Framing a National Strategy for the Appraisal and Selection of Geospatial Data

Nov. 17 and 18, 2010

Montpelier Room, LM-619 (6th floor, James Madison Building)
Library of Congress, 101 Independence Avenue SE, Washington, DC
Twitter hashtag: #locgeo2010

Attendees

Brett Abrams
Senior Archivist, Electronics and Special Media Records Division
National Archives and Records Administration
Brett.abrams@nara.gov

Martha Anderson
Director of Program Management, National Digital Information Infrastructure and Preservation Program
Library of Congress
mande@loc.gov

Colleen Cahill
Digital Conversion Coordinator, Geography and Map Division
Library of Congress
cstu@loc.gov

Lisa Dove
Deputy Assistant Director, Knowledge Services Group, Congressional Research Service
Library of Congress
ldove@crs.loc.gov

Robert Downs
Senior Digital Archivist
Columbia Center for International Earth Science Information Network
rdowns@ciesin.columbia.edu

Erin Engle
Digital Archivist, National Digital Information Infrastructure and Preservation Program
Library of Congress
eengle@loc.gov

Maurie Caitlin Kelly
Director, Pennsylvania Spatial Data Access, Pennsylvania State University

John Faundeen
Archivist
USGS Earth Resources Observation and Science Center
faundeen@usgs.gov

Pete Folger
Specialist in Energy and Natural Resources Policy, Congressional Research Service
Library of Congress
pfolger@crs.loc.gov

Michelle Gallinger
Digital Archivist, National Digital Information Infrastructure and Preservation Program
Library of Congress
mgal@loc.gov

Abigail Grotke
Lead Information Technology Specialist
Library of Congress
abgr@loc.gov

Bob Horton
State Archivist
Minnesota Historical Society
Robert.horton@mnhs.org

Steven Jackson
Technical Executive
National
Geospatial-Intelligence Agency
Steven.P.Jackson@nga.mil

Jan Johansson
Data Librarian, Congressional Research Service
Library of Congress
jjohansson@crs.loc.gov

Nancy Ritchey
Archive Branch Chief in the Remote Sensing Applications Division, NOAA National Climatic Data Center

pasda@psu.edu

Dan Kowal
IT Specialist, Data Management, NOAA
National Geophysical Data Center
Dan.Kowal@noaa.gov

Barney Krucoff
GIS Director, District of Columbia Office of the Chief Technology Officer
Barney.Krucoff@dc.gov

Butch Lazorchak
Digital Archivist, National Digital Information Infrastructure and Preservation Program
Library of Congress
wlaz@loc.gov

Bill LeFurgy
Project Manager, Digital Initiatives, National Digital Information Infrastructure and Preservation Program
Library of Congress
wlef@loc.gov

Steve Morris
Head, Digital Library Initiatives and Digital Projects
North Carolina State University Libraries
steven_morris@ncsu.edu

Jacquie Nolan
Cartographer, Geography and Map Division
Library of Congress
jnol@loc.gov

Mike Ratcliffe
Assistant Division Chief, Geocartographic Products and Criteria, Geography Division
U.S. Census Bureau
michael.r.ratcliffe@census.gov

nancy.ritchey@noaa.gov

Abby Rumsey
National Digital Information Infrastructure and Preservation Program
abby@asrumsey.com

Paul Schirle
Geospatial Information Systems, Knowledge Services Group, Congressional Research Service
Library of Congress
pschirle@crs.loc.gov

Joe Sewash
Services Program Manager at NC Center for Geographic Information and Analysis
joe.sewash@nc.gov

Michelle Torreano
Geospatial Metadata Coordinator, Environmental Protection Agency
torreano.michelle@epa.gov

Allan Wiley
Team Lead for the Geospatial Information Library, U.S. Army Geospatial Center
allan.s.wiley@us.army.mil

Peter Young
Chief, Asian Division
Library of Congress
pyou@loc.gov

Executive Summary

The National Digital Information Infrastructure and Preservation Program convened a variety of experts during November 17 and 18, 2010, to consider issues associated with geospatial data appraisal and selection. An earlier meeting to discuss framing a National Preservation and Access Strategy for Geospatial Data recommended a near-term focus on how collecting institutions should work together to decide which data sets merit preservation and how best to keep them accessible.

Institutions that keep geospatial data traditionally have used separate approaches to deciding what sources to add to their collections. Libraries often think about selection, a term based on institutional collection policies. Archives, on the other hand, use appraisal, which is a term rooted in recordkeeping mandates, including determinations of what constitutes permanent or archival records. Given the present ubiquity, use and importance of geospatial information, and also given the limited resources that all collection institutions have to manage holdings, it is worthwhile to take a higher-level view about choosing geospatial data sets for ongoing preservation and access.

Meeting participants were asked to address five questions:

* Are there any appraisal or selection policies that are good candidates for sharing across institutions?
* What aspects of the issue require more investigation?
* Are there shared service models to consider?
* Are current a/s policies and practices robust and adaptable enough?
* What are some next steps to advance the practice of geospatial data appraisal and selection?

The group identified a number of information resources that can be added to the NDIIPP-supported Geospatial Data Preservation Resource Center . Other potential next steps included sharing existing polices such as the NOAA Procedure for Scientific Records Appraisal and Archive Approval for other organizations to adapt for their own use, and exploring ways for archives and libraries to partner more effectively with data creation entities in the public and private sectors.

NDIIPP will continue to parse all the various issues associated with a national strategy for geospatial data preservation and access. A key operational vehicle for this work will be the Federal Geographic Data Committee Users/Historical Data Working Group, which has a number of Library staff in leadership roles.

Meeting Details

Laura Campbell opened the meeting and welcomed attendees to the Library of Congress.

Bill Lefurgy of the Library of Congress provided an overview and set the stage for meeting's goals (slides available). He discussed the themes that had arisen from the previous meeting in Nov. 2009:

- Need for a clearinghouse of information
    - Geopreservation.org
- A need for a further exploration of appraisal and selection activities
    - Appraisal is recordkeeping role
    - Selection is about policy and what to bring into the institution

Bill introduced the white paper, "Appraisal and Selection of Geospatial Data" prepared by Steve Morris of the North Carolina State University Libraries. The paper framed 5

questions that provided structure for the meeting:

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?

The meeting would address these questions, with the hope that the discussion would lead to concrete next steps and action items.

Comments from Initial Small Group Discussions of the Five Questions

- One observation was that the questions may be looking at the problem from the "wrong end of the telescope" in that they are framed from the perspective of cultural heritage stewarding organizations as opposed to the creating organizations.
- There are still limited resources available to do appraisal and selection properly. One solution is that the resources for appraisal and selection come from the providers themselves in some way.
- Appraisal is still largely done on a domain-specific basis without adequately taking into consideration appraisals for re-use purposes that cross domains. Consider ways to include perspectives on future use and analysis of geospatial data.
- It is useful to capture metadata in a variety of forms (including metadata about missing data) in that captured metadata might stimulate access to the data, leading to preservation actions.
- At what point in the OAIS Reference model (SIP, AIP, DIP) do preservation actions need to take place? Which copy is the archival copy when value-added additions to data have taken place?
- Current institutionalized appraisal practices have gaps and inconsistencies.
- Overemphasis on formats over information ecologies.
- Geospatial data is too important to make definitive appraisal decisions now, better to capture in quantity and use advanced discovery tools in the future.
- State and local governments are looking to the Federal government for high-level guidance on these issues. It is important to address these questions at the national level to minimize interoperability issues.
- How to deal with the issue of classified data?
- Terms of services and licensing contracts with data producers limit sharing among federal agencies.
- Do framework data themes provide a useful context to begin appraisal processes?

- Are data.gov and GOSS being archived?
- Go with the market standard – open source should not be a mandate. Formats become standards when they are market-driven.
- Collect materials with long-term value from data aggregators (Census, GeoEYE). There are many organizations that gather the data already.
- Open access to databases and allow the libraries and archives to pull the data directly from the aggregator (similar to web crawling; make it passive for the data producers and an easier "sell" to state and local archives).
- Need to more fully explore the role of data restrictions (IP, security, confidentiality, etc.).
- Records management training should include geospatial data.
- Use the NSF data sharing policy as an input to new appraisal approaches.
- Work to define the value of preserved geospatial data.
- Barriers to successful preservation include a lack of clear shared incentives along with multiple priorities and multiple stakeholders.

Presentations

Case Study: Current Appraisal and Selection Policies
Introduction: Butch Lazorchak, Library of Congress (slides available)

Nara Guidance, Brett Abrams, National Archives (slides available)
Provided examples of NARA guidance for geospatial data transfers and records scheduling. He noted that NARA currently has a limited knowledge on how to capture and schedule geospatial data, with a gap existing between producers, record keepers, appraisers – a "wobbly triangle" of relationships.

Learning from the Past: Geospatial Acquisitions at the Library of Congress
Colleen Cahill, Library of Congress (slides available)
She noted that the Library has two main sources of digital data: scanning of older, out of copyright materials; and acquisitions. They scan largely out-of-copyright materials on public request. This also includes staff and government requests and legal requests that may include in-copyright materials.

The born-digital acquisitions include materials requested by patrons and Congressional staff. Most of these are licensed, and they come in a great variety of formats. The Library collections duplicate other collections, but this is by choice, though questions remain on how to track archiving entities to avoid undesired duplication of effort.

In the current era of fiscal austerity collecting agencies need to learn how to do more with less. Crowd-sourcing activities such as the New York Public Library's Map Rectifier (http://maps.nypl.org/warper/) application help the agency take advantage of data sharing resources.

USGS Appraisal Process
John Faundeen, EROS Data Center (no slides)

He provided an overview of the types of data being archived and maintained at the USGS/EROS center, with a description of EROS' appraisal and section process, which has been in place since 2005. One of the policy requirements is to appraise all data for long-term preservation

The EROS review process is heavily documented. There are 42 appraisal questions, which are access- and mission-driven but also focus on economic considerations, including how to uncover how much the collection will cost over time. The appraisal process includes EROS staff as well as rotating contributions from independent scientists in the field. Under this process, 47 collections have been reviewed, 30 have been retained/accepted, and 17 rejected/disposed

DAY 2

Presentations

Current Issues in Appraisal and Selection: What specific aspects of appraisal and selection require more investigation?

Penn State Institutes of Energy and the Environment, Data Repositories Pennsylvania Spatial Data Access (PASDA)
Maurie Caitlin Kelly, PASDA (slides available)
PASDA is the official state geospatial data repository with thousands of data sets, 12+ million uses for their data in 12 months and 14 TB of data downloaded in 12 months. They incentives for data acquisition include a promise to create the metadata and through providing tools for people to create their own metadata. They crowd-source data quality through about 7,000 emails/phone calls each month regarding corrections to existing data.

They keep all imagery and provide access to all years of data imagery and provide an FTP site for direct download. The data is appraised by the people who create it (scientists, biologists, etc.). She discussed one particular project, Chesapeake View, a collaborative effort across the states and localities of the Chesapeake Bay basin to share information establish a central location for sharing information about the Chesapeake. She also discussed the Penn State Geospatial Commons, which is focused on storing data from Penn State research community activities supported by Federal granting agencies.

Current Issues in the Appraisal and Selection of Geospatial Data
Bob Downs, CIESIN (slides available)
Described the establishment efforts of the Geospatial Data Preservation online clearinghouse (geospreservation.org) being funded by NDIIPP.

He also discussed the work of SEDAC, a collaboration between Columbia University and the NASA Socioeconomic Data and Applications Center, on developing a long-term archive of data and documentation relevant to human dimensions of global change.

NOAA National Geophysical Data Center
Dan Kowal, NOAA (slides available)
Discussed the NOAA Procedure for Scientific Records Appraisal and Archive Approval guidance documents and the NOAA data center procedures for appraising records. He also discussed the NOAA use of the NASA Cost Estimation Toolkit as an appraisal tool to assistant them in estimating the future costs of a particular data set.

Users Perspective

Abby Rumsey (no slides)
She discussed examples of how humanists use geospatial data, such as the use of layers found on Google Earth to recreate Humboldt's travels, or plot the major events in the life of Jane Austen and her characters. She described this as an activity quite different from mapping in the sense the meeting participants conceive of it. What matters is that through environmental sciences we know about the interaction of humans and their physical environment.

Humanists are becoming increasingly comfortable with thinking of evidence as data. The convergence of environmental science and environmental humanities is that both communities like to show changes over time. Integrating historical and current sources will become more important over time.

She also urged the stewarding organizations to collect digital geospatial information in bulk while not worrying too much about the costs of delayed curation. Different research methods will be available in the future that will solve some of the cost issues.

Steven Jackson, NGA (no slides)
NGA is both a user and a generator of data, and much of the data they deal with is transient: they struggle with how you keep track of that data over time. NGA's most significant products are the Homeland Security Infrastructure Program DVDs that compile the best available Federal government and commercial proprietary data sets for the homeland security and defense communities. These were originally for Federal use only, but licenses have since been extended to include disclosures to state and local governments during disasters.

NGA has significant resources to store data and in practice is retaining most data while deferring long-term stewardship decisions.

Allan Wiley, US Army Geospatial Center (slides available)
The Army Geospatial Center is also a producer of data, but many of their products derive from others other products, so in that sense they are users. More and more their materials are almost entirely born-digital

Paul Schirle, Congressional Research Service, Library of Congress (slides available)
CRS serves Congressional committees and members of Congress and provide expert

assistance at every stage. CRS services and sources are authoritative, primary resources when available.

Geospatial representations can boil things down and make it easier for members to deliberate and make decisions. They convey complex and scientifically significant issues. They reveal consequences and impacts that might be concealed in textual documents. Legislation may not clearly depict the geography affected by a decision. Spatial analysis can help enumerate those issues.

As users, CRS recognizes that the source will always know more about the data and can identify the best resource. These sources could be clearinghouses of data, as long as they support standardized metadata and discovery that points to citable, trusted repositories.

Robert Horton, Minnesota Historical Society (no slides)
He identified three broad categories of users for their resources.

- K-12, who want very specific, focused materials.
- State and academic geospatial community who are looking for information for analysis, such as historical environmental and legal information.
- General public, largely through geo-enabling their photograph collections.

One issue facing state governments is an increasing reliance on project-based as opposed to program-based funding.

Models of Shared Service
Nancy Ritchey, NOAA, NCDC (slides available)
She discussed the activities within the NOAA data centers to supply tiered levels of stewardship service. There are some customers who just want to store their data in a deep archive and make it available to just a few people. At the top tier are customers using lots of data for many different purposes.

The quantity of data they've received at the NCDC has doubled in the past two years. What are the significant issues? Data volume, Data Heterogeneity, and Scientific Data Stewardship. Federation is a standards- and services-based framework which exploits multiple authoritative data sources that are separately administered. Each element can securely access data and metadata throughout the federation. Federation can best be deployed through rules-based distributed data management software.

Bob Downs, CIESIN (slides available)
He discussed the attributes of models for shared academic and government stewardship of digital research resources. These include:

- Organizational commitment for maintaining preservation and providing services.
- Sustainable infrastructure for managing geospatial data.
- Academic and government agencies that have a legacy of preservation and a history of partnerships together.

Brett Abrams, National Archives (slides available) and **John Faundeen, EROS on**

affiliated archives (no slides)
They described how an Affiliated Relationship with the National Archives allows the USGS to physically maintain records that legally transfer to NARA.

Peter Young, Asian Division, Library of Congress (no slides)
Discussed the Library's approaches to bringing in geospatial data and e-science materials. The three Cs of collaborative relationships:

- Coordination
- Cooperation
- Collaboration

"Cooperation" means handling your own money and your own resources. "Collaboration" involves the transfer of funds. Without mingling resources, infrastructure and funds, the rubric of partnerships changes the relationship from coordination to collaboration. You can choose cheaper, faster or better to make that move. The challenges we face today will get more complex and will require collaboration.

Identify Concrete Next Steps, Recommendations to the Library of Congress

- Explore a data rescue program through the ICSU CODATA Task Group on Data at Risk
- Embed data managers in the selection process
- Get primary use data creators to have rescue plans in place
- Build on the NOAA approach to create a template for general guidance for accomplishing all the steps in the OAIS process
- Get involved in the National Digital Stewardship Alliance Content Working Group (http://www.digitalpreservation.gov/ndsa/)
- Prepare an inventory of "definitive data" sets" based on Circular A-16 criteria.
- Work to secure agreement on a core set of appraisal criteria across organizations.
- Support the establishment of a clearinghouse of high-risk datasets.
- Find ways to engage with private sector data providers on the long-term preservation issues.
- Find ways to address the lifecycle of geospatial information.
- Explore the idea of trusted repositories.
- Explore ways for archiving institutions to partner more readily with data providers Partnering with your data providers is vital for long-term preservation.
- Provide coordinated input to the Library of Congress Geospatial Data Preservation Resource Center