

Research Data Curation Pilots: Lessons Learned

David Minor
The UC San Diego Library

Matt Critchlow
The UC San Diego Library

Arwen Hutt
The UC San Diego Library

Declan Fleming
The UC San Diego Library

Mary Linn Bergstrom
The UC San Diego Library

Don Sutton
The San Diego Supercomputer Centre

Abstract

In the spring of 2011, the UC San Diego Research Cyberinfrastructure (RCI) Implementation Team invited researchers and research teams to participate in a research curation and data management pilot program. This invitation took the form of a campus-wide solicitation. More than two dozen applications were received and, after due deliberation, the RCI Oversight Committee selected five curation-intensive projects. These projects were chosen based on a number of criteria, including how they represented campus research, varieties of topics, researcher engagement, and the various services required. The pilot process began in September 2011, and will be completed in early 2014. Extensive lessons learned from the pilots are being compiled and are being used in the on-going design and implementation of the permanent Research Data Curation Program in the UC San Diego Library.

In this paper, we present specific implementation details of these various services, as well as lessons learned. The program focused on many aspects of contemporary scholarship, including data creation and storage, description and metadata creation, citation and publication, and long term preservation and access. Based on the lessons learned in our processes, the Research Data Curation Program will provide a suite of services from which campus users can pick and choose, as necessary. The program will provide support for the data management requirements from national funding agencies.

Received 13 January 2014 | *Accepted* 26 February 2014

Correspondence should be addressed to David Minor, University of California, San Diego, 9500 Gilman Drive, #0699, La Jolla, CA, 92093-0699. Email: dminor@ucsd.edu

An earlier version of this paper was presented at the 9th International Digital Curation Conference.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: <http://www.ijdc.net/>

Copyright rests with the authors. This work is released under a Creative Commons Attribution (UK) Licence, version 2.0. For details please see <http://creativecommons.org/licenses/by/2.0/uk/>



Introduction

In 2008, a campus-wide survey designed to elicit feedback on topics related to contemporary scholarship and its use of cyberinfrastructure was conducted at the University of California, San Diego (UCSD). Among the many responses, one of the most common was the need for core facilities providing a range of data services. The most requested were management of active research data, long term preservation of research data, and tools for adequately describing research data.

The results of the survey were used as the basis of a publicly available blueprint (UCSD Research Cyberinfrastructure Design Team, 2009). This document lays out explicit direction, with concrete plans, for the infrastructure needed to support research, today and into the future. The blueprint designates six core functional areas: a high speed research network, centralized campus storage, campus co-location facilities, shared computing facilities, and digital curation. Based on the direction provided by the blueprint, a shared campus initiative was started, known as the Research Cyberinfrastructure Initiative, or RCI.¹ This group was tasked with operationalizing the services outlined, and providing them to campus in an efficient, cost-effective manner.

The ‘digital curation’ section of the blueprint has been realized in two ways:

1. The provisioning of a multi-year curation pilot program that is determining the needs of researchers on a spectrum of possible services;
2. The creation of a permanent Research Data Curation Program in the UC San Diego Library.

These two instantiations are working closely to create a viable data curation service for campus. In this paper, we will be examining the initial steps that have been taken under these efforts.

Data Curation Pilots

The pilot process started with a campus solicitation seeking applicants who would be active participants in the process. A range of possible curation services were proposed, with the understanding that not every researcher needed every service, and that some services would be created and tested during the pilot phase. The key service offerings are:

- Ingest of datasets and digital objects into the Library’s Digital Asset Management Systems (DAMS) for long term access, management, and discovery;
- Assistance with the creation of metadata to make data discoverable and available for future reuse;
- Ingest of data into the San Diego Supercomputer Center’s (SDSC) storage system, via high speed networks;
- Ingest of data into Chronopolis², a geographically-distributed preservation system;

¹ Research Cyberinfrastructure Initiative (RCI): <http://rci.ucsd.edu>

² Chronopolis: <http://chronopolis.sdsc.edu>

- Training classes on data management planning and the DMPTool³;
- Data object identifier services.

Five research groups were chosen as pilot participants. They were chosen based on several criteria, including the importance of their research collections and the degree to which they represent core research aims on campus. The five groups are:

- The Brain Observatory,
- The National Science Foundation OpenTopography Facility,
- The UCSD Levantine Archaeology Laboratory,
- Scripps Institution of Oceanography Geological Collections,
- The Laboratory for Computational Astrophysics.

Due to length restrictions, this paper will not be addressing the specific content of these research collections. In-depth analysis of this type can be found on the RCI website.⁴

UC San Diego Library's Digital Asset Management System (DAMS)

The UC San Diego Library's Digital Asset Management System (DAMS)⁵ is a technology framework for preserving, managing, and making digital objects and their metadata available, both within the university's current library systems and to external groups and applications. The DAMS is designed to accommodate the widest possible spectrum of digital media held by UC San Diego libraries, housing and delivering digital resources in ways that serve both the preservation of the data and its users in flexible, open-ended ways.

Adapting the DAMS for the RCI pilot process was both an exciting and challenging experience. It pushed the system toward new edges, challenged existing assumptions, and produced a stronger platform for research data going forward beyond the pilot process. Along the way, there were numerous lessons learned that we believe are broad in application and context.

The size of research datasets

Perhaps not surprisingly, the size of the research datasets we received in the pilot process, with few exceptions, were large. For example, the OpenTopography LiDAR datasets, stored as individual zipped .tar files, ranged in size from five gigabytes to over a terabyte. In addition, the San Diego Supercomputer Center (SDSC) was to be the host of these datasets on an OpenStack object storage instance. Library digital collections, by comparison, are stored on an EMC Isilon and managed using NFS and CIFS mounts. Our ingest process required uploading these very large files to the OpenStack instance, so we ran a series of tests to determine our best option (UCSD Library, 2012, November 14). In short, we identified that the OpenStack REST API was more performant than the Rackspace Cloudfiles API. Because of how the API worked, we also needed to segment files larger than five gigabytes in size. Finally, we had to pad the segments themselves with leading zeros, as the segments would otherwise be returned in an incorrect, lexical,

³ DMPTool: <https://dmp.cdlib.org>

⁴ RCI pilot project descriptions: <http://rci.ucsd.edu/data-curation/pilots.html>

⁵ DAMS: <https://github.com/ucsdlib/dams/wiki/DAMS-Manual>

ordering. Performance testing notes are also available in the UCSD Library blog post (2012, November 14), for those interested.

Complex object display

Moving past the data and toward the logical structure of the digital objects themselves, we encountered new challenges. The DAMS has always supported complex digital objects with nested structures. However, the required representation of the research datasets introduced complex objects of greater depth and breadth. Displaying objects of this complexity to end users presented numerous user experience challenges. How would a user want to logically navigate through an object with seven levels of depth, and a number of object components at each level of depth? How should a component of an object be displayed while still retaining the structural and metadata context of the parent? After a number of iterations, we came to a design that we have moved ahead with, but still feel has room for improvement. The design incorporates some design elements from traditional file browsing, with metadata displayed in a similar fashion to YouTube.

Data model changes and formalization

In addition to the design challenges of supporting complex dataset representation, we also needed to support and implement a new data model for the DAMS (UCSD Library, 2013). The new data model represents objects as linked data entities in an RDF triplestore, with very little metadata persisting in the object record itself (UCSD Library, 2012, November 2). While this allows for a very flexible database layer that provides a powerful, integrated corpus of data, we encountered difficulties implementing this linked data model in the Hydra⁶ framework. The level of effort to properly implement this was difficult to estimate, and it indeed took much longer than we had anticipated. However, after a successful consultation effort with digital curation experts, the Hydra framework has been supplemented to support the relationships and nested structure needed for our data model. This resulted in not only a successful implementation for UC San Diego, but the fundamental functionality supporting our implementation now exists in the Hydra framework itself and is already being utilized by other Hydra partner institutions.

Branding and complex collection display

A final lesson learned within the context of the DAMS in the RCI project has been branding and the hierarchical categorization of the datasets themselves. As the DAMS has traditionally only housed digital collections owned by the Library, the addition of research datasets raised questions of categorization and ownership. We want a user to be able to filter only research datasets, should they choose, and to enable broad searches across the entire corpus of data. Creating this categorization is difficult from a branding perspective and, to a lesser extent, a technical perspective. In the new data model, we introduced an Administrative Unit class, to which an object could belong. While this affords us the technical solution to this issue, exposing this distinction from a branding and user experience standpoint proved more problematic. We introduced an immediate filter on the main application landing page, allowing users to browse either research or library collections. From there, searching and faceting could be done within that scoped context.

⁶ Hydra: <http://projecthydra.org>

Metadata Processes

The metadata work we did in the pilots was largely based on a consultation model. This was a natural choice based on the investigative nature of our goals for the project. Our initial meetings with researchers included the entire RCI team, as well as all of the researchers' teams. These large meetings were a chance for us to provide the researchers with a brief introduction to the RCI team, as well as our goals and expectations for the pilot process. But more importantly, they were a chance for the researchers to tell us all about the work they are doing, talk about their data needs and their hopes for how RCI might meet some of these. As part of these initial meetings we used the Purdue Data Curation Profiles Toolkit⁷ to help structure our data gathering.

We continued discussions in subsequent meetings with smaller subsets of the whole group, allowing us to focus on specific aspects and issues of each pilot. Many of these on-going discussions were organized by one of two metadata analysts, each taking the lead on specific projects and working closely with those researchers. Other RCI team members participated as needed to address specific needs, such as data transfer, DOI assignment, etc. There was also a great deal of one-on-one work with specific members of the researchers' teams to hammer out specific details of data modelling, metadata workflow, and so on. Overall, the process was very iterative, collaborative and customized.

Services

The actual services we were able to provide during the pilots covered a fairly wide range, depending on the researchers and their needs. The biggest and most important part of our work focused on determining 'what is an object' in the context of each research project. This involved looking at the different aspects of their knowledge universe, their data, and their expectations for their data. Some of the questions we asked in trying to determine this included:

- What actually constitutes a discreet set of data? Where are the boundaries? What is required – be that files or metadata – for the data to be understandable, usable and reusable now?
- What's necessary for the data to be usable in the future? What's important for long term preservation and functionality of the data?
- What should be displayed and shared? What parts of the data are important for displaying in a repository and making available to users?
- What is important to be able to reference? Where should digital object identifiers be assigned to allow the appropriate parts of the data to be cited and thus give the original researchers full acknowledgement of their work.

This process of exploring the 'objectness' of the research materials involved in depth needs assessment, both for the research data as well as the researchers' expectations of the data curation process, and was incredibly enlightening.

Another aspect of our work included providing best practices and assistance with data organization. This included issues like:

- Collocation of files, data and metadata that are stored in different places, or if there was a need for distributed storage, making sure it's organized in a logical and intentional way;

⁷ Purdue Data Curation Profiles Toolkit: <http://datacurationprofiles.org>

- Development of identifiers or naming protocols for metadata and files which are unique within their context, and when necessary providing a simple mechanism for linking files and metadata together, or for showing relationships between files;
- General data clean up, addressing duplicate copies of data, versioning, etc.

Finally, we worked with the metadata itself, focusing on functionalities we wanted to enable. We looked at how existing metadata could facilitate discovery of the data, and what modifications or additions would enhance this. Once an item is discovered it is also important that it is understandable to a user. This is a much bigger challenge for data which isn't self-describing, and thus the more abstract the data is, the more important it is to have good supporting metadata. Also, when examining whether the data can be used, are there supporting programs or scripts necessary for rendering, processing or otherwise using the data? And as mentioned before, is there metadata to facilitate citation and citation tracking?

These functions are important for a host of reasons but they are especially important in the context of new funder mandates and expectations for how research data is handled.

The work we did to address these needs included the creation of metadata to facilitate these areas of functionality, identifying and using appropriate controlled vocabularies and value lists, formatting of metadata, and analysing workflows for metadata creation.

We learned a number of lessons from this work with the researchers:

- The researchers are the subject experts and we will need to continue to rely on them for that subject expertise. Our community of researchers is too broad, and library funding too overextended, to make it feasible for us as librarians to develop subject-specific metadata expertise for all the major programs in the university. What we can offer is more general data and metadata management expertise and advice.
- There are many similarities between different research data domains, and even between research data and our more traditional cultural heritage library materials. They often rely on many of the same major types of metadata for understanding and discoverability (e.g. topical subjects, geographic coordinate data, associated names with roles, etc.). Even though the contents are very different between, for example, the slices of a brain and the pages in a book, many of the core structural relationships are fundamentally the same (whole/part, parent/child). There are also often similarities in desired functionality – such as citation, faceting on specific data points, and coordinate based navigation.
- There is no one-size-fits-all solution. Despite the similarities, there is no single set of services that will meet all researchers needs. Everyone needs something a little different. While one group may be primarily interested in an enhanced display and delivery of their data, another was focused on preservation and storage, leaving display to specialized external systems. As a result, a high degree of specialization/customization for a specific collection or data set seemed unlikely to be of broad utility.

- Finally, consultation services can easily become overly time intensive. While in-depth consultations offer a wealth of information, they require a significant amount of the consultant's and researcher's time. Moving into production level services, we don't anticipate such in-depth consultations being scalable for us or for the researchers. Future production level consultation services will be more targeted, focusing on specific, established goals that can be achieved without overextending the researchers' or consultants' resources.

Data Ingest into the San Diego Supercomputer Center's (SDSC) Storage System

Technical staff from SDSC assisted in the movement of large amounts of complex data from a wide variety of locations on campus. This data movement also included the creation of BagIt⁸ manifests and checksums for data to verify the data at all stages. Time was also spent working with data owners on the proper ways to checksum digital objects.

A key objective in the data transfer process was to provide an easy and familiar transfer method from the data submitters perspective. With this in mind, several options were established:

- A cloud storage option provided via SDSC's newly developed OpenStack cloud storage system⁹. This system has an online drag and drop functionality to simplify data transfers. From the website it is also quite easy for both the data submitter and RCI staff to view and monitor uploaded data. Accounts were set up for those RCI projects wishing to use this method. Command line tools are available to bulk upload data to the SDSC cloud. Data objects over 2GB require the use of these command line tools, whereby the data object is automatically segmented into 2GB sections along with the generation of a manifest for segment management and reassembly. As this was a new cloud system some lessons were learned. In particular, we initially had problems working with filenames and directories that used special characters, and in working with data from Windows systems where backslashes instead of forward slashes are used for directory designation. In most cases software changes were made to handle these problems.
- Traditional POSIX disk space was also provided as an option. This storage option is part of a large storage system at SDSC running under UNIX/Linux OS. Many researchers, especially from the physical sciences where Linux operating systems are commonly used, were more comfortable with this option. A campus-wide account management system simplified access control to the disk space. For Windows users, a Samba interface was available to interoperate with the project storage disk space. When access was established, data submitters could transfer data much like they would within their own computer labs.
- In some instances, for smaller collections, simple email was used. In general we learned providing several transfer options paid off in terms of efficiency, particularly as if problems arose another option could be used.

A requirement for ingest into the RCI curation program is that the data submitter must provide a manifest listing all objects they are submitting, along with a checksum

⁸ BagIt: <http://en.wikipedia.org/wiki/BagIt>

⁹ SDSC Cloud Storage: <https://cloud.sdsc.edu>

for each data object. We recommended that the data submitter use the BagIt package format and at a minimum required a BagIt formatted manifest to be included during the upload process. We found generating this manifest to be problematic for some projects so we developed a BagIt manifest generating tool, which we made available online. This tool simplified this requirement.

The data transfers were often an iterative process as curation and project staff negotiated what objects and metadata to submit as part of their project. During the ingest process the inventory of objects that were part of a project would fluctuate over sometimes month-long periods. Once this inventory was finalized, RCI staff would use the provided manifest to verify that all objects were accounted for and validate their checksums. As a final step in the ingest process, this validated copy of the submitted project would be uploaded into Chronopolis for long term preservation.

Ingest of Data into Chronopolis

Chronopolis has been certified as a “trustworthy digital repository” that meets accepted best practices in the management of digital repositories by the Center for Research Libraries (CRL).¹⁰

The Chronopolis preservation network provides long term preservation services for valuable digital collections. It is an integral part of the RCI curation process in that the initial submitted collection as well as the final curated collection are ingested into Chronopolis. Many of the holdings are objects in the terabyte range, packaged using the BagIt format. The system currently has three nodes distributed across the county at SDSC (primary node), the National Center for Atmospheric Research (NCAR) and the University of Maryland’s Institute for Advanced Computer Studies (UMIACS). All nodes have complete replicates of all Chronopolis holdings. At each node the integrity of the data are continuously checked against authoritative manifests using ACE¹¹, an automated monitoring system. ACE uses rigorous cryptographic techniques and tokens stored in a centralized integrity management service to insure data sustainability. Chronopolis is also a first and replicating node within the Data Preservation Network (DPN)¹². DPN provides another option for long term preservation.

The movement of data into Chronopolis includes the transfer process from the data submitter, validation of this transfer against the provided manifest, and finally registration into Chronopolis and replication to its nodes. These steps occur at a collection level: the data submitter puts together a collection of digital objects and provides a manifest listing the objects and their checksums. This collection is managed as a single entity as it moves through the ingest process. A staging area is set up at SDSC for transfer from the submitter. Once complete, this transfer is validated and, if successful, the collection is moved into and registered into Chronopolis where a second validation occurs all using the provided authoritative manifest. As a final step of the ingest process, the collection is replicated across the Chronopolis data grid where validations at each individual node are also performed.

Chronopolis has a wide range of data, including the library digital holdings from many organizations, web crawls from political campaigns, atmospheric data, astrophysics simulations, oceanography samples and shipboard generated data, archaeological artifacts, LIDAR data and digitized brain scans. The collections are comprised of a wide array of data types, formats and organization principles.

¹⁰ See: <http://chronopolis.sdsc.edu/trac/index.html>

¹¹ ACE: <https://wiki.umiacs.umd.edu/adapt/index.php/Ace:Main>

¹² Data Preservation Network (DPN): <http://www.dpn.org>

In developing these data networks, many lessons have been learned. In particular, valuable knowledge has been gained in developing techniques for data transfers across heterogeneous systems and the necessary message management for systems like this to work in an automated fashion. A cornerstone of Chronopolis practice has been the use of an authoritative manifest generated by the data provider and used throughout the Chronopolis process to manage and validate data holdings. We have also found the BagIt packaging format to provide a simple and effective way to transfer and manage holdings at the collection level.

Training Classes on Data Management Planning and the DMPTool

In order to inform and educate campus researchers about basic research data management principles and best practices, librarians in the RCI curation program designed and delivered hour-long workshops that were presented regularly over the course of the academic year 2011/12 and 2012/13. Almost 200 people attended the workshops.

Because of the wide publicity about the National Science Foundation's (NSF) Data Management Plan (DMP) requirement in early 2011, and the associated flurry of concern and interest from a number of UC San Diego campus entities and research disciplines, the workshop was initially focused on providing information about the NSF DMP requirement.

The workshop content was structured around the stages of the Research Data Life Cycle model, and connections were established between the phases of the research data life cycle and the DMP sections specified in the NSF DMP General Guidelines:

- Describe data,
- Describe metadata,
- Policies for access and sharing, provisions for protection and privacy,
- Policies for reuse and distribution,
- Plans for archiving and preservation.

As more tools and resources were incorporated into the suite of research data curation services we offered, the workshop sessions provided an opportunity for discussion and hands-on exploration. For example, cards promoting the EZID digital object identifier service were handed out and, once the DMPTool was available, attendees were sent advance instructions on how to log in to the DMPTool so that they could access it readily during the workshop.

The workshops were one hour long. The schedule was promoted on the Library website, the RCI website, and was distributed to individual departments by liaison librarians. In addition to the librarian who taught the workshop, other RCI staff frequently attended the workshop. The RCI staff that were able to participate fielded specialized questions about metadata and storage, and got direct exposure to researchers' questions and concerns about research data management.

Over time the content and emphasis of the workshop shifted from a narrow focus on the NSF DMP requirement to a more comprehensive view of research data management planning. In its most recent iteration there were two broad workshop objectives:

1. Familiarity with the basic concepts of research data management and planning,
2. Awareness of the services and resources available to support this work.

A central tool presented in all of the workshops was the DMPTool. Modelled on DMPOnline¹³, a web-based tool produced by the Digital Curation Centre (DCC) to assist researchers in the United Kingdom in creating data management plans, the DMPTool provides context-specific guidance to researchers as they generate DMPs. The UC San Diego Library was among the founding members of the partnership that created this flexible online tool. The effort was in response to demands from funding agencies, such as NSF and the National Institutes of Health (NIH), that researchers plan for managing their research data. By joining forces the contributing institutions are able to consolidate expertise and reduce costs in addressing data management needs. The DMPTool guides users through the requirements of various funding entities and includes guidance to specific NSF Directorates and Divisions. UC San Diego researchers see information specific to this institution, such as language suggested by the Office of Contracts and Grants to be incorporated into UC San Diego DMPs, and examples of DMPs completed by other UC San Diego researchers. UC San Diego is among the heaviest users of the DMPTool.

There were two main lessons learned in our experience teaching workshops. First, the surge of interest in learning about the NSF Research Data Management Plan requirements subsided by the end of the 2011/2012 academic year and workshop attendance declined during 2012/2013. As researchers and labs become accustomed to meeting the DMP requirement, they no longer had the sense of urgency that filled the original workshops. However, other funding entities could become more explicit in data management planning requirements. There have also been anecdotal reports of NSF DMPs receiving specific critical comments during the review process, and it is possible that individual NSF Directorates, Divisions or the NSF as a whole will revisit and refine their requirements. Changes in the DMP requirements could revive interest in an interactive online course or an in person workshop. At this time there is growing attention on data citation and tracking the impact of data, so the topics of DOIs, metadata, and data citation are of more interest than the data management plan, per se.

The second major lesson learned is that in order to meet the needs of specific user communities – such as social sciences, arts and humanities, graduate students, specific departments or Organized Research Units (ORUs) – the workshop content and focus has to be customized. Compiling content in modules, as exemplified by the DataOne and the New England Collaborative Data Management Curriculum initiatives, would allow us to deliver workshops to meet the particular needs of a user community. Teaching in person is rewarding, productive and informative, but developing an online course would allow greater participation and more flexible access to the content.

Digital Object Identifiers: EZID

One service promoted heavily by the RCI Program is EZID¹⁴. Licensed with the University of California Curation Center (UC3), EZID is used to provide digital object identifiers (DOIs) for UC San Diego researchers. Staffing, policies and workflow for the assignment of EZID DOIs, both for data under curation the RCI program and by the general research community across the campus, were developed over several months.

In June 2011 the UC San Diego Library signed the EZID contract with UC3; the Library paid the licensing fee for the campus and continues to do so. A liaison librarian was designated as the EZID representative. One aspect of this Faculty liaison's role is to

¹³ DMPOnline: <https://dmponline.dcc.ac.uk>

¹⁴ EZID: <http://www.cdlib.org/services/uc3/ezid>

promote EZID, to register interested researchers, and to serve as a contact with UC3 EZID.

The process of registering users for EZID is straightforward and the researchers and research groups registered thus far have had prompt service and a positive experience. Over 170 DOIs have been generated by campus researchers. The process of setting up and managing EZID accounts is feasible with current staffing and software. Promotion and marketing of EZID DOIs remains a challenge, as this is a new service and the use of DOIs and the practice of data citation is not yet 'standard operating procedure' in the realm of publishing research results. There is growing support by national and international entities, touting the value of DOIs in ensuring the transparency and reproducibility of scientific results.

Conclusion

The curation pilot process is expected to be completed by mid-2014. In late 2013, the UC San Diego Library's Research Data Curation Program was chartered. This program will build on the work done in the pilot, and use the lessons learned in the pilot process to create a permanent researcher-focused enterprise. Much of the infrastructure components and services initially tested will be made production-worthy and offered to the campus. It is planned that other related offerings will be added as required, based on feedback from users.

References

- University of California, San Diego, Research Cyberinfrastructure Design Team. (2009). *Blueprint for a digital university*. Retrieved from http://rci.ucsd.edu/_files/Blueprint.pdf
- UCSD Library. (2012, November 2). New version of DAMS data model [Web log post]. Retrieved from UCSD Library Digital Asset Management System blog: <https://libraries.ucsd.edu/blogs/dams/new-version-of-dams-data-model/>
- UCSD Library. (2012, November 14). Open Stack Testing Notes [Web log post]. Retrieved from UCSD Library Digital Asset Management System blog: <https://libraries.ucsd.edu/blogs/dams/openstack-testing-notes>
- UCSD Library. (2013, March 5). About our data model [Web log post]. Retrieved from UCSD Library Digital Asset Management System blog: <https://libraries.ucsd.edu/blogs/dams/about-our-data-model>