Preservation Partnership (GeoMAPP). 2011. “GeoMAPP Key Findings and Best Practices.” Accessed July 2021. https://geomapp.com/docs/GeoMAPP_ProjectFindings_BestPractices20111231.pdf.

33. Hook, Les A, Suresh K. Santhana Vannan, Tammy W Beaty, Robert B Cook, and Bruce E Wilson. 2010. *Best Practices for Preparting Environmental Data Sets to Share and Archive.* Guidance Document, Oak Ridge, Tennessee, U.S.A.: Oak Ridge National Laboratory Distributed Active Archive Center (DAAC). Accessed October 16, 2019. doi:10.3334/ORNLDAAC/BestPractices-2010.

34. Ronsdorf, Carsten, Paul Mason, Jonathan Holmes, Urs Gerber, Andre Streilein, Marguerite Bos, Arif Shaon, et al. 2016. “GI+100: Long Term Preservation of Digital Geographic Information - 16 fundamental Principles Agreed by Nation al Mapping Agencies and State Archives.” Edited by Digital Curation Center. *International Journal of Digital Curation* (Universty of Edinburgh) 11 (2): 156-168. Accessed April 20, 2020. doi:DOI: 10.2218/ijdc.v.