Appraisal and Selection of Geospatial Data

Prepared for Library of Congress

November 2010

Table of Contents

Executive Summary 3

Introduction 5

- **Background** 5
- **Report Scope** 6

Establishing Criteria for Appraisal and Selection 7

- **Appraisal and Selection Challenges Imposed by Geospatial Data** 7
- **Impact of Organizational Focus on Appraisal and Selection** 8
- **Individual Criteria For Appraisal and Selection** 10

Models and Processes for Appraisal and Selection 12

- **Tools for Identification and Evaluation of Data Resources** 14
- **Triggers for Appraisal and Selection** 15
- **Retention and Disposition of Data** 16
- **The Motivation for Creative Disposition Solutions** 17
- **Looking Beyond Just "Data"** 18

Questions to Explore 19

Appendix I: Choosing the appropriate data version 21

Executive Summary

The Library of Congress National Digital Information Infrastructure and Preservation Program is developing a national strategy to collect, preserve and make available significant "born-digital" information. NDIIPP has sponsored three projects specifically addressing geospatial data. Currently the Library is considering the prospect of a national distributed collection of geospatial data, and has heard from stakeholders that appraisal and selection actions merit focused attention. The details below are largely drawn from a November 2009 workshop the Library of Congress held with geospatial community stakeholders.

Appraisal is often associated with government archival processes and is here loosely defined as identifying records or other information to determine which merit long-term or permanent retention. Selection is typically associated with library or other collecting institutions and is generally defined here as choosing materials for preservation because of their continuing value. The Library of Congress is considering both activities at this time, as they each appear to have value for defining geospatial content of enduring value to the nation.

Appraisal and selection of geospatial data are critical because of the limited resources that most collecting and stewardship organizations have for preservation. Given the very large (and growing) volume of geospatial data, it is a practical necessity to choose only the most important for long-term management. This management may involve resource-intensive attention such as metadata augmentation, file format conversion, storage media migration, and ongoing repository hardware and software maintenance.

While appraisal and selection are crucial, they present a fundamental challenge to creating and collecting institutions. As noted above, volume is a major issue; in some circumstances the scope and depth of geospatial information may be so great that knowledge of what even exists can be difficult. Other challenges include:

- Complex data structure and proprietary file formats
- Difficulty in describing data to provide for broad secondary use
- Institution-specific, siloed management policies and practices
- Uncertainty regarding the effectiveness of existing policies and practices

It seems clear that the challenge has two threads. One relates to the nature of geospatial data itself; the other to processes and polices that have been developed to manage data. It is necessary to understand and appreciate the first thread. But the second thread appears to be the most useful for focused

attention and consideration. A broad look at existing polices and practices might reveal areas where certain activities could be shared or adapted across institutional boundaries. The same examination might also uncover areas that need development, refinement or innovation.

The stakes are high. Geospatial data plays a role in a wide range of applications and industry sectors, supporting planning and decision making in the government, commercial, academic and not-for-profit spheres. While many applications are driven by the need for the "latest and greatest" data, there is increasing demand for older and superseded data to support historical and temporal analyses related to change in earth's natural and human landscape. Examples of applications that require historical data include study of climate change, disaster planning, environmental impact analysis, industry site location planning, and resolution of legal challenges.

Stewardship resources, on the other hand, are not keeping up with the volume of data, making it even more necessary for archives and libraries to make decisions about what to keep. While current library and archival processes need to explored, the data management and retention approaches within the data producing and data managing organizations--even if of shorter-term focus--should also be considered.

Up to this point an organizational focus has driven appraisal and selection decisions, with data producing agencies, data managing agencies, archives and libraries each making decisions according to their own individual needs. This will continue, but it is worthwhile to consider if a broader national (or multi-organizational) focus is useful. Key questions to address include:

1) Are current appraisal and selection policies and practices robust and adaptable enough to address geospatial data?

2) Which pieces of existing policies and practices are the best candidates for sharing across institutional boundaries?

3) What specific aspects of appraisal and selection require more investigation?

4) Are there models of shared services and cooperation between data managing agencies, on the one hand, and archives and libraries, on the other hand, around the long-term preservation of geospatial data?

5) What are some basic next steps that can be taken to advance the practice of geospatial data appraisal and selection?

Introduction

Background

Geospatial data plays a role in a wide range of applications and industry sectors, supporting decision making processes and planning efforts in the government, commercial, academic and not-for-profit spheres. While many applications utilizing geospatial data are driven by the need for the "latest and greatest" data, there is increasing demand for older and superseded data to support historical and temporal analyses. Examples of applications that require historical data include study of climate change, disaster planning and post-disaster analysis, analysis of land use change and environmental impacts, business and industry site location planning, and resolution of legal challenges. Geospatial data resources typically are supported by backup plans that are intended to ensure near-term retention of data, yet many such resources have not been addressed by archival plans that explicitly allow for longer-term retention of data, including superseded versions of current datasets.

In the year 2000, Library of Congress initiated the National Digital Information Infrastructure and Preservation Program, the mission of which is to develop a national strategy to collect, preserve and make available significant digital content, especially information that is created in digital form only, for current and future generations. To date NDIIPP has sponsored three projects specifically addressing the issue of long-term preservation of geospatial data:

- The North Carolina Geospatial Data Archiving Project (NCGDAP), which initially addressed the issue of preserving geospatial data at the state and local level.
- The Geospatial Multistate Archive and Preservation Partnership (GeoMAPP), which is currently working to integrate data archiving efforts with existing state geospatial data infrastructures.
- The National Geospatial Digital Archive (NGDA), which addressed geospatial data issues at a national level.

Currently the Library of Congress is focusing on a broader national consideration for appraisal and selection of digital information resources, an effort that includes a special focus on geospatial data.

Formal appraisal and selection processes, which include criteria and procedures for appraising data, provide a consistent way to assess future value of data resources and--keeping in mind constraints with regard to organizational

resources for data management--to limit commitment of geospatial data management resources to those resources deemed essential. From the perspective of a data producing or data managing agency, a variety of managerial, operational, and technical concerns can affect decisions about the retention of data. From an archival perspective, additional scientific, scholarly and historical dimensions can come into play. All of these factors guide development of formal processes for appraising geospatial data as electronic records.[1] In the library context, a similar set of factors guide the development of formal collection development policies that guide data acquisition efforts, which may take place without regard to agency origin. In the world of large commercial data firms a "keep everything" rule often applies, as the costs to add data storage may be lower than the costs in terms of time and resources necessary to determine what to delete. Even in this case de facto appraisal and selection may still come into play in the form of decisions about what data to make available for search, discovery, manipulation, and access.

Report Scope

Appraisal and selection will be considered from the point of view of earth-related, two-dimensional data that would typically be comparable to maps and charts but which may occur as a variety of geospatial data types including GIS data (raster, vector, and associated data formats), remote sensing data (including satellite imagery and digital aerial imagery such as digital orthoimagery), and similar data resources that are defined by geographic location or extent. Information resources that are associated with a particular geographic place or area but for which geographic representation are not the primary characteristic (e.g., a single document associated with a place) are not within scope of discussion.

Appraisal is a process by which archivists and records managers assign administrative, legal, research, and historical value to records in order to determine retention period, where records will be maintained, and when to transfer records to archives. **Selection** is the process by which an organization such as a library or data center makes decisions to add resources to their collections in order to meet the needs of their user population. Appraisal and selection may take place in different organizational contexts, addressing individual organizational needs, but these two approaches share some common elements, including: identification of resources to be considered for acquisition or retention, methods for assessing and assigning value to resources in accordance with organizational need, and triggers to initiate these processes. Excluded from discussion here are related issues such as metadata creation, ingest (or accession) of records into archives or repositories, data inspection, or development of access and discovery infrastructure.

---

[1] Center for International Earth Science Information Network (CIESIN). "Guide to Managing Geospatial Records, Version 1.0", June 2005. Available: http://www.ciesin.columbia.edu/ger/GuideToManagingGERv1Final.pdf

Establishing Criteria for Appraisal and Selection

Government records managers often appraise from the perspective of selecting for long-term retention those records that best document or capture the activities and information outputs of government agencies. In the case of geospatial data, it may be necessary to move beyond this approach in order to take into account the broad applicability of geospatial data, which contain valuable information of use in a variety of research areas that extend far beyond the intended use of the data.[2] The broad utility of geospatial data calls for appraisal and selection process that take into account not only intended use and organizational requirements, but also factor in the potential reuse value of the records for wider community.[3]

Appraisal and Selection Challenges Imposed by Geospatial Data

The domain of geospatial data comprises a wide range of information types, many of which exist in complex data or database formats. Unique characteristics of geospatial data that affect appraisal and selection include:

Frequently or continuously changing data: Feature data resources such as land records, street centerlines, and jurisdictional boundaries that are subject to frequent change will need to be addressed by appraisal and selection processes in such a manner that frequency or periodicity of capture is established. Since current data management practices among data producers often still involve overwrite of older versions of data it may not be possible to increase capture frequency retroactively after a capture approach has been implemented.

Commercial or proprietary data formats: Geographic information systems (GIS) is often created and managed in proprietary or commercial data formats. A decision to capture that data in its native format may introduce dependencies on proprietary technologies. On the other hand, efforts to capture that same data in a format that is based on an open standard may result in data loss or reduced data usability.

Spatial databases vs. individual datasets: Spatial databases, which comprise a combination of individual datasets in combination with relationships, behaviors, annotations and models, can be managed forward in time using complex

---

[2] North Carolina Center for Geographic Information and Analysis and North Carolina State Archives: Geospatial Multistate Archive and Preservation Partnership (GeoMAPP) Interim Report, 2007-2009. Available: http://www.geomapp.net/docs/GeoMAPP_InterimReport_Final.pdf

[3] Guy McGarva. Digital Curation Centre Briefing Paper: Curating Geospatial Data. April 4, 2006. Available: http://www.dcc.ac.uk/resources/briefing-papers/introduction-curation/curating-geospatial-data

technology that is available to the producing or managing data agency but which is not widely supported by libraries and archives. Database snapshots can be transferred to a receiving organization but may not survive long without active management using the appropriate database technology. Another data capture approach involves extracting individual datasets that may be transferred to an archive or library in a stable, more preservable form, but at the cost of omitting database content that is not amenable to such an extract method.

Complex, domain-specific metadata needed for appraisal and subsequent use: Many geospatial data resources will be difficult or impossible to use without associated, domain-specific metadata. The appraisal or selection process itself will also be difficult or impossible in the absence of the availability of this metadata, as it will be difficult to determine data origin and function.

Variety and complexity of data representation methods and data derivatives: The true counterpart to the paper map is not so much the dataset as it is usually a combination of datasets which have been synthesized and displayed in a particular manner. These derivative products represent an entirely different information object, the capture of which may stand as an additional objective which does not preclude or take the place of capture of the actual datasets (and other technology components) needed to create these outputs.

Scale and resolution: A data collecting organization may encounter situations where different dataset options for a given theme are available at different scales. A large-scale dataset will tend to capture more spatial detail, yet smaller scale datasets may offer associated information that is not available with large-scale alternatives, and may prove useful in applications where spatial detail is not needed. In the case of raster data, higher resolution data resources offer a higher level of spatial detail, but at the cost substantially higher local storage requirements and, potentially, adverse impact on application performance.

Data volume and capacity limitations: The sheer size, in terms of file size and number of files, of many geospatial data resources--especially in the case of raster data--in combination with limited nature of storage capacity found at acquiring archives and libraries may affect decisions with regard to disposition and archival acquisition of data.

The complexity of geospatial data, the complications imposed by dependencies on complex or proprietary technologies, and the relationships with ancillary information components need to be factored into decisions as to whether to take on a particular data resource.

Impact of Organizational Focus on Appraisal and Selection

Individual organizations will tend to appraise and select data resources from the perspective of organizational needs, although in some cases external advisory

groups may be engaged in order to gain a understanding of broader needs with regard to data retention. The retention strategies of data producing and data managing agencies may be developed with near-term rather than long-term retention considerations in mind, yet the practices of these agencies merit examination for two reasons. First, these data management practices are informed by a deep knowledge of the data and current use thereof, and as such may help to inform practices being shaped around longer-term retention. Second, data retention decisions early in the chain of data custody impact the quantity and nature data later available for consideration by archives and libraries.

Data Producing Agencies

Individual data producers include such entities as federal, state and local government agencies, not-for-profit organizations, universities, and commercial firms. Data producers are in the best position to document the data and to understand intended uses of the data, yet they may have more modest aims with regard to data archiving and preservation. Data retention efforts may be limited to activity such as backup procedures which are intended to ensure short- and medium-term access to older data versions but which are not designed to ensure longer-term access. Additionally, the data producer may be less aware of or less responsive to data uses that extend beyond initial intended uses. Limitations on both storage capacity and expertise to support long-term maintenance of data may also affect data producer decisions with regard to data retention.

Data Managing Agencies

A second class of data custodian includes data managing agencies that act as centralized data repositories at the level of a government unit (federal, state, or regional) or within a specific domain of interest, such as in the areas of climatic, oceanographic, or geologic data. These agencies maintain a thorough understanding of the broader availability of data within a specific geographic or thematic domain and have established relationships with individual data producers. Data managing agencies tend to have broader goals with regard to data custodianship and may be in a position to encourage, facilitate and perhaps even enforce creation of metadata and adherence to data content standards. These agencies typically are characterized by a well-established data management infrastructure, although in some cases data custodianship may be more focused on current uses (both intended and unintended) than on longer-term maintenance of content, and ongoing selection decisions may still in some cases be centered on meeting demand for the “latest and greatest data.” As aggregators of data, managers of relationships, and influencers of action, these centralized data repositories provide a convenient point of

contact for archives and libraries concerned with longer-term support for content.

Archives

Archives, and in particular government archives, whether at the federal, state, and local level, typically have data management aims that are longer-term in nature and broader in scope than those of data producing or managing agencies. These data management aims are conditioned by the organizational imperative of the archive in question, an imperative that may center on addressing the records of a specific set or class of government agencies. In order to appraise data, archives rely on metadata made available as part of data inventories, through data catalogs, or as otherwise provided in the context of record appraisal. Archives may also be subject to technical capacity limitations, both in terms of ability to process certain geospatial data types, such as spatial databases, and in terms of storage capacity. An immediate objective of an archival appraisal process will be to establish a retention and disposition plan for the appraised data resource.

Libraries

Like archives, libraries rely on inventories, catalogs, metadata, and data agency support in identifying and assessing data resources, and libraries may be subject to the same constraints as archives vis-à-vis limitations regarding technical capacity. A library's collection development will also be conditioned by it's own organizational mission, which may involve addressing a broader universe of information resources needed to address the information needs of the library's user population. In addition to government information, the collection development effort might directly address data from a variety of sources including commercial, not-for-profit, academic and international sources.

Individual Criteria For Appraisal and Selection

Data archives commonly delineate areas of data custodianship using statements of scope that clarify the geographic, organizational, or domain focus for data appraisal or selection efforts. In order to guide formal appraisal and selection processes, a standard set of questions may be employed to determine whether particular data resources meet the criteria for retention. Commonly used criteria include the following:

Is the data relevant to organizational focus?

In the case of the government archive the focus will be on the business of the agencies or set of agencies served by the archive, while a library may

focus more generally on information resources deemed useful to the audience in question. A data managing agency may acquire data within a given geographic focus (e.g., a state) or a given thematic area (e.g., climate or oceanographic data).

Is the data sufficiently documented to support identification, appraisal, and subsequent use?

Data must be documented in order to first identify the data (e.g., through a catalog or inventory), then to appraise the data and subsequently to support use. Appraisal processes might provide an opportunity to enhance metadata with additional descriptive, technical, and administrative metadata that supports ongoing management, discovery, and use.

Does the data address current, known research or application needs?

Data resources may be selected to meet ongoing organizational needs in terms of research or application development, or to address specific and emergent topics of interest, for example data associated with key legislation or associated with prominent natural disasters.

Is the thematic content of the data such that it will have a high propensity for use?

Data resources that lend themselves to use in a broad range of known applications, such as orthoimagery, land records or transportation data, have the potential to support an even broader set of unknown or unanticipated historical or research needs. For example, land records data that has initially been developed to meet the narrower needs of tax administration is also currently of use in a variety of applications related to real estate. This data might, in a historical context, support future analyses in the area of demography, ethnography, and site location analysis.

Is the geographic extent of the data in line with the needs of the targeted user populations?

Data resources that have a geographic extent that coincides with or encompasses the geographic scope of the acquiring organization (for example, at national extent or the extent of an individual state) may have greater potential for use by that organization’s user population.

Is the data in a form that is usable to the organization's user population?

> Closed, proprietary formats may pose a preservation risk longer term and in fact may limit use in the present use if current tool support is limited. On the other hand, transfer of that same data into a format that is an open standard may actually reduce data usability and introduce data loss.

Is the data reasonably acquirable?

> Data acquisition should not impose undue burden on the organization. Factors that might complicate acquisition include quantity of data (size, number of files), licensing or copyright restrictions that complicate data management or limit use, fees, or the presence of complex database technologies that require active management. Data transfer processes should lend themselves to streamlined, if not automated, processes.

Is the data at risk of impending loss?

> Data that is known to be at risk may receive higher priority for direct acquisition. This risk might be manifested in ways such as the following: versioned data overwrite, stated plans of a data custodian to discontinue retention of the data, or perceived need to take remedial action with regard to conversion of data out of formats that are no longer widely supported.

Is another organization archiving the data? If so, is the period of retention limited?

> Libraries or archives might consider data resources that are not directly addressed by archives as part of government records management programs.

Is there another data resource that provides some replacement value?

> The uniqueness of a data resource in terms of content may speak in favor of retention.

Models and Processes for Appraisal and Selection

Many data agencies, archives, and libraries are already conducting appraisal and selection of geospatial data and offer some working models, which include processes or functions such as the following:

- Initial review and subsequent re-review of data inventories and catalogs to identify resources.

- Review of metadata and other documentation to assess record value.
- Triggers for appraisal and selection (including mechanisms for response to impending record destruction).
- External review of, or contribution to, appraisal and selection decisions.
- Implementation of a disposition plan, which may result in destruction of the data, retention at the agency, or transfer to an archive.

The experience of some government archives, including those that are participating in the NDIIPP Multistate Geospatial Archive and Preservation Program (GeoMAPP), has been that existing selection and appraisal processes and records retention regimes can provide a basis for addressing geospatial data, but the unique nature of geospatial records makes necessary different disposition and capture processes.[4] In the government records context, the records retention process serves not only as a legal basis for data preservation, but also can be an organizing tool for development of preservation strategy. The most significant benefit from data acquisition efforts may be found in focusing on superseded and at-risk data. Chances for success in implementing archival programs have been increased by:

- Establishing close relationships with data agencies.
- Educating archival staff with regard to existing metadata standards in order to appraise, categorize and organize data as well as to support use.
- Capturing data from reliable consolidation points such as data managing agencies that act as central repositories.

Organizations that are implementing archive programs must decide whether to initiate new efforts from "day forward" perspective, as opposed to expending effort locating and acquiring older data.[5] "Day forward" approaches may involve lower start-up costs and may bring a higher likelihood of data agency participation, while successful identification and capture of older datasets may require additional effort on the part of the acquiring agency and current custodial agency.

---

[4] The Geospatial Multistate Archive and Preservation Partnership (GeoMAPP) Interim Report summarizes the individual appraisal experiences of the state archives of North Carolina, Kentucky, and Utah as well as overall findings with regards to appraisal of geospatial data: http://www.geomapp.net/docs/GeoMAPP_InterimReport_Final.pdf

[5] Guidance provided by the North Carolina Geographic Information Coordinating Council Archival and Long-Term Access Ad Hoc Committee in its Final Report, November 19, 2008. Available: http://www.ncgicc.org/Portals/3/documents/Archival_LongTermAccess_FINAL11_08_GICC.pdf

Tools for Identification and Evaluation of Data Resources

Selection and appraisal processes must be informed by some understanding of availability of data resources. Establishment of close working relationships with data agencies, and in particular agencies that act as centralized aggregators or repositories of data, will accelerate the process of data identification. In addition to contact with the data agencies themselves, other tools for identification of data resources include **data inventories** and **data catalogs**. Inventories, such as RAMONA, are sometimes used to track availability of data resources within a specific geography or thematic domain and provide an opportunity to assess how much data exists, current format, responsibility, creation date, and data origin, all of which are important in determining the extent and quantity of content is available.[6] Data inventories can provide an end-to-end, big picture view of what is available and what may be at risk in order to support acquisition priorities.[7] These inventories, if sustained over time, may be of value as a source of information to support evaluations of trends in terms of data availability, use of data formats and coordinate systems, and other data characteristics. Data catalogs, such as geodata.gov, which are intended to support data discovery by end users (and which may at least in part be populated by data inventories), are another resource for identifying data resources, yet these may not be configured to provide and end-to-end view of available data.

Data that has not been identified cannot be acquired, and yet many older datasets are not exposed through data catalogs or data agency portals. If data archive implementations involve retroactive identification of data resources then additional interactions with data agency staff may be needed to ferret out available data.

Metadata, which may both inform and derive from inventories and catalogs, supports current and future use of the data by providing the descriptive information needed to help satisfy search while also providing additional technical and administrative information needed to understand data limitations

---

[6] The Random Access Metadata Tool for Online National Assessment is produced by the National States' Geographic Information Council (NSGIC) as a tool for states and their partners. Its primary purpose is to track the status of GIS in US state and local government to aid the planning and building of Spatial Data Infrastructures. Available: http://www.nsgic.org/hottopics/ramona.cfm

[7] The value of data inventories in data archiving efforts is described in detail in the North Carolina Geospatial Data Archiving Project Interim Report, June 1, 2008. Available: http://www.lib.ncsu.edu/ncgdap/documents/NCGDAP_InterimReport_June2008.pdf. The GeoMAPP Interim Report discusses the use of data inventories by partner states in the selection and appraisal process.

and applicable uses.[8] In addition to metadata, a records analyst may obtain information about a dataset by working directly with data producing or data managing agencies.

Triggers for Appraisal and Selection

Upon engaging a data agency in a records appraisal process, an archive may initiate the appraisal process with a thorough review of that data agencies holdings, using some combination of existing inventories and catalogs, accompanying metadata, and consultation with data custodians. Once an initial appraisal has been completed a variety of triggers might activate subsequent record appraisal processes. Some common triggers for appraisal in the government records management context include:

1) Addition of a new dataset to data agency holdings: When a data producing agency creates and initiates maintenance of a new dataset government archives may set a retention schedule.

2) Removal of data from data agency holdings: A data agency decision to retire a dataset or remove a dataset from its data collection might trigger a "reappraisal" on the part of a partner archive.

3) Update or modification of a dataset: Initial appraisal of datasets that are subject to continuous update may, at the outset, result in establishment of a retention schedule that accounts for periodic capture, such as quarterly or yearly, on the other hand, relatively static datasets that are subject to infrequent and irregular update may need to be re-appraised on an as-needed basis.

In the sphere of libraries, collection development efforts may be shaped in response to additional triggers, including: emergence in interest in a particular topic as a result of a current events or key legislation; identification, through outreach, of valuable at-risk resources; establishment of a new area of organizational focus (e.g., a new research thrust area at a university); and establishment of a new organizational partnership for collection building.

In the commercial world, triggers for appraisal and selection processes might include: acquisition of new data that supersedes older data, exceeding limits on storage capacity, and acquisition of data holding companies.

---

[8] As part of the GeoArchives project, Maine's State Archives selected 16 FGDC metadata elements for direct use in the appraisal process. These elements were selected for likelihood of providing guidance about legal, informational, and evidential use of the dataset in question. Outlined in the "Creating the GeoArchives: Maine Archives of Geographic Information", August 3, 2006. Available: http://www.maine.gov/sos/arc/GeoArchives/geoarch.html

In order to fully inform appraisal and selection processes it may help to establish an advisory board or committee that represents a diversity of organizational interests. While an internal committee or review board may be more responsive to organizational practice and need in the short term, an external advisory board which represents a broader community of data users can provide outside perspective and expertise with regard to consequences and value of particular resources.[9] Public input may also be captured to help inform appraisal decisions.

Retention and Disposition of Data

In the sphere of government archives, an appraisal process might result guidance to a data agency to retain data either for a defined period of time or permanently. The disposition plan may also stipulate that data be transferred to the archives for retention. The size and complexity of the data may make it necessary to establish formal transfer processes. Data managing agencies that act as central repositories can act as a single point of capture, easing transfer from individual data producers to the archives. In cases where capacity limitations of the data producer do not make retention by that agency possible, the data managing agency might accept responsibility for retaining the data. In cases where technical limitations on the part of the archives might make transfer of data impossible, a data agency might continue to act as the custodian of the data.[10]

In the case of library collection development, the objectives of collection building for present day use and collection building for data preservation might diverge. In building data collections libraries have increasingly come to rely on connection to externally available web services as a substitute for direct data acquisition of data. This approach may satisfy needs for data access, but if the library wishes no ensure long-term access to the data it may be necessary to directly acquire the data.

---

[9] External advisory boards are described in the CIESIN "Guide to Managing Geospatial Records, Version 1.0". The Maine GeoArchives consults with an Archives Advisory Board, an entity independent of the Archives staff, which must approve any decision to destroy records.

[10] NARA and USGS cooperation in data archiving is outlined in a June 13, 2008 press release. Available: http://www.archives.gov/press/press-releases/2008/nr08-118.html. Similar arrangements involving retention of data by state GIS agencies, in cooperation with state archives, has taken place as part of the GeoMAPP multistate initiative.

The Motivation for Creative Disposition Solutions

The size and complexity of geospatial resources may require creative solutions to determining disposition of data due to the following factors:

- The large size of data resources, and in particular remote sensing data, the storage needs of which may exceed the technical capacity of archives, requiring an arrangement by which the data resides with a data agency.
- The complexity of data resources, notably spatial databases, which may need to remain under the active management of data agencies, while the archives themselves capture derivatives or snapshots that have reduced active support needs.
- Accommodation for rights concerns that might dissuade a data agency from making a data resource available to an archive for open access. Accommodations might include placing the data in a dark archive without general public access or agreeing to leave the data in the custody of the data agency.[11]

Disposition statements or data acquisition plans that result in the development of a distributed solution to data maintenance may call into need special provisions such as:

- Establishment of reappraisal triggers in cases where the maintaining data agency makes plans to remove the data.
- Development of close relationships between acquiring organizations and data agencies in order establish channels of communication to ensure that reappraisal triggers are acted upon and mutually workable solutions to distributed data management can be achieved.
- Development of dark archives with no access or limited public access; or in the case of government data, the cultivation of open data sharing arrangements that negate the need for archival access controls.[12]
- Providing archiving organizations with a degree of shared control over some portion of the content maintained within the technical infrastructure of a data agency.[13]

---

[11] The North Carolina Geospatial Data Archiving Project (NCGDAP) and the Kentucky component of the GeoMAPP effort each needed to address local agency concerns about public access, in the former case a dark archive was created and in the latter case data was left under the jurisdiction of the local agency.

[12] The NCGDAP project experience with restrictions on data access helped to inform the work of the North Carolina Geographic Information Coordinating Council Local/State/Regional/Federal Data Sharing ad hoc Committee, which published "Recommendations for Geospatial Data Sharing, revised November 7, 2007. Available: http://www.ncgicc.com/Default.aspx?tabid=156

From the perspective a library, the same issues that come into play in determining disposition of records in government archives context can also come to affect the notion of what it means to "build a collection", i.e., collection-building may involve a mix of physical acquisition of data with the formulation of a set of arrangements providing for discovery of and persistent access to data resources that reside with data agencies or perhaps even commercial firms. Collection building might in fact extend to capture of metadata--in lieu of capture of the data itself--in order to enhance discovery of externally managed resources, including possibly commercial data.

Looking Beyond Just "Data"

Data projects or data representations that combine one or more datasets with additional processing for presentation and analysis purposes provide added value content. These derivative information objects may be created by different agencies than those that create the source datasets, so in the context of a government records appraisal process these added value products may be appraised and scheduled separately than the underlying data. While many data projects may be of shorter-term value, others may be associated with topics that draw considerable public interest.[14]

Complexity and stability of data projects or representations is an important factor to consider when making retention or acquisition decisions. These complex arrangements contain both the original data as well as added functionality, yet these projects will be very difficult to retain over time as they may depend on short-lived proprietary technologies or external resources such as geospatial web services. Projects or representations may be desiccated into a form such as a geospatial PDF or an image representation, but at the loss of the original data and underlying functionality. A geospatial PDF (e.g., GeoPDF) will at least retain some subset of data intelligence that would be discarded as part of a simple image capture.[15]

---

[13] As part of the Maine GeoArchives project the state geospatial agency provided the State Archives with direct control of an archival portion of the enterprise geospatial database.

[14] In the GeoMAPP effort the Utah Archives found that most projects have a short-term value of only10 years or less, though occasionally a project will have significant public interest, in which case it was decided that those project files will be kept permanently.

[15] The Utah Automated Geographic Reference Center now generates GeoPDF versions of each GIS dataset transferred to the archives (in addition to Shapefile and File Geodatabase versions) in order provide an alternate, open representation of the data.

Questions to Explore

1) Are current appraisal and selection policies and practices robust and adaptable enough to address geospatial data?

- What gaps do formal records management processes and collection development policies leave?
- Can the retention decisions and strategies used by data agencies to address near-term retention help to inform the selection and appraisal practices in the archives and library community?
- To what extent do the technical processing capabilities of a receiving organization serve as a limiting factor in data acquisitions?

2) Which pieces of existing policies and practices are the best candidates for sharing across institutional boundaries?

- What is the best way to capture external perspectives in assessing data value (e.g., advisory boards, public input)?
- Do framework data themes provide a useful context for prioritizing data retention efforts?
- How important is current use for assessing long-term value?
- Should metrics associated with data download or use of geospatial web services serve as an indicator of need to target data resources for archival acquisition?

3) What specific aspects of appraisal and selection require more investigation?

- Is it possible to determine optimal frequencies of capture for different types of geospatial data?
- What balance should be struck between usability of data versus requirement for use of open standards?
- Can appraisal and selection processes be applied to data representations and data projects?
- Should libraries and archives capture administrative records, standards, and documents such as data inventories that relate to geospatial data?

4) Acquisition costs, capacity limitations, data complexity, and restrictions on data use or redistribution are all factors that may limit the ability of archives and libraries to physically acquire some kinds of data. Are there models of shared services and cooperation between data managing

agencies and archives and libraries around the long-term preservation of geospatial data?

- In situations where data agencies continue to maintain direct custody of the data by agreement with an archive, what triggers can be put into place to ensure re-appraisal if and when the data agency elects to cease retention of the data?
- Is it possible to arrange for joint administration of complex geospatial databases that require active management on the part of a data agency?
- Should archives or libraries consider acquiring metadata in the absence of acquiring actual data?

5) What are some basic next steps that can be taken to advance the practice of geospatial data appraisal and selection?

Appendix I: Choosing the appropriate data version

An individual geospatial data resource may be available in a bewildering variety of configurations that are optimized for different uses, different software tools, and different user populations. Some of the choices to consider when setting retention schedules or acquiring geospatial data include:

Raw vs. Processed Data

> Raw data are the manifestation of the data in its original, purest form, and as such are valuable as authentic records, but raw data may be less useable to broad audiences and may lack important value-added features found in processed derivatives (e.g., raw aerial imagery lacks the georeferencing, georectification, and QA/QC that makes a fully processed digital orthoimage widely useful). Delivery versions of a data resource may be more easily accessed and used due to compression, reduced size, or transfer into a widely used format, but may have incurred data loss (e.g., an access copy of an orthoimage may be subject to lossy compression).

Commercial or Proprietary Formats vs. Open Formats

> Proprietary formats introduce a technical risk by virtue of their closed nature yet they may support complex functionality that is not directly supported by available open formats. Some commercial formats, such as the Shapefile GIS format, are commercial yet openly documented and widely supported, yet may not support features found in more complex proprietary formats. Formats associated with open standards such as Spatial Data Transfer Standard (SDTS) and Geography Markup Language (GML) may not be supported by widely available software tools or may involve loss of content or functionality if used as a target format for conversion.[16]

Spatial Database vs. Individual Data Layers

> Spatial databases may store multiple datasets along with dataset relationships, behaviors, annotations, and data models, all of which are

[16] The GeoMAPP project came to the conclusion that "selection of a type of geospatial data format for preservation depends on the goals established for long-term preservation; priority emphasis should be placed on format openness (whether proprietary or not), community uptake, data portability, and the ease of data migration."

hosted in database system. These complex data resources have played an increasingly important role in data production and management, while dataset-oriented formats are often still used for data distribution. An acquiring organization may choose to capture periodic snapshots of the entire database, though the proprietary nature and unproven long-term sustainability of prominent database formats calls into question the long-term viability of these snapshots. An acquiring organization may also choose to capture individual datasets as extracts from the database. These extracts may prove to be more stable over time but do not provide a means to capture the entirety of the databases contents.

Compressed vs. Uncompressed

Raster data will often be made available in compressed form in order to reduce file size for ease of transfer and use or in order to enable selective decompression. Lossy compression methods result in changes that, while not necessarily impacting use of the data visually (by people), might negatively impact analytic functions that the data are intended to support.[17]

Coordinate System

An individual data resource may be available in more than one coordinate system. Individual coordinate systems may be more widely used in one community than in another, for example individual states may use one or more coordinate system tailored to their state. Long-term support of various coordinate systems is a challenge in and of itself.

Tiling Scheme

A given data resource may be available in more than one tiling scheme. Some common tiling schemes include county, state, national, USGS quadrangle, census unit, Landsat scene, river basin, and tax map unit. The convenience or usability of different tiling schemes may depend on user population and user technical environment (software and local storage).

Temporal Versions and the Frequency of Capture Problem

In the case of vector GIS data, some datasets such as land records, street centerlines, and municipal boundaries are updated fairly frequently while others

[17] Guy McGarva, Steve Morris, Greg Janee. Digital Preservation Consortium Technology Watch Report: Preserving Geospatial Data. Available: http://www.dpconline.org/advice/technology-watch-reports

more rarely. In the case of satellite imagery capture is typically more or less continuous. Granularity in change is lost if a version is not saved each time data are revised or updated, yet costs of acquisition and maintenance will be lower if capture is done less frequently in the form of periodic data snapshots. The data agencies themselves may maintain versioned databases that allow for maintenance of the characteristics of individual pieces of data over time, allowing for the recreation of a point in time, but an acquiring archive may not be able to maintain copies of such a database. The entire database, or components thereof, may also be captured as snapshots at static points in time. In some cases such snapshots might function as data releases, something like a data edition representing a present tense official view.[18] Alternatively, these snapshots can be captured not to meet present data access needs but rather as an explicit means of creating temporal views for later use.[19] At later points these snapshots provide a temporal view of periodically released data.

Factors that might favor of a higher frequency of capture include frequency of data change, absence of a data producer plan to retain (and make accessible) data snapshots, or existence of applications that have need for greater temporal granularity.

[18] David L. Brown, Grace Welch, Christine Cullingworth: Archiving, Management and Preservation of Geospatial Data: Summary Report and Recommendations. Available: http://www.geoconnections.org/en/resourcelibrary/keyStudiesReports

[19] The North Carolina Geospatial Data Archiving Project conducted surveys of data capture practice by local agencies in 2006 and 2008. Available: http://www.lib.ncsu.edu/ncgdap/documents.html