Introduction:

Rutgers University Libraries (RUL) are poised to offer exceptional research data services to Rutgers’ research community, including data deposit, archiving and preservation through RUcore, the institutional repository for Rutgers University. The Task Force on Research Data Implementation was formed in July 2014 in order to “establish an administrative and evaluation framework for the deposit of research data” in accordance with the Libraries’ and the University’s Strategic Plans. (See Appendix for Task Force Charge.) As of September 2014, the Task Force has completed a review of the administrative structure of data repositories and the evaluation process for technical, legal, and confidential issues. For the review, thirty seven institutions/repositories were evaluated including all of the members of the CIC. Review criteria were based on the ARL SPEC Kit 334 Systems and Procedures Exchange Center: Research Data Management Services and included five areas: Research Data Management Services (RDMS); Data Archiving Services; RDM Service Staffing; Partnerships; and Research Data Policy. The following executive summary provides an overview of the results from these reviews of external repositories.

Executive Summary: Brief Highlights

Our analysis revealed the following:

- There is considerable variance in processes and policies across the 37 institutions surveyed.

- Frequently, members of a community (center, department, etc.) act as the curators of their own data collections and make decisions about the appropriateness of items to be deposited in the repository.

- For many institutions, the principal investigator is responsible for the collection, management, confidentiality, retention, and finally sharing of research data in a repository.

- Self-deposit of data is an accepted approach that can lessen the overhead on library staff.

- Notably, one institution (Michigan) has started charging for extensive metadata, and Stanford recommends that grant proposals include IT costs.

- Most institutions do not specify storage allocations. For those who do, the no-charge allocation for a project varies from 10-100GB.
Unit Hosting the Repository. Just under fifty percent of the reviewed repositories appear to be managed and operated solely by the university library organization. Others included library collaborations with university offices, such as OIT.

Staffing. The Task Force found that the number of research data management service staff members was dependent on each institution’s funding and culture. Staffing numbers ranged in size, from one or two staff to as many as eighteen at one institution. It was not clear if these are FTEs.

Decision Process for Data Deposit. Most of the institutions delegated the decision to deposit data to the researcher. Typically, the researcher determines which data is appropriate to upload, and does this through a self-service interface. Different institutions provide varying levels of checks and assistance.

Funding. Data deposit is just starting out at most of the reviewed institutions, often under the umbrella of the existing institutional repository. These services appear to be funded by the Libraries or IT group hosting the repository, and most services are free to the researcher. A small number of institutions list fees that may be charged to support additional staff time for metadata creation, or for large amounts of storage space.

Project Types. Datasets included in the reviewed repositories are often associated with publications such as articles and dissertations. Few made any apparent distinction between grant-funded research and other scholarly research, and many of the reviewed institutions did not have statements about accepted project types.

Repository Infrastructure and Capacity. Only two of the reviewed repositories provided information directly related to their storage capacities. Some institutions appear to limit data deposits for individual datasets at 10 Gigabytes (GB), but other repositories utilized a tiered support system that specifies levels of support for various types of file formats.

Data Evaluation Criteria and Curation. Over half of the institutions reviewed provided some form of online information about their data evaluation criteria or collection development policies. Determination of suitability of data for deposit varies, with the PI having responsibility in many cases.

Statements Concerning Data Deposit. Six of the institutions reviewed have data-specific repositories, and have data deposit statements pertaining exclusively to data. A few are notable and are mentioned in the report.

Statements Concerning Confidentiality and Privacy. Thirteen of the reviewed institutions had stated data policies, six of which were considered exemplary by the Task Force.
Research Data Security. It appears that in most cases, the various institutions have relied on the IRB for specifying requirements for the protection of sensitive data, and in several cases have reiterated IRB procedures in a repository-related policy.

Training and Consulting Services. Several institutions provide exceptional research data management training and consulting, in the task force’s opinion, as part of ongoing library services. Most of the training and consulting services provided concerned the creation of data management plans, but some included data curation, data archiving and preservation, and data access and sharing.
**Methodology:**

Items one and two of the charge call for a review of the administrative structure of other data repositories; and of the evaluation processes involving technical, legal, and confidential issues surrounding data deposit at other institutions, that might serve as models. To begin this task, a list of institutions with institutional and data repositories, as well as some subject and consortial repositories, was compiled for review. All of the member institutions in the Committee on Institutional Cooperation (CIC) were examined; others were selected for review because they are leading research institutions, or are recognized as having established data repositories. Sources consulted were the Directory of Open Access Repositories (http://www.opendoar.org/index.html), and the Ranking of Web Repositories (http://repositories.webometrics.info/en). Most repositories we reviewed are located in the United States; other locations are in the UK, Canada, and Australia. Thirty-seven institutions/repositories were evaluated. They are:

1. University of California at Berkeley (UCData, Merritt UC system)
2. University of California at San Diego (iDASH, Merritt)
3. University of California at San Francisco (see Merritt)
4. Columbia University (Academic Commons)
5. Cornell University (Geospatial Information Repository, Digital Commons at IRL, and eCommons)
6. Dryad (“Dryad is a nonprofit repository for data underlying the international scientific and medical literature.” It was originally developed to archive evolutionary biology and ecology data.)
7. University of Edinburgh (DataShare)
8. University of Guelph (University of Guelph Research Data Repository, also Agri-environmental Research Data Repository, not reviewed)
9. Harvard University Dataverse (Archive for the Institute for Quantitative Social Science, collaborates with several organizations such as the Odum Institute. See http://thedata.org/book/collaborations for more information.)
10. University of Illinois at Urbana-Champaign (IDEALS)
11. University of Illinois at Chicago (INDIGO)
12. Indiana University (Indiana University Scholarly Data Archive, aka Knowledge Base)
13. Inter-university Consortium for Political and Social Research (ICPSR)
14. University of Iowa (Iowa Research Online)
15. John Hopkins University (JHU Data Archive, part of Dataverse)
16. University of Maryland (DRUM)
17. Massachusetts Institute of Technology (DSpace@MIT, also Geodata Repository, not reviewed)
18. University of Michigan (Deep Blue)
19. Michigan State University (only BRIC Biomedical Research Informatics Core, not reviewed)
20. University of Minnesota (University Digital Conservancy)
21. Monash University (Monash University Research Repository)
22. University of Nebraska-Lincoln (UNL Data Registry)
23. New York University (Spatial Data Repository)
25. Northwestern University (Northwestern Digital Images Repository- images only)
26. Odum Institute (An archive for the Odum Institute of UNC, for social science research. Odum is a member of the Dataverse Network.)
27. Ohio State University (Knowledge Bank)
28. University of Oregon (Scholars’ Bank)
29. University of Pennsylvania (Scholarly Commons- no data yet)
30. Pennsylvania State University (ScholarSphere)
31. University of Pittsburgh (D-Scholarship@Pitt)
32. Princeton University (DataSpace Repository)
33. Purdue University (PURR)
34. Stanford University (Stanford Digital Repository)
35. University of Texas (UT Digital Repository)
36. University of Virginia (Libra)
37. University of Wisconsin-Madison (MINDS@UW)

Of these thirty-seven institutions, there were two members of the CIC without institutional repositories for scholarly works; Michigan State and Northwestern University. Other institutions have an institutional repository but are not accepting data yet. In all, thirty-seven repositories were reviewed, counting the three hosted by Cornell.

The Task Force then developed a set of review criteria to categorize and analyze the Research Data Management (RDM) systems of these institutions. The review criteria were based on the Association of Research Libraries’ Systems and Procedures Exchange Center (ARL SPEC) Kit 334: Research Data Management Services (July 2013), which “surveys ARL member libraries on their activities related to access, management, and archiving of research data at their institutions.” The Task Force decided to base its review criteria on the SPEC Kit due to the comprehensive findings of the survey, and the broad range of ARL member libraries included, which we believe provide a robust foundation for the analysis of RDM systems. These criteria were reviewed based on publicly available information from the repositories’ and libraries’ websites. Information from data managers of some repositories is currently being sought via phone interviews to gather more details, which will be included in a final report.

The detailed review criteria cover thirty-four issues across five areas: Research Data Management Services (RDMS); Data Archiving Services; RDM Service Staffing; Partnerships; and Research Data Policy. The set of review criteria was itemized in a Google Spreadsheet and made available to all Task Force members via Google Drive for synchronous updating. Features in addition to those applicable to the first two items in the Task Force charge were also reviewed, because they were seen as key components of institutional research data services. A brief summary of some of these features is included at the end of this report. (See the Appendix for review criteria details).
Review of Data Repositories: Administrative Structure

Task Force members reported on the following criteria applicable to the administrative structure of the data repositories: the unit which hosts the repository, whether it is the library or another institutional office; the organizational structure of that unit; the staffing and qualifications of the members of that unit; which unit, department, or committee makes decisions regarding the evaluation of data to be deposited, and any funding methods for data deposit that were able to be determined.

Hosting Unit for Data Repository and Organizational Structure:

Among the institutions reviewed, just under 50 percent of the repositories appear to be managed and operated solely by the university library organization. At the same time, there are a number of notable examples of research data management support and repository initiatives that have developed through collaborative relationships, either within a single university, or as collaborations across multiple universities.

There are several notable examples of such partnership arrangements within the CIC. In general, the interaction with faculty and the process of preparing data and metadata for data collections is the purview of the library organizations. Wisconsin (MINDS@UW) and Penn State (ScholarSphere) both operate the infrastructure of their repository environments in cooperation with their respective central IT organizations. The University of Illinois at Urbana-Champaign (IDEALS) organization includes the central IT, and the Provost’s office along with the library. The well-known Purdue repository (PURR) is a joint initiative of the Library, central IT, and the Office of the Vice President for Research.

An interesting example of a partnership that has evolved into a more library-centric model can be found at the University of Virginia, where an original task force comprised of representatives from the OVPR, Library and ITS organizations came together to create a “virtual organization” known as the Scientific Data Consulting Group, that identified priorities for the support of research data management planning and sharing. This group has evolved into the present Data Management Consulting Group, staffed with library professionals who support RDM services and the Libra repository.

Discipline-focused repositories seem to be more likely to be operated without the involvement of the libraries. The UCSD iDASH repository, focused on biomedical data, is an initiative of multiple institutions. The ICPSR social science data repository is housed at the University of Michigan. The organizational structures of these subject-specific repositories are less relevant to the functioning of institutional repositories such as RUcore.

Staffing:

In order to address the unique needs of research in various disciplines, Research Data Management Services (RDMS) frequently include staff members from stakeholder groups across
the institution. Libraries, research offices, and IT departments are the organizational units most often involved in the provision of RDMS. Institutions with larger numbers of staff are assisting researchers in all phases of the data lifecycle, and those with smaller numbers often are only providing basic guidance for research data management plans. Funding and culture appear to play major roles in the staffing of RDMS. It is important to note that, although we were able to determine the number of staff associated with RDMS at many institutions from a preliminary review of their websites, we could not determine how many staff have RDMS as their primary job responsibility and how many contribute only small portions of their job portfolio to RDMS-related tasks. This is an area for further investigation; we plan to follow up our preliminary findings with phone interviews of repository managers or other relevant staff.

Nine of the thirty-seven repositories we studied employed between seven to ten staff members. These are University of California at Berkeley, Dryad, University of Iowa, University of Maryland, Massachusetts Institute of Technology, New York University, Pennsylvania State University, Purdue University, and Syracuse University. The following five institutions have as many as fifteen to eighteen staff members: University of California at San Francisco, Columbia University, Cornell University, Indiana University, and the University of Wisconsin at Madison. The remaining institutions reviewed did not provide specific staff information; however, they do provide a centralized contact for researchers to ask questions or to schedule a consultation.

The staff positions, where information was available, included but were not limited to data management librarians, subject specialists, business managers, and the staff who create and maintain all technical resources for both library and IT services. Their job duties included data storage and data migration tasks, verifying legal information, conducting financial activities, insuring data security, metadata creation and assignment, project management, and data preservation.

**Decision Making Process:**

For most of the institutions reviewed, the decision to deposit data is left to the researcher. Many institutions simply allow the addition of data files to their usual institutional repository environment, and in this case, it is likely that the deposit model remains the same as for articles. The researcher determines which data is appropriate to upload, and does this through a self-service interface. University of Michigan, Ohio State, Penn State, Purdue, University of Wisconsin-Madison, University of Texas, Berkeley, Columbia, Stanford, MIT, University of Maryland, University of Virginia, NYU, Edinburgh, and others follow this model. Maryland states that their staff reviews deposits for basic metadata accuracy, fulfillment of deposit criteria, files being correctly attached, and other routine checks. Presumably, most institutions do similar things behind the scenes. Columbia uses a decentralized repository model, where each academic department manages its own data and is responsible for final decision-making.

Other variations on this theme include UCSF, where only the PI is authorized to deposit data, and Virginia, which explicitly limits data deposit to public access datasets. At Stanford, the authority rests with the researcher, but guidance is provided on what kinds of data to submit, handling sensitive data, and other issues. Most institutions offer some form of assistance with deposit, even if it is just a contact.
Monash, Princeton, Johns Hopkins, and UNC offer mediated ingest. At Monash, the researcher completes an initial form, then the staff contacts them to continue the process. At other institutions, the process begins with the researcher contacting staff. The review process is not specified.

Services that operate beyond the boundaries of an individual institution naturally have a different mission. ICPSR accepts data that is in accord with its mission to preserve significant social science research data. Dryad accepts data that is associated with its partner journal publications. The Odum Institute at UNC also falls in this category. The Qualitative Data Archive at Syracuse mentions reviewing projects for compliance with human subjects and copyright issues. However, their archive is mission-focused and limited to qualitative data, which may be a reason for not placing responsibility with the PI.

**Funding:**

At most institutions, data deposit is only just starting out, often under the umbrella of the existing institutional repository. These services appear to be funded by the Libraries or IT group hosting the repository, and most services are free to the researcher.

At the mission-driven, non-institutional organizations (ICPSR, Dryad, Odum), a more explicit funding arrangement is specified. ICPSR funds its operations through grant-funding for major subject areas, with some chargebacks to users for deposit. Dryad has some tiers of pricing, with the lowest price set at $65 per deposit. Odum also charges fees tailored to each project.

Among university repositories, a few specify limits to the free service offered. The University of Edinburgh allows up to 500GB of space for each researcher for free. Michigan indicates there will be charges for extra metadata work by librarians. Stanford recommends that grant proposals include IT costs, and that it may charge for data over 10GB in the future (not at present). Syracuse also mentions an evolving funding model. Princeton and Berkeley charge strictly by size. Princeton charges 0.006/MB (or $6/GB) as a one-time charge. Berkeley charges $0.14/month for each GB stored.

Johns Hopkins model is unusual, partly because it was designed from the start to become self-funding once initial grants ran out. For small collections, a $1600 charge is standard. For large collections (2TB or more), 2% of the total grant funding is billed to support the data repository.

Finally, Purdue has the most detailed funding model. Central university funding pays for the following free allocations: 10GB for 3 years for trial projects, 1GB for 10 years for a small publication, and 100GB for 10 years for a grant-funded project or publication. Additional space is billed per GB on a yearly, or a 10-year basis. (See [https://purr.purdue.edu/about/pricing](https://purr.purdue.edu/about/pricing) for full details.)
Review of Repositories Evaluation Process: Technical, Legal, and Confidential Issues

Following are the review criteria from the ARL SPEC Kit for Research Data Management Services which were relevant to the evaluation processes for technical, legal, and confidential issues. These are: the project types accepted by the repository in that they may be guided by technical, legal, and/or confidential constraints; the repositories’ infrastructure and capacity; data evaluation criteria for acceptance and curation; statements concerning data deposit and whether deposit is mediated; statements concerning confidentiality and privacy as they relate to the evaluation of data for acceptance, and data security issues.

Project Types:

Twenty-two of the repositories reviewed were institutional repositories that took in data in addition to other scholarly materials, such as published articles, unpublished research, and dissertations and theses. Datasets are sometimes associated with these publications. Other project types include computer code, visualizations, and multimedia works. The institutional repositories we reviewed which accept data include the University of Michigan, Ohio State University, and University of Illinois at Urbana-Champaign. Many institutions state no specific project type restrictions, but generally accept the scholarly output of their researchers. There did not seem to be a distinction made between grant-funded research and other scholarly research. Subject repositories Dryad, ICPSR, Odum, and Qualitative Data Repository accept variously formatted data within their stated missions. Almost one third of the reviewed institutions did not have statements about accepted project types.

Repository Infrastructure and Capacity:

Only two of the reviewed repositories provided information directly related to their storage capacities. Indiana University’s Scholarly Data Archive has a 42 petabyte (PB) capacity, while Stanford University’s Digital Repository currently maintains 147 terabytes (TB) of items in its holdings. Other universities and institutions did not share total capacity information, but concerns about storage space are evident in deposit policies and storage services.

The University of Iowa offers free research data storage up to 3TB. Additional terabytes are available for purchase at $270/TB per year. They also offer a shared/dedicated storage option using high performance computing systems for a fee, and large capacity backup storage with command-line access for another fee scale (see http://its.uiowa.edu/researchstorage).

Other institutions top out data deposits at different storage capacities. A common limit for individual datasets at 10 Gigabytes (GB); one institution limits deposits to 1GB per dataset, and another limits project dataset sizes to 100 Megabytes (MB). Many institutions offer to allow deposit of datasets larger than specified, but for a yearly storage fee.

Formats of accepted data vary. Some institutions accept any and all types of data; others have lists of recommended file formats for deposit. These formats tend to be open-source or widely used and well-documented proprietary software suites (like MS Office applications).
Some of the larger repositories utilized a tiered support system that specifies levels of support for various types of file formats. The University of Michigan, University of Minnesota and the University of Illinois at Urbana-Champaign specify what kind of preservation activities they will supply for each level of service and which file formats are supported in each tier. The highest level of support generally includes maintenance of content, structure and functionality of files and migration to new file formats as technology evolves. This level of service is generally offered for non-proprietary, preservation-level formats with associated metadata such as PDF/A, .TIFF, WAV, .AVI and other well documented, widely used formats. Second level support includes support for preservation and transformation of at-risk files and associated metadata of compressed formats such as .JPEG, .MP3, .MPEG4 and widely used proprietary formats such as MS Office files. Third tier support includes metadata and file storage for proprietary and rare file formats with no guarantee of usability into the future (see Appendix for selected preservation policies).

**Data Evaluation Criteria for Deposit and Curation:**

Twenty-two of the thirty-seven institutions reviewed provided some form of online information about their data evaluation criteria or collection development policies. Four repositories, ICPSR, Qualitative Data Repository (QDR), Dryad, and Geospatial Information Repository, had well defined collection criteria due to their focused missions in specific subject areas.

Fourteen of the reviewed repositories are using DSpace, an open source, out-of-the-box repository platform. DSpace repositories are established on a framework of communities and collections. The communities are user groups organized around a particular discipline or department, research area, or institutional unit. A collection is a part of a community, and contains related materials contributed by members of the community. As such, the members of the community act as the curators of their own collections, and make decisions regarding the appropriateness of items to be deposited.

The review criteria for these communities were typically not available online; however two of these, IDEALS at the University of Illinois at Urbana-Champaign, and the University of Texas Digital Repository, had stated collection development policies. (See [https://services.ideals.illinois.edu/wiki/bin/view/IDEALS/CollectionPolicies](https://services.ideals.illinois.edu/wiki/bin/view/IDEALS/CollectionPolicies), and [http://repositories.lib.utexas.edu/policies_collections](http://repositories.lib.utexas.edu/policies_collections) ) For the University of Texas, this included defining the constituencies of communities that already exist. Both described generally the type of content that would be accepted, such as dissertations and peer reviewed pre-prints. They did not set criteria specifically for data, but did state that they were accepted for deposit. Others did not have policies, but some provided content guidelines regarding types of content, file format, and file size, as noted above. Interestingly, one DSpace repository, the Digital Conservancy at the University of Minnesota, stated that if a suitable subject repository was available for data, it should be the preferred archive.

Three repositories, Merritt at the University of California, San Francisco, Columbia’s Academic Commons, and Iowa Research Online at the University of Iowa, explicitly state that the PI is responsible for evaluating the appropriateness of data, where it was grant-funded; in others this
was assumed to be the case. An additional two repositories, Cornell eCommons and Carolina Digital Repository, stated that librarians would determine suitability of materials for deposit. At Cornell, the Research Data Management group is charged with this responsibility, and at Carolina Digital Repository it is the subject liaisons who act as curators within their disciplines. There were no criteria provided online, but often content guidelines were provided as described above.

**Statements Concerning Data Deposit:**

Among the institutions that we reviewed, six have research data repositories that are independent from separate institutional repositories and thus have research data specific deposit statements. Of these, the data deposit agreements for Harvard University, Princeton University, Purdue University, and the University of Edinburgh’s are simple but comprehensive (see Appendix).

There are usually three components of data deposit statements- those about the data depositor, those about the data repository, and those pertaining to the data itself. Statements concerning the depositor specify that they are the copyright owner or have the permission of the copyright owner to deposit the data, and that the depositor should have fulfilled any obligations to other existing contracts or agreements. The data depositor is required to indemnify the data repository against any legal actions or loss arising from the content of the data or misuse of the service. In all cases, the depositor retains the right to deposit data elsewhere, and has the right to be acknowledged in subsequent use of the deposited data.

Statements about the data repository specify that the data depositor grants permission to distribute, transform for preservation purposes, store, back-up, duplicate, rearrange, reproduce, promote, and describe the data, and to incorporate metadata into public catalogues. The repository is granted non-exclusive rights by the data depositor, and is not liable for loss or damage of the data, or for mistakes, omissions, or legal infringements of the data. The data repository typically retains the right to terminate or take down the data, with metadata left in the repository. Any restrictions to access are also specified.

Statements concerning the data that is deposited into the data repository specify that it should not violate confidentiality, privacy, copyright law, or other laws. It should have IRB approval if human subjects are involved.

**Statements Concerning Confidentiality and Privacy:**

Thirteen of the reviewed institutions had stated data policies, we considered six of these to be thorough and well-written. These were from Ohio State, Michigan State, Northwestern University, University of Wisconsin, Johns Hopkins University, and New York University. One of the most thorough data policies is from the Office of Research at Ohio State, and it covers definitions, policy details, ownership, collection and retention, data security, access, transfer in the event the primary investigator leaves, expert control, author disputes, and data access disputes. It is noteworthy that the principal investigator is responsible for the collection, management, and retention of research data for the periods required by the policy; to control access to research data; and to select the vehicle for publication or presentation of the data.
The commonalities in the best policies seem to be the following: 1) the university owns the data, 2) the principal investigator is responsible for making sure it is preserved, and 3) and protocols exist in the event the PI leaves the institution.

**Research Data Security:**

It appears that in most cases, the various institutions have relied on the IRB for protection of sensitive data, and in several cases have reiterated IRB procedures in a repository-related policy. For example, the University of Wisconsin has a risk assessment matrix that provides a practical approach for assessing risk related to human subject research. The University of Indiana requires that all researchers encrypt sensitive data before depositing it into the repository, while the University of Iowa offers recommendations for encryption of sensitive data. Other universities in the survey require that any sensitive information be stripped from data before the deposit. In cases where data is meant to be destroyed, all repositories surveyed that mention the destruction of data urge researchers to follow the protocols and requirements of granting agencies when destroying data.

Several institutions highlighted the importance of off-site backup, and ICPSR indicates that off-site backups should be encrypted. Excerpts are included as examples below:

University of Illinois at Chicago states the following relating to reliance upon IRB: “In its review of research, the UIC Institutional Review Board (IRB) considers whether adequate provisions exist for the security of research data, whether in paper or electronic form, throughout the research, including data analysis and retention. Investigators are responsible for ensuring that adequate controls, as described in the research protocol…” (See [http://research.uic.edu/sites/default/files/0927.pdf](http://research.uic.edu/sites/default/files/0927.pdf)).

Although access to data may be controlled by the repository, Penn State, as do other institutions, places the responsibility for security of data and confidentiality with the primary investigator: “Typically, when research is funded by federal or nonprofit granting agencies, the data are owned by the institution receiving the grant. The primary researcher or scholar receiving the grant has the responsibility for storage and maintenance of the data, including the protection of confidential or sensitive information… Scholars and researchers have a moral and professional responsibility to ensure that confidential or sensitive data is stored and released in a way that protects research participants.” ([http://www.research.psu.edu/training/sari/teaching-support/data-management](http://www.research.psu.edu/training/sari/teaching-support/data-management))

**Additional Topics- Training and Consulting Services:**

Our review of institutions included various additional aspects of research data management services that were not part of the above criteria. Several institutions appear to provide a very comprehensive level of service, ranging from data management training and consulting through data archiving and access, and we provide the following summary.
We found that with the exception of Dryad, almost all of the institutions reviewed provide training in the form of resources, workshops, or other types of programs. Eighteen institutions provided training through the library; an additional thirteen offered training through a collaboration between the library and offices such as the Office of the Vice President for Research or the Office of Information Technology. Other units involved in training included a Research Computing Center and Information Services.

Most of the training and consulting services provided concerned the creation of Data Management Plans, but some included data curation, data archiving and preservation, and data access and sharing. Some workshops appeared to be ongoing activities; others were scheduled by request. Staff providing the training were usually data services librarians or members of a research data management team. Several libraries’ webpages provided recorded workshops and presentations, as well as links to external sources of information. Most of the institutions providing training also provided consulting services. A few of the institutions with comprehensive training and consulting services are: Cornell, Johns Hopkins University, Michigan State University, Monash University, Ohio State University, Penn State, Purdue, Stanford University, University of California system, University of Edinburgh, University of Illinois at Urbana-Champaign, University of Minnesota, University of Virginia, and University of Wisconsin-Madison. For information about individual institutions’ training and consulting resources, please see the table provided in the Appendix.

**Summary and Conclusions**

Almost all of the institutions reviewed, whether they were accepting data or not, provided research data management training and consulting, typically in data management plan preparations. The institutions reviewed were at varying stages with regard to acceptance of research data. A few were not accepting data or had a limited number of datasets, but many had respectable quantities of data, and some were well established data exclusive repositories. Eighteen institutions had repositories operated by libraries, and many worked in collaboration with outside units and offices such as the Office of the Vice President for Research, and the Office of Information Technology. Staff responsible for the repositories’ activities varied by the size of the institution, with the largest data management teams at two institutions consisting of fifteen to eighteen members. Collaborative efforts with units outside of the libraries and a team approach in general seem to make the most sense for larger institutions.

Most repositories place the responsibility of the evaluation of data on the principal investigator, or in the case of DSpace repositories, on the Communities’ administrators, who represent department researchers. Only two repositories placed curation responsibilities exclusively with librarians, although others used teams including librarians. About half of the reviewed repositories allowed self-deposit of data or self-deposit with review, similar to the process used by RUcore in acceptance of research documents, albeit with additional forms and guidance. Data deposit agreements were common, and most shared a similar format. Depositors typically needed to agree that they were legally allowed to deposit the data for public access; that the data does not contain any personal or sensitive information; that the depositor holds the institution harmless from any liability incurred as a result of the deposit and public access of the data; and
that the repository may enact certain described operations in order to provide for data discovery, maintenance, and preservation.

Privacy and security issues were typically addressed by agreements wherein the depositor stated that the data was free of any confidential or sensitive data; and by stripping of identifying information, or in some cases by encryption. Responsibility for the protection of confidential data was placed with the principal investigator or the researcher depositor. Based on our findings, we believe that researcher responsibility for legal issues, and self-deposit make the most sense from a liability and efficiency standpoint, with case by case exceptions.

Information about repository storage capacity was limited. Restrictions to file sizes and file types were more prevalent, with offerings ranging from 10 – 100 GB free of charge; and acceptance of most standard file types associated with open source and widely used proprietary software was common. It seems as if funding models for storage and preservation of research data have not been established for most repositories, although a few did provide information about costs of services. Ongoing funding for data preservation and repository growth is an important aspect of data management services that will need to be addressed in order to create sustainable systems, particularly where staff time is a factor.

Services in addition to the technical, legal, and administrative aspects of data ingest, such as data management training and consulting, are flourishing in institutions even without data capabilities in their repositories. This is an area that can be leveraged for enhanced library visibility in anticipation of the acceptance of data into RUcore.

Areas for further study include a more in-depth review of staffing, specifically with regard to FTEs. Funding is also a concern, although fees for services were not prevalent among the institutions we reviewed. This may be due to the relative newness of data acceptance for many of these. An in-depth case study of a few well-established repositories with regard to these and other issues, such as confidentiality of sensitive data, would be beneficial.

References

Appendices

A. Task Force Charge

B. Selected Preservation Policies

C. Selected Data Deposit Agreements

D. Table of Research Data Training and Consulting Pages
Appendix A

Rutgers Libraries Task Force on Research Data Implementation

The task force on Research Data Implementation is charged with establishing an administrative and evaluation framework for the deposit of research data in RUcore. This implementation process will inform the development of a university data policy by the office of General Counsel working with our Libraries Copyright and Licensing Librarian.

The task force should involve other individuals as necessary to do its work, and engage at the outset with the office of the Vice President for Research and the Office of Research and Sponsored Programs to ensure that the implementation addresses issues of importance to the research faculty and appropriate administrative offices. The task force should also liaise with the Committee on Scholarly Communication through its chair, Laura Mullen.

Janice Pilch, Copyright and Licensing Librarian, will work separately on drafting a data policy. When your draft implementation plan is ready, Janice can review it from the copyright perspective. We believe this two part process will work effectively.

This is a Cabinet Task Force under the joint leadership of Grace Agnew and Melissa Just who will oversee and guide its work on behalf of Cabinet. We expect the plan to be completed no later than December 2014, and Cabinet would expect a progress report mid-way through the process.

The charge to the Task Force is to:

1. Review the administrative structure of other data repositories that might serve as models.
2. Review the evaluation process for technical, legal, and confidential issues involving data deposit at other institutions that might serve as models.
3. Consult with appropriate major stakeholders to ensure that RUL workflows and practices facilitate and do not conflict with policies and practices of those departments, especially the office of the Vice President for Research, and Research and Sponsored Programs.
4. Establish principles for prioritization of data deposit projects based on RUL strategic priorities. This should include a definition of various types of potential projects to ensure that we have the resources both to host and to sustain projects, i.e. federal grants, non-grant funded research, etc.
5. Develop a framework for evaluation for data deposit in RUcore that includes a questionnaire or series of questionnaires to be used for each data deposit, covering technical, legal, and confidential criteria (similar to the Digital Projects Evaluation Process approved by Cabinet in March 2013).
6. Develop a corresponding guide on evaluation criteria to provide clarity to subject librarians. (similar to the Deed of Gift Explanation in the RUL Deed of Gift).
7. Recommend assignments for functional responsibility in the area of data deposit.
8. Chart a workflow for the data deposit evaluation process.
9. Determine the major stakeholders at Rutgers who need to be familiar with the RUL data deposit process.
10. Consult with CISC to assess RUcore hardware and software infrastructure to support immediate, three year and five year needs.
Task Force Members:

Laura Palumbo, Chair
Ron Jantz
Yu-Hung Lin
Aletia Morgan
Minglu Wang
Krista White
Ryan Womack
Yingting Zhang
Yini Zhu

6/26/14
Appendix B

Selected Preservation Policies

University of Michigan: http://deepblue.lib.umich.edu/static/about/deepbluepreservation.html

University of Minnesota: http://conservancy.umn.edu/udc/pol-preservation.jsp

University of Illinois at Urbana-Champaign: https://services.ideals.illinois.edu/wiki/bin/view/IDEALS/PreservationSupportPolicy
Appendix C

Selected Data Deposit Agreements


Princeton University: http://dataspace.princeton.edu/jspui/about/DataSpacePnG.pdf

Purdue University: https://purr.purdue.edu/legal/termsofdeposit

University of Edinburgh: http://www.ed.ac.uk/schools-departments/information-services/research-support/data-library/data-repository/depositor-agreement
Appendix D

Research Data Management Training and Consulting

<table>
<thead>
<tr>
<th>1. University of Chicago (by RCC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tutorial on working with data (plans, storage, transferring, visualization, etc) - <a href="http://docs.rcc.uchicago.edu/tutorials/workingwithdata.html">http://docs.rcc.uchicago.edu/tutorials/workingwithdata.html</a> by RCC (research computing center).</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>2. University of Michigan (by library and ICPSR)</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>3. Ohio State University (by library)</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>4. University of California San Diego (by library)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training &amp; workshops on research data management offered by UCSD library <a href="http://libraries.ucsd.edu/services/data-curation/training-schedule/index.html">http://libraries.ucsd.edu/services/data-curation/training-schedule/index.html</a>.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>5. University of Illinois at Urbana-Champaign / at Chicago (by Library)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes, Workshops offered by library’s Scholarly Commons: <a href="http://www.library.illinois.edu/sc/services/">http://www.library.illinois.edu/sc/services/</a></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>6. Indiana University</th>
</tr>
</thead>
<tbody>
<tr>
<td>University site search function not working as of Sept 2, 2014</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>7. University of Iowa (by library)</th>
</tr>
</thead>
<tbody>
<tr>
<td>University Library's Digital Research &amp; Publishing Unit offers DMP consulting: <a href="http://www.lib.uiowa.edu/drp/services/">http://www.lib.uiowa.edu/drp/services/</a>.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>8. University of Maryland (by library)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Writing guides for data management plans for NIH and NSF offered by UMD libraries - <a href="http://www.lib.umd.edu/data/dmp">http://www.lib.umd.edu/data/dmp</a>.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>9. Michigan State University (by library)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Research data management guidance and training offered by MSU libraries with both regular scheduled workshops or upon request. <a href="https://www.lib.msu.edu/rdmg/training/">https://www.lib.msu.edu/rdmg/training/</a>. Consulting services on RDM at <a href="https://www.lib.msu.edu/rdmg/advising/">https://www.lib.msu.edu/rdmg/advising/</a></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>10. University of Minnesota (by library)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recorded workshops and slides on RDM, Data Management Course, and other university units' such as OIT and OVPR’s relevant trainings: <a href="https://www.lib.umn.edu/datamanagement/workshops">https://www.lib.umn.edu/datamanagement/workshops</a>. Other guidance information is at <a href="https://www.lib.umn.edu/datamanagement">https://www.lib.umn.edu/datamanagement</a></td>
</tr>
</tbody>
</table>
|   | **University of Nebraska-Lincoln (by library)**  
Workshops on data management plans offered by the Libraries - [http://libraries.unl.edu/data-management](http://libraries.unl.edu/data-management). |
|---|---|
| 12 | **Northwestern University (by library – 1 dmp information webpage)**  
Services on dmp offered by library [http://www.library.northwestern.edu/dmp](http://www.library.northwestern.edu/dmp)  
(Data retention and ownership and access policy: [http://www.research.northwestern.edu/policies/documents/research_data.pdf](http://www.research.northwestern.edu/policies/documents/research_data.pdf)) |
| 13 | **Penn State (by library)**  
PSU library’s RDMST’s DMP tutorial site: [https://www.e-education.psu.edu/dmpt/node/508](https://www.e-education.psu.edu/dmpt/node/508)  
Consultation tips - [http://www.libraries.psu.edu/psul/groups/rdmst.html#consultations](http://www.libraries.psu.edu/psul/groups/rdmst.html#consultations) |
| 14 | **Purdue (by libraries & OVPR)**  
Consultation on data management and data research [https://www.lib.purdue.edu/research](https://www.lib.purdue.edu/research) |
| 15 | **University of Wisconsin-Madison (by collaborative Research Data Services group)**  
Research Data Services is collaboration between UW-Madison Libraries, DoIT, the CIO office, the Graduate School, and the School of Library and Information Studies to assist researchers with data curation needs. [http://researchdata.wisc.edu/learn-more-about-data-management/](http://researchdata.wisc.edu/learn-more-about-data-management/).  
Services include help, training, consultation, and referrals - [http://researchdata.wisc.edu/help/our-services/](http://researchdata.wisc.edu/help/our-services/) |
| 16 | **UC Berkeley (by science libraries and D-Lab)**  
The science libraries offer services on data management and sharing - [http://www.lib.berkeley.edu/sciences/data/guide](http://www.lib.berkeley.edu/sciences/data/guide) and other related data services. [http://www.lib.berkeley.edu/sciences/data/](http://www.lib.berkeley.edu/sciences/data/).  
D-Lab offers consulting and advising services to faculty, staff and students in data intensive social sciences - [http://dlab.berkeley.edu/services](http://dlab.berkeley.edu/services) |
| 17 | **UC San Francisco (by library and CDL)**  
Provide links to tools and information for scientists to properly organize, manage, and document their datasets - [http://www.library.ucsf.edu/data-management](http://www.library.ucsf.edu/data-management). UC CDL Curation Center created the DMP Tool: [https://dmp.cdlib.org/](https://dmp.cdlib.org/) |
| 18 | **Columbia (by Academic Commons and libraries)**  
Academic Commons – How to videos: [http://academiccommons.columbia.edu/about/](http://academiccommons.columbia.edu/about/).  
|   | **Cornell (by RDMSG, collaborative campus wide organization)**  
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>eCommons encourage use of DMPTool. Consulting services cover training. The Research Data Management Service Group is a collaborative, campus-wide organization for DMS at <a href="http://data.research.cornell.edu/content/about">http://data.research.cornell.edu/content/about</a>. It offers a wide variety of support services spread across departments and centers. Seem to be discipline specific.</td>
</tr>
</tbody>
</table>
|   | **Dryad (a non-profit membership organization)**  
|   | No training or consulting services are offered. But FAQs on submitting data and using data are available. |
|   | **University of Edinburgh (by Information Services)**  
|   | IS provides guidance [http://www.ed.ac.uk/schools-departments/information-services/research-support/data-library/research-data-mgmt](http://www.ed.ac.uk/schools-departments/information-services/research-support/data-library/research-data-mgmt), workshops and courses including a non-credit, self-paced course. Created online course MANTRA for researchers to plan to manage digital data. Consultancy services are offered. |
|   | **Harvard (by Harvard Library and the Research Data Collaborative)**  
|   | Information on training and consulting with DMP is not found on the data management page - [http://isites.harvard.edu/icb/icb.do?keyword=k78759&tabgroupid=icb.tabgroup123330](http://isites.harvard.edu/icb/icb.do?keyword=k78759&tabgroupid=icb.tabgroup123330). However, info on how to write DMP is available at [http://isites.harvard.edu/icb/icb.do?keyword=k78759&pageid=icb.page407320](http://isites.harvard.edu/icb/icb.do?keyword=k78759&pageid=icb.page407320). |
|   | **ICPSR (Inter-university Consortium for Practical and Social Research)**  
|   | Provides training in data management as part of its Summer Program. Summer courses [https://www.icpsr.umich.edu/icpsrweb/sumprog/courses](https://www.icpsr.umich.edu/icpsrweb/sumprog/courses). Data management and curation information at [https://www.icpsr.umich.edu/icpsrweb/content/datamanagement/index.html](https://www.icpsr.umich.edu/icpsrweb/content/datamanagement/index.html). In addition, ICPSR hosted the Digital Preservation Management Workshops and Tutorial from 2007-2012. Provides consulting services related to disclosure risk assessment and mitigation. |
|   | **John Hopkins University (by libraries and JHU Data Management Services)**  
|   | **MIT (by library)**  
|   | Individual consultation service is also offered - [http://libraries.mit.edu/data-management/services/](http://libraries.mit.edu/data-management/services/). |
|   | **Monash University (by library)**  
<p>|   | Library offers several sessions for researchers and staff on research data skills <a href="http://monash.edu/library/researchdata/skills/">http://monash.edu/library/researchdata/skills/</a>. Advice and planning services offered by library and other relevant units such as Monash eResearch Centre, |</p>
<table>
<thead>
<tr>
<th>Institution</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>University of North Carolina (by library and Odum Institute)</td>
<td>Odum provides consulting, courses, certificate program, etc. Short video classes, video presentations from classes held at Odum, and video presentations regarding information on their services and tips on using specialized software are available. UNC libraries created a Research Data Toolkit - <a href="http://guides.lib.unc.edu/researchdatatoolkit">http://guides.lib.unc.edu/researchdatatoolkit</a></td>
</tr>
<tr>
<td>University of Oregon (by library and IT)</td>
<td>Library held training, but the page has not been updated since 2012. <a href="http://library.uoregon.edu/datamanagement/training.html">http://library.uoregon.edu/datamanagement/training.html</a>. Library offers guidance, both on web and in person, on DMP development. IT also offers assistance on NSF research data management plan - <a href="https://it.uoregon.edu/nsf">https://it.uoregon.edu/nsf</a>.</td>
</tr>
<tr>
<td>Syracuse University (by library)</td>
<td>Video tutorials for research data are posted in the Research Data Services libguide <a href="http://researchguides.library.syr.edu/datatraining">http://researchguides.library.syr.edu/datatraining</a>. Some external sources are also listed.</td>
</tr>
<tr>
<td>University of Texas (by Data Management, collaboration between UT libraries, TACC, ITS, OSP)</td>
<td>Data Management at <a href="http://www.lib.utexas.edu/datamanagement">http://www.lib.utexas.edu/datamanagement</a>, listing resources, news, DMP templates, instructions, etc. Consulting through partners, statistics department: <a href="http://www.lib.utexas.edu/datamanagement/resources">http://www.lib.utexas.edu/datamanagement/resources</a></td>
</tr>
<tr>
<td>University of Virginia (by UVA libraries)</td>
<td>Data Management Consulting Group (DMConsult) <a href="http://dmconsult.library.virginia.edu">http://dmconsult.library.virginia.edu</a> of UVA Libraries provide training sessions, data management workshop series - <a href="http://dmconsult.library.virginia.edu/training-sessions/">http://dmconsult.library.virginia.edu/training-sessions/</a>.</td>
</tr>
<tr>
<td>Princeton (by library and OIT)</td>
<td>Library and OIT provide resources to help researchers plan, organize, and preserve research data. Info on DMP training was not found. Engineering Library blog addressed data sharing - <a href="http://blogs.princeton.edu/engl/2013/10/sharing_research_data_when_in_what_form_with_whom.html">http://blogs.princeton.edu/engl/2013/10/sharing_research_data_when_in_what_form_with_whom.html</a>. ORPA has some info on data management plans.</td>
</tr>
<tr>
<td>NYU (NYU health sciences library’s DKI – division of knowledge Informatics)(by ITS – DM policy)</td>
<td>Health science library provides consultation for data management / sharing regulations and best practices - <a href="http://hsl.med.nyu.edu/content/consultation-requests">http://hsl.med.nyu.edu/content/consultation-requests</a>. DKI’s Bear &amp; Panda Data Management Videos: Data Sharing and Management SNAFU in 3 Short Acts - <a href="http://www.youtube.com/watch?v=3r2zP3sAtr-4">http://www.youtube.com/watch?v=3r2zP3sAtr-4</a>. University data management</td>
</tr>
<tr>
<td>policy</td>
<td><a href="http://www.nyu.edu/its/policies/datamgmt.html">http://www.nyu.edu/its/policies/datamgmt.html</a></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
</tr>
</tbody>
</table>