Business and Service Model for the University of Minnesota Libraries' Data Repository & Curation Service

A report from the University of Minnesota's Data Management and Curation Initiative

July 26, 2014

Project Sponsor:	John Butler, University Librarian for Data & Technology
Project Team:	Lisa Johnston (Project Lead), Jon Nichols (Technology Lead), Josh Bishoff, Steven Braun, Carol Kussmann, Francine Dupont Crocker, Kevin Dyke, Stephen Hearn, Alicia Hofelich-Mohr, Eric Larson, Erik Moore, Arvid Nelsen, Jon Nichols, Carolyn Rauber, Justin Schell, Bill Tantzen, Amy West

Abstract:
The University Libraries launched the Data Repository for the University of Minnesota (DRUM) (see http://z.umn.edu/DRUM) in November 2014. This service was developed and implemented by a group in the libraries called the Data Management and Curation Initiative (DMCI). The work products of the group presented here include the DMCI Business and Service Model that defines the draft policies, rational for data management and curation services, a proposed staffing model for distributed data curation, and initial first year budget. Other supporting documents for the launch of the repository include the Service and Functional Requirements that led the development of DRUM in the DSpace software (V 4.2) and the metadata schema, based on dublin core, for the data repository submission form and public access record.

Suggested Citation: University of Minnesota Libraries. (2015). Business and Service Model for the University of Minnesota Libraries' Data Repository & Curation Service. Retrieved from the University of Minnesota Digital Conservancy, http://hdl.handle.net/11299/171761.

View supporting documentation, a spreadsheet of functional requirements, and Dublin Core metadata schema, at http://hdl.handle.net/11299/171761.

Table of Contents

Introduction

The Data Management and Curation Initiative (DMCI) of the University of Minnesota (UMN) Libraries launched in February 2014. Building on the work of several library committees, user-needs assessments, and a 2013 pilot of data curation services, the Libraries developed this business model for the development and implementation of Data Management and Curation Services ("DMCI Services") targeted to launch in fall 2014 for University of Minnesota researchers.

The objectives of the Libraries' DMCI Services are to offer research data services that are scalable and appropriate for UMN research data, including data management plan consultation, metadata services/consultation, access, discovery and dissemination, and archiving and preservation.

The initial goal of the DMCI Services will be to offer a suite of services that will together provide access (open and by request), discovery, dissemination, and preservation for non-restricted digital research data created by UMN affiliates in order to meet data sharing needs and requirements. The strategies to implement this goal include:

- Infrastructure: Create repository infrastructure to enable data stewards to deposit UMN-affiliated digital research data for access, discovery and dissemination, and archiving and preservation of research data.
- Policies: Develop policies and procedures that will enable appropriate access, discovery and dissemination, and archiving and preservation of research data that incorporate policy-driven decisions for what data is collected and for how long the data is retained.
- Staff: Train new and existing library staff to
    a. provide consultation on data management plans,
    b. provide metadata consultation services,
    c. curate data (e.g. arrange, transform, and present the data for access, discovery, dissemination, and preservation according to established procedures).
- Partnerships: Partner with campus units, peer institutions, and network-level collaborations in order to create scalable solutions for data services for the large, complex, and ever growing corpus of digital research data at the UMN.

1.0 A Rationale for Developing Data Management and Curation Services on Campus

Over the last several years, researchers and administrators at the University of Minnesota have developed a growing awareness of and need for long-term access to digital research data. A recent and significant driver is the February 22, 2013 memorandum by the White House Office of Science and

Technology Policy (OSTP)[1], directing federal agencies to develop plans to ensure publications and research data are accessible to the public. Several federal funding agencies[2] already require investigators to include a plan for how will share research data, and this new requirement asks that resulting data are "publicly accessible to search, retrieve, and analyze."

Data management happens throughout the research data life-cycle and the Libraries are committed to ensuring that data are well-prepared, documented for contemporary use, and available for discovery and reuse through standards-based curation practices. The Libraries' existing role in experience with developing, managing, and preserving collections, along with its expertise on information discovery and access, and its campus-wide perspective and connections make it a strong candidate to take a leading role in coordinating these services. Furthermore, without an entity like the Libraries to provide across-the-academy guidance on use of standards and best practices, local practice will likely devolve into idiosyncratic methods. Taken together, a fragmented, disconnected overall state of data handling would likely ensue. We have yet to learn which services for data curation and management might best be executed at the institutional level and which via network-level collaborations.

As a result of exploring UMN research data management needs over the last several years, the library launched a Data Management and Curation Initiative (DMCI) in February 2014 to address this growing awareness and need for data sharing and to expand and increase our existing efforts. The DMCI had three primary objectives:

1. Develop data curation services that are scalable and appropriate for UMN research data, including, data management plan consultation, metadata services/consultation, access, discovery and dissemination, and archiving and preservation.
2. Better understand data management needs and best practices in order to provide researcher training, graduate student education, and outreach efforts at various stages of the research process.
3. Partner with University units/groups, peer institutions, and network-level collaborations (e.g., SHARE) to make appropriate strategic investments in programs and infrastructure that support deposit, discovery, access and stewardship services needed throughout the data life-cycle. Through working at local and network levels, refine our understanding of what services are best executed at the institutional level and which at the global level.

These objectives, and many of the strategies that the DMCI is employing to address our objectives, were drawn from the current environment and key recommendations outlined in three libraries' working group reports, (1) the 2012 ARL E-science Institute Capstone project report,[3] titled "Agenda

[1] http://www.whitehouse.gov/blog/2013/02/22/expanding-public-access-results-federally-funded-research
[2] https://www.lib.umn.edu/datamanagement/funding
[3] This report was published as "Developing E-science and Research Services and Support at the University of Minnesota Health Sciences Libraries." J Libr Adm. 2012;52(8):754-769. Available at http://www.ncbi.nlm.nih.gov/pubmed/23585706.

for Deepening Library Support for Research," (2) the 2011 unpublished report, "Near-Term Recommendations for Action from the Data Management, Access, and Archiving Working Group" (a subgroup of the Libraries' Research Support Services Collaborative), and (3) the 2009 unpublished report, "Data Stewardship Opportunities for the University of Minnesota Libraries' Recommendations from the Libraries' E-Science Data Services Collaborative (EDSC)."

1.1 Alignment with University Mission and Priorities

The DMCI Services will allow researchers and scholars to archive, share, and preserve the University's digital research data. By making this form of knowledge created at the University accessible to the citizens of the state, the nation, and the world, the DMCI Services directly support the University's mission to Research and Discovery, Teaching and Learning, and Outreach and Public Service. (See the Mission of the UMN)

At the University of Minnesota, the UMN Libraries have been actively engaged in understanding and supporting the evolving needs of researchers for the past several years. Numerous user-needs studies, organizational changes, service development activities, and collaborations signaled the extent and the involvement of the Libraries' role in the rapidly developing data environment. The Libraries have also been preparing the organization, through the creation of new roles and staff development activities for increasing support for research and data services. The Libraries' forging of intra- and inter-institutional partnerships, engagement in grants, and in the development of web environments and service programs in support of research has been abundant over the past six years.

Based on the extensive background and user-needs assessments that the libraries have led, members of the DMCI participated in a visioning process that allowed us to capture an agenda for developing infrastructure, services, and staff in the area of data management and curation.

1.2 Gap Analysis of Data Management Services and Policies

The Libraries have been formally and informally gathering researcher needs through a number of avenues. Primarily, our extensive workshop and training programs in data management[4] that have reached over 400 researchers and over 100 graduate students since 2011 have have revealed a need for developing services for research data management and data sharing in support of increased expectations from federal funding agencies. Our gap analysis (see Environmental Scan of Existing Policies and Services) of UMN services and policies plotted across the research data life-cycle found a surprising number of services for the active management and analysis of research data, but few comprehensive services that address the beginning and closing stages of the research life-cycle.

1. **Project Planning and Data Collection Gaps**: Researchers do not get adequate support for

[4] See University Library training and workshop offerings at https://www.lib.umn.edu/datamanagement/workshops

planning how they will plan create and manage their data during a project. The DMCI Services will address this gap by providing consultation and training for data management, particularly in the areas of data management plans and metadata creation.

2. **Active Research Data Management, Storage, and Analysis Gaps**: Covered by a wide range of services (e.g. OIT, MSI) that may, however, be difficult for users to navigate. Service catalogs and referral networks are recommended.
3. **Data Dissemination, Archiving, and Preservation Gaps**: With increasing demands on researchers to share their data from federal funders, researchers require data repository platforms that provide dissemination and access. This can happen at the disciplinary level (E.g. GenBank, Dryad) but many research disciplines do not have such options. Similarly, when a research project comes to a close, researchers are looking for support for storing, archiving, and preserving research data post-project. There are unmet needs here for the long-term archival storage and preservation of data.

The need for sharing data broadly is highlighted by the Office of Science Technology Policy (OSTP) Feb 2013 memo[5] requiring federal funding agencies with $100M/Year funding (includes the NSF, NIH) to develop a strategy that "improves the public's ability to locate and access digital data resulting from federally funded scientific research.... [the data] should be stored and publicly accessible to search, retrieve, and analyze." Specifically, the memo includes the several recommended objectives that the federal agencies implement (see Table).

Table: OSTP Recommendations for Federal Funding Agency Action Plans

Objectives for Public Access to Scientific Data in Digital Formats	**Strategies and Considerations (see more[6])**
Maximize open and free access to data created with Federal funds (when not in violation with confidentiality, privacy, proprietary and intellectual property rights).	● Ensure that the data are richly described with machine-actionable metadata. ● Ensure that data are complete, self-explanatory, and accurate (quality). ● Protecting confidentiality and privacy when making data available (remove identifiers, virtual data enclaves)
Ensure that all grants include data management plans for review.	● Follow the example set by the National Science Foundation in Jan 2011.

[5] Increasing Access to the Results of Federally Funded Scientific Research, http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

[6] ICPSR Guidelines for OSTP Data Access Plan, http://www.icpsr.umich.edu/icpsrweb/content/datamanagement/ostp.html

Allow the inclusion of appropriate costs in proposals.	● Account for the long-term access and preservation needs that go beyond the life of a grant.
Include mechanisms to ensure compliance with DMPs.	● Include evaluation of how the researcher complied with DMP.
Promote the deposit of data in publicly accessible databases.	● Identify and/or create trusted digital repositories to steward data over time.
Preserve intellectual property rights and other commercial interests.	● Embargos or delayed dissemination ● Creative commons licenses may be used
Encourage cooperation with the private sector.	● How to ensure that both the public and the organizations benefit?
Develop approaches for appropriate attribution to data.	● Proper citations for data to give credit to data producers.
Support training, education, and workforce development related to scientific data management, analysis, storage, preservation, and stewardship.	● How to grow the workforce to care for the data going forward?
Provide for the assessment of long-term needs for the preservation of scientific data.	● Ensure that formats are in standard, non-proprietary forms. ● Active ongoing management of digital files. ● Recognize that not all data should be available indefinitely by establishing selection or appraisal guidelines. ● How to fund data repositories for the long-term?

1.3 Results of the Libraries 2013 Data Management & Curation Pilot

Data curation involves the expertise of trained staff to select, ingest, describe, arrange, and prepare data for discovery, access, and reuse in ways that will preserve and protect the data over time. The

University Libraries ran a pilot of data curation services from May - December 2013[7].

The pilot's primary task was to develop and implement data curation workflows for 3-5 examples of research data. A call for proposals to participate in the pilot went out in the summer of 2013 and was open to researchers at the university whose data met a variety of criteria (including openness to the public). In response to the call, 16 proposals were received, and five were selected to represent a variety of disciplines and data types. Next, a detailed data curation workflow process was developed for each of the pilot datasets and formed the bases of the overall curation process. To accomplish this, the pilot involved the expertise of archival, digital preservation, and metadata and cataloging staff in the library, as well as data experts from the university, to curate the digital research data while utilizing existing tools, such as the institutional repository. The data were then successfully curated for discovery and reuse in the University's institutional repository, the University Digital Conservancy, at the persistent URL, http://purl.umn.edu/160292. To supplement this process, pre- and post-curation interviews took place with the participating data authors in order to determine the extent of their perceived need for data curation services and the resulting success or shortcomings of the final curated product.

The final report summarizes the steps taken to curate the datasets in the pilot. These steps will form the basis of the DMCI Service procedures for curating data ingested into the library. However, this pilot also highlighted several shortcomings to the library's pilot service. For example:

- Through the interview process, it became evident that several faculty were less concerned with archiving their data for others to access, or even meeting federal mandates, than with finding a permanent home for their data to "live on" with restricted access.
- As future services are developed, it will be important to consider the variety of software, and expertise to use the software, required for data curation. Important software for this study included statistical tools (SPSS, R) and GIS software (ArcGIS).
- The researchers interviewed did not have ready documentation to provide with their datasets. In several cases, readme.txt files were written by curation staff to supplement the data.
- All of the datasets included some aspect of ownership and intellectual property considerations, even though the pilot was explicit that all submissions were dataset ready for public consumption and reuse.
- Due to variables of scale, domain-specific data requirements, and diversity of domain culture and practices, the success of such a service will likely depend upon strong collaboration among interdependent service providers on campus.
- To be successful, significant capacities in areas of data management and curation, infrastructure, and domain knowledge must coalesce in operationally effective ways that minimize barriers and demands on researchers.

[7] A summary report of the University Libraries' 2013 Data Curation Pilot is available at http://hdl.handle.net/11299/162338

1.4 Ongoing and Future Needs Assessments

The University Libraries are currently gathering information on research data practices and issues of UMN researchers in order to continue to refine our understanding of user need. We are doing this in two ways:

1. Faculty Survey: The Libraries are partnering with the College of Liberal Arts' Data Management Specialist to survey faculty in several colleges, including the Academic Health Center (run in June 2014) and the College of Science and Engineering (forthcoming fall 2014). This survey was adapted by a successful survey designed and delivered by the CLA-OIT to CLA faculty in the Fall of 2013[8].
2. Review of Data Management Plans: The Libraries partnered with the Sponsored Projects Administration and Principal Investigators (PIs) to obtain and review the two-page data management plans for successful grants accepted to the National Science Foundation, required with all grant proposals since January 2011. Analysis of these plans (nearly 200 in total) should reveal current practices and highlight areas of need (results expected fall 2014).

2.0 Proposed Service Model

The University Libraries' DMCI Services, projected to be available to UMN researchers in the fall of 2014, will:

- enable data authors to deposit data into a library-hosted data repository, subject to library collection policies (outlined below);
- include curatorial actions for all data deposited to the library in order to best arrange, transform, and present the data for access, discovery, dissemination, reuse and preservation, according to established procedures;
- adhere to policy-driven decisions for how long the data must be retained and preserved by the library; and
- allow data authors to include the costs of the Libraries' DMCI Services into their grant funded project proposals.

The proposed DMCI Services can be organized into a suite of services that address key components of the data life-cycle; these span the planning stage (e.g., preparing a data management plan for grants), active data management, and post-project activities such as sharing and preserving data beyond the life of the project. In addition, the Libraries may offer our full suite of DMCI Services for researchers and PIs to incorporate into their grant-funded projects (appropriate pricing models will need to be determined as more experience in this area occurs).

[8] See the results of the College of Liberal Arts survey of Faculty data management needs at https://drive.google.com/file/d/0B1tV96Ef2Ic_S3dpQUZZTzllb3M/edit?usp=sharing.

Table: Suite of Data Management and Curation Services[9]

Plan	**Support for Writing Data Management Plans (DMP)** *No-cost Consultation*	Provide one-on-one assistance, templates, and consultation when writing DMPs, now required by some grant funding agencies (e.g. NSF and NIH). Benefits include: ● DMPTool online (x500 log-in) to help create a DMP. ● Timely DMP review and feedback to meet grant deadlines.
Manage	**Metadata Consultation and Data Management Training** *No-cost Consultation and training*	Data management consultants will help with topics such as: ● Metadata consultation. Larger projects may want to include library staff as cost-share on grant-funded projects (limited availability). ● Providing on-demand training and education modules. ● Policy advocacy and advice on complying with federal mandates for data sharing. ● Referral to campus services, such as storage, backup, analysis and tools for preparing sensitive data for sharing (e.g., de-identification services).

Share	**Repository and Curation for Your Data**
	
	Up to 100GB per project at no cost
		UMN researchers may deposit their data to the Libraries' data repository, subject to our collection policies. Data sets submitted to the data repository are reviewed by data curation staff to ensure that data is in a format and structure that best facilitates long-term access, discovery, and reuse. Benefits include:
		● Flexible access options:
		  ● **Open Data**: Publish data in our open access repository for sharing and reuse. Open Data are publicly available for access.
		  ● **Available via Request**: Control who has access to data (up to 2 years) before publishing. Requested data are accessible pending author approval.
		● Persistent DOIs for data citations.
		● Tracking of data downloads.
		● Bit-level preservation and archival storage for at least 10 years.

[9] See Recommendations and Justification of Service Limitations and Costs in Table below

Grant Partner	**Full Life-cycle Support for Data**	Offer a full suite of DMCI Services for PIs that seek support for data management, data sharing, and long-term preservation for data generated by funded grants. Costs determined per grant-funded project.

2.1 Definitions

The following terms are used throughout this document to describe various components of the DMCI Services proposed.

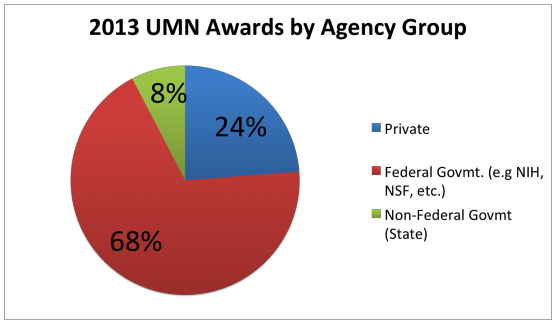
- **Authors:** Data stewards/creators that are the primary consumers/beneficiaries of the DMCI Services. UMN affiliates only.
- **Data:** The individual or package of electronic files associated with a data set. These will include raw, processed, and analyzed data and associated documentation and/or discipline specific-metadata files.
- **Data Curation:** A suite of management activities performed to maintain research data long-term such that it is available for reuse and preservation. Human activities may include:
    - Verifying that all the necessary components of the data are included in deposit.
    - Determining that the library is the appropriate repository for the type of data (vs. another disciplinary repository) and communicating this information to the author.
    - Evaluating the data documentation/description for completeness and suitability for reuse.
    - Verifying that data submission meets our collection policies (e.g., Check that data has no private or restricted information).
    - Identifying any corrupt or unusable files.
    - Asking for additional documentation from the author if needed.
- **Data Curation Staff:** Staff in the Libraries (or beyond) that provide and/or support the DMCI Services.
- **Levels of Access (what the user encounters when using the repository)**:
    - Open Data: All data and metadata are publicly available for users to search, discover, and download. This may also be referred to as “published” data.
    - Available by Request: Metadata are available to search and discover. However, the data are not available to the end-user without author authorization. This may also be referred to as “embargoed” data.
- **Metadata:** The information about the data set provided on ingest (by the author and/or library staff) and any metadata machine-generated and captured by the repository system (includes provenance, descriptive, and administrative).
- **Users:** An end-user of the available data and/or metadata. Can be any member of the public.

2.2 Target Audiences for DMCI Services

These services are intended to meet several stakeholder needs on campus, which will benefit in many ways.

- **Researchers and Data Authors**: The DMCI Services will enable PIs, researchers, and students the ability to appropriately manage, share, and preserve their research data through consultations, tools, and support.
- **Principal Investigators (PI)**: The DMCI Services will allow PIs to comply, and hopefully exceed, federal granting requirements, thereby making UMN research grants competitive in the face of constricting federal budgets.
- **UMN Administrators**: The DMCI Services will better enable UMN administrators to monitor how UMN digital data assets are released, archived, and reused after projects are complete.
- **Campus Partner Units (OVPR, OIT, others)**: The University Libraries will monitor the needs and demand for future services around research data and continue to work with our campus partners to best incorporate additional services for data on campus.

In 2013, there were 2,983 full-time faculty on the UMN Twin Cities campus, 570 part-time faculty, 2,090 full-time professionals, 866 part-time professionals, and 4,086 graduate assistants. Researchers in these categories brought in 1,464 federal grants in 2013, totaling $475,167,431 (68% of all UMN Awards, see graph). 80% of those federal dollars came from agencies with established data management requirements (HHS and NSF).



Researchers with federal grants are increasingly expected to make their data publicly discoverable, reusable and preserved. Plans for how a PI will do this are included with all grant proposals of NSF and likely many more agencies following the 2013 OSTP Memo (see Table above); therefore in order to be competitive, PIs must be able to carry out the sharing requirements for their data. However, the services must also take into account the tension that researchers have with releasing access to their data to broadly and thus should incorporate appropriate access controls (e.g open vs. mediated request).

The DMCI Services' primary audience will be principal investigators at UMN seeking to comply with federal agencies' data management, archiving, and sharing requirements. Secondary audiences will be

data creators at UMN (faculty, professionals, graduate assistants) who want colleagues to find, access, and use their research data; and data creators at UMN who want colleagues to find their research data, but who want to retain control over who can access and reuse them. Audiences external to the UMN who will use these services include researchers who will be able to discover and reuse data stored in the data repository.

2.3 Draft Service Policy

Policies subject to evolve over time as need and demand are better known. The Libraries' "data repository" (name and branding TBD) accepts research data that meet the following collection scope[10]:

- Data must be authored by at least one UMN researcher with an active x500.
- Data must be non-restricted data that DOES NOT contain any private, confidential, or other legally protected information (e.g., personal identifiable information). Link to University Data Security Policy and Definitions of private data.
- Data is contained in digital files EACH not exceeding 2GB per file. Larger files may be mediated through consultation with the library. The total size of a data collection can be up to 100GB per project without incurring additional costs.
- Data must include adequate documentation describing the nature of the data at an appropriate level for reuse and discovery. All data receive curation and data that is unusable may not be accepted in the repository.
- The data should be in a final or published state. For active or changing data, use a UMN storage solution listed here.
- Data should consist of original and unique data that cannot be easily reproduced or acquired elsewhere.

Table: Recommendations for DMCI Service Policy and Costs

DMCI Service	Recommendation	Justification of Service Limitation
Data Management Plan (DMP) Consultation	*Cost: No Cost*	This is a no cost service that the library has been providing since 2011. Consults take about an hour and is a part of the subject liaison job description. Also: ● Freely available tools outside of the library are currently available, such as the DMPTool[11]. ● Peers do not charge for this service.

[10] See Recommendations and Justification of Service Limitations and Costs Table

[11] The UMN libraries customized a version of the DMPTool for campus researchers beginning in 2012. https://dmptool.org/

		• Demand has been low. Since being launched in 2011, the Libraries received around 12 NSF plan consultation requests in the first year. Since, then 1-2 per year. However, we will monitor the demand closely, since other agencies may soon begin to require DMPs (per OSTP memo).
Consultation and training on Metadata and Data Management	*Cost: No cost, subject to limitations*	Training and education has been provided by the Libraries since 2009 and is growing with the number of library liaisons teaching the sessions. Pricing models for more intensive metadata consultation and grant partnerships need to be worked out per grant-funded project. For example, we need to account for grant-funded projects that may want to put a metadata expert on the team to help create a schema for data collected. For example, the NIH data informationist supplement[12]. For example, cost share of 1-100% FTE (varies by project) of a library staff member.
Data Repository and Curation Services for Your Data	*Cost: No cost up to 100GB per project*	The Libraries will continue to offer repository services (e.g., UDC, UMedia) as common good services, making every reasonable effort to avoid ISO, pay-as-you-go services. Therefore, the Data Repository and Curation service should accept research data deposits at no cost to researchers. However, the nature of big data may not allow this model to scale. Therefore, we should establish a maximum amount of files (size in bytes) that the data repository can accept per project (or collection in the UDC) at no charge. 100GB per collection is in line with our peers (most however do not specify a limit): • Stanford[13] has a max of 5GB per record (eg. one 5GB file or five 1GB sized files) and suggests contacting the staff for deposits of "large research data sets (10 GB or

[12] Supplement grant opporutnity issued by the NIH July 19, 2013 http://grants.nih.gov/grants/guide/pa-files/PA-12-158.html

[13] Stanford University benchmarks obtained May 6, 2014 from http://library.stanford.edu/research/data-management-services/share-and-preserve-research-data/data-preservation-stanford

		more)." ● Purdue[14] places limits based on how the space is used. For example, at no cost researchers get up to 100GB for project space and up to 10GB for published data. They charge $2/$14 per GB (respective of use) after that. ● MIT[15] allows for "initial project-based quotas may not exceed 200GB over the life of the project." ● Harvard[16] simply suggests that "If you plan to upload more than 1TB of data please contact us." ● Penn State[17] allows up to 1GB maximum total upload for multiple files. Pricing models per grant-funded project should be offered on a case-by-case basis. (see 4. Grant Partner justification below)
	Upload limit: 2GB per file	There are timeout limits on browsers and the DSpace-based submission form that necessitate the upload limit. 2GB would meet most data file size needs and be inline with our peers: ● Harvard and MIT have upload limits on their repository set at 2GB per file. ● Stanford has upload limits up to 5GB per record. ● Penn State has a 500MB upload limit per file but allows for mediated deposit of greater. ● Illinois[18] and Indiana[19] are set at 500MB and 150MB respectively. University Precedent: Our repository (the UDC) currently set at 1GB per file. This was raised from 250MB in 2011 due to research data submissions from the Minnesota Geological Survey (GIS data sets).
	"Available via Request" Limit:	Researchers will complete projects, change jobs, and/or leave the university over time. Therefore, they will not respond to

[14] Purdue University benchmarks obtained May 6, 2014 from https://purr.purdue.edu/about/pricing
[15] Massachusetts Institute of Technology (MIT) benchmarks obtained May 6, 2014 from http://libraries.mit.edu/guides/subjects/data-management/index.html
[16] Harvard University benchmarks obtained May 6, 2014 from http://isites.harvard.edu/icb/icb.do?keyword=k78759&tabgroupid=icb.tabgroup124911
[17] Penn State University benchmarks obtained May 6, 2014 from https://scholarsphere.psu.edu/
[18] Illinois University benchmarks obtained May 6, 2014 from https://www.ideals.illinois.edu/
[19] Indiana University benchmarks obtained May 6, 2014 from https://scholarworks.iu.edu/data/

	Up to 2 years	requests for access to their data indefinitely. A limit must be set on their ability to limit access to their data. Also: ● Repository software requires a date when objects will automatically be transferred to open access. ● Graduate students may be primary uploaders and are at the U for a min of 2 years. University Precedent: University Electronic Theses and Dissertations are currently embargoed for up to 2 years.
	Preservation Commitment: For a minimum of 10 years following deposit	The idea that the library can archive research data forever or indefinitely is not reasonable. It is important to clearly articulate a minimum time period that the repository will maintain the preservation and access to the deposit to set attainable, realistic expectations of service. University precedent for policy is lacking. Data retention policies[20] at the university only apply to university records (e.g., student grade data), and not research data. Funders have varying expectations[21] with 3 years being most common, but up to 10 years. A University-wide Data Management Policy Committee has been convened by the Vice President for Research (June 2014).
Grant Partner: Full Life-Cycle Support for Data	*Cost: One-time cost recovery fees incorporated into grant proposal budgets, negotiable per project.*	There are no well established cost models to build from that help to identify all of the ongoing costs of data curation. The libraries should tentatively provide these services and partner on grants, cautiously at first, in order to better establish the costs involved. Working with libraries financial office and the university, we may in time be able to present a model for cost recovery for our services that require significant effort above and beyond the benchmarks laid out above. Our peers appear to be in the same situation: ● Stanford has no cost models but states, “We are currently still in the process of working out the fee

[20] At the time of this report, a new policy committee was examining the need to define data policies related to ownership and retention. The existing policy is at http://www.ogc1.umn.edu/stellent/groups/public/documents/webasset/da_031145.pdf.

[21] The blog post How Long Should You Keep Data? (july 22, 2013) at DATA AB INITIO outlines several expectations. http://dataabinitio.com/?p=172

		structure." ● It is clear that Purdue encourages grant partnerships, but they do not indicate a cost structure on their site. ● Rutgers[22] repository service is free, unless there is significant, yet unspecified, effort involved. They state: "This fee can be accommodated through cost recovery charges in the grant budget, either as a data management fee or through the involvement of library faculty and staff as co-P.I.s or researchers on the grant, with associated line item cost recovery. This will be a one-time, cost recovery only fee that can be incorporated into the grant proposal budget. Data will be preserved and made accessible for the long term at no additional cost to the project beyond the one-time initial cost. However, that initial cost, although negotiable, will be based on the amount of work and effort anticipated for the life of the project." ● Johns Hopkins[23] does not clearly state their cost model for partnering on grant projects, however, their Data Conservancy services have been reported in webinars and conference presentations to be a flat percentage of the grant (estimated to range from 1-3%).

2.4 Partners

A comprehensive array of data management and curation exceeds the capacity of any single university unit or perhaps even single institution acting on its own. Therefore the Libraries will seek to work in coordination and partnership with the Office of the Vice President of Research and its program and service units (e.g., Minnesota Supercomputer Institute, University of Minnesota Informatics Institute), Information Technology, and other academic and support units. In areas where achievement of scale and aggregation are critical, the Libraries, with its University partners, will seek multi-institutional collaborations, likely in support of scaled infrastructure (e.g., networking, data storage, repository, etc.).

2.5 Investments and Costs

Through process mapping process (SIPOC) that evaluated the infrastructure, people, resources, and technology required to implement the DMCI Services (see results of the SIPOC in appendix), the

[22] Rutgers University benchmarks obtained May 6, 2014 from https://rucore.libraries.rutgers.edu/research/about/#fee
[23] Johns Hopkins Webinars on the Data Conservancy can be found at https://dataconservancy.org/education/webinars/

following infrastructure and costs were determined.

1. **Infrastructure Investments:** Initially, the DMCI Services will be build around the Libraries' existing repository infrastructure already in place for the University Digital Conservancy (UDC). The DMCI plans to develop a "Future Architecture Strategy" roadmap that may recommend different technologies in the long-term. This decision includes significant costs savings in the form of people and expertise needed to support the primary technological infrastructure for the DMCI Services. However, some modifications are required, including:
    a. **Upgrading Existing Repository Technology**: Migrating the University Libraries' existing digital repository infrastructure into the latest DSpace version (4.x), includes features such as the "available by request" functionality, versioning, etc. By depositing the data into the existing platform (via a custom submission interface and a web-based repository collection view), this will avoid multiple instances of DSpace for the Libraries to manage.
    b. **Data Storage, Backup, and Preservation**: Expanding the repository storage and backup capacity to accommodate the influx of digital research data. In addition, the data repository will be incorporated into the Libraries' digital preservation management system (currently being procured for implementation) and adhere to our existing digital preservation framework and preservation policies.
    c. **Notification System**: An electronic mechanism (email, RSS feed, etc.) to alert the appropriate library staff that a new data set has been submitted and is ready for curation actions. This notification should include the library liaisons.
    d. **Web Presence**: In addition to the "view" of the data collection afforded by the repository technology, there should be a separate identity and web presence for the Data Repository in order to promote to university researchers.

2. **Administrative Investments:** There are a few ongoing and new costs that are anticipated for the DMCI Services, such as a subscription with DataCite's DOI service either through Purdue or CDL. The 2012 estimate for this service was $2,500 per year for an institutional subscription. Also, and costs associated with implementing ORCIDs at the university, which are currently unknown.

3. **Staff Investments**: There are several recommendations for staff investments in support of the DMCI Services. Many of these staff are existing members of the libraries, however, several new requests are made (see budget below).

Table: Staffing Plan for DMCI Services

<table>
<tr><th colspan="4">Data Management and Curation Lead</th></tr>
<tr><th>Sevice Level</th><th>Coordination</th><th>DMCI Staff</th><th>Library Staff</th></tr>
<tr><td>DMP Consultation</td><td rowspan="4">Data Services Coordinator</td><td></td><td rowspan="2">Library Liaisons</td></tr>
<tr><td>Metadata and Data Management Support</td><td>Metadata Strategist</td></tr>
<tr><td>Repository and Curation Service</td><td rowspan="2">Data Curation Specialists (4-5 staff)</td><td rowspan="2">Digital Repository Developers</td></tr>
<tr><td>Grant Partner</td></tr>
</table>

- Leadership: One library staff member (Data Management/Curation Program Lead, 100%) that can give direction and leadership to the program, promote the service, and can strategically connect the DMCI Service to other library and non-library programs.
- Coordination Staff: One library staff member (Data Services Coordinator, 30%) that manages the day-to-day operations of the DMCI Services and respond to user-needs, questions, and concerns.
- Data Curation Staff: Two new graduate assistants (50%) will be hired and three existing library staff (20%) will reallocate part of their job to data curation services. The data curation staff will research and document procedures for curating data deposited into the library. These staff must have excellent software utilization skills and be able to implement new tools and follow procedures that use specialized software programs easily. Once established, these procedures will be used to re-tool and train other library staff as the workload is better understood. Learning from the 2013 data curation pilot, the following data types are expected to require specialized expertise prior to launch:
    a. Scientific Data: Hire two new graduate research assistants (Scientific Data Specialists, 50% x2) from the scientific/computing disciplines for a temporary appointment in the libraries. Overseen by the DMCI Lead, the RAs will help establish curation procedures for the variety of data types that we cannot anticipate. See draft position description in appendix.
    b. GIS/Spatial Data: Reallocate existing staff (Spatial Data Analyst/Curator, 20%).
    c. Social Sciences Data: Reallocate existing staff (Social Sciences & Professional Programs CLIR fellow, 20%).

   d. Digital Humanities Data: Reallocate existing staff (Arts & Humanities CLIR fellow, 20%).

- Metadata Staff: One dedicated library staff member (Research Metadata Specialist, 100%) and reallocate one senior library staff (Metadata Strategist, 20%) that will support the Metadata Consultation service. These staff are flexible to new and unknown researcher needs, service-oriented and have substantial expertise in metadata schemas, data models, and best practices that can consult on a variety of metadata and data organization methods.
- Technical and Repository Staff: The Data & Technologies unit provides overall support and development for the libraries' digital repository and preservation infrastructures. In addition to this support, the DMCI Services anticipate the following staff allotments:
   a. One repository administrator (Digital Repository Developer, 100%) that can manage the overall technical operations of hosting a data repository (web development, DSpace configuration, database support, etc.).
   b. A library staff member (Digital Preservation Analyst, 20%) to monitor and actively maintain the digital data held in the repository.
- Library Liaison Network: Liaison librarians should be trained to provide services for the Data Management Plan Consultation service and be supported with tools and guides. In addition, liaison staff should be able to consult with Curation Staff on incoming submissions to the library from their subject areas.

Table: Estimated Budget for DMCI Services

Recurring (annual)	Reallocation	New Investments
Staff Salaries/Benefits and Wages		
Data Management/Curation Program Lead (100%)	$ 92,400	
Data Services Coordinator (30%)	$ 20,400	
Senior Metadata Strategist (20%)	$ 19,000	
Research Metadata Specialist (100%)		$ 80,000
Repository Developer (100%)	$ 108,800	
Digital Preservation Analyst (20%)	$ 15,600	
Data Curation Staff		

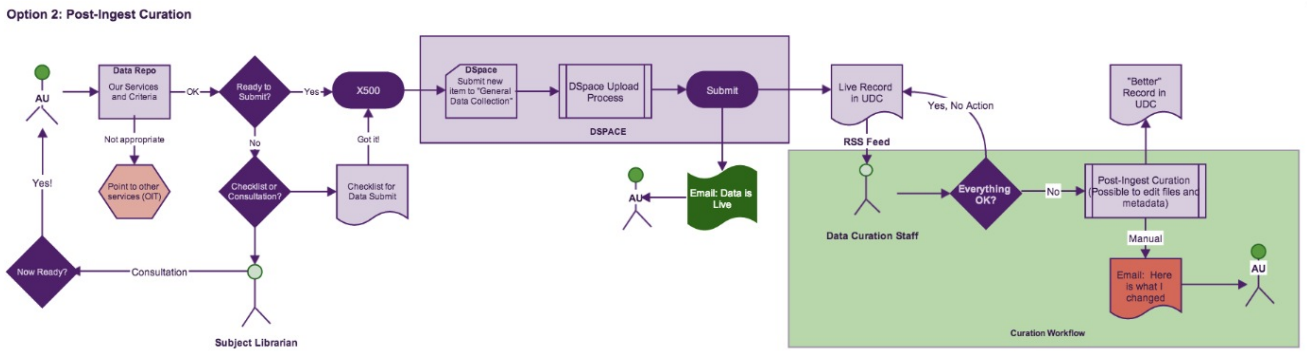
Scientific Data Research Fellow (50% graduate student, includes tuition benefits) x 2		$ 52,000
CLIR Social Sciences Data Research Fellow (20%)	$ 15,600	
CLIR Digital Humanities Data Research Fellow (20%)	$ 15,600	
Spatial Data Analyst/Curator (20%)	$ 15,600	
Administrative		
Annual license for DataCite DOI service	$ 3,000	
Infrastructure		
Expenses for data repository storage, backup, and preservation will be included in the Libraries' overall infrastructure expenses for digital asset management (access and preservation).	n/a	
	Reallocation	*New Investments*
Subtotals	$ 306,000	$ 132,000
Total Annual Estimated Costs	$ 438,000	

3.0 Technical Implementation

The University Libraries Data Management and Curation Initiative has been actively engaged in developing and preparing to launch the DMCI Services. This effort has involved significant staff time to meet the following objectives:

- Develop, test and release for use a repository explicitly designed to ingest datasets for curation and selection.
- Define related consultation services around data management plans, metadata creation, and data management training.
- Prepare library staff to provide services.
- Plan a communications strategy to market the new services to our primary audiences.

Graphic: Workflow for Data Repository and Curation Services



3.1 Timeline for Launching DMCI Services

The following milestones are projected in the launch of the DMCI Services.

- Service Development (Feb-May 2014): Develop a service model and business plan for the deployment of data management and curation services for campus.
- Development (May-July 2014): Repository technologies and related infrastructure are developed and tested. Train staff and refine operation workflow and procedures, etc.
- User Testing (July-August 2014): Various user testing and engagement with partners will help inform the refinement of the services.
- Partner Engagement: (July-October 2014): Establish referral networks and partnerships with units outside the libraries to help support researcher need beyond DMCI offerings.
- Soft Launch (September, 2014): The DMCI Services are included in the redesign and relaunch of the University Libraries' institutional repository, the University Digital Conservancy.
- Launch of DMCI Services (October 20, 2014): Promotion and marketing of services on campus.
- Post Launch (Nov-Dec 2014): The Services will be assessed and evaluated. Based on assessment, adjust, refine, and/or streamline services and processes.
- Future Development (2015): Develop and deploy enhancements to the repository infrastructure and service model.

4.0 Conclusions

This Business and Service Model presents the libraries with an overall plan to implement sustainable data management and repository services that include robust curation of all data ingested into the libraries.

Appendixes

A1. SIPOC Analysis of the DMP Consultation Service

A2. SIPOC Analysis of the Metadata Consultation Service

A3. SIPOC Analysis of the Data Repository and Curation Service

SIPOC DIAGRAM (PROCESS MAP)

Process Name: Data Management Plan (DMP) Consultatio

S	I	P										O	C
Suppliers	Inputs	Process										Outputs	Customer
			Step 1:		**Step 2:**		**Step 3:**		**Step 4:**				
Principal Investigator	Draft DMP		**Receive consultation request and schedule meeting with requestor**	>	**Meet with requestor and discuss best practices**	>	**Receive and review draft DMP**	>	**Deliver feedback to requestor**			Feedback in the form of notes in the requestor's DMP	Principal Investigator
Researchers (including non-grant-seeking researchers)	Abstract	**a.**	Provide point of contact (email, phone, online form, ServiceNow?	**a.**	Review the info that has been supplied by requestor.	**a.**	Receive the plan digitally.	**a.**	Email commented DMP to Author			Consent to share copy	Researchers (including non-grant-seeking researchers)
Grant coordinators	Grant instructions	**b.**	Receive contact in centralized place	**b.**	Identify questions or issues to consult with others.	**b.**	Read the Program Requirements	**b.**	Provide followup consultation as needed			Copy of DMP	SPA (indirect cust.)
Libraries (inputs #4-6)	Checklist (for UL eval)	**c.**	Assign request, initiate contact w/ actors	**c.**	Share with the user the relevant services.	**c.**	Read the draft plan	**c.**	Check-in with author to see if grant is funded			Data about missing elements, elements consistently included,	Grant coordinators
Other experts for advise and consultation	Checklist (for user?)	**d.**	Schedule consultation date/time (establish general TAT)	**d.**	Identify major DMP areas not covered.	**d.**	Use Checklist to identify any information that is missing	**d.**	Offer further assistance to implement plan			DMP Tool output (if available)?	
Funding agencies	Knowledge about best practices, etc.	**e.**		**e.**		**e.**	Use best practices for the field to identify anything incorrect.	**e.**				Early agreement for future data deposit?	
		f.		**f.**		**f.**	Note all recommendations and corrections in the draft	**f.**					
		g.		**g.**		**g.**	If time, send to another DMP reviewer	**g.**					
	Actors		DMP author, DMCI Coordinator, Subject liaisons		Liaison, DMP Author		Liaison, DMP reviewers		Liaison			**Actors**	
	Technology		Gmail?		Google calendar							**Technology**	
	Other Resources						Checklist, best practices					**Other Resources**	

See process details here >

SIPOC DIAGRAM (PROCESS MAP)

Process Name: Metadata Consultation (See Data Management Training bel

S	I	P								O	C
Suppliers	Inputs	Process								Outputs	Customer
			Step 1:		**Step 2:**		**Step 3:**		**Step 4:**		
Metadata specialist	Disciplinary standard		**UL staff receives inquiry or request for service.**	>	**UL staff and Researcher meet.**	>	**UL staff determines need.**	>	**Research and staff implement metadata system**	Metadata schema	Researcher
Researcher or graduate student	Example research data	a.	Provide point of contact (email, phone, online form, ServiceNow?	a.	Meet with the actors to establish need.	a.	Research available disiplinary standards	a.	Share metadata model with researcher.	Metadata implementation plan	PI
	Subject repository (if available)	b.	Receive contact in centralized place	b.	Gain access to example research data.	b.	Develop metadata standard to support researcher.	b.	Determine an implementation plan.		Grad student
		c.	Assign request, initiate contact w/ researcher	c.	Understand research process (use modified data curation profile)	c.		c.	Train research staff if needed.		
		d.	Schedule consultation date/time	d.		d.		d.			
	Actors				Researcher/group		Metadata Experts (UL staff)			**Actors**	
	Technology		Web form						Metadata software tools	**Technology**	
	Other Resources						Existing standards			**Other Resources**	

SIPOC DIAGRAM (PROCESS MAP)

Process Name: Data Management Training

S	I	P								O	C
Suppliers	Inputs	Process								Outputs	Customer
			Step 1:		**Step 2:**		**Step 3:**		**Step 4:**		
				>		>		>			
		a.		a.		a.		a.			

See process details here >

SIPOC DIAGRAM (PROCESS MAP)

Process Name: Data Repository & Curation Services

S	I	P								O	C
Suppliers	Inputs	Process								Outputs	Customer
			Step 1:		Step 2:		Step 3:		Step 4:		
Data Author	Data File(s)		Author submits data through web form	>	UL staff receive notification and assignment	>	Staff review and curate submission	>	Researcher reviews and data goes live.	Curated data set available online	Data author
Representative of Data Author (Graduate Student)	Documentation	a.	Navigates to web form	a.	Coordinator receives notifcation of new submission	a.	Curation Specialist logs into access Submission	a.	Specialist emails author with changes to submission	Record for the data in the repository	End user of data
	Metadata	b.	Selects "Data Repository" option	b.	Coordinator reviews submission using checklist	b.	Curation procceedure for data type are followed	b.	Author reviews submission	Updated procedure for that particular data type	Administration
	Proceedure for Collection Policy Checklist	c.	Locates, describes, and uploads data files	c.	Coordinator determines if criteria met	c.	Procedure is updated based on experience	c.	Specialists moves submission into "live" collection	Feedback from the author	Funders (compliance)
	IRB Agreement Consent Form	d.	Enters metadata	d.	Coordinator assigns submission to Curation specialist	d.		d.		Deposit agreement	
	Proceedure for Curation Specialist (by data type)	e.	Signs Deposit Agreement	e.	Curation specialist receives notication of new assignment	e.		e.			
	Submission Tracking tool	f.	Verifies entry	f.	Curation specialist adjusts schedule to complete request	f.		f.			
		g.	Clicks submit	g.		g.		g.			
		h.	Receives confirmation message	h.		h.		h.			
	Actors		Data Author		Coordinator, Curation Specialist		Curation Specialist, Data Author		Data Author, Curation Specialist	**Actors**	
	Technology		Repository upload form, email notification		Notification of new submissions to staff		Repository edit view		Repository public view	**Technology**	
	Other Resources				Tracking tool					**Other Resources**	