This proposal differs from the first draft reviewed by the Library Lab reviewers in that it combines two proposals:

- **Zone 1: First stage preservation storage for digital content**
- **IR2: Re-Thinking the Institutional Repository**

The question about the relationship to LOCKSS is addressed in Appendix A.

Zone 1: A Rescue Repository for Digital Content

Library Lab Proposal, April 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
Tom Rosko, Institute Archivist/Head, Institute Archives and Special Collections, MIT
Megan Sniffin-Marinoff, University Archivist, University Archives, Harvard University

Zone 1 is a project with three parallel activities:

1. Development of a prototype rescue repository at Harvard
2. 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
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

We are not aware of any similar project at Harvard or MIT.

Faculty, students and staff throughout universities produce, collect and rely on digital content in increasing amounts, but safe harbor is available for only a small portion of it.

Harvard's existing specialized storage solutions serve particular communities, offering different services for using and storing digital content. The Digital Repository Service (DRS) has over 400 TB of content, and provides access and long-term preservation services for permanently valuable content with custodial stewardship by Harvard organizational units. DASH, with over 5,000 items, provides storage and access to any content authored by a member of the Harvard community. IQSS Dataverse Network and the Murray Research Archive, with over 100 TB of content, provide access and preservation services for primarily social science research data. In addition, there are systems at Harvard that feed particular types of content into the DRS (close to 1 TB of Web content through WAX, and email through the under-development EAS). Likewise at MIT, DSpace@MIT (selected research and teaching output of MIT faculty), and Dome (curated digital library collections), both based on the DSpace software platform, contain 2+TB of data and 100,000+ files.

Yet there remains a lot of digital content at these universities that is not a good fit for these specialized repositories[1].
There are categories of content at Harvard[2] that would be better served in a general-purpose secure storage space that would act as a staging area[3]. The categories include 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[4]
- 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, e.g. the archived research in Figure 1.

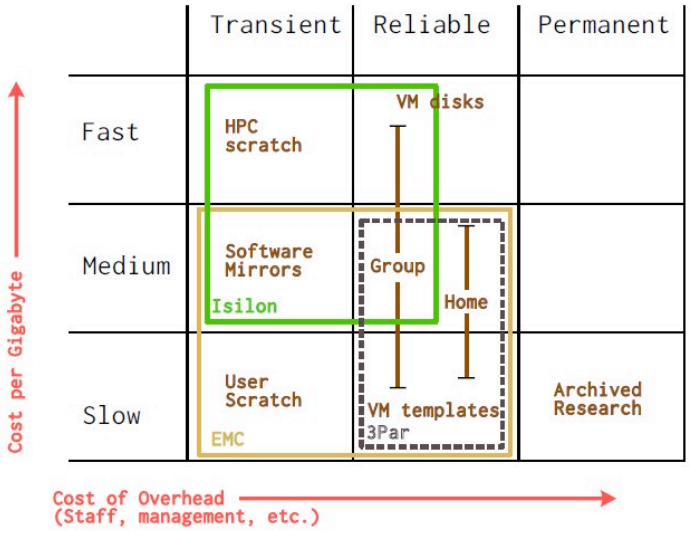

Figure 1: Storage Needs of Harvard School of Engineering & Applied Sciences (SEAS). The lack of a solution to handle the archived research (see lower right corner) has been described as a "painful hole" in their infrastructure.

Source: Matthew Miller, Senior Systems Architect, SEAS, April 1, 2011.

The Solution – Rescue Repository

We propose to solve this problem with the Zone 1 rescue repository, a new easy-to-use service for *all* of the Harvard community, for *any* digital content. This service would provide:

- easy ways to deposit content into the rescue repository, through commonly-used software, platforms and Web sites

[1] In addition to DRS, DASH, Dataverse Network and the Murray Archive; Harvard repositories include the HKS Research Report Online, and the HMS HMScholar.

[2] MIT has similar needs. Their repository infrastructure doesn't accommodate all the categories of digital content that they would like to manage.

[3] The records center portion of the Harvard Depository is an analogous staging area for non-digital staff and some faculty records.

[4] Learning objects are created by Presidential, Library, and Museum Instructional Technology Fellows (PITFs, LITFs, and MITFs) at Harvard each semester.

- a bit-level preservation[5] 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 at Harvard or elsewhere

The intention of the rescue repository is to provide the minimum infrastructure needed to rescue and secure digital content that does not already have a tailored storage solution. This repository is not intended to replace or replicate functionality provided by the DRS, DASH or any of the other existing repositories at Harvard. Rather it is meant to complement their functions by providing first-line preservation to any digital content while decisions are being made about its longer-term disposition. The rescue repository addresses the time-sensitive need to protect the content until it can be removed for scheduled destruction or fed into other repositories for more advanced services, including full preservation.

A prototype of the rescue repository will be built for this one-year project. It will focus on user interaction with the system – on developing easy-to-use methods for moving content into and out of the system, and reviewing content while it is in the system. The prototype will not include the back-end replication, integrity-checking and content recovery functions that have already been proven by systems such as the DRS, and that would be necessary in a production version of the rescue repository.

The following supporting activities are meant to inform the design of the prototype as well as any future production version of the rescue repository.

The Study

Concurrent with building the prototype rescue repository, the Zone 1 project will conduct a study of faculty records[6] at Harvard and MIT to identify any challenges and requirements for long-term management of the content. Electronic records -- both personal and professional -- created by faculty and

[5] 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.

[6] Faculty records include a broad array of research, teaching, personal, professional, and administrative documents. Among them are: course materials, biographical materials, correspondence (with colleagues, students, and administrators), internal and external/professional committee files, research files, drafts of published work, and data sets.

known to archivists as “faculty papers” are one likely category of content for the rescue repository. Faculty papers were chosen for this study because this content is likely to reveal multiple collection, preservation and access challenges including questions of ownership, privacy, IP and other legal considerations. The need for secure, general-purpose storage space is particularly acute for the faculty.

Faculty collections are among the largest, most highly prized, and most heavily used of the collections in the archives at both Harvard and MIT. The material has been mined for decades by scholars producing monographs and articles as well as graduate and undergraduate theses and dissertations. Within the institutional archives of our two organizations are thousands of feet of material comprised of millions of pages of documents recording, across several centuries, the lives, work, and thoughts of our faculty, key players central to the development of our institutions – among the most significant educational, scientific, and cultural organizations in the country, if not the world. Faculty collections provide a personal record of the intellectual and social history of the academic institution as well as valuable insight into the history of invention, creativity, and expertise.

In consultation with archives/library staff, a project archivist will analyze the structure, content and scale of faculty records generated by a sample of faculty at Harvard and MIT. The faculty materials selected for analysis represent a variety of disciplines and of tenures (ranging from early/mid-career to emeritus faculty). The project archivist will develop an appraisal methodology and tool for recording information about the faculty records, conduct an analysis of the content, summarize findings, and make recommendations.

This study will be led by the Institute Archives and Special Collections at MIT and Harvard. Other Harvard participants include Special Collections, Harvard Graduate School of Design; and the Center for the History of Medicine, Countway Library, Harvard Medical School. These partners were chosen because of the anticipated level of complexity in the records of these disciplines.

Policy Discussions

Over the course of the project a series of meetings will be held between Zone 1 project participants and others at MIT and Harvard to identify the critical policy issues needing attention and resolution before a production system can be fully implemented.

Project Benefit

Rescue Repository

The Zone 1 rescue repository prototype serves as proof-of-concept for a future production version of the repository which would close the gap in secure storage solutions at Harvard that currently exists. By lowering the deposit barrier and removing content eligibility requirements, Harvard greatly increases the chances that its valuable content won't be lost. It serves as a conduit of review by the Harvard community for potential re-use or long-term stewardship of the content and facilitates transfer to other repositories.

Any software created for the rescue repository prototype will be released under an open source license, which could be used by MIT or other universities with the same need.

The Study

The study will result in a greater understanding of the nature of digital faculty papers. For archivists and records managers, this understanding can be generalized to other types of content under their management.

It is expected that the analysis tool developed for the study will be useful to Harvard and MIT for analyzing content beyond the life of this project. In addition, the analysis tool may be helpful for the design of the rescue repository content review interface. Any software created by the study will be released under an open source license.

Policy Discussion

The policy discussions will lay the groundwork for moving the rescue repository prototype into a fully-implemented offering at Harvard, and for implementing a similar solution at MIT.

Project Work Plan:

Rescue Repository (two phases)

- Phase 1: (Months 1-3)
  - Define in more detail the prototype requirements
  - Identify content contributors, reviewers and stewards to help test the prototype, including faculty and staff
  - Design the prototype
  - Select prototype infrastructure (SW, HW)
  - Design review

- Project review (by Library Lab)
- Phase 2: (Months 4-12)
    - Build the prototype
    - Test the prototype
    - Evaluate success of the prototype
        - Functionality tests to determine if the prototype can support:
            - Deposit of content by content contributors
            - Nomination of content to receive long-term preservation
            - Review of content by potential long-term stewards
            - Review of content by potential re-users
            - Extraction of content by content contributors, content re-users and long-term stewards
            - Transfer of content to at least one external repository
        - Survey functionality testers (content holders and potential stewards)
            - Does the rescue repository fulfill an unmet need?
        - Evidence of future use
            - Identify community interest (content contributors)

The Study (two phases starting in the first month of the year-long project and ending after the sixth month)

- Phase 1: (Months 1-3)
    - Identify faculty members who will participate in the study
    - Develop content analysis tool
    - Conduct initial interviews with faculty
    - Conduct initial analysis of the faculty records
    - First draft of summary paper
- Project review (by Library Lab)
- Phase 2: (Months 4-6)
    - Complete interviews with faculty
    - Complete analysis of the faculty records
    - Final draft of a summary paper

Policy Discussions (two phases)

- Phase 1: (Months 1-3)
    - Two policy discussion meetings
    - First draft of a summary paper
- Project review (by Library Lab)
- Phase 2: (Months 4-12)
    - Four policy discussion meetings
    - Final draft of a summary paper

The timing of the three activities is summarized here:

Month 1: The rescue repository prototype work, study and policy discussions begin.
Month 3 end: Have completed the first milestone for each of the activities:

- Rescue repository: design review
- Study: first draft of summary paper
- Policy discussions: first draft of summary paper

Project review by Library Lab
Month 6 end: Study completed – final draft of summary paper
Month 12 end:

- Rescue repository prototype completed and evaluated
- Policy discussions completed and have final draft of summary paper

Project Resource Needs:

Rescue Repository

Staff

- Software developer/architect (1 FTE for 1 year)
  - To design, develop and test the prototype
  - This position can be contracted out or filled by term staff.
- Project manager
  - To provide direction on Zone 1 requirements, design and implementation
  - Andrea Goethals will serve in this role as part of her duties as manager of the digital preservation program.
- Liaison to content contributors, reviewers and stewards
  - To coordinate external testers of the rescue repository; to work with the study archivists, faculty and others to identify testers and potential future users
  - Wendy Gogel will serve in this role as part of her duties as manager of digital content and projects
- Reviewers
  - To review Zone 1 requirements, design and technology choices
  - Existing staff will be tapped for review on an as-needed basis. The particular individuals will vary depending on the nature of the review. At a minimum this will include a usability expert (Janet Taylor, OIS), software developers at OIS and throughout the university, and a representative from the study team (Skip Kendall, HU Archives).

Equipment and tools

- TBD based on design. May require a virtual machine and storage space.

The Study

Staff

- Project manager
    - To supervise the project archivist and provide direction on the study design and implementation
    - Skip Kendall will serve in this role as part of his duties as electronic records analyst and archivist in the HU Archives
- Project archivist – Part-time, 20 hours a week, 25 weeks
    - To manage the project, develop an appraisal tool, conduct the analysis of faculty papers and write the study report.
    - This position can be contracted out or filled by a temporary appointment
- Project advisors/reviewers
    - To advise the project manager and archivist; to review project process; to select faculty papers for the study
    - Megan Sniffin Marinoff, Harvard University Archivist
    - Virginia Hunt, Associate University Archivist for Collection Development, Harvard University Archives
    - Kathryn Hammond Baker, Deputy Director, Center for History of Medicine, Harvard Medical School
    - Mary Daniels, Special Collections Librarian, Harvard Graduate School of Design
    - Tom Rosko, Institute Archivist and Head, Institute Archives and Special Collections, MIT
    - Liz Andrews, Associate Head, Institute Archives and Special Collections, MIT

Equipment and Tools

- Computers, software and office equipment provided by Harvard University Archives

Policy discussions

Staff

- Project manager
    - To coordinate the discussions and write summary report
    - Skip Kendall will serve in this role as part of his duties as electronic records analyst and archivist in the HU Archives
- Discussion participants
    - This will include all of the Zone 1 project participants as well as others at Harvard and MIT

Equipment and Tools

- None needed

Total Funding Request:

Rescue repository prototype

- 1 FTE Developer for 1 year

- Equipment

Study
- Project archivist: 20 hrs/25 wks
- Appraisal tool database development

Policy discussions
- No funds needed

Appendix A: Relation to LOCKSS

This section is in response to a question that came out of the first review of the Zone 1 proposal, "… address how Zone 1 may relate to LOCKSS and if it could be built as an addition to LOCKSS. If this is not possible, then let the committee know why it's not a viable option and a new stand alone product is necessary."

LOCKSS is back-end replication, integrity-checking, and content recovery software. Since these functions have already been proven at Harvard by DRS, the experimental prototype for the Zone 1 rescue repository will not include them. However, a production version of the repository would need these functions which products like the LOCKSS software can address.

LOCKSS software is used in two ways:

1. It is used by a global network of libraries to collect and maintain bit integrity of copies of cooperating publishers' journals. In the event that this content becomes unavailable through the publishers' websites, the LOCKSS-stored copies can be made available to authorized users.

2. It has also been used to form private LOCKSS networks (PLNs). In contrast to the global LOCKSS network, PLNs preserve content that is of interest to their particular institutions, primarily special collections. For example, the MetaArchive PLN houses the Southern Digital Culture archive for six universities in the southern US.

In a LOCKSS-based network, content is replicated to multiple LOCKSS "boxes". The software performs comparisons of the copies distributed through the network to detect lost or corrupted copies. Its algorithm for detecting bad copies relies on multiple copies of the same content – the recommended minimum is seven copies. Although the storage required to keep seven copies of the same content can be prohibitively expensive for large-scale repositories such as the DRS, this solution is economical when the total amount of content stored in the network is not large. To illustrate, the MetaArchive PLN has 24 LOCKKS boxes, spread across 18 institutions, each institution maintaining one or two boxes. Although the number of participants in this network is large, it would take three times the size of the MetaArchive PLN to store seven copies of just the content stored in the DRS, not leaving any room for the other institutions' content. Although LOCKSS should not be ruled out as a possible piece of the production version of the rescue repository, the potential scale of the rescue repository may suggest a more economical solution.