Assessment Information

CoreTrustSeal Requirements 2017–2019

Repository:	UC3 Merritt
Website:	https://merritt.cdlib.org/
Certification Date:	07 August 2018

This repository is owned by: **California Digital Library**

UC3 Merritt

We have read and understood the notes concerning our application submission.

True

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:

CORE TRUSTWORTHY DATA REPOSITORIES REQUIREMENTS

BACKGROUND INFORMATION

Context

R0. Please provide context for your repository.

Repository Type. Select all relevant types from:

Institutional repository, Publication repository, Library/Museum/Archives, Research project repository

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept.

Comments

Merritt <https://merritt.cdlib.org/> is a general-purpose digital repository supporting curation and preservation services for the entire University of California community. It imposes no prescriptive eligibility requirements and holds significant collections of digital material in all genres: arts, humanities, sciences (life, physical, and social). Historically this content originated primarily from UC's libraries, archives, and museums, but in recent years, particularly with the advent of Merritt's companion data publication platform, Dash <http://dash.ucop.edu/>, Merritt also holds significant research data. It also manages preservation copies of all publications from CDL's e-Scholarship institutional repository <https://escholarship.org/> and the archival materials from CDL's Online Archive of California <http://oac.cdlib.org/>. The latter holds content from UC but also public libraries, archives, museums, and other local cultural memory organizations from around the state. While Dash provides the external appearance of being a repository in its own right, it is really just a lightweight overlay layer sitting on top of Merritt and providing alternative submission and discovery interfaces optimized for use by individual scholars and researchers, rather than the institutional librarians, curators, and archivists who continue to rely upon Merritt’s native interfaces. For purposes of this audit, Merritt and Dash are considered complementary components of a single comprehensive environment for ensuring long-term preservation of and access to the valuable digital content of the University of California, whether research data or otherwise.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept.

Brief Description of the Repository’s Designated Community.

Use of Merritt <https://merritt.cdlib.org/> and its companion Dash research data publication portal <https://dash.ucop.edu/> is open to all faculty, students, and research staff of the entire 10-campus University of California (UC) system for contribution, discovery, and retrieval. Thus, its primary designated community encompasses scholars from all disciplines, spanning the arts, sciences, humanities, and professions. The University also has a strong public service mission and Merritt manages cultural heritage materials on behalf of local public libraries, historical societies, and

other cultural memory organizations across the state of California. All of these stakeholders also constitute the repository's designated community, as are academic and public users external to UC for material in curatorially-designated public collections, and all Dash collections, which are explicitly public by stated policy. While Dash collections are inherently public, Merritt collections can be designated as public or restricted by their collection managers. Much of the Merritt corpus is restricted, although the material is largely available for retrieval through other CDL-supported discovery and access systems, such as the eScholarship institutional repository <https://escholarship.org/> for scholarly publications and the Calisphere portal <https://calisphere.org/> for archival materials. Non-UC-affiliated researchers in the earth and environmental sciences also can contribute data to ONEShare, a separately-branded collection underneath the Dash UI open for public submission, search, and retrieval <https://www.dataone.org/oneshare-member-node> and operated by CDL in collaboration with the DataONE project <http://dataone.org/> and the University of New Mexico Libraries <https://library.unm.edu/>; these researchers constitute another specialized designated community.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept.

Level of Curation Performed. Select all relevant types from:

B. Basic curation – e.g. brief checking; addition of basic metadata or documentation, C. Enhanced curation – e.g. conversion to new formats; enhancement of documentation, D. Data-level curation – as in C above; but with additional editing of deposited data for accuracy

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept.

Comments

Comments: While the scale of Merritt collections (currently over 2.7 million resources, 41.8 million files, and 93.2 TB) precludes detailed uniform curation review of all submissions, Merritt administrators do review, correct, and enhance contributed content as necessary. Dash supports a standard feature by which campus-based curators can and do review submissions for quality and enhancement.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept.

Outsource Partners. If applicable, please list them.

Merritt's external partnerships are shown the in diagram at <https://confluence.ucop.edu/download/attachments/180060250/Merritt-outsource-partnerships.pdf>. Merritt relies on other CDL services for identifier management and external service providers for storage and VM hosting. Dash also relies on external service providers for identifier management.

Merritt uses CDL's EZID service <http://ezid.cdlib.org> for assigning, managing, and resolving persistent ARK identifies for all managed objects. Dash DOI identifiers for datasets come from DataCite <https://www.datacite.org/>, an international non-profit membership organization, of which CDL is a founding member, for allocation and registration of DOIs. The DataCite statutes governing the contractual relationship between CDL and DataCite are available at <https://datacite.org/documents/DataCite_Statutes_officialTranslation_26February2016_final.pdf>.

As shown in the diagram <https://confluence.ucop.edu/download/attachments/180060250/Merritt-outsource-partnerships.pdf>, Merritt incorporates a storage-broker architecture that permits it to rely upon three third-party service providers for preservation storage. One is internal to the University of California (UC) system: the San Diego Supercomputer Center (SDSC)'s cloud storage <http://www.sdsc.edu/services/it/cloud.html>. SDSC has not been evaluated as part of any TDR assessment, but is routinely subject to Nessus scans <http://www.tenable.com/products/nessus-vulnerability-scanner>, a professional auditing service that probes for vulnerabilities and malware. The service level agreement defining the terms of the contractual relationships between CDL and SDSC is available at <https://www.cdlib.org/services/uc3/policies/SDSC-cloud-storage-SLA.pdf>.

Merritt also relies on a non-UC commercial service provider, Amazon web services (AWS), for preservation storage, using S3 <https://aws.amazon.com/s3/> and Glacier <https://aws.amazon.com/glacier/>, database hosting using RDS <https://aws.amazon.com/rds/>, and virtual server hosting using EC2 <https://aws.amazon.com/ec2/>. AWS complies with a number of regulatory and professional IT standards and certification programs <https://aws.amazon.com/compliance/>, including CSA, FERPA, FISMA, HIPAA, ISO 9001, 2701, 2017, SOC 1, 2, 3, and others. The service level agreements defining the terms of the contractual relationship between CDL and Amazon are available at <https://aws.amazon.com/agreement/>, <https://aws.amazon.com/s3/sla/>, and <https://aws.amazon.com/ec2/sla/>.

Merritt also relies on the University of New Mexico (UNM) for preservation storage for ONEShare collections <http://oneshare.cdlib.org/>. ONEShare is a data sharing portal freely available for public use by the earth and

environmental science research community <https://www.dataone.org/oneshare-member-node> developed by CDL in collaboration with DataONE <http://www.dataone.org/> and UNM. ONEShare is one of multiple UI-tenants supported by Dash. All ONEShare datasets are registered and indexed with the DataONE network for global high-level discovery <https://search.dataone.org/>. The guidelines governing the relationship between CDL and DataONE are available at <https://www.dataone.org/sites/all/documents/DataONE_MN_Partner_Guidelines.pdf>. The memorandum of understanding defining the terms of the relationship between CDL and UNM is available at <http://www.cdlib.org/services/uc3/policies/UC3-UNM-ONEShare-MOU-v2.pdf>

Dash relies on ORCID <https://orcid.org/>, an international non-profit membership organization, of which CDL is a member, for managing unique, persistent researcher identifiers, and Crossref's Fundref <http://www.crossref.org/fundingdata/>, for funding agency identifiers. The membership agreement defining the terms of the relationship between CDL and ORCID is available at <https://www.cdlib.org/services/uc3/policies/OCRID_Membership_Agreement_20151222.pdf>. Fundref data is available via an open public API <http://api.crossref.org> requiring no prior contractual relationship.

Curatorially-selected Merritt content, including all Dash managed datasets, are contributed to the Digital Preservation Network (DPN) <http://dpn.org/>, of which CDL is a founding member, where it is subject to additional replication to three geographically-dispersed and technologically heterogeneous repository nodes on the DPN network, selected from among the Academic Preservation Trust (APT) <http://aptrust.org/g/>, Duracloud Vault <http://duracloudvault.org/>, HathiTrust (HT) <https://www.hathitrust.org/>, Stanford Digital Repository (SDR) <https://library.stanford.edu/research/stanford-digital-repository> , and Texas Preservation Node (TPN) <https://www.tdl.org/dpn/> operated by the Texas Digital Library (TDL) consortium <https://www.tdl.org/>. (All Merritt and Dash content is already subject to Merritt-managed replication to two independent storage locations as described in response to Requirement R7.) The membership agreement defining the terms of the relationship between CDL and DPN is available at <https://www.cdlib.org/services/uc3/policies/dpn_depositagreement_cdl_2018.pdf>.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept.

Other Relevant Information.

Merritt currently manages over 2.7 million digital objects, represented in 3.5 million discrete states or versions, composed of 41.8 million individual files totaling 93.2 TB, expressed in 158 unique MIME types (as of 2018-06-26). This material is organized into 404 curatorial collections submitted by 35 curatorial units spread across all 10 University of California campuses, and is approximately 61% text, 27% application-specific, 11% static image, 0.2% sound, and 0.1% moving image.

Merritt is integrated with the DataONE network <http://dataone.org/> Merritt is represented on the DataONE grid as two Tier 1 member nodes, one allocated for UC-contributed data, the other forming the basis for the ONEShare open data repository jointly operated by CDL, DataONE, and the University of New Mexico, and open to public contributions <https://www.dataone.org/oneshare-member-node>.

While Merritt's companion research data publication portal, Dash <http://dash.cdlib.org>, gives the appearance of being a stand-alone repository in its own right, it is actually just a thin overlay layer integrated on top of Merritt and providing alternative, scholar-focused interfaces for publishing and searching research data collections. All data published in Dash is implicitly preserved in Merritt.

Merritt is registered with Re3data <https://www.re3data.org/repository/r3d100010747>.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept.

ORGANIZATIONAL INFRASTRUCTURE

I. Mission/Scope

R1. The repository has an explicit mission to provide access to and preserve data in its domain.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:

4 – The guideline has been fully implemented in the repository

Response:

The California Digital Library (CDL) exists to support the University of California (UC) community's pursuit of scholarship and extend the University's public service mission <http://www.cdlib.org/about/mission.html>. Within the CDL, responsibility for long-term digital curation, preservation, and research data management falls under the purview of the UC Curation Center (UC3) <http://uc3.cdlib.org/>. UC3 has offered preservation repository services (Merritt) to the UC community since 2005 and data publication services (Dash) since 2013, both in furtherance of the UC Libraries' strategic goal of open scholarship to "maximize discovery of and access to information resources" in order to "ensure that the cultural and scholarly record is preserved and accessible"<https://libraries.universityofcalifornia.edu/groups/files/about/docs/FY17-18_AnnualPlanAndPriorities_Final.pdf>. This is consistent with UC3's own mission "to provide transformative preservation, curation, and research data management systems, services, and initiatives that sustain and promote open scholarship" <http://uc3.cdlib.org/who/>.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept.

II. Licenses

R2. The repository maintains all applicable licenses covering data access and use and monitors compliance.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

Merritt and Dash terms of use conform to the general California Digital Library (CDL) terms of use <http://www.cdlib.org/about/terms.html>. Digital content submitted to Merritt through Dash is done so under the terms of a standard Creative Commons CC-BY license or CC0 public domain dedication, which prescribe minimal or no conditions on acceptable use. Material contributed to Merritt is covered by the terms of campus-level agreements granting CDL a non-exclusive, perpetual, but revocable license to store, copy, augment, federate, and, if so curatorially-designated, distribute for non-commercial use <https://wiki.ucop.edu/download/attachments/180060250/DPR_Submission_%20Agreement_%208-18-06_FINAL.pdf>.

Access to Merritt content is determined by the curatorially-assigned access control rules for the collection of which the content is a member, which permit designation for either: (1) restricted access by identified individuals (authenticated via Merritt user account information or campus institutional credentials and Shibboleth-conforming IdPs); or (2) unconstrained anonymous public access and use.

The Merritt terms of use <https://www.cdlib.org/services/uc3/policies/merritt_policies.html> state that use of Merritt constitutes acceptance by users of a commitment to abide by all applicable laws, regulations, policies, ethical concerns, and disciplinary best practices regarding the use of that content, including obligations regarding intellectual property rights, privacy, and accepted norms of scholarly discourse. CDL does not actively monitor usage to identify instances of noncompliance. However, users found to be exhibiting inappropriate behavior may be subject to loss of user privileges.

Dash datasets underlying articles being peer-reviewed may be designated for access restrictions for up to a six month period. During that time, no public data downloads are allowed, although certain minimal descriptive information - contributor(s), title, and date - will be presented on dataset landing pages.
Merritt and Dash are not appropriate repositories for managing content including clinical or personally identifiable information (PII) whose disclosure would constitute a violation of HIPAA/HITECH, FERPA, or other similar statutory, regulatory, or scholarly ethical regimes. It is the contributor’s responsibility to redact or anonymize content containing PII appropriately prior to submission to Merritt or Dash. Data contributed through Dash are monitored by campus-based curators for conformance to these restrictions and obligations.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept.

III. Continuity of access

R3. The repository has a continuity plan to ensure ongoing access to and preservation of its holdings.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

The University of California (UC) <www.universityofcalifornia.edu> has been in existence for 150 years, and is the nation's preeminent public institution of higher education, and among the world's best. The California Digital Library (CDL) is a core unit of the University's executive office of the systemwide President <http://www.ucop.edu/> with secure permanent funding. Merritt and Dash operate on a partial cost-recovery basis; while there is no service fee, data owners are assessed annually for the storage capacity that they have used. There is no pre-defined preservation period. The longevity of ongoing preservation management is contingent on payment of associated storage costs. As documented as part of its policies and procedures <https://www.cdlib.org/services/uc3/policies/merritt_policies.html>, Merritt and Dash content that is no longer being paid for is subject to de-accessioning or a transfer of curatorial responsibility to another UC entity at CDL's discretion. In the event that CDL is unwilling or unable to continue to offer Merritt or Dash as a service to the University community, it will work with content contributors and supervising curators to identify other curatorial institutions, within or outside the UC system, willing to take on future custodial responsibility. If that is not possible, CDL will return all content to its contributors at no added expense. Both of these contingency plans are publicly documented at <https://www.cdlib.org/services/uc3/policies/merritt_policies.html>.

External replicas of curatorially-designated collections, currently including all Dash data, are managed in the Digital Preservation Network (DPN) <http://dpn.org/>, a non-profit consortium of which CDL is a founding partner, providing distributed preservation services across a network of five technologically- and organizationally-independent repository nodes. The DPN business and sustainability plan is based upon an endowment pricing model intended to fund preservation management fully over a 20 year term. As stated in the published Merritt policies <https://www.cdlib.org/services/uc3/policies/merritt_policies.html>, at the end of that term, CDL will reassess the data in terms of its ongoing value and impact to determine its suitability for DPN renewal. In the event that CDL decides not to renew, DPN submission agreements <https://www.cdlib.org/services/uc3/policies/dpn_depositagreement_cdl_2018.pdf> include terms that provide for the legal transfer of stewardship responsibilities from members unable or unwilling to renew

for an additional preservation term to the DPN central organization or any of its individual members.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept.

IV. Confidentiality/Ethics

R4. The repository ensures, to the extent possible, that data are created, curated, accessed, and used in compliance with disciplinary and ethical norms.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

The CDL is a central service unit of the University of California (UC) and use of Merritt and Dash is open to community members on all 10 UC campuses. Primary responsibility for coordinating and reviewing the deposit of data into Merritt and Dash falls to individual campus library and research data management support staff. Dash supports features by which campus-based curators and RDM specialists can routinely review, assess, and enhance managed data. The Merritt and Dash terms of service <https://www.cdlib.org/services/uc3/policies/merritt_policies.html> highlight the obligation on the part of content contributors and consumers to comply with all appropriate legal, regulatory, institutional, and disciplinary requirements, policies, and ethical norms of scholarly best practice. Contribution of material to Merritt and Dash is taken as affirmation that the contributor accepts those terms of service obligations. CDL does not actively monitor contributions to Merritt to identify instances of noncompliance. However, users found to be exhibiting inappropriate behavior will be subject to loss of user privileges. Contributions to Dash are monitored by campus-based curators for conformance to

these restrictions and obligations, as well as for quality assurance review and curatorial enrichment.

Merritt and Dash currently are not appropriate repositories for data subject to FERPA. HIPAA/HITECH <https://www.hhs.gov/hipaa/> regulation of sensitive clinical or medical data, or other personally-identifiable information (PII) with disclosure risk. Merritt and Dash administrators actively remove inappropriate data as it is recognized.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept.

V. Organizational infrastructure

R5. The repository has adequate funding and sufficient numbers of qualified staff managed through a clear system of governance to effectively carry out the mission.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

The California Digital Library (CDL) is a central administrative unit of the University of California (UC) <www.universityofcalifornia.edu>, one of the world's premier public Universities. UC was founded in 1869 and now hosts over 264,000 students and 209,000 faculty and staff on 10 campuses, five medical centers, and three national laboratories, with an annual budget of $31.5 billion (FY 2016-2017), supported by federal and state government appropriations, public and private grants and contracts, endowment revenue, and tuition and fees <https://www.universityofcalifornia.edu/infocenter>. The CDL, founded in 1997, has an annual budget of $19.7M; the UC

Curation Center (UC3), the programmatic home for Merritt and Dash at the CDL, has an annual budget of $2.8M (FY 2014-2015). CDL budgeting is performed on an annual basis.

Merritt and Dash have no service fees, but Merritt users are billed for the recovery cost of preservation storage <https://confluence.ucop.edu/display/Curation/Merritt+Policies+and+Procedures>. Billing rates are set at the cost of provisioning storage by CDL's external storage providers at the San Diego Supercomputer Center (SDSC) and Amazon AWS, with an added 2.6% contingency surcharge to build up modest surpluses to be used in event of non-periodic or unanticipated expenses. The cost model underlying this billing structure is described at <https://confluence.ucop.edu/display/Curation/Cost+Modeling>. Preservation storage for the ONEShare public data sharing portal is provided to CDL at no cost by the University of New Mexico Libraries as part of its co-sponsoring agreement <http://www.cdlib.org/services/uc3/policies/UC3-UNM-ONEShare-MOU-v2.pdf>.

Merritt and Dash are supported by the UC Curation Center (UC3), one of four core programmatic units at the CDL. The UC3 team includes permanent, full-time, dedicated roles for Merritt and Dash service managers, a development manager, six frontend and backend programmers, and a DevOps/system administrator. UC3 also has access to the services of CDL's User Experience (UX) assessment and design team (8 FTE) <http://www.cdlib.org/services/uxdesign/> and Infrastructure and Application Support (IAS) team (4 FTE) <http://www.cdlib.org/services/infrastructure/>. CDL staff are widely recognized as international experts in their fields, and routinely participate in professional organizations, activities, and conferences, including the Research Data Alliance (RDA) <https://www.rd-alliance.org/>, Future of Research Communications and e-Scholarship (FORCE) <https://www.force11.org/>, CSV,conf <https://csvconf.com/> and PIDapalooza <https://pidapalooza.org/>, both co-organized by CDL, and Data Carpentry (DC) <http://www.datacarpentry.org/> and Library Carpentry (LC) <https://librarycarpentry.github.io/>, for which CDL is now hosting a US national coordinator position supported by a two-year grant from the Institute of Museum and Library Services (IMLS).

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept.

VI. Expert guidance

R6. The repository adopts mechanism(s) to secure ongoing expert guidance and feedback (either inhouse or external, including scientific guidance, if relevant).

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

Within the UC Office of the President, CDL falls under the purview of the vice provost for Academic Personnel and Programs. Systemwide accountability for CDL's operation is provided through frequent reporting to, and consultation with, the Council of University Librarians (CoUL), consisting of the 10 campus ULs and the CDL executive director <http://libraries.universityofcalifornia.edu/coul>, who meet monthly to discuss issues of systemwide policy and initiatives, and the Systemwide Library and Scholarly Information Advisory Committee (SLASIAC), consisting of representatives drawn from systemwide vice chancellors, vice provosts, CIOs, faculty, and the CDL executive director <http://libraries.universityofcalifornia.edu/slasiac>.

The CDL receives general guidance on its activities from the UC Libraries Advisory Structure (UCLAS) <http://libraries.universityofcalifornia.edu/about/advisory-structure> and its Direction and Oversight Committee (DOC) <http://libraries.universityofcalifornia.edu/doc>. CDL also relies on a number of systemwide Common Knowledge Groups (CKG) with broad campus participation by local library and RDM staff for more specific consultation and guidance, including the Born-Digital CKG <https://wiki.library.ucsf.edu/display/UCLCKG/Born+Digital+Planning+Group+CKG>, Data Curation CKG <https://wiki.library.ucsf.edu/display/UCLCKG/Data+Curation+CKG>, Digital Repository Metadata CKG <https://wiki.library.ucsf.edu/display/UCLCKG/Digital+Repository+Metadata+CKG>, and Preservation CKG <https://wiki.library.ucsf.edu/display/UCLCKG/Preservation+CKG>.

The CDL works closely with campus-based data librarians and data curation programs, particularly the Research Data Management group at UC Berkeley <http://researchdata.berkeley.edu/>, in which the CDL holds an advisory position, the UC Davis Data Science Initiative <http://guides.lib.ucdavis.edu/data>, the UC Santa Barbara data curation group <https://www.library.ucsb.edu/data-curation>, and the UC San Diego research data curation program <https://library.ucsd.edu/research-and-collections/data-curation/>.

The CDL holds institutional memberships in the Coalition for Networked Information (CNI) <http://www.cni.org/>, Council on Library and Information Resources <www.clir.org>, COUNTER <http://www.projectcounter.org/>, DataCite <http://www.datacite.org/>, Digital Library Federation (DLF) <http://www.diglib.org/>, Digital Preservation Network (DPN) <dpn.org>, EDUCAUSE <www.educause.edu>, HathiTrust <www.hathitrust.org>, International Coalition of Library

Consortia (ICOLC) <icolc.net>, International Internet Preservation Consortium (IIPC) <http://netpreserve.org/>, National Digital Stewardship Alliance (NDSA) <http://ndsa.org/>, National Information Standards Organization (NISO) <http://www.niso.org/>, OCLC <www.oclc.edu>, Open Content Alliance (OCA) <http://www.opencontentalliance.org/>, ORCiD <http://www.orcid.org/>, and Scholarly Publishing Academic Resources Coalition (SPARC) <http://sparcopen.org/>.

Specific guidance regarding the Dash open data publication service is provided by monthly meetings of UC data librarians (currently with 12 participating librarians) as well as a group of systemwide post-doctoral fellows and graduate students who routinely participate in testing and feedback exercises.
Merritt and Dash service managers routinely communicate with their designated communities through targeted email to the systemwide CDLINFO distribution list <https://www.cdlib.org/cdlinfo/>, email lists of campus collection administrators, and posts to the UC3 blog <http://uc3.cdlib.org/>

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept.

DIGITAL OBJECT MANAGEMENT

VII. Data integrity and authenticity

R7. The repository guarantees the integrity and authenticity of the data.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

All Dash datasets are automatically submitted to Merritt for long-term preservation. Content may be submitted to Merritt with optional cryptographic message digest values for individual file-level components <https://github.com/CDLUC3/mrt-doc/blob/master/doc/Merritt-ingest-service-latest.pdf>. If present, these digests are verified as part of ingest processing of Submission Information Packages (SIP). If individual files are not already associated with SHA-256 message digests, they are automatically assigned during the creation of the Archival Information Package (API). These digests are automatically verified following the transfer of the AIP from the ingest process to the archival storage process, by an immediate read-back. The digests are also verified on retrieval of the AIP for external distribution as a Dissemination Information Package (DIP).

All content managed in Merritt is replicated to distributed storage locations relying on heterogeneous technologies and media <https://github.com/CDLUC3/mrt-doc/blob/master/doc/Merritt-replication-service-latest.pdf>. A multi-threaded Audit process runs continually to verify the SHA-256 digests of all dispersed replicas <https://github.com/CDLUC3/mrt-doc/blob/master/doc/Merritt-audit-service-latest.pdf>. Canonical digest values for all versions of all dataset files are stored in Merritt's central Inventory database, which runs in a mirrored, cross-availability zone configuration in the Amazon AWS RDS cloud. Over the course of the past eight years of operation, bit-level damage was identified on less than half a dozen occasions and in all cases the damage was fully corrected by copying from verified replicas.

Individual replication sites also implement internal mechanisms for ensuring integrity and authenticity. Merritt relies on SDSC's UC private cloud service and Amazon's AWS S3 and Glacier commercial cloud services for preservation storage, which are federated through Merritt's storage broker architecture <https://github.com/CDLUC3/mrt-doc/blob/master/doc/Merritt-storage-service-latest.pdf>. SDSC Cloud Storage <http://www.sdsc.edu/services/it/cloud.html> implements an OpenStack Swift object store with three internal replicas on independent storage arrays, with ongoing MD5 verification of the validity of the replicas. AWS S3 and Glacier cloud storage maintain internal replicas and are designed to provide eleven 9's of reliability (99.9999999%) and sustain the simultaneous failure of any two internal copies. Merritt's Audit process provides external verification of all content independent of local internal methods at SDSC and S3. Glacier storage is not externally validated, however, as the AWS transactional pricing structure makes this financially prohibitive. All content stored in Glacier is also replicated at SDSC, which is subject to external validation.

Merritt is a strongly versioned repository. Any changes to data or metadata automatically results in the creation of a new version of the data object. Versioning relies on file-level backwards deltas to minimize duplicative file storage. Individual file-level components are never edited or replaced; new versions of files are added as components of the new dataset version. Similarly, files are never physically deleted; they are just de-associated from the newly created version, but not the prior versions to which they legitimately belong. All previous versions can be retrieved through the Merritt and Dash UI and API. Relationships between an object's various versions and files is maintained in Merritt's Inventory database <https://github.com/CDLUC3/mrt-doc/blob/master/doc/Merritt-inventory-latest.pdf>. Version provenance information -- effectuating agent, date/time -- is also stored in the database as well as being expressed in metadata files that form part of the object's archival (AIP) and dissemination (DIP) information packages. The database also holds each dataset's DataCite <http://schema.datacite.org/> and Electronic Resource Citation (ERC) metadata

<http://dublincore.org/groups/kernel/>; these metadata are also expressed in file form as part of the dataset's AIP and DIP.

Submission of new content to Merritt non-public collections requires prior authenticated login using Merritt user account credentials, which are managed in a three-node LDAP high-availability cluster <https://github.com/CDLUC3/mrt-doc/blob/master/doc/Merritt-LDAP-based-access-control-latest.pdf>. Submission of new data to Dash requires prior authenticated login using campus-issued credentials verified by campus Shibboleth-based identity providers (IdPs). Dash datasets are automatically submitted to Merritt in the context of Dash's administrative Merritt account, which is verified using LDAP-managed credentials over HTTP BasicAuth. As this communication occurs between servers behind the CDL firewall, these credentials are not at risk of public disclosure.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept.

VIII. Appraisal

R8. The repository accepts data and metadata based on defined criteria to ensure relevance and understandability for data users.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

By explicit policy <http://www.cdlib.org/services/uc3/policies/merritt_policies.html>, Merritt and Dash accept new data in any genre, format, or structure. CDL believes that the most significant impediment to the future use of preserved content is

not insufficiently-complete curation, but the lack of collection and management under an appropriate and proactive stewardship regime. Consequently, Merritt and Dash have been designed and are operated to maximize opportunities for self-service deposit of digital content. Once under secure management, this content is susceptible to ongoing review and enrichment by campus-based curators, collection managers, and RDM specialists to maintain and increase its curatorial value and provide a higher level of assurance of its ongoing availability and usability.

Dash submissions are accompanied by DataCite 4.0 metadata <http://schema.datacite.org/>. The Dash submission interface enforces the specification of only two of DataCite's mandatory elements: creator(s) and title, and one optional element: abstract; the other mandatory DataCite elements -- identifier, type, publisher, and publication date -- are added programmatically. Other optional elements -- funder, keywords, methodology, usage notes, related datasets/publications, and location (point, bounding box, or place name) -- may also be supplied, and their use is strongly encouraged. Dash's guidance on metadata is available at <https://dash.ucop.edu/stash/help#metadata>. Dash also endorses and encourages data contributors to follow the UK Data Service recommendations on formats <https://www.ukdataservice.ac.uk/manage-data/format/recommended-formats> and the DataONE recommendations regarding the form, structure, and description of data <https://www.dataone.org/sites/all/documents/DataONE_BP_Primer_020212.pdf>.

CDL provides general guidelines for recommended formats of content submitted to Merritt <https://www.cdlib.org/gateways/docs/GDO.pdf>. Merritt does not enforce any prescriptive eligibility requirements regarding associated metadata, but its submission interface provides the opportunity to supply Dublin Kernel metadata, internally serialized in the form of an Electronic Record Citation (ERC) <http://dublincore.org/groups/kernel/>, with optional elements for creator(s), title, date, and identifier(s). Additional metadata files may be freely included as part of the data package contributed to Dash or Merritt. Neither Dash nor Merritt perform any explicit check of dataset or metadata quality, beyond confirming the presence of required elements. Primary responsibility for developing appropriate data creation, acquisition, and description protocols and workflows is held by specialists at the individual campus library systems and RDM programs.

CDL's preservation policy <https://www.cdlib.org/services/uc3/policies/merritt_policies.html> obligates it to make reasonable efforts to provide managed content with the highest level of preservation assurance that is consistent with the form, structure, and packaging of the content, the degree to which that it is accompanied by authoritative and comprehensive metadata, the availability of appropriate tools, and other organizational priorities. This implies a continuum of preservation outcomes dependent upon the nature of the content. At a minimum, however, CDL is committed to providing bit-level preservation of all content. CDL offers 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

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:

Accept.

IX. Documented storage procedures

R9. The repository applies documented processes and procedures in managing archival storage of the data.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

Merritt's preservation policy is outlined in its Policies and Procedures statement <https://www.cdlib.org/services/uc3/policies/merritt_policies.html>. The Merritt repository is implemented in terms of a micro-services architecture <http://www.ijdc.net/index.php/ijdc/article/view/154>. The overall interoperation of the various micro-services are indicated in this architectural diagram <https://github.com/CDLUC3/mrt-doc/blob/master/img/diagrams/Merritt-Architecture.pdf>. Merritt relies on a primary strategy of replication to ensure the long-term integrity of managed data. All data is replicated to at least two geographically distributed locations and two heterogeneous technology stacks, currently at the San Diego Supercomputer Center (SDSC) <http://www.sdsc.edu/services/it/cloud.html>, which uses the OpenStack Swift platform, and Amazon AWS, which uses S3 and Glacier. Internally, the SDSC Swift cloud makes use of three independent replicas and its own internal digest-based auditing and self-healing capabilities. The S3 service description <https://aws.amazon.com/s3/> strongly implies that it also relies upon three independent replicas spread across availability zones with attendant internal fixity auditing and self-healing, and claims a 99.999999999% degree of durability.

The process of accepting SIPs and transforming them into AIPs for archival management is documented in terms of the individual micro-services implicated in that processing: Ingest <https://github.com/CDLUC3/mrt-doc/blob/master/doc/Merritt-ingest-service-latest.pdf>, which transforms the SIP into a conforming API; Storage <https://github.com/CDLUC3/mrt-doc/blob/master/doc/Merritt-storage-service-latest.pdf>, which dispositions the AIP to its final primary storage location at either the AWS S3 cloud or the San Diego Supercomputer Center (SDSC); Inventory <https://github.com/CDLUC3/mrt-doc/blob/master/doc/Merritt-inventory-latest.pdf>, which

retrieves newly acquired AIPs from Storage for parsing and populating the central metadata catalog and search index; Replication <https://github.com/CDLUC3/mrt-doc/blob/master/doc/Merritt-replication-service-latest.pdf>, which copies the AIP to its secondary storage location at either the San Diego Supercomputer Center or the AWS Glacier cloud; and Audit <https://github.com/CDLUC3/mrt-doc/blob/master/doc/Merritt-audit-service-latest.pdf>, which registers cryptographically-secure SHA-256 message digests for all replicas of all file-level API components, and subjects them to routine periodic verification. Reports of any discrepancies in the digests of stored replicas are automatically sent to Merritt administrators via email for investigation, triage, and if necessary, intervention. A reported discrepancy may not indicate actual bit-level damage; the temporary loss of online access to the remote cloud service providers is a legitimate trigger of nightly reports to administrators. The Merritt repository has been in operation since October 2010 without any data loss, and as of June 2018 manages over 2.7 million digital objects, represented in 3.5 million discrete versions and 41.8 million individual files totaling 93.2 TB. During that time, actual bit-level damage was identified on less than half a dozen occasions and in all cases the damage was fully corrected by copying from verified replicas.

All Merritt services (UI/API, Ingest, Storage, Inventory, Replication, and Audit) are provisioned with multiple instances running on independent VMs behind a front-end reverse-proxy load balancer. This configuration facilitates high-availability and high-performance operation and permits maintenance and upgrade activities to take place without any interruption to online availability. Individual service instances are sequentially removed from the load-balancer pool, modified, and then returned to the pool. Inter-service communication takes place by asynchronous messaging on subscribed-to message queues.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept.

X. Preservation plan

R10. The repository assumes responsibility for long-term preservation and manages this function in a planned and documented way.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

The Merritt terms of service <https://www.cdlib.org/services/uc3/policies/merritt_policies.html> make clear that the CDL is obligated to make reasonable efforts to provide managed content with the highest level of preservation assurance that is consistent with the form, structure, and packaging of the content, the degree to which that it is accompanied by authoritative and comprehensive metadata, the availability of appropriate tools, and other organizational priorities. There are no formal preservation “levels” per se; rather CDL assumes a continuum of preservation outcomes dependent upon the nature of the content. At a minimum, CDL is committed to providing bit-level preservation of all content. CDL offers 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.

Contribution to Merritt is taken as affirmative assent that the contributors are assigning to CDL the non-exclusive, perpetual, revocable right to save, copy, enhance, federate, create derivatives for purposes of long-term preservation, and provide access to contributed content, subject to curatorially-designated controls

Merritt maintains a complete change history of managed content as it may evolve over time. The repository relies upon a primary preservation strategy of replication of content to geographically-dispersed sites and technological heterogeneity. Merritt incorporates a process of continual verification of cryptographic message digests of all content replicas to detect and correct any bit-level damage. The Merritt repository has been in operation since October 2010 without any data loss, and as of June 2018 manages over 2.7 million digital objects, represented in 3.5 million discrete versions and 41.8 million individual files totaling 93.2 TB. During that time, actual bit-level damage was identified on less than half a dozen occasions and in all cases the damage was fully corrected by copying from verified replicas.

The design, implementation, and operation of Merritt are consistent with the community-accepted standard ISO 14721 Open Archival Information System (OAIS) reference model.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept.

XI. Data quality

R11. The repository has appropriate expertise to address technical data and metadata quality and ensures that sufficient information is available for end users to make quality-related evaluations.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

As described in response to R8, Merritt and Dash accept data in any genre, format, and package, as a consequence of CDL's belief that the most significant impediment to the future use of managed content is not insufficiently-complete curation, but the lack of collection and management under an appropriate proactive stewardship regime. Consequently the design and policy imperatives for Merritt and Dash place primary value on removing obstacles to content submission. However, UC3 fully recognizes the benefits accruing from added-value curation and works closely with data curators and RDM specialists on all ten UC campuses, who hold primary responsibility for consulting with affiliated researchers and reviewing and enhancing contributions to Dash from those affiliates. Dash offers a curation layer specifically to enable this type of distributed curation review. CDL has a pilot project underway with four campuses to formalize the campus-based data QA evaluation and augmentation by appropriate Library and Research IT curatorial experts. The lessons learned during the pilot will be applied in extending the program to the entire UC system. For deposit directly into Merritt, professional campus librarians and collections managers are responsible for upstream workflows to ensure metadata quality that meets their institutional retrieval requirements.

Dash supports the full DataCite 4.0 metadata schema <http://schema.datacite.org/>, which requires a descriptive abstract and provides opportunities to specify methodological statements, usage notes, and indicate external relationships to related datasets and publications.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept.

XII. Workflows

R12. Archiving takes place according to defined workflows from ingest to dissemination.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

Merritt and Dash are designed for self-service operation for content submission and retrieval. The nominal archiving workflow <https://github.com/CDLUC3/mrt-doc/blob/master/img/diagrams/Merritt-Architecture.pdf> for Merritt starts with transfer or one or more data files to the Merritt Ingest service <https://github.com/CDLUC3/mrt-doc/blob/master/doc/Merritt-ingest-service-latest.pdf>, which can occur through its online UI, REST API, or SWORD v2 protocol (for Dash-mediated deposits). Ingest processing is performed by the iterative invocation of individual ingest handlers, each focusing on a particular aspect of processing, including container disaggregation, digest verification, characterization, identifier assignment, and SIP-to-AIP translation. Ingest places messages on an asynchronous queue alerting the Storage and Inventory services of the availability of new content. The Storage service <https://github.com/CDLUC3/mrt-doc/blob/master/doc/Merritt-storage-service-latest.pdf> moves the content from a temporary staging space into its final primary storage location, either SDSC or AWS S3, under the direction of an ingest profile in which context the submission was made. The Inventory service <https://github.com/CDLUC3/mrt-doc/blob/master/doc/Merritt-inventory-latest.pdf> retrieves the content from Storage and parses its metadata files for inclusion in the Inventory database and search index. Part of the database representation includes a specification of the replication policy and fixity information. The Replication service <https://github.com/CDLUC3/mrt-doc/blob/master/doc/Merritt-replication-service-latest.pdf> runs continually and upon recognizing that the new content has not yet been replicated, will do so to its secondary location, either SDSC or AWS Glacier, again dependent upon the controlling submission profile. The Audit service <https://github.com/CDLUC3/mrt-doc/blob/master/doc/Merritt-audit-service-latest.pdf> also runs continually, cycling

through all registered content and retrieving and verifying SHA-256 message digests for all files of all versions of all replicas. Reports of any discrepancies are sent automatically to Merritt administrators for investigation, triage, and if necessary, intervention. All Merritt collections publish Atom feeds with the newest content on the first page. Feed metadata includes actionable links to managed content's landing pages. Content is also accessible through browsing or metadata search over contributor(s), title, and identifier metadata. A multi-threaded Audit process runs continually to verify the SHA-256 digests of all dispersed replicas <https://github.com/CDLUC3/mrt-doc/blob/master/doc/Merritt-audit-service-latest.pdf>.

Being an overlay layer, Dash interacts with the Merritt repository through APIs. Dash deposit packages, composed of researcher-contributed files along with DataCite metadata and additional structural/ administrative metadata in a Zip container, are sent to Merritt using the Simple Web-service Offering Repository Deposit (SWORD) v2 protocol <http://swordapp.org/sword-v2/>, where it is subject to standard processing as described above. Dash periodically retrieves metadata from Merritt using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) protocol <https://www.openarchives.org/pmh/>. This metadata includes the stable Merritt URLs for downloading dataset versions. Dash incorporates that information when displaying its own landing pages. DataCite metadata is indexed in Solr <http://lucene.apache.org/solr/> to support the Dash search function.

All code for Merritt and Dash services is managed in versioned-controlled GitHub repositories <https://github.com/cdluc3>. Maintenance and enhancement of the codebases relies on documented agile development practices with a user-centered design focus. Dedicated 100% FTE Merritt and Dash product managers gather stakeholder needs and stories and translate them into prioritized use cases, which are broken down into individual tickets managed in Pivotal Tracker (PT) <https://www.pivotaltracker.com/>. Work on the prioritized backlog is performed in two or three week time-boxed sprints, with incremental progress logged in the PT tickets.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept.

XIII. Data discovery and identification

R13. The repository enables users to discover the data and refer to them in a persistent way through proper citation.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

All objects managed in Merritt, including all datasets contributed through the Dash data publication service, are assigned unique, persistent ARK identifiers <https://tools.ietf.org/html/draft-kunze-ark-18> using CDL's EZID service <http://ezid.cdlib.org>. Datasets managed in curatorially-designated collections also receive DOIs from EZID, which relies in turn on DataCite <http://www.datacite.org>, for DOI allocation and registration. All Merritt object landing pages prominently display the object's actionable persistent identifier for use in citations. Landing pages in Merritt's companion Dash research data portal feature pre-formatted citations conforming to the 2014 FORCE11 Joint Declaration on Data Citation Principles <https://www.force11.org/group/joint-declaration-data-citation-principles-final>. For example:

Yu, Shengyang; Tward, Aaron; Knox, Sarah (2017), 10x Lacrimal Gland scRNA seq data matrix, UC San Francisco Dash, Dataset, https://doi.org/10.7272/Q6W37T8B

Dash provides search across all data based on indexed DataCite metadata: author(s), title, abstract, publisher, date, keywords, methods, usage notes, and geospatial points, bounding boxes, and place names. Search results can be refined by faceting on author, title, publisher, date, or type. Merritt provides search within collections based on indexed Dublin Kernel/ERC metadata: creators(s), title, date, and primary ARK identifier.

Merritt defines stable URL patterns for object landing pages, object downloads, version landing pages, version downloads, and file downloads of the general form:

https://merritt.cdlib.org/m/<objectid>
https://merritt.cdlib.org/d/<objectid>
https://merritt.cdlib.org/m/<objectid>/<versionid>
https://merritt.cdlib.org/d/<objectid>/<versionid>
https://merritt.cdlib.org/d/<objectid>/<versionid>/<fileid>

where <objectid>s are ARKs, <versionid>s are ordinal numbers 1,2,3…, with 0 a shortcut reference to the current version, and <fileid>s are pathnames (with colons ":" and slashes "/" URL-encoded). For example:

https://merritt.cdlib.org/d/ark%3A%2F13030%2Fm5qc02mk

https://merritt.cdlib.org/m/ark%3A%2F13030%2Fm5qc02mk
https://merritt.cdlib.org/d/ark%3A%2F13030%2Fm5qc02mk/3
https://merritt.cdlib.org/m/ark%3A%2F13030%2Fm5qc02mk/3
https://merritt.cdlib.org/d/ark%3A%2F13030%2Fm5qc02mk/3/cadwsap-s2400076-001-main.csv

are the URLs for a dataset landing page, dataset download, version 3 landing page, version 3 download, and main CSV download. Merritt ARKs (of the general form: http://n2t.net/<ark>) also resolve to Merritt object landing pages. For example:

http://n2t.net/ark:/13030/m5qc02mk

is the URL for the same dataset landing page. Dash DOIs resolve to Dash dataset landing pages. For example:

https://doi.org/10.7272/Q6W37T8B

Descriptive metadata for datasets associated with Dash DOIs is aggregated by DataCite where it is searchable from the central DataCite discovery service <https://search.datacite.org/>. Datasets in curatorially-designated collections, including the CDL/DataONE/University of New Mexico operated ONEShare public data repository for earth and environmental science, are automatically aggregated by DataONE coordinating nodes where it is searchable from the central DataONE discovery service <https://search.dataone.org/>.

Curatorially-designated content collections syndicate their holdings for automated harvesting via ATOM feeds accessible through a prominent button on collection landing pages. This function may be used by external value-added services building aggregated cross-repository collections and discovery mechanisms.

The Dash UI is optimized for search engine optimization (SEO) by including schema.org <http://schema.org/> JSON-LD <https://json-ld.org/> metadata as an HTML <script> block on dataset landing pages, as can be verified with Google's structured data testing tool, e.g., <https://search.google.com/structured-data/testing-tool/u/0/#url=https%3A%2F%2Fdash.ucop.edu%2Fstash%2Fdataset%2Fdoi%3A10.7272%2FQ6639MWG>.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept.

XIV. Data reuse

R14. The repository enables reuse of the data over time, ensuring that appropriate metadata are available to support the understanding and use of the data.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

In keeping with Merritt and Dash's explicit emphasis on removing barriers to more widespread adoption of self-service publication, prescriptive metadata requirements are minimal, although accompanied by opportunities for the inclusion of optional, but recommended, metadata. Dash supports the full DataCite 4.0 schema <http://schema.datacite.org/>. The mandatory contributor-supplied elements are title, author(s), institutional affiliation, email, and abstract. The DOI, type, publisher, and publication date are added programmatically. Other optional elements -- funder, keywords, methodology, usage notes, related datasets/publications, and location (point, bounding box, or place name) -- may also be supplied, and their use is strongly encouraged. Dash's guidance on metadata is available at <https://dash.ucop.edu/stash/help#metadata>. Dash endorses and encourages data contributors to follow the UK Data Service recommendations on formats <https://www.ukdataservice.ac.uk/manage-data/format/recommended-formats> [UK Data Service, 2018] and the DataONE recommendations regarding the form, structure, and description of data <https://www.dataone.org/sites/all/documents/DataONE_BP_Primer_020212.pdf>. CDL provide general guidelines for recommended formats of content submitted to Merritt <https://www.cdlib.org/gateways/docs/GDO.pdf>. Merritt does not enforce any prescriptive eligibility requirements regarding associated metadata, but its submission interface provides the opportunity to supply Dublin Kernel metadata serialized in the form of an Electronic Record Citation (ERC) <http://dublincore.org/groups/kernel/>, with optional elements for creator(s), title, date, and identifier(s). Additional arbitrary metadata, including domain-specific, may be freely incorporated as part of submitted Merritt or Dash data packages.

CDL's preservation policy <https://www.cdlib.org/services/uc3/policies/merritt_policies.html> does not prescribe any particular preservation strategy, which it believes must be carefully tailored to specific threats, use cases, and designated communities.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept.

TECHNOLOGY

XV. Technical infrastructure

R15. The repository functions on well-supported operating systems and other core infrastructural software and is using hardware and software technologies appropriate to the services it provides to its Designated Community.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

Merritt and Dash are hosted on Linux VMs in the Amazon AWS EC2 cloud environment. Their associated MySQL databases are hosted by AWS RDS. General operating principles for CDL's AWS usage is described at <https://wiki.ucop.edu/display/CUG/CDL+AWS+Environment+Orientation>, <https://wiki.ucop.edu/display/CUG/RDS+User+Guide>, and <https://wiki.ucop.edu/display/CUG/S3+and+Glacier+Storage+User%27s+Guide>. Consistent with CDL policy, all VMs are routinely updated to the latest stable AWS operating systems versions within six months of release, or within one week for critical security patches or 30 days for non-critical security patches.

Merritt and Dash are available on a nominal 24x7x52 high-availability schedule. All backend Merritt processing is performed by independent instances of the various Merritt services (UI/API, Ingest, Storage, Inventory, Replication, Audit) that are fully load-balanced so routine maintenance can occur without interrupting service availability. Infrequent maintenance on the one singleton process, the front=end load balancer, is scheduled with advanced two-week notification. The current status of Merritt and Dash availability can be found on the CDL system status page <http://www.cdlib.org/contact/system.html>.
Dash fully implements the DataCite 4.0 metadata community standard <http://schema.datacite.org/>. The OAI-PMH 2.0 metadata harvesting standard OAI, 2002] <https://www.openarchives.org/pmh/> is implemented without support for resumption tokens. The SWORD v2 repository submission standard <http://swordapp.org/> is implemented for creating a resource, and replacing a service, without support for continued deposit.

Merritt fully implements the Dublin Kernel/Electronic Record Citation (ERC) community metadata standard <http://dublincore.org/groups/kernel/>, OAI-PMH 2.0 metadata harvesting standard, and SWORD v2.

The Dash UI makes use of the community-supported open source Bootstrap framework <http://getbootstrap.com/> and Leaflet geospatial JavaScript library <http://leafletjs.com/>. Dash also incorporates the community-supported open source GeoBlacklight geospatial search engine <http://geoblacklight.org/>. Dash itself, as well as the Merritt UI, is a Ruby on Rails application <http://rubyonrails.org/> running on open source Passenger application servers <https://www.phusionpassenger.com/> behind NGINX proxy servers <https://www.nginx.com/> and Apache web server <https://httpd.apache.org/>.
Merritt incorporates the community-supported open source Jersey REST framework <https://jersey.java.net/> and Zookeeper coordination suite <https://zookeeper.apache.org/>. Individual Merritt micro-services are Java applications running on Apache Tomcat application servers <http://tomcat.apache.org/>; their source code and documentation are managed under revision control in public GitHub repositories <https://github.com/CDLUC3> and <https://github.com/CDLUC3/mrt-doc>.

Merritt infrastructure configuration is managed in a private Confluence wiki. Each service is documented in term of servers (hostname, CNAME(s), IP address(es), firewall rules), role accounts, file system layout, software inventory, process inventory, DB (if appropriate), scripts and cron jobs), log files, and monitoring regime. Server certificates are managed centrally by the CDL Infrastructure and Application Support (IAS) program.

CDL embraces agile work processes. The Dash development roadmap prioritizes features based on researcher needs and community standards <http://uc3.cdlib.org/2018/01/04/dash-2017-in-review/>. The roadmap items are provisioned through sprint-based two week releases <https://github.com/CDLUC3/dashv2/releases>. The entire codebase is made available under revision control in public Github repositories <https://github.com/CDLUC3/dash>, <https://github.com/CDLUC3/dashv2>, <https://github.com/CDLUC3/dash2-harvester>, <https://github.com/CDLUC3/stash>.

Reviewer Entry
Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept.

XVI. Security

R16. The technical infrastructure of the repository provides for protection of the facility and its data, products, services, and users.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

Merritt and Dash are deployed and operate on virtual machines in the Amazon AWS cloud. AWS is certified for ISO 27001 <https://d0.awsstatic.com/whitepapers/compliance/soc3_amazon_web_services.pdf> and BSI C5 <https://aws.amazon.com/compliance/bsi-c5/> security standards. AWS's comprehensive security controls, including those for both online and physical facility security, are described at <https://aws.amazon.com/compliance/data-center/controls/>.

All Merritt and Dash processes run on standard Amazon Linux VMs in the context of service-specific role accounts that are accessible only by individual UC3 staff members through explicitly-granted sudo privileges (sudo su - <role>). All grants of sudo privileges are logged and monitored by CDL system administrators. Personal login accounts on Merritt and Dash are limited to appropriate CDL staff and are accepted only via ssh from the IP range of CDL’s administrative office, its virtual private network (VPN), or by ssh from a hardened bastion server.

System deployments are controlled automatically using Puppet operating in a master/agent configuration <https://puppet.com/products/how-puppet-works>. Nagios <https://www.nagios.org/> is used for continuous monitoring of VM resource levels for disk space, load average, and swap usage. Disk space will trigger an email warning to system

administrators at 90%, and an alert when the disk is 95% full. Load average triggers a warning when the 1-minute load average is over 10, the 5-minute load average is over 15, or the 10-minute load average is over 20. Load average triggers an alert when he 1-minute load average is over 25, the 5-minute load average is over 35, or the 10-minute load average is over 45. Swap space triggers a warning when it is 90% used, and an alert when it is 95% used. Additional Nagios monitors are defined for key service functions to provide early notification in case of interruptions to online availability.

All AWS EC2 virtual servers are backed up nightly with a full instance snapshot retained for 35 days, with the first snapshot of each month retained for six months. Weekly machine image snapshots, including all data volumes and instance configuration information, are maintained outside of the primary AWS region. CDL central IT maintains backups of all AWS configuration information. Many resources, including IAM, security groups, S3 bucket configurations, Route 53 DNS records, and CloudFormation templates are under version control in a local git repository (itself backed up in an alternate region) <https://wiki.ucop.edu/display/CUG/CDL+AWS+Environment+Orientation>. MySQL databases, configured for mirrored operation in multiple RDS availability zones, are also backed up nightly and are retained for 35 days <https://wiki.ucop.edu/display/CUG/S3+and+Glacier+Storage+User%27s+Guide>. CDL-wide disaster recovery plans call for redeployment of services from image snapshots in the primary or another AWS region as dictated by necessity.

All CDL staff participate in mandatory annual cybersecurity training <https://security.ucop.edu/>.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept.

APPLICANT FEEDBACK

Comments/feedback

These requirements are not seen as final, and we value your input to improve the core certification procedure. To this end, please leave any comments you wish to make on both the quality of the Catalogue and its relevance to your organization, as well as any other related thoughts.

Response:

We very much appreciate the careful attention given to our application by the CoreTrustSeal reviewers, We have revised our application to address the comments and recommendations made by the reviewers.

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments: