

Stanford University
Report to the Digital Library Federation
April, 2005

TABLE OF CONTENTS

- i. Collections, services, and systems
- ii. Projects and programs
- iii. Specific digital library challenges
- iv. Digital library publications, policies, working papers, and other documents

I. Collections, services, and systems

A. Collections

Medieval and Modern Thought Text Digitization Project

The goal of Stanford's Medieval and Modern Thought (MMT) Text Digitization Project is to digitize on an ongoing basis reference works, source collections, and primary and secondary books in the broad area of medieval and modern thought. The Project's main purpose is to fulfill researcher needs for searchable text in support of ongoing research. Also, reference books are being added to the collection, including bibliographies, manuscript catalogs, and biographical lists. Content is drawn from the collections of the Stanford University Libraries and from other member libraries of the Research Library Cooperative Program. Digitization work is being done in-house, and at the end of calendar year 2004, approximately 75,000 pages had been digitized and converted to searchable PDF. The Smart Family Foundation has provided financial support for the project through the Allan Morgan Standish Book Fund, thus expanding the Fund's traditional role of purchasing library books to include also digitization of relevant materials.

<http://standish.stanford.edu>

Parker on the Web

Stanford is working with Corpus Christi College, Cambridge, to digitize the more than 500 manuscripts in the Matthew Parker Library and make them available through a rich and flexible online scholarly tool. We are near the end of an interim grant from the Mellon Foundation, through which a prototype website (populated with all page images of two Parker manuscripts) and techniques for production-scale imaging were developed.

The full project, if funded, will require several years of scanning and development. We expect that the platform to be developed will be adaptable to other collections of manuscript materials.

<http://parkerweb.stanford.edu>

Stanford Historical Photograph Collection

The Stanford Historical Photograph Collection is one of the Stanford University Archives' major collections, consisting of some 16,000 photographs from throughout Stanford's history, and covers architecture, events and personalia (including a comprehensive set of images of the Stanford family). Because of the collection's historical interest and high usage, SULAIR has had the desire to create high quality searchable digital surrogates for many years. This desire had been thwarted over the years by an insufficiently robust software and production infrastructure.

Beginning in 2004, renewed efforts for this collection have begun to bear tangible fruit: the digitized version of the Stanford Historical Photograph Collection has now been released in a beta version of approximately 500 images, and is being served as a Luna Insight image database. The production workflow put in place this year should enable completion and publication of the entire collection within a year or two.

<http://library.stanford.edu/depts/hasrg/hdis/aboutinsight.html>

Visual Resources Collection: Art 2 (Asian Art)

The Stanford Art Department Visual Resources Collection (VRC) has created an initial (alpha) version of one of the major components of its slide library, the Art 2 collection, which focuses on Asian art. The collection will use Luna Insight software to deliver its images for classroom use. This is the first Insight image collection created at Stanford outside the Libraries by an "independent" collection owner. Its inclusion in the general Stanford collections will greatly increase the potential use of the collection, beyond just the primary clientele of the Art Department; likewise, the suite of classroom presentation tools available in Insight will better serve the faculty and students of that department.

<http://vrc.stanford.edu/>

Novels of the Irish-American West

The Stanford Humanities Lab and SULAIR have jointly created a detailed author-title database and preliminary digital edition of largely forgotten novels of the Irish-American West. The database includes a large number of searchable fields specific to the situation of Irish-American writers and their work, including geographic locations and settings, biographical and bibliographic information, and abstracts. The full-text component consists of XML-encoded texts of primary works by Irish-American authors living west of the Mississippi, served via SULAIR's locally developed full-text workhorse, which is a PAT-based engine for web delivery of SGML.

<http://shl.stanford.edu:16080/IAW/>

The Fairchild Chronicles

A three-hour digital video documentary on Silicon Valley pioneer Fairchild Semiconductor is now available for sale. The Fairchild Chronicles is based on SULAIR's archival project "Silicon Genesis", a series of video oral histories of Silicon Valley. This is one of several Stanford collections on the history of Silicon Valley.

Through interviews with the people who made it happen, the Fairchild Chronicles tells the story of the company that invented the integrated circuit, describing the events that spawned the first generation of Silicon Valley technology companies. The DVD was co-produced by Rob Walker from Walker Research Associates of Menlo Park, and Kevin Bomberry from Panalta, Inc. of Palo Alto. It is available for \$39.95 from Panalta, Inc. 250 Emerson Street, Palo Alto CA 94301. All revenues go to SULAIR to continue chronicling the history of the semiconductor industry.

<http://silicongenesis.stanford.edu/>

<http://www.thefairchildchronicles.com/>

GATT Digital Library

With the support of an Institute for Museum and Library Services grant--and in close collaboration with the World Trade Organization (WTO)-- SULAIR has digitized and created an Internet presence for records of the WTO's predecessor, the General Agreement on Tariffs and Trade (GATT).

The GATT Digital Library website provides access to over 30,000 public documents and publications produced by this important international governmental organization during 1947-1994 as well as a substantial array of interpretive resources related to the organization's history. The website enjoyed a "soft-launch" in early March for the purposes of load and functionality testing as well as user feedback. A formal public announcement is planned for April in joint communiqué with the WTO. As the SULAIR/WTO collaboration evolves, the two organizations intend to expand significantly the scope of GATT-related public content made available over this site to include additional derestricted documents and archival materials.

<http://gatt.stanford.edu/>

B. Services

Lux

SULAIR is pleased to announce that the first phase of development on Lux, a full text search engine for homogeneous collections of XML, is complete. The software is written in Java, and uses the Apache Group's Lucene library as its underlying search implementation. Originally created to support the full text and metadata searching capabilities of the GATT Digital Library (see above in Collections), Lux allows librarians to create a searchable index from any collection of well-formed XML, without writing a

line of Java. The basic distribution includes Java libraries for searching Lux indexes, and a J2EE web application, built using the Struts framework, for searching Lux collections. SULAIR plans to use Lux to support the delivery of a variety of full-text collections. This coming summer, SULAIR plans to release the code for Lux to the open-source community, and we encourage our peers to use and further develop the software.

Find It @ Stanford University

Stanford University Libraries & Academic Information Resources (SULAIR) has completed implementation of the resource-linking technology SFX from Ex Libris. SULAIR was one of the first institutions to implement the new Version 3 of SFX. Initial installation was in mid-December 2004, and even with the subsequent two-week holiday closure, the target date of February 1, 2005 for the public roll-out of SFX was met. The implementation was achieved through the joint work of staff from the Acquisitions Department, the Digital Services Group's Systems Team, and Collections and Services. Shortly after the public roll-out, SULAIR also joined the first group of institutions testing the Google Scholar SFX pilot.

http://library.stanford.edu/catdb/help/sfx_faq.html

Stanford Grokker

In the Fall of 2004, the Stanford University Libraries and Academic Information Resources (SULAIR) teamed with Groxis to provide a customized version of their Grokker tool for the Stanford community. Grokker is an innovative research and information management tool that simultaneously searches many data sources, and presents results in a topically organized, visual map. Already, nearly 2000 Stanford faculty, staff, and students have downloaded the Stanford Grokker tool to their personal computers. Grokker is also available on public computer clusters and library kiosks throughout Stanford.

Grokker presents search results in a topically organized visual map, rather than in the long list of results typically provided by most search engines. Grokker's innovative mapping enables users to identify quickly and save relevant and valuable information, and to discover relationships among results. Grokker provides a form of federated searching by allowing users to search several resources simultaneously. The publicly available version of Grokker searches the Web, Amazon, and personal or shared hard drives. SULAIR staff have worked closely with Grokker to develop a customized Stanford Grokker that searches both publicly available resources and Stanford owned or licensed resources. The current version of Stanford Grokker provides a single point of access to Socrates (the Stanford library catalog), HighWire Press, Expanded Academic ASAP, Academic Search Premier, IEEE Xplore, RLG Union Catalog, the Library of Congress, and the Web. SULAIR and Groxis continue to work together to add new features to Grokker and to expand the number of research sources that Grokker can search.

http://library.stanford.edu/about_sulair/special_projects/stanford_grokker.html
www.grokker.com

The Sakai Project at Stanford

Stanford University has joined forces with three other institutions, the University of Michigan, Indiana University and MIT, to develop the next generation of course management tools. This landmark venture, called the Sakai Project, aims to create open-source course management tools and related software for the higher education community. It is being launched with a grant from the Andrew Mellon Foundation, with a commitment of resources and adoption from the core institutions that will swiftly integrate and synchronize the educational software.

Each of the partners in this consortium is contributing the work done on internally developed course management systems to create a new set of products that encompasses the best features of the individual efforts. The pre-integrated work products developed by the Sakai Project will greatly reduce the implementation costs of one or more of these tools at any institution. By synchronizing efforts, the four institutions are able to deliver more value to their own campuses than any one would by working alone. In addition, dozens of colleges and universities have joined the Sakai Educational Partners Program. Active pilots of the Sakai software are underway at several of the schools, and many more plan to adopt Sakai in the coming year.

The Sakai Project will provide Stanford with the next version of CourseWork, the popular course management system in use by thousands of Stanford faculty and students each quarter. In addition to all the features CourseWork now offers, the new environment will include tools to support project teams and other groups of people, and will have many new features as well. The new version of CourseWork will be tested at Stanford in the next academic year, as new features are developed, and will replace the current version of CourseWork.

<http://aboutcoursework.stanford.edu>

<http://www.sakaiproject.org>

C. Systems

The LOCKSS Program

The LOCKSS Program is continuing to build on its successes while looking to the future. A growing number of institutions are running LOCKSS machines with an increasing number of titles available. New software is released approximately once every six weeks. The system is a proven viable solution to addressing the risk of libraries losing their ability to own, access and preserve digital content. A vested community of partners is starting to form, and an infrastructure being built that can sustain and grow the program.

<http://www.lockss.org/about/users.htm>

<http://www.lockss.org/about/titles.htm>

In response to many requests for a simple demonstration of the capabilities of the LOCKSS system, we published the LOCKSS Winter 2005 Card. The Card contained a movie of the LOCKSS team, an excel spreadsheet, LOCKSS java software, and many other file formats. The card was available during February and March. It has now disappeared from the web. Fortunately, most of the LOCKSS machines around the world collected and preserved it. The readers at these institutions have perpetual access to this content. This simple exercise demonstrates the basic capabilities of the LOCKSS system:

- Content remains visible after it disappears from the publisher
- Access to preserved content is transparent - the Card will be visible via LOCKSS machines around the world at its original URL.
- The system is format agnostic - the Card includes a wide range of formats (HTML, PDF, Quicktime Movie, Microsoft Excel, gif, JPEG, XML, Java source, Java JAR files)

The LOCKSS system has designed and tested an initial implementation of format migration for Web content that is transparent to readers, building on the content negotiation capabilities of HTTP. This capability was demonstrated at a NARA workshop, November 2004, and appears to be the first time that a production digital preservation system has demonstrated transparent format migration of live content collected from the Web for end users.

<http://www.dlib.org/dlib/january05/rosenthal/01rosenthal.html>

The LOCKSS Alliance is a membership organization of those committed to advancing the LOCKSS Program to its next stage of evolution. The LOCKSS Alliance Board is finalizing details about membership fees, benefits and services, as well as governance and organization. The goal is to create a vibrant community of LOCKSS users that will share program costs and take full advantage of member benefits, including the leverage that a group of like-minded institutions can have on the marketplace.

<http://www.lockss.org/alliance/alliance.htm>

High-Capacity, Standards-Based Production-Repository-Delivery Workflow

In connection with work on the Stanford Historical Photograph Collection, created and digitized at Stanford and delivered via Luna Insight (see above, under Collections), we have developed a high-capacity, standards-driven production workflow, which has made possible the beta release of this complex collection, and which will become the basis for at least one predictable and dependable pipeline for creation of large image databases at Stanford in the future. While the workflow from production to preservation to delivery is far from seamless, we have been striving to make any necessary seams as smooth as possible. The basic workflow and associated technologies are these:

- A generalized, standards-based SULAIR metadata set for descriptive (Dublin Core-based), technical and administrative metadata, supported by the use of METS packaging
- A proprietary scanning and quality-control workflow program, which collects and binds this SULAIR standard metadata and image data at capture time, and stores it as database objects
- A script which ingests archival units of these metadata and images into the Stanford Digital Repository
- A script which exports these archival units from the Repository as XML METS packages containing MODS-encoded descriptive metadata for denoting the collections and the items within them, as well as mapping for the associated image files. The metadata also include a durable Repository ID to enable bi-directional interoperability between the Repository and the delivery system
- An XSL transformation which extracts the metadata from selected descriptive and technical fields from each object and inserts it as relational table data into a proprietary (Insight-aware) environment for further processing and delivery

Although this complex workflow is currently tailored for a particular set of metadata elements describing a particular collection, as well as for particular capture, repository and delivery systems, we are working to make it generalizable for different collections described with different subsets of our descriptive metadata set, and for different capture processes and workflows. A key component of the process is its ability to move data smoothly from a proprietary capture system to a proprietary delivery system -- and it is the standards-based middle portion of the process that makes this possible.

Stanford HighWire Update

In 2005, HighWire Press®, Stanford's electronic journal hosting service for the scholarly publishing community, celebrated its tenth anniversary.

As of April 2005, HighWire assists in the online production of 850 journals. From the start, HighWire worked with societies focused on research in the life sciences and medicine, the kind of journals that continue to be among the highest-impact titles. Starting in 2004, through a series of new publishing partnerships, HighWire has expanded its scope to include over 400 Social Science and Humanities journals. Within this broader context, HighWire continues to explore the best ways to support the provision of scholarly information and the scientific communication process.

In early 2005, HighWire helped create and launch GeoScience World, a new project by a group of leading geoscientific organizations, which offers a comprehensive Internet resource portal for research and communications in the geosciences.

For its exemplary work in online hosting and service, HighWire was recipient of the 2003 Association of Learned and Professional Society Publishers (ALPSP) Award for "Service to Not-for-Profit Publishing".

HighWire doesn't own or sell the content hosted on its website, but they do support librarian colleagues in other ways: helping their society partners 'hear' the librarian's voice on current issues, as well as enabling and encouraging publishers to free-up back issues. With nearly 850,000 free articles and counting, HighWire continues to be the largest repository of free full-text science available online.

There are a number of free searching and alerting services for end-users, including a customized home page, with user-selected links to "my favorite journals" and "my alerts"; and a 55,000-topic list of taxonomy categories establishing subject links directly to individual articles. With its portal, HighWire offers a series of librarian tools to tackle some common management tasks, such as multi-journal usage reports – in both detailed, and COUNTER compliant formats – and IP address maintenance across publishers.

Some of the HighWire-affiliated publishers are participating in a program called "Shop for Journals", where librarians can quickly find out how much a subscription will cost for their type of institution. In addition, there is now a simple FTP site where metadata (headers and abstracts) can be downloaded to registered users (the same feed as is provided to PubMed) to assist in searching. And, there is a feature on the "For Institutions" section of the HighWire website which makes looking up ISSN numbers, publisher addresses, and other FAQ's easy for librarians.

<http://highwire.stanford.edu>

www.geoscienceworld.org

<http://highwire.stanford.edu/institutions>

II. Projects and programs

A. Projects

Remote Hosting of Local Collections Pilot

SULAIR has been working with ARTstor to deliver one of its image collections, Antiquarian Maps of Africa, via the ARTstor interface. This collection is currently available worldwide from Stanford via the Luna Insight interface; the ARTstor pilot, when complete, will offer the same collection to Stanford users among its rich art resources.

While we are still far from the goal of complete image collection interoperability with this pilot project, ARTstor does offer one possible solution to the problem, and Stanford has been a participant in this ARTstor initiative.

<http://www.artstor.org/info/news/Hosting.jsp>

NDIIPP Award to Archive Geospatial Data Given to SUL and UCSB

The Library of Congress has selected Stanford and the University of California, Santa Barbara to develop one of eight major national initiatives for digital information preservation. The Stanford/UCSB team will form a National Geospatial Digital Archive (NGDA) with the goal of designing an infrastructure to collect and provide for long-term preservation of digital materials across the spectrum of geographic formats. The born-digital materials to be collected and preserved will range from LANDSAT imagery to other cartographic content from university, corporate and government resources, as well as Web sites. The Repository will preserve content vital for the study of history, science, environmental policy, urban and population studies, census construction and analysis, and other fields requiring U.S. geospatial information.

Once established, the Archive will allow Stanford Library staff to offer archival solutions to other organizations and individuals that have produced important digital geographical resources considered to be at risk. The University of Washington, the California Spatial Information Library, and noted collector and digital publisher David Rumsey are among those that have agreed to contribute digital resources to the Archives.

The Library of Congress announced the award of nearly \$3M to the Stanford/UCSB partnership in Washington on 30 September, 2004, culminating a nearly two-year effort to begin putting in place a series of cooperative networks of digital repositories. Julie Sweetkind-Singer, head of the Branner Earth Sciences Library and Map Collections and GIS/map librarian, will be the lead for the Stanford team, which will include up to a dozen individuals at any time during the three-year project.

www.ngda.org

B. Programs

Digital Services Group

Recognizing the ongoing need to address its readiness for the digital future, Stanford University Libraries and Academic Information Resources (SULAIR) reorganized to create the Digital services Group (DSG). The DSG operates the technology infrastructure for the libraries as well as for services directly used by the Stanford community. It also produces and supports applications for libraries and instructional settings. DSG projects build upon purchased systems, locally developed tools, and increasingly, public domain software. In these areas, it provides significant local integration and enhancement, plus ongoing support of these applications and their technical environments.

Specifically, the Digital Services Group supports enterprise applications; digital library projects for capture, description, storage, organization and access to information; Unicorn and the related library management tools; the nascent Stanford Digital Repository; and various academic applications. The DSG team is experienced in the implementation, integration and enhancement of purchased systems in the library realm, adding value to

the packages for use at Stanford. The same is expected to be true with incoming open-source applications such as DSpace, Greenstone, and ePortfolio.

Moving forward, the DSG has the following organizational goals:

- Build the technology infrastructure and collections that will comprise the digital library of the future.
- Stimulate and focus innovation and assessment: in context, with goals and evaluation processes
- Develop economies of scope and scale: consolidate expertise and projects; eliminate redundancy, enhance shared expertise and multi-use technologies and methodologies;
- Identify total cost and resources needed for development and ongoing support of digital collections
- Develop clear and supportable product and project life cycles
- Provide decision-making data and analysis to inform decisions and priorities.
- Provide leadership and stimulus in the Stanford technology community to develop new capabilities and technologies; inform technology directions and frameworks <http://library.stanford.edu/depts/dsg/>

III. Specific Digital Library Challenges

The Stanford Center for Excellence in the Knowledge Enterprise

This partnership between Sun Microsystems and Stanford University Libraries and Academic Information Resources (SULAIR) represents Sun's and SULAIR's continued commitment to innovation, collaboration with leading academic institutions, and the pursuit of new advances in networked education and federated information.

The overall objectives of the Center of Excellence (CoE) are to:

- Create best practices and models for the preservation and dissemination of information in academic research institutions;
- Lead both academia and the publishing industry in creatively addressing the related issues of access and preservation;
- Establish norms for institutional output storage policies and practices by example.

SULAIR is in the business of selecting, collecting, describing, disseminating, publishing, archiving and making accessible information for teaching, learning, and research. The Center of Excellence will further the state of the art in each of these functions in the digital environment. Specifically, we are deeply and fundamentally interested in:

- The technology of capture, description, delivery, storage and preservation of information on a massive scale;

- Growing the market within the academic industry for integrated hardware and software solutions suitable for local digital repositories, mirroring agreements, interoperating repositories, course management systems, distributed persistent digital caches, etc.;
- Creating and maintaining new kinds of communities among scholars in academic publishing.
http://library.stanford.edu/about_sulair/special_projects/Stanford_Sun_Center_Excellence.htm

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

- Keller, Michael A. "Casting Forward; collection development after mass digitization," March 2004.
http://digital.casalini.it/retreat/2004_docs/KellerMichael.pdf
- Keller, Michael A. "Commentary on NIH Notice on Enhanced Public Access to NIH Research Information," November 2004.
http://library.stanford.edu/staff/pubs/MAK_to_Zerhouni_041115.pdf
- Keller, Michael A. "Digitizing Literatures: Bringing the Library to Where People Search for Information," February 2005.
<http://library.stanford.edu/staff/pubs/Feb2005FacSenRept.ppt>
- Keller, Michael A. "Gold at the End of the Digital Library Rainbow: Forecasting the Consequences of Truly Effective Digital Libraries," December 14, 2004.
http://library.stanford.edu/staff/pubs/Forecasting_Effective_Digital_Libs.pdf
- Keller, Michael A. "Orphan Works and Research Libraries and Archives: A letter to the U.S. Copyright Office," March 18, 2005.
<http://library.stanford.edu/staff/pubs/OrphanWrksCopyrightLetter.pdf>
- Keller, Michael A. "Reconstructing Collection Development," Keynote address at the XXIV Annual Charleston Conference, Issues in Book and Serial Acquisition.
- Kott, Katherine. "Managing in Anxious Times: Thoughts about Neurobiology and Evolutionary Biology," A paper presented by Katherine Kott at the University of Arizona Library's Living the Future 5 conference, April 15-17, 2004.
- Lowood, Henry. "A Brief Biography of Computer Games." To appear in *Playing Computer Games: Motives, Responses, and Consequences*, ed. Peter Vorderer and Jennings Bryant. (Lawrence Erlbaum Associates, exp. 2005)
- Lowood, Henry. "Electronic Game." *Encyclopædia Britannica*. 2004. *Encyclopædia Britannica Online*. 16 July 2004. < > With side-bars on "Zork,"

- "Pac-Man," "The Legend of Zelda," and "DOOM." To appear in 2004 print edition. <http://search.eb.com/eb/article?eu=1566>
- Lowood, Henry. "Gosu Game Studies" (Hard Core Column). DIGRA-Online. (Digital Games Research Association) 11 Jan 2005. <http://www.digra.org/article.php?story=20050111124812120>
 - Lowood, Henry. "High-Performance Play: The Making of Machinima." To appear in *Computer Games and Art: Intersections and Interactions*, special issue of *Anomalie*, ed. Grethe Mitchell and Andy Clarke.
 - Lowood, Henry. "It's Not Easy Being Green: Real-Time Game Performance in Warcraft." In preparation for: *Videogame/Player/Text*, eds. Barry Atkins and Tanya Krzywinska. (Manchester Univ. Press, exp. 2006).
 - Lowood, Henry. "The Obstacle Course: Documenting the History of Military Simulation," in: *America's Army PC Game: Vision and Realization*, ed. Margaret Davis (Monterey, Calif: U.S. Army and MOVES Institute, 2004): p. 18.
 - Lowood, Henry. "Real-Time Performance: Machinima and Game Studies." To appear in: *Journal of the International Digital Media and Arts Association* (March 2005).
 - Lowood, Henry. "Technology and Leisure," "Computer and Video Games" and "Computers-Personal." *Encyclopedia of 20th-Century Technology*, ed. Colin Hempstead. (Routledge, 2004).
 - Lowood, Henry. "Video Games in Computer Space: The complex history of Pong." In preparation for: *Ludologica Retro, Volume 1: Vintage Arcade (1971-1984)*, eds. Ian Bogost & Matteo Bittanti (Edizioni Unicopli, exp. mid-2005).
 - Lowood, Henry. "Virtual Reality." In preparation for *Encyclopedia Britannica*, due April 2005.
 - Worthey, Glen. "Digital Delivery of Interlibrary Loan and Democratic Digital Collection Development at Stanford." *Against the Grain*, v.16, no.4, pp.48-52.