

**Requirements Document for the Rescue Repository**  
**Rescue Repository Requirements Task Force**  
Yale University Library, May, 2004

**I) Background, Charge and Methodology**

In March 2004, the Rescue Repository Requirements Task Force (RRRTF) was created by the Integrated Access Council (IAC). A background is best provided by excerpting from a draft proposal written in February 2004 (Meg Bellinger, Frederick Martz, David Gewirtz, Audrey Novak):

*"An increasing number of projects in the Yale University Library are generating or acquiring digital content (images, video, audio, full text, data sets) along with varying levels of metadata and public interfaces for access to this material. Most of these projects urgently need a long-term (or even a short-term) digital preservation strategy..."*

*"...the lifecycle management of this content has become a clear necessity as the volume continues to grow. The digital masters for much of this material are in immediate danger of permanent loss through media decay, physical damage, technological obsolescence, or difficulties in archival management..."*

*"...in the interim, we propose a flexible and agile/quick short-term solution..."*

*"...The concept behind the Rescue Repository is to provide a centrally supported system for the short-term safekeeping of these endangered materials. This repository will constitute the first step toward the evolution of a long-term, OAI-compliant digital preservation archive..."*

*"...a principal objective of the repository would be to pool expertise and resources in order to provide a managed environment for the protection of digital masters that will be needed to create derivative objects for current and future user-oriented applications..."*

The RRRTF was duly charged as follows:

*Develop a systems requirements document that describes the functional requirements for a rescue repository/storage environment for readily supported file format types that will be procured, installed (ITS AM&T), tested and with a first collection selected and ingested by the end of summer 2004. The RR must be functional without major intervention for 36 months.*

The complete Rescue Repository plan involves two groups: the Rescue Repository Requirements Task Force to gather and document user requirements, followed by an Implementation Team which will select, acquire and implement the system based on these requirements.

This document is the outcome of the RRRTF's work. Regarding methodology, the RRRTF approached the problem from an operational/functional perspective (explained below). An initial discussion of issues and questions led to the concept of the development of a questionnaire as a tool for both organizing these issues and gathering responses. Also, several issues revealed themselves as non-questions (e.g. does the system need to be secure?); these we noted as "first principles" and they are included among the requirements. After the development of this questionnaire was completed it was filled out by members of the task force, and distributed to and filled out by representatives of several other candidate electronic collections not represented in the group.

These results were compiled and distilled from the range of answers on each issue to a single user requirement. On several issues we have also recorded strong preferences, in an effort to help guide the Implementation Team more accurately in its selection process.

The operational perspective mentioned above drove the organization of issues and questions in that they were conceived within the framework of what the group considered the target system's main functions, or operations. This perspective purposefully excludes any strictly technical or implementation considerations which fall more appropriately to the implementation team (e.g. systems maintenance, operating system, disk redundancy techniques, cost sharing model) but rather limits itself to requirements and expectations from the point of view of the user of the Rescue Repository.

The main operations considered as important for this group to address are:

- **Ingest:** the transfer of digital materials into the repository. This aspect includes consideration for both an initial ingest and ongoing ingest as collections grow over the course of the system's life span.
- **Storage and Backup:** this area addresses questions of storage capacity, based on current collection sizes and expected growth. Also, questions regarding backup and recovery expectations are addressed here.
- **Retrieval:** this area addresses questions regarding how materials may be retrieved, by what means and by whom.

## II) Requirements and Preference Recommendations

This section contains the requirements, along with some preferences as noted above, for the proposed Rescue Repository as determined from our discussion and survey. They are organized into a “first principles” section, sections for the ingest, storage and backup and retrieval functions, and a last section regarding access control. Any preferences are noted in italics. This section also contains information regarding activity expectations within the operations which may be helpful to the implementers – these are underlined and prefixed with “Note:”. Last, any significant outliers are embedded per issue as well – these are underlined and italicized.

### A) First Principles

- 1) The Rescue Repository must be functional without major intervention for 36 months. (directly from the RRRTF's initial charge).
- 2) The Rescue Repository is not meant to replace the online systems intended to deliver content to readers (taken from draft proposal for Rescue Repository). Nor is it meant to replace automated backup systems; a notable difference is that the ingest process is a formally initiated user-driven process rather than the automated “pull”-oriented behavior implicit in an automated backup system.
- 3) Along with master content data, metadata may be stored as a user convenience, however there is no expectation that stored metadata will play an active role in file retrieval. The Rescue Repository is not meant to provide a framework for a metadata structure.
- 4) System will be generally secure based on
  - Net-ID-based validation via CAS or Windows Active Directory, or equivalently secure mechanism
  - Access control lists for ingest, deletion and retrieval functions
- 5) Repository will retain owner, collection and optionally sub-collection level of organization upon ingest in that:
  - Access to collections will be controlled by collection owners
  - Individual file retrieval can be directory-based
  - Bulk retrieval would be facilitated based upon this
  - Duplicate filenames *across* collections would not cause overwriting or error
- 6) Ingest function will catch duplicate filenames *within* a collection without overwriting.
- 7) Changes to Rescue Repository data will be limited to ingest and deletion. Changing a particular file will be effected by deleting from the Rescue Repository and re-ingesting from source system.
- 8) There will be an initial ingest, aided perhaps by a third party, and ongoing local ingest.

- 9) Ingesting data into the Rescue Repository will not involve an irreversible conversion from one data format to another.

## B) Ingest Requirements

- Materials on portable media may be transferred from their home collection to an on-campus ingest site, but must be returned to the home collection within one week. The ingest site must be on the Yale campus More than one collection expressed a need for an additional ingest mode based on electronic file transfer.
- There are no other conditions required regarding materials movement. *One collection expressed preference that initial ingest be performed during academic downtimes.*
- There are no special requirements regarding timeframe for the initial ingest.
- User feedback regarding the ingest process must include a detailed log showing all items, and whether the transfer was successful or failed. *There is also a preference for showing the results of some verification/validation testing, such as successfully opening the file, for the more common formats (e.g. TIFF, PDF, JPG, XML).*
- For the ongoing ingest, along with the log noted above, an audit must be supplied showing filenames, timestamps and a person identifier.
- The ongoing ingest function should be available from 8:30 AM-5PM, Mondays through Fridays. The electronic transfer mode should be available nearly 24x7; short daily outages and reasonable exceptions are acceptable. If this is implemented in a two-stage mode (i.e. submission to a holding area followed by actual ingest into the repository proper) then the interval between these stages must be timely (within 24 hours).
- A request-based interface for the ingest function is adequate. *There is a preference for a rudimentary automated interface. The electronic file-transfer mode implies a rudimentary interface for submission of materials.*
- Note: Frequency for the ongoing ingest function is envisioned as between “sometimes” (about once a month) and “infrequently” (less than once per month)

## C) Backup & Storage Requirements

- If all candidate collections are included, and expected growth rates prove accurate, the Rescue Repository will need storage capacity on the order of 30-35 terabytes. See Appendix C for projected storage requirements by file and media formats. *It would be wise to make allowance for the possibility of currently unforeseen collections as well.*
- Data backup procedure must include offsite backups, onsite backups and redundant drives. “Offsite” is defined here as outside of Yale and maintained by a professional secure facility. The Beinecke Rare Book Library expressed a strong interest for an ingest process which would write to at least two separate and independent media (either simultaneously or within a period of a few hours) in order to reduce the vulnerability inherent in a single-write/daily backup model.
- In case of system failure, a maximum downtime of 72 hours is required, with reasonable exception if retrieval from offsite backup is required. *A 24 hour maximum downtime is preferred.* Downtime is defined as a system state in which the operations of ingest, storage and/or retrieval are not available to a user.

#### D) Retrieval/Removal Requirements

- The retrieval function should be available nearly 24x7; short daily outages and reasonable exceptions are acceptable.
- The interface for the retrieval must be automated, with access authenticated via CAS, Windows Active directory or an equivalently secure mechanism, as noted in the “first principles’ section. It must present a hierarchical (directory or tree-like) structure to allow access to collection materials based on owner, collection, optionally sub-collection or project, optionally media (e.g. CD or tape) #, and Image ID# or filename.
- For removed items an audit must be supplied showing filenames, timestamps and a person identifier.
- Note: Frequency for the retrieval of materials from the Rescue Repository is envisioned as between “infrequently” (less than once per month) and “only in emergencies”
- Note: Frequency for the removal of materials from the Rescue Repository is envisioned as between “infrequently” (less than once per month) and “almost never”
- Note: The Rescue Repository is envisioned as a “safe haven” for digital materials, as opposed to acting as the primary content storage site for this collection’s day-to day-operations, or anywhere in between.

#### E) Access Control Requirements

- The system must base access to the Rescue Repository on three permission sets – one for each type of access:
  - Ingest
  - Deletion
  - Retrieval
- Note: For each collection surveyed, the number of people expected to need access to the Rescue Repository ranges between two and five.

#### F) Conclusion

In addition to the principles, requirements, outliers and specific notes listed above, two general observations are offered in conclusion:

- The ingest process, because of varying physical media, file formats and file-location systems will most likely require individually customized implementation by collection. Even with specialized hardware to help cope with the high numbers of CD’s involved, the optimal automation of this process, and the establishment of its attendant workflow, are likely to be major concerns in the implementation phase of this project. It should be emphasized that the value of this initial ingest is a key component of the Rescue Repository project, and is realized not only upon its establishment, but also with any subsequent data migration to future platforms able to use the Rescue Repository as a data source.
- Although group response was virtually unanimous in declaring the the purpose of the Rescue Repository to be essentially a “safe haven” for data, and expectations for retrieving materials ranged for the most part from “only in emergencies” to “infrequent”, it should be noted that none of the requirements in this document mandate this low level of retrieval service. The Rescue Repository is likely to play a larger role in normal retrieval operations for at least some collections if it is eventually perceived to be a more convenient data source. To the extent that this could constitute added value and will depend on the actual implementation, it should be among the considerations as the system is being designed and selected.