

## **NGDA External Advisory Board Meeting**

November 29 - 30, 2006  
San Francisco, California

November 29, 2006

Technical progress:

### **I. Larry Carver presented a report on the technical progress of the grant.**

He displayed the architecture with 5 points of interaction:

- 1) Format registry
- 2) Ingest services and tools
- 3) Access services built on top of federated archives
- 4) Archive data model to facilitate data migration over time
- 5) Export function

He presented a prototype of the NGDA browser using Google Maps as a backdrop. It is expected to go live in the spring of 2007 with 2,000,000 searchable records from NGDA content and the Alexandria Digital Library (ADL) records. It was noted that accuracy of the display hinges on the specificity and accuracy of the coordinates.

It was suggested that icons (thumbnails) be created to give a sense of what kind of resource it is.

A possibility to a challenge for file searching, for instance- Stanford has the Stanford geological survey with very rough Lat/Long- often gave county designation, which makes it hard to search by location

### **II. Nancy Hoebelheinrich discussed Stanford's work on their Repository.**

The first phase has been completed and the repository is ingesting materials.

Phase II includes the addition of online and offline storage. Online is backed up on campus, offline is backed up in SAL locations. *Honeycomb* software manages the backup levels. Highly desired content is stored closer to improve speed performance. Phase II also involves the federation of the SU & UCSB repositories. Stanford is in process of evaluating ADL requirement. Research projects are continually underway and a security audit was carried out last fall.

The process by which content is ingested using conversion software, a Transfer Manifest, and ingest software which creates a universally unique id algorithm.

Present research projects:

Storage subsystems

Workflow management software

Looking to use Fedora as the first access layer

Evaluating software for xml access - *Lucene* is one that Stanford is already using

### **III. Mary Larsgaard presented Collection Development Policy.**

Collection policy: Mary, Julie and Tracey developed a policy that details positions of digital map collections rather than the traditional CDP's for paper maps (Link to: [NGDA Collection Development](#))

[Policy](#) ). This policy describes what the digital format means to the goal of collecting US data. Julie Sweetkind-Singer explains legality of collecting data related to the mission of the university.

#### **IV. Tracey Erwin presented Progress on Contracts and Depository Agreements.**

Stanford requires making depository agreements for data that is not in public domain, thus the reason to create copyrights. In addition, potential depositors need to know what is being done with their data. Legal agreements will give potential depositors a chance to trust NGDA.

It is determined that two distinct agreements are to be put in place:

- 1) A Node-to-Node Agreement (in development) for NGDA participants.
- 2) Node/Content Provider Agreement (Link to: [Draft 3](#), [Draft 3, Exhibit B](#))

- a. The depositor will hold copyright
- b. The archival side will be dedicated to preserving the data using current technology to best of their ability and have good backup systems (an element subsequently stricken as void of measurable details and unnecessarily committal)
- c. Each group has the right to terminate Stanford archiving the data (also subsequently stricken as overly burdensome and conflicting with the concept of an “archive”)

Factors related to development of contracts:

Copyright owned by data provider

NGDA will preserve the data; backup / archival copies made. Library of Congress will get a copy of it.

Not intended to be a dark archive, so be able to provide access.

Reserve right to use data for scholarly purposes

Right to decide not to preserve content.

Discussion Questions:

Q: Why allow people to pull the content out? (Vicky Reich)

A: It is intended to make it hard to pull the data, but want to allow it to be pulled for instance if there is a copyright issue. Also if a business is bought out and new owners change their mind. NGDA intends to make as difficult as possible to pull out to avoid people using the Archive for "free" storage.

Make it dark? A- Difficult to do.

Abby Smith mentions a resource for guidelines about what to do with data that poses a national threat.

Abby also suggests talking with Mary Rasenberger to get language about conditions for withdrawing materials from the Archive. LoC has lots of experience.

What about treating the data as how libraries treat “gifts”? i.e., libraries can do as they wish (more or less). There are some advantages to working collaboratively, such as among the university of California institutions. UCSB approached UCB about VTM (Vegetation Type Mapping) project – Jim Thron was their contact.

Q: What is the policy of what we collect? (Mike Goodchild)

A: Purposely state any US data, to be fuzzy and open to the variety of data available, but such a fuzzy definition can only be done until it becomes unmanageable. Presently, there are many unknowns and NGDA is still looking for nodes, which will help determine needs/interests.

He also asks about if we were asked to house all EROS data—what to do with that? Larry is saying that we want to give data in a library environment... he gives example about MODIS- NOAA will likely house it- but that is a TON of data

M. Goodchild: suggests an approach to prioritize by archiving product of scholarship. Mary refers to the need to emphasize that this project is to preserve US culture of data, but international is open and flexible.

Abby mentions that instructional material is not supported by national collections; they are only a university interest.

Vicky: What if the data is too big for repository? She likes the *clear commons* agreement.

Julie took the agreement they created thus far to ESRI conference and received other librarian input. A traditional librarian concern is the need to copy for archival purpose, which often leads to problems.

Abby Smith discussed Library of Congress role. In 2000, Library of Congress was mandated by congress with a task to identify national network of universities for providing access and maintaining content, however the partnership is not limited to universities. NGDA is currently in the process of partnering with commercial entities. Stewardship is KEY to Library of Congress and they seek groups that will take on the necessary responsibility in preserving data.

Abby Smith also recapped that this project is a cooperative agreement and not a grant. It is a permanent funding, but contingent on what groups like NGDA contribute. The focus is collecting data valued to the nation. A possible role might be for NGDA to educate others about a repository is more than simply backing up data.

Copyright needs to be met head-on. The copyright Office is part of LoC. There are specific tasks needing to be accomplished. Section 108. May not be a report unless all are in agreement and it will address preservation and access, not one or the other.

EROS -- Associate archive of NARA as part of 5 step process to establish it's stewardship, including tour of facilities. Was pilot for a process; NOAA next (from Dianne).

Abby: Library of Congress is looking at these efforts as model. Unclear how/who these other efforts will work together with NDIIPP.

A possible emerging issue is geospatial data by scale. NC State considers this same issue of cultural materials that are geospatial. For example, Katrina mashups that are geographically based.

Many have misconception that data on servers are synonymous for repository. It is suggested that a NGDA role might be to educate others about a repository is more than simply backing up data.

Abby passes our handouts with a graph of partners list with categories of partnerships.

Julie: Asks if there are there any additional suggestions for levels of partnership?

Larry: mentions groups that are producing for near term usefulness, an example is Navteq. He suggests that NGDA might want to snapshot that data on a yearly basis, but it was suggested that snapshots are not very in sync with archival interests. David Rumsey states that Navteq is very interested in preservation; he could make contact with them.

An interesting project is at the University of Wisconsin in which they are building longitudinal views of census data. It might be useful to talk with them about what they're doing.

Financial consideration for potential partners: could be tax benefits for donating, but also other financial incentives. The economic sustainability group is looking at this consideration. Library of Congress is in contact with those in the entertainment industry and are interested in preserving. LoC could then make use of them later (those who have copyright).

An issue arises in terms of what is being archived is intellectual property rights akin to real estate. Maybe parts of a data resource can be used / re-used. It was suggested that it is important to consider common

good in these discussions. It would be important to be able to deal with all kinds of IP owners & what they want.

Creative commons licenses useful here. Also, make arrangements to allow "dumber" data available to the public, e.g., government box policies.

Ray McDowell suggests a way to get data would be to for instance resample 1ft data to 2ft data.- digital data gives the opportunity to fuzz information and allow donors to select particular aspects as viewable.

CERES goal – intends to setup an archive of dataset snapshots...but there is no way to promise since they are not considered stewards. They are mandated to give data to state library, but not sure what state library does with it. CA does not have a geospatial officer (Jon Ellison is acting), but tries to stay involved with NISGIC (state cartographic representatives)- pieced together with partnerships, but no funding.

Q: Is NGDA interested in older data?

A: Depends on format, technology & how it can deal with the resolution -- moving target. Satellite companies might be worth considering, but just some- Landsat 1 = 1972 is not so useful because of resolution. Older aerial photography such as those from 1930s, 1940s is still useful.

AFTER LUNCH:

Q: What are reasons for people to join the network? What is NGDA looking for from nodes? (Julie)

A: M Goodchild suggests data are important, but services might be more important, or MD & additional MD such as how the data are, can be used. He mentions below sites:

[Geospatial One Stop](#)

[ESRI Geography Network](#)

These provide access to services AND data.

Selling points:

Abby: suggests rather than soliciting the content based on the services, but as a place for preservation services which is the value of what is being offered. Such aspect might provide a competitive advantage to those who are depositing.

Diane Garcia: Might also be an advantage to be a source for compiling information such as feedback for the number of hits on the data when in it is stored in the archive, which could provide another competitive advantage to depositors.

Larry: Do we need to think about how to provide means to pay for the archive? ADL uses this for aerial photographs by providing services on top of these as downloads. Approx \$40,000 for 11,000 downloads per month.

David Harris: Another possibility can be using the public library model as well as to work both ends.

David R. 95% of users are 3 commercial firms who use these data, but not for public site. Using Google textual ads that are relevant to the type of document such as 15,000 page views per day, generate 3% = \$3,000. Wikipedia is looking at this as way to fund itself.

Mike: biz model needs to be based on preservation decisions long ago for ADL. Need to think about what present decisions we should be making now.

Abby: Needs to be tied to collection development decisions as well as the economic options for sustaining that. For example, long term preservation of data is a value in itself. Abby reads to the group a document titled "collection of element policy". She mentions much of the geospatial considerations.

David H: Need to go where people have the passion for what they're collecting, see what kind of buzz is generated from that, and then deal with the economic models.

Julie: Interest has grown about the stewardship of data, how to get partners that are reliable one too. How many partners does NGDA need? A. No one is sure how many partners would be best for NGDA to have. It was suggested that they should wait until needs are determined. What would it take to become a node? Provide best practices as means of educating people about what needs to be done and how have we faced the issues that other archives will need to address.

Larry: suggests going after commercial data instead of government data for a shorter time period. Library of Congress is interested in commercial data- they are more likely to be high risk than govt. Abby suggests bringing one group at a time and focus on how they might want to be involved.

Mike: geo-demographics snapshots (yearly basis) would be very useful 30 years from now.

Need to consider static data vs dynamic data.

Governance structure:

Abby: Governance structure will change depending upon the stage of maturation of the network.

Julie: Is it NDGA's interest to include other nodes? Do we really know that the other group is a valid archive repository? Or will the data we chose to include will be an accurate dataset? M Goodchild asks why expand the network because would dilute the brand? Another question was asked- what do you gain by expanding the network? We already have two well endowed universities, so why include Reno, for instance?

Criteria to consider for other nodes: Institution with already strong geospatial expertise and good technical abilities. Possibilities mentioned: ASU, Harvard, and Princeton.

Vicky describes LOCKSS:

LOCKSS (lots of copies keep safe)-peer-peer digital preservation- it is unique by in a large network – gives a report as to the status of integrity of data. CLOCKSS- controlled LOCKSS- private LOCKS network among 12 huge commercial and 7 libraries- founders are board members- everything is done by consensus. Monetary contribution is the same for each node.

Maybe we should have an active oversight board. Not just say yea or nay.

Action item:

Larry wants to get on phone with David and Mike and decide who might be the first to approach for private party nodes. He suggests that it is best of Abby is present and maybe Laura (?). Jack Dangermond at ESRI is a good start.

Mike: we should have a talk with military intelligence, might be able to know more about it through Jack Dangermond. Military intelligence is important data producer, presently their interest could be broad enough for historic preservation. Stanford has a military intelligence connection with somebody named Andrew (Nancy mentioned him).

What is happening internationally?

JISC has a project. Abby knows.

CITAS--- Chinese GIS, look at University of Michigan which has established a node. That would be humanities related project.

European Economic Community?

Open Geospatial Consortium: looking at geo preservation.

MAPAID -- disaster relief group in U.K. Geodata sets behind Katrina. At UCSD?

Australia is very active

Edinburgh - grey project

THURSDAY- November 30, 2006:

Julie: shows the third draft of NGDA definition of stewardship. It states that NGDA is allowed to make copies for backup- for preservation purposes only.

Abby: wants to make clear that NGDA is not a legal entity. She suggests making it explicit as a custodian of NGDA to let the institution take on the responsibility. She thinks the legal agreement Julie displayed is very strong and the library of congress will likely use it as a model. For working with commercial partners, it might be good to make the agreement to look more palpable. Also, language for best practices might need to change.

Julie: Brought up requirements for bringing on network nodes. Mary wants to make sure it is clear that there is one distinctive difference between two repository- ADL requires geographic coordinates- for points or bounding box(for bounding box easier and preferred) while Stanford does not have this requirement. All agree that there is a need to make sure quality metadata accompanies the data.

David R. mentions that Oxford received a large donation of 1930s air photos of northern Africa. – Mashetz? – Might be a good project for finding a group that would georectify the photos through an automated process and NGDA in return will store the data.

Larry suggests NGDA's role to attempting to get money for processing data and not just storing data.

Abby: suggest to take advantage of Microsoft competition with Google and they might georectify African photos. It could possibly be a pilot project, which would be good for Microsoft and good for us.

Larry suggests that this example for something for Mellon Foundation.

Ray asks about National Geographic- many geographers consider it too commercial and are steering clear from them. Question about their digital preservation methods- Larry thinks they might be fairly analog. Allen Carroll is someone contact there...

We know that aerial photos in general are of great interest to many- might need to be a focus to preserve since it is likely that it will be valuable 30-40 years into future.

#### Model ideas

1- Establish something with a group like Navteq in which they would give data to Library of Congress; it would be dark ten years and then goes public through NDGA.

2- Each node has an area of responsibilities. For instance, Stanford could be responsible for roads data and Reno could be responsible for mining and minerals data. Thus specific data goes forward to a specific node.

Julie sees us in the infancy of creating strong search techniques for geospatial data. David Rumsey suspects Google will become a search engine for geospatial data and when this happens, it will probably be very sophisticated.

Larry: what would be possible interferences to NGDA goals? Ray thinks trust to custodians could be potential issue. Larry suggests LoC could be good for neutral territory. All agree there is a need to fully prepare how to approach potential partner groups and therefore, a need to know what we want to collect. David Rumsey stresses Google and ESRI are very different and NGDA needs to know how to interact with them.

Service potentials have been discussed and Larry finds them all interesting-

Access federation layer- analogous to a catalog

Tools area- facilitate a tools library

A good offices environment- to put people and money together  
missed this one

Conferences in the area that might be good to know more about other groups:

Location Intelligence

Where 2.0 – Tim O'Reilly group – May 29-30, 2007

The idea to team up with tool builders was discussed. There would be a need to see if tool builders are interested in the synergy with the libraries. A starting point could be contacting Microsoft or others.

**Invitees:**

Eric Frost, Co-Director, San Diego State University Visualization Center

Mike Goodchild, National Center for Geographic Information and  
Analysis, Department of Geology, UCSB

Diane Garcia, Science Information and Library Services,  
US Geological Survey, Menlo Park

David Harris, CERES Program, California Resources Agency

Joe Langdon, Chief, Science Information and Library Services,  
US Geological Survey, Menlo Park

Ray McDowell, CERES Program, California Resources Agency

Vicky Reich, Director, LOCKSS Program, Stanford University

David Rumsey, Cartography Associates

Abby Smith, NGDA Project Liaison with The Library of Congress

**NGDA Team:**

Larry Carver, UCSB  
Mitch Englander, UCSB  
Mary Larsgaard, UCSB

Julie Sweetkind-Singer, Stanford  
Tracey Erwin, Stanford  
Rachel Gollub, Stanford  
Nancy Hobelheinrich, Stanford  
Mindy Syfert, Stanford