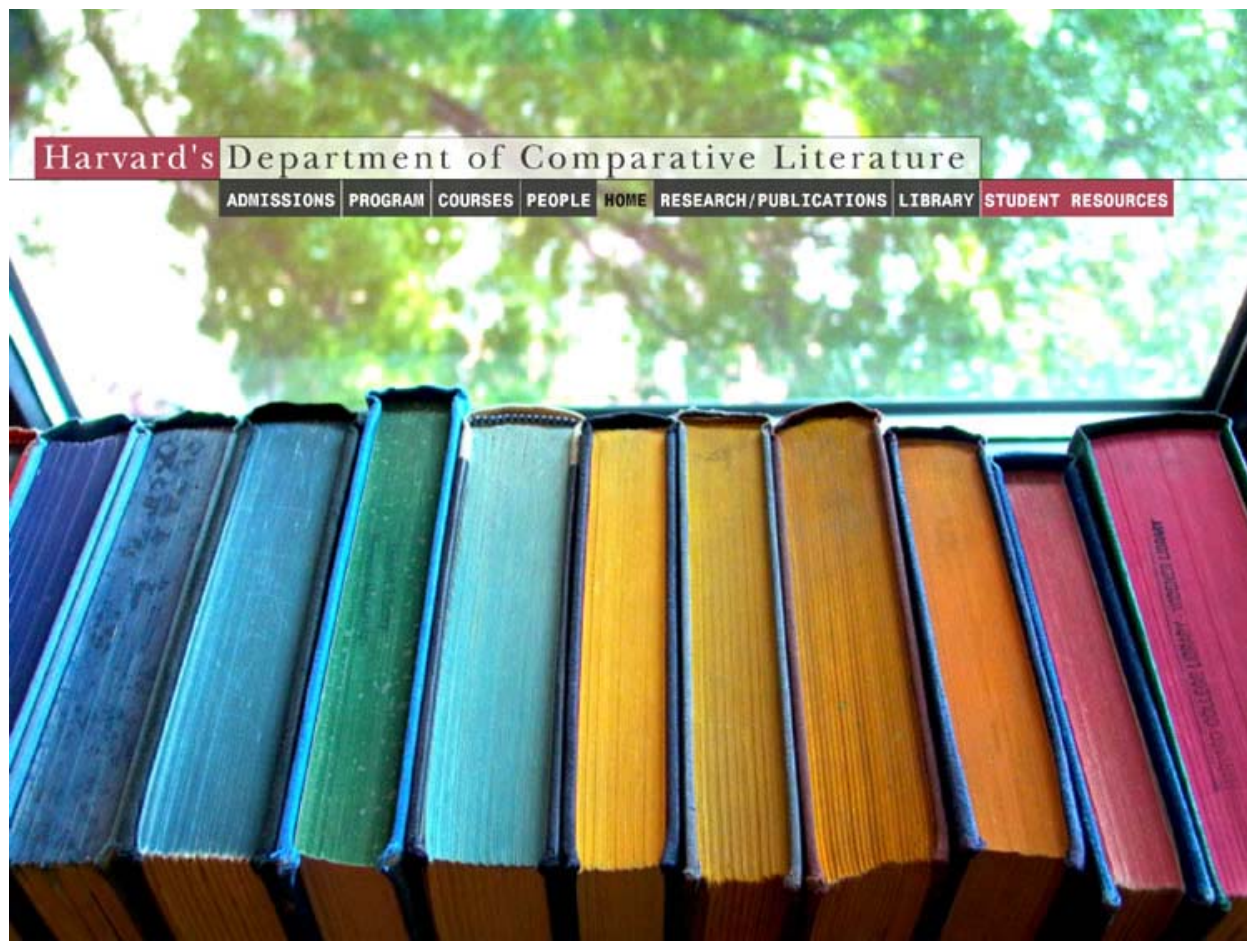


A-Sites: Archived Harvard University Web Sites

A WAX Collection of the Harvard University Archives



Final Report
November 6, 2009

Project Manager
Skip Kendall, Senior Electronic Records Analyst/Archivist

Introduction

At the core of the Harvard University Archives' mission is the preservation and provision of enduring access to historical records. Increasingly, such records are found in electronic form. In response, the Archives (HUA) must expand its capacity to acquire and maintain such born-digital content. Over the past 15 years, mission statements, annual reports, academic program offerings and events, and faculty profiles have moved from printed reports, catalogs, and newsletters to web sites.

HUA undertook this project to advance its experience with and capacity to acquire and preserve born-digital history. As a significant place to start, the 40 web sites of the Degree-Granting Departments and Committees of the Harvard University Faculty of Arts and Sciences were selected. With the completion of the project, some of this historical record has been acquired and a model has been created for future acquisitions.

WAX is a system designed to manage the harvesting of web sites, perform quality assurance checks on the resulting harvests, and transfer acceptable harvests to the Digital Repository Service. From the curatorial perspective, WAX has two parts: WAXi – the interface through which curators manage sites, harvesting, and harvests, and the public interface, WAX's window to the world.

Through this project, the Archives has explored, and has achieved, a greater understanding, of the issues surrounding the preservation of web sites, including appraisal considerations, cataloging requirements, harvest frequencies, workflow impact, budgeting, and intellectual property issues. Beyond web sites, this experience is valuable as we approach the acquisition of other types of electronic content.

Participating Staff

Over the course of the project, more than half of the Archives' staff worked on this project. All were involved in quality assurance work, which involved reviewing harvested web sites for problems with display, structure, or content. A smaller group were involved in project planning and policy and procedural considerations. Another group, primarily Collections Services staff, put in significant time doing descriptive work. The following people worked on the project:

Project Management

Skip Kendall

Advisory Support

Kate Bowers, Collections Services Archivist

Robin McElheny, Associate University Archivist for Collections and Public Services

Barbara Meloni, Public Services Archivist

Megan Sniffin-Marinoff, University Archivist

Quality Assurance

Michael Austin, Processing Archivist

David Best, Associate University Archivist for Records Management Services

Kate Bowers, Collections Services Archivist

Pat Donoghue, Descriptive Data and Processing Archivist

Dominic Grandinetti, Processing Archivist

Claudia Holguin, Administrative Fellow
Virginia Hunt, Associate University Archivist for Collection Development
Jennifer Jacobsen, Collection Development/ Appraisal Archivist
Juliana Kuipers, Special Materials Cataloger/Processing Archivist
Robin McElheny, Associate University Archivist for Collections and Public Services
Barbara Meloni, Public Services Archivist
Julie Revak, Processing Assistant

Descriptive Metadata

Kate Bowers, Collections Services Archivist
Julie Revak, Processing Assistant

Public Interface

Barbara Meloni, Public Services Archivist
Kate Bowers, Collections Services Archivist
Robin McElheny, Associate University Archivist for Collections and Public Services
Megan Sniffin-Marinoff, University Archivist

Special notice for significant time and thought spent on the project should go to Kate Bowers, Robin McElheny, Barbara Meloni, and Julie Revak.

Office for Information Systems

Special thanks is also due to the following staff at OIS, without whom nothing would have happened.

Chris Vicary, Software Engineer
Andrea Goethals, Digital Repository and Preservation Manager
Wendy Gogel, Digital Projects Program Librarian
Robin Wendler, Metadata Analyst
Vitaly Zakuta, Digital Project Librarian



Research

Directory

[Individual Faculty Pages](#)

[Group Pages](#)

[Research Areas](#)

[Publications](#)

[Affiliated Institutes and
Departments](#)

Research Areas

Atmospheric: [Anderson](#)
[Friend](#)

Bioorganic: [Knowles](#)
[Liu](#)
[MacBeath](#)
[Schreiber](#)
[Verdine](#)
[Shair](#)
[Saghatelian](#)

Biophysics: [Shakhnovich](#)
[Xie](#)
[Zhuang](#)

Chemical Biology: [Kahne](#)
[Liu](#)
[MacBeath](#)
[Schreiber](#)
[Shakhnovich](#)
[Verdine](#)
[Shair](#)
[Saghatelian](#)

Inorganic: [Holm](#)
[Whitesides](#)

Materials: [Aspuru-Guzik](#)
[Friend](#)
[Gordon](#)

Figure 1 – Sample page from the Chemistry Department: links to faculty profiles

Project Timeline and Summary

The WAX project involved seven phases. The first phase began before the first project meeting as we selected web sites to capture. Next, the curatorial participants began discussions with the Office for Information Systems (OIS) on the functionality of the system to be developed. We then worked with the Office of General Counsel on the significant issue of intellectual property. Actual web harvesting came next, followed by testing and further development of the new system. Finally, descriptive metadata was created and the public interface was developed and tested.

Web site selection

Work began in October 2005 with the selection of web sites to include in the project. For HUA, the degree-granting departments and programs of the Faculty of Arts and Sciences seemed an ideal choice: they constituted a clearly-defined group of sites, had enough content and technical variety to be a good test case, and are of core historical importance to the University. The initial list of sites included 32 departments and 8 degree-granting programs. All sites were publicly accessible and maintained on Harvard servers.

System planning

In March 2006, the Archives began meeting with the other project partners (Office for Information Systems, Schlesinger Library, and Reischauer Institute) to identify goals for curatorial and end-user functionality, including

- parameters for setting up harvests
- harvest reviewing
- ability to distinguish harvested web sites from live versions
- standard web navigation
- full-text search

With the establishment of these functional goals, OIS proceeded to design and build the web archiving system, to consist of curatorial and public interfaces and extensive behind the scenes functionality.

Intellectual property review

In April, Archives staff began meeting with the Office of General Counsel (OGC) to discuss intellectual property issues. No significant problems were anticipated for two reasons: the selected web sites were already publicly accessible (harvests are conducted from non-Harvard servers to avoid restricted to Harvard content) and harvests would be targeted exclusively to information stored on Harvard servers. Additionally, since the system does not yet have the capability of harvesting password-protected content, no limited-access information could be harvested. There turned out to be little on the servers for which the intellectual property rights did not belong to the University. OGC and the Archives agreed that such content would either not be harvested or would be removed upon request. Over the next several months, work continued on the system's functional requirements and images of the proposed system interface were available by October 2006.

Web harvesting

While OIS started building the WAX system, web harvesting began using a temporary system, the Web Curator System (WCT). WCT is an open-source tool for managing web harvesting. The first WCT harvests were conducted in February 2007 and continued through August 2007. Although WCT did not meet all of the project's functional requirements, it enabled us to learn practical lessons about harvesting and quality assurance (QA) while obtaining harvests that would eventually become part of the permanent collection.



Figure 2 - Selection from the Music Department home page (2007)

System development and implementation

WAXi, the curatorial interface for the new system, became available in September 2007. As expected, WAXi proved to be a significant improvement over WCT. In WAXi, curators can enter root URLs ("seeds") for selected web sites, along with summary descriptive information about the sites. Curators can then set the scope and frequency of harvests. WAX allows for fairly granular limitations on harvesting, so that specific sections of web sites can be avoided. Post-harvest quality assurance is conducted in WAXi, as well. Thanks to custom development by OIS, harvesting in WAX had a much higher success rate than in WCT.

The introduction of WAXi initiated a long period of harvesting web sites and reviewing them, both of which are now part of the Archives' collection development and collection management routines. Regular harvests were scheduled for April and November in order to capture information added specifically for the Fall and Spring terms but before changes had been introduced for the next term. When problems arose with harvests, adjustments were made, either to the system or the harvest parameters, and extra harvests were scheduled.

As mentioned earlier, a number of Archives' staff members were involved in the QA of harvested web sites. Reviewers were assigned particular web sites and reviewed them as time

allowed. QA results were communicated to the project manager, who kept track of which harvests had been reviewed and what problems were found. Reports of problems were forwarded to OIS for action.

Once a harvest has been determined to be sufficient for long-term retention, the files are transferred, through WAXi, to the Digital Repository Service (DRS). In the DRS, web sites are assured of long-term physical preservation (preserving the bits) while OIS continually works to combat format obsolescence. Format obsolescence - when a file can no longer be read by current software - is a significant problem in the preservation of electronic records. While no one has figured out a perfect solution, the DRS compares well with other digital preservation systems in use.

System development and QA of harvested web sites continued until February 2008, with the introduction of the "collections" functionality for curators and preliminary public interface.

The system's collections functionality was designed to give curators the ability to group web sites based on shared content: a web site found in each of two different makes the collections related. HUA has initially opted to implement this to create the appearance of hierarchical collections: A-Sites, which represents all of the Harvard web sites that HUA will acquire, and FAS, a sub-collection, which consists of the sites that HUA is currently harvesting from the Faculty of Arts and Sciences. In the future, we may want to create other collections within A-sites, such as Registrars, which could include the FAS Registrar's web site with registrar's web sites from the other Harvard faculties. The flexible relationships available in WAX offer the promise of new ways to present archived web sites and to facilitate public access and discovery.

Archived copy of www.ves.fas.harvard.edu/
 Why was this page archived?
 Archived on: March 02, 2007 13:55:00 Archived versions of this web page
 Collection: A-Sites: semi-annual captures of selected Harvard web sites
 Archived by: Harvard University Archives, Harvard University Library

Harvard University Library
 Web Archive Collection Service

Cite This Resource

ves
 DEPARTMENT OF VISUAL AND ENVIRONMENTAL STUDIES HARVARD UNIVERSITY

Carpenter Center
 24 Quincy Street
 Cambridge Massachusetts
 02138-2604
 617-495-3261
 617-495-8197 fax

about VES for VES Concentrators courses academic calendar faculty events/exhibitions

Upcoming at the Carpenter Center:
 Carpenter Center Lecture by Victor Burgin
 "The Responsibility of the Artist"
 Thursday, March 8, 6 pm
 Reception for the exhibition TELL ME!...a secret...
 Thursday, March 15, 5:30-6:30 pm
 Gallery talk with Hans Tutschku
 Friday, March 16, 5 pm
 Carpenter Center Lecture by Carol Dunham
 cosponsored by the Harvard Advocate
 Thursday, April 12, 6 pm
 eflux screenings w guest curators

Hans Tutschku
 TELL ME!...a secret...
 March 8-April 13, 2007
 Main Gallery

EVR: E-flux Video Rental
 through April 13, 2007
 Sert Gallery
 Monday-Saturday 1-8 pm

The Little House
 Installation by Victor Burgin
 March 8-April 13, 2007
 Sert Gallery

© 2006-07 President and Fellows of Harvard University Photo: Timothy Pittman

Use of WAX is subject to our Terms of Use Questions and Comments Help

Figure 3 - Visual and Environmental Studies home page as seen through the WAX public interface (2007)

User interface development

OIS formed a working group in order for all project participants to work on the public user interface (UI) with Barbara Meloni representing HUA. Significant features decided upon by the working group are free-text search, framed harvests (to ensure a distinction from live sites), browsing, citation support, and many other features. Their product stands as a very useful web site in its own right.

A key element to the success of the UI was the creation of descriptive metadata for the sites. As work was done on the public interface, descriptive metadata was created for sites and collections, with Kate Bowers doing the lion's share of the work. Because the systems are cross-referenced, simultaneous data entry was necessary in WAX and HOLLIS. Users accustomed to research in HOLLIS will now find these harvested web sites while those entering through WAX can find further information on the organizations represented by navigating to HOLLIS records.

Descriptive metadata

In order to maximize public access and discovery, HUA staff decided to catalog both HUA collections (A-Sites and Faculty of Arts and Sciences) and each "seed" web site in HOLLIS. HOLLIS provides the controlled names and indexing terms essential to prudent management of electronic resources as an important part of our overall holdings. HOLLIS data is also indexed by Google books and WorldCat, making our resources more prominent to searchers of the World Wide Web, now almost universally targeted as the first stop in preliminary research by Harvard students and faculty.



[Spring 2007 Office Hours](#)

[Course Catalog](#)

[The Virtual Staff Assistant](#) (for current faculty, students, and staff only)

Announcements

New Spring Courses

Japanese Literature 124. The Tale of Genji in Word and Image - (New Course)
 Catalog Number: 2181
 Melissa M. McCormick
 Half course (spring term). Th., 2-4. EXAM GROUP: 16, 17

Classroom location: Canaday Hall, B Entry

This undergraduate seminar introduces students to The Tale of Genji, often called the world's first novel, authored by the court lady Murasaki Shikibu around the year 1000 CE. In addition to a close reading of the tale, topics for examination include Japanese court culture, women's writing, and Genji's afterlife in painting, prints, and the Noh theater. The class will include visits to art collections and the viewing of a Noh performance.

New Spring Course Offered by Visiting Faculty

Yang Lu, Lecturer on East Asian Languages and Civilizations, Fall and Spring

Chinese History 248. Introduction to Archaeology of Medieval China (400-1000) - (New Course)
 Catalog Number: 0948
 Yang Lu
 Half course (spring term). W., 1-4. EXAM GROUP: 6, 7, 8
 This seminar surveys major archaeological finds in the past 50 years that help shed light on life in medieval China. The materials include finds excavated outside of China proper, such as in Inner and Northeast Asia.
Prerequisite: Reading knowledge of modern Chinese required. Reading knowledge of modern Japanese recommended, but not required.

News

New Books



Figure 4 - Announcements of the East Asian Languages Department (2007)

Lessons Learned

Communication is critical

Good communication between the curators and the IT professionals at OIS was critical to the success of the project. There were many extended discussions occasioned by misunderstandings of concepts and terminology that one group thought would be obvious to the other. Shared terminology between the archives world and the information technology world (with slightly different meanings) was one source of misunderstanding. Another was between library terminology and archives/special collection terminology, where differences in meaning were even more subtle.

The project required the full range of expertise found in the Archives

No matter their experience, no single person at the Archives could have adequately covered the areas of researcher expectations, metadata, web knowledge, and administrative necessities. This was truly a group effort.

The amount of time necessary to review a harvest is nearly impossible to estimate

The time required for QA work on a harvest, from initial review through the reporting and resolution of problems can take from 5 minutes to more than 10 hours. The complex variables that go into determining how long it takes to review a web harvest make general time estimates for reviewing very difficult to determine. These variables include the size of a site, the complexity of site design, and the degree of troubleshooting required to determine what went wrong when a harvest is not quite right. More QA experience is needed to provide reliable benchmarks for estimating ongoing staffing requirements.

Essentially, the system will always be in development

WAX does a good job of handling the many different ways there are to build web sites. At the same time, the Web is always changing and new methods of creating web sites are always being developed. For the Archives, this means that we will always have to be on the lookout for changes that break the current system. We cannot assume that, just because we successfully harvested a site last time, the new harvest will also be successful. For OIS, this means some level of ongoing development as the Archives, and other WAX participants, will always be sending them more problems to solve.

As we expand beyond the initial sites, the workload and storage requirements could grow dramatically

The web sites selected for this project are being harvested twice a year. Each round of harvests takes 3 gigabytes of storage space and significant staff time for review. As we add sites to the collection, we will not only be adding their storage and workload requirements, but will be biannually repeating those for sites already being harvested. These factors alone are likely to lead to exponential growth. Additionally, new sites require setup work and can require significant work with the site owners. Although we find it difficult to come up with exact estimates of the impact of this collection on HUA resources, we know that it will be significant.

Appendix

Numerical Summary (through 8 February 2009)

- 1 primary (A-Sites) and 1 non-primary (Faculty of Arts and Sciences) collection
- 42 department and program web sites, harvested from 48 different URLs (seeds)
- 147 harvests stored in the Digital Repository Service
- 6 harvests presumed successful but not yet in Digital Repository Service
- 111 Gigabytes harvested

Sites with associated seeds

- Faculty of Arts and Sciences departments
 - African and African American Studies
 - <http://www.fas.harvard.edu/~afroam/>
 - Anthropology
 - <http://www.fas.harvard.edu/anthro/>
 - <http://www.fas.harvard.edu/~bioanth/>
 - Astronomy
 - <http://cfa-www.harvard.edu/hco/astro/>
 - <http://www.cfa.harvard.edu/ast/>
 - <http://www.cfa.harvard.edu>
 - Celtic Languages and Literatures
 - <http://www.fas.harvard.edu/celtic/>
 - Chemistry and Chemical Biology
 - <http://www.chem.harvard.edu/>
 - Classics
 - <http://www.fas.harvard.edu/classics/>
 - Comparative Literature
 - <http://www.fas.harvard.edu/~complit/>
 - Department of Stem Cell and Regenerative Biology
 - <http://www.scrb.harvard.edu/>
 - Division of Engineering and Applied Sciences
 - <http://www.deas.harvard.edu/undergraduate/>
 - Earth and Planetary Sciences
 - <http://www-eps.harvard.edu/>
 - East Asian Languages and Civilizations
 - <http://www.fas.harvard.edu/ealc/>
 - <http://www.ealc.org/>
 - Economics
 - <http://www.economics.harvard.edu/>
 - English and American Literature and Language
 - <http://www.fas.harvard.edu/~english/>
 - Germanic Languages and Literatures
 - <http://www.isites.harvard.edu/icb/icb.do?keyword=k4326>
 - Government
 - <http://www.gov.harvard.edu/>

- History
 - <http://history.fas.harvard.edu/>
- History of Art and Architecture
 - <http://www.fas.harvard.edu/~hoart/>
- History of Science
 - <http://www.fas.harvard.edu/~hsdept/>
- Linguistics
 - <http://www.fas.harvard.edu/~lingdept/>
- Mathematics
 - <http://www.math.harvard.edu/>
- Molecular and Cellular Biology
 - <http://www.mcb.harvard.edu/>
- Music
 - <http://www.music.fas.harvard.edu/>
- Near Eastern Languages and Civilizations
 - <http://www.fas.harvard.edu/~nelc/>
- Organismic and Evolutionary Biology
 - <http://www.oeb.harvard.edu/>
- Philosophy
 - <http://www.fas.harvard.edu/~phildept/>
- Physics
 - <http://www.physics.harvard.edu/>
 - <http://library.physics.harvard.edu>
- Psychology
 - <http://www.isites.harvard.edu/icb/icb.do?keyword=k3007>
- Romance Languages and Literatures
 - <http://www.fas.harvard.edu/~rll/>
- Sanskrit and Indian Studies
 - <http://www.fas.harvard.edu/~sanskrit/>
- School of Engineering and Applied Sciences
 - <http://www.seas.harvard.edu/>
- Slavic Languages and Literatures
 - <http://www.fas.harvard.edu/~slavic/>
- Sociology
 - <http://www.wjh.harvard.edu/soc/>
- Statistics
 - <http://www.stat.harvard.edu/>
- Visual and Environmental Studies
 - <http://www.ves.fas.harvard.edu/>
- Degree-granting programs
 - Environmental Science and Public Policy
 - <http://www.espp.fas.harvard.edu/>
 - Folklore and Mythology
 - <http://www.fas.harvard.edu/~folkmyth/>
 - History and Literature
 - <http://www.fas.harvard.edu/~histlit/>

- Literature
 - <http://www.fas.harvard.edu/~litconc/>
 - <http://isites.harvard.edu/icb/icb.do?keyword=k38927>
- Social Studies
 - <http://www.fas.harvard.edu/~socstud/>
- Special Concentrations
 - <http://www.specialconcentrations.fas.harvard.edu/>
- Studies of Women, Gender, and Sexuality
 - <http://www.fas.harvard.edu/~wgs/>
- Study of Religion
 - <http://www.fas.harvard.edu/~csrel/>

Timeline

October 2005	Work begins on the Archives' proposal for participating in a web archiving project
March 2006	Project meetings begin
April 2006	Archives begins working with the Office of the General Counsel to discuss intellectual property and other legal issues
October 2006	System mockups of what would become WAXi became available
February 2007	Harvests begin using the Web Curator System (WCT)
August 2007	Final harvests using WCT are conducted
September 2007	WAX becomes available for harvest and quality assurance
November 2007	Archives conducts first harvests in WAX
February 2008	Public user interface and collections functionality become available
February 9, 2009	WAX is formally announced to Harvard's library community