Data Repository and Curation Services, How Do We Compare? A Snapshot of Six Academic Library Institutions

Authors: Lisa R. Johnston, Jake Carlson, Patricia Hswe, Cynthia Hudson-Vitale, Heidi Imker, Wendy Kozlowski, Claire Stewart

Abstract
Many academic and research institutions are exploring opportunities to better support researchers in sharing their data. As partners in the Data Curation Network project, our six institutions developed a comparison of the current levels of support that we provide for researchers to meet their data sharing goals through library-hosted data repository and curation services. Service areas analyzed include repository services for data, technologies used, policies, and staffing in place. Though this process we aim to better define the current levels of support offered by our institutions as a first step toward meeting our project's overarching goal: to develop a shared staffing model for data curation across our multiple institutions.

Background
Funder requirements, institutional and journal data sharing policies, and new trends in research reproducibility are signaling that academic research will become increasingly more open in the coming years. And we believe that data curation is critical to ensuring that this movement is fully actualized. Our six institutions are beginning to dedicate some level of resources towards data curation services. In doing so we are interested in leveraging our individual progress to contribute to the greater data curation community. The six academic library-run data repository services compared here are participants in the Alfred P. Sloan Foundation-funded Data Curation Network project (https://sites.google.com/site/datacurationnetwork). The goal of the Data Curation Network project is to bring together institutions individually providing local support for data repository deposit and curation in order to develop a shared, cross-institutional staffing model for applying expert-level curation across more disciplines than any one institution could offer alone. This analysis captures our current levels of siloed support, which are still growing and evolving. This comparison will help the team design a shared service that fits within the existing scope of our institutions’ capacities, yet broadens our ability to curate a wider variety of digital data for researchers across our network. Our analysis is also intended to help others who are at the beginning stages of institutional data repository development and are scanning for examples of what peer institutions have implemented.

Despite this progress, data-specific curation activities are relatively new to academic libraries, and it’s clear to us that individually we have much to learn. In these early stages, formation of the Data Curation Network serves as a way for us to learn from each other about how to best curate datasets. However, moving forward we hope the Network will begin to enable

Follow our progress at https://sites.google.com/site/datacurationnetwork
the community to pragmatically and effectively provide added value when publishing datasets. Most importantly, by intentionally structuring our efforts to coordinate as a Network that can be grown and incorporate new institutions over time, we hope to play a role in empowering the larger data curation community through sharing of experiences and provide a platform for continued dialog and discussion in this area.

Method
To understand the baseline levels of service currently provided for data repository and curation the following six repositories were examined: the Data Repository for the University of Minnesota (DRUM), the eCommons at Cornell University, the Illinois Data Bank at the University of Illinois Urbana-Champaign, Deep Blue Data at the University of Michigan, ScholarSphere at Penn State University, and the Digital Research Materials Repository (DRMR) at Washington University at St. Louis. This is a sample of convenience which is based on the institutions’ involvement with the Data Curation Network project. A project team member from each institution (authors) was asked to write a summary report and address specific questions (presented here as tables) based on their own knowledge and experience. Following the self-reporting exercise, each team member gave a 20-minute webinar presentation to the project team to further clarify their responses. The results of this exercise were captured and presented in this report for sharing with peer institutions. This is not intended to be a scientific comparison or a comprehensive representation of existing data repository and curation services in the field. Rather, this review is a snapshot in time--the six institutional service offerings represented here will change and grow in the future. Additionally, this report focuses primarily on data repository services and their parameters. Issues around the mechanics of data curation and specific steps taken to prepare data for sharing will be addressed in greater depth in future reports by the Data Curation Network project.

1.0 Data Repository Services Overview
Each of the six institutions currently provide data repository services. They do so either as a service of the traditional institutional repository or IR (Minnesota, Cornell, Penn State, Wash U) or via a dedicated data repository (Illinois, Michigan). All of the repositories make content available on an open access basis, meaning the data housed in these repositories are publicly accessible for search, retrieval, and download.

Data Repository for the University of Minnesota (DRUM)
Institution: University of Minnesota, Twin Cities (Minneapolis, Minnesota)
URL: http://hdl.handle.net/11299/166578

Follow our progress at https://sites.google.com/site/datacurationnetwork
Launched: November 15, 2014
Data Holdings: 58 data records in DRUM and 315 in the IR as of June 23, 2016
The University of Minnesota (U of M) Libraries has been providing research data management services for a number of years, including support for writing data management plans, educational training and workshops, and consultation (see http://lib.umn.edu/datamanagement). The Libraries launched the Data Repository for the University of Minnesota (DRUM) in 2014 for U of M researchers to self-deposit their data for long-term open access and reuse when no other discipline-appropriate data repository exists. DRUM resides within the existing institutional repository service, the University Digital Conservancy, as a sub-collection with a custom metadata schema and submission workflow. An example dataset in DRUM is shown in Figure 1. All data submitted to DRUM undergo curatorial review by one of five library data curators who collaborate with the data author to ensure that the data are in a format and structure that meet our policies and best facilitate reuse.

**eCommons at Cornell University**
Institution: Cornell University (Ithaca, New York)
URL: http://ecommons.cornell.edu
Launched: Fall 2002
Data holdings: 106 data records as of June 29, 2016
The purpose of eCommons is to provide stable, long-term public access to digital content produced by members of the Cornell University community and its sponsored associates. Because policies and submission processes are the same for datasets as other content, our approach to providing open and persistent access to research results is to accept all forms of “scholarly output” in Cornell Library’s institutional repository. We encourage use of eCommons for data, particularly when there are no appropriate discipline-based repositories available, or when a researcher doesn’t wish to incur a cost associated with their deposit. Data submitted to eCommons are assigned a type “dataset” for discovery purposes, and can be added to the organizational collection of the submitter’s choice. As of 2015, datasets undergo a discovery metadata review, and some receive an additional curation of science metadata and data file format and structure. Most science metadata are submitted as readme files, but standardized metadata are accepted as item files. If a researcher rejects suggestions of the curator, data are still accepted to the repository. eCommons at Cornell University launched in the fall of 2002, and the first dataset was deposited in 2005. An example dataset from eCommons at Cornell is displayed in Figure 1.

**ScholarSphere**
Institution: Penn State University (State College, Pennsylvania)

Follow our progress at https://sites.google.com/site/datacurationnetwork

---

URL: https://scholarsphere.psu.edu/
Launched: Fall 2012
Data Holdings: 439 public data records as of June 23, 2016
ScholarSphere is a self-deposit repository service through which faculty, students, and staff at Penn State are able to share their work, including research data sets, on a worldwide scale and be assured of its long-term preservation and thus access. The main impetus behind designing ScholarSphere was to help researchers comply with research data management requirements, as well as with increasing requirements from publishers to link research articles to the data sets associated with them. At the same time, until ScholarSphere, Penn State had not had an institutional repository capturing the scholarly record of its faculty, students, and staff for preservation and access purposes. (There has been an electronic thesis and dissertation service since the mid-2000s, but the university perceived a need for a service accepting a broader array of scholarship - hence, the decision for ScholarSphere to take in both data sets and conventional scholarly publications.) The University also has a stand-alone, mediated-deposit data repository, DataCommons, more specifically geared toward earth and environmental sciences, including geosciences. We connect our researchers to data repositories beyond Penn State as needed via consultation and via a LibGuide for research data management services, which points users to re3data, an online index of data repositories, and to repository services known to accept data sets, such as figshare and Zenodo. Data submitted to ScholarSphere do not undergo any curatorial review, apart from an automatic audit of the files for preservation purposes. An example data record in ScholarSphere is shown in Figure 2.

Deep Blue Data
Institution: University of Michigan (Ann Arbor, Michigan)
URL: https://deepblue.lib.umich.edu/data
Launched: February 29, 2016
Data Holdings: 18 data records as of June 23, 2016
Deep Blue Data is a repository offered by the University of Michigan Library that provides access and preservation services for digital research data that were developed or used in the support of research activities at U-M. Deep Blue Data is a component of a suite of services provided by the U-M Library designed to broadly disseminate the intellectual contributions in research, teaching and creativity made by the University of Michigan community and to ensure its longevity. It is a companion repository to Deep Blue, which serves to provide access to papers, presentations, reports and other human readable scholarship from the University of Michigan. Our primary goal in offering research data services is to connect researchers to resources that are best suited to support their specific needs for their data. In cases where subject-based data repositories and services are available that meet a researcher’s needs we will consult with the researcher and the repository to assist with the submission process as appropriate. However, researchers in many fields do not yet have a data repository devoted to their
needs, or in some situations the disciplinary repository is not a viable option. The Deep Blue Data repository was developed to provide these researchers with the means to satisfy requirements and take advantage of the benefits that sharing and curating data affords. As we continue to develop the capabilities of Deep Blue Data our intent is to go beyond providing a place to put data and create more of a platform for others to interact with the data in ways that add value. An example data record from the beta interface of Deep Blue Data is shown in Figure 2.

**Illinois Data Bank**
Institution: University of Illinois at Urbana-Champaign (Illinois)
URL: [https://databank.illinois.edu/](https://databank.illinois.edu/)
Launched: May 16, 2016
Data Holdings: 7 data records as of June 23, 2016
The Illinois Data Bank's mission is to centralize, preserve, and provide persistent and reliable access to the research data created by affiliates of the University of Illinois at Urbana-Champaign, such as its faculty, academic staff, and graduate students. The Illinois Data Bank is intended to be responsive to the Illinois research community, is supported by the University of Illinois at Urbana-Champaign, and endeavors to be both durable and sustainable. The Illinois Data Bank is a platform for making datasets created from research projects by University of Illinois at Urbana-Champaign researchers publicly accessible by seeing that the research data is both widely discoverable and linked to associated works, such as journal articles, source code, or data deposited elsewhere. During consultations we may point to alternative repositories and encourage depositors to reconsider if a more appropriate repository is available. We elected to go with development of a web application that interacts directly with our preservation system in order to leverage that system's functionality and allow us to focus our long-term efforts on centralizing our preservation efforts. Depositing research data into the Illinois Data Bank is voluntary. An example data record in Illinois Data Bank is shown in Figure 3.

**Digital Research Materials Repository**
Institution: Washington University in St. Louis (Missouri)
URL: [http://openscholarship.wustl.edu/data](http://openscholarship.wustl.edu/data)
Launched: January 5, 2015
Data Holdings: 3 data records as of June 23, 2016
The purpose of the digital research materials repository (DRMR) is to provide a long-term, institutional home for research data and supplemental materials produced at Washington University in St. Louis (WUSTL). A free service of the University Libraries, DRMR curates data and the supporting documentation used to verify or support research, including any analysis scripts, data dictionaries, and domain metadata. The DRMR at WUSTL is a companion collection within our institutional programs.

Follow our progress at [https://sites.google.com/site/datacurationnetwork](https://sites.google.com/site/datacurationnetwork)
repository, Open Scholarship, which serves to provide access to dissertations, theses, and other scholarly output of the university. DRMR provides a data archiving solution for the WUSTL community who do not have an appropriate discipline or domain repository available to them, or do not want to incur the costs of deposit. Once submitted to DRMR, datasets and submitted materials undergo archival processing and curation treatments. Curators work directly with WUSTL researchers to enhance records and documentation for reuse and accessibility. An example data record in DRMR is shown in Figure 3.
Figure 1: Example dataset in DRUM (left, http://dx.doi.org/10.13020/D6PK5C) and in Cornell eCommons (right, http://hdl.handle.net/1813/43783) which both use the DSpace 5.x software.
A report from the Data Curation Network Project released June 30, 2016


Follow our progress at https://sites.google.com/site/datacurationnetwork
A report from the Data Curation Network Project released June 30, 2016


Figure 2: Example dataset in Penn State’s ScholarSphere (left, https://scholarsphere.psu.edu/files/m900nt50p) and Michigan’s Deep Blue (right, https://deepblue.lib.umich.edu/data/concern/generic_works/707957642) both using Hydra with Fedora.

Follow our progress at https://sites.google.com/site/datacurationnetwork
A report from the Data Curation Network Project released June 30, 2016


Figure 3: Example dataset in the Illinois Databank (left, https://doi.org/10.13012/J8PN93H8) and Wash U’s Digital Research Materials Repository (right, https://doi.org/10.7936/K7J6F60).

Follow our progress at https://sites.google.com
1.1 Data Curation Workflows

The comparison of curation workflows (illustrated in Table 1) demonstrate how a “data set” typically flows through the curation process prior, during, and post-ingest to the local data repository services offered at by the six institutions. Data sets are defined by each institution as:

- **Minnesota**: A dataset is one or multiple files that comprise a collection of facts that may be further analysed, interpreted, or understood. The information may be presented as number, text, image, video, audio, or other medium.

- **Cornell**: eCommons has not formalized the definition of a dataset. Because submitters can self-assign types, some “datasets” are in the form of PDFs or other tables. Recently, submissions typed as “dataset” undergo a curatorial review (see Services Overview), and submitters are encouraged to allow a change of the assigned type if a curator deems the data to be in a format that cannot be easily evaluated (contain appropriate metadata) and re-used (are provided in an accessible format).

- **Illinois**: An intentional collection of Data Files and/or Metadata Files together with the associated Descriptive Metadata. Note: A Dataset may be as simple as a single Data File and/or Metadata File and the associated Descriptive Metadata. A Dataset could also be a package of any combination of raw, semi-processed, or aggregate Data Files, relevant processing and analysis scripts and Metadata Files such as a readme file. (other def here https://databank.illinois.edu/policies)

- **Michigan**: For the purposes of Deep Blue Data, research data are defined as representations of observations, objects, or other entities used as evidence of phenomena for the purposes of research or scholarship. In practical terms, Deep Blue Data will accept data that were developed or used in the support of research activities of U-M faculty, students and staff.

- **Wash U**: A dataset is the commonly accepted material(s) used to verify research claims. Data is an inclusive term that comprises images, coordinates, field sheets, numbers, metadata, and any other documentation that is the result of scholarship.

Our comparison found that each curation workflow is based on a self-submission model allowing researchers to deposit their data at will. All but one repositories (Minnesota) automatically accepts the data once deposited. All but one repository (Penn State) provides post-ingest curatorial review of the deposited files and metadata. Persistent identifiers in the form of a DOI are added in various ways. These similarity are encouraging and may allow our model to scale data curation work across the institutions in a similar post-ingest manner. Four institutions provided illustrative diagrams that depict this curation workflow process and they appear as Fig. 4 (Minnesota), Fig 5 (Cornell), Fig. 6 (Illinois), and Fig. 7 (Wash U.).

Follow our progress at [https://sites.google.com/site/datacurationnetwork](https://sites.google.com/site/datacurationnetwork)

Follow our progress at https://sites.google.com/site/datacurationnetwork
Table 1: Comparison of the data curation workflows at the six institutions

<table>
<thead>
<tr>
<th>Service Workflow Steps by Institution</th>
<th>Pre-ingest Curation?</th>
<th>Mediated vs Self-deposit?</th>
<th>Accept/Reject Stage?</th>
<th>Public</th>
<th>Post-ingest curation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Consult only</td>
<td>Staging Area for deposit</td>
<td>Mediated deposit</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Minnesota</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Cornell</td>
<td>X</td>
<td>X*</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Illinois</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Michigan</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Wash U</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Penn State</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

* On request  ** When possible

Follow our progress at https://sites.google.com/site/datacurationnetwork
Follow our progress at https://sites.google.com/site/datacurationnetwork
Follow our progress at https://sites.google.com/site/datacurationnetwork

Figure 5. Cornell University Data Curation Workflow Diagram (Large View)

Follow our progress at https://sites.google.com/site/datacurationnetwork
A report from the Data Curation Network Project released June 30, 2016


Follow our progress at https://sites.google.com/site/datacurationnetwork
A report from the Data Curation Network Project released June 30, 2016


Figure 7. Washington University St. Louis Data Curation Workflow Diagram (Large View)

Follow our progress at https://sites.google.com/site/datacurationnetwork
1.2 Tracking Data Curation Activities

Data curation services may involve augmentation to the metadata, file format transformations (e.g., preservation friendly file formats), and documentation added to the record. Each repository tracks these changes to the data deposit in a variety of ways.

- **University of Minnesota**: Before making any changes, curators create a working copy of the submission and store the original files and metadata as a back-up copy, in case reversion is needed. During the curation process, staff keep a text-based curator's log file detailing all changes made during the curation process. The curators also (manually) capture all relevant correspondence with author (e.g., email exchanges) regarding the changes made and save with the log. This log file is archived with the dataset in DRUM but not made publicly available.

- **Cornell University**: Prior to submission, the curator documents all interactions, either in person, or via email, on an internal wiki; no strict format/standard yet in place. Once submitted, changes are tracked by DSpace in a basic provenance record (date, time, user), and the curator logs any additional, relevant information to both the discovery and science metadata.

- **Penn State University**: Versions are automatically tracked in ScholarSphere, so if there are metadata changes, then the system is monitoring these; the depositor can backtrack to the earlier version(s) as needed and select the one that he/she would like to expose publicly. There is no notification to the repository service manager when deposits are made to ScholarSphere.

- **University of Illinois at Urbana-Champaign**: We've implemented a ticketing system (OTRS). All deposits automatically create a ticket. After the curation review, depositors get an email documenting changes (even if none) or asking questions as needed. Metadata changes available as changelog; file changes would occur as versioned datasets.

- **University of Michigan**: RDS Manager & Data Curation Librarian are sent an email notifying of a new deposit. Data Curation Librarian contacts appropriate Subject Liaison and they review submission. Any issues, questions or suggestions for improvements with the deposit are documented and depositor is contacted via email and asked if action should be taken. Interactions with depositors are currently tracked in a spreadsheet, and use of a ticketing system is being explored.

- **Washington University in St. Louis**: Treatment action text files are created and email exchanges documented and stored with with AIP along with an image of the original submission, Bitcurator reports, and checksums generated. Additionally, a spreadsheet tracks the processing of the submission as it travels along the curation workflow.

Follow our progress at [https://sites.google.com/site/datacurationnetwork](https://sites.google.com/site/datacurationnetwork)
Tracking data curation activities will be a key aspect of the resulting Data Curation Network model in order to measure the levels of curation staffing needs for particular disciplines, to monitor the time involved, and to demonstrate efficiencies gained by each Network participant.

### 2.0 Data Repository Technologies

Each of the institutional repositories use software to manage the digital assets in their data repository service. Two systems use DSpace\(^1\) (Minnesota, Cornell), two use or intend to use Sufia\(^2\) running on a Hydra/Fedora platform (Michigan, Penn State), Illinois runs a custom solution known as Medusa\(^3\), and Washington University at St. Louis uses Digital Commons by BePress\(^4\). The specific software versions, upload limitations, features, metadata schemas, and support for external services are compared in Table 2. It will be critical for the Data Curation Network to work across a variety of technology solutions and this cross-section provides an excellent base from which to build on.

#### Table 2: Comparison of the technology, features, and limitations for the six institutions’ repositories for data

<table>
<thead>
<tr>
<th>Data Repository for the Univ of Minnesota (DRUM)</th>
<th>eCommons at Cornell Univ</th>
<th>Illinois Data Bank at Univ of Illinois</th>
<th>Deep Blue Data at the Univ of Michigan</th>
<th>ScholarSphere at Penn State Univ</th>
<th>Digital Research Materials Repository (DRMR) at Wash U St. Louis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technology Platform</td>
<td>DSpace 5.5</td>
<td>DSpace 5.5</td>
<td>Custom-built Ruby on Rails webapp as a microservice to Medusa, a local preservation repository</td>
<td>Hydra/Fedora Sufia 7</td>
<td>Hydra/Fedora (soon to be Sufia 7)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

\(^1\) DSpace is open source repository software offered by the DuraSpace organization at [https://wiki.duraspace.org/display/DSDOC5x/Release+Notes](https://wiki.duraspace.org/display/DSDOC5x/Release+Notes).

\(^2\) Sufia is an open source repository front-end application for Hydra and Fedora available at [http://sufia.io/](http://sufia.io/).

\(^3\) Medusa is a digital preservation repository developed at the University of Illinois at Urbana-Champaign, read more at [https://wiki.cites.illinois.edu/wiki/display/LibraryDigitalPreservation/Medusa+FAQ](https://wiki.cites.illinois.edu/wiki/display/LibraryDigitalPreservation/Medusa+FAQ).

\(^4\) Digital Commons is a fee-based hosted digital repository solution provided by BePress at [http://digitalcommons.bepress.com/](http://digitalcommons.bepress.com/).

Follow our progress at [https://sites.google.com/site/datacurationnetwork](https://sites.google.com/site/datacurationnetwork)
<table>
<thead>
<tr>
<th>Upload limits</th>
<th>Self-deposit up to 2 GB per file. Larger files must be mediated. Total size per project per year is 10GB.</th>
<th>Self-deposit up to 2 GB per file. Larger files must be mediated. Total size per project per year is 10GB.</th>
<th>Self-deposit up to 2 GB via Box.com. Larger files may be ingested via a mediated mechanism.</th>
<th>Self-deposit up to 2 GB per file. Larger files via Dropbox (1.9 GB) or Box (5 GB). Up to 100 files and totaling less than 1 GB in size.</th>
<th>Recommended 2 GB per file (not a hard limit - up to 10-20 GB).</th>
</tr>
</thead>
</table>
| Features       | - Open Access  
- Versioning (mediated)  
- API  
- OAI/PMH feed | - Open Access  
- Versioning (mediated)  
- Related material linking  
- OAI/PMH | - Open Access  
- Descriptive metadata editing  
- Dataset Versioning (mediated)  
- Related material linking within published metadata (mediated) | - Open Access  
- Related Material linking (next deployment) | - Open Access  
- Versioning (mediated)  
- OAI/PMH feed  
- Related material linking (mediated) |
| Service/Software Add-ons | - DataCite DOI (mediated and only upon request) | - DataCite DOI  
- Box.com integration  
- ORCID integration | - DataCite DOI  
- Box.com integration (next deployment) | - Dropbox  
- Box.com integration | - DataCite DOI |
| Discovery Services | - Web indexing: full-text and metadata  
- Data Citation Index (WoS)  
- Datacite.org  
- re3data.org  
- SHARE | - Web indexing: full-text and metadata | - Web indexing: metadata only  
- Datacite.org  
- re3data.org  
- SHARE (forthcoming) | | -Web indexing  
- Datacite.org  
- SHARE |
| Metadata Schema | Dublin Core | Dublin Core | Compatible with DataCite Metadata Schema 3.1 | Dublin Core | Dublin Core |
| Published schema | [http://hdl.handle.net/11299/171761](http://hdl.handle.net/11299/171761) | Not yet published | [https://www.ideals.illinois.edu/handle/2142/91019](https://www.ideals.illinois.edu/handle/2142/91019) | Not yet published | Not yet published |

Follow our progress at [https://sites.google.com/site/datacurationnetwork](https://sites.google.com/site/datacurationnetwork)
3.0 Policy Comparison

Policy development is a critical component of developing data repository and curation services. The institutions all have publically viewable policies for deposit, access, documentation, and preservation (compared in Table 3). However, several challenging policy limitations and themes emerged from our discussions.

- **Undefined documentation requirements**: Several institutions (Minnesota, Wash U, Michigan, Cornell, Penn State) described their policies for what constitutes adequate documentation for a data deposit to be vague. Michigan said “Expected documentation is only loosely defined in the policy.” and Penn State said “We could define our documentation requirements, period.”

- **Difficulty in determining who can deposit**: All six institutions require at least one author be a institutional affiliate to deposit their data. However, our Illinois team member reports “There are lots of collaborations and infrastructure projects at our university, so some asking to allow data deposit where an Illinois affiliate is not always an author. Similarly, some centers and projects want to be labeled at the data author or the long-term contact (e.g. organization as author).”

- **Sensitive data concerns**: None of the repositories allow data deposit that contain private data. Illinois mentioned “Lots of issues around sensitive data, third party data and Data Use Agreements (DUAs)” Michigan said “We frequently encounter researchers with sensitive data issues who would like guidance on how to share their data. We are still learning how we can respond effectively.”

- **Overlapping or competing data repositories**: If the institution houses other data repositories, scope can become an issue. Minnesota said “We have a large medical school with separate clinical data repository and a do-it-ourselves ethic limits our outreach in this side of campus.” and Penn state reported “There are two other repository services at Penn State, in addition to ScholarSphere. These are DataCommons and Penn State Law eLibrary. Depositors would benefit from a clearer, more explicit expression of our policies, particularly around the scope of our collections.”

- **Access control**: Some institutions provide authors the ability to embargo or temporarily restrict access to their data deposits (Minnesota, Cornell, Illinois, WUSTL). Cornell said “We do get submitters who want to control access (either to Cornell community, or only “upon request”).”

The Data Curation Network must consider conflicting policy issues, build a shared understanding (e.g., memorandum of agreement), and create a governance model that addresses the unique needs and restrictions in place at each institution.

*Follow our progress at* [https://sites.google.com/site/datacurationnetwork](https://sites.google.com/site/datacurationnetwork)
Table 3: Comparison of policies for data repository and curation service at the six institutions

<table>
<thead>
<tr>
<th></th>
<th>Data Repository for the Univ of Minnesota (DRUM)</th>
<th>eCommons at Cornell Univ</th>
<th>Illinois Data Bank at Univ of Illinois</th>
<th>Deep Blue Data at the Univ of Michigan</th>
<th>ScholarSphere at Penn State Univ</th>
<th>Digital Research Materials Repository (DRMR) at WUSTL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Link to Deposit License Agreement</td>
<td>[<a href="https://conserva">https://conserva</a> ncy.umn.edu/pages/drum/policies/#deposit-license](<a href="https://conserva">https://conserva</a> ncy.umn.edu/pages/drum/policies/#deposit-license)</td>
<td><a href="https://ecommons.cornell.edu/page/policy#license">https://ecommons.cornell.edu/page/policy#license</a></td>
<td><a href="https://databank.illinois.edu/policies#deposit_agreement">https://databank.illinois.edu/policies#deposit_agreement</a></td>
<td><a href="https://deepblue.lib.umich.edu/data/agreement">https://deepblue.lib.umich.edu/data/agreement</a></td>
<td><a href="https://scholarsphere.psu.edu/agreement/">https://scholarsphere.psu.edu/agreement/</a></td>
<td><a href="http://openscholarship.wustl.edu/data/policies.html">http://openscholarship.wustl.edu/data/policies.html</a></td>
</tr>
<tr>
<td>Data Type Collection Scope</td>
<td>General, all data</td>
<td>General, all data (but recommend disciplinary repos when available and appropriate)</td>
<td>General, from either research project or with an expectation to be used for research purposes (but recommend disciplinary repos when available)</td>
<td>General, all data that conform to our collections policy (we assist researchers in identifying disciplinary repositories if desired)</td>
<td>General, all data (but we recommend other data repositories as relevant)</td>
<td>All data</td>
</tr>
<tr>
<td>Deposit</td>
<td>U of M affiliates (Shib log in, no auth required)</td>
<td>Cornell affiliates (via Shib log in or upon request to admins after justification)</td>
<td>Illinois affiliates (log in - restricted to fac, grad student and staff groups)</td>
<td>Michigan affiliates (log in)</td>
<td>Penn State affiliates (Shib login, no auth required)</td>
<td>WU affiliates (log in)</td>
</tr>
<tr>
<td>Private data (PII)</td>
<td>Not accepted, run Identity Finder to be sure</td>
<td>Not accepted</td>
<td>Not accepted</td>
<td>Not accepted</td>
<td>Not accepted</td>
<td>Not accepted</td>
</tr>
<tr>
<td>Access</td>
<td>Default open access (option to approve individual requests for access for up to 2 years)</td>
<td>Default open access; can withhold access until related publication released (mediated; not advertised as an &quot;embargo&quot;).</td>
<td>Default open access but option during deposit to embargo either the entire dataset or just the files for up to 1 yr</td>
<td>Open access. No embargos at this time.</td>
<td>Open Access. No embargos currently.</td>
<td>Open Access. Possible to embargo for up to 2 years.</td>
</tr>
<tr>
<td>Restrictions</td>
<td>Data must have &quot;adequate&quot;</td>
<td>No confidential information,</td>
<td>no private, confidential, or</td>
<td>no private, confidential, or</td>
<td>n/a</td>
<td>no private, confidential, or</td>
</tr>
</tbody>
</table>

Follow our progress at [https://sites.google.com/site/datacurationnetwork](https://sites.google.com/site/datacurationnetwork)
A report from the Data Curation Network Project released June 30, 2016  

<table>
<thead>
<tr>
<th>Documentation and be original, unique</th>
<th>Proprietary information of others or export controlled information.</th>
<th>Other legally protected information</th>
<th>Other legally protected information. Data must have research value.</th>
<th>Other legally protected information</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Documentation</strong></td>
<td><strong>“Adequate for reuse”</strong></td>
<td>None required but strongly encouraged (and assistance offered)</td>
<td>None required (outside of some basic metadata) though &quot;A detailed description of a data’s origins, purpose, and use” is strongly encouraged.</td>
<td>n/a</td>
</tr>
</tbody>
</table>

| Real-time, data, active and "streaming" | Not accepted | Not accepted | Not accepted | n/a |

| **Preservation Commitments** | Files preserved at least 10 years via Rosetta | "...committed to preserving the binary form of the digital object..."; no commitment to format migration | Minimum of 5 yrs via the preservation repository (Medusa) | Minimum of 10 years. 3 tiers of commitment depending on format. |

| Creative Commons License | Author-specified (optional) | Author-specified, but optional. CC0, and CC-BY, CC-BY-ND, CC-BY-SA, CC-BY-NC, CC-BY-NC-ND, CC-By-NC-SA; Other licenses (eg. ODC) can be added via metadata) | CC0 and CC BY encouraged, licence.txt allowed | Author-specified (Required) CC0, CC BY, CC BY NC (other licenses will be considered if requested) |

| **End-User Terms of Use** | Appropriate attribution required for use, | None | | |

Follow our progress at [https://sites.google.com/site/datacurationnetwork](https://sites.google.com/site/datacurationnetwork)
4.0 Staffing for Data Repository and Curation Services

Of the six institutions’ reported staffing levels, one commonality was the heavy reliance on partial or shared staff that dedicates only a percentage of their time to data repository and curation services. In fact, for the six team members, this was the case for each of our positions. Table 4 describes the levels of staffing for the six services and is followed by a brief description of the organizational oversight and staffing structure in each case. The implications for this baseline metric are key for the Data Curation Network. A shared staffing model across the Network will provide each of our services with a an infusion of expert staff that will increase the collective capacities for offering data curation services and allow our individual services to scale.

Table 4: Comparison of staffing levels for data repository and curation services

<table>
<thead>
<tr>
<th></th>
<th>Minnesota</th>
<th>Cornell</th>
<th>Illinois</th>
<th>Michigan</th>
<th>Penn State</th>
<th>Wash U</th>
</tr>
</thead>
<tbody>
<tr>
<td># full-time staff (service / technical)</td>
<td>0 / 0</td>
<td>0 / 0</td>
<td>1 (developer)</td>
<td>1/0</td>
<td>0/0</td>
<td>0/0</td>
</tr>
<tr>
<td># of shared staff (service / technical)</td>
<td>1.5 FTE / .5 FTE</td>
<td>Difficult to separate “data” staff from IR staff.</td>
<td>1.4 FTE (service)</td>
<td>Difficult to quantify as data being</td>
<td>.75/1.5 - but may be changing</td>
<td>0.65FTE (service)</td>
</tr>
</tbody>
</table>

Follow our progress at https://sites.google.com/site/datacurationnetwork
<table>
<thead>
<tr>
<th>1 Coordinator (20%)</th>
<th>Totals: ~1 FTE / ~0.5 FTE</th>
<th>2 data curators (33%)</th>
<th>incorporated into liaisons responsibilities and tech staff assignment is somewhat fluid based on library needs.</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 data curators (10%)</td>
<td>1 Coordinator (30%)</td>
<td>5% Digital Preservation Librarian</td>
<td>RDS 2.25 FTE</td>
</tr>
<tr>
<td>1 data curator (50%)</td>
<td>1 collection-specific data curation (liaison responsibility)</td>
<td>15% of Life Sciences Data Librarian</td>
<td>1 Developer significantly invested</td>
</tr>
<tr>
<td>----------------------</td>
<td>----------------</td>
<td>---------------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>1 preservation analyst (10%)</td>
<td>1 data curator / science metadata librarian (5-10%)</td>
<td>20% of Metadata Librarian</td>
<td>1 project manager</td>
</tr>
<tr>
<td>2 developers (unknown %)</td>
<td>1 metadata librarian (~5%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 coordinator (30%)</td>
<td>1 developer (~25%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 admin (~25%)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### List any non-library staff
1. 1 data curator from College of liberal arts (ad hoc)
2. 5% of a central IT person for UX (volunteer)

### 4.1 Organizational Approaches to Data Repository and Curation Services
Each institution has a unique approach to how data curation services fit within the broader campus landscape. Understanding these relationships will aid in developing clear incentives for joining the Data Curation Network that reaches stakeholders both within and external to the library.

#### University of Minnesota Organizational Structure
Link: n/a
1. University Oversight: n/a
2. Library Oversight: AUL for Research and Learning/Libraries Cabinet
3. Lead/Director for Data Curation Services: Director of DRUM (Lisa Johnston, 30%)

---

Follow our progress at [https://sites.google.com/site/datacurationnetwork](https://sites.google.com/site/datacurationnetwork)
a. Direct Report Curation Staff: Scientific Data Curator, a graduate research assistant at 50% time (Fall/spring only)

b. Non-Direct Report Curation Staff:
   i. DRUM Coordinator (Repository Archivist, 10%) → reports to University Archivist
   ii. Health/human subjects data curator (10%) → reports to Dir of health sciences unit
   iii. Social sciences data specialist (CLIR fellow, 10%) → reports to Dir of social sciences unit
   iv. Spatial/GIS data curator (10%) → reports to Dir of Map Library
   v. Digital humanities data specialist (10%) → Reports to Dir of arts & humanities unit

4. Unit of Library the service belongs: Launched as a “library initiative” reporting directly to AULs. Now sits under Research and Learning AUL (reference and subject liaisons group).

5. Related groups: Research Data Services Team (outreach and training); DSpace Management Team (technical development/roadmapping); UDC Management Team (IR governance)

6. Related non-library groups: Data policy implementation team (based in OVPR), informal CoP for Research Data Management

Cornell University Organizational Structure

Link: https://www.library.cornell.edu/sites/default/files/CULibrary_20160600_OrgChart_Kenney.pdf

1. University Oversight: n/a

2. Library Oversight: Coordinated by CUL’s Scholarly Communication Librarian, who reports to Associate University Librarian (AUL) for Scholarly Resources and Preservation Services, who is part of the Library Executive Group (University Librarian+AULs)

3. Lead for Data Curation Services: Wendy Kozlowski, Data Curation Specialist
   a. Direct Report Curation Staff: 0
   b. Non-Direct Report Curation Staff: 0.2 (collection-specific data curator / subject librarian)

4. Unit of Library the service belongs: Scholarly Communication Services (Unit of Digital Scholarship and Preservation Services)

5. Related library groups: Metadata Services (Unit of Library Technical Services); eCommons Advisory Board

6. Related non-library groups: n/a

Penn State University Organizational Structure

Link: https://libraries.psu.edu/file/psul-organizational-chart

1. University Oversight: n/a

Follow our progress at https://sites.google.com/site/datacurationnetwork
2. **Library Oversight:** Associate Dean for Technology and Digital Strategies

3. **Lead/Director for Data Curation Services:** (still being determined)
   - Direct Report Curation Staff: Patricia Hswe (Head, ScholarSphere User Services), Rob Olendorf (Science Data Librarian), Nathan Piekielek (Geospatial Data Services Librarian)
   - Non-Direct Report Curation Staff: n/a

4. **Unit of Library the service belongs:** Technology and Digital Strategies and Research, Collections, and Scholarly Communications

5. **Related library groups:** ScholarSphere Service Team, Digital Scholarship Working Group, Data Services Action Working Group (strategic implementation group), Research Data Management Team

6. **Related non-library groups:** Services and Solutions in ITS, Institute for CyberScience, Research IT Advisory Council, Data Governance Working Group

---

**University of Illinois at Urbana-Champaign Organizational Structure**

Link: [http://cms.library.illinois.edu/cms/staff/orgchart/](http://cms.library.illinois.edu/cms/staff/orgchart/)

1. **University Oversight:** Office of the Vice Chancellor for Research (for RDS specifically - otherwise it's the Provost)

2. **Library Oversight:** AUL for Research

3. **Lead/Director for Data Curation Services:**
   - Direct Report Curation Staff: RDS Director
   - Non-Direct Report Curation Staff: n/a

4. **Unit of Library the service belongs:** Office of Research

5. **Related library groups:** Repository Team, Digital Preservation, Archives, Scholarly Commons (data discovery and use), Subject Specialists, Cataloging and Metadata

6. **Related non-library groups:** Research IT (e.g. storage and security + campus IT professionals), IRB, Research Ethics, Technology Transfer, Graduate College, supercomputing center (NCSA)

---

**University of Michigan Organizational Structure**

Link: n/a

1. **University Oversight:** Provost (loosely)

2. **Library Oversight:** AUL for Research & AUL for Library Information Technology

3. **Lead/Director for Data Curation Services:**
   - Jennifer Green, Head of Sci/Eng/SAND Libraries and Director of RDS
   - Sebastien Korner, Head, Architecture and Engineering, Library Information Technology

---

Follow our progress at [https://sites.google.com/site/datacurationnetwork](https://sites.google.com/site/datacurationnetwork)

---

29

b. Direct Report Curation Staff: Jake Carlson, Research Data Services Manager;

c. Non-Direct Report Curation Staff: Amy Neeser, Data Curation Librarian (reports to RDS Manager)

d. Unit of Library the service belongs: Research Unit, Library Information Technology (LIT)


5. Related non-library groups: Advanced Research Computing (ARC), IRB, Office of Research and Sponsored Programs (ORSP), MIDAS (Data Science Initiative), Consulting for Statistics, Computing and Analytics Research (CSACR)

Washington University in St. Louis Organizational Structure

Link: http://library.wustl.edu/about/orgchart/

1. University Oversight: Provost

2. Library Oversight: AUL; Aaron Addison, Director of Scholarly Services

3. Lead/Director for Data Curation Services: Cynthia Hudson-Vitale, Data Services Coordinator

   a. Direct Report Curation Staff:

   b. Non-Direct Report Curation Staff: Jennifer Moore, GIS Projects Manager (reports to Director of Scholarly Services), domain point people, SL’s, repository librarian

4. Unit of Library the service belongs: Data & GIS Services

5. Related library groups: Digital Library Services, Publishing Production Services, Archives, Subject Specialists

6. Related non-library groups: research computing infrastructure (HPC, storage, and data transfer), Research Office, Center for Biomedical Informatics

4.2 Position Descriptions and Job Duties

By reviewing position descriptions for research data curation staff and other library staff with data repository and curation responsibilities we aim to better understand the skills needed and the encompassing roles already expected from the staff that our Network model will draw from. Here are some experts from the partner institutions’ position descriptions. See also the recent report⁵ from the Joint Task force on Librarians’ Competencies in Support of E-Research and Scholarly Communication.

Lead/Director for Data Curation Services. Example duties include:

Follow our progress at https://sites.google.com/site/datacurationnetwork

---

Collect, manage, curate, provide access to and assist in the discovery of research data; refer researchers to disciplinary repositories as appropriate.

Provide consultation services for researchers and liaisons to enhance the ability of others to manage, preserve, and conduct new research using digital data collections.

Develop innovative methods for data discovery to enhance the library’s delivery and discovery environment.

Work with faculty, graduate and post-doctoral students, academic and administrative units, and research centers to enable them to better manage, describe, archive, preserve, and make available university research data.

Work with researchers to identify, recruit, ingest and deposit data into repositories, including the library's digital repositories, adhering to local policies and national and international standards and best practices for data management, public access and preservation.

Data Curators. Example duties include

- Serve as primary expert contact for new users inquiring to submit content to the data repository, authorizes new submitters, and answers questions to assist during the upload process for distributed content providers.
- Process submissions for deposit and archive datasets in the digital repository; Research data-related repository activities, workflows, and policies.
- Collect, manage, curate, provide access to and assist in the analysis of research data related to [specific subject discipline]; refer researchers to disciplinary repositories as appropriate.
- Engage with [disciplinary] data producers at the University, as well as at the state and local government levels, to acquire and build a corpus of digital spatial data for access and preservation.
- Perform data curation actions for [disciplinary] data contributed to the data repository or other appropriate repositories.
- Apply data management and data curation techniques for a variety of digital formats (text, code, images, video, etc.).

Library Staff/Subject liaison. Example duties that related to data repository and curation services include:

- Work closely with faculty and students in [subject area] to understand and respond to their changing workflows and patterns of research, research dissemination, and management and preservation of research data.
- Educate and inform faculty, students, and campus administrators about scholarly communication issues such as author’s rights agreements, open access publishing models, and discipline repositories for publications and data.

Follow our progress at https://sites.google.com/site/datacurationnetwork
5.0 Conclusions

The data repository and curation services at the six institutions represent a snapshot in time for library-based activities in this area. By comparing side-by-side services, policies, technology, and staffing levels, our team holds a better understanding of the similarities and contrasting approaches underway so that we may move forward in our goal of developing a shared staffing model for providing data curation services across our institutions. Throughout our discussions it became clear that many of our service goals were well aligned. Even differences in the data repository policies were not described as fundamental divergences, but rather, as policy gaps that should have or will be addressed. It was common to hear a team member say “No, our policy does not say that, but it probably should.” Additionally, in our parallel yet separate implementations of the data repository technology, each using a variation of one of three software approaches, we found much common ground in the workflows and design of how data interacted with the service. These commonalities are thanks in large part to the institutional repository model that each of our systems are either based-on or emulating for the use case of research data. Finally, the staffing models had strong similarities, even though the lack of stable full-time staff was the underlying theme. Yet, as the primary goal of the Data Curation Network is to approach a shared staffing model for data curation services, it is this lack of staffing resources that fuels our project. Therefore the next phase of the project will develop a model for how the Data Curation Network will function, including how data will enter and flow through the service in ways that match our shared expectations, and also how the Network will be administered and sustained.