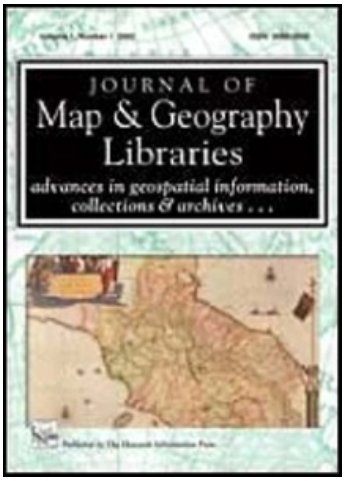

Journal of Map And Geography Libraries
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t792306932

The National Geospatial Digital Archive: A Collaborative Project to Archive Geospatial Data

Tracey Erwin [a]; Julie Sweetkind-Singer [a]
[a] Branner Earth Sciences Library and Map Collections, Stanford University, Stanford, California, USA

Online publication date: 31 December 2009

To cite this Article Erwin, Tracey and Sweetkind-Singer, Julie(2010) 'The National Geospatial Digital Archive: A Collaborative Project to Archive Geospatial Data', Journal of Map And Geography Libraries, 6: 1, 6 — 25
To link to this Article: DOI: 10.1080/15420350903432440
URL: http://dx.doi.org/10.1080/15420350903432440

Journal of Map & Geography Libraries, 6:6–25, 2010

ISSN: 1542-0353 print / 1542-0361 online
DOI: 10.1080/15420350903432440

Routledge
Taylor & Francis Group

The National Geospatial Digital Archive: A Collaborative Project to Archive Geospatial Data

TRACEY ERWIN and JULIE SWEETKIND-SINGER
Branner Earth Sciences Library and Map Collections, Stanford University, Stanford, California, USA

The National Geospatial Digital Archive is a collaborative project between the University of California at Santa Barbara and Stanford University. The project was funded by the Library of Congress through their National Digital Information Infrastructure and Preservation Program (NDIIPP). The goal of the collaboration was to collect, preserve, and provide long-term access to at-risk geospatial data. The project partners created preservation environments at both universities, created and populated a format registry, collected more than ten terabytes of geospatial data and imagery, wrote collection development policies governing acquisitions, and created legal documents designed to manage the content and the relationship between the two nodes.

KEYWORDS geospatial data, GIS, long-term preservation, archiving, format registry, collection development, contracts

INTRODUCTION

The National Geospatial Digital Archive (NGDA) project began in November, 2004, as a partnership between the University of California at Santa Barbara (UCSB) and Stanford University (SU). Under the auspices of the Library of Congress's National Digital Information Infrastructure and Preservation Program (NDIIPP), the NGDA mission has been to investigate the long-term preservation and archiving of at-risk geospatial data and imagery. The NGDA cooperative agreement was one of the eight original NDIIPP awards made

Address correspondence to Tracey Erwin, Geospatial Librarian, Branner Earth Sciences Library and Map Collections, 397 Panama Mall, Stanford, CA 94305-2211, USA. E-mail: terwin@stanford.edu

to various institutions to further the knowledge of digital preservation of diverse materials. The process, successes, pitfalls, and results of the NGDA effort to address geospatial preservation are presented here.

The initial plan called for the creation of two parallel archives, one at UCSB and one at Stanford. Each archive targeted and ingested different collections. We planned to build a federated search mechanism to show the breadth of content collected. The early inquiries included technical questions such as how best to federate, practical questions of what to collect, and legal questions such as how to craft rights agreements.

The technical, practical and legal issues all presented various challenges—some that were anticipated, others that were not.

What Makes Geospatial Data Different?

Long-term preservation of digital data is becoming increasing familiar both in the cultural heritage space and in the world at large. As storage becomes less expensive and digital output expands, archives are proliferating. However, the requirements for archiving geospatial data are different from those for other types of digital content. As observed by NGDA principal investigator at UCSB Greg Janée (2009), "Whereas a multimedia document typically resides within a single file, geospatial data may reside in complex, multi-file objects. Whereas the interpretation of a PDF document may be defined by the format label 'PDF', and in turn by an entry in a central format registry, geospatial data may require extensive, product-specific context to interpret. Whereas a thesis or journal article is fixed upon publication, geospatial data can remain dynamic indefinitely due to the lifetime of the generating program and the need to be periodically reprocessed."

Janée goes on to state that there are a number of characteristics that define geospatial information. He notes that there is no uniform data model. This is due to the way in which these data are organized with different applications and file formats supporting different data types. Many geospatial formats are proprietary and are linked directly to the program in which they were designed to work. Geospatial data vary widely in the amount of information they show, for example, an individual feature may be depicted or a nationwide thematic layer may be displayed. Increasingly these data are being stored in relational geodatabases requiring sophisticated storage and archiving schemes. Geospatial imagery datasets are often quite large, with some satellites alone transmitting information at the rate of terabytes per day. Geospatial data may be produced over time, with satellites collecting information for decades, perhaps based on outdated technology or software systems. Metadata may be voluminous, but knowledge of the technology used and how it has changed over time is often difficult to find or not included with other information about the file itself (Janée 2009). In addition to these aspects of geospatial data, the datasets are extremely large relative

to nongeospatial datasets. These elements—file size, file complexity, format, metadata considerations, and the serial nature of the data—combine to create a variety of challenges. With these challenges, some of which became more apparent as we moved forward into the project, each institution began to build out repository architectures. These differences informed the process all along the way.

Overall Achievements of the Project

The results of the NGDA effort included both technical and nontechnical solutions for managing, preserving, and providing access to the content. Technically, two working repositories have been created based upon different components and goals. The Stanford team built a repository designed to ingest any content procured by the library system. The UCSB team built a repository specifically to manage geospatial content. UCSB developed a federated search engine using the Alexandria Digital Library (ADL) Globetrotter software allowing for metadata searching across both collections. Both groups worked extensively with the Library of Congress (LC) on content transfer methods for efficiently moving materials to LC's dark archive. A registry for the description of formats was created, which housed both the information about geospatial formats and the accompanying specifications. More than ten terabytes of content have been collected and ingested with more left to accession. Each of these results will be described from the vantage point of the respective nodes.

Nontechnical outcomes included a set of legal agreements designed to govern the content under license or copyright coming into the federation and a node-to-node agreement binding those groups that agree to collect and preserve content into the future. Collection development policies were written to govern the content collection in a systematic fashion. One policy is overarching for the network, and two individual policies have been written that direct each university's collecting efforts.

Each of these results will be described from the vantage point of the respective nodes. It should be noted that the legal agreements and creating a format registry were added to the scope of work as we went forward. These elements greatly strengthened the final results. In hindsight, it is hard to imagine the project without each of these components. It is also evidence of just how much the NDIIPP experience has been one of learning by doing.

Technical Efforts

The partnership between UCSB and Stanford proved to be mutually beneficial given the nature of our historic collecting strengths. UCSB has long been recognized for the work done to build and populate the ADL with its rich aerial photography collection, digital raster graphics, and Landsat imagery. Stanford's collecting interests centered on both raster and vector data

of California and particularly the Bay Area. Also, we crafted an agreement with map collector David Rumsey to archive his extensive scanned map collection, which would form the core of the initial targets for the archive.

The design of the projects was to build repositories at each institution and federate the metadata. Santa Barbara redesigned the ADL interface and released it under the moniker, Globetrotter, to allow for geographic selection of the content. The Globetrotter interface uses the Google Maps application program interface (API), which allows one to position the pointer over a certain location to reveal items containing those coordinates, such as the georeferenced maps in the David Rumsey collection at Stanford or the digital orthographic quarter quadrangles held by UCSB.

At UCSB, the NGDA repository is its own entity, standing alongside the ADL but separate from it. At Stanford, the NGDA content is ingested into the Stanford Digital Repository (SDR), a content-agnostic repository built to manage digital resources now and into the future. The conceptual framework for the repository is "trust over time." Garnering this trust is achieved by maintaining security, transparency, and proof (Johnson, 2007).

Long-Term Preservation

Over the course of the project considerable thought went into what preservation means. That thinking coalesced around several key concepts. We envision preservation as a relay over time in which threat mitigation and context preservation are the key elements. An additional element, the option to do nothing, is also part of a realistic preservation strategy.

The Relay Model

If the goal is to preserve information for a century or longer, any archive system, no matter how well designed or well supported or preservation supporting, is destined to become obsolete and unsupportable long before the century mark. Instead, long-term preservation is more likely to resemble a series of shorter term curatorships interspersed with handoffs, and an archive system supportive of such a relay must focus both on curating the information over its (the archive's) lifespan, and on facilitating the handoff to the next archive system. To maximize flexibility, this handoff ability should be supported independently at the institutional, repository, and storage levels.

Threat Mitigation

During the life of any archive in the phase between handoffs, its main job is to minimize the chances of data loss or corruption. Thus, for digital preservation to be successful, the threats to that process must be mitigated. Threats against bit preservation include bit rot such as memory checksum

errors, hardware and software failures in the form of hard-drive failure or faulty tape backups, geographic disasters like earthquakes or tornados, and internal and external security breaches by hackers or disgruntled employees. Renderability may be lost through bit corruption, losing the data as when a hard drive goes missing, or not knowing the specifications of a file format to read the data that can occur when a commercial entity does not write a specification for a certain format. Other threats include the lack of a proper renderer for a file format (e.g., having no access to a copy of Microsoft Word 1.0) or no longer possessing the physical devices to read the content, as with new machines lacking 5.25-inch floppy drives.

To combat these threats, the strategy of both UCSB and SU, in line with many other digital repositories, is to create multiple copies that are decorrelated across as many threat vectors as possible. UCSB has explored two technologies for providing redundant, reliable storage: an Archivas storage cluster that combines RAID (redundant array of independent risk) technology with active integrity monitoring; and Logistical Networking, a distributed storage technology. Stanford attempts to address most of these threats, but cannot mitigate all of them to the same degree because of cost and time constraints. For example, the SDR does not have appropriate file renderers for all content nor full specifications for all file formats. Some threat vectors are not entirely independent as well. While the SDR makes three tape copies of all content, this is carried out on the same software platform using the same tape library hardware/media.

Preserving Context

Preserving any type of information necessitates preserving both the information itself and sufficient context surrounding the information to render it intelligible in the future. That way, as the information's present context invariably changes or disappears over time, the preserved context can be referred to by future custodians. For geospatial data, the problems of preserving enough of the data's context, and of capturing it in the first place, are especially challenging. Whereas knowledge of the PDF format is sufficient to render PDF documents, and therefore usable by people, geospatial data can require much more, and more complex, contextual information. For example, using remote-sensing imagery in scientific modeling requires detailed knowledge of platform and sensor characteristics, and in many cases calibration and processing steps as well. Strictly speaking, such contextual information constitutes metadata, but in practice, being voluminous, it is not handled as such (for example, it is not stored in metadata records bundled with the data). Thus an archive of geospatial data can't simply rely on an external format registry to supply and preserve context; it must take on those tasks itself by archiving the context along with the data.

Doing Nothing

As is the case with many types of digital information, the amount of geospatial data and the rate at which it is being produced are both increasing. Owing to the high value of historical geospatial data in, for example, long-term climatological studies, there is a strong desire to preserve our entire heritage of climate data records. If we are to preserve even a fraction of this information, then preservation must be as inexpensive as possible. But beyond simply being inexpensive—always a goal in and of itself—an archive system must support a fallback preservation mode. Or, as Clay Shirky (2006) once observed, an archive must have "the option to do nothing." For any given piece of information, we cannot assume that the information will be maintained in a baseline usable state, let alone fully curated, at every point over a century or more. The perceived value of information changes over time, archive resources inevitably change over time, and there may be periods of time during which the upkeep of the information cannot be supported or justified. Furthermore, the risk of insufficient resources is acutely significant at handoff points, particularly handoffs between institutions. This risk can be mitigated by allowing the information to drop into a low-cost, unusable, or fallback state (e.g., a state incurring the cost of bit storage only) with the proviso that sufficient context is preserved to allow programmers and domain specialists of the future to resurrect usability as desire and resources permit.

The Archives

The final products of this conceptual framework are two working repositories that meet the needs of their respective institutions. As noted previously, UCSB chose to create a geospatial-only repository while the SU repository is content agnostic, containing many different data types. A summary of the architecture of the two data models follows.

The Stanford University Data Model

The SDR holds the data collected by Stanford as part of the NGDA. Originally conceived in 1997, the SDR is a content agnostic repository that is primarily focused on digital preservation, which includes two key components: bit preservation and renderability. Bit preservation requires that the content, in the form of bits, remain intact in their original form, while renderability ensures that the content be understandable by people.

The SDR has three components: the data model, ingest/validation, and storage (see Figure 1).

SDR's data model uses a Metadata Encoding Transfer Schema (METS) metadata format to encapsulate a digital object. A digital object is any grouping of files, such as the scanned pages of a book, all the tracks of a CD,

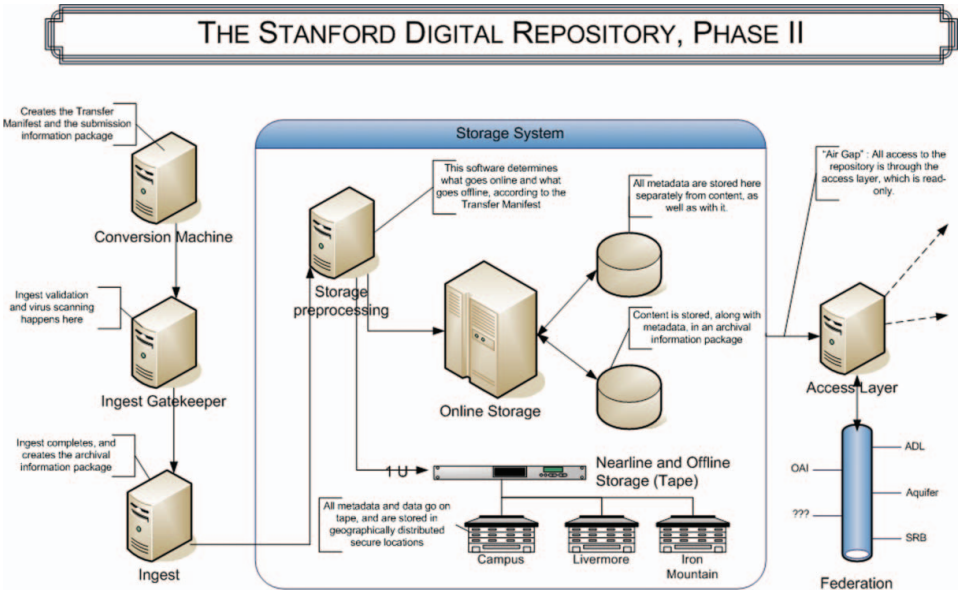

FIGURE 1 SDR technical architecture in 2006/2007.

or all bands of a satellite image. METS holds all types of metadata for the digital object including technical, descriptive, and administrative. Access to any file is through the METS metadata file itself, and there is a one-to-many, METS-to-individual-files mapping.

This means that one METS metadata file can point to many individual files; however, each individual file can belong to one—and only one—METS metadata file.

The ingest/validation step examines the METS package and validates the digital object. It validates the METS metadata file against an XML schema and then verifies some technical metadata, checksums, and file formats using the Harvard Object Validation Environment known as JHOVE. It inserts some data into the package and assigns an SDR unique identifier for each individual file.

Storage then makes multiple copies of the digital object and sends it to various storage systems and media. Currently, the object is sent to three tape copies and one disk copy. The METS metadata file is sent to a metadata server. After the initial ingest, all copies are immediately read back and their checksums are verified. Copies are then audited by use of checksums on a regular basis to prevent bit rot. The tapes are regularly sent to offsite locations to protect against geographic disasters. In the future, copies will be migrated to new media at regular intervals, and certain file formats will be migrated to new formats.

The development of the repository was accomplished by a team of two programmers, a metadata librarian, a metadata specialist during part of the

grant period, a product manager, a geospatial librarian, and the coprincipal investigator. The SDR has been ingesting content since 2006. SDR 2.0 is currently being designed.

The UCSB Data Model

UCSB's archive system is built around an open archive information system (OAIS)-compliant data model. There are two parts to the data model: a logical (or abstract) data model that defines a uniform, self-contained representation of archival objects, object semantics, and interobject relationships; and a physical data model that defines the representation of logical archival objects as directories, files, and XML manifests in a file system hierarchy. The logical data model defines a means of archiving context alongside data. The physical data model provides a fallback representation of archived content (beyond the archive's access mechanisms) and provides another level at which content can be handed off.

In terms of components, a central archive server builds and validates archival objects and associates data objects with semantics-defining objects (see Figure 2). The server is built on top of a minimal storage API that virtualizes the underlying storage system. On top of the archive server sit a number of components: an ingest crawler that crawls provider content

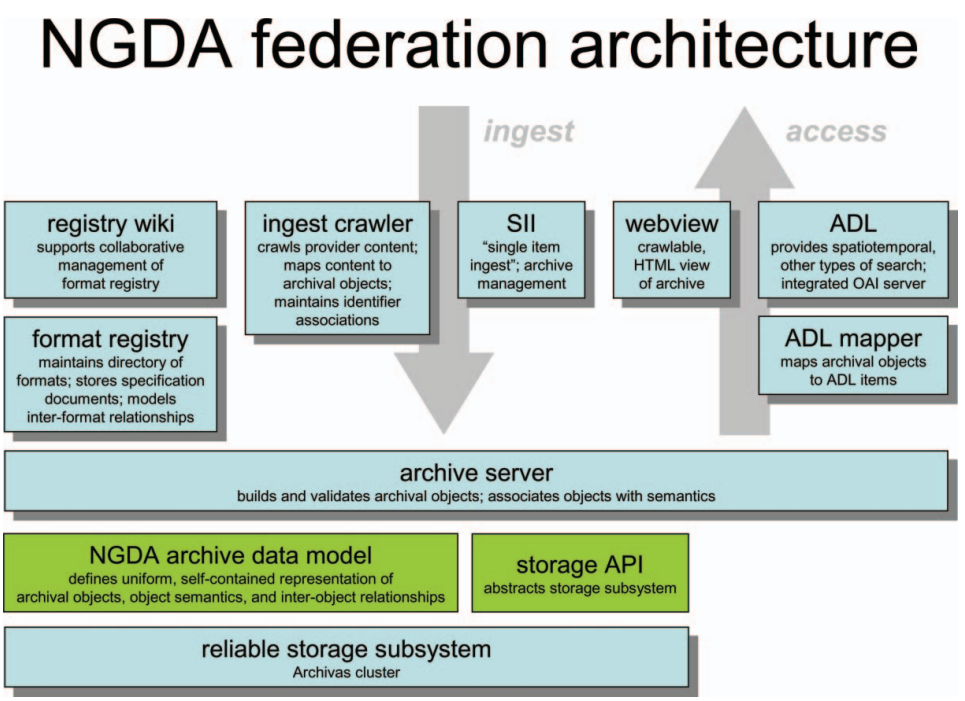

FIGURE 2 The NGDA architecture at the University of California at Santa Barbara.

and maps the same to archival objects, a format registry with wiki front end for building up and managing format definitions, a Web access mechanism that makes the archive appear to be a crawlable Web site, and a mapping mechanism to the ADL federated search system.

Format Registry Efforts

The development of a registry for geospatial formats was a natural outgrowth of the data ingest process. As we began acquiring data in various geospatial formats, it became evident that the knowledge base for geospatial formats was incomplete. Format registry and format sustainability work has been started by three different organizations: Harvard, developing the Global Digital Format Registry funded by the Mellon Foundation; the National Library in the UK with their development of PRONOM; and the Library of Congress's work on the Sustainability of Formats Web site. None of these entities has fully constructed format definitions for geospatial formats, but they have focused on more widely used specifications such as Adobe PDF files and Microsoft Word documents. Research carried out by the NGDA team during the first phase of the grant demonstrated that "the use of format registries is an implicit and important part of the metadata strategy for most archiving and preservation institutions" (Hoebelheinrich and Munn 2009). The NGDA team broke the work down into two spheres, with UCSB focusing on the technical infrastructure and Stanford working on populating the registry with content. Stanford subsequently subcontracted with Content Innovations to carry out the registry population.

Content Innovations began their work by analyzing the data model that the NGDA had created and compared it with the three registries noted above. They analyzed twenty- three geospatial formats and thirteen format subtypes in the registries to understand how well each registry's data model would fit our data types and to make recommendations as to how the NGDA model should be expanded. The results of Content Innovations' work show that the NGDA data model needed to be revised to accommodate the complex container structures and parent/child relationships so common with geospatial data. For example, as noted by Hoebelheinrich (2009), an Environmental Systems Research Institution (ESRI) shapefile contains numerous files that must travel together for the file to be read, stored, and archived.

The next phase of Content Innovations' work will be to build out registry definitions for all the formats collected; nine formats have been completed so far. NGDA will collect the documentation related to format as well as capture Web pages that explain the formats, specifications, and contextual information such as technological and software systems used to create the data in the first place. The output is being stored as XML. It will be disseminated to the newly formed Universal Digital Format Registry (formed by the consolidation of the GDFR [Global Digital Format Registry] and PRONOM)

when they are ready to accept it, and to both the UCSB and Stanford format registries. Format registry output will also be added to the Library of Congress's Sustainability of Digital Formats Web site. It is not expected that the NGDA will continue to house their own format registry indefinitely but rather it is expected that they will work in conjunction with the UDFR to add new formats and revise the old ones as necessary.[1]

Format registries are an important part of the long-term preservation equation intended to ensure long-term access to data as formats become obsolete. Format registry work will also enhance format migration efforts, such as the KEEP (keeping emulation environments portable) project, an initiative of eight European institutions in five countries.

Collections and Contract Efforts

Collection Development Polices

Geospatial data collection development policies (CDPs) differ from traditional paper-map policies in a number of ways. Larsgaard, Sweetkind-Singer, & Erwin (2006) provided a detailed description of the type of data to be collected to cover a broad area of digital data:

> The scope of collecting is solely in the realm of geospatial digital data. The term, "digital geospatial data," is defined as digital items, displayed as graphics that are georeferenced or are geographically identified. These are primarily composed of: digital maps; remotely sensed images (e.g., aerial photographs; data collected by satellite sensors); datasets (e.g., shapefiles, layers, geodatabases, etc.); atlases; globes (celestial and terrestrial); aerial views (e.g., panoramas); block diagrams; geologic sections; topographic profiles; etc."

Within this framework, both institutions agreed that we would focus on United States data and imagery, with UCSB collecting specifically for southern California and Stanford for northern California. The size of some of the datasets, especially for high resolution imagery, precluded either library from formally agreeing to collect and preserve a broader range of data and imagery than required by the research needs of each campus.

The NGDA collection development policy also discusses other aspects to consider when gathering digital data and imagery. Geospatial data are subject to versioning because of updated information being made available or to correct past errors in the data. When accessioning datasets, one should think about whether the data are versioned, and if so, how often this data should be collected to accurately portray change over time. Minimum core data elements should be collected with the information itself. It is recommended that the following fields be included: geographic extent, type, format, projection or coordinate system, scale or resolution, title, date, and

issuing body. The CDP discusses file formats, noting a general preference for open source, nonproprietary formats that may be manipulated in standard image processing or geospatial software. Widely used proprietary formats, such as the ESRI's shapefile, will also be readily accepted due to the ubiquity of use. One must also consider if the collecting node will take embargoed content that may be deposited as "dark" for a number of years before being released for public use. The NGDA nodes also decided to focus on three levels of data collection. The first level is to procure information created at the national, state, and local levels of government. The secondary focus will be on content from commercial firms, and the third level, collecting data created by individuals.

Three collection development policies have been created. The first policy is designed to be a general statement to be used as a model for other institutions. The other two are specific to each archiving node. Given the nature of these data, it is important for geospatial collecting at one's institutions to be aligned with current research interests and areas of expertise. In the long term this strategy will support more breadth to the archive as well as leverage the strengths of each institution. A near-term goal is to recruit other NGDA nodes that will collect in alignment with research and teaching needs of their respective institutions and, hence, their geographic region.

Legal Agreements

The NGDA project has created a contract for accepting copyrighted and licensed data, as well as agreements governing the interactions of the archiving nodes. The crafting of these legal documents was a collaborative process between the two universities. Lawyers at both institutions suggested that rather than having them craft an agreement, the librarians should create agreements to reflect the goals of the partnership and the needs of both content depositors and the universities for a clear understanding of the roles and responsibilities of each. Only after creating agreements in lay terms would the lawyers create the formal versions. This strategy proved to be quite time intensive. However, it did create a forum for many discussions that clarified the vision, mission, operations, and goals of long-term geospatial preservation.

Content Provider Agreement

Two types of data were identified for collection from a legal perspective: those in the public domain and those that are copyrighted or licensed. Although many of the data sets ingested into the two repositories are in the public domain, the copyrighted or licensed data require an agreement between depositor and collector.

The content provider agreement is designed to address copyrighted and licensed materials. The agreement consists of three parts: the main

body, Exhibit A, and Exhibit B, all available at the NGDA Web site at http://www.NGDA.org. The main body of the agreement specifies that the NGDA is copyrighted or licensed by the depositor and that specific rights are granted to the custodians of the archive. The grant of license states, "Content Provider hereby grants to Custodians a paid-up, non-exclusive, world-wide, transferable to reproduce, prepare derivative works of, distribute, perform publicly, display publicly, digitally transmit and otherwise use the Licensed Materials at no cost in any media no known or hereinafter created in accordance with the terms of the agreement" (Larsgaard, Sweetkind-Singer, & Erwin, 2007). Other provisions allow for removal of the content in rare circumstances and what should happen to the content if the agreement is terminated.

Exhibit A is written in conjunction with the depositor, and details the corpus of materials to be deposited into the archive. Information in this section includes a description of the content, whether it will be versioned, the metadata available, any conditions of use above and beyond those described the main body or Exhibit B, how the materials will be transferred, and the rights issues surrounding the collection. The frequency of communication, as well as methods of and appropriate subjects for communication, are described.

Exhibit B delineates the authorized users of the archive, the authorized uses of the data, and the management of the copyrighted and licensed materials by the nodes. The reasoning behind this structure was that the main agreement would remain unchanged through time, while Exhibit B could be amended as necessary to reflect technological or policy changes. Thus, rather than requiring depositors to sign new agreements each time there is a change of any kind, notification would be given to them when changes are made. Such changes, if unsatisfactory to the depositor, could allow them to request the removal of their copyrighted/licensed data from the archive. However, it is believed that depositors who are committed to long-term preservation of their materials will approve of our evolutions in policy and procedures. As noted by Sweetkind-Singer, Erwin, & Larsgaard (2009), "This section was the most difficult to craft and took the most negotiation between the UCSB and Stanford library staff and lawyers. All wanted this section to be acceptable to both a public and a private university. As we were writing this agreement, anything the group considered essential for all nodes had to go into this section. Only if all existing nodes and future nodes agreed to these provisions could content be shared across the collecting network."

Four classes of users are included in the agreement: those at the institutions initially archiving the content, walk-in patrons to the same institutions, users of the Library of Congress (which functions at present as a dark archive for all NDIIPP content), and those of the general public allowed by copyright and license provisions. The nodes agree to use the materials in

accordance with current copyright law. The nodes are allowed to make multiple ephemeral copies for preservation purposes. Other provisions allow the universities to use the content for teaching and research, including using it in course packs and sharing small amounts of the data with fellow researchers. The custodians agree to use best practices when storing and preserving the materials. We will also give credit to the copyright or license holder and let that holder know where their data are being stored.

Taken together, this three-part document is designed to allow for flexibility at each institution while creating a structure that allows for the sharing of content across the network to ensure redundancy of the materials.

Content Collection Node Agreement

A multipart structure was also adopted as the framework for our later agreements: the Content Collection Node Agreement and the accompanying procedure manual. The goal of the Node Agreement is to allow a legal framework whereby nodes within a network may share data knowing there are provisions in place that specify the standards by which the materials will be managed. Duplication of content across numerous storage platforms and in geographically distributed and secure preservation environments provides the highest likelihood that the content will be available in the near term as well as the distant future. The agreement makes it explicit that members accept the provisions of Exhibit B in the Content Provider Agreement. Depositors must know that no matter where their content is stored, it will be stewarded under the same provisions and with the equal care.

The agreement lays out the expectations and obligations of any institution that is or wants to be a part of the collecting network. In order to be a node, the new member must agree to the following: create a collection development policy, acknowledge that it has an institutional mandate to collect digital content, archive the data they collect, and agree in writing to be a part of the network. The agreement goes on to define the governance structure, responsibility of members, how nodes are indemnified vis-à-vis the other nodes, how content is removed from a node, and the process by which a node leaves the network.

The procedure manual provides specific details on the governance structure and describes communication between nodes, meeting frequency, administrative responsibilities, how potential nodes are identified and vetted, how new nodes join or depart, transfer of content in the event of discontinuance of a node, and the acquisition or removal of content.

Collection Efforts

More than twelve terabytes of data have been collected by the two nodes, including born digital data and scanned maps. The mandate from the Library

of Congress was to collect only digital content. In terms of the born digital data, the libraries have collected the following:

- California Spatial Information Library (CASIL): The state of California's geospatial content including scanned and georeferenced United States Geological Survey (USGS) topographic quadrangles; Landsat imagery; thematic layers for school districts, boundaries, and transportation networks; high resolution imagery; and more.
- National Map: Content has been retrieved from the USGS clearinghouse for geospatial data and imagery, including high-resolution orthographic imagery of the Bay Area, National Agricultural Imagery Project (NAIP) imagery, and the National Elevation dataset (NED).
- National Atlas: Layers of interest have been downloaded from this government site that focuses on national-level content of general interest to the public, including thematic data on geology, climate, and water.
- Miscellaneous other content: These data have been collected from a variety of Web sites and vendors, including scanned aerial pilotage charts, Landsat 7 imagery for the state of California, world shoreline data, and Vector Map–Digital Chart of the World (VMAP) Level 0 database.

Two collections of scanned data have also been accessioned into the preservation repositories.

- The David Rumsey Historical Map Collection: David Rumsey is a private map collector who spent decades creating a large collection of cartographic content related to the history of the United States in the eighteenth and nineteenth centuries. In the past decade, his focus has been on scanning these materials, and there are now more than 20,000 images available on his Web site: www.davidrumsey.com. The imagery is delivered a few times a year to the SDR for long-term retention.
- The Stanford Geological Survey: This survey spanned nearly 100 years from 1903 until 1995. Students learned field mapping over the summer and produced maps and reports as part of their work. This output was scanned many years ago, and the imagery is now part of the NGDA.

UCSB recently received a donation of the Citipix aerial imagery collection. This collection contains nearly half a million original, color-stereo, negative images of more than sixty-five cities across the United States. The ground resolution of the images is 6 inches. The content, when ready, will become part of UCSB's ADL and NGDA content.

Lessons Learned

Technical

Metadata are crucial when ingesting content for preservation. Both UCSB and Stanford found that acquisition and creation of high quality, usable metadata were time-consuming processes; datasets that purported to have full metadata did not, and others required rearrangement of existing metadata. At times, there was no standardized way to obtain further metadata, which led to discussions with data providers and extensive Internet searches for more complete information about a particular dataset.

Both repository staffs spent numerous hours preparing metadata prior to ingest. For example, Stanford's repository requires a transfer manifest (TM) to ingest data. The TM can be thought of as a packing slip designed specifically for each type of data that will be ingested. The TM is a METS XML file containing many types of metadata including descriptive, rights, technical, and structural metadata. Before the creation of the manifest, a preconversion step occurs in which all of the parts of the manifest are created. Once the parts are assembled, they are converted into a complete TM. SDR then takes the final package, validates it, and stores it along with the content. The creation of these transfer manifests is the most time-consuming part of the process at present. SU is rethinking this process to make this movement of data smoother.

Stanford also realized the need for and the importance of an administrative database. Currently, to locate an ingested item with the user's filename requires searching a log file to find the identification tag corresponding to that filename. Instead, there should be an external database that shows this connection and is easily searchable. A digital object repository is being built using Fedora that will be separate from the SDR.

UCSB is continuing active research investigating how their preservation architecture can be implemented using a third-party repository system such as Fedora or DSpace. In addition, the technical team will continue to develop a "crawl" interface that allows one to programmatically discover and download desired archival objects. Finally, UCSB will continue to develop the ADL infrastructure to allow for automation of the collection process.

While both UCSB and Stanford were building their own repository systems, commercial and open-source digital repository systems and software have evolved. The opportunities today are far greater than they were four years ago, such that institutions like the National Library of Medicine had numerous choices when they evaluated repository systems last year (Marill & Luczak, 2009). It is imperative that each technical team periodically review their choices for hardware and software, maintaining up-to-date systems and services.

Legal Agreements

The process of creating depositor-and-node agreements proved to be more time intensive than had originally been envisioned. This was due to several factors. The team wanted to explore and understand a variety of possible situations in which data would be taken in, retained, and used. Over time, numerous "what if" discussions took place between the team members, which resulted in the crafting of complex contingencies for potential situations. In several cases our legal counsel eliminated sections we had labored over, viewing them as unnecessary or because other clauses covered the same points. Ultimately, our lack of legal background slowed the process considerably. In retrospect, we might have been more efficient had we been given boilerplate agreement language and fitted our goals to existing legal language.

We also had trouble gaining continuous access to legal counsel. The lawyers on both campuses provided their expertise on an as-needed basis, which suited the team's needs, but often with long delays due to other priorities of the universities that superceded our ongoing work. We hope that the work done on these documents will provide the basic language and framework for others crafting these kinds of agreements.

Collection Development

The creation of collection development policies required a rethinking of the traditional type of collection policies. We decided that new sections had to be written to govern the specifics of digital curation (metadata standards, file formats) and likely producers of geospatial content. We quickly realized that there had to be limits to what each library collected and that more complete coverage of the United States would have to come from increasing the number of nodes in the network, striving for geographic diversity.

Ongoing difficulties remain in quickly and easily procuring data. Geospatial datasets are often large when desired in their aggregate rather than in small tiles. This necessitates the transfers of content via FTP or hard drives. There is no doubt that these processes will become commonplace in the future as both the libraries and the data creators learn to work together more effectively.

The Meaning of "At Risk"

The Library of Congress mandated from the beginning of the project, that the partners needed to collect content that was at risk. Throughout the life of the project, it has been difficult to define at-risk qualities. The team began with the idea that *at risk* meant materials for which there was only one digital

copy. Also, at-risk content included the output of small labs and institutes that lack funding for anything more than simple backups. Yet the more we tried to pinpoint the meaning of at risk, the more it became clear that everything digital potentially fits under this heading.

For example, elevation data for the United States is created by the United States Geological Survey. It is served out from the Earth Resources Observation and Science (EROS) data center. EROS maintains aerial, map, elevation, satellite, and land cover datasets and, as indicated on their Web page (EROS, M2009) is "home to the US National Satellite Land Remote Sensing Data Archive." The NGDA collection team decided to acquire elevation data for several western states including California. We considered these data to be of ongoing, high value to our researchers and believed the data set had a place in the NGDA. Yet, was it truly at risk as it was held at a national data archive? In an informal conversation between coprincipal investigator Julie Sweetkind-Singer and John Faundeen, archivist at EROS, Mr. Faundeen expressed appreciation for NGDA's efforts in this area. He pointed out that, while there is replication of data, all servers are at the EROS facility in Sioux Falls, South Dakota. Given tornado activity in the area and the lack of geographic distribution of the EROS data, he considers this data at risk. Indeed, according to readily available data (http://www.city-data.com/city/Sioux-Falls-South-Dakota.html), Sioux Falls historical tornado activity is not only well above the rest of South Dakota, it is 171% greater than the U.S. average. As previously noted redundancy as threat mitigation, in this case potential natural disaster, is significant for long-term preservation.

Creation of a Digital Data Workflow

Data are also at risk as long as there is no routine method for processing them once they have been acquired. The NGDA has acquired data via download, on hard drives, from CD ROMs, and in-house servers. There is no single point of entry nor is there a routine process for the acquisition of such data.

The life cycle of paper materials is well understood, and it is replicated across thousands of libraries around the world. Typically, books and paper journals are purchased through a well-known set of publishers or vendors with internal library systems set up to identify, purchase, catalog, and pay invoices for those materials. This process is in place and requires little-to-no monitoring by the ordering librarian.

The same cannot be said for digital materials. The life cycle of digital materials in a library is still in a state of flux, especially for nonelectronic journal and book content. Stanford's experience is a case in point. For example, the geospatial librarians decided that they want to acquire high resolution orthographic imagery of the San Francisco Bay area from the National Map. The acquisition process for obtaining the content was handled through a

series of e-mails directly between the Branner librarian and a contact person at the EROS Data Center. This process took months owing to delays in identifying the correct person, and making sure the desired data were available. When all was agreed upon, a hard drive was mailed from Branner Library to Sioux Falls. The hard drive was mailed back with an invoice, which was sent to the payments department. Data then had to be checksummed and backed up onto a server for redundancy. This entire process was handled by Branner librarians including cataloging of the content. At this point, Stanford does not have a way to search for or serve these data through a spatial data catalog, so access is on a case-by-case basis through the GIS librarian. Finally, the imagery will be ingested into the SDR with appropriate metadata downloaded from the National Map Web site.

It is obvious that at this point the digital workflow requires intervention in nearly every step of the process by those acquiring the data or imagery in the first place. The vendors for the content are dispersed. The process to procure the content can be laborious and slow. The need to duplicate the content in a robust manner is immediate (and perhaps more challenging because of the size of the datasets). The display, access, and use of the content presents unique challenges for geospatial data because typical library OPACs are not set up to handle the complexities of geospatial data, that is, multiple data sets on a single CD/DVD/hard drive. In addition, long-term preservation is more likely to succeed with thorough metadata, which is not always provided.

It is clear that strategies for managing data from beginning to end will emerge as more and more libraries collect these data. For now, the approaches are piecemeal and fraught with delays and hurdles. Simply bringing the data in-house does not mean it is no longer at risk if good data management practices have not been put in place. It is obvious that a great deal more research and thinking must go into this area.

Looking Ahead

In the four and a half years since the project's inception, our thinking about long term preservation has evolved. Where the NGDA mission was to build an archive and focus on access at a later date, it is now understood that access is inextricably linked to preservation. On the technology side, scalability is now and will continue to be an important issue. While not being considered at this point by UCSB or Stanford, cloud computing may be the next step in an evolution that will allow scalability and access to coexist more readily. As geospatial data sets are notoriously large, the scalability alone offered by cloud computing is compelling.

Perhaps the most pressing issue facing digital preservationists is who will pay for preservation. While UCSB and SU built their repositories with the aid

of the NDIIPP project funds, many midsized and smaller institutions may never have the financial clout to build robust repositories.[2] Both UCSB and SU have current top-level administrative support for the ongoing existence of their repositories, and storage gets less expensive over time; however, the political will at each institution to maintain these repositories remains necessary.

New Partners

The original vision of the NGDA was a robust collecting network of archiving nodes across the United States. In the latter half of 2009, the NGDA plans to expand its archiving network by attracting new partners to join in the mission of geospatial data preservation. Although it would have been appealing to add new partners at intervals over the past four years, the team realized that expanding the network in the midst of the work of developing the archives, policies, and procedures was an overly ambitious goal. In the remaining six months of the granting period the team will focus on increasing the size of the federation. The NGDA partnership is currently engaged in preliminary discussions with several potential nodes. This expansion, part of the original mandate from the Library of Congress, is important to achieving the goals of forming an active network of partner institutions, geographic distribution of archiving sites, and increased scope. The potential new nodes include other academic institutions collecting geospatial data, as well as other map libraries and state data clearinghouses. These potential partners and allies have significant geospatial collections, technical expertise, and a desire to join our existing structure.

CONCLUSION

The results of the NGDA experience are multifaceted. In practical terms, the successful ingestion of data into working repositories is the most significant outcome. In addition to two functioning repositories, the team continues to increase its knowledge regarding how to structure metadata for ingest and improve all aspects of the archiving process. Also, the group has learned a great deal about selection and has cultivated the ability to identify and collect valuable datasets. The creation of formal legal agreements for obtaining copyrighted data and agreements governing the relations of archiving partners are significant achievements. Another important outcome is our format registry work, which will aid in the long term understanding of and access to data collected today.

Whether the future holds archiving in the "cloud," and how the economics of digital preservation eventually resolve, the next steps in digital archiving of geospatial data will be built upon the work accomplished by

projects like the NGDA and those of all the NDIIPP partners who have laid the foundations of digital preservation.

NOTES

1. More complete registry analysis work can be found in the *Report to National Geospatial Digital Archive regarding geospatial treatment in data format registry efforts* by Content Innovations at http://contentinnovations.com/NGDA/NGDAFindings05042009withexcel.pdf.

2. "Sustaining the Digital Investment: Issues and Challenges of Economically Sustainable Digital Preservation." Interim Report of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access, December 2008. http://brtf.sdsc.edu/biblio/BRTF_Interim_Report.pdf

REFERENCES

City-Data.com. 2008. Sioux Falls, South Dakota. http://www.city-data.com/city/Sioux-Falls-South-Dakota.html (accessed July 21, 2009)

EROS Data Center. National satellite land remote sensing data archive. http://edc.usgs.gov/archive/nslrsda/overview.html (accessed July 21, 2009).

Hoebelheinrich, N., and N. Munn. "Assessing the utility of current format registry efforts for geospatial formats." In *Archiving 2009, Final Program and Proceedings*, 30–34. Springfield, VA: Society for Imaging Science and Technology, 2009.

Janée, G. "Preserving geospatial data: The National Geospatial Digital Archive's approach." In *Archiving 2009, Final Program and Proceedings*. 25–29. Springfield, VA: Society for Imaging Science and Technology, 2009.

Johnson, K. 2007. Stanford Digital Repository. http://library.stanford.edu/depts/dlss/collections/sdr.htm (accessed July 21, 2009).

Larsgaard, M. L., J. K. Sweetkind-Singer, and Erwin, T. "Collection development policy for the National Geospatial Digital Archive," 2006. http://www.ngda.org/research.php#CDP. (accessed July 10, 2009).

Marill, J., and Luczak, E. "Evaluation of Digital Repository Software at the National Library of Medicine." *D-Lib*, May/June, 2009. http://www.dlib.org/dlib/may09/marill/05marill.html/ (accessed June 23, 2009).

Shirky, C. "Presentation related to partnership building." Speech given at Digital Preservation Partners Meeting, January 9–11, 2006, in Berkeley, CA.

NDIIPP Partners Meetings, Washington, DC.

Sweetkind-Singer, J., T. Erwin, and M. Larsgaard. "Legal agreements governing archiving partnerships: The NGDA approach." In *Archiving 2009, Final Program and Proceedings*. 11–15. Springfield, VA: Society for Imaging Science and Technology, 2009.