Zone 1: A Rescue Repository for Digital Content

Library Lab Final Report, November 15, 2011

Submitted by:

Andrea Goethals, Digital Preservation and Repository Services Manager, Harvard Library Office for Information Systems (OIS)
Wendy Gogel, Digital Content and Projects Manager, Harvard Library OIS
Skip Kendall, Senior Electronics Record Analyst/Archivist, University Archives, Harvard University
Tom Rosko, Institute Archivist/Head, Institute Archives and Special Collections, MIT
Megan Sniffin-Marinoff, University Archivist, University Archives, Harvard University

Introduction

Zone 1 is a project with three parallel activities:

1. Development of a prototype rescue repository at Harvard
2. Re-thinking faculty papers - An in depth study by Harvard and MIT of one category of content expected to be deposited to a rescue or long-term preservation repository—faculty records (the report on this activity begins on page 7)
3. A series of discussions at Harvard and MIT of policies that would need to be addressed if the rescue repository were to become a production system at Harvard, or a similar system were implemented at MIT

Prototype rescue repository

Project summary

The project objective was to develop a prototype rescue repository that could be used by Harvard and other institutions with a similar need, such as MIT. It would provide temporary, secure storage for content that is not a good fit for other more specialized content repositories. This includes content:

- at immediate risk of loss, e.g. on degrading media such as magnetic tape
- with temporary value, e.g. for university records retention requirements or classroom use
- not yet supported by the existing repositories, e.g. learning objects that support coursework
- of undetermined long-term value, e.g. unprocessed collections
- identified as having likely permanent value by content contributors and researchers, but currently without a solution to preserve the content long-term

The repository would provide the minimum infrastructure needed to rescue and secure the digital content while decisions were being made about its longer-term disposition. Specifically the repository would provide:

- easy ways to deposit content into the rescue repository, through commonly-used software, platforms and Web sites
- a bit-level preservation[1] storage solution that keeps the content safe from unauthorized changes, deletions, media degradation, transfer errors and disasters
- a mechanism for content contributors and reviewers to recommend or propose the content for long-term preservation
- a mechanism for potential stewards (e.g., records managers and collection managers) to review and select content in advance of taking long-term preservation responsibility for it
- a mechanism to allow review by others (e.g., teachers or researchers) for potential reuse
- easy ways to extract content out of the rescue repository, for sharing, destruction, reuse or transfer into other repositories within the institution or elsewhere

The project was implemented through a collaboration between OIS and Berkman staff. It included:

- Sebastian Diaz, Berkman Center (manager of development)
- Laura Miyakawa, Berkman Center (project manager)
- Dan Collis-Puro, Berkman Center (advisor to developers, quality assurance, code reviews)
- Andrea Goethals and Wendy Gogel, OIS (functional requirements, providers of test data and metadata)
- Third-party developers managed by the Berkman team

Accomplishments

Within the first month of the project, Sebastian, Dan, Andrea and Wendy defined the functional requirements for the prototype. The requirements included:

- An identification of the core entities to model in the system: files, accounts, groups (for access control)
- The metadata to record in the system for files, accounts and groups
- Specification of the policies that would need to be configurable in the system (e.g. who can deposit content) and default settings for the prototype
- General system-wide requirements
- Activity-specific requirements for:
    - Account creation and management
    - Content and metadata management
    - Deposit of content into repository
    - Ingest of content into repository
    - Retrieval of content from repository
    - Review of content by potential long-term stewards
- Glossary of terms
- Items to defer for a potential follow-up Library Lab project

[1] There are two digital preservation levels: (1) bit-level keeps the file bits safe from changes; (2) full-level keeps the information usable over long periods of time even as technology changes – this requires a greater deal of organizational commitment as well as cost, effort and technical skill.

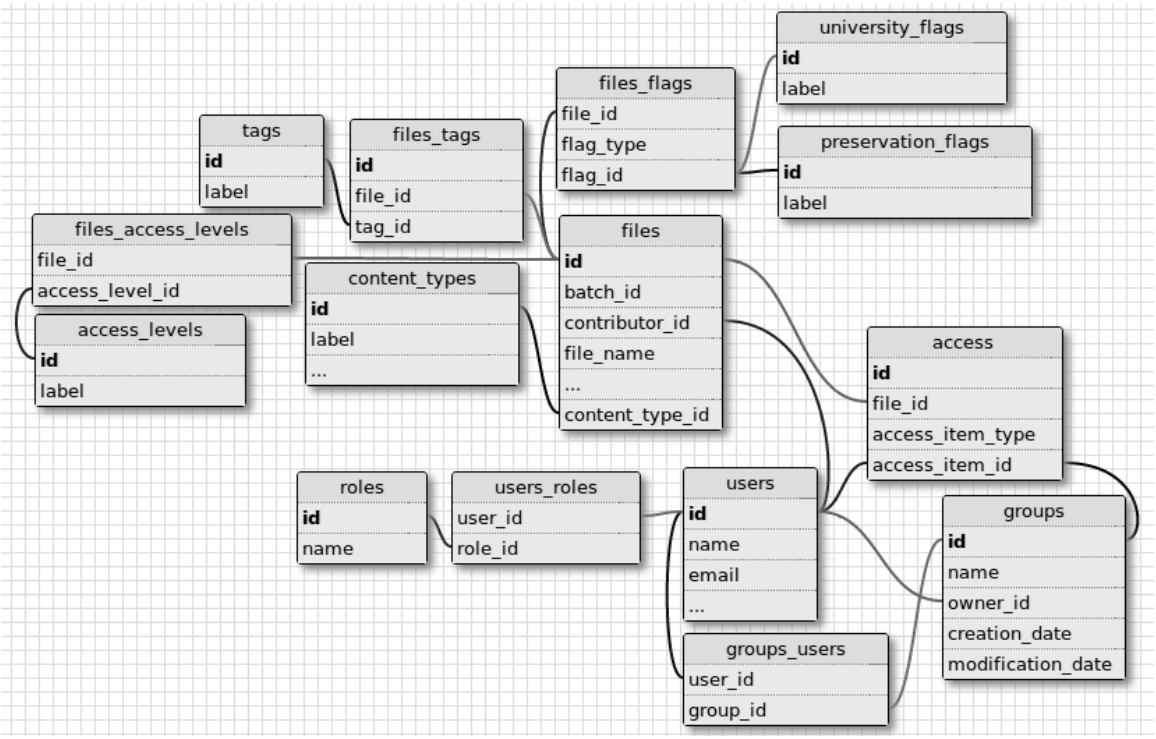
Wendy and Andrea put together a test suite of content and metadata to assist in the development of the system. The test suite included a variety of public domain and other content in many different formats (images, videos, office documents, audio recordings).

Collection	Short description
Live at the Waterworks	Color photographs of the Metropolitan Waterworks Museum opening taken on a iPhone
Blues Maker	1969 film on Mississippi blues singer Fred McDowell
Giant Kelp	Scientific image of a Giant Kelp
Hubble Images	Images of the universe taken with the Hubble telescope
File Format Obsolescence Project Proposal to NEH	Andrea Goethals' papers related to an NEH grant proposal
Unified Digital Format Registry	Project papers, technical documentation, meeting minutes, presentations and grant papers for the UDFR project
Voices from the Dust Bowl	Historical recording of Arthur Clyde, a guitar player, in the early 1940's
Backstage	A silent film from 1919
The Wonderful World of Oz	Earliest surviving version of The Wonderful World of Oz from 1910
Recycling Plant	Wendy Gogel's color photographs taken at a Florida recycling plant
Window Tableaux	Wendy Gogel's experimental color photographs
Tori Amos	Images taken at a 2009 Tori Amos concert from a phone
Gary's Room 1968	Scan of Gary Gogel in dress uniform
Catalonia is not Spain	PowerPoint slideshow of Wendy Gogel's 2010 trip to Spain
Bongo and Pongo	Wendy Gogel's pet videos
Images	Wendy Gogel's RAW images

The development work consisted of defining the key tasks, designing and implementing the data model and database schema, writing the back end code, developing user interface wireframes, developing the user interfaces and system testing.

The key tasks were identified as: integration of FITS (a format identification tool), user authentication, uploading via a Web interface, uploading via SFTP, batch management, tagging functionality, support for file access levels, preservation flagging, university record flagging, item search, item browse, group management, QA and testing, and performance optimization.

The database was implemented using PostgreSQL. A partial display of the data model is shown below.



The developers created code in Perl, Java, Ruby, JavaScript and Shell. The code is hosted on the Berkman Center's GitHub site: https://github.com/berkmancenter/zone1

The Zone 1 code uses a variety of open source software including nginx + passenger (Ruby web application deployment and server), devise (authentication), acl9 (role-based authorization), plupload (Ajax-upload tool), solr and sunspot (search), acts_as_taggable_on (custom tagging), jquery and jquery UI (Web interfaces), will_paginate (pagination), FITS (format identification), and resque + redis (creating and managing background jobs).

The code has support for creating and managing accounts, managing content and metadata, depositing content (through upload and sftp), ingest of content, content retrieval, content tagging and browsing. While it doesn't support all the functional requirements within each of these categories it does support the major functions.

The following wireframes were created to specify the Zone 1 user interface:

- Homepage
- Browse a list
- Browse by thumbnails
- Edit files individually and in bulk
- Upload files individually and in bulk
- Manage groups
- System configuration
- Rights configuration
- Flag configuration

- Role configuration

The developers are expected to finish developing the Zone 1 user interface based on the wireframes as part of this phase of the project. The Berkman team will do some usability testing, system and performance testing.

Challenges

Because this is open source software that could be used by other institutions, and because the Harvard policies have not yet been defined, the policies behind the system (who can do what) needed to be highly configurable. The system needed to be implemented in a way that policies could be configured (and changed over time) according to the institution's policies and experience with the system. These "rights requirements" proved to be challenging for the developers. Instead of being able to use an existing framework as-is they had to extend a framework in order to meet the requirements.

Next steps

Some of the functions were not coded in this round of the project. Most notably the content exports, transfer to other repositories and content reporting need to be implemented. In addition, once these functions are implemented, there needs to be testing and evaluation of the system by potential users of the repository. In the hopes of gaining funds to complete these tasks and testing, a follow-up Library Lab proposal will be submitted.

Budget spent

Total development hours: 1130.21
Total development cost (including salary, benefits, overhead, and administration): $120,423.79

Project Publicity and Presentations

October 27, 2011: Zone 1 table, Library Lab Showcase

> Skip, Wendy and Andrea represented the project at the showcase at Lamont which was very well attended. It was a great opportunity to talk about the project with all the people from Harvard and MIT who stopped by the Zone 1 table.

September 9, 2011: “Shelving Bits is Harder Then it Looks”, Presentation on Library Lab Projects, Sue Kriegsman and Sebastian Diaz, ABCD Committee Meeting

August 4, 2011: Library Lab Project Showcase, Lamont

Andrea, Wendy and Skip spoke about the project in this open meeting.

June 23, 2011: “Digital Pioneer: Andrea Goethals” - Blog posting by Mike Ashenfelder <http://blogs.loc.gov/digitalpreservation/2011/06/digital-pioneer-andrea-goethals/>

Excerpt: “*Harvard has several repositories dedicated to specific purposes. But a lot of other valuable content is drifting loose around the university, such as faculty research, student work and drives full of unprocessed data. Much of it is on unstable media or in danger of being lost during a move. So Goethals is a co-creator of Zone 1, a catch-all “rescue repository” for homeless content. To enable access by a wide range of users, Zone 1 will deliberately be easy to use and will require only a small bit of metadata. Zone 1 users could evaluate the content and have it moved to the appropriate long-term preservation repository.*”

Library Lab Final Report

Zone 1 – The Study: Outcomes, Findings, and Deliverables

November 15, 2011

The Harvard University Archives actively collects faculty archives, an important component of Harvard's institutional and intellectual history. Likewise, the MIT Institute Archives has a long history of collecting faculty archives. Changes in the composition of these records, however, now make it necessary both to re-examine and re-conceptualize what constitutes records created by faculty and to assess the challenges associated with their collection and preservation.

Zone 1 – The Study is designed to achieve two primary goals: (1) to collect hard data to answer basic questions about the changing nature of traditional faculty paper collections and to identify ingest, storage, and preservation complexities and (2) to develop a database survey tool to consolidate donor information, record information, and produce reports on the composition of faculty collections surveyed by collection development staff. The principal organizers of the project were the Harvard University Archives and the Institute Archives and Special Collections at MIT. Other Harvard partners will include, by the end of the project, special collections at the Graduate School of Design, Business School, and Medical School. In addition to Project Fellow Alexandra Bisio and Project Manager Skip Kendall, project participants, to date, include:

- Harvard Graduate School of Design, Loeb Library – Mary Daniels, Ann Whiteside
- Harvard Medical School, Countway Library of Medicine, Center for the History of Medicine – Kathryn Hammond Baker, Giordana Mecagni
- Harvard University Archives – Virginia Hunt, Megan Sniffin-Marinoff
- MIT, Institute Archives and Special Collections – Elizabeth Andrews, Thomas Rosko

As this report is submitted, we are winding up the project that we expect will be completed by mid-December 2011. With a slow start in identifying and hiring the project fellow, we received permission to keep the fellow on the project until November 30. In addition, in late October we were contacted by the Baker Library Historical Collections at the Harvard Business School (HBS), asking if it was possible to be included in the project. With extra financial support from the HBS, we expanded the scope of faculty to be surveyed by the fellow until December 16.

We project that the final [Library Lab] funding for staff salary and benefits will be $13,583.64. Choosing to work with a local HUL staff programmer, we will not expend the sum we set aside for programming assistance.

Appraisal and Technology: Identifying Challenges in Faculty Records

Though archivists are well aware of the abstract challenges digital media pose to processing and preservation of faculty records, little has been done to identify and quantify the problems to address when attempting to create appraisal policies. Also, changes in the way academics communicate and conduct their work are challenging the notion that archivists continue to have a firm understanding of what constitutes a collection of faculty records. During the course of this project we examined: the contents of a variety of faculty records; the technical problems digital elements will pose and at what magnitude; and, most importantly, how to prepare to face challenges related to access, obsolescence, ingest, and the cost of preserving and maintaining digital materials.

Collection Development Survey Tool

In addition to collecting data to enhance an understanding of modern faculty records, we commissioned a database tool, with the code to be shared with all participants, that will not only be used to store information accumulated for this project, but will also be used for future for surveys of donor and faculty records. The database will be used to hold various pieces of information, including textual information, various overlapping category designations, and volume. Originally, we identified six attributes as essential to the functionality of the tool:

- An easy to use interface
- The ability to be used "in the field" and without an internet connection at time of data entry
- Good reporting capabilities
- Flexible export capabilities
- Web-based interface
- Open-source

When complete, the database will consist of two parts – one to hold information taken from both the physical inventory-style survey and the other information from the verbal interview-style survey conducted by the project fellow. We anticipate that this form of data gathering from individual faculty members will be common going forward. The first section of the Zone 1 "Study" database is in the early testing stages, and the second is still in production, with a final product anticipated in December 2011, after we have added information for drop-down menus and finished the work with the addition of the HBS faculty. Description and screenshots of the database are included in the appendix to this report.

Methodology

The project fellow conducted interviews with faculty members and surveyed their records. The interviews and surveys take place in two or more meetings: the first to explain the project, allow the faculty member to ask any questions regarding the project, and to to conduct the verbal, interview-style survey and the following meeting(s) to examine both digital and paper records held in various offices belonging to faculty members and record the information in measurable ways in the database tool created for this project.

Verbal Survey

Initially, a draft of the verbal survey was created to be a complement to the physical survey of the subject's records, intended to "enhance the data gathered as well as to simply not overlook anything that might not seem obvious to the faculty member." In August, after speaking with some faculty and staff, the research fellow began to reconstruct and reorganize this survey through consolidation and significant expansion. The final survey is approximately five pages long and takes around thirty to forty-five minutes to complete (depending on the detail of the subject's answers). The seven sections of the survey are:

- Personal and Professional History
- Digital History
- Nature of Records
- Location of Records
- Structure of Records
- Use of Records
- Access to the records

Physical Survey

Physical surveys are conducted in the subject's primary place of work, generally a university office, and in any other location in which records maybe kept that the subject will allow us to view, such as an outside professional office, home office, or laboratory. The research fellow makes a detailed inventory of all records, paper and electronic, available to her in each space. The information gathered as part of the survey includes:

- Subject of records
- Recording Categories
- Storage
- Form

Study Sample

The sample for this project consisted of the records of faculty at both Harvard and MIT, representing a number of disciplines (ranging from the hard sciences to the applied arts), and at different stages of faculty careers. A range of individuals (who for this report will need to remain anonymous) were selected in order to gain a full understanding of the composition of records for various disciplines and departments.

- The Loeb Library at the Graduate School of Design selected three candidates for this study from three disciplines: urban planning and design; architecture; and landscape architecture
- The Countway Library of Medicine at Harvard Medical School selected three candidates for this study, who in addition to being medical researchers also hold/held key administrative positions
- The University Archives selected two members of the Faculty of Arts and Sciences for this study: a computer scientist and historian of science
- The MIT Institute Archives focused on faculty in science and engineering, as well as architecture and urban planning

The Baker Library at the Harvard Business School library has just become involved in the project and will select faculty members within the month.

Initial Results

As of this report, the initial results of the Zone 1 surveys are limited to data collected though verbal surveys conducted by the project fellow with the faculty member whose records are to be examined. Although other appointments are expected to be scheduled with faculty by the end of November, one faculty member has been able to accommodate the project staff's request to complete the full survey request to view her records. Though more specific data will be attained though the physical survey, we have been able to collect faculty members' expanded descriptions of technology usage, record creation, record storage, data sharing, and how the records they have created and stored may reflect their careers at Harvard and MIT.

Findings: Similarities among the faculty surveyed

Hardware and Removable Storage Media

- Most faculty members have at least one laptop that they carry between locations. For some, these laptops are their primary computers at home as well as at their university and professional offices. At least two faculty members keep all documents, whether professional, personal, academic, or university administrative, on one laptop.
- All of the faculty members use some type of removable media for storage or for transporting files from one location to another, the most commonly used being USB flash

drives. While the faculty members used other common tools, including external hard drives and some optical disks, there were some instances of obsolete media present in the collections of faculty whose work was started on early computer technology. For example, one faculty member who started his career as a computer scientist had a multiplicity of obsolete material including cartridges, two gigabyte digital tapes, TRS 80 cassettes, magnetic tape, and paper punch cards.

File Types

- Though faculty tend to create a great variety of different data forms and formats, there were some common to all faculty members regardless of discipline. Several faculty members have created large collections of images, both professional and personal (although the line between the two seemed to be somewhat blurred for some of the GSD faculty), in various file formats including high resolution TIFFs. Two GSD faculty members have created large collections of architectural slides that have recently been digitized and that they plan to donate to Loeb Library Special Collections. Another common file category consists of large collections of PDFs, generally articles saved as reference and general reading material. At least one faculty member explicitly stated in his verbal survey that he did in fact annotate the digital copies of PDFs, which he stored in an application called Papers that allows him to extract metadata from PubMed and efficiently organize his collection. Some faculty stated that they sometime printed out these PDFs to make reading copies.

Communication

- All of the faculty members use some sort of digital tools to communicate with other scholars and professionals, even if in the most limited sense. At the very least, all faculty members used email for almost all correspondence. Some faculty use commercial services such as Google Docs and Dropbox to share work with collaborators. One faculty member, for example, uses Google Docs to communicate and share documents with the other members of a journal editorial board. A few other sites, such as Net Temp, Doodle, and Web X, are used by faculty for scholarly communication and storage. One faculty member uses Skype quite extensively as a tool for international communication. A majority of the

faculty surveyed use Harvard iSites to communicate with students, though most iSites tend to be managed by teaching assistants and technology professionals.

- A significant number of faculty tended to use an email account other than their University-provided address as their primary email account, Gmail being the most common service used. Most used their University address as an alias account that forwards messages directly to their Gmail address. Indeed, a meeting with Kevin Lau, Head of Library Information Systems and Instruction Technology at the Frances Loeb Library, revealed that though many different services are offered by Harvard technology offices, Harvard faculty do not always use them. Goliath, a system providing faculty with remote access to personal GSD server space, works very much like the commercial service Dropbox, but is much more secure. Still, one surveyed faculty member relies heavily on Dropbox for storage and remote access to files. In fact, none of the faculty interviewed thus far regularly use the shared server space provided to them by the university, and only one uses the personal storage space as a back-up for his University files, most of which are saved on the hard drive of his computer. He did not, however, back up any of his personal files at all with any service or on any device. The fact that faculty tend not to store their materials on spaces easily accessible to the various collecting archives will make ingest of data at a later date more difficult.

Findings: Differences found among the faculty surveyed

Technology Usage

- Within the faculty group survey there was a wide range of technological skill; some of the faculty had been using computers since the early 1970s, where one has been using a computer for less than eight years. Early adopters tend to be those whose areas of study rely heavily on computing, where those whose work is less dependent on computers tend to adopt technologies later.

- A professor at the GSD, who heavily relies on many different kinds of technology in both his work and personal life, is the only professor who had been trained to set up his own iSites and who does so every semester.

Work Life

- Though the faculty surveyed thus far have been both professionals and academics, those that still had professional offices tended to have more fragmented records. Only one professor at the GSD used cloud computing to sync his files to multiple computers in multiple locations. Most of those with professional offices used one laptop for all their work.

- The faculty at the GSD tends to do more work internationally than do either of the subjects interviewed at the HMS. Indeed, one professor at the GSD had his records split between a home in Boston and a home in Mumbai.

Project Challenges

During the course of this study, the project team ran into a few problems that impeded the completion of ***Zone 1 – The Study*** by October 31. (The extension to November 30 will resolve most issues.) These problems included:

- Scheduling

  We have encountered many problems when scheduling faculty members to meet with the project fellow for both initial and follow-up meetings. The beginning of the semester is, generally, a very busy time for faculty and most were unable to accommodate us in the initial time frame given for the Zone 1 project. Meetings had to be scheduled later in the semester than originally expected.

- Full Access

  There have also been some problems with faculty members allowing the project fellow full access to their records. In one physical survey, the faculty member maintained control over the computer and led the overview of the documents, making it difficult to collect all of the data required. In addition, though we had first intended to survey all of a faculty member's work places, many seem to be comfortable only allowing access to their university offices.

- Equipment

  We have had a few technical problems with the equipment needed for Zone 1. While work on the Zone 1 survey tool has been steady, a late start has prevented its use in production. Also, the archives fellow attempted to use a new file analysis application distributed by NARA for a physical survey, which failed and resulted in the loss of some data.

Presentations

Tom Rosko and Megan Sniffin-Marinoff have proposed a session to discuss the project at the upcoming annual meeting of the Society of American Archivists. Once all of the data is collected, we will consider other options to share the findings.

Continuing Zone 1

Between now and the middle of December we will continue interviews with faculty members to whom we have already begun the interview process. HUA and MIT Institute Archives staff will continue to complete any physical surveys that cannot be fully accommodated by faculty schedules – our greatest challenge – before the project fellow departs. The project fellow will continue to test the first completed section of the Zone 1 tool by entering data collected from faculty. Dee Dee Crema of the Harvard Library will continue development of the survey tool over the next five to six weeks. Discussions of the policy implications of the Zone 1 Repository are ongoing. A number of issues have been identified and a Harvard/MIT group is being organized to discuss them.

Alexandra Bisio, Project Fellow, Harvard University Archives

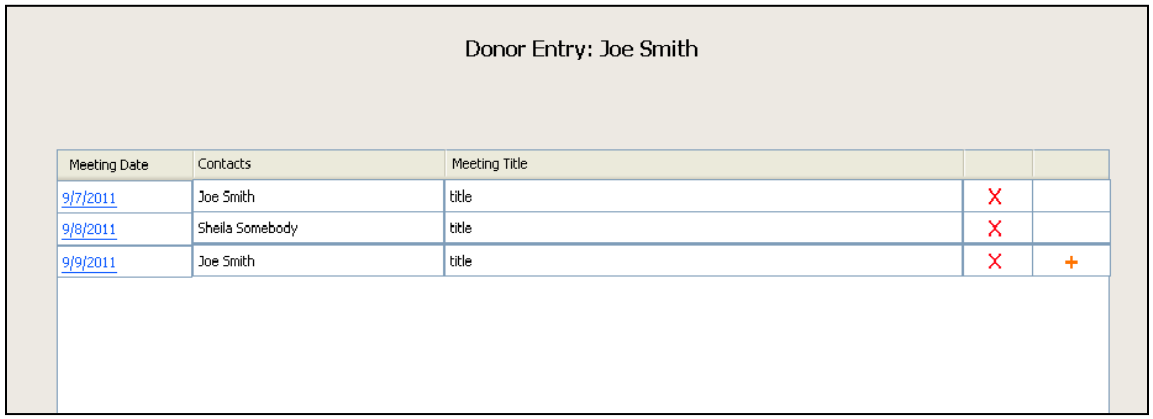
Skip Kendall, Project Manager, Harvard University Archives

Thomas Rosko, Head, Institute Archives and Special Collections, MIT

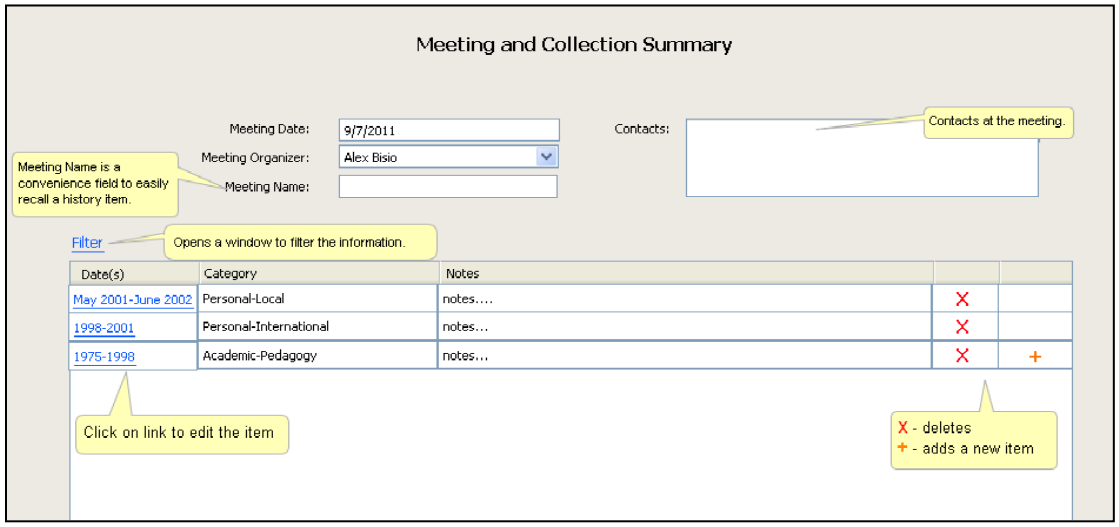
Megan Sniffin-Marinoff, University Archivist, Harvard University Archives

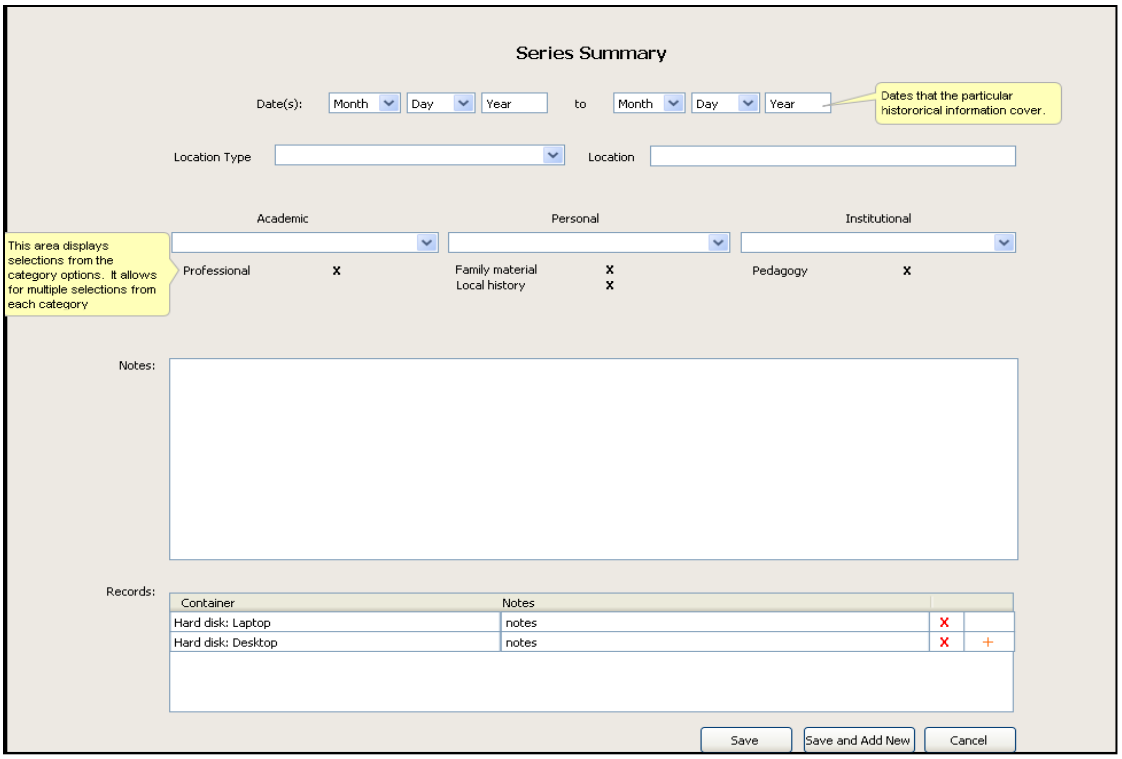
Appendix

Zone 1 Collection Development Tool



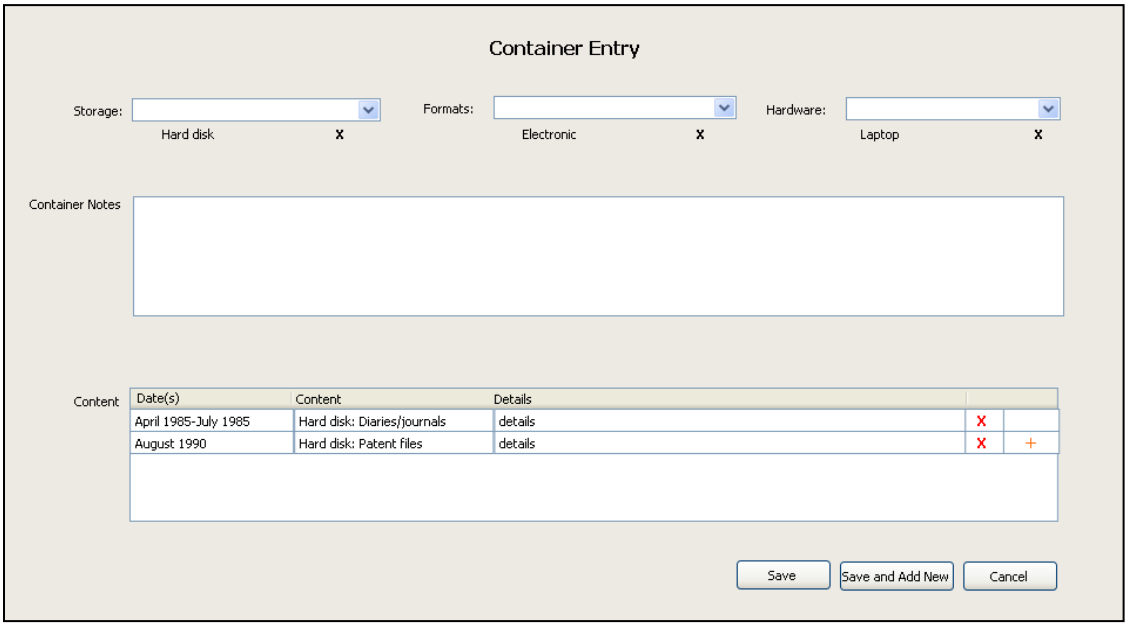
Donor Entry: The donor entry page displays a list of meetings generated though a search by date, contact name, or meeting title. From this page you can select a date to open a past or current meeting with a faculty member.

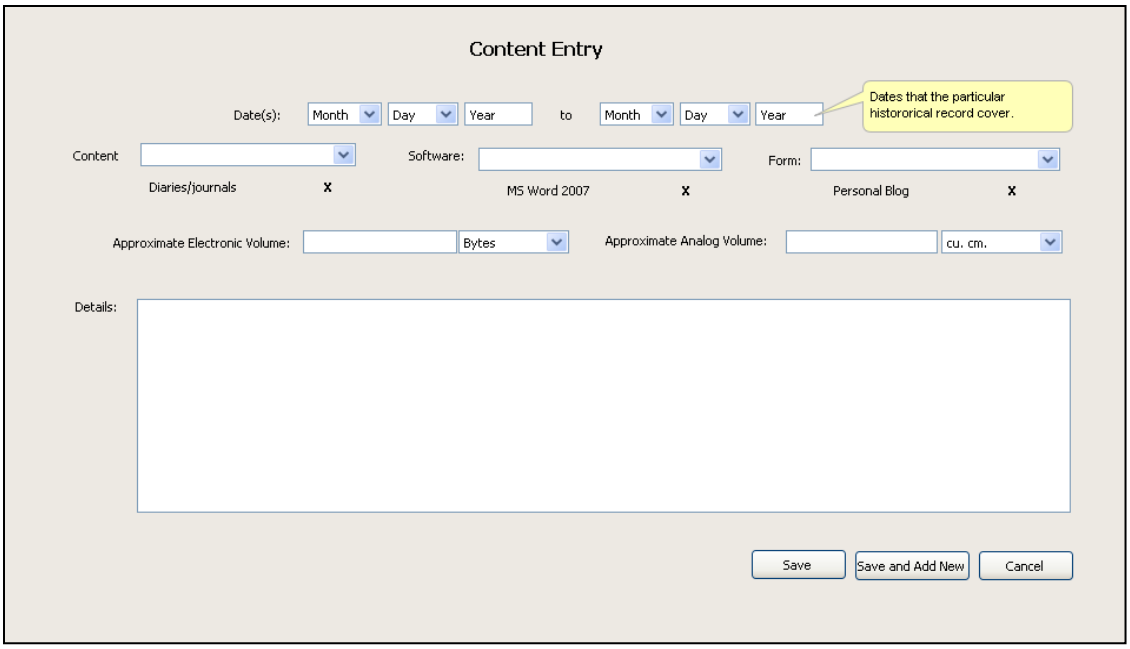




Meeting and Collection Summary: This page allows the archives fellow to enter contact information and describe the meeting in more detail. On this page the fellow will record a list of series found in a faculty member's collection.

Series Summary: This page allows for more detail to be entered regarding the subjects found in each series within a donor's collection.





Container Entry: This page allows for the entry of specific information regarding the various containers, both digital and analog, used by a faculty member to store items entered in the content entry.

Content Entry: On this page the fellow will enter information about the content and form of the items found within the containers. Volume and software used for creation will also be noted here.