Final Report of the **National Geospatial Digital Archive (NGDA)** and **Federated Archive Cyberinfrastructure Testbed (FACIT)** Projects

December 17, 2009

Prepared by:
Greg Janée, University of California at Santa Barbara
Julie Sweetkind-Singer, Stanford University
Terry Moore, University of Tennessee at Knoxville

We would like to acknowledge these early and significant contributors to the project:

UCSB:
Chris Barteau, Larry Carver, Angus Forbes, James Frew, Mary Larsgaard,
Catherine Masi, Justin Mathena, Adam Ross

Stanford:
John Banning, Tom Cramer, Tracey Erwin, Rachel Gollub, Nancy Hoebelheinrich,
Keith Johnson, Natalie Munn

Tennessee:
Micah Beck

Vanderbilt:
Alan Tackett

1 Background and scope

The National Geospatial Digital Archive (NGDA) project was one of eight initial projects funded by the Library of Congress's National Digital Information Infrastructure and Preservation Program (NDIIPP)[1]. The principal NGDA participants were the University of California at Santa Barbara, Stanford University, and, later, the University of Tennessee at Knoxville and Vanderbilt University. The project commenced in late 2004. A second phase of funding, which we named the Federated Archive Cyberinfrastructure Testbed (FACIT) project, carried the project through to the end of 2009.

The overarching goal of the project, as set out in our first-year roadmap[2], was to answer the question: *How can we preserve geospatial data on a national scale and make it available to future generations?*

Our focus was specifically on *geospatial* data, by which we refer to the wide variety of scientific and government-produced datasets that have a geographic component and that can be viewed as representing a portion of the Earth's surface in some way. This class of information encompasses remote-sensing imagery, aerial photography, maps, data produced by both fixed and mobile geographically-embedded sensors, and data created and processed by GIS (Geographic Information System) tools. We excluded from consideration the broader class of *georeferenced* information, which includes geotagged photographs and textual documents containing geographic references by name. While these other information types merit preservation in many cases, we felt that the geographic aspect of their preservation would be subsumed by the considerations demanded by geospatial data. Nor did we want to be sidetracked by the problems inherent in georeferencing/geotagging.

Our focus was specifically on *long-term* preservation of digital information. By "long-term" we refer to a period of time far exceeding the lifetimes of the applications, platforms, and people involved in the information's creation. Although much of our activity on the project was necessarily concerned with the present and took place in the present—we created archives, ingested data, and otherwise took steps to address the preservation of selected geospatial data *now*—we wanted to make sure that we also identified general design principles, best practices, and, if possible, software architectures that have a chance of carrying archived information through a century or more of unforeseeable technological and social change.

Our focus was on preservation on a *large scale*. We wanted to avoid applying unsustainable and unrepeatable amounts of resources to a few "cherry-picked" collections. Instead, we wanted to define a minimum level (or standard) of preservation that has a high chance of being achieved over the course of a century, without interruption, such that the information remains either as useful as when it was first created or, failing that, *potentially* as useful with some resurrective work.

---

[1] http://www.digitalpreservation.gov/

[2] http://www.alexandria.ucsb.edu/~gjanee/ngda/roadmap.html

And our focus was on making preserved data *available*. We took the position at the outset of the project that, to be both immediately useful and politically and financially sustainable, any archive must make its content available online and in a form that is immediately useful to end users. We were not interested in creating so-called "dark archives."

2 Activities and accomplishments

Our work can be divided into six broad areas: preservation research; archive development; format registry research and development; collection development and ingest (i.e., actual archiving); federation formation; and storage interoperability research. As shown in Figure 1, these work areas were interrelated and informed each other throughout the project.

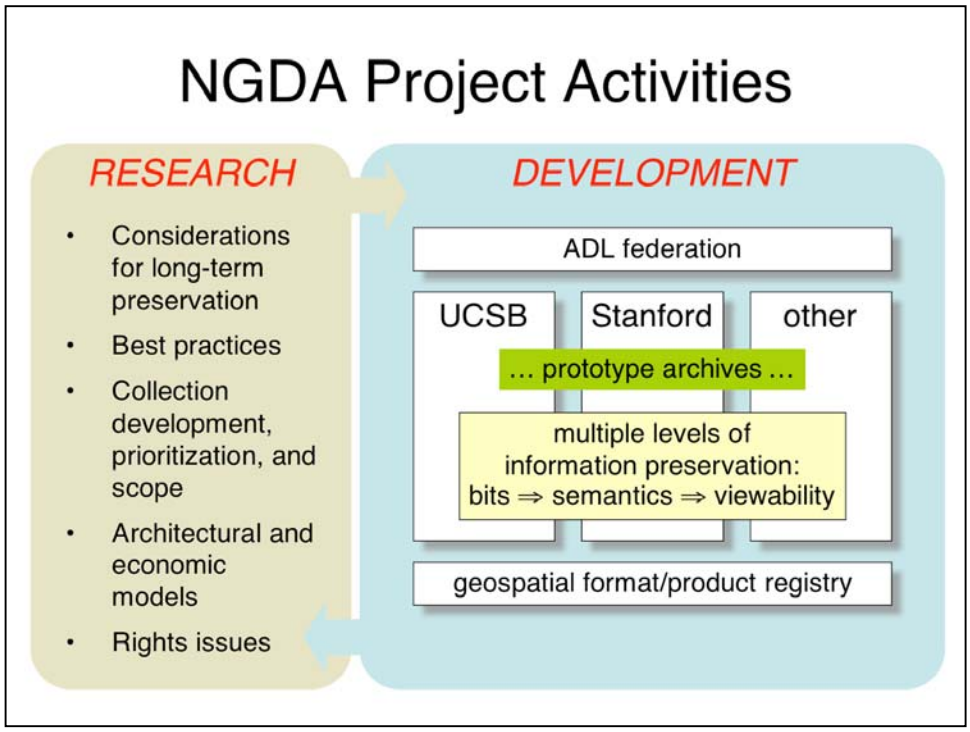

Figure 1. Project activities.

Preservation research. We investigated the characteristics of geospatial data that impact preservation [7] and, beyond that, we investigated the *contexts* in which geospatial data, particularly geospatial science data, is produced [10] [34]. We published best practices regarding the preservation of geospatial data and delineated the many technical, organizational, and scalability challenges that continue to bedevil geospatial data preservation efforts [14].

We researched software architectures supporting long-term preservation. Observing that turnover and handoffs will necessarily occur over the course of a century or more, and will occur at all levels (across storage systems, repository systems, curators, and even across institutions), we developed the principles of preservation relays and handoffs, and of fallback and resurrection [11]. We analyzed where and how these principles

can/do/should manifest themselves in repository architectures [12] and we developed an architecture for the UCSB archive that implements these principles [9].

Late in the project, recognizing commonality between some of the ideas and interfaces in UCSB's archive architecture and the Curation Micro-Service[3] work of the California Digital Library (CDL), we briefly collaborated with CDL, experimenting with data handoffs between NGDA and CDL and exercising the micro-services and helping to refine their specification.

Archive development. UCSB and Stanford each developed archive systems comprising storage components, virtualized storage system interfaces, repository components, ingest and workflow components, and access components. In UCSB's case the archive development was initiated largely from new, relying on the previously-developed Alexandria Digital Library (ADL)[4] as an access mechanism. In Stanford's case the archive took the form of a geospatial extension to the Stanford Digital Repository, an institutional repository already under development. Federation within the project (specifically, cross-archive searchability) was achieved using ADL's federation capabilities.

We experimented with and developed graphical end-user clients for searching over and accessing archived geospatial content [6].

Format registry research and development. In developing our respective archives we immediately encountered the need for a system that could store format specifications, format metadata, and other semantic specifications. The only well-known format registry in existence at the start of the project (GDFR, the Global Digital Format Registry[5]) was not far enough along in its development for us to use, nor in any case did it have the capabilities we required. Thus, as has been true for a number of preservation projects, we developed our own registry. We researched data models and metadata elements for formats and format relationships. We researched geospatial formats specifically, and created registry entries for geospatial formats, dependent formats, and other, related formats. We experimented with "desiccated" representations of format specifications (e.g., GIF screen captures of textual documents) for greater survivability.

Given the immediately apparent burden of building and maintaining a format registry, we explored ideas in collaborative registry building. We created a prototype registry interface that has the features of a wiki (and was in fact implemented using MediaWiki[6]), but is integrated, under curator/librarian control, with an underlying registry and archive system.

Collection development and ingest. We wrote three collection development policies governing the types of materials each institution would collect. The first policy [18] is a general overarching policy discussing the scope of materials to collect, the geographic

---

[3] http://www.cdlib.org/inside/diglib/

[4] http://www.alexandria.ucsb.edu/middleware/

[5] http://www.gdfr.info/

[6] http://www.mediawiki.org/

extent of the materials, and ancillary considerations such as metadata standards. The other two policies are individual policies, one governing UCSB [19] and one governing Stanford's collecting [20]. The two specific policies were written with the knowledge that our collections must align with the research and pedagogical needs of the Universities [4]. Following the policies, we ingested and archived several terabytes of geospatial data. Sources included the California Spatial Information Library (now Cal-Atlas)[7]; the Global Land Cover Facility (GLCF)[8]; the David Rumsey Historical Map Collection[9]; elevation data and high resolution orthoimagery from the National Map[10]; and collections supplied directly by the UCSB and Stanford libraries. Ingested data types included scanned maps, remote-sensing imagery, aerial photography, and GIS datasets.

Federation formation. Recognizing that preservation is too large a problem for any one organization to handle on its own, we formed a federation of geospatial data archives, data providers, curators, and other interested parties. We created the legal framework for the federation in the form of a Content Provider Agreement [21], Content Collection Node Agreement [22], and Content Collection Node Procedure Manual [23]. These documents were vetted by the legal departments at UCSB and Stanford, and both institutions signed them, making UCSB and Stanford the first two members of the federation. Given the legal analysis and level of scrutiny given to these documents, we believe that other institutions will find them acceptable.

Storage interoperability research. After the first phase of the project, we realized that the issue of storage, or more to the point, storage interoperability, is integral to any architecture based on handoffs and one that we needed to address. Therefore, in the second phase of the project—the Federated Archive Cyberinfrastructure Testbed (FACIT) project—we brought in the University of Tennessee at Knoxville and Vanderbilt University to explore the use of Logistical Networking (LN)[11]—a distributed storage technology based on open protocols—in an archival setting. In this phase, in addition to continuing the previously-mentioned project activities, we demonstrated the integration of UCSB's archive with LN-based storage services and storage "depots," and we demonstrated that LN's architecture of limited-duration bit leases could be extended to archival objects and archival time periods. In addition, we demonstrated that an archive can achieve great I/O performance gains by streaming data from multiple storage depots simultaneously, an important consideration for any archive so fortunate as to be faced with the "problem" of being popular [40].

3 Looking forward

At the conclusion of the project the archive implementations at UCSB and Stanford are still very much in development; neither can be considered complete in any sense, and in

[7] http://www.atlas.ca.gov/

[8] http://glcf.umiacs.umd.edu/

[9] http://www.davidrumsey.com/

[10] http://seamless.usgs.gov/

[11] http://loci.cs.utk.edu/

fact both archives are currently undergoing significant "second generation" redesigns. In addition, both archives have a backlog of data in their ingest queues. This was and is to be expected. The goal of the NGDA project was not to "achieve" and be done with preservation via any one-time action, but rather, to establish procedures and system interfaces that have the potential of transcending the inevitable turnover in software components, personnel, and even institutions. We believe we have gone a long way towards accomplishing that goal.

UCSB and Stanford are members of the NGDA federation and are committed to remaining so and, furthermore, are committed to supporting and expanding the federation for the foreseeable future. Stanford will be taking the lead in federation activities and outreach. While the federation is still in its formative stage, we believe that the need for what the federation can provide—a forum where resources can be shared, where gaps can be discovered, where potential partners can be identified, where needs can be aired and met—is great enough that the federation has a good chance of expanding in the future.

The results of our research and additional details of our development activities can be found in the journal articles, conference proceedings, and project reports listed in the bibliography; they will not be repeated here. In the remainder of this report we focus instead on a number of observations and lessons learned that have not been published previously, and that we believe may be of interest to the Library of Congress and other organizations in informing future preservation-related research, development, and funding. These observations are listed in no particular order, except that we've grouped them into: general observations; those that arose more out of one institution's experience than another; and those that arose out of the FACIT work.

4 General observations and lessons learned

4.1 Temporally indeterminate requirements

Long-term preservation generates requirements that are difficult to characterize. We know that certain actions must be taken sometime in the future, but there's usually no *specific* time or point by which these actions must be taken. For example, we might know that the risk of a particular proprietary file format becoming unsupported in the future is all but assured, but when will that occur, and how will we know it? More specifically, by what points should we capture information about that format, or migrate files from that format to another? The only thing we can say with certainty is that, at some point, it will be "too late" to do anything, that is, the cost of maintaining information in that format or migrating information out of that format will have become prohibitively expensive.

A consequence of such "temporally indeterminate" requirements is that it becomes difficult to justify spending resources (money, personnel, computing resources) on preservation actions, particularly when resources are limited and there are other, competing requirements that are far more definite, visible, and immediate, and for which there are clear rewards/benefits and indicators of success.

A specific example of NGDA's: we spent much effort researching the formalization of information semantics. For example, following OAIS[12] principles, we created logical and physical data models for representing and archiving the semantics of archived information; we formalized format relationships and semantics-defining chains, and defined principles of interpretability; and, as mentioned previously, we created and populated a format registry. It would require significant resources and effort to complete this work and fully populate the registry. Would it be worthwhile? And if so, by when? While this work clearly addresses a key concern of preservation, it must be noted that, of the data that NGDA has archived to date, none is facing format obsolescence or provider demise, nor is it likely to in the foreseeable near-term future. NGDA's format registry represents a kind of insurance against preservation risks that have not yet revealed themselves, and perhaps never will.

An argument can be made for delaying preservation work as long as possible in what we might call a "lazy evaluation" approach to preservation. With this approach, much work may be averted. The NGDA format registry might be populated only as needed. In the extreme case, archived data may be deaccessioned before anything need be done to it or on its behalf.

But an opposite argument can be made that preservation actions should be taken when the content is still well-understood, when the original provider still exists and can be consulted if necessary, and when formats are still well-supported. Performing preservation actions on old material is likely to be difficult. Consider, for example, the difference in difficulty between migrating a collection of textual documents to PDF when the source format is currently supported (e.g., Microsoft Word) versus archaic (e.g., VisiWord).

There is currently no rigorous characterization of preservation risks and costs, nor is there a simple calculus we can use to balance them. Instead, it appears that preservation organizations will individually, and on a case-by-case basis, need to analyze potential risks and possible actions, and will need to formulate *ad hoc* arguments to justify spending resources.

In NGDA's case, we note that the emphasis over the course of project shifted from 1) trying to capture complete chains of semantics, to 2) capturing and archiving only the metadata that is uniquely associated with the archived content, and establishing placeholders where metadata may be added in the future as necessary.

4.2 Forms of technical federation

At the beginning of the project we considered how our respective archive-building efforts could possibly and most productively interoperate. We looked at storage sharing at the filesystem level, but even though both of our archive systems incorporated storage virtualization interfaces, we found that the lack of storage management tools, coupled with the buy-in into our respective (and very different) storage systems, created a hurdle

[12] Consultative Committee for Space Data Systems (2002). Reference Model for an Open Archival Information System (OAIS). CCSDS 650.0-B-1, Blue Book (January 2002). http://public.ccsds.org/publications/archive/650x0b1.pdf

too great to cross. We also investigated sharing archival objects, but the differences in our object models precluded that option. We ended up using the Alexandria Digital Library (ADL) to provide a cross-archive search capability.

ADL federation can be characterized as "unidirectional view mapping" in that each participant in the federation maps native representations to a common structure for the specific purpose of supporting a central service. In NGDA's case, each archive maps internal representations (archival objects and metadata) to a common, ADL-specified structure. The mapping is unidirectional because participants maintain their native representations for local purposes. This form of federation is common and can be seen in systems that provide discovery over metadata mapped to Dublin Core, for example, and in systems that provide access to data mapped to delivery protocols such as OPeNDAP[13] and WMS[14].

Sharing/swapping of storage requires a different kind of federation, what we might call "bidirectional representation sharing." In this form, participants both send and receive agreed-upon representations. We believe that this form of federation is likely to be supportable only if the shared representation is coincident with native representations, or if good tools are available for performing bidirectional mappings. The MetaArchive Cooperative[15], another NDIIPP-funded project, was able to implement this form of federation because all members agreed to a particular storage technology (LOCKSS[16]) beforehand. UCSB and Stanford found it too difficult because we shared no representations, at the file level or at the archive object level.

The Library of Congress requested copies of our archived data, and in doing so mandated a particular representation, BagIt[17]. But BagIt was not a native representation for either of our archives, and therefore transfers could be accomplished only with custom programming and *ad hoc* transfer mechanisms. This situation introduces a number of problems: repeated transfers to the Library of Congress will be equally difficult; reverse transfers from the Library of Congress, should they occur in the future (and if we don't plan for that possibility, what's the point?), will require reverse conversions back to native representations; and, most significantly, there are no mechanisms for inventorying Library of Congress objects or synchronizing or updating objects between our archives and the Library of Congress. While the Library of Congress's offsite storage represents an invaluable service for NDIIPP partners, and while the BagIt specification is quite easy to implement, we remain unconvinced that this form of federated storage sharing will be supportable over the long term, or as valuable as it could possibly be, unless and until better tools are developed.

---

[13] http://www.opendap.org/

[14] http://www.opengeospatial.org/standards/wms

[15] http://www.metaarchive.org/

[16] http://www.lockss.org/

[17] http://www.cdlib.org/inside/diglib/bagit/bagitspec.html

We believe that the costs and benefits of technical federation are contingent on whether views are being mapped or representations are being shared; what the representations are; and on the availability of appropriate tools.

4.3 Formats and format registries

The major format registry efforts (GDFR[18], PRONOM[19], UDFR[20], the Library of Congress's own Digital Format Sustainability website[21]) have all adopted what we have termed the "portal" view of registries, which is to say that the registry does not itself serve as a repository for format specifications, but instead refers to specifications and other format-related artifacts on the Web. Facilities for storing simple files in these registries are primitive at best; facilities for storing and curating complex compound objects are nonexistent.

Because formats play such a vital role in long-term preservation, NGDA attempted to address the preservation of formats by adopting an "archive" view of registries. In this approach, formats are themselves represented as archival objects and are stored in an archive alongside the data objects that reference those formats; data–format and format–format relationships tie everything together.

In starting to populate NGDA's own format registry, we anticipated encountering difficulties in identifying definitive specifications and in navigating copyright and proprietary issues. But in practice we found registry population to be unexpectedly difficult for different reasons:

- Format specifications can be quite complex. A specification that is a single document, a PDF file say, poses little difficulty since the PDF file can be stored as a component of the format's archival object, and an "interpretation defined by" relationship can be used to transitively tie the PDF file to another archival object representing the PDF format. But a specification that is a set of linked HTML documents begets the much more difficult tasks of identifying the boundaries of the specification and of archiving what can turn out to be entire websites. Much work has been done in archiving the Web, of course, but the implication here is that a format registry may need to replicate the structure and access mechanisms of a Web archive such as the Internet Archive.
- We speak of formats as being defined by specifications, and they are in a theoretical sense, but in the real world they are ultimately defined by existing software, particularly popular and ubiquitous software. There is a specification for the TIFF format[22], for example, but a TIFF file is ultimately that which well-known applications and services that claim to accept TIFF, in fact accept.

---

[18] http://www.gdfr.info/

[19] http://www.nationalarchives.gov.uk/PRONOM/

[20] http://www.udfr.org/

[21] http://www.digitalpreservation.gov/formats/

[22] http://partners.adobe.com/public/developer/en/tiff/TIFF6.pdf

Examples of the differences between specifications on the one hand, and software behavior and interpretation on the other, are numerous and well-known. In such a situation, archiving format-related software, particularly reusable, open source software libraries (e.g., libtiff[23] in the case of TIFF), can be extremely useful in understanding how the format has actually been interpreted in practice. But this in turn opens up the challenges of archiving software.

- Associated with many formats are maintenance groups that assert jurisdiction over a format. These groups may periodically update specifications and other format-related artifacts on the Web. Format-related software, particularly open source software, is often stored on SourceForge and like systems, where the software may be undergoing active development. There are no established protocols or mechanisms for coordinating these artifacts on the Web and the artifacts stored in the archive, or even of taking snapshots.

Additional findings from our research into geospatial formats and format registries [2]:

- Format specifications can disappear with little warning. For example, we researched Jet Navigation Chart (JNC), Operational Navigation Chart (ONC), and other chart formats. On November 18, 2004, the National Geospatial Intelligence Agency (NGA) announced its intent to remove aeronautical information from public sale and distribution[24]. Some specifications and data related to this action are no longer available to the public via their previously published URLs.
- Geospatial formats are well-supported by popular industry software such as ArcGIS[25] and conversion utility packages such as GDAL[26] and Safe Software[27], but as a rule they are very poorly represented in format registries.
- Containers and parent-child relationships play an important role in geospatial formats, and to properly accommodate the formats, format registries must have good support for these concepts.
- Examples of conforming sample datasets and files can be essential to understanding geospatial file format relationships.
- The lack of contribution/ingest web interfaces (both interactive and programmatic) to format registries discourages format contribution from knowledgeable science, government, and industry parties who may be active in developing and enhancing formats, but are likely too far removed from exposure and participation in format registry contribution efforts.

We conclude that format registries require continued research and development.

---

[23] http://www.libtiff.org/

[24] *Federal Register*, vol. 69, no. 222, pp. 67546–67547. DOCID fr18no04-31.

[25] http://www.esri.com/software/arcgis/

[26] http://www.gdal.org/

[27] http://www.safe.com/

4.4 Data workflow issues

During the life of the cooperative agreement we obtained data from many sources. This gave us the opportunity to reflect on the whole process, from selection and acquisition to management and preservation of the data. What we observed is that this is currently a somewhat chaotic, relatively unstructured process. In particular, the process is far more labor intensive than it might initially appear and it requires direct intervention at many stages.

Data should be considered at risk as long as there is no routine method for managing and processing the content once it has been acquired. The NGDA has acquired data via download from various Web sites, on hard drives, from CD- and DVD-ROMs, and from in-house servers. There is no single point of entry nor is there a routine process for the acquisition of such data. Data are collected on an *ad hoc* basis based upon the scope of the Collection Development Policy and the needs of the faculty and students at the University.

The life cycle of paper materials is well understood and 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 automatic and requires no monitoring by the ordering librarian.

The same cannot be said for digital materials. The life cycle management of digital materials in a library is still in a state of flux, especially for materials that are outside the norm of electronic journals and book content. Stanford's experience is a case in point. As an example, the geospatial librarians decided to acquire high resolution orthographic imagery of the San Francisco Bay area identified on the National Map Seamless Server[28] web site. The acquisition process for obtaining the content was handled through a series of emails directly between the Geospatial Librarian and a contact person at the EROS Data Center[29]. This process took months due to delays in identifying the correct person and then making sure the data we wanted was available. When all was agreed upon, a hard drive was mailed to Sioux Falls, South Dakota. 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. The hard drive was then cataloged in the library's online catalog. Nearly all of the steps in the process were handled by Stanford librarians. In the paper-based world, the vendor would have been known, the Acquisitions department would have created the order, and there would have been no need to checksum or back up the data onto library servers. At this point, Stanford does not have a way to serve these data through a spatial data catalog, so access is managed by the GIS librarian working directly with the patron. Ultimately, the imagery will be ingested into the Stanford Digital Repository with appropriate metadata downloaded from the National Map web site.

---

[28] http://seamless.usgs.gov/

[29] http://eros.usgs.gov/

At this point the digital workflow requires intervention in nearly every step of the process by those acquiring the data and imagery in the first place. The vendors for the content are dispersed. The process to procure the content is often laborious and slow. The need to duplicate the content in a robust manner is immediate and perhaps more challenging due to the size of the datasets. The display, access, and use of the content presents unique challenges for geospatial data as typical library OPACs are not set up to handle the complexities of geospatial data (e.g., multiple datasets 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 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.

4.5 Creating legal agreements

The NGDA team created two legal agreements and a manual over the course of the project. The Content Provider Agreement governs the deposit of copyrighted material. The Content Collecting Node Agreement and accompanying Procedure Manual spell out how the partners will work together as a team to acquire, house, and provide long term preservation and access to materials.

Although both partner institutions had the same overall goals, this process was far more complex and time intensive than we had anticipated. The reasons for this were many. The hindrances that were encountered included a lack of formal legal training for the committee members, difficulty in consistent access to legal counsel, and different collection priorities for each collecting node.

A decision was made early on that Stanford's General Counsel would take the lead on helping the NGDA craft the legal agreements. The General Counsel for UCSB would be brought in at a later date to review the work done and ensure that the agreements met the needs of that campus. Stanford's General Counsel suggested that the agreements be crafted in layperson's language and capture what the committee thought were critical issues to address. For example, in obtaining copyrighted or licensed material, if the committee decided to include a section on rights and responsibilities of the content provider and the content collector, the group would lay out in detail those rights and responsibilities as best they could. Once this preliminary document was created, the General Counsel would then transform the wording into legally acceptable language. This latitude left the group with the ability to consider a wide variety of scenarios that could arise.

Some legal knowledge. While a lack of legal knowledge had implications, the reverse was also true. One committee member had some familiarity with the law and often questioned whether certain concepts would be legally viable and, if so, how to ensure that we were not introducing any ambiguity with our word choice.

Access to Legal Counsel. We found it difficult to schedule face to face time with our legal counsel to review sections of the agreement. While we did have a number of such

meetings, in some cases we had moved on to new sections while awaiting feedback on earlier work. We eventually began to present larger sections of work rather than waiting for input on smaller sections.

Different collection priorities. The first agreement completed was the Content Provider Agreement. This agreement governed the deposit of copyrighted materials into the archive. Stanford had a verbal agreement to archive the map images of the David Rumsey Historical Map Collection. As such, we needed a written agreement to formalize the arrangement. Stanford also anticipated the possibility of acquiring other copyrighted collections. UCSB, by contrast, had chosen to focus on public domain data such as the holdings of the California Spatial Information Library. With this focus, it was difficult for UCSB to see the value in having an agreement for copyrighted material.

5 UCSB observations and lessons learned

5.1 Filesystem-level repository interoperability

As part of its research into architectures supporting long-term preservation, NGDA described the relay principle of preservation: the idea that preservation resembles a relay across storage and repository systems and across curators and institutions, and that the ability to handoff archived content from one repository to the next or from one institution to the next is therefore as or more important than the preservation actions taken within a system or institution.

Entirely unexpectedly, UCSB encountered a handoff situation of its own near the end of the NGDA project. In the middle of redesigning our archive and transitioning from our first-generation repository system to a second-generation system, we were simultaneously faced with an unexpected cut in project duration (effectively, a budget cut as well) from the Library of Congress; the retirement of several senior project personnel; the departure of key technical personnel; and an unprecedented University-wide budget crisis that precipitated a total hiring freeze. Remaining library staff taking over the archive and coming on to the project were committed to preserving the data gathered to date, but they had no direct experience with the older NGDA repository system, with the redesign in progress, or with the archived data itself.

Fortunately, our archive system's data model was designed to store all information as XML manifests and simple files in a filesystem according to a standard structure. In transitioning the data to the new archive system, the new NGDA staff could entirely ignore the old system and focus solely on filesystem content. Only minor scripting was required to massage XML manifests and files and directories as required by the new archive system's ingest facilities. Our later work with the CDL Curation Micro-Service specifications, which are also filesystem-based, echoed this experience.

Had it been necessary to resurrect the old archive server, the handoff task would have been much more difficult. The first-generation server was not running, and getting it to a running state again may have been impossible due to software dependencies. In general, the dependencies of any piece of software on operating systems, applications, programming languages, libraries and other third-party software, patches, etc., can quickly break (due to changes, lack of support, etc.) and then compound, so that it

becomes either impossible to return a piece of software to a running state, or at least prohibitively expensive.

Relational databases pose similar difficulties, and it is for this reason that UCSB assiduously avoided any use of databases in its archive data model. Databases can be saved to and restored from snapshots only, but snapshots are difficult to preserve because their formats are proprietary and because they are highly dependent on database versions, database plugins, and other contextual specifics. In addition, even if one is able to restore a database, if the goal is to simply extract (meta)data into a new system, there is then the hurdle of having to understand and reverse-engineer application-specific schemas.

We believe that the idea of filesystem-level repository interoperability is promising, but it requires further research and development. To gain experience, it also requires greater adoption by the preservation community, and by the major repository systems in particular.

Repository interoperability is not a new subject, of course, but we note that most efforts to date have been aimed at interoperability between running systems, a focus that is echoed in OAIS as well. As seen above, the situation when a handoff must be made from an older, defunct archive system is one that has and will be encountered. Furthermore, we believe that filesystem-level transfer of content is likely to remain a standard means for representing and transferring data. We note that the Library of Congress's AIHT[30] experiment began with exactly such a data transfer, and that the Library of Congress is today transferring and storing data from NDIIPP projects represented as BagIt packages, i.e., as files in filesystems[31].

Ultimately, we believe that long-term preservation would be best assured by the creation and adoption of a *standard* for filesystem-based structure and layout of archival objects. Although NGDA developed its own specifications[32] along these lines, we believe that the CDL Micro-Service specifications, Dflat[33] in particular, provide a more robust foundation for these concepts. It remains to be seen whether repository systems such as Fedora[34] or DSpace[35] could be modified to operate natively on Dflat objects and, if so, if they would be willing to.

5.2 Data portability

At the beginning of the project UCSB purchased an Archivas ArC storage cluster. Similar to many other products on the market, it redundantly stored information in a

---

[30] http://www.dlib.org/dlib/december05/shirky/12shirky.html

[31] http://www.digitalpreservation.gov/news/newsletter/200908.pdf

[32] http://www.alexandria.ucsb.edu/~gjanee/ngda/data-model/

[33] http://www.cdlib.org/inside/diglib/dflat/dflatspec.pdf

[34] http://www.fedora-commons.org/

[35] http://www.dspace.org/

RAID-like configuration, but the Archivas unit had a number of additional desirable features that caused UCSB to choose it:

- Its architecture was based on low-cost, generic boxes running an open source operating system (Linux).
- It was fairly hardware agnostic, and could even operate on dissimilar hardware, such as different hardware types that might be acquired over time.
- It was easily extensible.
- It could associate and store arbitrary, additional metadata with objects.
- It continuously and actively validated files, and would automatically correct any detected corruption.
- It could enforce policies related to stored content.
- It assiduously avoided proprietary lock-in by supporting multiple, standard interfaces for storing and accessing content (NFS, WebDAV[36], etc.).

As such, the unit appeared to be an ideal platform to support NGDA's long-term bit preservation. The only question was one of performance: given the tradeoffs in engineering the unit, I/O throughput was not the primary design consideration, and there were questions as to whether such a storage cluster could directly host an active archive containing extremely large objects.

By the end of the project, Archivas no longer existed as a company (it had been acquired by Hitachi), and the ArC product was no longer manufactured or supported. Furthermore, UCSB's unit had developed some maintenance issues and so, with some urgency, the data was transferred to another storage system. And in this way the question of performance was rendered entirely moot in a span of just a few years.

It has been widely observed that storage is a commodity characterized by a large number of competitors offering functionally equivalent products and by rapid turnover in competitors and products, but what is clear from NGDA's experience is that it is a commodity at all levels: not only at the levels of disk drives and boxes and racks, but at the level of entire storage systems. In such an environment, information portability must be the primary consideration. In Archivas' case, that portability was delivered in the form of the public, standard interfaces that the ArC unit supported that allowed NGDA content to be transferred off. Even failing those interfaces, data portability was achieved by the fact that the content was stored straightforwardly in a Unix filesystem.

5.3 Validation

Late in the project we discovered that some of our archival objects created early in the project contained some structural and metadata errors. The source of the problem was a combination of classic snafus: the specification for our data model changed at one point, the archive server that implemented the data model had some bugs, and we failed to

[36] http://www.webdav.org/

reprocess some objects. The errors were minor, but the episode nevertheless illustrated the importance of validation to detect such errors.

We believe that independent validation tools would be most valuable in these situations. An independent validation tool should be created for every persistent representation: for data structures, file formats, filesystem and directory structures, etc. The validation tool should check for all constraints mentioned in the relevant specifications. Thus, for example, a validator for a Fedora repository object should check, not just for the well-formedness and validity of the object's FOXML[37] manifest, but for the validity of the object's datastreams and relationships as well. (We in fact created such a tool.)

We are not recommending that independent validation tools supplant the validation that occurs during normal processing. We expect that an archive system will still validate inputs at ingest time. But we see three major benefits to expressing validation as an independent tool as well:

- The tool can be run at any time, whereas validation performed at ingest time is performed only once: at ingest.
- If the specification changes, the tool can be changed and rerun.
- If the validation tool itself has bugs, the tool can be corrected and then simply run again, whereas if an archive server's ingest validation is buggy, it is typically not possible to simply reingest the archive content.

Furthermore, our recommendation is that validation tools be developed in parallel with data structures and file formats. Test-driven Development[38], an outgrowth from the Extreme Programming software development methodology, advocates that test suites/harnesses be developed in parallel (and even in advance of) mainline software. Our recommendation is similar, but applied to persistent structures as opposed to software.

We emphasize that the need for validation is not limited to "complex" objects and specifications. While working with the CDL Micro-Service specifications we examined the need for validation and the role that validation plays. These specifications are conceptually some of the simplest in this domain, yet in developing a validation tool for just the DFlat[39], Checkm[40], and Namaste[41] specifications, we needed to print more than 60 unique error messages.

---

[37] http://www.fedora-commons.org/download/2.0/userdocs/digitalobjects/introFOXML.html

[38] http://en.wikipedia.org/wiki/Test-driven_development

[39] http://www.cdlib.org/inside/diglib/dflat/dflatspec.pdf

[40] http://www.cdlib.org/inside/diglib/checkm/checkmspec.html

[41] http://www.cdlib.org/inside/diglib/namaste/namastespec.html

6 Stanford observations and lessons learned

6.1 Staffing challenges

The NGDA project was originally conceived as a three year award, then extended an additional two years with the no-cost extension. During this time staff changes at both Stanford and UCSB presented challenges to the ongoing production of work.

Almost immediately after the agreement began, Stanford Libraries reorganized the technical side of the organization and created the Digital Library Systems and Services Group. This reorganization and creation of the Stanford Digital Repository team took months to complete. The realignment of staff created a robust, targeted organization, but delayed the start of technical work on the repository. Throughout the agreement, there were numerous personnel transitions both in positions paid by the grant and in positions provided by cost-share. The economic downturn in Fall 2008 had a large impact on the project with the loss of cost-share staff due to layoffs at Stanford, including key project members who had expertise on data transfers and metadata standards and implementation.

6.2 Data transfer experience

We transferred geospatial data from the Stanford Digital Repository (SDR) over two phases, using two different approaches and different versions of the BagIt protocol [3].

In phase I, the BagIt encoding process was integrated within reconstruction, which necessitated making difficult, internal code adjustments, thus complicating the process. (Reconstruction is a predecessor process to BagIt that involves pulling files from storage and reassembling them in their original deposit form, checksumming, and comparing the checksum generated to the original checksum.)

In phase II, the BagIt protocol had been improved based on feedback from phase I. The changes made the BagIt protocol much more modular, which enabled the BagIt encoding process to be decoupled from the internal SDR reconstruction process, and which made adjustments to the workflow much easier. Improved BagIt tools for validation, including the "verifyvalid" operation which checks for fixity failures in the bag, made the BagIt transfer process easy. Performing the "verifyvalid" check when the bag is created on the Stanford side, and re-running the check on the Library of Congress side once the bag has been pulled from the Stanford server using rsync, indicates a successful transfer.

This project has provided an excellent opportunity for the SDR to exercise a complete ingest, reconstruction, and transfer workflow. Improvements to BagIt between phase I and phase II made the transfer process itself completely seamless. Separation of BagIt from reconstruction enabled us to see the SDR workflow clearly and identify improvements to implement in the next version of the repository.

6.3 Stanford Digital Repository

Due to its length, this section is available as an ancillary report [1].

7 FACIT observations and lessons learned

7.1 Robust replication management

One major lesson learned during FACIT concerned the recognition that more robust and flexible replication management was required for decades-long preservation of digital objects as physical encodings. These ideas arose through conversations with the LOCKSS team and a review of the papers associated with their work. They make a persuasive case that we are currently profoundly ignorant with respect to what is needed for truly long-term bit preservation. They show that there is an alarming mismatch between known threats to long term preservation, which they catalog, and the design philosophy of many current storage systems. When a simple but reasonable model of long-term storage failures is taken into account, what we know about the physical preservation of digital objects can be summarized by three simple propositions:

1. Make copies: The more copies that are made, the better the chance that the data will survive.
2. Decorrelate copies: The higher the number of correlations between existing copies of a given data object (e.g., same geographic location, same system, same *type* of system, same administrator, etc.), the higher the chance of a shared failure that will destroy them all. Avoid single points of failure by decorrelating copies.
3. Audit and maintain copies regularly: The more frequently copies of a given data object are audited, the more quickly latent failures will be detected and repaired, which in turn lowers the potential for correlated failures with other copies.

The problem of replica creation, decorrelation, and maintenance now confronting the long-term preservation community calls for interorganizational cooperation of unprecedented breadth and form. Decorrelated replication, in particular, represents a serious problem for any stand-alone organization because it is expensive, especially when the amount of data is large and continuously growing as it is with remote-sensing imagery, for example. If one wants to have replicas in two widely separated locations, facilities at both locations have to be acquired, provisioned, and connected. An obvious way to address this problem is by partnering with others and sharing resources; absent such partnering, it is not clear that any scalable alternative is available. Moreover, some forms of decorrelation, e.g., having copies under different administrations, cannot be achieved at all without partnering with other organizations. But current storage infrastructure is not generally designed for the kind of cross-organizational deployment, scalability, and resource sharing that is necessary to enable a community to carry out such a cooperative and highly distributed replication strategy. These difficulties are only made worse when questions of replica auditing, maintenance, and low-latency access are taken into account.

7.1.1 Problem for FACIT in replication management

This requirement for decorrelated data replication highlights the value of FACIT's low-level storage architecture, which is based on Logistical Networking (LN) technology and which is explicitly designed to scale across such organizational boundaries and platform heterogeneity. But this newfound emphasis on decorrelated replica management also

exposed a weakness in the Logistical Distribution Network (LoDN)[42], the service that FACIT used for this purpose. Since the storage substrate that FACIT builds on is essentially passive, it requires a client that can direct its activities on the basis of user goals. This layer, which we think of as a kind of routing service, must take responsibility for decisions regarding where data should be stored and moved (e.g., during replication), and must apply services to ensure that the result is reliable and that the data is correct. As the work of the LOCKKS group and others show, the problem of choosing a set of physical locations at which an object will be stored in order to maintain access and to preserve that access over time is a difficult one.

In FACIT, the module within LoDN known as the "dispatcher" manages automated data movement and replication. The version of the dispatcher that was available when we started FACIT worked by very simple rules; it sought to maintain a copy of the data at each of a number of "sites," where a site is defined by a specific list of depots, and the LoDN dispatcher put the decision in the hands of an end user, thus relying on manual direction. But storage used to preserve data over the long term may have very different characteristics and properties from buffers used to distribute it conveniently and quickly. And these properties are not stable over time, so they must be monitored and data placement decisions reevaluated both periodically and when exceptional events occur. It became clear that it would be important to make progress in automating the placement of data and guaranteeing properties of stored data with some degree of confidence.

7.1.2 Solution for flexible replication management

Addressing this limitation of LoDN required us to redesign the "data dispatcher mechanism" for more flexible and automated use. LoDN is implemented in two parts: a back end server that is accessed programmatically by clients using the LoDN library, and a Web-based front end that exposes an interface similar to a typical FTP client interface. In addition to the conventional file management operations, the LoDN Web interface includes high-performance parallel upload and download tools that run on the client using the Java Web Start framework. The necessary improvements to the robustness, flexibility and programmability of LoDN's data dispatcher required a complete reimplementation of its back end services. The new version of LoDN is already being tested in a prototype on REDDnet[43], will be in beta testing and available to the NDIIPP community in the early Spring 2010, and will be in general release by the summer.

It should be noted that another limitation of LoDN we uncovered during the FACIT project is that the dispatcher does not probe the contents of the replicas it manages, and so cannot address the problem of whether they are identical and, if not, then which data should be treated as authoritative. Thus, it is quite possible that after the dispatcher has managed data for a long time, the copies may be inconsistent. To the extent that the metadata maintained by the dispatcher includes checksums, there is a problem in the replication and management of the metadata. But even when checksums can be used to identify corrupted data, this cannot help in retrieving the original data if all copies are

[42] https://ln.eecs.utk.edu/

[43] http://www.reddnet.org/

corrupt. Thus, identification of corrupt data and creation of new, valid replicas must go on constantly.

The LOCKSS project has developed technologies for identifying authoritative copies of distributed data, and plans for the future evolution of FACIT incorporate the use of the LOCKSS anti-entropy protocol into the LoDN dispatcher.

7.2 Networking monitoring

Another general area where our experience working with FACIT suggested the need for innovation was in the area of network monitoring in order to make distributed replica networks easier to manage. As many parts of the NDIIPP community know, getting high-performance networks to work well on large, wide area data transfers is no simple matter. This is particularly important for FACIT, which relies on LoDN's data dispatcher to do automated data movement based on user-determined policy. Improving this situation, for both manual and automated use, requires a much closer integration between FACIT's storage infrastructure technology and network monitoring and management tools. Consequently, we have undertaken an effort to make that change.

Through collaboration between the FACIT team and others in the REDDnet community, we will bring out the Data Logistics Toolkit (DLT), which we believe will help the entire NDIIPP community on this front. The DLT combines the software technologies that FACIT uses for shared storage with tools for network monitoring (e.g., perfSONAR[44]) and enhanced control signaling (e.g., Phoebus[45]) for more efficient use of wide area links and dynamically allocated circuits. Integration with perfSONAR will enable the DLT data movement/replication tools to adapt dynamically and automatically to network topology and conditions; these tools can then use Phoebus to optimize the use of network resources. This integration will enable distributed storage infrastructures, such as FACIT's, to automatically tune protocol and network settings and to dynamically rebalance active data flows or seek alternate paths to maximize throughput, allowing overlay multicast for massive data flows to be scheduled along highly efficient paths.

Taken individually, the value of the components to be included in the DLT has been demonstrated through a variety of research and infrastructure projects. But to achieve dramatic improvements in production environments, such as those involved in various NDIIPP projects, they must be made to work together seamlessly and to leverage each other's capabilities. They must also be thoroughly tested and hardened for production use, and must be packaged for easy delivery, installation, and configuration.

8 Bibliography

Filenames in angle brackets (<>'s) below refer to files submitted to the Library of Congress at the end of the grant.

[44] http://www.perfsonar.net/

[45] http://damsl.cis.udel.edu/projects/phoebus/

8.1 Ancillary reports

[1] Cramer, Tom. The Stanford Digital Repository: Lessons Learned — A Report for NDIIPP. December 2009. http://www.ngda.org/docs/SDRLessonsInPreservation.pdf. <SDRLessonsInPreservation.pdf>

[2] Munn, Natalie. Report to National Geospatial Digital Archive Regarding Geospatial Data Treatment in Data Format Registry Efforts. December 15, 2009. http://www.ngda.org/docs/CINGDAfindingspub121509.pdf. <CINGDAfindingspub121509.pdf>

[3] Pande, Alpana, and Kott, Katherine. NGDA Phase II Data Transfer Report. http://www.ngda.org/docs/PhaseIIDataXfer.pdf. <PhaseIIDataXfer.pdf>

8.2 Publications

[4] Erwin, Tracey, Sweetkind-Singer, Julie, and Larsgaard, Mary L. (2009). The National Geospatial Digital Archive—Collection Development: Lessons Learned. *Library Trends*, vol. 57, no. 3, Winter. doi:10.1353/lib.0.0049. <Pub_LT_Erwin_09.pdf>

[5] Erwin, Tracey and Sweetkind-Singer, Julie (2009). The National Geospatial Digital Archive: A Collaborative Project to Archive Geospatial Data. *Journal of Map and Geography Libraries* (in press). <Pub_JMGL_Erwin_09.pdf>

[6] Forbes, Angus, and Janée, Greg (2007). Visually Browsing Georeferenced Digital Libraries. *Geoinformatics 2007 Conference* (San Diego, California; May 17–18, 2007). http://www.alexandria.ucsb.edu/~gjanee/archive/2007/geoinformatics-abstract.html. <Pub_GEOI_Forbes_07.pdf>

[7] Hoebelheinrich, Nancy, and Banning, John (2008). An Investigation into Metadata for Long-lived Geospatial Data Formats. NGDA technical report. http://www.ngda.org/reports/InvestigateGeoDataFinal_v2.pdf. <Pub_LOC_Hoebelheinrich_08.pdf>

[8] Janée, Greg, and Frew, James (2005). A Hybrid Declarative/Procedural Metadata Mapping Language Based on Python. *Research and Advanced Technology for Digital Libraries: Proceedings of the 9th European Conference (ECDL)* (Vienna, Austria; September 18–23, 2005): 302–313. doi:10.1007/11551362_27. <Pub_ECDL_Janee_05.pdf>

[9] Janée, Greg, Mathena, Justin, and Frew, James (2008). A Data Model and Architecture for Long-term Preservation. *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)* (Pittsburgh, Pennsylvania; June 16–20, 2008): 134–144. doi:10.1145/1378889.1378912. <Pub_JCDL_Janee_08.pdf>

[10] Janée, Greg, and Frew, James (2008). Preserving the Context of Science Data. *Eos, Transactions, American Geophysical Union*, vol. 89, no. 53, Fall Meeting Supplement, abstract U13D-06. http://www.alexandria.ucsb.edu/~gjanee/archive/2008/agu-abstract.html. <Pub_AGU_Janee_08.pdf>

[11] Janée, Greg (2009). Preserving Geospatial Data: The National Geospatial Digital Archive's Approach. *Archiving 2009: Final Program and Proceedings* (Arlington, Virginia; May 4–7, 2009): 25–29. http://www.alexandria.ucsb.edu/~gjanee/archive/2009/archiving-2009-paper.pdf. <Pub_Arch09_Janee_09.pdf>

[12] Janée, Greg, Frew, James, and Moore, Terry (2009). Relay-supporting Archives: Requirements and Progress. *International Journal of Digital Curation*, vol. 4, no. 1. http://www.ijdc.net/index.php/ijdc/article/view/102. <Pub_IJDC_Janee_09.pdf>

[13] Janée, Greg (2009). Digital Curation. *Encyclopedia of Database Systems* (Ling Liu and M. Tamer Özsu, eds.) (Springer, 2009): 816–817. doi:10.1007/978-0-387-39940-9_879. <Pub_EDS_Janee_09.pdf>

[14] McGarva, Guy, Morris, Steve, and Janée, Greg (2009). Preserving Geospatial Data. *Digital Preservation Coalition (DPC) Technology Watch Series* report 09-01. http://www.dpconline.org/docs/reports/dpctw09-01.pdf. <Pub_DPC_McGarva_09.pdf>

[15] Sweetkind-Singer, Julie, Larsgaard, Mary Lynette, and Erwin, Tracey (2006). Digital Preservation of Geospatial Data. *Library Trends*, vol. 55, no. 2, Fall. doi:10.1353/lib.2006.0065. <Pub_LT_Sweetkind_06.pdf>

[16] Sweetkind-Singer, Julie (2009). Uncharted Territory: Building a Network for the Archiving of Geospatial Images and Data. *Against the Grain*, vol. 21, no. 2, April. http://lib.stanford.edu/files/ATG_Sweetkind_032309.pdf. <Pub_Sweetkind_ATG_09.pdf>

[17] Sweetkind-Singer, Julie, Erwin, Tracey, and Larsgaard, Mary Lynette (2009). Legal Agreements Governing Archiving Partnerships: The NGDA Approach. *Archiving 2009: Final Program and Proceedings* (Arlington, Virginia; May 4–7, 2009): 11–15. http://lib.stanford.edu/files/Sweetkind_Archiving2009.pdf. <Pub_Sweetkind_Arch2009 _09.pdf>

8.3 Federation documents

[18] Collection Development Policy for the National Geospatial Digital Archive. November 1, 2006. http://www.ngda.org/docs/NGDA_Collection_Development_Policy.pdf. <NGDA_Collection_Development_Policy.pdf>

[19] Collection Development Policy for the National Geospatial Digital Archive: Node, Map and Imagery Laboratory (MIL), Davidson Library, University of California, Santa Barbara. May 2007. http://www.ngda.org/docs/UCSB_CDP_05_07.pdf. <UCSB_CDP_05_07.pdf>

[20] Collection Development Policy for the National Geospatial Digital Archive: Node, Branner Earth Sciences Library and Map Collections, Stanford University. May 2007. http://www.ngda.org/docs/SU_CDP_5_07.pdf. <SU_CDP_5_07.pdf>

[21] NGDA Content Provider Agreement, Stanford University. http://www.ngda.org/docs/Stanford_NGDA_Contentprovider_102307final-1.pdf. <Stanford_NGDA_Contentprovider_102307final-1.pdf>

[22] NGDA Content Collection Node Agreement. March 19, 2009. http://www.ngda.org/docs/NGDA_NodeAgreement_March2009.pdf. <NGDA_NodeAgreement_March2009.pdf>

[23] NGDA Content Collection Node Procedure Manual. March 19, 2009. http://www.ngda.org/docs/NGDA_ProcedureManual_March2009.pdf. <NGDA_ProcedureManual_March2009.pdf>

8.4 Presentations

[24] Erwin, Tracey. Collection Development Policies for Geospatial Data. *ESRI Education Users Conference* (San Diego, California; June 16–19, 2007). <Pres_Erwin_ESRI_07.pdf>

[25] Erwin, Tracey, and Sweetkind-Singer, Julie. Collecting and Preserving Geospatial Content: The NGDA Experience. *ESRI Users Conference* (San Diego, California; July 13–17, 2009). <Pres_Erwin_ESRI_09.pdf>

[26] Frew, James. Relay-supporting Archives: Requirements and Progress. *4th International Digital Curation Conference* (Edinburgh, Scotland; December 1–3, 2008). http://www.alexandria.ucsb.edu/~gjanee/archive/2008/frew-idcc-presentation.pdf. <Pres_Frew_IDCC_08.pdf>

[27] Janée, Greg. National Geospatial Digital Archive. *UCSB Institute for Computational Earth System Science (ICESS) seminar*. May 31, 2005. http://www.alexandria.ucsb.edu/~gjanee/archive/2005/icess-seminar.pdf. <Pres_Janee_ICESS_05.pdf>

[28] Janée, Greg. National Geospatial Digital Archive. *National Digital Information Infrastructure and Preservation Program (NDIIPP) partner meeting*. July 13, 2005. http://www.alexandria.ucsb.edu/~gjanee/archive/2005/ndiipp-partners-meeting.pdf. <Pres_Janee_NDIIPP_05.pdf>

[29] Janée, Greg. A Hybrid Declarative/Procedural Metadata Mapping Language Based on Python. *Research and Advanced Technology for Digital Libraries: 9th European Conference (ECDL)* (Vienna, Austria; September 18–23, 2005). http://www.alexandria.ucsb.edu/~gjanee/archive/2005/ecdl.pdf. <Pres_Janee_ECDL_05.pdf>

[30] Janée, Greg. National Geospatial Digital Archive. *Digital Curation Centre seminar* (University of Edinburgh; September 27, 2005). http://www.alexandria.ucsb.edu/~gjanee/archive/2005/dcc.pdf. <Pres_Janee_DCC_05.pdf>

[31] Janée, Greg. National Geospatial Digital Archive. *Workshop on Maintaining Long-term Access to Geospatial Data* (National e-Science Centre, Edinburgh, Scotland; October 27, 2006). http://www.alexandria.ucsb.edu/~gjanee/archive/2006/ngda.pdf. <Pres_Janee_NESC_06.pdf>

[32] Janée, Greg. Long-term Preservation as a Relay. *National Digital Information Infrastructure and Preservation Program (NDIIPP) partner meeting*. June 27, 2007. http://www.alexandria.ucsb.edu/~gjanee/archive/2007/handoff-architecture.pdf. <Pres_Janee_NDIIPP_07.pdf>

[33] Janée, Greg. A Data Model and Architecture for Long-term Preservation. *8th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)* (Pittsburgh, Pennsylvania; June 16–20, 2008). http://www.alexandria.ucsb.edu/~gjanee/archive/2008/jcdl-presentation.pdf. <Pres_Janee_JCDL_08.pdf>

[34] Janée, Greg. Preserving the Context of Science Data. *AGU Fall Meeting* (San Francisco, California; December 15, 2008). http://www.alexandria.ucsb.edu/~gjanee/archive/2008/agu.pdf. <Pres_Janee_AGU_08.pdf>

[35] Janée, Greg. Preserving Geospatial Data: The National Geospatial Digital Archive's Approach. *Archiving 2009* (Arlington, Virginia; May 4–7, 2009). http://www.alexandria.ucsb.edu/~gjanee/archive/2009/archiving-2009-presentation.pdf. <Pres_Janee_Arch09_09.pdf>

[36] Larsgaard, Mary. The National Geospatial Digital Archive (NGDA). *ALA MAGERT Map Collection Management Discussion Group* (Washington, D.C.; June 21–27, 2007). <Pres_Larsgaard_MAGERT_07.pdf>

[37] Sweetkind-Singer, Julie. NDIIPP: The National Digital Information Infrastructure and Preservation Program. *Association of Jewish Libraries* (Oakland, CA; June 19–21, 2005). <Pres_Sweetkind_AJL_05.pdf>

[38] Sweetkind-Singer, Julie, and Banning, John. Long-term Archiving of Geospatial Data: the NGDA Project. *ESRI Users Conference* (San Diego, CA; Aug 7–11, 2006). <Pres_Sweetkind_ESRI_06.pdf>

[39] Sweetkind-Singer, Julie. Digital Collections Decisions: an NGDA Perspective. *National Digital Information Infrastructure and Preservation Program (NDIIPP) Symposium* (Washington, D.C.; June 25–27, 2007). <Pres_Sweetkind_NDIIPP_07.pdf>

8.5 Screencasts

[40] Moore, Terry. FACIT and Globetrotter demo. *Fall 2008 Internet2 Member Meeting* (New Orleans, Louisiana; October 13–16, 2008). http://www.ngda.org/2009/video/globetrotter_demo/

[41] Moore, Terry. LoDN Download Tutorial. 2008. http://www.ngda.org/2009/video/lodn_tut/