UC3, Merritt, and Long-Term Preservation

Rev. 2012-03-01

The University of California Curation Center (UC3) believes that in order for long-term digital preservation to be efficient, effective, and sustainable, a comprehensive programmatic approach is necessary, matching significant human competencies, analysis, and decision making with a robust technical infrastructure.

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 implemented by a granular set of small, independent, but highly interoperable micro-services. Using the micro-services approach, Merritt supports all of the desirable characteristics of a preservation infrastructure, providing high service availability, responsivity, reliability, efficiency, adaptability, and sustainability (see Table 1).

Service provider:	University of California Curation Center (UC3) at the California Digital Library (CDL)
Curatorial eligibility:	Open to all University of California and external content managers
Content types:	No prescriptive requirements; any content in any form is eligible
Submission:	Single object and batch submission via UI or API
Persistent identifiers:	ARK or DOI, resolved through N2T
Content discovery:	Full-text search of indexed metadata or direct access via identifier resolution
Access control:	Curatorially-designated public or restricted access
Collections:	Curatorially-defined collections
Versioning:	Full 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 BSD license with fully documented specifications
Online availability:	Operational on high-availability clusters with automated failover, nightly backup, and 24x7 monitoring
Support:	Online help and consultation with service managers

Table 1 – Summary of Merritt repository features

Data modeling

Merritt is based on a flexible data model capable of representing the widest possible 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 of arbitrary complexity can be submitted to the Merritt Ingest service using a variety of protocols and workflows designed to minimize technical barriers.

Online availability

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 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 the 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, and geographic replication between the UC administrative datacenter and campus datacenters at UC Berkeley and UC San Diego. 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 a BSD open source license. (See Table 2.) 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.

Curation	*Adding value*	***Annotation***		*In planning*
		Notification		*In progress* (ATOM)
	Providing utility	***Access***		*In progress* (XTF)
		Transformation		*In planning*
		Index/search	✓	*In production*
		Ingest	✓	*In production*
Preservation	*Maintaining context*	***Characterization***		*In progress* (JHOVE2)
		Inventory	✓	*In production*
	Protecting state	***Replication***	✓	*In production*
		Fixity	✓	*In production*
		Storage	✓	*In production*
		Identity	✓	*In production* (EZID/N2T)

Table 2 – Merritt micro-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.

UC Curation Center (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. Digital curation encompasses curatorial management, preservation, and access, which are complementary activities: preservation ensuring access *over* time while access depends upon preservation *up to a point* in time.

The Merritt infrastructure is named for Lake Merritt, a prominent landmark close to the UC3 offices in Oakland, California. Lake Merritt was the first official wildlife refuge in the United States and is a designated National Historic Landmark.

http://www.cdlib.org/uc3
http://merritt.cdlib.org/
uc3@ucop.edu