MINNESOTA GEOSPATIAL ADVISORY COUNCIL

Minnesota Geospatial Advisory Council - Archiving Implementation Workgroup Final Report

Introduction

The Minnesota Geospatial Advisory Council (GAC) partners with a cross-section of organizations that include city, county, regional, state, federal and tribal governments as well as education, business and nonprofit sectors, and other stakeholder groups that benefit from geospatial technology to further the coordination among the Minnesota geospatial community.

Data archiving of geospatial information encompasses a wide range of considerations and practices for preserving public geospatial records and historical materials. In 2018, the GAC authorized the creation of an Archiving Workgroup with the purpose of defining the guidelines, best practices, and procedures for archiving geospatial data in Minnesota so that a wealth of valuable geospatial data can be preserved and available for future use. The Archiving Workgroup aimed to engage with data stewards and stakeholders at various levels of government, academic institutions, private sector interests, non-profit organizations and citizens of the state, and to collaborate with the Minnesota Geospatial Information Office (MnGeo) to propose datasets and methods for geospatial data archiving. In August of 2019, the Archiving Workgroup submitted its report, which included a recommendation for an Archiving Implementation Workgroup. Over the course of 2020, this group worked to define recommendations for Minnesota's future geospatial data archive. The explorations and recommendations of the Archiving Implementation Workgroup both complement and extend the reports and recommendations of the Archiving Workgroup. The Archiving Workgroup operated on a higher, overview level of planning, which the Archiving Implementation Workgroup was designed to take a deeper dive into specific aspects of developing a geospatial archive in Minnesota.

The work plan was divided into five subgroups that each contributed to this report:

- Outreach & Education - build support for archiving geospatial data and engage with data creators at various levels of government, academic institutions, and other relevant stakeholders
- Program Design - recommend governance, staffing, and a coordination strategy
- Technology- determine technical infrastructure needs, file types, and workflows
- Pilot Exploration - develop a pilot project for archiving geospatial data
- Funding - explore funding strategies and develop recommendations

Summary of Recommendations

Program Design

The program design recommendations include a governance framework, staffing needs, and a coordination strategy for the archive. The **governance framework** consists of a GAC Archiving Committee, an Operations Group, and a Working Team. **Staffing needs** will be ongoing and include at least one full-time Archivist. Recommendations also include at least one Archival Assistant and a software developer − full time at first, part-time after launch. The **coordination strategy** includes an outline of the benefits for the Minnesota geospatial community, as well as specific roles for Data Providers, Data Consumers, and Project Sponsors.

Technology

The technology recommendations include details on file formats, metadata, and infrastructure. Regarding **file formats**, the archive will store the original formats, but may also choose to create and store alternative formats. The **metadata** necessary for archiving includes descriptive metadata, structural & technical metadata, and administrative metadata. In the area of **infrastructure**, there are recommendations for storage, discovery platform, and exit strategy. For storage, the archive will need to balance the needs of vector data, raster data, and LiDAR data with regards to storage space, as well as preservation and access copies. The discovery platform will likely be similar to many existing data portals. However, due to the nature of the resources in the archive, the interface may need additional functionality not typically available in other portals, such as temporal searching and filtering, a schema for defining complex item relations, the ability to parse large datasets, disclaimers about the nature of historical data, and digital object identifiers for data citations and long-term access. Consideration was also given to the need for a detailed exit strategy, as technology eventually needs to change and migrate.

Workflows

A key aspect of the archive implementation will be establishing workflows for acquiring archival data. Building on the previous work outlined in the Archiving Strategy Report, the group devised two paths for adding geospatial data to the archive: one path for **data ingested from the Commons** and another path for **items added directly from data providers**. All Commons data will by default be eligible for archiving and will be added to the archive on an annual basis or more frequently as deemed by the Archivist or data provider. Items added directly from data providers would include data that an organization stores locally on internal servers or physical media, as well as data from counties, cities, and other organizations that distribute resources on their own portals instead of the Commons. This data will receive curatorial review before being accepted, and the Archivist will work with the data provider to prepare files for consumption. Although this method will require more manual processing,

it may ultimately serve to facilitate the ability for new data providers to contribute to the Commons. The workflow recommendations also note that the archive system will require continual administration and management. These **internal preservation system activities** include data management practices, fixity checks, regular backups, accessibility considerations, and systems reviews.

Funding

With regards to a **funding strategy**, the recommendation is to pursue a legislative appropriation while also exploring the potential for grant funding during the implementation phase. If it is determined that the archive should be hosted at the University of Minnesota, then the leadership in the Libraries and Office of Vice President for Research will need to be engaged in order to get their support for a funding request.

Next Steps

1. Create an **Archiving Pilot Workgroup** to:
    a. Evaluate and test a range of potential archive technologies
    b. Create a proof of concept with a sample set of data
    c. Continue to perform community outreach
2. Continue to pursue **funding strategies** in order to build the foundation for a funding ask, likely in the 2023 legislative session.

Context

Benefits of Archiving

Early areas of exploration for the Outreach & Education team were to show which user groups and communities have a need for historical geospatial data, what would be the risks and costs of *not* having an archive, and what are the benefits of a statewide geospatial archive.

At a high level, needs for historical geospatial data include:

- Historical and cultural site investigations for assessment and/or compliance
- Environmental monitoring over time
- Field research
- Academic research

Risks of *not* archiving geospatial data include:

- Data loss − storage media and formats change over time
- Ephemeral data - distinct versions of data may be overwritten with updates and lost
- Duplication of effort across government agencies
- Barriers to access − data requests are time-consuming and reduce overall usage
- Cost − agencies use staff time to provide data and satisfy retention schedules

Rewards of archiving geospatial data include:

- Save time and effort for data producers
- Create a single access point for historical geospatial data
- Support a broad community of users with shared needs
- Ensure timely, equitable access to historical geospatial data
- Increase usage of historical geospatial data
- Improved return on investment over time
- Facilitate an increase in the amount of free and open data in the Commons by providing support for contributors

There are efficiencies for data providers that would be gained by establishing an archive. For example, some existing resources in the Minnesota Geospatial Commons (Commons) such as the MetroGIS regional parcel dataset, may be managed differently once an archiving system is established. This data is currently saved as a yearly snapshot inside the Commons. Instead, the archive could ingest the older parcel datasets, relieving MetroGIS of maintenance needs. Another example is data that is held at the county or city level and is only available through local data portals or websites. This may be because of the relatively high threshold for metadata validation required to join the Commons. As these data providers opt into the archiving process, detailed metadata will be created with the assistance of the archive staff. This will allow these data providers to more easily submit their resources to the Commons, making them available for a broader audience of data consumers.

Recommended Timeline

Phase I: Research (2019-2020)

- **Archiving Workgroup (2019)**
  - Define guidelines, best practices, and high-level procedures
  - Engage data providers and stakeholders
  - Research funding possibilities

- **Archiving Implementation Workgroup (2020)**
  - Develop detailed recommendations for governance, technology, and workflows
  - Educate the community
  - Determine funding strategy

Phase II: Pilot & Proposal (2021-2022)

- **Archiving Pilot Workgroup (2021)**
  - Evaluate and test a range of potential archive technologies
  - Create a proof of concept with a sample set of data
  - Continue to perform community outreach

- **Archiving Proposal (2022)**
  - Draft a State of Minnesota legislative proposal for the 2023 session
  - Determine if a grant proposal is required for Phase III: Implementation

Phase III: Implementation (2023)

- Convene a GAC Archiving Committee
- Assemble Operations Group and Working Team
- Build technology infrastructure
- Ingest initial set of items

Phase IV: Ongoing Operations (2024 and beyond)

- Maintain and grow the archive
- Troubleshoot and upgrade technology and process as needed
- Evaluate continuing staffing needs
- Perform regular outreach to stakeholders

Outreach & User Education

The goal of the Outreach & Education team was to explore how archiving data would help support key governmental initiatives and statutory requirements, as well as continuing to build support for the archiving effort within stakeholder communities through educational messaging. To that end, between July 2020 and January 2021, the group devised, drafted, and distributed communications related to:

- Informational overview of archiving
- Benefits of archiving & the risks of not archiving
- Public policy and historical geospatial data
- Statutory requirements for records retention of GIS data

The Outreach & Education team also identified a list of stakeholder communities for targeted communications, including MnGeo, government agencies at all levels, academic researchers, U-Spatial affiliates, students/teachers/historians through the MNHS Library, all users of historical geospatial data, non-profit organizations, private sector, tribal nations, and State agencies / County administrators (non-GIS leadership positions). The main venues for distribution of the communications were the GovDelivery list and the MN GIS/LIS E-Announcements. Those messages were then forwarded to more specific communities.

In addition, in order to both promote archiving and gather user input, the team devised a survey (Archiving Historical Geospatial Data). Responses to the survey are ongoing and will be utilized in making a case for a geospatial data archive in Minnesota.

The Archiving Pilot Project Workgroup will carry forward the work of the Outreach & Education team (see Next Steps).

Program Design Recommendations

The Program Design Subgroup outlined a governance framework, staffing needs, and a coordination strategy for the archive. Due to ongoing uncertainty related to the COVID-19 pandemic, the recommendations do not include any specific designations of a host organization. Instead, the focus is on creating a general framework that provides the oversight, staffing, and coordination needed to meet the needs of data providers, archive users, government recordkeeping requirements, and more.

Governance Framework

The group researched distributed governance frameworks and determined that a structure similar to the one used by the Minnesota Digital Library (MDL) would fit the needs of the GIS community and meet government records management requirements. It includes a governing committee for oversight and strategic planning, an operational group to act as a source of subject matter expertise, and a working group of dedicated staff to handle day-to-day operations. This structure would provide the

archive with the right combination of resources, technical expertise, statutory oversight, and input from stakeholders to meet the needs of its users.

Table 1: Recommended governance framework

GAC Archiving Committee	**Responsible for:** programmatic decisions and strategizing about the future **Members:** representatives from the archive host institution, state archives, and statewide GIS community (state, county, city, academia, private sector, non-profit, tribal)
Operations Group	**Responsible for:** technology decisions about archive infrastructure and the discovery interface **Members:** archive staff and external advisors with specialized knowledge about the technology stack, analytics, and spatial data
Working Team	**Responsible for:** day to day communications, operations, and workflows **Members:** archive staff (Archivist, Archival Assistants)

The GAC Archiving Committee members and the Operations Group will dedicate time on a regular but limited basis. This labor will most likely be provided in-kind or as unpaid professional contribution. The Working Team will be composed of the archives paid staff (discussed in more detail below) and therefore will communicate on a regular basis.

Staffing Needs

The staffing needs of the archive will be continual, as its collections will require constant support in perpetuity. To meet these ongoing needs, the archive will need a full-time Archivist, one or more Archival Assistants, and one or more part-time technical support positions. During Phase III: Implementation, supplemental staffing will be needed to establish the administrative framework, build the technical infrastructure, and ingest the initial collections.

The Archivist, a single full-time position, will serve as the primary contact person regarding the archive and its administration. They will coordinate and attend meetings of the Advisory Committee and the Operations Group, liaise with data providers, provide reference services to researchers, make curatorial decisions as needed about what enters the archival collections, direct the work of Archival Assistants, collaborate with technology support providers, and work directly to process and manage

the archive's collections. The background knowledge needed for this position is primarily centered around digital archiving and metadata, with GIS experience strongly preferred but not required.

One or more Archival Assistants will support the Archivist's work. These employees will process and manage collections, add needed metadata, and perform assigned work under the Archivist's direction. They may be professionals, graduate students, or interns in Library and Information Science, GIS, or similar fields. Recommendations regarding whether or not the Archival Assistants should work full- or part-time and be permanent or temporary can be developed after a forthcoming pilot phase, which will assess the scope, time, and skill level required for the labor. This decision should also be regularly re-evaluated by the GAC Archiving Committee and will depend on the goals, timeline, and priority of the archive within the state's portfolio of projects.

Technological support will be necessary to develop the archive's discovery interface and back-end storage system that includes preservation actions. This work may require a dedicated developer initially, which may be a temporary contracted position. Additional ongoing support for infrastructure maintenance and troubleshooting will also be needed. If the archive is hosted by an institution with its own IT department, some or all of this support may be provided in-kind by the host institution.

Table 2: Description of the proposed positions

Archivist	• One person, full time • Archiving background required, with GIS background preferred • External face of the archive to data providers, users, and other stakeholders • Participate on Advisory Committee and Operations Group • Make curatorial decisions about archival collections, work directly with collections • Direct work on archival Assistants, collaborate with Technological Support
Archival Assistants	• 1+ people, part or full time • Either GIS or archiving background preferred, professional degree optional • Work directly with collections and metadata
Technology Support	• 1+ people; full time during development, part or full time thereafter • Previous experience with software chosen and GIS preferred • May be provided in-kind by archives' host institution

Coordination Strategy

A centralized archive for geospatial data will require additional coordination between stakeholders beyond what currently exists. At present, Minnesota data providers are the sole stewards of their data and may opt to distribute it to the public by self-submission to the Commons or a self-hosted portal. Data providers are fully responsible for metadata, publication, updates, and removal.

In contrast, the archive will be the steward of its contents. This stewardship may entail enhanced metadata, maintenance, and accessibility of data. A benefit of this service will be to ease the burden on data providers, who are currently responsible for retrieving information after the end of the data's active life cycle. The Archivist can also provide metadata assistance and quality assurance for data providers at the point of submission to the Commons, a service that is currently not available in the Commons model.

However, this level of centralized stewardship will be somewhat new to the Minnesota geospatial community and will require clear communication about the goals, processes, and capabilities of the archive. The archive will need to coordinate with several groups of stakeholders, including data providers, data consumers, and project sponsors.

- **Data providers** will be organizations that include (but are not limited to) government agencies, nonprofit and commercial groups, and academic institutions. Their interactions will include:
    - submitting their data to the archive indirectly via the Commons process that they already use
    - submitting their data to the archive directly by working with the Archivist
    - consulting with the Archivist for advice about metadata, file storage, and update frequency

- **Data consumers** will include data provider organizations, academic researchers, businesses, non-profit organizations, and the general public. The archive should be usable by people with a wide range of GIS abilities, from professional to novice. Data consumers will access the resources via a public-facing discovery platform. However, they may need to contact the archive to request resources, such as large imagery sets or other materials that cannot be delivered through the discovery platform.

- **Project sponsors** will be organizations that are involved in the creation and maintenance of the archive. These groups will likely include the Minnesota Geospatial Advisory Council (GAC), the Minnesota Geospatial Information Office, the University of Minnesota, and the Minnesota Historical Society/Minnesota State Archives. Members from these groups will sit on the GAC Archiving Committee, and they may provide staffing personnel for developing and operating the archive.

Technology Recommendations

The technology recommendations outlined in this report were developed by a collaborative effort between the Technology, Program Design, and Pilot Exploration Subgroups. Our recommendations cover file formats, metadata, and infrastructure.

File Formats

The archive will always store the original file formats of the data. To improve accessibility, the Archivist may choose to create and store alternative formats as well. This may be necessary in cases where the original format has become obsolete (e.g., Coverage files) or when a dataset needs to be partitioned (e.g., image services). The Archivist should consult format recommendations such as those released by the Library of Congress to assist with developing best practices.

Metadata

Descriptive Metadata

Descriptive metadata is primarily to aid users in finding and interpreting resources. Data providers should supply this metadata upon submission, but the Archives Working Team may need to add additional or updated information. This type of metadata is particularly important to the longevity of GIS data, as it includes projection information, data field definitions, and other information vital to the use and re-use of the resource. The Minnesota Geographic Metadata Guidelines (MGMG) will serve as the archiving system's descriptive metadata schema, and all data, regardless of type, will need to conform to this schema. Examples of descriptive metadata include:

- Title: the name of the resource; it may need to be updated when ingested into the archive to specify a version, dates, or other conditions
- Date Issued: when the resource was originally published
- Subject: a controlled vocabulary is recommended; the Commons uses ISO Topic Categories

Structural & Technical Metadata

MGMG contains numerous fields for structural and technical metadata that are needed for object storage. If this information is not already present, it may be captured during the ingest process or as part of the long-term management of the data. Examples of technical metadata include:

- File format(s): the original file format of the data and transformed format, if applicable
- File count: the number of files that make up a resource
- File size: individual file size and the total size of a group of files

Administrative Metadata

The most extensive addition to the resource's metadata will be in the form of administrative metadata. Management of preservation actions are done primarily by logging activity to show who did what and when within the system. These logs, also called administrative metadata, help ensure the archive's accountability and trustworthiness over time as well as the long-term accessibility of the archive's holdings. For example, if the Archivist opens a set of files and updates the descriptive metadata, the system should record who made the change, when it was made, and what was changed. Generally, administrative metadata is created automatically in the background of a system, but staff may also need to add or edit it periodically (e.g., to add information about the file transfer before ingest or to add notes about a server crash). Examples of administrative metadata include:

- Checksum value: used to verify that no changes have happened to the files over time
- Date of ingest: when the resource was added to the archive
- Software: the application and version used to process the files for the archive
- Access and usage rights: see the Archiving Agreement report for proposed values

Note: resources from the Commons will have some associated administrative metadata in a file called *dataresource.xml* that can be used as a starting point.

Infrastructure

Storage

The most important consideration for estimating storage needs is understanding the file sizes associated with GIS data types and formats.

- **Vector data:** Most geospatial data is distributed in vector formats, such as shapefiles, which have a relatively small file size. For example, the Geospatial Data Resource Site (GDRS) is a platform that hosts most of the data found in the Minnesota Geospatial Commons. Although there are hundreds of datasets in the GDRS, it is currently using only about 195 gigabytes (GB) of storage.

- **Raster data:** By contrast, the data that has been cited as the highest priority for the Minnesota geospatial community is imagery. This type of data uses a raster format, which can have a comparatively large file size. For example, the National Agriculture Imagery Program (NAIP) data for Minnesota is 3.5 terabytes (TB) for the year 2019 alone. Additionally, it must be understood that future annual collections of imagery data will be larger in size as the demand for imagery with greater detail increases.

- **LiDAR:** A third data type is LiDAR, which may be stored as vector, raster, or both. The state's current LiDAR collection is a little over 3 TB.

To archive the majority of this data, the archive should be implemented using servers with 20 TB of storage. This would allow for thorough testing of vector and raster datasets with multiple years (2 or 3 imagery collections each). Storage solutions, including virtual, cloud, or locally dedicated servers, should be evaluated to determine the overall cost of storage. An ideal technology stack for the archive would use two servers: one server to hold preservation copies of data and one server for access copies via a discovery website.

Discovery platform

An essential component of the archive will be a front-facing discovery platform for the public to find and access the archived resources. The discovery platform will have many features in common with existing data portals, including the ability to use free text searches, filter by metadata values, such as organization or category, and view summaries of the full MGMG metadata within the interface. However, due to the nature of the resources in the archive, the discovery website may need additional functionality not typically available in other portals.

- **Temporal searching and filtering**: In order to differentiate between versions of the same data theme across time periods, users should be able to filter temporally.
- **Item relations:** The site would also need a more complex schema for defining item relations. This would enable users to find both current and archived versions of data that may require moving back and forth between the discovery platform for the archive and external clearinghouses like the Commons.
- **Parsing large datasets**: The archive will likely provide access to large datasets, which may require subsetting, tiling, and providing download file size warnings.
- **Disclaimers:** Archived data is a static snapshot, and users may not realize they are using old data. There should be a disclaimer to that point added to the metadata or download page and, ideally, a link to the current version of the data.
- **Digital Object Identifiers (DOI):** Archived items will have a persistent identifier to reference for data citations and long-term access.

The discovery platform for the archive may have a different audience than existing data portals across the state. GIS professionals typically are searching for the most up-to-date resources. In contrast, visitors to the archive are more likely to be historical researchers or journalists looking to compare social or physical changes over time. It follows that many of the archive users may not have any desktop GIS skills or ready access to the software. To reach this audience, archived resources could be made available in alternative formats, such as non-spatial tabular formats or web services.

The discovery platform can be built with one of the many software platforms available, including custom, open-source, or proprietary options. One possibility is to use the same technical infrastructure as the Commons, which is built with an open-source application called Comprehensive Knowledge

Archive Network (CKAN). The CKAN interface is highly customizable, allowing for flexibility in the website design and data handling, which could allow the archive to have the same look and feel as the Commons. This approach has the advantage of providing a familiar user experience and the opportunity to seamlessly integrate the discovery platforms of the Commons and archive.

Exit strategy

Ultimately, any technology stack implemented will eventually need to change, whether through partial upgrades (such as new servers) or a complete sunsetting or migration of the system. A successful archive will offer multiple exit strategies that allow archives staff to migrate both data and metadata easily and accurately. It is particularly important to consider exit strategies with hosted storage and proprietary systems; costs, timelines, data transfer methods, file formats, and protocols for the deletion of remaining data should all be addressed in the vendor contract.

Data migrations to new systems should always be planned carefully and documented as fully as possible in the administrative metadata. Copies should also always be verified by using checksums to ensure that no files were lost or changed during the move.

Metadata migrations, particularly of administrative metadata, should use open and documented formats and transformations (also known as "crosswalks") wherever possible; they should further take advantage of any metadata validation tools that exist. To facilitate metadata migrations, open formats (such as plain text, XML, etc.) should be used wherever possible. When transforming metadata between formats (e.g., when exporting from a proprietary database), it is best practice to create several test exports to ensure that metadata is transformed in a way that conforms to metadata standards.

Workflows

Workflow Paths

There will be two paths for geospatial data to be consumed or added to the archive.[1]

- Path 1 will be for data ingested from the Commons.
- Path 2 will be for items added directly from data providers.

Path 1: Data ingested from Commons

All Commons data will by default be eligible for archiving. The GAC Archiving Committee and staff should perform regular outreach communications to the state geospatial community to make sure that contributors to the Commons understand that their public data may be archived. If a data provider does not want their resources archived, they may submit a request to opt-out of the default process.

Commons data will be added to the archive on an annual basis or more frequently as deemed by the Archivist or data provider. The schedule may also vary depending on the importance of the data, frequency of data updates, its temporal nature, historical significance (e.g., based on an event), or the Archivist's needs. If a retention schedule exists for the materials, it should be reviewed and followed. It may be useful to add a metadata field relating to the retention schedule of the materials. At the discretion of the Archivist, data that is deemed low value or physically corrupted may be excluded.

Once the data is determined suitable for archiving, there will be a grace period before the data is transferred to the archive. This is because errors are occasionally noticed after the resource has been posted in the Commons and are then quickly corrected. The grace period will help eliminate low-quality information from being ingested into the archive and allow data providers to feel confident that the intended dataset is being archived.

After the grace period, the data provider will be notified and the data will be ingested into the archive. The management of the archive system metadata and data transfers could be largely automated with a set of scripts modeled on the GDRS. A trigger field could be added to the Commons Geobroker to indicate that a resource should be archived, further automating and documenting archiving decisions. By integrating with the existing Commons infrastructure, Path 1 provides a reliable, low cost, high volume return for archiving geospatial data.

Path 2: Items added outside the context of the Commons

Path 2 would be used for data that an organization stores locally on internal servers or physical media. Typically, this data is not known to the public or is only available upon request. Path 2 can also be used for data from counties, cities, and other organizations that distribute resources on their own portals instead of the Commons. However, this path would primarily be for one-time data ingests of historical

[1] These paths correspond to the two data categories identified in the Archiving Strategy Report.

or unique resources. Resources that are regularly updated should be submitted to the archive via the Commons (Path 1).

All Path 2 data will receive curatorial review before being accepted. Data that is incomplete or not ready for reuse may not be accepted. If the data does not conform to a standardized format, it may need to be transformed or converted to a format that can be accepted by the archive. For example, data may be in an obsolete file format, or the data may be spatial in nature but not in a GIS format (e.g., a table of information with latitude and longitude).

In most cases, the Archivist will work with the data provider to prepare files for consumption, and there will likely be circumstances where the Archive Working Team does the majority of the preparation. Examples of manual work that the Archivist would need to complete include standardizing or converting file formats, writing configuration files, and creating directory structures. It is also likely the archive staff will need to assist the data provider with meeting the metadata requirements.

Although Path 2 will require more manual processing, it may also provide an opportunity for the archive staff to educate data providers on data management and metadata. Path 2 may ultimately serve to facilitate the ability for new data providers to contribute to the Commons and eventually use Path 1.

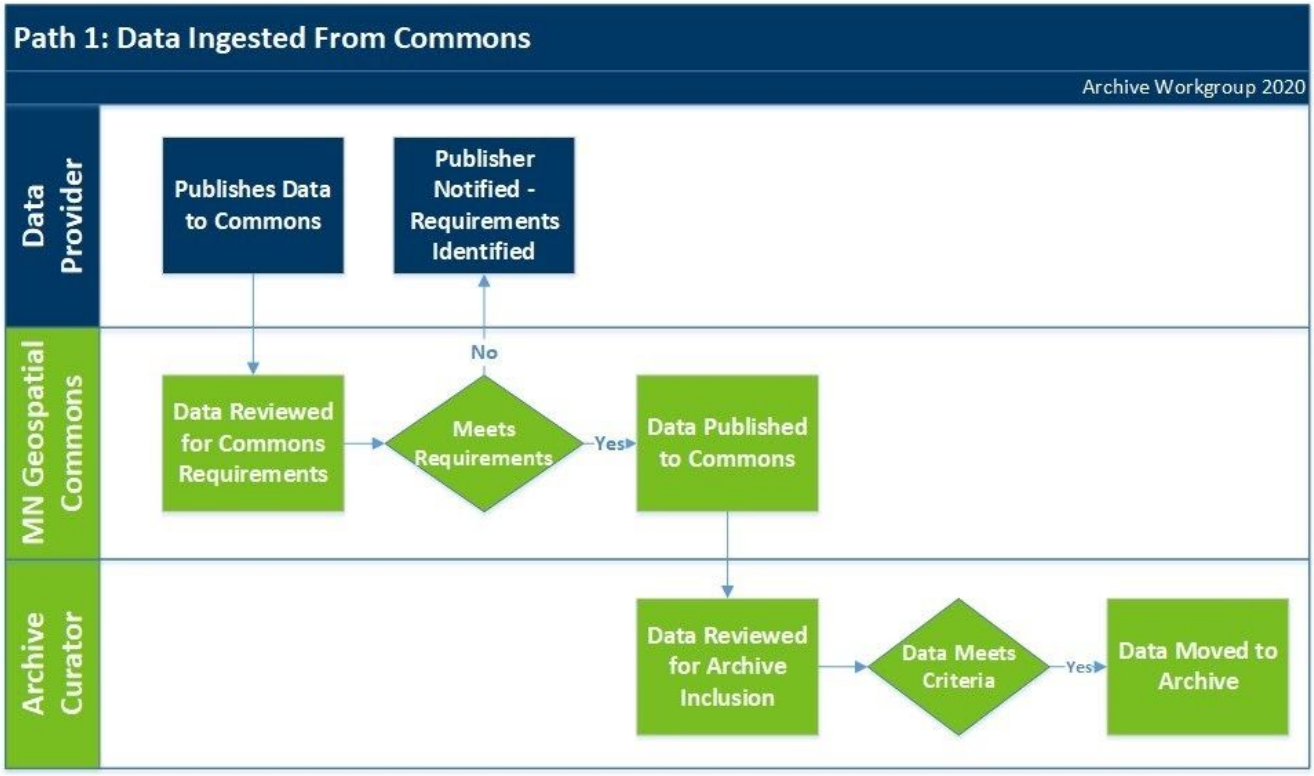

Figure 1: Overview of Workflow for Path 1

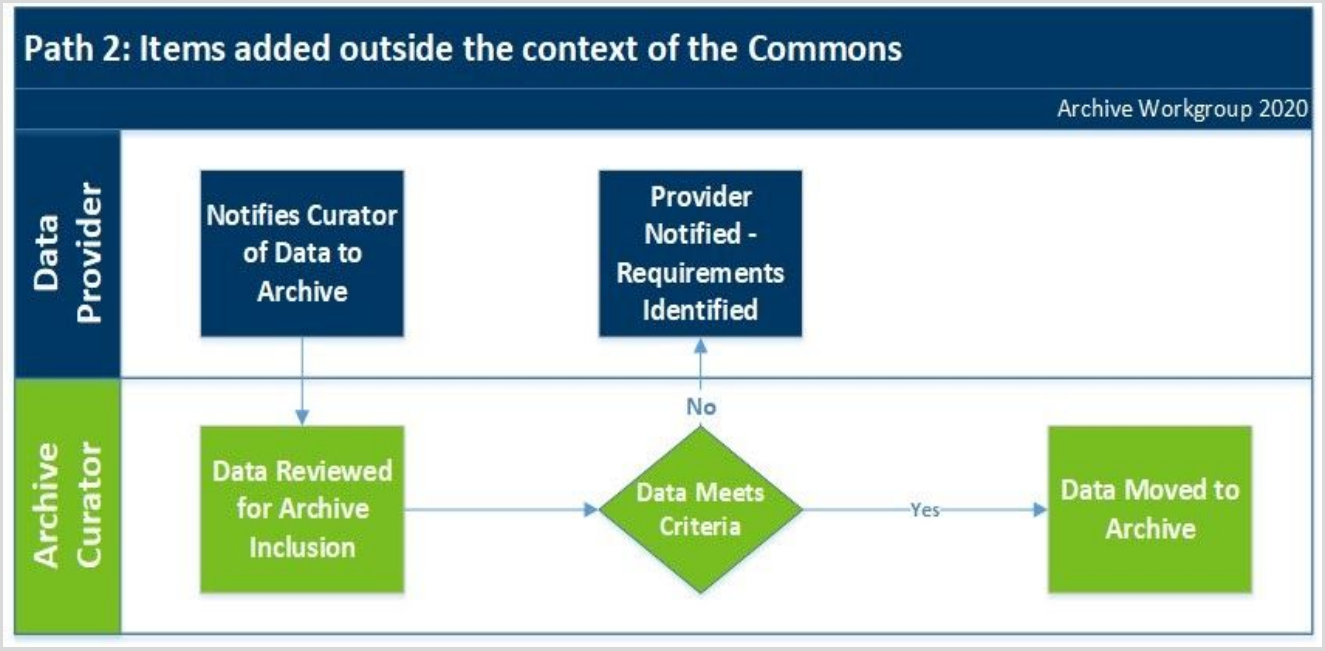

Figure 2: Overview of Workflow for Path 2

Internal Preservation System Activities

The archive system will require continual administration and management. These tasks are designed to ensure accurate preservation and continuous accessibility of the collections. Key activities include:

- Data management: Maintain a system registry to assist with understanding what datasets or files have been ingested into the system.
- Fixity: Perform annual reviews of the collections in the archives using checksums to spot bit-rot and file corruption.
- Backups: Create archivally sound backups of collections on a regular basis (multiple copies in multiple places for preservation purposes). Every time data is moved, the full resource should be copied, including metadata and layer files each time, not just the data. This has implications for the underlying data structure, so that each version of each resource would be a subfolder including the date of archive. To ensure authenticity, a checksum process will validate that new files always match the original files.
- Accessibility: Verify that file formats are still accessible and migrate obsolete formats to newer ones when needed and if possible.
- Systems Review: Perform ongoing reviews of processes and workflows to ensure they are as robust as possible and are keeping with evolving best practices in the field of digital archiving.

Funding Strategy

The Funding subgroup recommends a strategy involving a legislative appropriation (through the University of Minnesota if the archive is hosted there) while also exploring the potential for grant funding during the implementation phase. For a legislative ask, further exploration is needed to determine which University unit/department/college would apply for the funding. It will be important to engage leadership in the Libraries and Office of Vice President for Research to get their support for the archive. One step towards building support is to gather anecdotal evidence of the value of the archive from faculty at the University and other higher education institutions as they will be large users of the archive for research purposes. The subgroup realizes that this funding approach is a multi-year effort, further hampered by COVID budget constraints at all levels of government agencies.

Another strategy explored was to seek operating funds for the archive from various State agencies and other data producers (i.e. Metropolitan Council, counties). While this reduces risk of losing operation funds in the future by distributing the funding among several entities, it adds greatly to the amount of oversight needed in collecting yearly funding commitments. This strategy would also leave the funding stream in jeopardy if one or more of the agencies was no longer able to contribute.

Next Steps

The next step is to begin Phase II: Pilot and Proposal. The first part of this phase will begin in 2021 and includes creating an Archiving Pilot Workgroup to evaluate and test a range of potential archive technologies, create a proof of concept with a sample set of data, and continue to perform community outreach.

The second part of the phase will be to pursue funding strategies. The outcome of this work is anticipated to be a grant proposal to cover startup costs and a Minnesota Legislative Proposal for ongoing operations.

List of Workgroup Members

Sarah Barsness - Minnesota State Archives
David Brandt - Washington County
Jennifer Corcoran - Minnesota Department of Natural Resources
Jon Hoekenga - Met Council
Melinda Kernik - University of Minnesota Libraries
Len Kne - University of Minnesota
Leanne Knott - Goodhue County
Carol Kussmann - University of Minnesota Libraries
Brent Lund - MNIT / MnGeo
Karen Majewicz - University of Minnesota Libraries (Vice Chair)
Andra Mathews - Minnesota Department of Transportation
Ryan Mattke - University of Minnesota Libraries (Chair)
Nancy Rader - MNIT / MnGeo
Soren Rundquist - Environmental Working Group
Zeb Thomas - MNIT / Minnesota Department of Natural Resources
Ben Timerson - Minnesota Department of Transportation
Brandon Tourtelotte - Pro-West & Associates

Appendix: Glossary of Working Definitions

This list of working definitions was created to assist readers with developing a common understanding of the terms, concepts, and acronyms used throughout this report. Terms are listed in alphabetical order.

ArcInfo Coverage: An ESRI ArcInfo Coverage is a georelational data model that stores vector data. It is a legacy format and is no longer supported.

Checksum: A value that represents the exact bitstream of a file or set. Checksums are used to verify that data has not changed during a transformation or migration.

Commons: The Commons, sometimes known as the GeoCommons, a shortened name for the Minnesota Geospatial Commons, is a collaborative space for users and publishers of Minnesota's geospatial resources. A variety of resource types, which come from many sources associated with geographic locations, are available in the Commons. The Commons is used by researchers, cartographers, web and application developers, journalists, planners, and other citizens who need GIS data for a project.

GDRS: Geospatial Data Resource Site. A system of networked sites that enables organizations to share and regularly update data and applications. Under this model, an organization publishes its data and applications (resources) to its own GDRS node, and those resources are distributed to all participating networked organizations through a set of tools. When resources are shared in the GDRS, they can then be exposed to the public via the Commons; publishers authorize that exposure through the Commons "Geobroker" application. The organization supporting a full GDRS node maintains all publishing rights to its folders on the node. This method is generally preferred for state agency publishers.

GIS: A Geographic Information System is a digital collection of structured data and analytical tools related to space and place. GISs are created and used across the state for various reasons (e.g., streets, watersheds, archaeological sites, population data, and more). Their structure allows for a wide variety of uses, including visualizations, data analysis, and interactive services (dashboards, interactive maps, etc.). As official government records and as repositories of valuable information, GISs have enduring value, and many must be retained and accessible permanently.

GovDelivery: Email list that distributes information about work by MnGeo and their GIS partners at Minnesota state agencies.

Government Records: Government records include materials used, made, or received by an officer or agency of a government entity (including state, county, and municipal governments). Government records come in many formats such as ledgers, photographs, reports, maps, and GISs; regardless of their format or storage medium, government records should be managed, retained, and disposed of according to records retention schedules.

MN GAC: The Minnesota Geospatial Advisory Council is a coordinating body for the GIS community in Minnesota and includes representatives from state/county/local government, education, non-profit, tribal government, and business sectors.

MN GIS/LIS E-Announcement: Regular email newsletter from the Minnesota GIS/LIS Consortium.

MDL: The Minnesota Digital Library is a shared digital platform run by Minitex for accessing digital collections from cultural heritage organizations across Minnesota. In addition to working with contributing organizations, MDL has partnerships with several organizations, including the University of Minnesota, the Minnesota Historical Society, and the Digital Public Library of America.

MGMG: The Minnesota Geographic Metadata Guidelines are an official state guideline, adopted by MNIT Services, based on a federal standard called The Content Standard for Digital Geospatial Metadata, and help GIS professionals have a consistent approach to documenting and describing their work. This metadata ensures that a GIS is credible, usable over time, and reusable in other contexts.

MnGeo: Minnesota Geospatial Information Office is part of Minnesota IT Services and was established in May 2009 as the first state agency with legislatively defined responsibility for coordinating GIS within Minnesota. MnGeo succeeds the Land Management Information Center, which was exclusively devoted to providing GIS services within state government.

MNHS: The Minnesota Historical Society is a non-profit organization that provides government records services to the State of Minnesota, including the housing and management of the State Archives and acting as the secretary for the Minnesota Records Disposition Panel

OAIS: Open Archival Information System is a model used to describe how an archive, particularly an archive that collects and shares digital collections, should operate. It defines roles, workflows, and data types that serve as a framework for any institution seeking to store and share information in the long term.

Retention Schedule: Records retention schedules are formal documents that determine what records an institution keeps and how long they are kept. Records are organized by content rather than format, as formats frequently change; for example, plat books, maps, and GISs may contain similar information. Retention schedules for government entities are approved and managed by the Records Disposition Panel.

UMN: The University of Minnesota is a large and diverse educational institution and hosts many entities involved in building, maintaining, and/or staffing a GIS archive, such as the Libraries or U-Spatial.