# CDL Merritt Preservation System

## Operational Excellence Award submission

CDL's Merritt Digital Preservation System provides long term digital preservation for content from the ten UC campuses, their affiliated organizations and for the Dryad data repository. Since its inception in 2012, the Merritt team has strived to provide a low cost, bit-level preservation solution that enables the UC community to archive, manage and share its digital content. However, beginning in 2019, a significant series of changes to the system have accelerated the success of the repository such that in 2022 it now manages close to two and a half times the amount of data it did three years ago – and it does so through more *reliable, transparent, efficient and cost-effective* means, all purposed to both mitigate the risk of data loss and provide content access in perpetuity.

The list that follows outlines these changes, associated initiatives and their results.

- **Merritt's usage cost lowered from $650/TB to $150/TB, while the number of replicas for every digital object in the system increased from two to three copies.**
    - At the beginning of 2019, the majority of collections in Merritt incorporated two object copies and the storage cost per TB of data was set at $650. Through evaluation of new cloud storage providers and technology, the team decided to migrate Merritt's content off of OpenStack Swift storage to a newer, more economical Qumulo storage solution provided by SDSC. We then replicated the corpus to another low-cost cloud storage provider, Wasabi. Both Qumulo and Wasabi storage adhere to industry-leading data durability metrics.
    - Given the above two actions, plus an existing content copy in AWS Glacier, the total cost per TB for storage (provided at-cost) is now $150 for three object copies – a savings passed directly to end users.
    - The combination of the above factors allows Merritt to attain the gold standard of risk mitigation strategies for digital preservation – three object copies stored with three different service providers, across two separate geographic regions (U.S. East and West coast), with at least one copy being in nearline or offline storage technology.
- **The standing file audit cycle was reduced to a third of the time necessary to fixity check all data in the repository.**
    - A critical operation in all bit-level digital preservation repositories is file fixity checking. This ensures detection of bit rot before it's too late to address. With three digital object copies in the cloud, the system's cycle time for auditing content had grown to 150 days. Through an innovative approach that examines bitstreams in memory, we've reduced that cycle to approximately 50 days (66% decrease), even with the increased growth in holdings.
- **The rate of new deposits has increased significantly since 2019, such that holdings have more than doubled.**
    - Since 2019, holdings in the repository have increased from 120TB for a single object copy, to 295TB as of May 2022.
    - Taking into consideration all three object copies that the system manages at a file level, Merritt is now managing close to 1PB of data.

- **Reduction of the number of local data copies in play across cloud storage and microservices has increased the rate of new content ingest.**
    - As a system, Merritt incorporates nine microservices, each of which runs in a high availability configuration. By introducing a shared cloud-based file system, running fixity calculations on-the-fly and moving specific operations directly into the cloud, we've reduced the number of local host copies utilized by the microservices. Altogether these changes have made Merritt's ingest process 4 times faster than it was, as the system can now move more than 4TB/day.
- **Increased sustainability achieved through implementation of system assertions and added system transparency.**
    - As a digital preservation system, Merritt is focused on the sustainability of content produced by the UC community. The Merritt Team undertook a number of initiatives to improve the sustainability of the system itself to ensure it can be managed and maintained by the current team as well as by future team members.
    - To this end, an administrative layer that sits over the system has been implemented. It enables the team to monitor and interact with recently processed and processing content respectively. It has allowed us to build a series of assertions, or tests, that run nightly and deliver a system consistency report each morning. Highly granular data at byte, file and object level is automatically presented for review. These data display the results of assertions associated with system health while queueing, ingesting, replicating, and auditing content, as well as with its ability to provide object access.
    - Beyond these daily reports, any individual on the team can also accomplish a wide range of tasks with Merritt's admin layer – from more easily diagnosing operational issues, to correcting those that may arise through automated actions, to providing up-to-date content metrics for our users, to extending the repository's configuration by setting up new collections and adjusting collection configuration. All of these interactions play key roles in promoting the sustainability of the content that the system manages.
- **Steps taken toward system scalability, including the streamlining of microservice configuration.**
    - A longer term goal for the Merritt system entails an ability to autoscale its microservices accordingly, in times of both heightened and reduced submission loads. Two steps have been taken toward this goal, each of which have important side effects of bolstering system security.
    - Merritt's microservices make use of extensive configuration data. By moving this data into a central, cloud-based parameter store the team has not only streamlined configuration and enabled dynamic, runtime configuration changes, but it has also introduced enhanced security for the data itself.
    - Builds for Merritt's microservices have been refactored in order to enable simplified and more frequent application of updates across the system, thus increasing our ability to readily deploy security enhancements among the many dependencies used throughout Merritt's codebase.