Member Node Description: UC3 Merritt Repository

Version 1.1 2/1/13 John Kunze

General

Name of resource: UC3 Merritt Repository
URL(s): http://merritt.cdlib.org
Institutional affiliation(s): University of California Curation Center, California Digital Library
Primary geographic location: Oakland, California, USA
Project Director & contact info: Trisha Cruse, patricia.cruse@ucop.edu, 510-987-9016
Technical Contact & contact info: Stephen Abrams, Stephen.abrams@ucop.edu, 510-987-0370
John Kunze, jak@ucop.edu, 510-987-9231
Age of resource: Since March 2010
Funding support: Recharge
Unique Identifier: urn:node:CDL

Content

Content description/collection policy (1 paragraph, domain and spatial/temporal coverage, uniqueness of content, exclusions, as applicable):
A cross-domain repository for data and metadata provided by the California Digital Library at the University of California.

Types of data (complex objects, text, image, video, audio, other):
All data types accepted (spreadsheet, image, video, text, etc.).

Data and metadata availability (rights, licensing, restrictions):
Public or private.

Option for embargo (yes/no, duration):
Yes, duration stipulated by service-level agreement.

Size of holdings (number and size of datasets, mean and median granules (files) per dataset):
17,000 data objects, 1 GB.

Please describe recent usage statistics, if known, including information on annual data product downloads, annual number of users, annual number of data products used in publications:
Too early to be meaningful.

User interactions

How does a user contribute data? (what can be deposited, how are data prepared, are specific software required, documentation/support available)
The Merritt data model permits deposit of any dataset that can be represented as one or more versions of a file hierarchy. A serialized form of this representation is submitted for deposit, at which time it is assigned a persistent identifier (unless it already has one). Minimal metadata is strongly recommended, but not required. Deposit can occur via the Merritt user interface or a RESTful API that
can be operated with standard command line tools (e.g., “curl”). These interfaces are fully documented at the Merritt URL.

**How does a user acquire / access data?**
Any data object, or versions, or files thereof, can be retrieved by users at any time, subject to curatorially-defined access control rules.

**What user support services are available (both for depositing and accessing/using data)?**
Generic tools are provided to deposit and access the data. Metadata is indexed via both an inventory and (next six months) search service.

**How does the resource curate data at the time of deposit?**
Upon deposit, a dataset (“object”) is processed to extract metadata, compute file digests, generate an OAI-ORE resource map (to represent structural relationships), and an object manifest. Objects are replicated, as well as scanned to generate characterization information (e.g., format validation and technical metadata).

**Technical characteristics and policies**

**Software platform description, incl. data search and access API(s):**
UC3-authored, community-supported, portable open source software based on the CDL micro-services architecture. APIs and command-line interfaces provide a superset of all functionality available in the user interfaces. All APIs are RESTful.

**Service reliability (including recent uptime statistics, frequency of hardware refresh, if known):**
24 x 365, 1-4 scheduled downtimes per year.

**Preservation reliability (including replication/backup, integrity checks, format migration, disaster planning):**
RAID 6, tape backup of all system components, one extra replica of all content, fixity service running every 3 days, high-availability server clusters, standard PCI data center protocols for physical and electronic security.

**User authentication technology (incl. level of create/modify/delete access by users):**
HTTP Basic over SSL, LDAP-based account management, create/modify/delete operations only available to depositor.

**Data identifier system and data citation policy, if available:**
ARK and DOI using EZID service (http://n2t.net/ezid).

**Metadata standards (including provenance):**
DataCite, Dublin Core, Dublin Kernel.

**Capacity/services to DataONE**

At what functional tier will you initially be operating? (see http://bit.ly/MNFactSheet for definitions)
- [ ] Tier 1: Read only, public content
- [x] Tier 2: Read only with access control
- [ ] Tier 3: Read/write using client tools
- [ ] Tier 4: Able to operate as a replication target
If you can host data from other member nodes, what storage capacity is available?  
n/a

Can you provide computing capacity to the broader network? If so, please describe.  
In principle, yes, but would have to be subject to review for alignment with CDL priorities and UC computer usage policies.

Other Services

What other services or resources (such as expertise, software development capacity, educational/training resources, or software tools) can be provided of benefit to the broader network?  
We plan to develop and release a pluggable open-source Merritt software stack adapter that will permit any instance of a Merritt repository to comply with DataONE software protocols.