2. Abstract

Geospatial information has played an important role in the history of the United States. From the first colonial maps to the satellite imagery of the 21st century, cartographic information has helped define and frame our view of the United States. The University of California, Santa Barbara (UCSB) and Stanford University intend to collect materials across this broad spectrum of cartographic history. Early maps in digital format of the United States, Washington, and Georgia will be preserved along with the most up-to-date imagery from the Landsat and MODIS satellites. Preserving such a broad range of digital geospatial materials in distributed archives will form the basis for structured access to the content and trend analysis vital for the study of history, environmental policy, urban and population studies, census construction and analysis, redistricting at federal and state levels, government aid determinations, land use policy, etc.

UCSB and Stanford (the Partners) propose jointly to lead the formation of the National Geospatial Federated Digital Repository, a collecting network for the archiving of geospatial images and data. This network, led by these two major institutions, will recruit additional partners and content providers to assure persistent access to a critical mass of digital information that might otherwise perish.

Objectives:

• Create a new national federated network committed to archiving geospatial imagery and data.

• Investigate the proper and optimal roles of such a federated archive, with consideration of distant (dark) backup and migration, directly serving content to users, vs. referring requestors back to the originators of the data for copies or assistance, active or passive quality/integrity monitoring, application of metadata, federated searching, dissemination of metadata, etc.

• Collect and archive major segments of at-risk digital geospatial data and images.

• Develop best practices for the presentation of archived digital geospatial data.

• Develop partner communication mechanisms for the project and then ongoing.

• Develop a series of policy agreements governing retention, rights management, obligations of partners, interoperability of systems, exchange of digital objects, etc.

Collections: The materials would not be archived in a single location, but rather retained at Stanford, UCSB or, eventually, elsewhere. UCSB would focus largely on born-digital collections, such as those it currently collects, e.g., LANDSAT imagery and other born-digital content from university, corporate and government sources, and images and reference data accessible at independent web sites. Stanford would focus on historically-

oriented geospatial collections and outreach to organizations (particularly professional societies and state and local agencies) unlikely or unable to assure preservation of their own resources, including digital materials converted from analog sources.

Collaborators: The Partners will seek additional collaborators from university, government, professional society and corporate sectors through the life of the grant and beyond. Several indicative key players are already committed, but we emphasize that these are merely the starting point for a broad outreach to and engagement with the geospatial community. Probable collaborators will include partners such as NASA, NIMA, USGS, several state geological surveys, the Universities of Georgia, Maryland and Washington, Washington State U., and the UC San Diego Supercomputer Center, ESRI, Cartography Associates, Maps.com, and Roelef Odden.

Grant administration: The University of California Santa Barbara is the lead institution for this proposal. There will be two co-Principal Investigators, Sarah Pritchard for UCSB and Michael A. Keller for Stanford.

3. TABLE OF CONTENTS

4. Work Plan

The University of California Santa Barbara (UCSB) and Stanford University propose jointly to lead the formation of a new national collecting network for the archiving of geospatial images and data. This collaborative network would form a new umbrella organization led by these two major institutions (the Partners) with the intention of attracting additional partners and content providers to assure persistent access to a critical mass of historical and contemporary digital information on the United States.

The materials will be a rich mix of file formats, and the Partners will carefully capture, link, or create metadata information for each digital object as it is processed, or ingested, into the Partners' respective digital environments. The target material is vast: according to Dr. Michael Goodchild – internationally recognized expert in GIS – the annual global rate of acquisitions of geospatial data is one petabyte. Also, digital geospatial data forms are technically complex to manipulate. Therefore, it is essential to have at the outset two partners, each with considerable and complementary geospatial/digital-data expertise. UCSB will continue to grow and develop the collection development, preservation and delivery aspects of the Alexandria Digital Library (ADL). Concurrently, Stanford will work with a growing number of content provider partners to extend and develop the utility, security, and coverage of several digital geospatial collections. Lastly, by incorporating the project materials in the Stanford Digital Repository, which is founded on the basis of permanent retention, refreshment, and readability, the materials gathered under the aegis of the NDIIPP program are slated to remain viable as a part of a virtual national digital repository indefinitely.

Proposed deliverables for the joint project will be:

- A national network of heterogeneous federated geospatial repositories of at least four nodes
- 12 to 15 terabytes of preserved geospatial data with associated metadata capable of being transferred to or replicated at the Library of Congress
- Publication of Best Practices papers, including metadata for geospatial files, work flows for ingestion of geospatial files, interoperation between repositories
- A model partnership agreement for digital – particularly geospatial – distributed archives
- A “gap analysis” report on future needs for geospatial digital archiving
- A collaborative website for the National Geospatial Federated Digital Repository that provides archives, links, best-practices documents, FAQs, and contacts information.

A few preliminary comments on the project will introduce the specific project phases.

Relationship among the project phases

The project will consist of four phases, which will run concurrently: Content Identification and Selection; Partnership Development; Content Acquisition; and Retention. Content Identification and Selection have already been started. We have identified a number of content-provider partners and their collections. Partnership Development will consist of working on developing partnership agreements, of mutual benefit to collection providers and to the Repository. We will solicit and accept offers of content from many new partners through the life of the project. Content Acquisition is already present in our organizations. The acquisition of the information, and its preparation for retention, will be a collaborative effort. The focus of acquisition will be on generating reliable metadata – specifically technical and administrative metadata - of

use primarily for the ongoing management of the Repository. The Retention phase will include the development of tools for maintaining a heterogeneous distributed repository.

Cost Sharing

The Partners' cost sharing will be primarily in two forms: expert high level staff focused in depth in this project; and a significant proportion of available storage and processing technology dedicated (permanently in the case of storage media) to the content to be acquired and preserved. Staff involved are listed in the section on Staffing; UCSB and Stanford will average 3.0 and 2.4 FTE per year, respectively. UCSB will contribute some $30,000 per annum in acquisition funding for content. Stanford will also contribute the time and effort of its general counsel and Intellectual Property counsel in the development of agreements with the content providers. Content provider partners will be making available a great deal of valuable information. Their contribution should be noted, though not included in the cost share. Moreover, each content provider will engage in the several stages of data transfer and process management, format discussion and clarification, metadata transfer, linking, proofing, and discussion and negotiation (as appropriate) of rights and terms with the Partners.

Content Identification and Selection

The Partners propose to ingest a broad range of digital content on the Geography, Earth Sciences, and related sciences of North America. Much of the information gathered is historical in nature, and thus provides insight to the American experience, of broad use to the public as well as to researchers. Other information is very up-to-date (e.g., current GIS files) and of vital importance for policy, administration, and legislation; over time, it also will acquire historical value. Overall, the assembled sets of collections will become a deep, broad and growing concentration of digital representations of the American landscape. As the grant proceeds, the Partners expect to coordinate these holdings with those of several additional concentrations, to achieve optimal coverage in the area of American cultural, environmental, and physical geography - materials are of present use to our libraries' users. In addition, the content can help legislators and policy-makers understand contemporary events. Public policy requires research employing geospatial data. Examples include census construction and analysis, redistricting at federal and state levels, government aid determinations, urban and population studies, environmental policy, and land-use policy.

The Partners will continue to identify and select content suitable for addition to their respective repositories. UCSB will continue its focus largely on at risk born-digital collections, revised digital content, historical LANDSAT imagery and other born-digital content from university, corporate and government sources, and images and reference data accessible at independent web sites. Stanford will continue its focus on historically-oriented geospatial collections and outreach to organizations (particularly professional societies and state and local agencies) unlikely or unable to assure preservation of their own digital resources, including digital materials converted from analog sources.

Together, the Partners will preserve a broad and continuously growing body of geospatial information in many forms, of value to the research, legislative and policy communities. By the end of the grant period, we anticipate acquiring, processing, and preserving about 12 to15 terabytes of geospatial information on behalf of the public and education .

Much of the digital information to be collected and preserved through this project is at imminent risk of disappearance. As new revisions of content are created, older versions are abandoned by content creators. Migrating data off outdated media is one of the challenges born-digital preservation. Migrating to large disk stores is one part of the answer, but there is little or no assurance of long-term maintenance or management of the data over time, even when datasets are backed-up or archived to tape. In several prospective cases, the very agency responsible for the data is in imminent danger of being defunded and the data stand a very real possibility of disappearing at the moment the agency is disbanded or absorbed. Information not of interest to the new agency may be destroyed. This has happened to one state's geological survey, where complete destruction of the files occurred. The same fate is very likely in at least four other states.

The Partners understand that digital information held uniquely at any one institution, with or without nominal backups, is authentically at risk. Risk factors include:

- smallness of an institution and the meagerness of its budgets
- direct management of the data by its creators
- relative distance of the data from professional archival specialists
- age of the data
- maintenance of historical archives of different versions of born-digital datasets

Thus, many of our types of targets – local, state and even some federal agencies, professional societies, and isolated projects – are understood to be very much at risk.

In terms of topical content parameters, the program will acquire geographical, geological, environmental, resource management, and related information pertaining mainly to the United States, both past and present. Most of this will have been professionally published either in analog or digital form. Much of the information will be government-produced (at local, state or federal level); but it will also include university, professional society and commercially produced information. The project is intent on developing and negotiating rights agreements that will assure cooperation with the producer/owners, whatever their nature or business models. User input on selection of data, interfaces, and the website for the archive will be via user log files and emails from users.

The quantity of digital information to be captured through the proposed project is estimated at twelve to fifteen terabytes. The usable quantities of information from known partners are not finalized, and new content partners with new bodies of information will be enrolled during the project. The rate at which data can be processed - which is dependent on, among other variables, the quality and quantity of metadata provided by partners - will provide a practical limit on the total amount of information captured during the project's performance period. A main focus of the Content Acquisition phase

will be increasing the speed of processing; this is one of the operational mandates of the project that will lead to documented practices useful to other like efforts.

Selection will concentrate on those collections that are either not copyrighted or where the rights holder(s) will allow: storage by the Stanford or UCSB libraries; access by the public to a usable level of data (e.g., if not full data, then to MrSID images); and transmission to the Library of Congress (if LC requests). Targeted content will tend to be resources produced by:

- government agencies at every level
- not-for-profit organizations, including professional societies
- libraries generally for digital-library purposes
- for- profit agencies – to a relatively limited extent compared with the text world

Examples of content from which data will be selected include:

- as examples of large datasets from agencies that archive digital geospatial data as a part of their charge, the products of federal agencies prominent in producing digital geospatial data, both born-digital (the U.S. National Aeronautics and Space Administration's remote-sensing images such as AVHRR – Advanced Very High Resolution Radiometer; the U.S. Geological Survey's collections of publications that are now produced first in born-digital form and increasingly often available only over the Web, e.g., the Open-File Report series) and scanned (e.g., the topographic maps of the United States at various scales, which have been scanned by different agencies, including libraries).
- well-respected geospatial-data Web pages such as Oddens' Bookmarks (http://oddens.geog.uu.nl/index.html) and Map History (early maps; http://www.maphistory.info/aboutim.html) as a source of URLs for both new and retrospective collection of digital geospatial data available over the Web; this will provide content pools from which to select data – of countries all over the world – created by many different producer sources; in the past, UCSB has concentrated its digital-data acquisition on large datasets covering the state of California, and with the exception of downloading from the U.S. Geological Survey Website all the Open-File Reports dealing with California, it has not in the past ventured to work on the conundrum of dealing with archiving digital geospatial data that is at some point in its lifetime on the Web.
- the products of state and local agencies, which range from born-digital maps available over the Web (e.g., the California Geological Survey's Seismic Hazard Mapping Program; http://gmw.consrv.ca.gov/shmp/) to GIS (geographic information system) layers available perhaps only from a county or city planning department.

UCSB content will focus on acquiring and preserving digital geospatial data and raster-based imagery. UCSB collection standards are that content must be in well documented generic forms, preferably in open formats with freely available source code. Proprietary formats will not be archived unless they are manipulable by non-proprietary software.

Individual files will range in size from about 20 megabytes to about 600 megabytes with some more isolated instances of files of 1.5 gigabytes.

Stanford will continue the active development of relationships with content-provider organizations and will make every effort to accept information with as few barriers (such as format, metadata standards) as feasible. This approach will doubtless produce challenges of several kinds, but will correspondingly provide important experience for eventual robust, real-world repository management. Confirmed and prospective content sources are listed in the first appendix.

Total collection size will start with approximately three to five terabytes at the end of the first year, to eight terabytes by the end of the second year, to twelve to fifteen terabytes by the end of the third year.

Access control and restrictions will be an important area of negotiation with collection owners. It seems likely that collection-owner requirements will range all the way from owners who not only permit but prefer that the archive both store and serve out data, to owners who permit access to metadata but who insist that potential users must directly contact the owner.

Content Acquisition

The Partners will each acquire content in ways best suited to the nature and source of the content and existing technical structures, as discussed below. Over time, we anticipate convergence on some of the approaches as well as process refinement, so the following is best understood as the Partners' respective initial states, rather than as a definitive plan.

UCSB

Technical specifications and standards for capture mechanisms

UCSB will concentrate on working with geospatial repository maintainers. In addition to the ADL content, UCSB will work with ESRI, NASA, National Geographic Society and other partners to identify content of interest to the general educational and public audiences. ADL is capable of working with many data standards. A modified structure of MARC – FDGC is the basis of the metadata repository. Metadata crosswalks exist for all major data and metadata type within the ADL operational framework. ADL captures its content in many distributable forms, e.g., computer-to-computer; CD's, DVD's, tape and directly scanned output. While this is the present state of the ADL system, we anticipate that the Partners will develop tools to replicate content to heterogeneous storage environments of the participating repositories.

Technical protection mechanisms

Protection of the terabyte stores of data is the highest priority of operational issues. ADL has infrastructure supporting three levels of data and metadata backup: first, RAID

level 5 disks, so that if any disk fails, it can be rebuilt; second, on-site tapes that store and backup all data at regular intervals plus incremental segments daily, with tapes rotated to off-site disaster proof storage and testing performed to make sure recovery from backup tapes can be done; and third, a copy transferred at regular intervals to the San Diego Super Computer Center for deep storage.

Content authentication approach

At the time of data load or transfer, check-sums are performed to be sure of data integrity. The process of maintaining data-integrity over time in a participating archive is of major research interest to the computing community. Methods are being examined by ADL researchers on self-maintaining digital objects. Currently file integrity is done by manual monitoring of ADL's data store. Automating methods of archiving objects will be a focus of UCSB and partners.

Security considerations

The ADL middleware fully supports pluggable security in every service interface. From experience, UCSB knows that content requiring rights and security considerations should be stored in a system where security is inherent. Collections that require security can be stored in SDSC's Storage Resource Broker (SRB), or in a system such as Stanford's Digital Repository. ADL-Operational's security infrastructure is by means of firewalls, Internet isolated data stores, tripwire server integrity, log monitoring, and storing of data in SRB. The campus infrastructure monitors intrusion attempts by sampling network activity for known security vulnerabilities.

Vendor tools and services used to acquire content

Currently ADL uses no vendor tools for securing digital content. Direct action between the library and government or private suppliers varies depending on storage media and security issues. Each association is unique and transfer processes are customized. For the purposes of this grant, Oddens' Bookmarks and Map History will be used as sources for URLs of data providers.

Level of effort required for automatic metadata capture

ADL has developed and implemented several automated methods for processing content. ADL has developed and is currently using metadata web forms, backed by databases, for collection and item level metadata. It has also developed methods for expediting data handling and preparing for long-term access, as well as various deliverable formats for the Internet. These methods, production processes and scripts will be shared with those wishing to become a geospatial archive or node in the national network. Data objects containing metadata take unique processing for which ADL has much experience. Data mining techniques will be used as appropriate. ADL is currently building a metadata harvester and collection discovery service.

Stanford

In cooperation with the content provider partners and the Library Systems department, project staff will see to the secure ingestion of contributed data into the Stanford Digital Repository (SDR) in batch operations, depending on the means of transfer available to the provider partner and the extant form(s) of the content. Stanford intends not to acquire much of its content through crawling and harvesting of websites. Details of existing processes are as follows:

Capture mechanisms

The Stanford Libraries currently gather digital assets and metadata in a number of ways: secure FTP download, transfer from some other digital medium (CD-ROM, tape, etc.), or even as email attachments. Regardless of the collection mechanism, the assets and metadata are assigned an intermediate storage location as soon as they arrive. This is usually the Network Attached Storage (NAS) storage array.

Currently Stanford has active processes and tools in place for TIF images, PDF files, and ASCII files (both plain text and marked-up XML). Because of the modularity of our quality control procedures, the only modification needed to add a new format of digital asset is to identify a suitable tool to scan that format. Quite often the scanning tool has been the standard image delivery tool itself.

Ingestion Processing and Quality Control

Once the assets and metadata are stored, the Stanford Digital Repository team members invoke a quality control process. The assets are inventoried and an exception report is generated if anything is missing. A statistically significant number of randomly selected digital assets and metadata are checked for accuracy.

Once the assets and metadata are deemed ready, the process of ingestion into the repository is begun. Stanford has invested considerable programming effort to automate each phase of ingestion, e.g. asset ingestion, creation of structure map, ingestion of metadata, and so on. Each of these scripts generates a status/exception report as well as returning an actual status upon execution.

A master script is run each night. This script first checks the intermediate storage area to see if anything is ready to be ingested. If so, it invokes each of these phase scripts in sequence, checking the status of each phase as it completes and determining whether the next phase should proceed. This master script also creates a report indicating what happened as it was run. These scripts are then emailed to appropriate departments so that, if something goes awry, the appropriate set of people know immediately and can take action. Once the corrections have been made the master script will detect that the set of assets and metadata are ready for ingestion again.

An assessment of the assets' format(s) occurs at this time, to determine if the assets need to be further transformed (e.g. aggregating TIF's into a PDF) for inclusion in the repository, as well as to determine if the format is one of those supported by Stanford's repository.

If the metadata are received in digital format, they are also stored on the NAS at this time, then examined to determine format, quality, and completeness. This in turn leads to one or more subtasks to enhance or refine the metadata and/or convert them to a format which allows easy loading into the intermediate database.

In principle, and usually in fact, the content provider is the best arbiter of the completeness and quality of transferred and ingested data. Effective communication - before, during and after the data are received and processed – is a vital key to the effectiveness of the process and the eventual quality of the data preserved.

Security considerations

The intermediate storage arrays, as well as the permanent storage of the repository, are on servers that are maintained in a central campus data operations environment where they are monitored 24 hours a day, 7 days a week. Incremental backups are made on a daily basis, with full backups occurring weekly. Copies of each backup are kept both on-site and at a remote site.

These servers are also protected from tampering by a hardware firewall as well as a software user authentication which allows only certain identified IT operations personnel and Stanford staff members to access the files on the server.

Partnership Development

The proposal Partners intend to provide a distributed digital repository for geospatial data for the general good, on behalf of many communities, in addition to their own behalf, indefinitely into the future. We intend that the Repository fill an active and cooperative role in the nascent national program of digital repositories as envisioned by the Library of Congress through NDIIPP. We will work together closely - to share practices and experience, to foster standards, to explore effective interoperation, and to recruit additional partners. The steady growth of the network is critical to its success.

Clear and open written agreements will be crucial to the content provider partners. Broadly, there is little trusted third-party digital archiving tradition or established practice (ignoring in this context corporate data warehousing or remote storage of business records); this project - and others within NDIIPP - will be forging the relationships and supporting instruments for such a practice. Having worked with about 150 publishers participating with the HighWire Press and/or LOCKSS, Stanford has considerable experience in the respectful treatment of publisher rights, interests and concerns, on the basis of which no serious difficulties are anticipated in establishing the Partners as trusted third-party repositories for digital geospatial content.

Publishers, university libraries, and even government agencies often believe themselves to have unique needs, interests and situations, and value making their own decisions based on their own analysis. However, it is also clear that there is a concurrent strong interest in reaping the benefits of cooperative effort. This suggests that success will breed success for the proposed geospatial repository program. Once some have agreed to work with the Partners on this project, it is expected that others will join in time.

Content Provider Partners

Content provider partners are necessarily in different states of readiness to act, whether for technical or deliberative reasons. Several providers of content are ready now (at the time of submitting this proposal) to provide content. Still, the plan below suggests the general sequence of events.

- Solicit a pool of early adaptors from several sectors (September 2003 – June 2004)

- Universities

- Professional / Scholarly societies

- Federal agencies, esp. USGS, NASA, NIMA

- State Geological Surveys

- Municipal and county planning agencies

- Other agencies and data holders

- Adapt, negotiate and execute several standard versions of a repository agreement (April – June 2004)

- Retained university IP counsel will coordinate this effort

- Negotiate transfers of sample data from partners (April 2004 – ongoing)

- Content objects as well as metadata (in whatever form)

- Clarify format and linking aspects of partner data

- Develop production parameters and schedule for each partner - batches and/or ongoing capture (April 2004)

- Perform detailed tallies, analysis and quality control of provided content (Ongoing once data have begun to arrive in production mode)

- Renegotiate with provider as necessary

- Work with provider to clarify or solve content-specific issues
- Report to provider
- on completion of discrete bodies of information
- on agreed schedule for ongoing ingestion

The work plan for provider partners may be affected by discussions with the Library of Congress, particularly with regard to standards and expectations.

Growth Plan for Addition of Repository Peer Partners

The goal for this aspect of the project will be to recruit to the partnership several well-established digital repositories with geospatial content in the United States. These partnerships may include ingesting imagery into the Repository, building technical partnerships for experimentation with interoperability, establishing best practices for metadata standards, or creating models for migration of complex data structures, such as banded imagery or GIS data files. Potential partners include:

- The National Geospatial Data Clearinghouse (http://130.11.52.178/gateways.html): a collection of over 250 spatial data servers. Downloadable content will be harvested ad archived from participating sites.
- The Geography Network (http://www.geographynetwork.com/data/index.html): ESRI's data clearinghouse and map serving Web site. This may need to be a set of replicated database, or their content.
- CIESIN (http://www.ciesin.org/): The Center for International Earth Science Information Network at Columbia University.
- The USGS EROS Data Center (http://edcwww.cr.usgs.gov/products/satellite.html): Includes aerial photography, satellite imagery, elevation and land cover data, and maps.
- CUGIR (http://cugir.mannlib.cornell.edu/): Cornell University Geospatial Information Repository, an FGDC Clearinghouse Node for New York State.

Content Retention and Transfer

All content acquired during and under the auspices of the project by the Partners will be ingested by one or another partner, with the intention of storing it permanently. The goal of the Retention phase is to utilize replication to heterogeneous storage repositories. If content can no longer be retained by one partner, the other partner(s) will make every effort to accommodate the content in jeopardy. In the further event that this transfer cannot be accomplished securely, then such content will be transferred to the Library of Congress, as required under the NDIIPP conditions, using transfer protocols and format standards negotiated with the Library of Congress at that time.

The Partners have distinct technologies and programs in place for preserving digital content as described in the following sections. Over time, it is expected these approaches will benefit, though they will not necessarily converge, from the interaction and cooperation among the Partners and staffs.

Policy and legal issues will be worked out with each collection owner so that there is a clear understanding by the collection owner and by the Partners as to what are each party's responsibilities and expectations

Economic and technical issues encountered - judging from previous experience in dealing with terabytes of digital geospatial data - will in the main have to do with planning and budgeting for sufficient computer-technical staff and hardware to keep up both with adding new collections and metadata and with maintaining and providing access to existing collections.

UCSB

The Map and Imagery Laboratory (MIL), Davidson Library, UCSB, first began working with digital geospatial data in the mid-1980s, initially with a MicroVAX with two 90-megabyte disks and what was then a high-end IBM personal computer. Since then, MIL has dealt with tapes, diskettes, cassettes, CDs, DVDs, and increasingly larger amounts of hard-drive disk storage. MIL currently has approximately seven terabytes of digital geospatial data (about 290,000 files) and servers as appropriate, and by early 2004 will have an additional five terabytes of storage available.

Throughout this time period, MIL has understood and lived by the value of backups. Currently all digital geospatial data is regularly backed up on tape and the tapes stored at the library's off-campus storage facility. This pattern of regularly scheduled backup is one that during the grant period will be followed for the archiving of digital geospatial data. In addition, data will be backed up at SDSC. Protection from unauthorized use will be the work of the authentication procedures.

No one place can ever be the modern-day equivalent of ancient Alexandria for archiving digital geospatial data. Therefore, MIL's efforts during the grant are to archive multiple terabytes of different kinds of digital geospatial data and to generate a best-practices document for such archiving, thus encouraging as many institutions as possible to participate in this endeavor. During the grant period, MIL will be archiving many different types of geospatial data, of many different geographic areas, as a part of the work on determining how archiving may best be done. Thus when there are data archived at Santa Barbara during the grant period that may more appropriately in the stewardship of a shareholder other than MIL, MIL will work at turning over the data to appropriate agencies. For example, MIL has a user base for which digital geospatial data of California is of supreme importance and even that data exists in such large amounts that it is common sense to share out the archiving work with other shareholder institutions in California, e.g., government agencies and libraries.

Stanford

Stanford University is strongly committed to the long-term preservation of selected digital information. During a period of fiscal hardship, the University has made a substantial investment of resources into making the Stanford Digital Repository a reality. Stanford has every intent and prospect of building and growing its repository—it is highly unlikely that Stanford would ever find itself unable to preserve the content acquired as part of proposed project.

The current focus for the SDR staff is to save bits and bytes as they arrive in a secure, "enterprise-level" managed environment, assure the integrity of the data received, and rigorous extraction and development of metadata at the point of ingestion. The further vital aspects of digital preservation will be addressed as the state of the art advances and as necessity dictates, once it is assured that the data objects and associated metadata are secure. Stanford Libraries preservation staff has closely monitored discussions of forward-looking issues such as format migration, emulation and encapsulation, but understand that there are more immediate challenges. Much like their peers at UCSB, Stanford staff closely follow, indeed are active participants, in the development of standards and best practices, and will adopt new standards and guidelines as they evolve.

Internally, Stanford has been developing a program of service level agreements for locally owned content that would depend on the form of content being delivered, such that "canonical formats" would be guaranteed to be kept readable as well as in bit-perfect replica of the original submission. Given the variety of geospatial data types and formats to be gathered for this project, Stanford assures at least that bit-perfect files will be retained. Where practical and possible, as part of the larger SDR effort, derivatives of files (either forward migrated or "canonicalized") will be available. Stanford will work with the LongNow Foundation through the project to devise, test, and establish remote secure storage of copies of Repository content.

5 Staffing and Institutional Capacity

UCSB and Stanford are both active participants in the Stanford / UC Map Libraries Group, a cooperative purchasing agreement that has made possible the acquisition of many critical geospatial data sets. This experience has paved the way for the more profound partnership proposed herein. For more information, see http://library.ucsc.edu/maps/ucsmg/.

UCSB Environment

The Libraries at the University of California, Santa Barbara, consist of the Donald C. Davidson Library and the Arts Library. The UCSB Libraries have a collection of approximately 2.7 million volumes, 5,000,000 federal, state, and foreign government publications, and over 315,000 audio recordings. The Library subscribes to over 22,000 serials publications, provides access to over 5,000 electronic journals, and has over 3.6

million microforms. Through the Libraries' Web site, UCSB users anywhere in the world can access online catalogs, databases of articles and books, complete electronic journals, and other scholarly sources. The UCSB Libraries serves a campus of 22,417 students and 1,015 faculty.

The UCSB Libraries have significant experience with managing grant funds and special projects. UCSB just received an IMLS National Leadership Grant for the digitization and preservation of 6,000 wax cylinder recordings. The library has successfully completed several LSTA grants for cataloging of maps and sound recordings, and University of California grants for the digitization of visual and textual materials in the California Ethnic and Multicultural Archives. A current grant from the Mellon Foundation funds investigation in informatics systems for faculty research data. UCSB is an active participant in the digital initiatives of the University of California's California Digital Library including the Online Archive of California and eScholarship.

The Alexandria Digital Library

Our largest and most notable success in this field is the Alexandria Digital Library (ADL). Established as a collaborative initiative at UCSB in 1994 with multi-million dollar funding from the National Science Foundation, its goal was to engineer and build a geospatial digital library. In 1998 ADL became an operational service of the Davidson Library, integrated into regular budget and services and with Internet distribution to all. Since that time it has developed into a significant community resource, serving the needs of instructors and researchers on the campus, in the University of California generally, in the region, and around the world. Many of its features are now fully operational and available to general users. It also continues to evolve, and is a platform for substantial research efforts in geography, computer science and other fields. The UCSB Library – while proud of ADL's accomplishments - is painfully aware that many major improvements are needed, for example: improved procedures for speedy ingest of data and metadata; more user-friendly interface; easy loading of ADL software so it may be readily used by other geospatial data libraries and agencies; and new functionality to address archiving problems (e.g., content verification, automatic migration, access authorizations).

ADL has an operational Catalog (ca. 2 million metadata records) and Gazetteer (ca. 5 million place-name records) in public service. The Catalog was constructed mainly from non-MARC metadata, with digital data primarily from U.S. federal and state agencies. The Gazetteer database was constructed from two existing sources, the Geographic Names Information System (GNIS) for U.S. place names, and for foreign names the database generated by the U.S. Board on Geographic Names (BGN); these entries were merged and redefined via a locally constructed feature-type thesaurus, using a locally constructed content standard. The gazetteer is integral to the cataloging of spatially referenced materials.

The long-term mission of ADL is focused on research and operational issues critical for the construction of distributed digital libraries of geospatially-referenced, multimedia materials; development of technologies necessary to support such a library; design, construction, and evaluation of test-bed systems based on research and development results; and resolution of organizational and technological issues underlying the transition from test-bed system to operational digital library over the long term. This work continues.

Remote-Sensing Images, Maps, and GIS

The Map and Imagery Laboratory (MIL) has collections of maps, aerial photography, satellite imagery and other spatial data of about 4.5 million items. The hard-copy collection is composed of: remote-sensing images (ca. 4 million items) composed of 2.8 million air photos (mainly of areas of California) and 1.2 million Landsat MSS (Multi-Spectral Scanner) satellite images, with the latter being world archival coverage for the time period 1972-1978; a map collection of ca. 480,000 sheets, principally medium and large-scale topographic sheets; ca. 48,000 microforms (mainly fiche); ca. 8,000 atlases and reference books; and ca. 20 globes. The digital-data collection is composed of: ca. 3,500 CDs, DVDs, diskettes (offline); and ca. 292,000 files (on-line) of ca. seven terabytes, growing at the rate of ca. one terabyte per year. In 1992, MIL was ranked the number one spatial-data collection in the top 100 member libraries of the Association of Research Libraries. MIL has three PCs loaded with ArcInfo, ArcView, and ERDAS Imagine and two staff specializing in working with users of digital geospatial data.

UCSB Staffing

The proposed project requires the participation of several library personnel, with an emphasis on computer engineers who build content (data and metadata) of the Alexandria Digital Library. The head and the assistant head of MIL will be heavily involved. Percentages given below for time given to proposed project are consistent through all three years of the proposed project.

Library management:

- Sarah Pritchard, University Librarian (5%) - PI
- Larry Carver, Director of Library Technologies & Digital Initiatives; head of Map and Imagery Laboratory (20%) – Project Director
- Mary Larsgaard, Assistant Head of Map & Imagery Laboratory (30%) – selection, cataloging, and user needs
- Marilyn Moody, Assistant University Librarian (5%): Collections and Information Services input

Library systems:

Alexandria Digital Library-Operational (ADL-Op):

- Computer and network tech III: Catherine Masi (60%); in charge of ADL-Op
- Computer and network tech II: Dave Valentine (60%); lead programmer for ADL-OP
- Computer and network tech II: Greg Hajic (60%); digital-data/digital-data software specialist
- Computer and network tech II: Cian Phillips (20%); Web programming

Library/MIL systems:

- Computer and network tech III: Clay Burnham (10%); head, Library/MIL systems
- Computer and network tech II: Kim Park (20%); assistant head, Library/MIL systems
- Lab assistant II: Carolyn Jones (20%); metadata preparation manipulation for ADL-Op
- Library assistant IV: Ann Hefferman (20%); MIL office manager and student-assistant supervisor

The core project team of current personnel will be: the Director of Library Technologies & Digital Initiatives (i.e., the project director); the assistant head of MIL; the head of ADL-Op; the lead programmer for ADL-Op; and the digital-data manager for MIL and ADL-Op. Job descriptions and resumes are in the appendices.

Staff to be hired as part of the project team are two programmers (one for data, and one for metadata), an information scientist to lead the enhancement of the ADL Gazetteer, and a site (project) manager. The latter is especially important for coordinating the partnership activities. Dr. Michael Freeston (NSF-supported researcher) will kindly serve as a consultant.

The proposed project requires the participation of several library personnel, with an emphasis on computer engineers who build content (data and metadata) of the Alexandria Digital Library. The head and the assistant head of MIL will be heavily involved. In the digital as in the hardcopy world, intelligent, informed selection of resources is a key to success. In the world of digital geospatial libraries, selection is done by collection managers with many years of answering reference questions in map libraries, and a considerable knowledge of geospatial data. For the purposes of this grant, selectors will be the head and the assistant head of the Map and Imagery Laboratory, Davidson Library, UCSB; the two have a combined experience of selecting and working with geospatial data of 72 years. These selectors will consult with data producers and data users, the latter mainly via email. Percentages given below for time given to proposed project are consistent through all three years of the proposed project.

Stanford Experience

To serve its mission, Stanford University Libraries &Academic Information Resources (SULAIR) employs digital information technology creatively to improve scholarly practices, to make them more productive, more creative, more expressive, timelier, and

more affordable. SULAIR has created or enfolded several important university enterprise organizations and projects to aid these processes: Media Solutions, the Stanford University Press, and the HighWire Press. It has also pioneered Open Software development efforts, most significantly, CourseWork (an OKI-compliant Course Management System) and LOCKSS (an Open Source, decentralized preservation technique for e-journals and other material). We believe the academic world is just at the delta point for great changes in the way scholarly information is produced, shared, used, published, and archived, and we are moving aggressively in all these areas. We believe we have the duty as well as the opportunity to provide services, models, and best practices to academia, information industry, business, and others.

SULAIR has been developing, exploring, purchasing, and making available digital information for many years, much as have other major research libraries. However, it has uniquely applied itself to digital publishing, as demonstrated by HighWire Press, and a variety of like projects, including Knowledge Environments ™ and numerous other efforts. For almost five years in this blended organization, its professional staff has been about equally divided between librarians and technologists, all dedicated to the use of information in teaching, learning, research, and publishing.

The Stanford Digital Repository

Digital preservation is among the greatest concerns and interests of SULAIR. Particularly as SULAIR's HighWire Press holds and serves some two terabytes of customer data – representing about 350 electronic journals belonging to over 120 publishers (most of which are scholarly societies) – assuring digital continuity is a constant issue. As SULAIR understands itself to be a service for future generations, and mindful of the fragility of digital information and the certainty that a significant fraction of critical information created in the past decade (or longer, as in the case of NASA tapes), achieving a stable platform for digital preservation is a very high priority.

The proposed effort fits integrally into the SDR concept and operation. If the partnerships develop as intended, the volume of geospatial material will approximate one-third the volume of information in the SDR through the grant period. The challenge is to learn to acquire and ingest at ever faster rates, and the project's geospatial data will help the learning process, much as other experience over the past few years will inform the proposed geospatial project.

SULAIR has been working on SDR for three years, concentrating on exploring middleware management tools, developing metadata models, and planning service level agreements for content holders. The SDR is understood to be integral to the organization, rather than an appendage. Collections and Services and Technical Services divisions each have decision-making bodies to guide the SDR, which operates under the aegis of the Preservation department. SDR staff work closely with the Systems department and the Chief Information Architect.

SDR is the bottom layer of a three-tiered endeavor. At the top layer, the collection program provides for the selection of content, as well as for the intellectual interface between our digital holdings and prospective users of them. At the "action" layer in the middle sits our preservation program. Here resides the Stanford research library's permanent commitment to its never-ending "remember and remind" role. The preservation layer is SDR's service interface to the collection layer above it. In turn, the preservation program shapes and drives the objectives for SDR's layer of storage, refresh, and delivery technologies below it.

In terms of SDR's technology, we provide three levels of accessibility, with online proxies being readily accessible to the extent that IP, security, and ownership requirements permit. For those types of resources where ready access to full copies of source objects is important (e.g., art objects, pages from manuscripts, historical maps), SDR keeps near-line copies ready for robotic transfer to an online cache. For other resources, the source objects move to an off-line storage environment from which they can return to an online cache for use by the next business day.

Maps, Earth Sciences, and GIS

Branner Earth Sciences Library and Map Collections houses all of the maps dated after 1840 for the Stanford libraries. They recently took possession of the Hoover and East Asia map collections, which are now being integrated into the map collections. Branner's collections include maps and aerial photography (paper and digital), satellite imagery and spatial data in numerous formats. The map collection has about 270,000 sheets with strong subject emphasis on geology, topography, and the environment; hundreds of atlases with world-wide coverage; aerial photography focused on Santa Clara and San Mateo counties; and Landsat and DOQQ imagery bought in conjunction with the University of California Map Libraries. Branner is also the main point of service for all GIS needs on campus, including software, data, and technical expertise. The direct involvement of Branner staff in this project will be pivotal in several respects: contacts with content providers, identification and evaluation of potential content, management and training of the project staff, technical evaluation, and setting of metadata standards specific to geospatial material.

Stanford Staffing

The proposed project will be very prominent in the SULAIR organization; the University Librarian and several of his direct reports will be involved to assure the success of the project. A highly qualified and experienced tier of senior managers engaged intimately in the project, round out the expertise from the several areas of the organization:

- Mike Keller, University Librarian – PI
- Julie Sweetkind-Singer, GIS & Map Librarian– Project Director
- Assunta Pisani, Associate University Librarian – Collections and Services engagement
- Catherine Tierney, Associate University Librarian – Technical Services

- Jerry Persons, Chief Information Architect – overall system design; standards implementation; integration of technical components
- Gerry Smith, Manager of Systems – operational system implementation and data streams
- Connie Brooks, Head of Preservation and Manager of the Stanford Digital Repository – direct responsibility for the growth and operation of the SDR
- Meredith Williams, GIS Manager – validation of formats, file descriptions, compatibility and software issues
- Vicky Reich, Manager of LOCKSS Project – publisher content agreements, cooperative digital archiving development
- Katherine Kott, Head of Cataloging and Metadata – assurance of integrity in metadata harvesting and development, adherence to local and national emerging standards
- Nancy Hoebelheinrich, Metadata Librarian – monitoring emerging geospatial and other metadata specifications, harvesting and transforming metadata, testing of access services
- Walter Henry, Lead Analyst, Media Preservation – archival review, coding
- Cathy Aster, Head, Media Preservation – preservation planning, emerging standards
- Paul Zarins, Head of the Digital Library Program – integration with concurrent systems and efforts, testing of access technologies
- Christa Easton, Head of Acquisitions – digital rights management

Job descriptions, time commitments and resumes are to be found in the appendices. Time commitments of the Stanford staff vary over the course of the project and total over seven FTE years.

There will be a core project team, comprising the GIS & Map Librarian (i.e., the project site director), the Digital Repository Manager, the Metadata Librarian, and the Manager of Systems, and the Digital Repository Project Manager. This group will meet regularly and confer with the Partners as appropriate. Other staff will be involved, ranging from the systems staff of the HighWire Press, to the software development team in Academic Computing (another division of SULAIR). The Digital Repository Project Manager, reporting to Connie Brooks, will be hired before the project would begin (job posted October 2003).

Finally, staff will be hired through the grant at Stanford to perform several production functions of the project. The hired staff will include a geospatial librarian who will serve as site manager and project liaison; a metadata assistant to assure adherence to appropriate standards and "map" across schema (in conjunction with the Metadata Librarian); staff trained to review, correct, and add metadata, to track content and media in the ingestion stream, perform quality control, and conduct the communications associated with permissions, contracts, and related partnership documentation; and a programmer/analyst to handle and manipulate digital files and any necessary software.

Lastly, but by no means least, Stanford intends to retain the input, engagement and council of outgoing Earth Sciences Library Head Charlotte Derksen on a contract basis.

Advisory Group

We intend to establish an advisory group large enough to accommodate the many kinds of interest and stakeholders involved in the field, but small enough to function as a working group. There will annual or semi-annual face-to-face meetings of the group and ongoing communication via an email listserv. While the group will not have final authority to make decisions, it will be expected to engage deeply in the operational and policy level decision making of the project. The topics brought to it will include technical, legal, research, organizational, strategic and business issues. A prospective list of initial members of the advisory group is provided as an appendix.

The advisory group will include, at any one moment, between 12 and 18 members, generally representative of the several communities involved. It will include individuals from other universities, repository peers, state agencies, professional societies, and commercial publishers. There will not be positions or quotas reserved for a certain interest. Members will be nominated at first by the project staff and thereafter by the group and ratified by the Principal Investigators. It is hoped that the Library of Congress will allow a senior staff member to serve on the group throughout the project; in particular, input from the Geography and Maps division would be most welcome, as well as program management in NDIIPP itself. Terms will be for two years, but the terms of the more productive members will probably be renewed.

Conclusion

The proposed National Geospatial Federated Digital Repository will fulfill, for a broad class of critical information, the basic objective of the NDIIPP: a national network of cooperating repositories, populated with a significant body of digital information that would otherwise be at risk of loss, operating with sound protocols and best practices, under the aegis of the larger program led by the Library of Congress. The Partners will commit to working toward a model of universal access to its contents. At the end of the three-year grant period, the project will have produced:

- A national network of heterogeneous federated geospatial repositories of at least four nodes
- 12 to 15 terabytes of preserved geospatial data with associated metadata capable of being transferred to or replicated at the Library of Congress
- Publication of Best Practices papers, including metadata for geospatial files, work flows for ingestion of geospatial files, interoperation between repositories
- A model partnership agreement for digital – particularly geospatial – distributed archives
- A “gap analysis” report on future needs for geospatial digital archiving

- A collaborative website for the National Geospatial Federated Digital Repository that would provide archives, links, best-practices documents, FAQs, contacts information.