DIGITAL LIBRARY
FEDERATION

Stanford University

Report to the Digital Library Federation
Spring, 2004

TABLE OF CONTENTS

I. Collections, services, and systems

A. Collections

Antiquarian Maps of Africa

Maps of Africa features antiquarian maps from the collections of the late Dr. Oscar I. Norwich and the Stanford University Libraries, dating from the late 15th to the early 20th century. They are a majestic tribute to centuries of exploration, cultural interaction, mapmaking, printing, and collecting, as well as a scholarly source for a wide range of researchers. Dr. Norwich's well-known collection was documented in a print volume, *Maps of Africa: an Illustrated and Annotated Carto-Bibliography* (1983; 2nd ed. 1997), and was recently acquired by the Stanford University Libraries' Department of Special Collections. These maps, more than 300 in number, along with more than 200 antiquarian maps previously held in Stanford's collections, were captured as high-resolution digital photographs, cataloged, and joined into a single digital collection, which is served via Luna Insight. The acquisition and digitization of the Norwich maps, and the digitization of the Stanford maps, are part of the William Jacobson Africana Program. http://library.stanford.edu/depts/hasrg/hdis/aboutinsight.html

Stanford Geological Survey

The Stanford Geological Survey (SGS) existed for 100 years, from 1895 until 1995. During this time, students and faculty from Stanford's School of Earth Sciences, went into the field to survey and map parts of California, Nevada, Idaho and Utah. From September 2001 until October 2003, Branner Earth Sciences Library and Map Collections received grant funding from the California State Library (LSTA monies), and matching funds from Stanford University Libraries, and the School of Earth Sciences. The grant funding allowed the project staff to catalog, scan, and display the manuscript maps, field notebooks and field reports from this collection. All of the work produced by the Survey has been cataloged in Socrates, Stanford's online catalog. The Web site, with a portion of the map imagery, is now online. The user must download the Luna Insight software Java client in order to access the collection. Questions about the project can be directed to Julie Sweetkind-Singer (sweetkind@stanford.edu). Instructions and the client are available at: http://gill.stanford.edu/depts/branner/SGS_home.html

Survey of Race Relations on Pacific Coast Digitization

In a collaborative project at the Stanford University Libraries and the Hoover Institution Archives, the Survey of Race Relations Records, 1924-1927, a pioneering social science undertaking using life-history and survey research techniques, is nearing completion. The original archive comprises 37 manuscript boxes and additional material from the University of Oregon, the University of Chicago and University of Southern California may be included. The archive contains reports, correspondence, interview transcripts, questionnaires, and printed matter, relating to the social and economic status of Chinese, Japanese, other Oriental, Mexican, and other minority residents of the Pacific Coast of the U.S. and Canada, and to race relations on the Pacific Coast. A digital finding aid to the collection may be found at http://findaid.oac.cdlib.org/. Public access should be available in the early spring, 2004.

B. Services

CourseWork and the Sakai Project

CourseWork, Stanford’s course management system, was developed by Academic Computing with two goals: to meet the diverse needs of the teaching program at Stanford, and to develop a sustainable, interoperable platform that could host new features as they were envisioned by the faculty. CourseWork was designed to meet the pedagogical needs of the faculty. The original design focused on two major teaching programs, undergraduate large lecture classes in the sciences and language teaching, as well as commonly requested features from faculty across the curriculum. Originally funded by the Andrew W. Mellon Foundation as part of the Open Knowledge Initiative, CourseWork is in use today by over 40 percent of the faculty in all seven schools.http://aboutcoursework.stanford.edu

The Sakai Project is Academic Computing’s next development effort. While CourseWork is successful, as with all technology, we must plan for upgrades and new features. In collaboration with three other institutions—Indiana University, the University of Michigan, and MIT—Stanford is planning the development of a new system that will replace the course management systems now in use at all four institutions. The project, called Sakai, intends to capitalize on the significant intellectual property of these institutions to produce a “best of breed” course management and portal environment that is truly modular, scalable, and interoperable. By combining the resources of the four institutions, we will achieve more than any one of us could individually. The products of the Sakai collaboration will be moved into production at Stanford in Fall 2005, and will be available in open source. http://www.sakaiproject.org/

Discovering Dickens: Stanford University Libraries Contribute to Innovative Community Reading Project

December 2003 marked the beginning of the second year of a Stanford University "community reading" project called *Discovering Dickens*, to which the University Libraries continue to make major contributions of rare content and digitization expertise. The project is an invitation not only to the Stanford community, but also to remote readers, "to reenter the world of serial publication and of family reading circles," by reproducing and distributing Charles Dickens' own original serial publication of one novel each winter, as both paper facsimiles and downloadable (and searchable) PDF surrogates, which are released in the same weekly segments as originally published. The program is marked by public dramatic readings and other events; the on-line publication is enhanced with community-focused biographical, historical and scholarly notes, maps and other digitized images, and a "Key to Allusions" related to each serialized segment. The texts are scanned from original copies of Dickens' weekly journal, *All the Year Round*, using an evolving set of library technologies and equipment. Having presented the community with a unique opportunity to read *Great Expectations* "in the original" during 2002-2003, we are now offering *A Tale of Two Cities* in the same way. http://dickens.stanford.edu

Robotic Book Scanning

The Stanford University Libraries' Digital Library Program is well into its first year of utilizing the 4DigitalBooks Digitizing Line robotic book scanner for the digitization of bound library materials. Seven months into live production, the Digitizing Line has scanned approximately 400 books constituting over 270,000 page images. The formats, book structures and content of materials scanned on the DL vary greatly, ranging from books published recently by the Stanford University Press and Center for the Study of Linguistics and Information to books in the public domain published prior to 1923. The lab is currently digitizing a series of Congressional hearings on the Joint Commission on Atomic Energy, an effort to increase SUL's online collection of government documents. For a white paper describing the lab, see:
http://www-sul.stanford.edu/depts/diroff/DLStatement.html

Web Site Redesign

A completely redesigned Web site (http://library.stanford.edu) was implemented in January 2004 for the Stanford University Libraries and Academic Information Services. Stanford University Media Solutions (http://mediasolutions.stanford.edu/) designed and built the new Web site, in consultation with SULAIR staff. Media Solutions redesign included emphasis on a more user-centered information design; improved ADA compliance; improved, quick access to research resources and to information about SULAIR services; new How To sections with links to instruction and policies; and many other improvements.

C. Systems

Data Extraction Web Interface (DEWI)

DEWI is a web-based data search and extraction system for easy access to over 100 social science numeric datasets in the Stanford collection. Built to leverage the software and computing resources available to faculty, staff, and students, DEWI acts as a variable discovery tool where extract files can be easily created for use in research or instruction. These files can be saved in a variety of statistical software formats, such as SAS, SPSS, or Stata, and downloaded directly to a user's computer for further analysis. Links to codebooks and other online resources are provided for each dataset.
Examples include: the General Social Survey Cumulative File; American National Election Studies Cumulative File; Correlates of War Project: International and Civil War Data, 1816-1992; Slave Sales and Appraisals, 1775-1865; and Latino National Political Survey 1989-1990. DEWI is ideal for use in the classroom and has been used in introductory courses on social science computing and quantitative methods. If you would like to learn more about DEWI or would like to take DEWI out for a test drive, send a message to Ron Nakao (ronbo@stanford.edu).

HighWire Press Update

HighWire hosts the full text content of 346 journals (some material dates back as far as 1891). Searching on HighWire means querying over 14 million articles – not just abstracts, but also the full text of the articles hosted on HighWire’s servers. HighWire’s free collection has grown substantially over the last several years, and as of mid-December 2003 offered 644,000 free full text articles (http://highwire.stanford.edu/lists/freeart.dtl).

The American Society for Biochemistry and Molecular Biology (ASBMB) has published a series of articles about the HighWire site in the monthly ASBMB

Newsletter. This series of articles provides a good template for libraries and other societies to adapt for their own use to identify the features and functions available to users of HighWire. http://highwire.stanford.edu/inthepress/asbmb

Recently released search functions include MatchMaker and Instant Index. MatchMaker is a tool for providing matching articles based on a weighted set of topics. This weighted set can be generated from a single citation or from a search result. The resulting MatchMaker patterns can also be manipulated graphically to emphasize one element of an article over another. A link to MatchMaker can be found to the right of many articles on a HighWire search results page. Instant Index, using the Vivisimo Clustering Engine, allows readers to find and retrieve information organized into hierarchical folders. This enables them to tell at a glance the types of information available on any topic in a search result. It also helps results that might otherwise be buried deep in later pages to rise to the top. Since the clusters are created on the fly, they are as current as the content itself. On the HighWire search results page, Instant Index is displayed on the top right under the section called "Subject Indexes."

HighWire offers a variety of tools for managing subscriptions. These include viewing usage reports across all journals administered, and the ability to maintain IP addresses in one place for multiple pages. A recently released feature is "Shop for Journals." A group of HighWire-affiliated publishers have agreed to conform to a common pricing model and a standard set of Guidelines for Internet access to journals for their institutional subscribers. http://highwire.stanford.edu/institutions

LOCKSS

The LOCKSS Program (Lots of Copies Keep Stuff Save) open source software provides a peer-to-peer, distributed, persistent access preservation system for web delivered information. The LOCKSS technology provides persistent access to materials that are delivered through HTTP; published in serial; have a reasonably stable URL structure; have or can be made to have an authoritative version [http://lockss.stanford.edu/projectdescbrief.htm].

In 2003, the LOCKSS program, which includes a growing number of participating libraries and publishers, [http://lockss.stanford.edu/projectstatus.htm] made substantial progress towards implementing a robust, sustainable, distributed persistent access repository.

Libraries are beginning to collect and preserve leased electronic journals purchased through individual institution subscriptions or consortia licenses and important freely available electronic titles [http://lockss.stanford.edu/titleregistry.html]. Coordinated collected development efforts are underway in the Humanities, government documents http://lockss-docs.stanford.edu/ (see below for more information) and a for a few geographically based area studies. The collection development process is briefly outlined at [http://lockss.stanford.edu/collectionswork.html]

The LOCKSS software turns a personal computer into a preservation tool by transforming it into persistent web cache [http://lockss.stanford.edu/techoverview.html]. The production version of the software is on schedule for release early 2004. Emory University prototyped integrating LOCKSS caches into institutional networks so readers are delivered information when the publisher is unavailable [http://lockss.stanford.edu/lockssnetworkintegrationview.html] Indiana University, with input from the community, built a user interface librarians can use to manage the preserved content and the LOCKSS computers [http://documents.lockss.org/uidemo/].

The LOCKSS Alliance was established to support ongoing work as we transition from grant to community support. The Alliance, a fee-based organization within Stanford University Libraries is overseen by a Board of Advisors. Libraries need not join the alliance to run LOCKSS caches; publishers need not join the Alliance to have their materials preserved on the system. However, the Stanford team will begin to offer selective support and services to Alliance members beginning summer 2004.

For further information, contact Vicky Reich, Director LOCKSS Program, Stanford University vreich@stanford.edu

DOCKSS: LOCKSS for Government Documents

Stanford received an NSF planning grant in Fall 2002 to explore the functional requirements for development of the LOCKSS technology to address web-based Federal government documents. The one-year grant covered the costs for several stakeholder meetings and one 2-day workshop during which government information librarians at nine partner libraries addressed ways in which the LOCKSS technology might be adopted by members of the Federal Depository Library Program (FDLP) to support their roles in preserving and providing access to government information. Managers at the Government Printing Office (the grant’s sponsoring agency) actively participated in the project effort; the GPO identified ways in which the LOCKSS technology might be adopted within the agency to achieve agency goals related to capturing, authenticating, and preserving web-based government information. A Needs Assessment document incorporating these findings is posted on the project website at: http://lockss-docs.stanford.edu

Luna Insight: Expanding and Integrating Image Offerings, Both Local and Remote

Stanford has been using Luna Imaging's Insight® image management and delivery software since late 2001. 2003 saw a significant increase in the number and variety of image collections created and served through this system, as well as the first all-Insight course taught, at Stanford. Insight was instituted in 2001 with two collections unique to Stanford: the Athanasius Kircher Correspondence, and Chicana Art. The former was launched in conjunction with a major exhibition of Kircher first editions and an international conference at Stanford in 2001; and the latter, a project of Professor Yvonne Yarbro-Bejarano of Stanford's Department of Spanish and Portuguese, was the principal source and teaching tool for two instances of an undergraduate course in Chicana art forms during 2003. At the same time, local digitization and collection building efforts produced two major unique collections, the Stanford Geological Survey and Maps of Africa (see separate descriptions in the "Collections" section of this report); the addition of several shared image collections (primary among them the David Rumsey antiquarian map collection); and our first foray into a subscribed collection delivered through Insight, the AMICO Library. All of these collections are now available to the Stanford community through a single, cross-searchable interface. By the same token, several of our local collections (Kircher, African Maps, and Stanford Geological Survey) are available as shared, cross-searchable collections to any other institution using Insight. (In fact, the general public is able to access Stanford's non-restricted collections via Insight; other Insight institutions are able to offer them seamlessly alongside their own.) We anticipate continuing growth and use our Insight image collections.
http://library.stanford.edu/depts/hasrg/hdis/aboutinsight.html

II. Projects and programs

GATT Digital Archive

Work is proceeding on the IMLS-funded portion of this project involving the digitization of approximately 17,000 documents of the General Agreement on Tariffs and Trade (GATT) published between 1947-1994. Over half of the projected 500,000 digital images have been created from microfiche and paper sources. In addition to the base TIFF files, raw OCR text and TEI-level markup files have been created. Scripts for ingestion of these files and associated metadata into the Stanford Digital Repository (SDR) are being created. Specifications for the web interface which will be implemented by the Media Solutions group at Stanford are complete. More details are available at the project’s website: http://gatt-archive.stanford.edu

The Machinima Archive

The Machinima Archive will be a historical collection of machinima files and information built by a collaboration of parties and hosted by the Internet Archive (http://www.archive.org) in a manner that secures long-term preservation of a significant body of work in the new medium of machinima (animated movies made using game software). The project is coordinated by Henry Lowood, Curator for History of Science & Technology Collections in the Stanford University Libraries, as part of the How They Got Game Project, Stanford Humanities Laboratory. An initial selection of machinima movies has been made, a brief metadata scheme developed, and a design for the website -- which will be a moving image collection of the Internet Archive -- approved. We expect to begin loading movie and software files in January 2004.

III. Specific Digital Library Challenges

Stanford Digital Repository

Various elements of the current development of Stanford's Digital Repository (SDR) could be, for the purposes of this report, described under Systems, Services, Projects. We have presented this information in the Digital Library Challenges section because the development of the SDR illustrates and intersects with many of the key current challenges. Many thanks to Jerry Persons, Chief Information Architect for Stanford University Libraries and Academic Information Resources, for providing the following summary and analysis of the SDR and related issues.

SDR Context

A large research institution, like many corporations, typically holds a vast amount of unmanaged or under-managed digital information. A considerable portion of the digital information resources coming to campus arrive in concert with curation and service programs of Stanford's several library programs. In addition we know that there are hundreds of servers on campus, all of which lie outside the purview of libraries, many of which contain significant information of permanent interest to the institution, every single one of which will fail or be replaced within a small number of years. The culture of digital decentralization presents the likelihood of wholesale loss of important, mission-central digital documents. The SDR (Stanford Digital Repository) offers Stanford a means to gather and protect every facet of the University's organizational memory and intellectual capital.

SDR Scope

Working in concert with Stanford Libraries' Preservation program, the SDR will manage content in the Libraries' own born-digital collections (i.e. purchased, created, or donated) as well as the products of numerous, large-scale digital conversion projects (e.g., ca. 20 million pages of pre-digital content from 350+ journals serviced by HighWire Press). It will encompass the born-digital resources produced by the Stanford Libraries' two publication programs – HighWire Press (http://highwire.stanford.edu/about) and Stanford University Press (http://www.sup.org). The Repository also will partner with schools and departments to gather and preserve the output of Stanford's teaching and research programs (digital working "papers," published articles, models, animations, conference proceedings, course materials, etc.). Furthermore, the SDR will seek opportunities to provide publishers and owners of digital content that is germane to Stanford's research and education programs with a permanent, secure, cost-effective home for materials in any and all digital formats.

SDR Services

Foremost among all the capabilities and services provided by the SDR is the Stanford Libraries' commitment to the permanent preservation of content in digital formats – this being the very same commitment, with its long history of successful accomplishment, that underlies every research library's program to build and preserve collections of knowledge in any pertinent format, be it "traditional", digital or "post-digital."

At a detailed level, permanent preservation requires an ongoing commitment to services that can successfully assess and capture the salient features of every type of new content presented for ingestion into the SDR. Included here must be capturing appropriate characteristics of the digital objects themselves as well as documenting the intent/will of the content owner with respect to the long-term use of objects. Such work includes judgment calls about what can and cannot be preserved—work that is a continuous balancing act weighing "perfect preservation" against practical constraints like costs and technical feasibility in order to set reasonable, achievable long-term goals for each new collection of content.

From an operational point of view, the products of such analysis shape the scope and focus of metadata encoding; they drive both policy and operational decisions for SDR data migrations; and they serve to define the provisions of the service agreement between data providers/owners and the SDR.

Stated in terms of the OAIS Reference Model (http://www.ccsds.org/documents/650x0b1.pdf) the products of the productions of these analytical services drive both the SDR's Preservation Planning and its Administration support components--which themselves serve to shape and control the Ingestion, the Data_Management/Archival_Storage, and the Access services delivered by the SDR. Negotiation of a service agreement for each collection also governs the extent to which the Repository's archival information package for content is populated with preservation description information. The SDR is designed to allow ingestion of content with accompanying preservation metadata that ranges from records that fully conform to established standards and best practices all the way down to service agreements that promise nothing but bit refreshment for objects with file names as the sole metadata.

SDR and Canonicalization

From a strategic point of view, the Stanford Digital Repository serves to ingest, preserve, and deliver access to the most stable, faithful representations of digital content that can be had at affordable costs using the most appropriate technologies and the best practices of the day. Several years ago, Clifford Lynch provided the framework on which the SDR's policies and practices are modeled in his article on "canonicalization" (http://www.dlib.org/dlib/september99/09lynch.html). We make special note of this facet of the SDR's intent and capabilities in order to set our efforts apart from those whose work includes both research and applications of software and hardware emulation as a means of long term preservation of digital formats (http://www.clir.org/pubs/reports/rothenberg/contents.html). We based our choice of approach on what we view to be the practical requirements of making substantive progress toward protection of large amounts of digital content at affordable costs using reasonable transformations and translations of content into formats that have the best prospects of long-term usability regardless of the computing resources applied to those formats.

IV. Digital library publications, policies, working papers, and other documents

Derksen, Charlotte, and Julie Sweetkind-Singer. Forthcoming. "Accessing and preserving field maps and notebooks: the Stanford Geological Survey Map and Field Notebook Access Project." In *Proceedings*: Geoscience Information Society.

Frost, Hannah. 2002. "A comparative analysis and evaluation of EAD implementation guidelines." *Journal of Archival Organization* 1 (3).

———. Forthcoming. "Waiting to happen: lessons from preserving disaster-afflicted electronic media in an archival collection." Paper read at Preservation of Electronic Records: New Knowledge and Decision-making, Canadian Conservation Institute, Ottawa, September 2003.

Keller, Michael A., Victoria A. Reich, and Andrew Herkovic. 2003. "What is a library anymore, anyway?" *First Monday* 8 (5). http://firstmonday.org/issues/issue8_5/keller/index.html

Lerner, Heidi. 2003. "Historical Jewish periodicals/newspapers on the Web." *Perspectives: the newsletter of the Association of Jewish Studies* Fall/Winter 2003.

Lerner, Heidi, and Seth Jerchower. 2002. "Metadata, digitization, and the Cairo Geniza: issues in user access and retrieval." Paper read at Association of Jewish Libraries Annual Convention. Summary at http://www.lib.cam.ac.uk/Taylor-Schechter/GF/45/#catalogue and complete text (for Association members only) at http://aleph.lib.ohio-state.edu/www/ajl.html

LOCKSS Team. 2001-2003. LOCKSS Papers and Technical Reports. http://lockss.stanford.edu/peerreviewedpapers.htm

Lowood, Henry. 2001. "The hard work of software history." *RBM: A Journal of Rare Books, Manuscripts, and Cultural Heritage* 2 (2):141-161.

———. 2002. "Archives and online sources." In Atushi Akera and Frederik Nebeker (eds), *From 0 to 1: an authoritative history of modern computing*, (pp. 207-214)). Oxford University Press.

———. 2002. "Shall we play game: thoughts on the computer game archive of the future." Paper read at BITS OF CULTURE: New Projects Linking the Preservation and Study of Interactive Media, October 7, 2002, at Stanford University. http://www.stanford.edu/class/sts145/Library/shall_game.rtf

———. 2003. "Playing art with games: computer games in the museum." Paper read at CIMAM Conference, UC Berkeley Art Museum and Pacific Film Archive, San Francisco Museum of Modern Art, November 13, 2003 session at Yerba Buena Center for the Arts.Online access forthcoming.

———. 2004 Forthcoming. "The obstacle course: documenting the history of military simulations." In *America's Army for Game Scenes/Bang the Machine Exhibition Catalog*: Yerba Buena Center for the Arts.

Sweetkind-Singer, Julie, Mike Powers, and Charlotte Derksen. 2003. "Stanford Geological Survey Access Project." *Information Bulletin* 34 (3):145-157.