MINNESOTA
GEOSPATIAL ADVISORY COUNCIL

Minnesota Geospatial Advisory Council Archiving Workgroup

Archiving Strategy Subgroup Report

The Archiving Strategy Subgroup was charged with creating an archiving strategy that includes guidelines, best practices, and procedures to be reviewed by the geospatial community.

Introduction

The topic of geospatial data archiving has been explored for over a decade, notably with several National Digital Information Infrastructure and Preservation Program (NDIIPP)[1] funded projects in the late 2000s, such as the Geospatial Multistate Archive and Preservation Partnership (GeoMAPP) and the National Geospatial Digital Archive (NGDA).[2] Since that time, a few states have developed ongoing initiatives to archive geospatial data, such as North Carolina.[3] Additionally, a few private academic institutions have also built spatial data infrastructures to accommodate archiving, including Stanford[4] and Princeton[5] Universities.

Despite this research and a few known successful projects, most public agencies in the United States have not implemented systematic geospatial data archiving. When it has happened, it frequently featured a one time investment, capturing geospatial data from one year or for a specific project before funding runs out. Although the initial startup costs of creating the Archive may be higher than subsequent years, a review of past projects demonstrates that the most important anticipated challenge of this initiative will be to establish plans for **technology, curation**, and **funding** that will be sustainable.

This report describes recommendations based upon research and discussions carried out in 2019. The recommendations are not intended as binding solutions, and they may be altered or adjusted by the GAC and subsequent committees or workgroups.

[1] http://www.digitalpreservation.gov/
[2] see appendix I for details on these projects
[3] http://digital.ncdcr.gov/cdm/home/collections/gis-data
[4] https://earthworks.stanford.edu/
[5] https://maps.princeton.edu/

Table of Contents

Introduction **1**
Table of Contents 2

1. Purpose, Goals, and Objectives **3**
Statement of Purpose 3
Goals 3
Objectives 3

2. Technology Plan **4**
Data Repository 4
Discovery Platform 4
Storage 5
Staffing 6
Tools 6

3. Curation Plan **7**
Classifying Data 8
Contributing Data 8
Ingesting Data 9
Preservation 9

4. Funding Strategy **10**
Estimate of Costs 10
Funding Sources 10

Appendices **11**

1. Purpose, Goals, and Objectives

Statement of Purpose

The Archive will be a repository for geospatial data with permanent historical/archival value.

Goals

- Create a repository where publicly available geospatial data are discoverable, accessible, and downloadable.
- Archive and preserve geospatial data that is at-risk for no longer being made publicly available by any other means.
- Make historical geospatial data available for planning, research, and teaching − for example: longitudinal studies, case studies, and impact analyses.
- Provide historical geospatial data to a cross-section of organizations that include city, county, regional, state, federal and tribal governments as well as education, business and nonprofit sectors, and any other stakeholder groups that benefit from geospatial technology.
- Develop a program for archiving Minnesota geospatial data.

Objectives

- Build a sustainable spatial data infrastructure, including discovery/access layer, archival storage, and (ideally) geospatial web services.
- Create mechanisms for identifying at-risk geospatial data.
- Develop a robust communications plan.
- Promote the use of historical geospatial data for research, teaching, and planning.
- Create an Archiving Committee for prioritizing data archiving activities, to include a cross-section of organizations that benefit from geospatial data in Minnesota.

2. Technology Plan

Data Repository

The foundation of the Archive will be a data repository that ensures that the resources will be safely and systematically preserved. This will require a system that features multiple copies, tape copies, geographic distribution, versioning, fixity checking, and preservation metadata.

The Workgroup recommends that the data repository be placed under the stewardship of an existing infrastructure initially, such as the University of Minnesota Libraries. Although the collection scope of the Libraries does not yet include government data, the University is the best example of an organization that already maintains an applicable technical infrastructure and relevant staff expertise.

Discovery Platform

Users will find and access the archived resources with an online discovery platform. The workgroup recommends the Minnesota Geospatial Commons ("Commons")[6] as the primary discovery platform. The data would be managed in the Commons as a new organization node. This will keep access to Minnesota geospatial resources in one site and allow seamless searching for current and archived datasets.

Storage

The storage requirements for the Archive can be approximated by examining the candidates for inclusion identified by the Workgroup and by making broad estimates.

The GDRS

A large portion of the future archived data will come from the Geospatial Data Resource Site (GDRS), the data structure that supports the Commons.[7]

- As of July 2019, the GDRS has 817 records and a total size of 123 GB.
- About 132 (about 20%) of the records in the GDRS are from counties or universities. The Priority Datasets Subgroup determined that the archive should primarily focus on state level data, at least initially.[8] Omitting county and university data leaves 685 records. Subtracting 20% from 123 GB reduces the size of potential candidates down to about **100 GB**.
- Analyzing the difference in total file storage in the GDRS between December 2017 and July 2019 shows a potential annual growth of about **1 GB**.[9]

[6] https://gisdata.mn.gov/

[7] https://gisdata.mn.gov/dataset/gdrsmanager

[8] https://www.mngeo.state.mn.us/workgroup/archiving/Priority_Datasets_Report.pdf

[9] GDRS sizes: December 2017 was 122 GB (56k files). July 2019 was 123 GB (63k files).

Historical Vector Layers

Historical vector layers that are not in the GDRS will also be archived, but a systematic inventory of this data has not yet been performed. A review of annually issued MnDOT vector datasets from the 1990s are about 500 MB. Similar layer collections from other agencies could be predicted to be of a similar size. A generous estimate for allotting space for 20 years of historical vector layers could then be estimated at **20 GB** total.

Aerial Imagery

Another source for data will be the Minnesota Geospatial (MnGeo) Image Service, which has 1.4TB of imagery and grows by 70GB annually.[10] If MnGeo begins to archive services at the same rate as its growth, the Archive would correspondingly require 70GB of additional storage every year.

Combining these estimates yields a storage requirement of about **200GB** for the Archive for its first year, with growth rates anywhere from **1-70GB** annually. See Table 1.

Source	Total Size (2019)	Eligible data size	Annual snapshots	Annual Growth	Archive - Storage Year 1	Archive - Annual Growth
GDRS	123	100	25	1	100	26
Other Historical Vector Layers	20	20	-	-	20	0
MnGeo Image Services	1400	unknown	-	70	70	70
				Estimated Storage Needs	200	96

Table 1: Estimated Storage Needs by Source and Size (all values in GB)

[10] Minnesota Geospatial Image Service Sustainability Plan Proposal, 2018

Staffing

The Workgroup recommends that one full time staff member should be hired on a permanent basis as the Archive's curator. This position will:

- Develop the specific procedures needed to efficiently transfer data into the Archive.
- Coordinate with state agencies for scheduled submission of priority datasets.
- Consult with state agencies about submitting historical datasets and unique materials.
- Curate data submitted to the archive, including the curation activities of ingesting and preserving data.

The Workgroup also recommends one temporary student research assistant to process resources during the first year, when the inventory of submitted data is expected to be the highest.

Tools

The Archive's curator will need to use **GIS desktop software** to evaluate and potentially transform data.

The curator will also need **metadata editing tools**. Currently, the suitable tools are ArcCatalog and the Minnesota Metadata Editor.[11] Both of these applications have limitations. ArcCatalog will automatically generate technical metadata, such as bounding boxes and attribute table names, but it is cumbersome for finding the Minnesota Geospatial Metadata Guidelines (MGMG) elements.[12] The Minnesota Metadata Editor was designed specifically for MGMG and is easier to learn. However, it does not generate technical metadata, so all values must be manually entered. Neither of the applications offer batch editing techniques. The establishment of the Archive may be an opportunity to develop a new metadata editing tool, particularly if the existing tools are shown to be a barrier to contributing or processing resources.

[11] http://www.mngeo.state.mn.us/chouse/mme/index.html

[12] https://www.mngeo.state.mn.us/committee/standards/mgmg/metadata.htm

3. Curation Plan

Classifying Data

The Priority Datasets Report identified datasets by theme that have been cited by the Minnesota geospatial community as the most important to archive.[13] The datasets on this list vary in terms of format, creation date, and metadata quality. This will present different levels of work required to collect and process the layers. These datasets can be divided into two broader categories by grouping them in terms of their lineage. The categories may require different submission processes, metadata workflows, and publishing criteria.

Category 1

This category is for data that is compatible with the Archive in its current structure. This consists of a file format that functions in currently available applications and is accompanied with valid standards metadata, such as MGMG, the Federal Geographic Data Committee's Content Standard for Digital Geospatial Metadata, or the International Standards Organization 191xx series. These resources would likely have been recently available through a data repository, such as the Commons, and would need minimal processing to be submitted and ingested into the Archive.

Although Category 1 data may not be deemed "at-risk" or highly sought out by contemporary users, it will gain value over time. This concept was illustrated by Dyke et. al (2016) arguing that the newest remotely sensed imagery and decades old imagery are the most sought after layers. Archiving and preserving recent imagery is the only way to build its value for the future. See image 1.

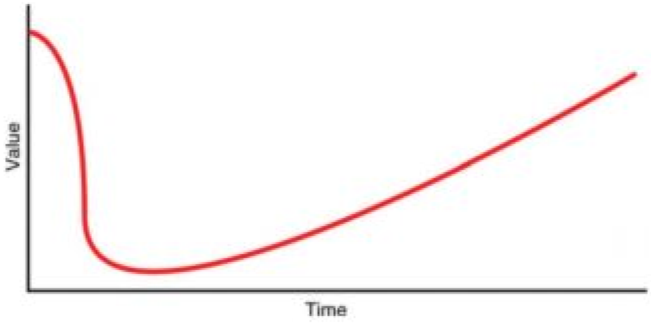

Image 1: Perceived value of remotely sensed imagery over time.[14]

[13] https://www.mngeo.state.mn.us/workgroup/archiving/Priority_Datasets_Report.pdf

[14] Kevin R. Dyke, Ryan Mattke, Len Kne & Shawn Rounds (2016) Placing Data in the Land of 10,000 Lakes: Navigating the History and Future of Geospatial Data Production, Stewardship, and Archiving in Minnesota, Journal of Map & Geography Libraries, 12:1, 52-72, DOI: 10.1080/15420353.2015.1073655, Image from page 68.

Category 2

This category is for historical or esoteric data that will require specialized appraisal and curation processes to determine its inclusion in the Archive. It may need to be converted to a supported file format and may need to be documented with complete metadata. This data would likely pre-date current metadata standards and discovery platforms.

Category 2 data will require more expertise and be slower to process than Category 1 data, but it will likely be highly valuable to users, particularly for longitudinal studies. However, it would not be practical to focus exclusively on Category 2 data at the outset. While the Archive may see more use of older historical data, processing it would also require more time to launch.

A possible timeline to mitigate the labor required for Category 2 data could be to process it in stages. The first stage could be to ingest the data with the original file formats and metadata and make them available to the public immediately as is. The second stage could be to improve the accessibility of the data by converting them and augmenting the metadata.

Contributing Data

Agencies will be responsible for contributing digital data to the Archive. This data must be opened and inspected for quality and integrity before submission. Unknown or legacy data formats may be accepted on a case by case basis. For data that is regularly updated, agencies will need to work with the Archive to develop a plan for interval of deposit that avoids hosting duplicate resources in the same discovery platform.

As described in the Archiving Agreement, data for deposit should be accompanied by metadata records that adhere to MGMG, or a successive format as approved by the Geospatial Advisory Council. Other metadata formats or completeness thresholds may be considered at the discretion of the Archive's curator. See the Archiving Agreement for additional specific requirements for contributing data.[15]

[15] http://www.mngeo.state.mn.us/workgroup/archiving/Archiving_Agreement_Report.pdf

Ingesting Data

The Archive's curator will perform initial curation activities to ingest the data into the Archive, including assessment, remediation, and publishing procedures. This may be an iterative process involving correspondence with the dataset's Contributor. See Table 2.

Action	Category 1	Category 2
Verify data transfer was error free	x	
Evaluate data quality		x
Assess metadata validity and quality	x	x
Create metadata in the approved standard		x
Assign a persistent identifier that will always point to the object and/or its metadata.	x	x
Deposit item to the data repository	x	x
Publish item to the discovery platform	x	x

Table 2: Proposed list of Curator Ingest Actions by Data Category

Preservation

After the data has been added to the Archive, it will require ongoing maintenance to preserve its discoverability and usability. This involves long term preservation activities, such as:

- Maintaining provenance records and other preservation metadata to support accessibility and management over time
- Providing secure storage and backup
- Ensuring periodic migrations to new storage media
- Ensuring routine fixity checks using proven checksum methods
- Undertaking strategic monitoring of file formats
- Planning and performing migration to a succeeding format upon obsolescence[16]

[16] File format transformation will be considered primarily for formats at risk of obsolescence. The Archive may consider improvements to accessibility through more open file formats, at a later time as interest and funding allows.

4. Funding Strategy

Estimate of Costs

The following chart lists cost estimates by category. Estimates are preliminary and will need to be researched more thoroughly by a subsequent workgroup.

Proposed Expenses	Year 1 Expenses	Year 2 Expenses
Personnel		
UMN Curator (full time)	$80,000	$82,400
UMN Graduate Assistant (1 GA, ½ time: Fall & Spring)	$40,000	-
Subtotal	**$120,000**	**$82,400**
Technology and Infrastructure		
Computer hardware (desktops/laptops)	$4,000	-
Software Licenses	$500	$500
Technical Infrastructure and Systems Support	$10,000	$10,000
Software Development (metadata tool)	$30,000	$5,600
Subtotal	**$44,500**	**$16,100**
Promotion & Outreach		
Travel funding for conferences and consultations	$3,500	$3,500
Recruitment, education, and other outreach activities	$500	$500
Subtotal	**$4,000**	**$4,000**
Grand Total	**$168,500**	**$102,500**

Funding Sources

The Workgroup recommends applying for national and state grants for initial development costs. However, for sustainability of the Archive, operational funds should be supplied by ongoing Legacy funding or through the state’s legislative budget. One model to follow could be NC OneMap, the Archive for the state of North Carolina, which is government funded as part of the state geospatial office.[17]

[17] http://www.nconemap.com/Portals/7/documents/GIS_Study_Implementation_Plan_FINAL.pdf

Appendices

Appendix I: Multi-institutional projects that established best practices for geospatial data archiving

Geospatial Multistate Archive and Preservation Partnership (GeoMAPP) - Library of Congress and state geospatial and archives staff from North Carolina, Kentucky, Montana, and Utah
Archived project site

National Digital Stewardship Alliance (NDSA) - Geospatial Data Stewardship - an initiative of the National Digital Information Infrastructure and Preservation Program of the Library of Congress
https://ndsa.org/working-groups/content/geospatial-data-stewardship/

National Geospatial Digital Archive (NGDA) - A multi-year project supported by NDIIPP at the Library of Congress that explored the development of a collecting network for archival geospatial information. http://www.ngda.org/reports.html

North Carolina Geospatial Data Archiving Project (NCGDAP) The joint project of the North Carolina State University Libraries and the North Carolina Center for Geographic Information and Analysis focused on collection and preservation of digital geospatial data resources from state and local government agencies in North Carolina. (2004-2010.) http://www.lib.ncsu.edu/ncgdap/

Appendix II: Best Practices Documents

State of Utah Business Plan for Archival Preservation of Geospatial Data Resources (Example budget plans for archiving - Budget Plan is Section 6.3)

Geoarchiving Business Planning Toolkit - tips for designing a business case for archiving geospatial data including Cost-Benefit Analysis Guidance and Use Case Guidance

Best Practices for Geospatial Data Transfer for Digital Preservation - provides suggestions for data transfer procedures including how to get everyone on the same page about what they should be

Interstate Data Transfer Design Template - Outline (i.e. concise summary) of the Geospatial Data Transfer for Digital Preservation report

Best Practices for Archival Processing for Geospatial Datasets - suggested workflow for archival organizations processing geospatial datasets. Some of this is more detail than we need given that we will be building from an existing archive infrastructure - but p.10-52 discussing ingest process and quality assurance would be a useful reference for developing the Data Transfer procedures.