Next-Generation Technical Services (NGTS)
Power of Three Group 1, Lightning Team 1B

Merritt Gap Analysis

Final Report

August 30, 2012

POT 1, Lightning Team 1B Membership

Todd Grappone, UCLA (POT 1 Member and LT 1B Convener)
Eric Milenkiewicz, UC Riverside (POT 1 Member)
Stephen Abrams, California Digital Library
David Minor, UC San Diego

Table of Contents

Merritt Gap Analysis

Executive Summary

Merritt is a comprehensive preservation and access repository from the University of California Curation Center (UC3) at the California Digital Library (CDL). Merritt is used by many of the UC libraries and other campus content managers. The charge of Lightning Team 1B was to perform a gap analysis of Merritt relative to its use as: 1) the preservation repository of a UC Libraries systemwide DAMS with a discovery and display system, and 2) other functions as determined by POT 1, Lightning Team 1A. Additionally, LT 1B was charged to develop an inventory of current DAMS with discovery and display technologies utilized by UC campus libraries.

Given that LT 1B was asked to complete a gap analysis prior to the identification of a systemwide DAMS solution by POT 1, the team decided to use the Trustworthy Repositories Audit & Certification (TRAC)[1] guidelines as the foundation for analysis. Our intention was not to conduct a detailed review, but rather to focus on a high-level inquiry of the gap between Merritt and TRAC guidelines in two specific areas: technical analysis and policy analysis with a focus on Merritt as a preservation repository.

The conclusion of LT 1B is that Merritt would be an effective preservation repository for a UC Libraries systemwide digital asset management system. The team notes that while many UC campus libraries currently utilize Merritt, two campus libraries have preservation repositories in place that may also be suitable, including Chronopolis at UC San Diego (TRAC certified) and Fedora at UCLA. Worth noting is that the Merritt development team is currently engaged in integrating Merritt with two common CMS/DAMS frameworks: UC Berkeley's Research Hub, which is based on the Alfresco CMS; also, in collaboration with UCLA and Discovery Garden, CDL is working to integrate Merritt with Islandora accompanying the Fedora repository underlying the Drupal CMS.

Merritt Technical Analysis

TRAC certification is based on a repository's ability to manage a digital object from ingest through storage and preservation. In our investigation, the team interviewed members of the Merritt technical team based at the CDL, including Margaret Low (UC3 systems engineer) and John Ober (Manager, Infrastructure and Application Support). The questions focused on developing a picture of how Merritt operates, what the component pieces are, and how they map to TRAC. [See Appendix C for additional information regarding the Merritt infrastructure hardware and software.]

The conclusion of LT 1B is that Merritt meets the technical requirements for the preservation repository of a systemwide DAMS. [See Appendix E for a summary of Merritt preservation features.]

The team recommends that CDL increase its geographic replication by establishing a third repository copy outside of the Bay Area. (Currently, Merritt is automatically replicated between the UCOP administrative data center in Oakland and the UC Berkeley data center.) Furthermore, a formal verification of Merritt component systems discussed with the LT 1B was not within the scope of the team's charge. Based on its analysis, LT 1B feels that Merritt would be an effective preservation repository for a systemwide digital asset management system. If that decision is ultimately made, an ITIL best practice

[1] Trustworthy Repositories Audit & Certification: Criteria and Checklist
http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf

that should be followed is to conduct an Independent Verification and Validation process as the first step in that implementation.[2] [See Appendix A for additional information regarding the Merritt technical analysis.] Note that UC3 is performing a transparent TRAC self-audit. The preliminary results are available at https://confluence.ucop.edu/display/Curation/TRAC.

Merritt Policy Analysis

The TRAC criteria evaluate a repository's organizational infrastructure in terms of five facets: 1) organizational viability, 2) structure and staffing, 3) accountability and policy, 4) financial sustainability, and 5) contractual. Information about Merritt relative to all five facets was provided to LT 1B by the UC3 Merritt management team. Our analysis of the scope and detail of policy documentation currently available for Merritt as compared to that required for TRAC certification indicates a lack of formal documentation although most of the necessary policies and practices are actually in place. The additional policy documentation that would be needed for formal TRAC certification has been identified by the UC3 management team and is under development. During the process of this team's work with UC3 on the Merritt gap analysis, the Merritt management team made significant progress on developing policy and documentation using the TRAC checklist. [See Appendix B for additional information regarding Merritt policy.]

Our analysis indicates that Merritt currently lacks the complete formal, vetted policy required for full TRAC certification. Nonetheless, while there is a gap between Merritt TRAC readiness with respect to policy, basic information technology and preservation management practice used by Merritt is sound, e.g., backups occur, data loss is monitored, and security scans happen at regular intervals.

The conclusion of LT 1A is that the basic policy structure is in place for Merritt to be employed as a preservation repository for a UC Libraries systemwide digital asset management system. We encourage CDL's UC3 group to complete the administrative and policy structure for TRAC certification, and to create a succession plan on how data can be easily shared outside the Merritt preservation store.

Current Preservation Strategies

To further develop a picture of the preservation requirements for UC Libraries, LT 1B drew upon information gathered from the 10 UC campuses by POT 1 LT 3A regarding the need for a long-term preservation system. Additional information regarding Merritt provided by CDL's UC3 group was also used in the analysis presented below. Preservation solutions for digital collections currently in use across the UC Libraries include:

Campus	Preservation System	Notes
UCB	None	Select content in Merritt
UCD	None	Plan to use Merritt (especially for ETDs)
UCI	DSpace & Merritt	UCI actively uses Merritt, but Merritt doesn't completely meet their needs
UCLA	None	Migrating from a "home grown" system to one based on Islandora (Drupal+Fedora). UCLA library collaborating with UC3 and Discovery Garden to integrate Merritt with Islandora

[2] Independent Verification and Validation: https://en.wikipedia.org/wiki/Verification_and_validation

UCM	None	Local servers used for file backup; Merritt used for ETDs and select special collections
UCR	None	External hard drives used for file backup
UCSD	Chronopolis,	Potentially tied to Chronopolis; Merritt used for ETDs and integration with local DAMS facilitates automated transfer
UCSF	None	Local servers used for file backup; Merritt used for ETDs
UCSB	None	Local servers used for file backup; Merritt used for ETDs
UCSC	None	Local servers used for file backup; Merritt used for ETDs, select special collections, and Grateful Dead Archive

Note that while many campuses utilize Merritt for ETDs, few use the service to manage/preserve other digital collections and therefore are not viewed to be currently using Merritt as a preservation system.

All of the 10 UC campuses libraries demonstrated a need for the long-term preservation of digital content. Given that most of the campus libraries do not have a local solution in place, nearly all showed interest in utilizing a centrally deployed system. UC San Diego is the only campus currently tied into an existing solution other than Merritt (i.e., Chronopolis) and would need to consider carefully the benefits before moving to different system. While UC Irvine is using a combination of DSpace and Merritt, they feel that this combination does not completely meet their needs and they indicate being open to exploring additional options. Potential barriers to moving to a centrally supported system include cost, resources required to export from the current system, and network issues related to the transfer of large amounts of data. However, the campuses also pointed to several factors why a centrally deployed system would be desirable including:

- lack of trust in local system
- increased efficiency
- cost effectiveness
- flexibility

The findings of the POT 1 LT 3A survey make it clear that: 1) the UC Libraries are interested in a long-term solution for the preservation of digital content, 2) utilizing a centrally deployed system is an option, and 3) with some additional enhancements, Merritt can fulfill campus preservation needs. While most campuses pointed to Merritt as a viable solution, some respondents expressed that without certain enhancements Merritt would not fully meet their requirements. Features that the campuses would like to see in a long-term preservation system include:

- file format migration
- TRAC certification
- global editing
- drag-and-drop ingest
- statistical reporting
- easily accommodate multiple objects
- browse and search functionality
- ease of use for general user
- cost effective

In an effort to understand the Merritt development path, LT 1B requested further information from UC3 about the desired features and functionality surfaced by survey respondents. See Appendix D for the UC3 response and information relative to the Merritt development path.

Conclusion

The primary charge of Lightning Team 1B was to perform a gap analysis of the Merritt preservation and access repository relative to its use as the preservation component for a UC Libraries systemwide DAMS. Using the Trustworthy Repositories Audit & Certification (TRAC) guidelines as the foundation for technical and policy analysis, the conclusion of POT 1 LT 1B is that Merritt would be an effective preservation repository for a UC Libraries systemwide digital asset management system. From the user perspective, a survey conducted earlier this year by POT 1 LT 3A surfaced several features and functionalities that if developed hold the potential of significantly enhancing the Merritt user experience. UC3 is actively engaged in development efforts to address these concerns. There were no further requirements as identified by LT 1A.

APPENDIX A

Merritt Gap Analysis: Technical Analysis
Rev. 06/29/2012

TRAC certification is based on the ability of a repository to manage a digital object from ingest through storage and preservation, including: (1) fixity, (2) AIPs, (3) access security, (4) copies, (5) versioning, and (6) change management.

1. Fixity

In a preservation environment, there needs to be a way to monitor the status of an object – whether it has changed over time or not. The most common way of doing this is via fixity checking. This is most often done by computing and comparing checksums or hashes. The preservation system will generate a checksum or hash at an agreed-upon time in the ingest process. This checksum or hash will then be re-run and checked at specific intervals for the life of the object in the repository.

Merritt does fixity checking as part of its services. It has a separate Fixity micro-service that runs within the repository as described above. All objects processed by the Ingest micro-service are automatically registered with the Fixity service. Any discrepancies in checksums (none of which have occurred in over two years of production operation) are reported to Merritt managers in a nightly summary report.

2. AIPs

AIP is the acronym for Archival Information Package. This is a concept that comes from the OAIS specification (ISO 14721). It describes an object, along with accompanying information that can be preserved over the long term. It can be functionally or structurally different from the ingest and dissemination versions of the object.

Merritt includes the concept of AIPs in its preservation store. Merritt stores submitted objects in their original form, but augments that form with additional administrative and technical metadata produced during Ingest processing.

3. Access security

A good data center or repository will have tight control over who can access the content contained within.

All Merritt services and storage are hosted at the UC central administrative data center and the UC Berkeley data center, both of which confirm to industry standards and best practices for physical and information security.

4. Copies

Having multiple copies of objects (particularly AIPs) is an important facet to digital preservation. It is important that these copies are constantly checked and verified to make they have not changed individually and that they are appropriate replicas of each other.

Merritt does maintain multiple copies of preserved objects, and stores them in multiple data centers at different sites. Currently these sites are in geographic proximity to each other in the northern California bay area, however, and this is a concern for long-term preservation. CDL is working towards adding an additional replica site in southern California, and possibly in a commercial cloud, to address this concern.

5. Versioning

Versioning is the process of assigning unique names or numbers to unique states of objects. Versioning is often used for keeping track of incrementally different versions of electronic information, allowing for a number of functions, such as rolling back to a previous version of data.

Merritt is a strongly versioned repository system. Any change to object state, whether its data or metadata, automatically creates a new, uniquely identified version. All previous versions are available for retrieval through the Merritt UI or API.

6. Change management

Change management refers to a structured method of managing organizations, systems and people. Its goal is to produce a well-understood, auditable, and clearly delineated environment where changes are done according to clear plans. It is particularly important in the digital preservation environment where close track must be kept for the management of objects for the long term.

CDL has demonstrated that they have change management in place within their technical environments, for software development as well as object management. Responsibility for managing hardware and system-level software is shared by UC3, CDL central IT, and UCOP central IT. These groups have established procedures for coordinating their activities

APPENDIX B

Merritt Gap Analysis: Policy Analysis
Rev. 06/29/2012

The TRAC criteria [2] evaluate a repository's organizational infrastructure in terms of five facets: (1) organizational viability, (2) structure and staffing, (3) accountability and policy, (4) financial sustainability, and (5) contractual.

1. Organizational viability

Merritt [6] is a service offered to the UC community by the UC Curation Center (UC3) [8], one of five programmatic units of the California Digital Library (CDL) [2], a centrally-supported organization under the UC Office of the President (UCOP) [10]. The mission of UC3 is to "help researchers and the UC libraries manage, preserve, and provide access to their important digital assets." Merritt is part of a comprehensive suite of interoperating services provided by UC3 to ensure that the valuable digital resources supporting and resulting from UC's research, teaching, and learning activities remains available, usable, and authentic throughout the scholarly lifecycle – planning, acquisition/creation, preservation, publication, and discovery – for use, and re-use, now and into the future.

A formal succession plan for Merritt is not available at this time, but is in a formative stage of development by UC3. Merritt is based on a flexible data model in which all content information, both data and metadata, is fully expressed in a self-documenting manner on a replicated file system or cloud storage. The complete record of Merritt's holdings, and all of the information managed by Merritt about those holdings, can be retrieved by a full traversal of the archival file system using commonplace operating system command shell tools. The source code for the Merritt system is available under the BSD open source license and could be reconstituted by a third-party service provider.

Since UC3 is primarily a service provider with no direct curatorial responsibilities, it does not have a collection policy for Merritt in the traditional sense. However, Merritt is fully agnostic to content genre, type, and structure, and imposes no prescriptive eligibility requirements for submission. Thus, Merritt is an appropriate curation environment for all content acquired or created through the implementation of local campus collection policies and practices.

2. Structure and staffing

Merritt is a service offered to the UC community by the UC Curation Center (UC3), one of five UC3 is currently staffed with 15 FTE filling the following roles: administration, outreach, service/product management, project management, metadata analysis, software development, and operational support. UC3 relies on CDL-provided service units for marketing, user experience design, assessment, and infrastructure support; and the UCOP central Information Technology Services (ITS) group [9] for system and storage administration, and service hosting in the UCOP administrative data center.

UC3 staff are widely recognized for their expertise by the preservation and curation community, particularly in the areas of sustainability, identifiers/citation, metadata, representation formats, data curation, and web archiving. They publish widely and are regular participants in international conferences, symposia, and other relevant events [7]. UC3 is a corporate member of several important curation organizations, including the DataCite consortium, the International Internet Preservation Consortium (IIPC), the National Digital Stewardship Alliance (NDSA), the Preservation and Archiving

Special Interest Group (PASIG), and PrestoCentre; and at the CDL-level, the Coalition for Networked Information (CNI), Council on Library and Information Resources (CLIR), Digital Library Federation (DLF), EDUCAUSE, HathiTrust, International Coalition of Library Consortia (ICOLC), National Information Standards Organization (NISO), OCLC/RLG, Open Content Alliance (OCA), and the Scholarly Publishing Academic Resources Coalition (SPARC).

3. Accountability and policy

The primary designated community for Merritt is the University of California, including its libraries, museums, archives, academic departments, and research centers and laboratories. UC3 also makes Merritt available for use by non-UC governmental, commercial, non-profit, and private institutions and organizations.

Merritt's preservation policy can be summarized as an obligation by UC3 to expend its best efforts towards providing the highest level of preservation service, as defined by commonly-accepted community standards and best practices, that is consistent with the form, structure, and packaging of the managed digital content, the degree to which that content is accompanied by authoritative and comprehensive metadata, and the availability of appropriate tools. Note that this implies a continuum of preservation outcomes dependent on the nature of the content submitted by campus curators and collection managers, although at a minimum Merritt will always provide bit-level preservation of all content as a baseline practice. The open ended nature of this preservation policy is the natural consequence of Merritt's eligibility policy of not enforcing any prescriptive requirements for content submission. However, UC3 does provide consultation and guidance on ways to acquire or create digital content in a manner that is most amenable to the highest level of future preservation service [3].

Merritt operational procedures are documented on an internal UC3 wiki that retains full version history. The software components of the Merritt system are managed in a source code repository that similarly retains full version history. Changes to system hardware are tracked by the CDL Infrastructure and Application Services (IAS) group and the UCOP ITS, which manages the data center in which Merritt is hosted.

UC3 documents its Merritt-related activities with full transparency, relying on a mixture of a public website [6], wiki [5], blogging, webinars, and frequent email contact with its customers and stakeholders.

Merritt insures the bit-level integrity of its managed digital assets through a combination of storage redundancy and replication and file-level message digests verification throughout all of its workflows. All archival storage arrays used by Merritt are configured at RAID 6 in order to be tolerant of the simultaneous failure of two independent disks. All Merritt content is automatically replicated to a second, independent data center, and UC3 is investigating the possibility of expanding that replication to a third data center. Merritt's content submission API accepts an optional submission package digest from the contributing agent. All internal transfers of content into and out of the archival storage micro-service are accompanied by digest verification. Furthermore, all submitted content is automatically registered with the fixity micro-service, which performs a comprehensive periodic validation of all file-level digests on a two week cycle. Any bit-level damage that is uncovered (none of which has occurred to date) is repaired by copying from the relevant replica.

4. Financial stability

Historically, UC3 has funded its activities as a line item on the CDL budget, augmented with significant income from external granting agencies for specific research and development efforts. In the future, UC3 will be shifting its production services to a partial cost recovery model. In order to provide flexibility to its users, UC3 is investigating both pay-as-you-go and paid-up pricing structures for Merritt [1]. While the full analysis is not quite complete, it is likely that Merritt pricing for UC customers will be based primarily on recovering storage costs; other aspects of operational costs with be subsidized from CDL sources, while development activities will be supported through external grants.

5. Contractual

UC3 secures and maintains agreements with contributing campus units in terms of digital asset submission agreements (DASA) and inventories (DASI) that formally identify the institutional unit accepting curatorial responsibility, assert copyright status and other related intellectual property rights, indemnify UC in cases of third-party rights infringements, and establish the precise scope of UC3's curation obligation, possibly including the preservation, access, and redistribution of contributed content. The form of these agreements is currently under review by UC3; a revision offering a reformulated statement of reciprocal rights and obligations of all parties is anticipated shortly.

UC3 will respond to intellectual property rights challenges by establishing the bona fides of the claimant and if necessary working with the responsible campus curatorial unit to comply with valid access restrictions or takedown requests.

References

[1] Abrams, Stephen, Patricia Cruse, and John Kunze, "Total cost of preservation: Cost modeling for sustainable services," *CNI Spring 2012 Membership Meeting*, Baltimore, April 1-3, 2012, http://www.cni.org/topics/digital-curation/pay-once-preservation-forever/.

[2] California Digital Library, *California Digital Library*, http://www.cdlib.org/.

[3] California Digital Library, *CDL Guidelines for Digital Objects (CDL GDO)*, August 2011, http://www.cdlib.org/services/dsc/contribute/docs/GDO.pdf.

[4] Consultative Committee for Space Data Systems, *Audit for Certification of Trustworthy Digital Repositories*, CCSDS 652.0-M-1, Magenta Book, September 2011, http://public.ccsds.org/publications/archive/652x0m1.pdf.

[5] University of California Curation Center, *Curation Wiki*, https://confluence.ucop.edu/display/Curation/Home.

[6] University of California Curation Center, *Merritt*, http://www.cdlib.org/services/uc3/merritt.

[7] University of California Curation Center, *Publications, Presentations, and Webinars*, http://www.cdlib.org/services/uc3/resources/index.html.

[8] University of California Curation Center, *University of California Curation Center*, http://www.cdlib.org/uc3.

[9] University of California Office of the President, *Information Technology Services*, http://www.ucop.edu/irc.

[10] University of California Office of the President, *University of California Office of the President*, http://www.ucop.edu/.

APPENDIX C

Merritt Gap Analysis: Core Infrastructure Hardware and Software

Background:

On May 18, 2012 POT 1 LT 1B interviewed members of the Merritt technical team based at CDL, including Margaret Low (UC3 systems engineer) and John Ober (Manager, Infrastructure and Application Support). The purpose of the meeting was to gather information relative to Merritt's core infrastructure hardware and software.

Meeting Attendees:

Todd Grappone, UCLA (LT 1B member, meeting convener)
Colby Riggs, UC Irvine (LT 1B member, meeting notes)
Eric Milenkiewicz, UC Riverside (LT 1B member)
Stephen Abrams, California Digital Library (LT 1B member)
David Minor, UC San Diego (LT 1B member)
Margaret Low, California Digital Library
John Ober, California Digital Library

Merritt core infrastructure hardware and software:

What OS?

- Solaris 10 and SLES (Linux) 11

What hardware?

- Sun-Fire-V490
- Sun-Fire x4600
- Sun-Fire x4500

What other core software is used (i.e. Apache web server, Solr)

- Apache 2.2.22 - entry point to the services, load balancing
- Apache-ant-1.8.2 - building code
- Apache-maven 3.0.4 - building code
- Apache-tomcat 6.0.35 - container for Java application
- Java jdk1.6.031 - " " " " "
- Jenkins 1.457 - source control
- Mercurial 2.1 - " "
- Monit 5.3.2 - monitoring applications
- MySQL 5.0.95 – database for automated fixity checks
- Nexus 1.9.02 - version control
- OpenDS 2.2 (ldap) - authentication
- Ruby 1.8.7 - user interface
- Postgres 9.0.4 - data storage
- scala - scripting
- Zookeeper 3.3.1 - asynchronous message queue
- 4store 1.0.4 - –RDF quadstore for semantic metadata catalog

What software is used for backups?

- Backup software: Tivoli Storage Manager (TWM) and rsync 3.0.9
- Backup a distinction between object content and metadata
    - All information known to the repository is expressed in the file system.
    - Due to its size, object content is not amenable to the traditional backup, so it is replicated between storage arrays at the UCOP data center and the UC Berkeley data center using rsync
    - All other aspects of the systems subject to Tivoli Storage Manager (TSM) backup
        - Weekly full backup including cold backup of the MySQL, Postgres, and 4store databases
        - Nightly incremental backup
        - TSM as of 2 months ago going to a virtual tape library and UC Berkeley (2 copies) and San Diego Supercomputer Center (SDSC) (2 copies)
    - Merritt TSM backups on spinning disks with no tape involved but a virtual tape library
        - Object content running on enterprise quality storage arrays with RAID 6 (currently 2 copies, but would like more)
    - When does rsync happen?
        - A cron job runs daily
          Action: Confirm the frequency of the running of the cron job related to rsync - Stephen
        - What happens when two changes are made within the hour?
          Both changes will be captured, the versions will stack up there is no replacement

Location of backups and number of copies?

- TSM backups in virtual tape libraries (VTL) in the UCOP and UCB data centers
- replicated content on hokusai.cdlib.org:/dpr2repl/repository in the UCB data center

Where are the physical storage locations?

- Physical storage at UCOP and UCB data centers

What is the procedure for synchronization of copies? Is this procedure automated?

- Fixity verification for new content at the point in ingest
- RESTful message passing - each main Merritt service is an independent process that have very well documented APIs
- Ingest handlers that are involved iteratively (https://confluence.ucop.edu/display/Curation/Ingest)
  How are the changes in metadata handled?
    - Introducing a change in the primary content or metadata will automatically create a new version which is integrated into the storage service. To apply a change to an object must send the entire object. The object is sent to the storage service, which recognizes that it already exists so automatically creates delta files from old to new versions, only storing deltas between the versions. New files are automatically registered in the fixity service. Will enhance this process in August 2012 to submit only the change to streamline to process.

How is data loss handled? Who is notified?

- Fixity services continually running - All file level checksums are SHA-256
- Nightly statues reports are generated and distributed to Merritt managers
- When a problem is encountered is a report distributed immediately?
    - No, a problem would be disclosed in the nightly report

- A problem has never occurred

- Checksums may be supplied as part of object submission. If the object does not come in with one, a SHA-256 is automatically calculated
- At the storage array level do you get reports (e.g. bad sectors)?
  - Yes, review and screen the daily reports
    Action: In the storage array status reports what elements in the report are routinely monitored? Low-level reports are monitored by storage administrators in the UCOP center IT group. High-level summaries, including notification of any significant problems, are monitored by the CDL Infrastructure and Application Support group.
  - Handled by a storage groups at the Office of the President data center and there is also a collocation agreement with UC Berkeley

Is there a change management policy and procedure?

- Change testing is passed through an acceptance procedure which tries to keep the systems the same
- There is a change management policy for non-urgent changes which occurs 2 times a year
- Approval for changes occur in committee meetings, the Tech Council

How are changes to the hardware and software tested?

Software

- The developer pushes upgraded code to mercurial code repository, hg.cdlib.org. For some micro-services this will trigger an automatic rebuild and deployment of the executable onto a development machine using the Jenkins server. For other micro-services, the developer will build and deploy using the Jenkins server to a development machine. Testing/evaluation of the update are accomplished by the Merritt project managers and developers.
  Once upgrades are approved in development, the upgrades are promoted to the Merritt stage instance where additional testing is performed by the Merritt team and the Merritt community. The same procedures are taken to promote the code to the production instance.
  In order to standardize third-party binary code, Merritt uses a local Maven repository (Nexus) to manage and store artifacts. This allows the Merritt team to control the versions of the binary code as well as control Merritt artifacts shared across the micro-services.
  - Who approves the code?
    - The project lead Perry Willett - Users from outside the group will also review code on a staging server
    - Code development managed via user stories via Pivotal Tracker
  - For the micro-services there is a developer for each service which Perry (services) and Stephen (technical) manages the high level coordination
  - The tech managers meet every two weeks to work through conflicts and scheduling

Hardware

- Replacement on a five year cycle or the expiration of a vendor service contract
- Components are routinely swapped out
- Service architecture built in a high availability mode using two techniques 1) Veritas Cluster Server (VCS); and 2) Non-VCS with worker instances of services running on separate services monitored by the Apache load balancer
- Typically have 3 to 5 instances running on a server farm but are migrating to virtual machines and using Monit by the end of the year but will continue to run the Apache front-end to send worker processes.
- What are the triggers for hardware replacement?
  - Perform annual capacity planning and also through vendor support/service agreements

    - The System Admin checks the logs daily

How are security and critical updates applied?

- Upgrades are triggered by security concerns and Berkeley routinely performs security scans and passes along the vulnerabilities to John's group which reviews the vulnerabilities for urgency. UCOP data center is not performing routine security scanning
- IT security follows a strong set of practices including intrusion detection for system files password rules, no outside logins so must access via VPN, everything locked down with firewalls and ports are restricted.

Is there a hardware inventory and how is this managed?

- Excel spreadsheets in a file share
- Use Groundwork
- Use Tripwire

Are there documented disaster recovery procedures? Who is responsible?

- John's group responsible for disaster recovery
- Within the last year they have created disaster recovery process but it is only a quarter completed
- At a high level using UC Ready tool and Kuali ready tool for models
- Have not gotten to the part of creating procedures for one component or service
- Upstream dependencies have established disaster recovery plans

Has load testing been done?

- Everything in Merritt is expressed into a file system - ZFS Solaris file system
- Elasticity is built into the architectural design which allows a quick response

What is the connectivity capacity for the network?

- Office of the President on core node on CENIC and CALREN with 10 gigabit in and out

APPENDIX D

Requested Merritt Features: In an effort to understand the Merritt development path, LT 1B requested further information from UC3 about the desired features and functionality surfaced by LT 3A survey respondents.

LT 3A Survey Findings: Desire for Additional Merritt Features and Functionality	UC3 Response: Merritt Development Path
No file format migration	While no specific format migration workflows are currently implemented, UC3 is prepared to perform migrations if it is determined that content in Merritt is at risk of format obsolescence. All content files are characterized at ingest and technical metadata include MIME media types. UC3's general expertise in the area of digital formats is evidenced by leadership in the development of the JHOVE2 format characterization tool and the UDFR format registry. UC3 believes that it is premature to engage in detailed migration workflow design or implementation in advance of specific and credible preservation threats. Delaying implementation facilitates the use of evolving tools and community best practices.
No TRAC certification	CDL will be pursuing formal TRAC certification. UC3 is currently performing a transparent self-audit of Merritt documented at https://confluence.ucop.edu/display/Curation/TRAC.
No global editing option	As Merritt was initially designed as curation repository rather than a content management system, it does not support direct editing of cataloging metadata. However, UC3 has two current projects to integrate Merritt with external CMS/DAMS systems: (1) working with the UCB ResearchHub project on integration with the Alfresco CMS; and (2) working with UCLA and Discovery Garden on integration with Islandora, which will replace Fedora with Merritt as the repository underlying the Drupal CMS.
No drag-and-drop ingest	Merritt currently does not support drag-and-drop ingest, but this has been identified as a priority for the UCSF DataShare project. UC3 is in the requirements gathering phase of a project to add this functionality.
No statistical reporting	While statistical information cannot currently be requested directly through an administrative interface, internally Merritt does log relevant administrative and usage statistics.
Laborious when dealing with multiple objects	While Merritt does support batch submission of multiple objects, the structural packaging of batches is complicated. As mentioned above, UC3 is engaged in a project to simplify all aspects of the submission process, including drag-and-drop behavior and easier batch deposit.

Lacking browse and search functionality	Merritt supports object browsing and searching via metadata keywords. In an effort to provide a higher level of user experience for content discovery, UC3 is working with the CDL Publishing program to integrate XTF to provide a highly intuitive faceted discovery environment. The initial application of this new interface is for the UCSF DataShare project, which seeks to encourage reuse and sharing of biomedical datasets. Subsequently, the interface will be applied to all public Merritt collections.
Too complex for general user	UC3 is continually working on improving and simplifying the user experience for Merritt as well as training materials for its use. The concern regarding complexity is usually focused on content submission, particularly for batch submission. As noted above, UC3 is engaged in development work that promises to enhance and simplify the submission process significantly.
Not cost effective	While the use of Merritt is currently available at no charge, at some point later this year UC3 will start to operate Merritt on a partial cost recovery basis. For UC users, the service fee will be based on the amount of preservation storage that is consumed. UC3 is working with the SDSC to provide storage at the lowest possible price point for enterprise-quality storage. The anticipated price promises to be quite low. UC3 is also developing a paid-up price model that relies on a one-time, up-front price for a fixed term of preservation service.

APPENDIX E

UC3, Merritt, and Long-Term Preservation

Rev. 2012-02-10

The University of California Curation Center (UC3) believes that long-term digital preservation requires a comprehensive programmatic approach in order to be efficient, effective, and sustainable. The UC3 preservation repository is called Merritt. Merritt was developed using a new design paradigm known as micro-services, in which a comprehensive body of preservation functions are devolved into a granular set of small, independent, but highly interoperable micro-services. Using the micro-services approach, Merritt is able to support all of the desirable characteristics of a preservation infrastructure, providing high service availability, responsivity, reliability, efficiency, adaptability, and sustainability.

Merritt Preservation Summary

Organization	University of California Curation Center (UC3) of the California Digital Library (CDL)
Who can deposit	University of California and external content managers
Allowable content types	All content types, no prescriptive requirements; any content in any form is eligible
Submission methods	Single object and batch submission via UI or API
Persistent identifiers	ARK or DOI, resolved through N2T <http://n2t.net>
Discovery	Full-text search of indexed metadata or direct access via identifier resolution
Collections	Curatorially-defined collections
Versioning	version history is maintained; all prior version are directly retrievable
Storage	Multi-site replication between RAID-6 storage arrays
Fixity	Ongoing verification of cryptographic hashes
Architecture	Micro-services architecture
Codebase	Open source with fully documented specifications
Online availability	Operational on high-availability clusters with automated failover, nightly backup, and 24x7 monitoring
Certification	Launching an “Open TRAC” community certification process
Support	Online help and consultation with service managers

Data modeling

Merritt is based on a flexible data model capable of representing the widest range of digital objects and contextual metadata describing those objects. The data model is strongly versioned; any change in object state results in the creation of an entirely new version of that object, preserving the object’s chain of provenance over time. Any previous version can be easily re-instantiated upon request. Objects can be assigned to collections defined to meet various administrative and curatorial purposes. All information objects in the Merritt repository are provided with unique and persistent URLs by which they can be interrogated and retrieved. Digital content can be submitted to the Merritt Ingest service using a variety of protocols and workflows designed to minimize technical barriers.

Reliability

The Merritt infrastructure places user-facing interfaces and key shared resources, such as databases and storage, on high-availability, multi-node server clusters with automated failover; all other Merritt processes run as multiple load-balanced instantiations on an elastic server farm. This architecture ensures high overall service availability and, at the same time, high service performance, since the server farm can be quickly augmented in response to increased user demand. All Merritt services operate on servers in the UC administrative data center, with redundant power, cooling, and network connectivity. The services are subject to round-the-clock monitoring; any service interruption automatically triggers notification to the data center Server Operation Center and UC3 staff for triage and appropriate intervention.

The primary strategy for ensuring Merritt service reliability is the use of redundancy to avoid potential single points of failure. The adherence to redundancy extends across all aspects of Merritt system design and operational practice. The source code for the Merritt services is managed in a distributed source code repository with automated scripts for continuous integration and deployment. UC3 development practice emphasizes the use of standard programming languages and platform independent design patterns. All of the working file systems for Merritt services, with the exception of Storage service, are backed up nightly to tape as a contingency for disaster recovery and business continuity. The Merritt Storage service makes use of both localized and global redundancy in the form of RAID storage arrays, dynamic mirroring between arrays, and geographic replication. Every file managed within the Storage service has an associated cryptographically-secure checksum that is periodically recalculated by the Merritt Fixity service to detect bit-level corruption. If damage is discovered, it can be repaired by copying the necessary data from a verified replica.

Architecture

Long-term technical sustainability depends upon the ability of the infrastructure to evolve gracefully over time in response to changing conditions. The micro-services approach places a strong emphasis on service modularity and clean public interfaces. Adherence to these principles facilitates both the incremental enhancement and wholesale replacement of system components without impinging on overall service availability or established workflows. Since each micro-service is small and self-contained, they are collectively easier to implement, maintain, and enhance. Although the scope of any given micro-service is narrow, complex global behavior is nevertheless an emergent property of strategic combinations of these services. All of the Merritt micro-services will soon be publicly available for download, evaluation, and deployment under an open source license. The specifications for all services and their subcomponents, also publicly available, have undergone significant community review. An important validation of the Merritt approach has been demonstrated by a number of independent implementations of key specifications and services.

Preservation planning and support

As mentioned previously, positive preservation outcomes require more than just technical systems; enduring preservation solutions rely on significant human expertise and actions. Merritt preservation activities include the publication of best practice guidelines for preservation management, with recommendations on content creation and identification, and the use of preservation amenable formats, metadata practices, and packaging standards; the development of preservation action plans for dealing with the myriad potential risks to the long-term usability of preserved content; ongoing technology watch to proactively identify incipient obsolescence and other disruptive changes in the wider technological environment; and stakeholder engagement to keep abreast with the evolution of user expectation and

practice. Consultation is available to help assess user requirements and design appropriate solutions in all areas of digital content creation, management, preservation, and use.

Through 2012/2013, UC3 will also undertake an “open community audit” of Merritt following the Trustworthy Repository Audit and Certification (TRAC) checklist. We will use a public wiki to post documents, allowing the community to chart our progress and comment on our policies and practices. With this open process, we will provide an up-to-date view of the policies, resources, infrastructure, and technology that comprise Merritt services.

UC3

UC3 staff are internationally recognized for their leadership in the preservation field, with particular depth in persistent identifiers, metadata, formats and format characterization, organizational and programmatic sustainability, trust and certification, and web archiving. Staff members actively participate in a number of important national and international organizations, initiatives, and standardization efforts.

The University of California Curation Center is a creative partnership bringing together the expertise and resources of the California Digital Library, the ten UC campuses, and the international curation community. Together, the UC3 partnership provides innovative curation solutions to its campus constituencies and external partners.

http://www.cdlib.org/services/uc3/docs/UC3-Merritt_Preservation.pdf
http://www.cdlib.org/uc3
http://merritt.cdlib.org/
uc3@ucop.edu