Unchartered Territory: Building a Network for the Archiving of Geospatial Images and Data

by Julie Sweetkind-Singer, Librarian, Branner Earth Sciences Library & Map Collections, Stanford University

Librarians working in the realm of geospatial information routinely live in a 20% world. When librarians collectively talk about what systems will be set up to handle content, this typically means books and journals, the 80%. I have found this to be true dealing with either paper-based maps and aerial photography or digital data and imagery. Procedures for paper-based content are well formulated. It is what libraries have learned to do over the last few hundred years. But, lifecycle management of digital content is not fully understood, especially when dealing with that 20% of non-standard content. Over the last four years, librarians and technologists at Stanford University (Stanford) and the University of California, Santa Barbara (UCSB) have worked together to learn how to address the challenge of digital lifecycle management, especially focusing on the last component in that cycle, long-term preservation. As is often the case, what we thought we needed to do to understand long-term preservation of digital geospatial data and imagery was the tip of an iceberg that was much larger and more complicated than we imagined.

Funding the Project

In December 2000, the United States Congress authorized nearly $100 million to fund a national effort to "set forth a strategy for the Library of Congress in collaboration with other federal and nonfederal entities, to identify a network of libraries and other organizations with responsibilities for collecting digital materials that will provide access to and maintain those materials."[i] The program was to be administered through the National Digital Information Infrastructure & Preservation Program (NDIIPP). Congress mandated the money be used to develop policies, protocols, and strategies for the long-term preservation of "at-risk" materials. Stanford and UCSB were in the first round of funding announced in September 2004, which included eight awards totaling nearly $14 million. Stanford and UCSB proposed the creation of the National Geospatial Digital Archive (NGDA). The goals of the NGDA were to create a national federated network committed to archiving geospatial imagery and data, to investigate preservation strategies, to collect "at-risk" content across a spectrum of formats, and to develop policy agreements governing retention, rights management and obligations of the partners. Along the way, we have had to build two archival storage systems, create collection development policies, content provider agreements, partnership agreements, a format registry, and an interface to federate the materials through an online catalog. This paper will focus on the non-technical parts of the work we have done.

The NDIIPP agreement clearly stated that these awards were specifically for archiving digital data. While we were not able to allocate money towards digitizing paper collections, we could archive previously scanned materials. The geospatial data and imagery we chose to collect spanned a wide array of content types and formats including scanned historical maps from the David Rumsey Collection and the United States

Geological Survey, to satellite imagery such as LANDSAT, digital aerial photography, and data layers created to provide information about the earth's surfaces and features including elevation, ocean depths, land use, transportation, and weather, to name a few. Increasingly geospatial content is being used to inform decisions both in the private and public sector in areas ranging from population studies and census construction to land use policy and government aid determinations, and as such, it is valuable data to retain for future generations.

Data Unlike Any Other

Digital geospatial data are different from other types of data in significant ways, which affected the way we thought about and dealt with the content. First, the amount of data being created is massive. A single satellite may send down a terabyte of raw data per day. Second, the data are often released in time slices requiring decisions to be made early on as to the frequency of capture. For example, MODIS satellite data are constantly collected and then aggregated into 16- and 32-day composites. MODIS satellites capture data in 36 spectral bands, which can then be used to study large-scale changes in climate and land, ocean, and atmospheric processes. Third, proprietary software makers, such as ESRI, dominate the marketplace resulting in file formats that are ubiquitous and, at times, less well understood than their open source counterparts. Fourth, there are a large number of file formats, many of which require contextual information in order to be understood in the future. Finally, the data structures are often quite complex with multiple files creating a single "layer" of information, meaning they always need to travel together in order for the file to be read.

Rules of Engagement

The issues regarding massive amounts of content immediately made us realize that we would need to write Collection Development Policies (CDPs) detailing what would and would not be collected by each NGDA member, called a node. Choices would have to be made about what to collect and we wanted to elucidate why we were deciding one way or another. While both subject specialists, Mary Larsgaard at UCSB and I, had CDPs governing our paper map collections (with a nod toward digital materials), neither of us had written any specifically for our digital collections. With the help of Tracey Erwin from Stanford, we ended up writing three policies: an overarching policy that would apply to any node that joined in our collecting effort, and then one for each campus that was specific to that university's research needs. The CDPs include the typical topics such as collection purpose, selection criteria, and scope. They then continue with additional sections on date/chronology, formats, copyright, metadata recommendations, sources for collecting data, and a glossary.

Once we knew what we wanted to collect, we needed to ensure that if the collections were not in the public domain there was an agreement with the content provider as to the rights and responsibilities of each entity detailing how the information would be stored, used, and distributed. A Content Provider Agreement (CPA) was drafted by the relevant working group with the help from the legal staff at Stanford and UCSB. The agreement is structured in three parts. First, the main section of the agreement describes the nature of the NGDA, the grant of license allowing the university to hold the data/imagery, the distribution and use of the materials, and how the contract may be terminated. This section may be amended as a node sees fit to meet the needs of its specific institution. Exhibit A provides space to describe the content and any procedural matters relating to that content. Finally, Exhibit B lists in detail the authorized users and uses of the licensed materials as well as the management of the materials by the "custodians" of the content. This section of the contract is required to be a part of any agreement signed by the content owner regardless of the node in which the content is deposited. Having all of the universities (or other archiving entities) agree to the terms of Exhibit B allows us to share the data and the metadata as needed for preservation purposes. This provision also makes it clear that no matter which node originally receives the content, it will be treated in same way.

The next step was to create a contract between the collecting institutions who agreed to participate in the NGDA. We worked to create a contract that does not violate any provisions of the Content Provider Agreement, allows the participating institutions to adapt to new circumstances and technologies over time, and gives the content providers a say if there were to be large-scale sweeping changes in the way we decide to do business. The decision was made to create a highly structured and yet general contract that clearly laid out the expectations and obligations for participation. We set up a governance structure, noted each member's responsibilities, laid out how to remove content from a node no longer able to host it, and specified how a node would leave the organization. The specifics for how processes would be handled are filled out in the Procedure Manual. This two-part structure allows us to change the Procedure Manual as necessary without the need to get the main agreement between the partners re-signed. For example, the main contract states that the nodes will convene "as provided in the Procedure Manual," to discuss topics such as the acquisition of new content, adding new nodes, and operating procedures. What the contract does not do is state how often this will happen, who will pay for it, who will host it, and if the meeting must be in person. All of these particulars reside in the Manual, which is much easier to change. It is hoped that this structure will lessen bureaucracy and allow us to adapt quickly to changes over time.

Collaborative Collecting

Content collection began in earnest from the start of the award period. Both universities had content identified from the start. UCSB ingested the geospatial content from the California Spatial Information Library (CASIL), which included scanned topographic maps, LANDSAT imagery of the state of California, thematic data layers including transportation, boundaries, elevation, farming, and structures. Stanford accessioned the David Rumsey Collection of 18th and 19th century scanned historical maps and the output (maps and field notebooks) of the Stanford Geological Survey. The collections continue to grow rapidly with UCSB acquiring the Citipix aerial imagery collection of 65 metropolitan areas across the United States with over half a million images. Stanford has collected high resolution imagery of the San Francisco Bay Area, elevation data, data layers from the National Atlas, coastline data, and scanned aeronautical charts.

One of the current problem challenges we at Stanford are addressing is setting up a structured workflow for the data life cycle. For example, we acquired imagery and elevation data from the United States Geological Survey's EROS Data Center. It was delivered on a hard drive. The data then had to be reliably duplicated on another storage medium in case the hard drive failed. Metadata was not included and so had to be pulled from the USGS National Map Seamless Server. Now that the metadata and the content are in place, decisions have to be made about how the content will be stored in the archive – as a whole collection or in its individual parts. The data and imagery then must also be brought into the library workflow for patron use with cataloging, display options, and the ability to download the files of interest. There are many pieces to the puzzle with potential failure points in numerous spots along the way; our approach is piecemeal and not yet fully formed. The goal, by the end of the agreement with the Library of Congress (August 2009), is to have a comprehensive workflow for our digital acquisitions that is as seamless as the process for our paper-based materials.

Finally, a format registry is being created as a joint effort by both universities to maintain technical information about the formats being archived. The registry will house specifications, standards, white papers, and ancillary information about the formats in order to increase the likelihood that they will be understood and usable in the future. It has been a complicated process to decide exactly what should be kept, where it should be housed, and when to say enough is enough in terms of the amount of information collected. We have been watching the developments of similar projects at Harvard's Global Digital Format Registry[ii] and the United Kingdom National Archives' PRONOM[iii] projects as we would eventually like to pool our registry information.

Conclusion

The work on the NGDA project has been challenging, interesting, and critical to the success of the geospatial collections at both schools. While it is easy to grab digital content and bring it in house, it is entirely a different matter to make sure that access is provided now and into the future as securely as any book we pull off our shelves. It is our hope that the work we have done to address and resolve some of the issues inherent in geospatial data collection will be of use to others in the field. At our Web site, www.ngda.org, we have posted the collection development policies, contracts, the NGDA interface to view a sample of the collections, articles and publications, tools, and technical architecture specifications.

[i] U.,S. House of Representatives Report 106-1033 - Making Omnibus Consolidated and Emergency Supplemental Appropriations for Fiscal Year 2001. Accessed March 18, 2009 at http://frwebgate.access.gpo.gov/cgi-bin/getdoc.cgi?dbname=106_cong_reports&docid=f:hr1033.106.pdf .

[ii] GDFR: Global Digital Format Registry. Accessed March 23, 2009 at http://www.gdfr.info/.

[iii] The Technical Registry PRONOM. Accessed March 23, 2009 at http://www.nationalarchives.gov.uk/PRONOM/Default.aspx.