Digital Research Data Curation:
Overview of Issues, Current Activities, and Opportunities for the Cornell University Library

A report of the CUL Data Working Group

May 2008

The Cornell University Library (CUL) Data Working Group (DaWG)

Formed in December of 2006, the Cornell University Library's Data Working Group's purpose is to exchange information about CUL activities related to data curation, to review and exchange information about developments and activities in data curation in general, and to consider and recommend strategic opportunities for CUL to engage in the area of data curation. The Data Working Group has discussed publications and activities related to data curation, and has hosted presentations by (or discussions with) DaWG members and Cornell faculty and staff. This white paper presents an overview of the current landscape and issues surrounding data curation, and includes recommendations for CUL in this area.

Members of the Data Working Group: Paul Albert, Kristine Alpi, Pam Baxter, Eli Brown, Kathy Chiang, Jon Corson-Rikert, Peter Hirtle, Keith Jenkins, Brian Lowe, Janet McCue (LMT liaison), David Ruddy, John Saylor (co-chair), Rick Silterra, Leah Solla, Gail Steinhart (co-chair), Zoe Stewart-Marshall, Elaine L. Westbrooks.

Table of Contents

Acknowledgments

A great deal of research went into this collective effort. We've made extensive use of published literature, project web sites, and personal contacts to produce this report. We have made every effort to represent the work of others accurately, and the responsibility for any errors rests with the Data Working Group.

The DaWG would like to acknowledge the following people for contributing to or commenting on portions of this report: Nan Hyland, Steve Kelling, Joy Paulson, Paul Redfern, and Oya Rieger.

Executive Summary

The purpose of the Cornell University Library's (CUL) Data Working Group (DaWG) is to exchange information about CUL activities related to data curation, to review and exchange information about developments and activities in data curation in general, and to consider and recommend strategic opportunities for CUL to engage in the area of data curation. This white paper aims to fulfill this last element of the DaWG's charge.

Advances in computational capacity and tools, coupled with the accelerating collection and accumulation of data in many disciplines, are giving rise to new modes of conducting research. Infrastructure to promote and support the curation of digital research data is not yet fully-developed in all research disciplines, scales, and contexts. Organizations of all kinds are examining and staking out their potential roles in the areas of cyberinfrastructure development, data-driven scholarship, and data curation.

There are three primary (and related) motivations for developing a robust data curation infrastructure: enabling new discoveries by exposing data for use in data-driven research, ensuring access to and preservation of scholarly output, and meeting existing or forthcoming requirements of funding agencies or institutions regarding data management, retention, and access. Libraries have demonstrated expertise in several areas that could be productively applied to the practice of data curation, and in some cases, cyberinfrastructure development.

This report of the Data Working Group offers five broad recommendations for ways in which the Cornell University Library might engage in data curation and related activities.

1. Seek out and cultivate partnerships with other organizations.

The challenges of digital data curation are significant, and in many cases, are best handled in cooperation with other organizations. The problem encompasses storage and network capacity, integration with cyberinfrastructure development, best practices related to data archival and metadata, user support, and more. Effective collaboration within Cornell and across institutions is essential to meeting the challenges of data curation.

1a. Identify opportunities to collaborate with other Cornell data curation efforts. Collaborators might include the Center for Advanced Computing (CAC), Cornell Information Technologies (CIT), Computer and Information Science (CIS), the Cornell Institute for Social and Economic Research (CISER), and other departments, research groups, or organizations. Also identify promising opportunities for involvement with domain-specific data curation initiatives.

1b. Participate actively in forthcoming initiatives, such as those described in the Association of Research Libraries Joint Task Force report on library support for e-science and libraries (2007), and other initiatives as they arise.

2. Provide services to Cornell researchers in several areas.

This group of recommendations is focused on providing services and information (as opposed to dedicated infrastructure) to support Cornell researchers, particularly with respect to existing or forthcoming data management, sharing or archival requirements. Whenever possible, information – both educational resources and information on CUL services - should be made accessible on the Internet.

2a. Assist with the development of data management plans in grant proposals. This affords CUL opportunities to engage researchers at the very start of a research project, enabling us to better forecast and prepare to meet data curation needs.
2b. Collect and provide information on best practices for data management, archival, and preservation. This area may lend itself to collaboration with other individuals or professional organizations. Offering such guidance is practical to the extent that information is generalizable across disciplines. In cases where significant subject area expertise is required, it may not be possible for the library to provide guidance in all subject areas.
2c. Educate researchers about intellectual property issues as they pertain to research data.
2d. Refer researchers to appropriate resources dealing with the protection of confidential and private information. Institutional Review Board and HIPAA requirements prohibit the release of personally identifiable information, and researchers who collect such information in the course of their work may be faced with reconciling requirements to share data with the protection of confidential information. CUL staff should be cognizant of the issues involved.
2e. Participate in the formulation of institutional policies on data retention. CUL should also provide support for researchers in meeting the requirements of such policies.

3. Assess local needs and develop local infrastructure and related policies.
These recommendations suggest that CUL become more than an advisory partner in data curation by providing infrastructure in contexts that make sense within the university, collaborating with other units as much as possible.

3a. Examine the feasibility and desirability of creating a local repository infrastructure for depositing digital research data. A digital data audit may be appropriate. Collaboration with the DISCOVER RSG group's development of a cyberinfrastructure roadmap is also advisable.
3b. Further refine the guidelines presented in this report for the appraisal and selection of data to be curated locally.
3c. Investigate roles for CUL with respect to the curation of data related to publications.
3d. Continue to participate actively in the curation of "small science" data.

4. Cultivate a workforce capable of addressing the new challenges posed by data curation and cyberinfrastructure development.
Expanding current data curation and cyberinfrastructure activities and embarking on new ones will require investment in professional development for CUL staff, and possibly the creation of entirely new positions.

4a. Identify new skills that will be needed among CUL librarians to support activities in the area of data curation and cyberinfrastructure, as well as the positions where those skills would be needed and applied.
4b. Identify what new positions might be needed to support new CUL activities in the area of data curation and cyberinfrastructure.
4c. Make professional development activities available to CUL librarians – including reorganizing the Data Working Group as suggested below.

5. ***Form a Data Curation Executive Group and reorganize the Data Working Group.***
One of the core purposes of the Data Working Group, articulated in its charge, was to educate itself in the areas of data curation and cyberinfrastructure, and to develop a set of recommendations for CUL. Having largely completed these tasks, two needs emerge:

advancing the recommendations made in this report, and continuing education for CUL staff. We recommend the following:

5a. Form a Data Curation Executive Group (DataExec) to advance the recommendations outlined in this report. Data curation and e-scholarship should also become part of the portfolio of an AUL-level position, and that individual should sit on the DataExec.
5b. Reconstitute the DaWG as a smaller steering committee, operating much as the Metadata Working Group steering committee does.
5c. Publish this report in eCommons, and consider publishing an abbreviated version in D-Lib magazine or elsewhere.

Curation of digital research data is, with a few exceptions, a new area of activity for CUL, as are activities related to the development of cyberinfrastructure. As should be evident from the introductory sections of this report, the pace of development is brisk, and organizations are working to identify and claim roles for themselves. Implementing some of the recommendations of the DaWG would require significant resources. Term funding from research grants is not likely to be sufficient to provide for long-term stewardship of research data; institutional commitments are needed. Nonetheless, our sense is that the time is right to articulate roles for CUL in this arena, and that we have a solid foundation of both new and established activities to demonstrate our competence in this area.

I. Introduction

Advances in computational capacity and tools, coupled with the accelerating collection and accumulation of data in many disciplines, are giving rise to new modes of conducting research. In addition, funding agencies increasingly recognize the need to document and archive the data that result from funded research, and in some cases, mandate that researchers prepare and implement data management plans to meet this need. Infrastructure to promote and support the curation of digital research data, however, is not yet fully-developed in all research disciplines, scales, and contexts, and organizations of all kinds are examining and staking out their potential roles in the areas of cyberinfrastructure development, data-driven scholarship, and data curation. While cyberinfrastructure development may be more advanced in the sciences than in the humanities, it is an emerging area of activity across a broad array of disciplines (ACLS Commission on Cyberinfrastructure for the Humanities and Social Sciences 2006; Borgman 2008). This report focuses on issues related to digital data curation, primarily in the sciences, but it also inevitably touches on cyberinfrastructure and e-scholarship.

There are three primary (and related) motivations for developing a robust data curation infrastructure: enabling new discoveries by exposing data for use in data-driven research, ensuring access to and preservation of scholarly output, and meeting existing or forthcoming requirements of funding agencies or institutions regarding data management, retention, and access. The publication in 2003 of the report "Revolutionizing Science and Engineering Through Cyberinfrastructure" (widely known as the "Atkins report"), by the National Science Foundation's Blue-Ribbon Advisory Panel on Cyberinfrastructure, and the subsequent creation of the Office of Cyberinfrastructure at NSF, were critical events that launched the current wave of funding for research and development in the U.S. for cyberinfrastructure (Atkins et al. 2003; Gold 2007a). A 2005 report from the National Science Board urged the National Science Foundation (NSF) to develop a strategy for protecting its research investments by supporting digital data collections, emphasizing the critical importance of digital data in supporting cyberinfrastructure-enabled research (National Science Board 2005). A 2006 report from a joint NSF-ARL workshop on data stewardship examined the roles of research libraries and other potential partners in this arena, and urged the NSF to adopt a requirement for data management plans as a part of all grant proposals (ARL Workshop on New Collaborative Relationships 2006). Many other papers and reports have been published recently on these topics (see Gold 2007a for a concise review).

The Role of the Academic Research Library

There are conflicting views over potential roles for academic research libraries in the realm of digital data curation. One view holds that libraries, given their preservation mission, are “natural heirs” for data (e.g. Lord and Macdonald 2003). One could make the “natural heir” case based on the library’s role in the academy as a trusted and neutral organization, providing access to and serving as a reliable steward of information. A contrasting view holds that libraries lack the necessary subject expertise, and information technology skills and resources, and that such activities belong more appropriately in discipline-based communities (Lord and Macdonald 2003; Lyon 2007). The ARL workshop report describes the data problem as a distributed one, requiring an “ecology of institutional arrangements among individuals and organizations,” and recognizes the need for cooperation and collaboration among libraries and the domain sciences, advocating for a “tight partnership” between libraries and disciplines (ARL Workshop on New Collaborative Relationships 2006). Skeptics might argue that the ability of academic libraries to participate effectively in this arena is potentially complicated by the fact that their constituent communities are local (within the institution in which the library is embedded), while research teams and activities tend to be trans-institutional (Joint Task Force on Library Support for E-Science 2007; Messerschmitt 2003), and that existing commitments may hamper libraries' ability to adapt to this rapidly evolving landscape (Joint Task Force on Library Support for E-Science 2007). Allocating funding away from local services and infrastructure to support shared services may present a significant challenge (Arms and Larsen 2007), but Messerschmitt (2003) argues that libraries' traditional, local focus need not preclude them from contributing and participating in this arena.

In fact, research libraries are becoming more engaged with research processes, particularly scholarly communication, by hosting e- and pre-print repositories and developing new publishing modes (Joint Task Force on Library Support for E-Science 2007), activities which transcend institutional boundaries. These activities may represent one logical extension point from which libraries can engage in digital research data curation, and libraries have much to offer. The NSF has made it plain in its DataNet solicitation that it would like to see the combined expertise of librarians, archivists, computer and information scientists, and domain scientists brought to bear on data curation issues (National Science Foundation Office of Cyberinfrastructure 2007). Program officers have also expressed an expectation that institutions will play a role in the curation of their own data (Greer 2007). The NSF views the movement of data between types of collections (research, community, and reference; see Appendix A for definitions) to be a normal occurrence, and research institutions might reasonably be expected to be responsible custodians of research-level data collections (National Science Board 2005).

Libraries have demonstrated expertise in several areas that could be productively applied to the practice of data curation, and in some cases, cyberinfrastructure development. Some of these areas include:

- *Principles and policies related to scholarly communication.* Libraries already operate institutional repositories, and in some cases, domain repositories, and have been active in debates surrounding changing practices in scholarly communication (Joint Task Force on Library Support for E-Science 2007). Intellectual property issues are important and sometimes contentious; for libraries this is already familiar terrain and they are well-positioned to help formulate policy as well as assist researchers in understanding and applying it (Messerschmitt 2003)
- *Description and discovery.* Metadata is essential for discovery and reuse of digital research data, and libraries are expert in the creation and application of metadata standards. Participating in the development of standards in support of digital research data, as well as providing services to data owners, would be natural extensions of current library activities (e.g. Gold 2007b; Joint Task Force on Library Support for E-Science 2007; Messerschmitt 2003).
- *Interoperability.* Libraries have experience in developing and implementing tools for facilitating interoperable access to information; examples include federated searching, metadata standards, and projects such as MIT's SIMILE (Joint Task Force on Library Support for E-Science 2007).
- *Digital preservation.* Technology will continue to change at a rate that presents a constant challenge to digital preservation capabilities (e.g. Thibodeau 2007); this is an active area of research and engagement for many libraries.

- *Selection and appraisal*. While criteria for selection and appraisal of digital research data are not yet fully-developed, archivists have significant expertise that can be applied in this area (e.g. Harvey 2007).
- *User support for information retrieval, computer and Internet applications*. A commonly stated goal in cyberinfrastructure development is an unmediated experience for users, yet this is seldom the result, and libraries are accustomed to providing this type of user support (Messerschmitt 2003). Many academic libraries already have extensive experience with the provision of services related to geospatial, social science, and bioinformatics data (Gold 2007b). Such support might also be extended to data owners, to facilitate the deposit of data into trustworthy data archives. Various studies have shown that local support for participants is important in ensuring the success of data sharing efforts (Glover et al. 2006; Karasti et al. 2006; Lord and Macdonald 2003).
- *Business models.* Research libraries have worked to develop and partner in a variety of projects and business strategies aimed at the long-term preservation of digital materials, including Portico, LOCKSS, and the National Digital Information Infrastructure and Preservation Program (NDIIPP) (Joint Task Force on Library Support for E-Science 2007).

There are additional (perhaps newer) areas where important opportunities may exist for libraries as partners in data curation efforts. One such opportunity is in curating “small science” data, which some predictions hold will exceed the total volume of “big science” data sets such as those generated by climate, astronomy, or genomics research (Carlson 2006; Gold 2007b). Facilitating the work of publishing data sets related to publications is another area that libraries might successfully participate in, as models for supporting this activity are not yet fully developed (Gold 2007b; Joint Task Force on Library Support for E-Science 2007; Lynch 2006). Finally, while Messerschmitt (2003) argues that repositories supporting cyberinfrastructure development should generally be based in their respective disciplines, he does allow that some libraries may have sufficient subject expertise in a particular discipline to warrant consideration as hosts for discipline-specific repositories.

II. Environmental Scan – Beyond Cornell

This section is intended to serve as a very high-level and selective view of data and e-science activities in progress outside of Cornell, which are too numerous and diverse to survey thoroughly. We include selected information on leading national and international coordination and research and development efforts, activities of selected U.S. universities and academic libraries, and existing training opportunities for librarians. Additional information on selected discipline-specific efforts related to data curation is included in Appendix B.

National and International Activities

CODATA, ICSU, and the World Data Centers

CODATA, the Committee on Data for Science and Technology of the International Council for Science (ICSU), works to improve access to high quality scientific data (CODATA 2008). It serves a coordination role by sponsoring conferences and workshops, and convening committees and task forces to address specific issues such as data access for particular disciplines, and access to and preservation of data in developing countries. CODATA is also the sponsor of the peer-reviewed Data Science Journal. National committees of CODATA exist to coordinate national activities; the United States National Committee for CODATA is affiliated with the National Research Council of the National Academies (National Academy of Sciences 2008). ICSU established the World Data Center (WDC) system in support of the International Geophysical Year (1957-1958). Currently there are more than 50 WDCs in operation around the world, serving the earth and space sciences, accepting data from a broad array of programs as well as individual scientists, and making data available at the cost of distribution (National Geophysical Data Center n.d.).

United Kingdom

The UK is a leader in the area of digital data curation, in part because there is some coordination of efforts, led by the Digital Curation Centre (DCC). The DCC was established in 2004 and is a joint project of the Joint Information Systems Committee (JISC) and the Research Councils e-Science Core Program. The DCC is charged with leading research and development efforts in digital curation for research data and publications, providing strategic leadership, influencing national and international policy, providing expert advice to practitioners and funders, creating high quality resources and tools, engaging in staff development for curators, and strengthening networks and partnerships in this area (Hockx-Yu 2007).

Liz Lyon, Associate Director for Outreach at the DCC, produced an important report in 2007 that articulates the "roles, rights, responsibilities and relationships of institutions, data centres and other key stakeholders who work with data" (Lyon 2007). While focused primarily on activities and issues in the UK, it contains a number of observations and recommendations of interest to a more general audience:

- Research funders should create, implement, and enforce data management, sharing, and preservation policies.
- Research project proposals should include data management plans and these plans should be reviewed.
- Higher education institutions should have institutional data management, sharing, and preservation policies, and the policies should recommend deposit in an open access repository.
- A study should be conducted to identify the properties of data sets that facilitate reuse, and best practice guidelines should be developed.
- Roles and responsibilities for universities include the right to be offered a copy of research data, to set internal data management policy, to manage data in the short term (meeting best practices standards), to promote local repository services, and to provide training and support for scientists.
- Institutional repositories are emerging as an alternative for depositing research data, although not all groups of stakeholders view institutional repositories as suitable for research data. Some stakeholders expressed skepticism that institutional repository staff, as generalists, have the needed skills for digital data curation. Existing institutional repositories tend to be focused on a specific discipline (e.g. eCrystals for chemistry, GRADE for GIS data). Lyon also noted that researchers tend to identify more closely with disciplinary communities than their institutions, and that this barrier must be addressed for institutional repositories to be successful. In spite of these issues, there is also some recognition that institutional repositories are more likely to be stable than their discipline-based counterparts over the long term.

European Union

DRIVER, the Digital Repository Infrastructure Vision for European Research, is an international partnership aimed at facilitating the development of infrastructure to make all types of scientific information, whether papers, reports, data, or other types of information, publicly accessible across Europe. Like the proposed Australia National Data Service (see below), a principle aim of DRIVER is to develop a network of institution-based repositories, whether subject-based or not (DRIVER 2008).

Australia

In 2006, the Prime Minister's Science and Engineering Innovation Council (PMSEIC) was charged with developing recommendations for a national strategy on the management of scientific research data (Working Group on Data for Science, PMSEIC 2006), and is expected to strongly influence national policy and investment in this area. The proposed Australian National Data Service (ANDS) is intended to address many of the needs identified in the report to the PMSEIC, and its areas of activity include the development of a national framework program (policies and agreements in support of a national network of repositories, to comprise the Australian "Data Commons"), a utilities program aimed at developing needed technologies to support the network, a repositories program intended to help existing repositories make needed improvements, and a researcher practice program intended to assist researchers and data managers obtain needed skills (ANDS Technical Working Group 2007). ANDS itself is expected to play a coordinating role in all of these related efforts. Noteworthy about the ANDS report is the emphasis on the role of institutions in managing and hosting their own data, with ANDS facilitating the development of an "overlay" style data commons. The report asserts that institutions have inherent interests in managing their own data for a variety of reasons, including maintaining the scholarly record of work undertaken at the institution, to retain data to support published research claims, to ensure access to data for commercial applications of research, and supporting re-use of data. The report acknowledges that some types of data, very large and established data sets or high-volume data sets produced from large research instruments or facilities, may exist outside of the institution-based framework described in the report.

The activities at Monash University are noteworthy in terms of utilizing institutional repositories for research data. One interesting aspect of their work is the definition of a continuum for data curation that recognizes that research data curation needs may not be met satisfactorily in one repository environment. Treloar et al. (2007) describe an active "collaboration domain," characterized by more and larger data sets, often dynamic, and with perhaps less metadata, a more mature "publication domain," as well as a "curation boundary," where the work of migrating and preparing content for publication occurs.

United States

Federal agencies have formed the Interagency Working Group on Digital Data (IWGDD), with the aim of ensuring that all scientific data generated by federal agencies is publicly accessible (Butler 2007). Funders, such as the National Institutes of Health (NIH) and the National Science Foundation (NSF), are interested in maximizing the return on their research investments by encouraging or requiring data sharing. The NIH requires recipients of grants of more than $500,000 to share data, but the conditions for doing so are flexible and include an option to share data "under the auspices of" the principal investigator (National Institutes of Health 2007). Some programs of the NSF have data sharing requirements; the Division of Ocean Sciences is one such program (Division of Ocean Sciences 2003). Unlike some other federal agencies, NSF as a whole does not have consistent policies, nor does it provide systematic support for data sharing and archiving. NASA and NOAA, in contrast, do provide more significant and systematic support for maintaining accessible digital data collections. NASA, for example supports nine Distributed Active Archive Centers (NASA 2007).

The NSF Office of Cyberinfrastructure (OCI) is actively encouraging the development of infrastructure to support data sharing, issuing a program announcement in September 2007 entitled "Sustainable Digital Data Preservation and Access Network Partners (DataNet)". Total funding for this new program is $100,000,000. DataNet partners are expected to "serve as component elements of an interoperable data preservation and access network," providing capabilities in preservation, access, analysis and visualization. The NSF expects that new types of organizations will have to be formed to meet this challenge and expects DataNet partners to incorporate expertise from multiple domains, including cyberinfrastructure, library and archival sciences, computer and information sciences, and domain sciences (National Science Foundation Office of Cyberinfrastructure 2007). In addition to the DataNet program and other OCI programs, NSF also has also issued a number of calls for proposals for discipline-specific cyberinfrastructure and data curation activities, including the areas of plant science, environmental observatories, arctic research and the International Polar Year (2007-2008), and others (National Science Foundation 2007).

Finally, commercial entities are expressing an interest in partnering in activities related to data curation and distribution. A few examples include:

- A partnership between Fedora Commons and Sun Microsystems to create a petabyte-scale data storage system that combines Sun's StorageTek 5800 system with the Fedora repository platform (Fedora Commons 2008).
- Google's 2005 announcement that it had signed a memorandum of understanding with NASA to collaborate on large-scale data management and other activities (Google 2006), and a January 2008 announcement that Google intends to host open-access scientific data (Madrigal 2008).
- Microsoft is developing a research outputs repository (Dirks and Parastatidis. 2008).

Universities and Academic Libraries

The San Diego Supercomputer Center (SDSC) continues to make significant contributions in the area of data distribution and management, developing such applications as the Storage Resource Broker (SRB, middleware for managing distributed collections), and iRODS, data grid software that supports the application of management policies to digital collections via a rule engine. SDSC is a partner in many data curation and distribution activities, including some described elsewhere in this report (the NEES data network, and the Knowledge Network for Biocomplexity, for example). SDSC is also a partner in developing cyberinfrastructure to support shared research facilities and sensor networks, including those in development by the Consortium of Universities for the Advancement of Hydrologic Science (CUAHSI), and the National Ecological Observatory Network (NEON). Data Central (http://datacentral.sdsc.edu/) is a relatively new unit at SDSC, and provides data hosting and management services.

In academic libraries, new positions and organizations are being created. Purdue University Libraries is home to the Distributed Data Curation Center (D2C2), which aims to support the curation of "small science" data by developing a distributed institutional repository system, and addressing some of the research questions that arise in this particular environment. Along with a director, two professional positions support the D2C2: an interdisciplinary research librarian, and a research data scientist (Purdue University Libraries 2007).

The University of Washington Libraries announced in 2007 its decision to appoint Neil Rambo, then Associate Director of Health Sciences Libraries, to the new position of Director of Cyberinfrastructure Initiatives and Special Assistant to the Dean of University Libraries for Biosciences and e-Science. Simultaneously, he was appointed to a half-time position with the Association of Research Libraries as a program officer, tasked with working with ARL staff and members on e-science issues (Association of Research Libraries 2007).

Johns Hopkins University announced the creation of the Digital Research and Curation Center (DRCC), to be headed by Sayeed Choudhury, and intended to “manage, preserve, and provide access to the mounting digital scholarship generated by faculty and researchers at the university” (Johns Hopkins University 2007).

Training and Careers in Data Curation

The field of data curation is still quite new and few opportunities for training exist. The need for a career path for “data scientists” is widely recognized (e.g. National Science Board 2005). The need for training for librarians is also recognized (e.g. ARL Workshop on New Collaborative Relationships 2006), although there are already some programs and courses worth noting. Indiana University’s School of Informatics offers undergraduate and gradate degrees in informatics, requiring BS students to have a disciplinary focus in addition to coursework in informatics. Masters degree options include bioinformatics and cheminformatics. The University of Illinois at Urbana Champaign has developed two programs to address the need for trained professionals in this area, a Biological Information Specialist Master of Science degree, and a concentration in Data Curation within the Library and Information Science MS program. A 4-day Summer Institute on Data Curation (for practicing librarians) is also planned for 2008, and will address the topics of digital preservation, technical aspects of data repository systems, appraisal and selection of digital data, and resource requirements for a data curation program. The School of Information and Library Science at the University of North Carolina (Chapel Hill) is working to develop a curriculum in the more general area of digital curation, and offers a graduate course in cyberinfrastructure, as well as a certificate of specialization in bioinformatics. Syracuse University offers a course in science data management, and includes scientific metadata in its general course on metadata. While not necessarily aimed at training research data curators, many library and information science programs offer relevant courses such as health or medical informatics, database design, information architecture, and systems analysis.

III. Environmental Scan – within Cornell

This section is intended to serve as a selective (but representative) view of data curation activities that are primarily centered at Cornell. In compiling these examples, particularly those outside the library, we encountered a broad range of activities. Some organizations have multiple, significant activities related to data curation, or may collaborate with other organizations at Cornell. There are also individual data collections scattered throughout the university. We chose to focus here on those organizations and units with significant activity in this area. We also describe in Appendix C selected individual collections managed outside the units described in this section, but it is worth noting here that some of those activities have ceased, generally due to lack of funding. Overall, the diversity of initiatives implies substantial challenges in meeting campus needs, such as financial sustainability, appraisal and selection, preservation of data, and cooperation among Cornell units. Some of these issues will be taken up in Section IV of this paper.

Data Curation within the Cornell University Library

CUL manages a broad array of digital collections (Cornell University Library 2007). While most of these would not be considered data collections in the sense generally used throughout this paper (digital data collected or generated in the course of conducting research), there are some noteworthy efforts involving text that we chose to include because they hold the potential to support cyberscholarship, particularly in the humanities. We offer here brief descriptions of those text-oriented efforts, followed by other data initiatives, and finally a summary of some of the research projects currently underway.

Text-based Collections

Large Scale Digitization Initiative (LSDI)

CUL entered into agreements with Microsoft in 2006 and with Google in 2007 to digitize materials from its collections. The Microsoft effort is focused primarily on out-of-copyright materials and will result in the digitization of approximately 100,000 volumes in the first year of production. Google will digitize up to 500,000 volumes from CUL and is not restricting digitization to works in the public domain. Both projects have the potential to begin to enable novel forms of scholarship, particularly in the humanities, by making text mining of a large number of works possible. The volume of data for the first year of the Microsoft project is estimated to reach 40TB, making it the largest collection managed by CUL. The development of an OAIS-compliant repository, based on the aDORe architecture is also a new effort for CUL.

eCommons - http://ecommons.cornell.edu

eCommons is Cornell University's institutional repository and provides members of the Cornell community with a means of storing and distributing digital materials that are useful for educational, scholarly, research or historical purposes. Five years after its 2002 launch, it contains slightly more than 8000 items in 355 collections. Under-utilization by faculty is an ongoing issue (64% of items were batch-uploaded as a result of deliberate collection development efforts), although eCommons does have some staunch advocates among faculty, and the Dean of the Graduate School supports a requirement for deposit of electronic theses and dissertations. We are also beginning to see eCommons being used as a generic platform for distributing small data sets. This includes data in support of publications (see for example "Vascular Plant Species of the Cayuga Region of New York State," a data set related to a paper published in the *Journal of the Torrey Botanical Society*, http://hdl.handle.net/1813/9413), as well as stand-alone data sets (for example those in the collection of the Agricultural Ecology Program: http://ecommons.library.cornell.edu/handle/1813/7659/browse-title).

arXiv - http://arxiv.org/

Initiated in 1991 by Paul Ginsparg (then at Los Alamos National Laboratory), the arXiv was conceived as a means for high-energy particle physicists to share pre-prints. Now maintained by CUL and supporting additional disciplines, the number of items in the collection is approaching half a million. The arXiv is viewed as one of the unmitigated success stories of the open access movement. Future plans for arXiv include support for distributing data sets related to publications in arXiv.

Data Collections

United States Department of Agriculture Economics, Statistics and Market Information System (USDA ESMIS) - http://usda.mannlib.cornell.edu/

The USDA ESMIS is a collaborative effort including Mann Library and five economic agencies of the USDA, begun in 1993 and supported with funding from the USDA. The system provides the public with fast and free electronic access to information covering U.S. and international agriculture and related topics, with a specific focus on commodity information. Participating agencies issue reports daily, weekly monthly or annually, which Mann Library disseminates within 5 minutes of their release via the web interface or email. More than 2 millions reports are downloaded annually. Mann Library support for the system is mainstreamed into normal library operations and is a part of the job descriptions of one librarian who serves as the project manager, two technical services staff, and two programmers. Reference staff handle questions from users at the reference desk.

Cornell University Geospatial Information Repository (CUGIR) - *http://cugir.mannlib.cornell.edu/*

CUGIR provides free, open access to geospatial data and metadata for New York State, with special emphasis on those natural features relevant to agriculture, ecology, natural resources, and human-environment interactions. As a participating node in the National Spatial Data Infrastructure (NSDI), CUGIR metadata are harvested for inclusion in Geodata.gov, a federal portal for geospatial data. Created in 1998 for the purpose of distributing U.S. Census TIGER/Line data online, CUGIR now hosts data sets from a range of data providers, including federal agencies (such as the U.S. Census Bureau and the U.S. Geological Survey), state agencies (such as the New York State Department of Environmental Conservation and the New York State Department of Agriculture and Markets), local county and city governments, and individual research groups (usually affiliated with Cornell). Data are generally published exactly as received from the provider, although on occasion CUGIR staff will consult with the data provider and make modifications to data to make it more widely usable. CUGIR staff may also provide significant assistance with the preparation of metadata that conforms to the FGDC standard (Federal Geographic Data Committee's Content Standard for Digital Geospatial Metadata). Since 2007, CUGIR has begun to include metadata records for externally-hosted data sets that fall within its collection scope, to facilitate discovery of data for New York State from multiple sources. The total size of the collection is approximately 10GB, with over 600,000 downloads from 2001-2007. As with the USDA-ESMIS system, CUGIR is just one element of the job descriptions for the library staff that support it (2 librarians, 1 programmer, and 1 support staff).

Blackout archive - http://metadata.library.cornell.edu/blackout/

The purpose of the Blackout Archive is to archive and preserve technical and other data related to the northeast U.S. electric power blackout of August 14, 2003, collected by the blackout investigation team. The data (including thousands of transcripts, graphs, generator and transmission data and reports) were procured under a confidentially agreement, archived "as is," and made available to technical experts and the blackout investigation team to determine the cause of the blackout and to make recommendations to avoid similar power failures in the future. Implementation of the archive presents the opportunity to create new standards for information exchange among groups in the power industry, although the project is on hiatus until further funding is secured.

Research and Development Activities

Data Intensive Science Organization for Virtual Exploration and Research - Research Service Group (DISCOVER RSG)

The Office of the Vice Provost for Research issued a call for proposals in Fall 2007 for Research Service Groups (RSGs). These RSGs are intended to enable Cornell's Center for Advanced Computing (CAC) to support cyberinfrastructure development and to respond to significant requests for proposals in collaboration with the Cornell research community. The proposed DISCOVER RSG's purpose is to create a research service group oriented toward data-driven scientific discovery across a variety of disciplines at Cornell. Collaborators include the Cornell University Library and the Information Science (IS) program, as well as faculty from multiple disciplines – including astronomy, crop and soil science, physics, and ornithology. The goals are to develop a system that enables storage and curation of diverse data sets as well as tools for access, discovery, and analysis that promote cross-disciplinary studies. The DISCOVER group will assess current and projected needs of Cornell research teams, create a roadmap for developing the needed infrastructure and resources, initiate pilot projects in selected strategic areas, work with staff in the Library, the CAC, and IS to help integrate various components and make recommendations for needed infrastructure, and respond to calls for proposals from various funding sources. Funding is expected to be available for two years beginning in 2008.

Data Staging Repository (DataStaR) and Library-Laboratory Collaboration (LiLaC)

DataStaR, a project funded by the Office of Cyberinfrastructure of the NSF, is a logical extension of a Small Grant for Exploratory Research (SGER) awarded to Mann Library in 2004. In the SGER, the project team developed a conceptual model for library-laboratory collaboration (LiLaC) and explored the feasibility of large research libraries offering data curation services to researchers. The project team worked with two Cornell research groups: the Cornell Language Acquisition Laboratory (CLAL), and the Upper Susquehanna River Basin Agricultural Ecology Program. The DataStaR grant, awarded in October 2007, extends this work by developing an institutionally-based data staging repository whose function is to facilitate the documentation and transmission of research data sets from a variety of disciplines to domain-specific repositories and/or institutional repositories. Using the staging repository, a researcher will be able to create preliminary metadata for research data sets; share preliminary data publicly, or only with selected colleagues; complete a more detailed metadata record using a form-based editor; optionally upload completed data sets to the staging repository; export metadata in any number of domain-specific formats; re-use elements of existing metadata records in the creation of new metadata records; and obtain assistance with any of these processes from librarians with domain-specific or general curatorial expertise. The project's proposed metadata management scheme leverages Mann Library's Vitro web application, the software that underlies VIVO, CUL's integrated source of information on life sciences activities and resources at Cornell (http://vivo.library.cornell.edu/).

Other Cornell University Organizations with Significant Activities Related to Data Curation and Cyberinfrastructure

Center for Advanced Computing - http://www.cac.cornell.edu/

The Cornell Center for Advanced Computing (CAC) is a leader in high-performance computing system, application, and data solutions that enable research success. As an early technology adopter and rapid prototyper, CAC helps researchers accelerate scientific discovery. CAC serves Cornell faculty researchers from dozens of disciplines, including biology, behavioral and social sciences, computer science, engineering, geosciences, mathematics, physical sciences, and business. The Center operates more than 2,000 processors in Linux, Windows, and Mac OS X configurations. CAC technical staff has experience and expertise in high-performance computing systems and data storage; application porting, tuning, and optimization; database system design and deployment; computer programming; and Web portal design. CAC is also recognized for its excellence in online training and, as part of an NSF-funded team, is developing new training on topics such hybrid programming with OpenMP and MPI and large scale data visualization. CAC has played a leading role in the Northeast Lamda Rail consortium; as a result, Cornell has access to a high-capacity, high-speed (10-gigabit per second) research network known as the National Lamda Rail. CAC has also connected to the national TeraGrid and participates in the TeraGrid Science Gateway program.

Sample CAC collaborations include:

- CAC and Arecibo: Searching for Pulsars in Very Large Databases - CAC brings Arecibo telescope data to Cornell for analysis by astronomers from around the world. A 3-hour observing session generates ~1 TB of data. CAC developed a data pipeline from the observatory to Cornell, where signal processing code transforms the data prior to loading into a large relational database where it becomes accessible online for further analysis in the search for pulsars. Online plotting tools range from simple histograms, scatter plots, and line plots, to customized tools for more specific types of analysis (TeraGrid site: http://www.teragrid.org/programs/sci_gateways/projects.php?id=54, Cornell University Center for Advanced Computing n.d.c)
- CAC and the College of Agriculture and Life Sciences (CALS): Improving Weather Data Accuracy and Accessibility - As part of the Cornell Computational Agriculture Initiative, CAC provides Web services to support query and extraction of temperature and precipitation data for agricultural, environmental, and water resource models requiring a high level of accuracy at small spatial scales (TeraGrid site: http://www.teragrid.org/programs/sci_gateways/projects.php?id=62, Cornell University Center for Advanced Computing n.d.b).
- CAC and John Bunge: Estimating Life's Diversity on Land and at Sea - Cornell researchers such as John Bunge, Chair of Social Statistics, use CAC database design and development expertise to manage vast sets of data such as microbial and DNA data. Shortening the time it takes to process data and to get from the collection to the interpretation stage is a strength of CAC (Cornell University Center for Advanced Computing n.d.a).

CAC also collaborates with the Database Research group in Computer Science, led by Johannes Gehrke. This group is developing general-purpose data management and analysis tools which can be adapted to data problems that cut across multiple disciplines.

Cornell Information Technologies (CIT) - http://www.cit.cornell.edu/

Cornell Information Technologies provides the information technology infrastructure, services, and security that undergird Cornell's teaching and research mission. Current initiatives and strategic goals include building a new university facility to provide data center services, data storage, and computational resources; upgrading Cornell's network connections with the outside world; and providing improved collaboration tools and facilities to link all Cornell units together and to scholars around the world.

As a participant in regional and national consortia focused on high- speed research networks for higher education, including the New York State Grid, NYSERNet, Northeast LamdaRail, Internet2 and National LamdaRail, Cornell will continue to encourage regional and national infrastructure consolidation and collective approaches to provide robust, seamless connectivity. CIT will also continue its collaborations with the Center for Advanced Computing (CAC) to explore improvements to Cornell's cyberinfrastructure including computational resources, grid computing collaborations, data acquisition, campus networking, storage, and backup (Cornell Office of Information Technologies 2007). CIT has also collaborated with CUL to specify, purchase, and configure a mass storage service to support the library's large-scale digitization projects.

Cornell University Program in Information Science - http://www.infosci.cornell.edu/

Cornell's Faculty of Computing and Information Science (CIS) brings together experts in computing with researchers and scholars from computer science, operations research, human computer interaction and the social sciences. Cornell has played a leadership role in several national and international digital library initiatives including the National Science Digital Library and the Fedora Commons open-source repository project. Other current research includes the Open Archives Initiative Object Reuse and Exchange (OAI-ORE) project, a specification for describing compound digital objects (http://www.openarchives.org/ore/). The Cornell Web Laboratory is a joint project of the Information Science program and the Computer Science department, with computing facilities based at the Center for Advanced Computing. The project builds on the historical collections of the Internet Archive to provide researchers in the social sciences and information science a database and tools with which to analyze the collection of the Internet Archive in greater depth than is possible through the Internet Archive's Wayback Machine. A data analysis cluster provides support for crawling subsets of the collection and interactively exploring the web graph of a selected set of initial pages; results can be retained for further analysis or comparison across time or domain of inquiry.

Cornell Institute for Social and Economic Research (CISER) Data Archive - http://ciser.cornell.edu/info/about.shtml

CISER’s current focus is on providing infrastructure support for social science research at Cornell University, including a cluster of multi-processor servers and attached file server for general research computing, as well as the Cornell Restricted Access Data Center (CRADC), a secure computing environment for use of restricted-use data. The manager of the CRADC works closely with the Office of Sponsored Programs and the Institutional Review Board (IRB) to insure compliance with existing human subject research regulations, procedures, and ethics in restricted data plan consultation and management. CISER is also home to a U.S. Census Research Data facility, one of 9 such units in the country, which affords a unique opportunity to work with confidential data produced by federal statistical agencies.

CISER serves as the Cornell liaison unit to Inter-University Consortium for Political and Social Research (ICPSR) and Roper Center for Public Opinion Research and maintains these memberships on behalf of Cornell, as well as memberships in the Association of Public Data Users (APDU) and Consortium of Social Science Associations (COSSA). CISER staff collaborate with several units across campus, including the Cornell Statistical Consulting Unit, Center for Advanced Computing, Cornell University Library, National Data Archive for Child Abuse and Neglect, Institute for the Social Sciences, and the Survey Research Institute.

The data archive is a focal point for social science dataset acquisition, maintenance, and delivery. Collection development is largely driven by the needs (both expressed and anticipated) of CISER’s primary user groups. Holdings consist of both public data produced by federal and state statistical agencies, datasets purchased or licensed for use by Cornell researchers, and materials acquired from memberships. The archive maintains a modest number of datasets produced by Cornell researchers and also encourages faculty to submit their data products to ICPSR for distribution via its publications archive. CISER leadership is pursuing a dynamic approach to research data archiving with the Survey Research Institute, urging clientele to consider long-term disposition of their data products.

Cornell Lab of Ornithology - http://www.birds.cornell.edu/

The Cornell Laboratory of Ornithology is a world leader in research and conservation of bird population studies (http://www.avianknowledge.net). One facet of the Lab's mission is to engage citizens in participatory science, and it is recognized as the leader in citizen science. The Information Science Program at the Lab is analyzing tens of millions of observations of birds, combining them with GIS and remote sensing information, and using advanced computing techniques to model the patterns and trends in bird populations at hemispheric scales. The Conservation and Bioacoustics Research Programs at the Lab are developing a network of nocturnal flight call detectors with the aim to quantify bird migration, and are generating tens of terabytes per year of digital frequency-time acoustical data. The storage and accessibility requirements for this data are very similar to the astronomical surveys, and algorithms for transient signal detection in acoustical data may have relevance to astronomy - in both cases, detection algorithms operate on the frequency-time plane after filtering in the spatial domain. The goal of this project is to develop accurate species detectors that will allow the Lab to quantify and identify the occurrence of bird species during migration at the continental scale.

The Lab is also home to the Macaulay Library (http://www.birds.cornell.edu/macaulaylibrary/; formerly the Macaulay Library of Natural Sounds), which maintains the world’s largest collection of animal sounds and video. Much of the collection has been digitized and is being made available online.

Weill Cornell Medical College

Weill Cornell Medical College provides researchers with 22 Core Facilities (http://www.med.cornell.edu/research/rea_sup/). These are partial cost-recovery units, and each core facility serves a unique user-driven function in the research life of the College. The sophistication of data management schemes, services, and infrastructure between cores varies.

All the cores present their clients with digital data even if it requires converting analog data into digital data. Cores deliver data with their clients in a variety of ways (CD or DVD, dedicated server, with or without discipline-specific software for use on core servers). Data storage also varies by core. The majority retain copies of data they have generated, though the mechanism and commitment to length of storage varies and may not be specified. At least one core, Nuclear Magnetic Resonance, requires users to be responsible for their own data. Certain cores such as Mass Spectrometry generate a high volume of data, up to 10GB of data output per 1-hour analysis.

Data management solutions do not seem to be in demand from researchers. The Biostatistics and Research Methodology Core advises researchers on data management plans when grant proposals require them. The Computational Genomics Core's current policy is to refer questions on data management to the Institute for Computational Biomedicine (http://icb.med.cornell.edu/). In addition, Cornell University researchers have access to the Columbia Genome Center, which manages a Bioinformatics Core Facility (http://amdec-bioinfo.cu-genome.org/) funded by AMDeC, a consortium of New York's medical schools, academic health centers, and medical research institutions.

Five core administrators cited a national repository for their discipline: DNA Sequencing and Gene Therapy – GenBank; Microarray – Gene Expression Omnibus (GEO); Computational Genomics – GEO; Nuclear Magnetic Resonance - Protein Data Bank. The Mass Spectrometry core administrator indicated that a repository with non-proprietary standards is on the horizon and should be widely accepted within the year.

Beyond the Core Facilities, the Office of Research and Sponsored Programs (RASP) contracts with an off-site facility to maintain the Research Data Archive for data and related documentation (http://www.med.cornell.edu/research/rea_com/res_arc_sto.html).

IV. Data Curation Issues

Organizations involved in data curation face several important challenges. Most are not entirely unique to digital research data, although the rapidly increasing volume of research data may exacerbate certain challenges, such as financial sustainability, selection and appraisal, and digital preservation.

Financial Sustainability

As all kinds of organizations begin to assess their roles and responsibilities with respect to digital research data, they must necessarily consider the costs of supporting data curation and preservation. The newness of these activities for academic libraries can make forecasting and planning quite difficult. Some data centers have developed practical business models to sustain the work of curating digital data over the long term, but in general, such models are lacking.

Grant-based funding is not a long-term solution to the problem, and NSF itself acknowledges that the “vast majority of NSF support carries with it no long-term commitment.” (National Science Board 2005) Not surprisingly, there are several current and recent efforts to explore models for sustaining digital data curation. Perhaps the newest significant one (announced in September, 2007) is a blue ribbon task force on economically sustainable digital preservation and access, funded by the NSF and the Andrew W. Mellon Foundation, and co-chaired by Fran Berman of the San Diego Supercomputer Center and Brian Lavoie of OCLC (Blue Ribbon Task Force on Sustainable Digital Preservation and Access 2008; OCLC 2008). A joint conference in 2005, sponsored by the Digital Curation Centre and Digital Preservation Coalition, explored digital curation cost models, and concluded that there are “very few concrete answers,” and that “much more work must be done in determining useable cost models” (Digital Preservation Coalition 2005). The 2006 ARL-NSF workshop on long-term stewardship of digital data also made economic sustainability one of their core issues for discussion, exploring the topics of what to sustain, how to get value, payment approaches, persuading funders, capacities required, capabilities required, and economic roles. Some of their recommendations were to “involve economics and social science experts in developing economic models,” to experiment with different repositories, to “use the NSF program process to help the research and academic library community take more responsibility for the stewardship of scientific and engineering research data” (ARL Workshop on New Collaborative Relationships 2006).

In her review of business issues related to digital repositories, Alma Swan (Swan 2008), presents five general repository business models:

- Institutionally-owned: responsibility for operating a repository is assumed by an institution whose goals are advanced by the repository;
- Publicly-owned: a public organization sponsors the repository for the public good, services may not be suitable for revenue generation, and an institutional or community-based model is not appropriate;
- Community-supported: the repository is sustained by the community it serves;
- Subscription-based: services are sold on a cash basis to paying customers;
- Commercially-supported: revenue is generated by commercial means (other than subscription), such as the sale of advertising

Data centers operated by the U.S. federal government, such as those maintained by the Census Bureau and the National Agricultural Statistics Service, are examples of publicly-owned data repositories operated for the public good (but note also the mandates that require making government information available, as described earlier in this report). The federal government also operates data centers which recoup the distribution costs for at least some of the data they provide; the National Oceanic and Atmospheric Administration’s (NOAA) National Oceanographic Data Center and the National Climate Data Center are two examples. Celera attempted to commercialize the distribution of human genome data, but abandoned the effort when it could not compete with the free, public distribution of the data derived from the publicly-funded Human Genome Project. The Inter-University Consortium for Political and Social Research (ICPSR), is an example of a subscription-based model operating in service of a particular community (social scientists), with institutions paying a membership fee to ICPSR.

The National Aeronautics and Space Administration’s (NASA) effort to develop a cost-estimation tool for data systems is noteworthy in its detail and comprehensiveness. The study team quantified the cost of each of the various functions of a repository (for example, ingest), by breaking each function down to the component costs (e.g. FTE for management, software development or customization, etc.) associated with a defined level of service (Hunolt 2002, Hunolt 2003). The approach is a functional one and requires that organizational structures and services be known in great detail – something that is quite difficult for an organization just beginning to work in this area.

If Cornell University in general, or CUL in particular, were to choose to host a data repository for the Cornell community, one option is to simply support it with no expectation of revenue generation, and to justify this on the basis that such support serves the best interests of the university. While some of the other options for recovering at least partial operating costs may be politically very difficult to effect, they include charging depositors (on the assumption that such costs might be written into grant proposals), obtaining funding from facilities and administrative (F&A) charges to grants, and offering value-added services to data distribution to generate revenue. It will take some time for CUL to assess what its role is in this area, what services it should offer, and how best to fund them.

Appraisal and Selection

There are two main schools of thought regarding the need for selection and appraisal of data sets. Proponents of a “technological deterministic future” argue that the cost of computer storage will continue to decrease, making it practical to forego appraisal and selection and to keep everything (Harvey 2007). Harvey goes on to assert that this “approach is limited just to bit preservation and therefore considered not appropriate for knowledge preservation;” it does not account for additional costs such as metadata creation and higher-level digital preservation costs such as format migration, which may be far greater than the cost of storage alone. A more reasoned approach would mandate the selection and appraisal of data before making a long-term commitment to their preservation. This approach may significantly alter the traditional role of the archivist/curator, requiring an appraisal/selection process at the point of creation, or the beginning of a data set’s life cycle. The archivist can no longer wait “passively at the end of the life cycle for records to arrive at the archives when their creators no longer wanted them – or were dead” (Cook 2000).

Digital data sets pose some special challenges when it comes to selection and appraisal. Unlike public or government records, which are considered evidential (i.e. transactions of legal, fiscal and administrative value), and may be managed with methodologies such as macro-appraisal. Research data sets are considerably more heterogeneous and defy the application of such general standards. They may be repurposed in many ways by multiple users. This is the challenge of creating cohesive standards for research data archives. Strategies must be developed that ensure the preservation and authenticity of born-digital records, whose characteristics are more transient and ephemeral than analog materials.

Appraisal decisions must take into consideration the mission and goals of the institution, intellectual property considerations and rights, and legal and contractual obligations. Beyond that, additional criteria for selection and appraisal might include:

- Are the data important or relevant to other constituencies in the discipline?
- For what purpose were the data originally created?
- Are the data significant, unique and influential?
- What is the potential for re-use of the data in secondary scholarship and research?
- How significant is the principle investigator/faculty member in his or her discipline?

- Is there intellectual duplication in other areas within or outside the institution?
- Are the data complete and authentic?
- What would the consequences be if these data were not available?
- Are the data accessible/readable?
- What are the resources required for the long-term storage, preservation and dissemination of these data sets? Can the institution maintain the technologies required to use the data?
- Is there documentation supporting the creation, provenance, and maintenance of these data sets? Is the descriptive and preservation metadata associated with the data sufficient to provide access and long-term technical support and sustainability?
- Are there legal requirements for the retention of this data? Does the institution have a mandate to preserve the data as a condition of receipt of federal funds?
- Do issues of privacy, through identifiable personal information, restrict access to the data?
- Do intellectual property rights or copyright considerations restrict access or use of the data?

Acquisition of data sets may require archivists to be more creative in understanding the needs of potential users; indeed the application of traditional descriptive archival standards may not provide optimum access for the re-use of these records.

Digital Preservation

Digital preservation is not a new area of concern for libraries, and the library community has accomplished a great deal by collaborating on the development of standards and best practices for digital preservation. Among the successes are the development of the OAIS reference model, the subsequent Trustworthy Repositories Audit & Certification checklist, and PREMIS (PREservation Metadata: Implementation Strategies). Within CUL, one of the 2005-2006 priority objectives included a three-year project to build an OAIS-compliant system for managing Cornell's digital assets. Planning was the focus of the first year of work, with construction of the system taking place in the second and third years. While that goal has yet to be fully realized, progress has been made. The CUL Digital Preservation Framework (2005) articulates the objectives of CUL's digital preservation program, and specifies the scope of effort, operating principles of the program, and roles and responsibilities for digital preservation, very general guidelines for selection and acquisition, and anticipated challenges. The Common Depository System was established; work began on developing preservation processes for selected collections (arXiv, MathArc), and development of an OAIS-compliant repository for the Google and Microsoft book scanning projects is nearly complete. The library also tested the demand for and feasibility of a file format and media migration service; a pilot service launched in 2004 and handled more than two dozen requests for file and/or media migration for thousands of files (Entlich and Buckley 2006). That service was eventually discontinued, and the authors argued that faculty should be encouraged to deposit data in institutional repositories, thus enhancing "the survivability of digital content by putting it into a well-managed, centralized environment where it can be subject to state-of-the-art technological and organizational digital preservation techniques."

Early digital preservation efforts in libraries were focused on creating and preserving digital versions of analog materials. Digital research data pose some additional, unique challenges for preservation, including variety and complexity of file formats, as well as sheer volume, which continues to increase rapidly. Preserving the means to access and use digital data is a universal challenge, but the challenge may be even more significant for research data. Messerschmitt (2003) points out that in addition to the obvious candidates for preservation (data, and descriptive and structural metadata), it is also necessary to preserve "the logic, processes, and algorithms" as well as the software that execute models and simulations, or enable data analysis. As an example, in the field of archaeology alone, respondents to a survey conducted by the Archaeology Data Service in the UK identified more than 50 software packages used for collecting and analyzing data. The fact that many of these software packages and their associated file formats are proprietary makes the work of preservation even more difficult (Austin and Mitcham 2007).

One example of the specialized needs for data preservation is that of modern database management systems. Database preservation was the topic of a conference held in Edinburgh in March of 2007 (http://homepages.inf.ed.ac.uk/hmueller/presdb07/). Database systems facilitate query-based access to data rather than to the raw data itself, and general characteristics of databases that challenge current preservation practice include their dynamic nature, as well as structure and integrity constraints. Curated databases such as molecular biology databases also present unique challenges. Workshop participants considered what aspects of databases require preservation and shared different possible strategies for preservation (Christophides and Buneman 2007); many of the proposed solutions rely on "flattening" databases, preserving data as single tables in a non-proprietary format, and storing information about database structure separately.

The Library of Congress' National Digital Information Infrastructure and Preservation Program (NDIIPP) includes some partners whose focus is on digital data. Examples of partner projects include: investigations of data provenance, creation of an archive of Moderate Resolution Imaging Spectroradiometer (MODIS) data, and preserving geospatial data (Library of Congress 2008). Another current effort related to the preservation of digital data is Chronopolis, a collaboration of the San Diego Supercomputer Center, UCSD libraries, the National Center for Atmospheric Research, and the University of Maryland's Institute for Advanced Computer Studies. The focus of the project is on providing redundant data storage by creating a preservation "grid" for data. Two NDIIPP partners, the California Digital Library and ICPSR, will provide data for the project (University of California San Diego 2008).

In addition to the preservation of standalone data sets, the possibility for enabling access to the data that support a publication, more directly and in ways that go beyond simple citation to a static data set deposited separately in a repository, presents new challenges to preserving something as "traditional" as a journal article. While these enhanced publications offer new and exciting ways to communicate research findings, the act of reading such an article may also require more complex software and potentially complicates preservation efforts (Lynch 2007).

Finally, the problem of data preservation suggests a number of areas of debate and uncertainty that warrant further research:

- What types of data sets should be preserved? (see also the section on Appraisal and Selection in this report)
- What aspects or elements of data sets (broadly defined, including related software, documentation, etc.) can and should be preserved?

- At what point should the preservation process be initiated for data sets? Early in the research process? Upon publication of results?
- For how long should data sets be preserved?
- Who should be responsible for data preservation and who should pay for it? The funder, the researcher, or the institution? (see also the section on Financial Sustainability in this report)

In short, digital preservation is a dynamic and evolving field, with significant challenges specific to research data. CUL has already expressed a general commitment to providing stewardship for digital content; developing the capability to preserve research data is a logical extension of this commitment.

Intellectual Property

Data curation initiatives intended to foster e-science at Cornell must be cognizant of intellectual property issues. Different types of intellectual property may be at issue in a data deposit. They include:

- Copyright. Copyright protection does not extend to factual data. Compilations of data, however, can be copyrighted if the organizing principle behind the compilation is itself creative and original. Copyright protection can also extend to the person who recorded the data (a photographer, for example, or to a sound engineer who makes a recording) if they contribute something original to the recording.
- Licenses. While the data itself may not be protected by copyright, the creator and/or owner of the data may impose license conditions comparable to copyright rights when depositing the material with a data archive.
- Patents. Some of the data deposited in an e-science archive may be the basis for a patent application. It would be important to preserve the contextual metadata that establishes when and by whom the data was created. And some of the data could itself be patented – most notably genomic data.

Ownership of research data is a complex issue, even if we limit the data archive to the products of members of the Cornell community. In collaborative projects, especially those involving researchers at other institutions (the likely norm for e-science), ownership issues become very murky. As one prominent researcher recently noted in his summary of a meeting on the topic, "**no one** understood the legal aspects of data very clearly, **no one** could figure out an algorithm for when copyright applied and when it didn't, and **everyone** wanted a solution" (Wilbank 2007)

The default position for most academic research activity at Cornell is that any copyright created belongs to the faculty member who conducts the research. If the research is conducted by non-academic staff, copyright rests with the University. The University also claims an ownership interest in digital materials created with a substantial contribution of equipment or other resources from the University (Cornell University 1990). Note, however, that grant funding may stipulate copyright ownership that supersedes the default University policy. And if the data can be patented, then all intellectual property rights (including copyright) associated with the data belong to the University (Cornell University 2008).

The interactions between intellectual property and data archiving are many:

- Before research data can be added to a preservation and access system, it must be determined that the depositor is either the owner of the data or has the authority to make the donation.
- In order to preserve research data, it will be necessary for Cornell to replicate the data in archival and backup systems. In addition, it may be necessary to reformat or massage the data in new ways. This implicates the exclusive rights of the copyright owner to reproduce a copyrighted work and to prepare derivatives of that work.
- Data will be preserved for re-use. It is imperative that explicit permissions regarding the re-use of deposited data be secured prior to deposit. Any limitations on re-use of data, because of privacy concerns or the desire to protect trade secrets, should be clearly spelled out and enforced.

While the issues listed above may seem daunting, there are several elements that work in the favor of archiving data. Most importantly, as is noted above, there is no copyright in facts themselves. Much of the material in a data archive, therefore, will be copyright-free. Physical ownership of the data sets becomes less important in the academic environment that fosters the sharing of data and research results. In most cases a simple authorization from the data owner, similar to that used in the eCommons system, that Cornell can preserve donated data and that the depositor has the right to grant this authorization, will be enough.

More explicit explanation on how others may use the data would be desirable, however. Recently Science Commons created a Protocol for Implementing Open Access Data as a method of ensuring that scientific databases can be legally integrated with one another (Science Commons 2007). An Open Data Commons Public Domain Dedication and License has been created to implement the protocol, making it simple for researchers to dedicate their data to the public domain (Open Data Commons 2008). The Open Access to Knowledge (OAK) Law Project in Australia is also making great progress in explicating legal issues associated with data sharing (Fitzgerald et al. 2007). Initiatives such as these may make it easier to properly manage intellectual property rights in data archives, though the efficacy of these approaches is highly dependent on the social and communicative structures of disciplines (Burk 2006).

In sum, a Cornell data archiving system would have to be built with intellectual property issues in mind, but they should not be viewed as “show-stoppers.”

Confidentiality and Privacy

A long-standing concern in the collection and use of social science and medical data collected about individuals is how to protect respondent privacy and retain the data’s richness for statistical analysis. Advances in data mining techniques and computing technologies have exacerbated the issue with respect to public-use microdata products (data about individual respondents, entities, or events) and interest in restricted-use or confidential microdata.

Two strategies gaining acceptance among practitioners of quantitative analysis are creation of data enclaves for use of these confidential data products (either on-site, or using remote access technologies), and development of synthetic data products, such that the synthetic versions reflect the statistical properties of the “real” confidential data (Abowd 2004).

Researchers employing qualitative methods have additional challenges regarding disclosure risk. Although tools such as QualAnon (http://www.icpsr.umich.edu/DSDR/qualanon.html) can assist with anonymization of qualitative document files, audio, and video, such software does not address all of the challenges that researchers face. The legal and ethical risks posed by inadvertently sharing personally identifiable information, as well as the effort involved in attempting to mitigate such risks, can be a significant deterrent to data sharing.

Participation by Data Owners

Enabling new modes of research and new discoveries is perhaps the most significant benefit to sharing data; this is a benefit that extends to entire research communities and disciplines, and other potential users such as educators, policy makers, and the general public.

The decision to share data, however, rests with individual researchers, and a variety of factors may enter into a researcher's decision to deposit data. The cultural norm within a discipline may or may not operate to encourage data deposit. So-called "big science" research efforts, those organized around very costly instruments or research facilities that gather data for use by an entire community, tend to have data management plans and dedicated infrastructure from the outset, and there is a shared expectation that data will be available to the community. In some individual disciplines sharing of pre-publication materials in general is already relatively widely accepted; examples include the high energy physics community's early adoption of arXiv.org for sharing pre-prints, and bioinformatics disciplines, where deposit of sequence data in advance of related publications is often a requirement (Davis and Connolly 2007).

An important barrier to sharing data is simply the work involved. Researchers may not feel they have time, nor the knowledge or skills, for formatting and preparing data for deposit and preparing metadata, although local support appears to enhance the success of data sharing efforts (e.g. Karasti et al. 2006; Lord and Macdonald 2003). That research and development in documenting, sharing, and archiving data is ongoing, makes standards and best practices a moving target for data owners (Karasti et al. 2006).

Intellectual property and the protection of confidentiality are two additional issues that complicate data sharing. Researchers are understandably concerned with retaining data until they've fully exploited it for their own purposes. It can be difficult to anticipate future uses of data, complicating the decision as to when a researcher is "done" using their own data. This concern is even more acute for researchers whose work has the potential to be commercialized. Even when researchers are willing to share data, they may wish to be asked or notified when their data are used by others, and for what purpose (e.g. Helly et al. 2002). While some metadata standards include usage or rights statements where a data owner might specify such conditions for use, we are unaware of any existing data repositories that can enforce such conditions automatically, although some repositories such as ICPSR do support embargoed release of data. Privacy and confidentiality are discussed in the previous section; we note here only that lack of tools and expertise to ensure the protection of privacy and confidentiality.

There are some potential benefits to depositing data, although to researchers they may not outweigh the perceived costs or disadvantages. Such benefits include establishment of precedence in a research area, off-site back up of data, the potential for error-catching and subsequent correction and improvement of data as a result of re-use by others, and a convenient mechanism for sharing data with colleagues (Helly et al. 2002). Professional recognition for creating and sharing a data set is more complicated; current tenure and promotion processes focus more on publication of papers in high profile journals than data sets, and standardized methods of citing data sets (and particularly portions of a database) are only now in development.

The nature of a data center or repository itself can also influence a researcher's willingness to deposit data there. Quality and quantity of repository content, credibility and trustworthiness of the repository, curatorial services and quality control, access restrictions, and usability of the repository interface, and other repository characteristics may factor into a researcher's decision to deposit in a particular repository (Smith 2007). It's worth making special note of some of the hurdles institutional repository managers have encountered in trying to encourage deposit by faculty members. In addition to the issue of perceived effort, researchers often express a strong preference for depositing their works in subject repositories rather than institutional ones (Davis and Connolly 2007).

In addition to the benefits and barriers to data sharing discussed above, mandates to share or deposit data may affect researchers' decisions and behaviors. Some of the policies of several of the major US federal funders of research are described earlier in this report (see Section II). Note that there is considerable variation in existing policies, and that in some cases, a requirement for a researcher to share data may carry with it no specific requirement to deposit data in a publicly accessible repository. Individual institutions may have data retention policies in place, but the motivations behind these policies typically have little to do with data preservation or access; more often these policies are motivated by concerns related to compliance with policies and laws governing the responsible conduct of research.

V. Recommendations and Conclusions

The decision to undertake selected activities in relation to data curation has important strategic implications for Cornell and the Library. Provost Biddy Martin's February 6, 2008 address to the CUL academic assembly (Martin 2008) and Anne Kenney's (2007) address delivered to the all staff meetings held in November 2007 provide outlines of Cornell's and CUL's current high level goals and priorities. Some of these are directly supported by the recommendations that follow. Two goals from Provost Martin's address, which focused on the goals articulated in Cornell's consolidated planning document, are supported by undertaking data curation efforts within CUL. The first of these is Goal III: "Enable and encourage the faculty, their students and staff to lead in the preservation, discovery, transmission, and application of knowledge, creativity and critical thought." Exercising responsible stewardship of the outputs of research, including digital research data, would support that priority. Goal IV - "Extend our leadership in the use of research and education to serve the public good in fulfillment of Cornell's land-grant mission and its long-standing commitment to capacity building in communities in the United States and around the world" – is also supported by developing data curation services and infrastructure, by enabling the sharing and reuse of research data by members of other institutions and the public at large.

In support of the overarching goals for Cornell, Anne Kenney asserted that CUL is a logical leader to develop "an academic information infrastructure to support preservation, discovery, transmission, and application of knowledge, creativity and critical thought." Kenney described elements of such an infrastructure; several of these would be supported by providing services and infrastructure in support of data curation. These include "access to scholarly information resources at the point and place of need," "cutting-edge facilities and services to support research, learning, and scholarly communication across disciplines," and perhaps most importantly, "stewardship of the University's intellectual assets." Kenney also mentioned specific needs related to cyberscholarship, including the development of repositories in partnership with CIT and the CAC, and assuming responsibilities for preservation of digital content. More general goals that would also be served include expanding and enhancing research collaborations with faculty and supporting staff development to build on existing skills within CUL.

With these priorities in mind, the Data Working Group recommends activities in the following areas: providing services to Cornell researchers, developing local infrastructure and related policies in several areas, cultivating partnerships with other organizations, staff development, and revising the structure and charge of the Data Working Group.

1. Seek out and cultivate partnerships with other organizations.

The challenges of digital data curation are significant, and in many cases, may best be handled in cooperation with other organizations. The problem encompasses storage and network capacity, integration with cyberinfrastructure development, best practices related to data archival and metadata, user support, and more. Expertise and resources to address these component challenges are distributed among several units at Cornell. In addition, some data curation activities are most appropriately situated within or handled in cooperation with disciplinary communities or other institutions. These recommendations are offered in recognition of the fact that effective collaboration within Cornell and across institutions is essential to meeting the challenges of data curation.

1a. Idenitfy opportunities to collaborate with other Cornell data curation efforts – including the areas of usability and user support, metadata and discovery, intellectual property issues, digital preservation, and the curation of "medium-sized" data sets in the sense of the DISCOVER RSG proposal. Collaborators might include CAC, CIT, CIS, CISER, other departments or research groups, or other organizations. Also identify promising opportunities for involvement with domain-specific data curation initiatives, particularly in disciplines where either CUL or Cornell faculty have particular strengths.

1b. Actively participate in forthcoming initiatives, such as those described in the ARL Joint Task Force report on library support for e-science and libraries (2007), and other initiatives as they arise. Many of the ARL recommendations are focused on education and communication of new opportunities to the research library community. Specific ARL recommendations that might be of interest to CUL include:

- Establish an ARL e-science working group. CUL should take advantage of any opportunities for representation or participation in this group.
- Organize programs to explore e-science issues and share the programs with a broader audience. This included two suggested program types, both of which might be of interest. The first is a panel of "significant players" – Cornell is listed as one such player. The second is workshops for "self-selected teams" around a particular topic or project. One possible example would be to convene a workshop for librarians who are already attempting to work in this area to articulate some of the challenges it poses, and suggest development needs (such as for specific standards, software, policy and best practice documents, etc.).
- Among the recommendations for education and communication resources is one to create an inventory of discipline-based e-science centers and large-scale projects. This is something CUL could collaborate on, possibly using the VIVO architecture.

2. Provide services to Cornell researchers in several areas.
This group of recommendations is focused on providing services and information (as opposed to dedicated infrastructure) to support Cornell researchers (students, faculty, and staff), particularly with respect to existing or forthcoming data management, sharing or archival requirements. Maintaining credibility with researchers is important, and significant effort may be required to ensure or develop an adequate level of knowledge and skill to support these activities. Whenever possible, information – both educational resources and information on CUL services - should be made accessible on the Internet. This might be accomplished by developing new web-based resources, enhancing or expanding existing ones, directing users to other (external) existing resources, or some combination.

2a. Assist with the development of data management plans in grant proposals. As noted in Section II of this report, some funders (such as NSF) that do not already require the inclusion of data management plans in research proposals are likely to do so in the near future. For many researchers this will be new and unfamiliar territory; assistance is not only likely to be welcomed by them, but it also presents CUL with opportunities to engage researchers at the very start of a research project, enabling us to better forecast and prepare to meet data curation needs.

2b. Collect and provide information on best practices for data management, archival, and preservation. Researchers are accustomed to managing their own data for their own use, but the practices that may serve them well enough in those circumstances may not be adequate to meet requirements for submission to a data center. This area may lend itself to collaboration with other individuals or professional organizations; we are aware of at least two efforts to provide such guidance to faculty at other institutions. Ross Harvey (Charles Sturt University, Australia, and a contributor to the manuals on digital curation being produced by the Digital Curation Centre), and Anne Graham (MIT) both noted efforts at their institutions to create such guidelines (personal communications). Offering such guidance is practical to the extent that information is generalizable across disciplines; in other cases significant subject area expertise is required. While CUL has a number of subject experts who may be able to serve in this capacity, it may not be possible for the library to provide guidance in all subject areas.

2c. Educate researchers about intellectual property issues as they pertain to research data. Researchers are understandably concerned with issues such as ownership of data, and means of protecting their interests in the data they create (copyright, licensing, and attribution issues). This information might be added to the existing library website on scholarly communication (http://www.library.cornell.edu/scholarlycomm/), or the copyright information center (http://www.copyright.cornell.edu/), or be linked to those pages if it resides elsewhere.

2d. Refer researchers to appropriate resources dealing with the protection of confidential and private information. Institutional Review Board and HIPAA requirements prohibit the release of personally identifiable information, and researchers who collect such information in the course of their work may be faced with reconciling requirements to share data with the protection of confidential information. While the standards and means to meet such requirements are not yet fully developed, CUL staff should be cognizant of the issues involved. As tools and options for removing such information, or secure venues for sharing that fulfill requirements (such as the Cornell Restricted Access Data Center) become available, CUL staff should be prepared to refer researchers to them.

2e. Participate in the formulation of institutional policies on data retention. The Office of the Vice Provost for Research is in the process of developing a policy on the retention of research data. The Data Working Group was invited to comment on a draft of this policy, and CUL should remain involved in the development of this policy. When and if the policy is implemented, CUL should provide support for researchers in meeting the requirements of the policy.

3. Assess local needs and develop local infrastructure and related policies.
These recommendations suggest that CUL become more than an advisory partner in data curation by providing infrastructure in contexts that make sense within the university, collaborating with other units as much as possible.

3a. Examine the feasibility and desirability of creating a local repository infrastructure for depositing digital research data – whether it is eCommons, one dedicated to digital data only, or some combination. A digital data audit may be appropriate. Existing, completed surveys related to data may be available and should be reviewed. Collaboration with the DISCOVER RSG group's development of a cyberinfrastructure roadmap is also advisable.

3b. Further refine the guidelines presented in this report for the appraisal and selection of data to be curated locally, recognizing that individual projects might require different guidelines.

3c. Investigate roles for CUL with respect to the curation of data related to publications (this could be done in conjunction with a digital data audit). In light of the role of library in projects such as arXiv and Euclid, this is a natural extension of current activities.

3d. Continue to participate actively in the curation of "small science" data.

4. Cultivate a workforce capable of addressing the new challenges posed by data curation and cyberinfrastructure development.
Expanding current data curation and cyberinfrastructure activities and embarking on new ones will require investment in professional development for CUL staff, and possibly the creation of entirely new positions.

4a. Identify new skills that will be needed among CUL librarians to support activities in the area of data curation and cyberinfrastructure, as well as the positions where those skills would be needed and applied.

4b. Identify what new positions might be needed to support new CUL activities in the area of data curation and cyberinfrastructure.

4c. Make professional development activities available to CUL librarians – including reorganizing the Data Working Group as suggested below.

5. Form a Data Curation Executive Group and reorganize the Data Working Group.
One of the core purposes of the Data Working Group, articulated in its charge, was to educate itself in the areas of data curation and cyberinfrastructure, and to develop a set of recommendations for CUL. Having largely completed these tasks, two needs emerge:

advancing the recommendations made in this report, and continuing education for CUL staff. We recommend the following:

5a. Form a Data Curation Executive Group (DataExec) to advance the recommendations outlined in this report. Data curation and e-scholarship should become part of the portfolio of an AUL-level position, possibly the AUL for scholarly communication and collections, and that individual should sit on the DataExec. Other members might include the Chief Technology Strategist, and other individuals whose primary responsibilities include data curation.

5b. Reconstitute the DaWG as a smaller steering committee, operating much as the Metadata Working Group steering committee does. Their activities might include organizing DaWG forums open to all interested CUL staff (or Cornellians, for that matter), a journal club, or organization of workshops or training sessions for staff. There is the potential for significant overlap in interests between a re-formed Data Working Group and the Metadata Working Group, and it's worth considering whether there should be a DaWG liaison to the MWG, in order to coordinate efforts, such as joint sponsorship of visiting speakers.

5c. Publish this report in eCommons, and consider publishing an abbreviated version in D-Lib magazine or elsewhere. Some of the content will age quickly, so this should be acted on as soon as possible.

Conclusions

Curation of digital research data is, with a few exceptions, a new area of activity for CUL, as are activities related to the development of cyberinfrastructure. As should be evident from the introductory sections of this report, the pace of development is brisk, and organizations are working to identify and claim roles for themselves. Implementing some of the recommendations of the DaWG would require significant resources. Term funding from research grants is not likely to be sufficient to provide for long-term stewardship of research data; institutional commitments are needed. Nonetheless, our sense is that the time is right to articulate roles for CUL in this arena, and that we have a solid foundation of both new and established activities to demonstrate our competence in this area.

Appendix A. Definitions

Cyberinfrastructure – Cyberinfrastructure (CI) is the combination of hardware, software, networking and communication technologies, data, and human expertise that are required to enable data-driven research (Atkins et al. 2003).

Data - The report of the National Science Board (2005) describes data as any information stored in digital form, regardless of format. This report concerns itself primarily with digital data collected or generated in the course of conducting research. Data can further be classified as observational, experimental, and computational, derived, or processed data.

Data collections - The National Science Board report (2005) defines data collections not just as aggregations of data, but also the infrastructure and staff required to maintain and provide access to the data. The report further defines three levels of data collections, the distinction being based primarily on the size and breadth of the prospective user population. The report notes that a collection may evolve and move between categories over its lifetime.

- *Research data collections* are typically the result of one or more research projects, serve a specific group (usually immediate collaborators), are not often organized or documented to community standards, do not generally have support for curation or preservation, and may not be maintained beyond the life of the project(s) that generated them.
- *Resource (or community) data collections* serve a single-disciplinary community. They may drive the selection or creation of community standards for data. Budgetary support is intermediate and of variable duration.
- *Reference data collections* serve very broad segments of diverse research and education communities, use well-established standards, and tend to be financially well-supported over the long term.

Data curation - The Digital Curation Centre (http://www.dcc.ac.uk/about/what/) defines digital curation as the "activity of managing the use of data from its point of creation to ensure it is available for discovery and re-use in the future," and for the purposes of this report, this definition is appropriate. It is worth noting, however, that at times data curation is conflated with digital preservation, and sometimes with archiving. Both archiving and preservation may be thought of as subsets of activities related to data curation, as well as areas of research and professional practice in their own right (e.g. Lord and Macdonald 2003), but it's worth noting that these terms have different meanings to different communities (Beagrie 2006). In some contexts (the curation of NCBI databases, for example), curators add meaning to data by means of annotations or linking to related data, further adding to the possible confusion surrounding this term.

Data-driven research (data-driven science, data-driven scholarship, e-science, e-research, e-scholarship) - This set of terms refers to emerging research practices that are made possible by cyberinfrastructure. Specific examples or techniques include data and text mining, data visualization, advanced simulation modeling, and remote collaboration.

Medium-sized data sets – Medium-sized data sets, in the sense of the DISCOVER RSG proposal (see Section III of this report), are those data sets that are too large to be easily manipulated or downloaded in their entirety as a single digital data file, but too small to require high-end computational resources. Databases of moderate size, where a relational database structure is important to access and make use of the data, fall into this category.

Small science – We define small science to mean research endeavors undertaken by individuals or small to medium-sized groups. Data files tend to be small and can be used and manipulated on personal computers, and a single file may contain an entire data set (as opposed to storing large amounts of data in a database).

Appendix B. Data Curation Activities in Selected Disciplines – Beyond Cornell

There are significant, and in some cases, long-established activities or organizations supporting the curation of research data generated within specific disciplines. Selection of discipline-specific activities for description in this appendix was based primarily on the availability of librarians with relevant subject expertise, and this overview is by no means exhaustive. In addition to providing an overview of activities in the disciplines that follow, we attempt to note where there is participation by members of the Cornell community, including whether Cornell researchers are either consumers or producers of data.

Social Sciences

As a whole, the social science disciplines have been served by high-profile and long-standing data preservation efforts. Principal among these is the Inter-university Consortium for Political and Social Research (http://www.icpsr.umich.edu), and Cornell was one of the 21 charter members when ICPSR was organized in 1962. Its original purpose was to collect, archive, and distribute quantitative data for secondary research. Its upcoming strategic plan champions support for qualitative and mixed-methods research data. It has taken the lead in promoting best practices for data preservation and description standards.

Since 1968, the National Archives and Records Administration, specifically its Electronic and Special Media Records Services Division (http://www.archives.gov/research/formats/electronic-records.html) has been at the forefront of efforts to migrate and preserve data products generated by federal statistical and information agencies. In this country, many research institutions inaugurated data archives in the late 1970s and early 1980s. Their focus was to provide a service environment to match data and needs of their clientele, although a few (such as Wisconsin) have long histories of archiving data generated by researchers on their campuses. The University of Wisconsin's Data and Program Library Service was begun in 1966, and as early as 1980, it archived data files collected by Wisconsin researchers, including (but not limited to) National Science Foundation grant projects. Until 2007, DPLS was funded by the College of Letters and Science. In 2007, DPLS merged with data services provided by the Center for Demography and Ecology and the Center for Demography of Health and Aging, forming the Data and Information Services Center (DISC).

There have been a large number of multi-institutional (and as described later, multi-national) collaborative efforts to locate, preserve, and provide information about data products. The Data-PASS collaboration, organized in 2005 and funded by the Library of Congress, seeks to identify, acquire, describe, and preserve social science research data, with emphasis on at-risk datasets (http://www.icpsr.umich.edu/DATAPASS/). The Harvard/MIT Data Center, a Data-PASS participant, incorporates the holdings of the Murray Research Archive, a significant repository supporting both qualitative and quantitative research (https://www.hmdc.harvard.edu). Harvard/MIT is also home to the Dataverse Network Project (formerly the Virtual Data Center), a customizable data discovery application that transcends physical data depositories (http://thedata.org). Dataverse has been implemented in a variety of contexts, including a small number of journal replication archives.

Historically, these trends have parallels in other countries, particularly in Western Europe. For over forty years, the UK Data Archive has collected, preserved, and disseminated data in social sciences and humanities disciplines (http://www.data-archive.ac.uk/). The Steinmetz Archive in the Netherlands (since absorbed into DANS http://www.dans.knaw.nl/) was founded in 1962 to provide wider dissemination of social science research data to the public. The Council of European Social Science Data Archives is a collaborative organization of European social science data archives and has recently inaugurated CESSDA Portal to identify (and in some cases, analyze) datasets held by its members (http://www.nsd.uib.no/cessda/home.html).

As one might expect, data description standards have been numerous and evolved over time. The Data Documentation Initiative (http://www.icpsr.umich.edu/DDI/) has emerged as the standard (or at least one that is widely recognized), having been adopted by significant data distributors (including ICPSR and CESSDA members). DDI is supported by a variety of utilities and toolkits. Final release of version 3.0 is due Spring 2008.

Bioinformatics

In the United States, the National Center for Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov/) at the National Library of Medicine, part of the National Institutes for Health, is the primary bioinformatics data repository. NCBI's GenBank database is an archival database of publicly available primary sequence data. GenBank exchanges data daily with its two partners in the International Nucleotide Sequence Database Collaboration (INSDC): the European Bioinformatics Institute (EBI) of the European Molecular Biology Laboratory (EMBL), and the DNA Data Bank of Japan (DDBJ). Nearly all sequence data are deposited into INSDC databases by the labs that generate the sequences, in part because journal publishers generally require deposition prior to publication so that an accession number can be included in the paper. Over 2600 papers by Cornell authors indexed in PubMed include links to data deposited in GenBank.

The bulk of data curation at NCBI comes from RefSeq. A portion of the RefSeq dataset, data related to viral, vertebrate, and some invertebrate organisms, is supported by NCBI curation staff. Most bacterial, plant, and fungal records are provided either by collaboration or by processing the annotated genome data submitted to GenBank; however, a small number of bacterial genomes are annotated and curated by NCBI staff. Viral genomes are curated in consultation with a viral board of advisors. The RefSeq database is the result of data extraction from GenBank, curation, and computation, combined with extensive collaboration with authoritative groups. Each molecule is annotated as accurately as possible with the organism name, strain (or breed, ecotype, cultivar, or isolate), gene symbol for that organism, and informative protein name. Collaborations with authoritative groups outside of NCBI provide a variety of information ranging from curated sequence data, nomenclature, feature annotations, and links to external organism-specific resources. When no collaboration has been established, the NCBI staff assembles the data from GenBank. Each record has a comment, indicating the level of curation that it has received and attribution of the collaborating group.

Many of the other data repositories at NCBI allow for private data deposit and embargoed public release. NCBI also repurposes and makes links between their data and data drawn from other major repositories, such as the Protein Data Bank. The Worldwide Protein Data Bank (wwPDB) consists of organizations that act as deposition, data processing and distribution centers for PDB data. The mission of the wwPDB is to maintain a single Protein Data Bank Archive of macromolecular structural data that is freely and publicly available to the global community. The founding members are RCSB PDB (USA), MSD-EBI (Europe) and PDBj (Japan). The Research Collaboratory for Structural Bioinformatics (RCSB) (http://home.rcsb.org/) is a non-profit consortium of Rutgers, University of California, San Diego, and the University of

Wisconsin-Madison dedicated to improving understanding of the function of biological systems through the study of the 3-D structure of biological macromolecules.

Chemistry

There are two general types of users of chemical data- practicing chemists (and scientists in related disciplines) and cheminformaticians. Chemical data itself has multiple facets- there is a large variety of complex data types and much of it is collected experimentally in numerical and textual form, much of it derived from theoretical calculations; but much of what is useful to both user types is the information derived from the data and combined and presented in a myriad of ways ultimately to characterize chemical compounds, their complex relationship to the natural world and their potential utility to humanity. Historically, and even now, much of this information is first distributed in the published literature, which has an exceedingly high potential value with regard to its depth and richness, and also economic potential. This has prompted a number of approaches to storing, manipulating, processing, controlling and mining the information and data behind it. Practicing chemists rely on the resulting databases to inform their chemical research, from chemical foundation to experimental methodology to establishing priority. Cheminformaticians mine the data and literature sources for new types of information and new combinations of data to identify chemical compounds of potential value for further study by practicing chemists. The strong need of both groups for rigorous methods of storage, organizing and presenting chemical data and information has resulted in a number of proprietary and open approaches to these issues for well over 100 years, the number of organizations involved and styles of approach almost as complex as the data itself.

Chemical databases and data sources can be classified into several categories based on general purpose, including: literature databases, factual databases, structure and reaction databases, molecular biology databases and numerous less formal instances of Internet-based data storage and sharing, including wiki based lab notebooks, interactive blogs and Web 2.0 communities. The curation of databases in each category varies considerably from large, long-standing publishers with subscription based products to communities of sub-disciplines collectively managing open deposition and content to individual research labs with wikis and blogs. Patents are another large source of chemical data, openly accessible on the Internet for many nationalities. The large corporate R&D sector in chemistry considers chemical data and information produced in-house as valuable intellectual property and companies maintain highly secure intranets of their data. The cheminformatics branch shares much of its approach with the bioinformatics community, utilizing many of the same data sources, methodology and tools; complementing the large biological molecule focus of bioinformatics with similar approaches to small molecules that play an increasing role in society especially in the formulation of drugs and agricultural chemicals. Another closely related area from the physics and computer sciences angle is computational chemistry which is primarily concerned with theoretical calculations but shares computational methodologies with cheminformatics. One starting place for several open databases used by cheminformaticians is hosted by http://cheminformatics.org/. Societies involved in chemical data and information issues include the American Chemical Society Divisions of Chemical Information (http://www.acscinf.org/) and Computers in Chemistry (http://membership.acs.org/c/Comp/), the Royal Society of Chemistry (http://www.rsc.org/Publishing/Journals/ProjectProspect/index.asp), and the Cheminformatics and QSAR Society (http://www.ndsu.edu/qsar_soc/), among others.

Background for this summary – including references to specific databases, data sources and cheminformatics initiatives – is more fully represented in T. Engel's (2006) extensive review article.

Astronomy

The study of astronomy is based on observation, collection, interpretation and mining of immense catalogs of data from ground- and space-based telescopes. Since the advent of digital data storage it has been critical for the astronomy research community to develop strategies to manage, distribute, integrate, and archive rapidly expanding data collections. Discussions of data storage limitations have continuously haunted the astronomy literature and conference proceedings moving more and more towards collaborative initiatives to enable astronomers to find, retrieve, and analyze terabytes of observational data. One current approach is through the International Virtual Observatory Alliance (http://www.ivoa.net/) which integrates astronomical archives and computing applications from over 16 nationalities into an interoperable system available worldwide. The focus of the alliance is primarily on the continued development and refinement of standards to promote the usability and preservation of data. In the US, NSF and NASA currently fund the Virtual Astronomical Observatory (http://us-vo.org/) which "provide access to data sets, create and maintain data protocols and standards, and provide analysis tools and services to the astronomical research and educational community" This initiative is the successor to the NASA Astronomical Data Center (http://adc.gsfc.nasa.gov/) which archived and distributed data collections for 25 years until it closed in 2002; the content is still available online.

Geospatial Data

Geospatial data are used across a broad range of disciplines, but because there is some coherence in the approaches used to document and share geospatial data, we include a description of selected activities in this section on domain-specific efforts.

The United States has several federal agencies that produce enormous quantities of geospatial data, most prominently the United States Geological Survey (USGS), the National Aeronautics and Space Administration (NASA), and the National Oceanic and Atmospheric Administration (NOAA). Although these organizations have always managed large quantities of data, in recent years they have become increasingly adept at efficiently distributing their data online to the public. Many of these datasets are commonly used as base layers for maps and analysis.

USGS produces several major datasets that cover nearly the entire country, including the National Elevation Dataset, the National Hydrography Dataset, the National Land Cover Dataset, Digital Orthophoto Quadrangles, Digital Raster Graphics (scanned images of the famous USGS topographic maps), and Digital Line Graphs (vector data derived from the topographic maps). Seamless delivery is now available for many of its datasets, meaning that users can download data for any arbitrary region of interest, limited only by the total download size (http://seamless.usgs.gov/).

NASA operates satellites that continuously collect enormous quantities of remote-sensing data. Raw data is then analyzed and processed to create derivative data, such as land surface temperature, vegetation indices, or burned areas. Through the Land Process Distributed Active

Archive Center (LP DAAC, http://lpdaac.usgs.gov/), users can access a variety of raw and processed data, which may be routinely collected or on-demand, current or archival, and may incur fees or not, depending on the specific data product. Other satellite data, some dating back to 1959, is now distributed from USGS Earth Resources Observation and Science (EROS, http://edc.usgs.gov/). These are all large, raster-based datasets that provide snapshots of specific places and times. NASA also runs the Shuttle Radar Topography Mission (SRTM), which has collected precise elevation data for most of the globe.

NOAA administers the National Data Centers for Climate, Geophysics, Oceans, and Coasts. Each of these collects and distributes large quantities of data from a combination of satellite- and ground-based instruments.

There are many other, smaller organizations and research projects that create geospatial data, and it would be impossible to list them all. But it is worth noting that some Cornell researchers are involved in these efforts. For example, the Cornell Institute for Resource Information Sciences (IRIS) does work for the New York state Agricultural Districts Mapping program; several Cornell researchers were involved with the New York portion of the national Gap Analysis Program (http://gapanalysis.nbii.gov/); and Steve DeGloria (faculty, Crop and Soil Sciences) is currently working with the nascent National Geospatial Development Center (http://ngdc.wvu.edu/).

Beyond these data-generating organizations, there are some notable efforts to collect and organize geospatial datasets from a variety of sources. There are many state-based portals, but the quantity and quality of data varies widely. The National Atlas (http://nationalatlas.gov/) originally started in 1874 as a printed volume of maps, but now distributes maps and data online, organizing datasets from over 20 federal agencies into thematic groups such as the "environment,” "history,” "people,” or "water". A younger effort, Geospatial One-Stop (GOS, http://geodata.gov/), does not host any data directly, but instead harvests metadata from existing geospatial repositories within the United States, and provides a searchable interface into this metadata.

Internationally, the Food and Agriculture Organization (FAO) of the United Nations (UN) has developed the open-source GeoNetwork software (http://www.fao.org/geonetwork/), which it uses to organize geospatial datasets of interest to various UN agencies and non-governmental organizations working around the world.

Both GOS and GeoNetwork rely on standards-compliant metadata (either FGDC or ISO-19115), and the quality of their portals is dependent upon the quality of the underlying metadata. Both are also starting to focus on web services, which allow users to link dynamically to remote datasets, rather than downloading a complete copy of the data to a local computer. In the future (at least in certain environments) we are also likely to see greater use of transactional web services, where users can submit edits to a remote dataset, using a standard protocol that can be implemented across different systems.

Biodiversity Informatics

Biodiversity informatics efforts are concerned with bringing together information on the collections of natural history museums worldwide, as well as other recorded observational data, and making it available online. Most existing specimen collection catalogs had analog origins, and contemporary catalogs are generally not interoperable. Global, online availability of specimen information in catalogs would make it easier for investigators to examine questions about species distributions, past and present, and to make use of the data in modeling efforts aimed at addressing issues such as the impact of global climate change. Networked species collections are often organized around a particular taxonomic group; by way of example, the Cornell Museum of the Vertebrates collections are included in ORNIS (birds, http://olla.berkeley.edu/ornisnet/), HerpNET (reptiles and amphibians, http://www.herpnet.org/), FishNet (fish, http://www.fishnet2.net/index.html), and MaNIS (mammals, http://manisnet.org/).

The Global Biodiversity Information Facility (GBIF, http://www.gbif.org/), a consortium of national governments and international organizations promoting development of a global biodiversity information infrastructure by supporting projects in the areas of database interoperability, digitization of natural history collections, compilation of species name information, and outreach and capacity building, represents an important effort in this area. There are also many other active groups, including the Biodiversity Information Standards group (formerly known as the Taxonomic Database Working Group, TDWG), major natural history museums such as the Smithsonian and the Missouri Botanical Garden, and academic research centers such as the bioinformatics group at the University of Kansas' Natural History Museum.

Some standards and tools in use include the protocol for Distributed Generic Information Retrieval (DiGIR, http://www.specifysoftware.org/Informatics/informaticsdigir/), a protocol for providing federated searching of natural history collections, and two data specifications for encoding information about museum specimens and species occurrences: Darwin Core (http://wiki.tdwg.org/twiki/bin/view/DarwinCore/WebHome), and Access to Biological Collections Data (ABCD, http://wiki.tdwg.org/ABCD/). An important issue in documenting species information is species name authority; the Integrated Taxonomic Information System (ITIS, http://itis.gov/) is one source of this information.

Ecological and Environmental Informatics

Two prominent sets of efforts aimed at making ecological and environmental data available are the collaborative projects led by the ecoinformatics group at the National Center for Ecological Analysis and Synthesis (NCEAS, http://www.nceas.ucsb.edu/) working in partnership with the Long-Term Ecological Research Network (LTER, http://www.lternet.edu/), and the National Biological Information Infrastructure (NBII), managed by the United States Geological Survey.

NCEAS and its collaborators have developed several important tools and protocols that are in use for describing, distributing, and analyzing ecological information. Many of these are deployed as part of the Knowledge Network for Biocomplexity (KNB, http://knb.ecoinformatics.org/index.jsp), a national network intended to facilitate the discovery and use of ecological data. As of this writing, Cornell researchers have contributed 49 data sets to the KNB, primarily associated with the Hubbard Brook Experimental Forest, an LTER site in New Hampshire; Cornell researchers have been associated with Hubbard Brook since well before it joined the LTER Network. The remaining Cornell data sets in the KNB are associated with a research group working on nutrient and sediment cycling in the Upper Susquehanna River Basin (see section III of this report). KNB developments include Ecological Metadata Language (EML) which was initially developed to support the data documentation needs of LTER researchers, but increasingly, unaffiliated researchers and projects also make use of it. Metacat, a data and metadata database was developed for use as a data storage system (Berkley et al. 2001), and Morpho is a client that serves as both a metadata editor and an interface for searching and submitting data to Metacat installations (Higgins et al. 2002). Kepler is an open-source scientific workflow application that evolved from the Ptolemy project allows researchers to retrieve, transform, and analyze data, in a documented and reproducible fashion (Michener et al. 2005).

The National Biological Information Infrastructure (NBII, http://www.nbii.gov/) is intended to provide comprehensive information of all kinds on the United State's biological resources, including biological databases, publications, organizations, and more. More than a dozen geographic and thematic nodes comprise the network. The NBII maintains a metadata clearinghouse (http://mercury.ornl.gov/nbii/), hosted by the Oak Ridge National Laboratory; data sets from that Cornell Laboratory of Ornithology are listed, as are others attributed to Cornell researchers (note that the NBII harvests KNB metadata, so there is some duplication). Important development efforts in which the NBII has played a role include the Biological Data Profile, a profile to the FGDC-Content Standard for Digital Geospatial Metadata (CSDGM) intended for use with biological data whether it is geospatial or not, and the Biocomplexity Thesaurus (http://thesaurus.nbii.gov/), developed initially in partnership with Cambridge Scientific Abstracts.

Data and Publications in Ecology and Evolutionary Biology

The Ecological Society of America (ESA) maintains a data registry (http://data.esa.org/) for data sets associated with papers published in the society's journals, but as of this writing it only contains 15 data sets. The ESA's Ecological Archives (http://esapubs.org/archive/) also publishes materials in support of ESA publications; Cornell researchers have published supplemental materials to the archives. The ESA publishes peer-reviewed data papers, with abstracts published in printed journals. The society has also organized workshops on data sharing, including representatives from many other professional societies in allied disciplines (http://www.esa.org/science_resources/datasharing.php). In spite of some publisher and professional society, support for the publication of data sets with articles, the issue is a contentious one (Cassey and Blackburn 2007; Parr 2007)

A newer effort in this area is DRIADE, the Digital Repository of Information and Data for Evolution (DRIADE, http://ils.unc.edu/mrc/driade), a repository in development for hosting underlying data for journal articles in evolutionary biology publications. This is a joint project of the National Evolutionary Synthesis Center in Durham, NC, and the Metadata Research Center at the University of North Carolina – Chapel Hill, and has the support of the editors of many of the major journals in the field. CUL librarians Oya Rieger and Gail Steinhart participated in a May 2007 DRIADE workshop entitled "Digital data preservation, sharing, and discovery: Challenges for Small Science Communities in the Digital Era".

Appendix C. Other Data Collection and Curation Activities at Cornell – Disciplines and Independent Collections

This section describes selected activities within disciplines at Cornell, as well as individual data collections that are not managed by one of the Cornell units with substantial data curation activities. This section is meant to illustrate the diversity of data collections on campus, and is not comprehensive; the personal data collections of individual researchers are not described here, and represent possibly the bulk of the research data generated at Cornell. Some of the activities described here are centered around shared research facilities that generate data, and may or may not have a common data sharing or archival infrastructure. Some of these independent collections are managed primarily for the use of an active research group ("research data collections"; see Appendix A for definitions); others are managed with the intent of providing public access and may have a broader range of contributors and/or users ("resource" and "reference" data collections). We've made no attempt to classify the collections described here.

Disciplines at Cornell

Physics

Cornell has a strong research program in high-energy accelerator physics, long supported by the Cornell Electron Storage Ring (CESR). As the field has advanced, the facility has seen many upgrades; the current program is CLEO-c (http://www.lepp.cornell.edu/Research/EPP/CLEO) and is dedicated to research in charm physics. The CLEO grant is in its final year and currently there is no plan for preserving these data, nor is there any other place in the world to acquire similar data. Two new proposals are underway for the next generation CESR: one for a test accelerator for instrumentation development; the other for the development of an Energy Recovery Linac, an ultra-high brilliance x-ray source. These would upgrade the companion x-ray program at CESR, the Cornell High Energy Synchrotron Source or CHESS (http://www.chess.cornell.edu/). CHESS is funded by NSF and utilizes the synchrotron radiation from CESR as a high-intensity X-ray source for a wide variety of research projects in physics, chemistry, biology, environmental science and materials science, with users from universities, national laboratories and industry across the United States. According to the CHESS website, "each year, 500-600 scientists, graduate and undergraduate students visit CHESS to collect data." The facility generates a large amount of data and a variety of data analysis tools are provided on-site, but individual research groups are responsible for their own data storage and preservation once their work at the facility is complete. MacCHESS, funded by NIH, is a macromolecular diffraction facility at CHESS that supports protein crystallographic studies. MacCHESS data become the user's responsibility and may be transported to the researcher's home lab via a network connection or user-supplied media.

Chemistry

The Cornell approach to chemical data and information is varied and there are a number of research and computing facilities maintained by Cornell that generate and maintain data for both local and international use. These include the Northeastern Collaborative Access Team (NE-CAT), which operates a crystallographic research facility at the Advanced Photon Source at Argonne National Laboratory (http://necat.chem.cornell.edu/), and The National Biomedical

Research Center for AdvanCed Electron Spin Resonance Technology (ACERT, http://www.acert.cornell.edu/). Work stations with commonly used software are available for data processing at these facilities but users must bring their own external disk drives to archive their data. Additional analytical facilities for the local Cornell community include the X-Ray Diffraction Facility, the NMR facility, and the Chemistry Research Computing Facility that maintains research group based computer clusters for calculating molecular simulations. There are no current plans for local data storage.

Data Collections at Cornell

Sol Genomics Network (SGN) - http://www.sgn.cornell.edu/

The SGN contains genomic, genetic and taxonomic information for species in the Euasterid clade, which includes several agronomically important species such as tomato, potato, tobacco, eggplant, pepper, and the ornamental *Petunia hybrida*. Data include expressed sequence tags (ESTs), unigenes, genetic maps and markers, and other tomato gene sequencing information. Data are contributed from researchers around the world. Funding is provided by the NSF, the USDA, Nestle Corporation, and the Binational Agricultural Research and Development Fund. The Computational Biology Service Unit at Cornell also provides support by making computing clusters available.

Spacecraft Planetary Imaging Facility (SPIF) - http://astrosun2.astro.cornell.edu/facilities/SPIF.php

In addition to activities described in the main body of this report (see Section II, Center for Advanced Computing), the Department of Astronomy jointly sponsors and operates (with NASA) the SPIF, one of several Regional Planetary Image Facilities. SPIF maintains a collection of planetary image data and associated information, including maps and mission documentation. The collection is comprised of over 100,000 images resulting from the U.S. planetary exploration program, including the most recent images obtained by the Mars Rovers. Images are stored as hard copy or on digital media, and the SPIF houses computers and software for image display and processing, as well as equipment for viewing analog images.

Network for Earthquake Engineering Simulation (NEES) - http://nees.cornell.edu/ and http://nees.org/

Cornell is one of 15 partners in NEES, a national, networked earthquake simulation resource. The Cornell Large Displacement Soil-Structure Interaction Facility is used for experimental testing, evaluation, and analysis of soil-structure-foundation interaction in critical lifeline facilities, as well as the seismic performance of highly ductile above-ground structures. Cornell's facility supports networked data acquisition and is connected to the NEES data network, hosted at the San Diego Supercomputer Center. The NEES consortium is funded by NSF for 2005-2014.

National Data Archive on Child Abuse and Neglect (NDACAN) - http://www.ndacan.cornell.edu/

A project of the Family Life Development Center in the College of Human Ecology, the mission of NDACAN is to facilitate the secondary analysis of research data relevant to the study of child abuse and neglect. NDACAN acquires microdata from leading researchers and national data collection efforts and makes these datasets available to the research community, along with free user support. NDACAN evaluates the archiving suitability and feasibility of submissions on a case-by-case basis, and requires adequate documentation with submissions. Recipients of research funds provided through the Child Abuse and Prevention Treatment Act (CAPTA) are required to archive their data with NDACAN.

Insect Flight Database

In collaboration with the CAC, the Physics Department's complex fluids group is creating a database to share video, data, analytical tools for the study of insect flight. The project entails coupling state-of-the-art experimental and numerical techniques with the computational capacity of the CAC to collect, archive and analyze insect flight data. Physicist Itai Cohen has developed a state of the art visualization facility, with fully-automated data collection, allowing for a dramatic increase in the number of videos obtained for a given flight maneuver in a particular species of insect. Theoretical and Applied Mechanics Professor Jane Wang, an expert in unsteady aerodynamics, and her group are developing advanced computational tools for analyzing insect flight.

Additional projects and collections

Many data collection and distribution activities are initiated with grant funding, and receive little or no further support once the original funding period ends. There are several examples of data collections that have or may soon be "orphaned" in this manner:

- *Data and Story Library* (DASL, http://lib.stat.cmu.edu/DASL/DataArchive.html) is an online library of data files and stories intended for educational use by statistics instructors for the preparation of lecture materials and assessment tools, or by students wishing to engage in individual study of statistical issues. Developed by Paul Velleman (ILR), DASL was funded by the NSF from 1992-1996. No new content has been added since then, although the site is still currently used, with copies at both Cornell and Carnegie Mellon.
- *Solid Earth Information System* (SEIS, http://atlas.geo.cornell.edu/nsdl/nsdl.html) – The aim of the SEIS was to compile comprehensive GIS data sets for the National Science Digital Library (NSDL), and to make them and related tools for access and analysis available online. Coverage was intended to be global, with current data sets available at scales ranging from regional to global. Future funding for this effort is uncertain.

References

Abowd, John, and Julia Lane. 2004. New approaches to confidentiality protection: Synthetic data, remote access and research data centers. In *Privacy in Statistical Databases*. ed. J. Domingo-Ferrer and V. Torra. Berlin: Springer-Verlag, 2004, pp. 282-289.

ACLS Commission on Cyberinfrastructure for the Humanities and Social Sciences. 2006. *Our cultural commonwealth: The report of the ACLS commission on cyberinfrastructure for the humanities and social sciences.* New York, NY: American Council of Learned Societies. http://www.acls.org/cyberinfrastructure/index.htm (Accessed 12/19/2006).

ANDS Technical Working Group. 2007. *Towards the Australian data commons: A proposal for an Australian national data service.* Canberra: Australian Government - Department of Education, Science and Training. http://www.pfc.org.au/twiki/pub/Main/Data/TowardstheAustralianDataCommons.pdf (Accessed 01/31/2008).

ARL Workshop on New Collaborative Relationships: The Role of Academic Libraries in the Digital Data Universe. 2006. To stand the test of time: Long-term stewardship of digital data sets in science and engineering.

Arms, William Y., and Ronald L. Larsen. 2007. *The future of scholarly communication: Building the infrastructure for cyberscholarship. report of a workshop held in phoenix, Arizona April 17-19, 2007.* National Science Foundation and the Joint Information Systems Committee. http://www.sis.pitt.edu/~repwkshop/NSF-JISC-report.pdf (Accessed 10/08/2007).

Association of Research Libraries. 2007. ARL Appoints University of Washington's Neil Rambo as ARL Visiting Program Officer. Association of Research Libraries, Washington, DC. http://www.arl.org/news/pr/neilrambo.shtml (Accessed 1/31/2008).

Atkins, D. E., K. K. Droegemeier, S. I. Feldman, H. Garcia-Molina, M. L. Klein, and D. G. Messerschmitt. 2003. *Revolutionizing science and engineering through cyberinfrastructure: Report of the national science foundation blue-ribbon advisory panel on cyberinfrastructure.*

Austin, Tony and Jenny Mitcham. 2007. Preservation and management strategies for exceptionally large data formats: 'Big data'. EH Project No: 3984, Arts and Humanities Data Service - Archaeology Data Service. http://ads.ahds.ac.uk/project/bigdata/. (Accessed 04/18/2008).

Beagrie, N. 2006. Digital curation for science, digital libraries, and individuals. *International Journal of Digital Curation* 1(0). http://www.ijdc.net/ijdc/article/view/6 (Accessed 02/05/2007).

Berkley, C., M. Jones, J. Bojilova, and D. Higgins. 2001. Metacat: A schema-independent XML database system. Paper presented at 13th International Conference on Scientific and Statistical Database Management (SSDBM 2001), Jul 18-20 2001.

Blue Ribbon Task Force on Sustainable Digital Preservation and Access. 2008. Blue ribbon task force on sustainable digital preservation and access. http://blueribbontaskforce.sdsc.edu/ (Accessed 04/01/2008).

Borgman, Christine L. January 2008. Data, disciplines, and scholarly publishing. *Learned Publishing* 21(10):29-38.

Burk, Dan. 2006. Intellectual Property in the Context of E-Science (August 18, 2006). Minnesota Legal Studies Research Paper No. 06-47. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=929479. (Accessed 04/11/08).

Butler, Declan. 2007. Agencies join forces to share data. *Nature* 446(713) 354.

Carlson, Scott. 2006. Lost in a sea of science data. *Chronicle of Higher Education* 52, (42): A35-7.

Cassey, Phillip, and Tim M. Blackburn. Reproducibility and repeatability in ecology. *Bioscience* 56(1): 958-9.

Christophides, Vassilis, and Peter Buneman. 2007. Report on the first international workshop on Database Preservation (PresDB'07). SIGMOD Record 36(3): 55-58.

Clarke, Roger. 2004. Open source software and open content as models for eBusiness. Paper presented at the 17th International eCommerce Conference (Slovenia). http://www.anu.edu/people/Roger.Clarke/EC/Bled04.html. (Accessed 04/01/2008).

CODATA. 2008. CODATA: Committee on data for science and technology. http://www.codata.org/ (Accessed 04/01/2008).

Cook, Terry. 2008. Beyond the screen: the records continuum and archival cultural heritage. Australian Society of Archivists Conference, Melbourne, Australia.

Cornell University. 1990. Cornell University Copyright Policy. Cornell University Copyright Policy. http://www.policy.cornell.edu/cm_images/uploads/pol/Copyright.html. (Accessed 04/11/2008).

Cornell University. 2008. Policy 1.5, Inventions and Related Property Rights. http://www.policy.cornell.edu/vol1_5.cfm. (Accessed 04/11/2008).

Cornell University Center for Advanced Computing. n.d.a Estimating life's diversity on land and at sea. http://www.cac.cornell.edu/about/studies/Bunge.pdf (Accessed 04/01/2008).

———. n.d.b Improving weather data accuracy and accessibility. http://www.cac.cornell.edu/about/studies/NRCC.pdf (Accessed 04/01/2008).

———. n.d.c Searching for pulsars in very-large databases. http://www.cac.cornell.edu/about/studies/Arecibo.pdf (Accessed 04/01/2008).

Cornell University Library. 2005. Cornell University Library Digital Preservation Framework. http://commondepository.library.cornell.edu/docs/cul-dp-framework-0405_main.pdf (Accessed 04/18/2007).

———. 2007. Cornell University Library Registry of Digital Collections. http://rdc.library.cornell.edu/. (Accessed 04/01/2008).

Cornell University Office of Information Technologies. 2007. Business Plan for Fiscal Year 2008 and Five-year Strategic Goals. http://www.cit.cornell.edu/oit/strategicplan.pdf. (Accessed 04/04/2008).

Davis, Philip M., and Matthew J. L. Connolly. 2007. Institutional repositories: Evaluating the reasons for non-use of Cornell University's installation of DSpace. *D-Lib Magazine* 13, (3/4) (Apr). http://www.dlib.org/dlib/march07/davis/03davis.html (Accessed 04/01/2008).

Digital Preservation Coalition. 2005. Report for the DCC/DPC Workshop on Cost Models for Preserving Digital Assets. http://www.dpconline.org/graphics/events/050726workshop.html (Accessed 3/24/2008).

Dirks, Lee and Savas Parastatidis. 2008. Research-Output Repositories - An Overview of Microsoft Initiatives. OR2008 Publications. Third International Conference on Open Repositories, University of Southampton, Southampton, UK. http://pubs.or08.ecs.soton.ac.uk/84/ (Accessed 04/04/2008).

Division of Ocean Sciences, National Science Foundation. 2003. Division of ocean sciences data and sample policy. NSF 04-004. http://www.nsf.gov/publications/pub_summ.jsp?ods_key=nsf04004 (Accessed 04/04/2008)

DRIVER. 2008. DRIVER: Networking European scientific repositories. http://www.driver-repository.eu/ (Accessed 04/01/2008).

Engel, T. 2006. Basic overview of cheminformatics. *J. Chem Inf. Model.* 46: 2267-2277.

Entlich, Richard, and Ellie Buckley. 2006. Digging up Bits of the Past: Hands-on with Obsolescence. RLG Diginews 10(5). http://digitalarchive.oclc.org/da/ViewObjectMain.jsp?fileid=0000070513:000006282602&reqid=8139#article1. (Accessed 04/18/2008).

Fedora Commons. 2008. Sun and Fedora Introduce a Petabyte-scale Object Store. http://www.fedora-commons.org/about/news.php#petabyte (Accessed 04/04/2008).

Fitzgerald, Anne, OAK Law Project, Legal Framework for e-Research Project. , Dept. of Education, Science, and Training., and Kylie Pappalardo. 2007. Building the Infrastructure for Data Access and Reuse in Collaborative Research: An Analysis of the Legal Context /

Pappalardo, Kylie. Brisbane, Qld. : Open Access to Knowledge (OAK) Law Project : Legal Framework for e-Research Project.

Glover, David M., Cynthia L. Chandler, Scott C. Doney, Ken O. Buesseler, George Heimerdinger, J. K. B. Bishop, and Glenn R. Flierl. 2006. The US JGOFS data management experience. *Deep-Sea Research Part II: Topical Studies in Oceanography* 53(5-7): 793-802.

Gold, Anna. 2007a. Cyberinfrastructure, data, and libraries, part 1: A cyberinfrastructure primer for librarians. *D-Lib Magazine* 13, (9/10) http://www.dlib.org/dlib/september07/gold/09gold-pt1.html (Accessed 09/18/2007).

———. 2007b. Cyberinfrastructure, data, and libraries, part 2: Libraries and the data challenge: Roles and actions for libraries. *D-Lib Magazine* 13, (9/10) http://www.dlib.org/dlib/september07/gold/09gold-pt2.html (Accessed 09/19/2007).

Google. 2006. NASA takes Google on journey into space. http://www.google.com/press/pressrel/google_nasa.html (Accessed 04/01/2008).

Greer, Chris. 2007. Funders' perspectives. DigCCurr2007, an International Symposium on Digital Curation. April 18-20, 2007, Chapel Hill, NC.

Harvey, R. 2007. *DCC | digital curation manual: Installment on "Appraisal and selection".* Digital Curation Centre, , http://www.dcc.ac.uk/resource/curation-manual/chapters/appraisal-and-selection/ (Accessed 02/06/2007).

Helly, John J., T. Todd Elvins, Don Sutton, David Martinez, Scott E. Miller, Steward Pickett, and Aaron M. Ellison. 2002. Controlled publication of digital scientific data. *Communications of the ACM* 45, (5) (05//): 97.

Higgins, D., C. Berkley, and M. B. Jones. 2002. Managing heterogeneous ecological data using Morpho. Paper presented at Proceedings of 14th International Conference on Scientific and Statistical Database Management, 24-26 July 2002, .

Hockx-Yu, H. 2007. Digital curation centre: Phase two. *International Journal of Digital Curation* 2, (1), http://www.ijdc.net/ijdc/article/view/30/33 (Accessed 08/01/2007).

Hunolt, G. 2002. SEEDS Working Paper Four: Data Service Provider Model, Model Parameters. http://esdswg.gsfc.nasa.gov/pdf/WP4_ModelParm.pdf. (Accessed 03/31/2008).

———. 2003. SEEDS Working Paper Five: Data Service Provider Model, Requirements and Levels of Service. http://esdswg.gsfc.nasa.gov/pdf/WP5_ModelRqmts.pdf. (Accessed 03/31/2008).

Johns Hopkins University. New center created to manage digital scholarship. in Johns Hopkins University. Baltimore, MD, 2007. http://www.library.jhu.edu/about/news/releases/pressrel07/drcc.html (Accessed 01/23/2008).

Joint Task Force on Library Support for E-Science. 2007. *Agenda for developing E-science in research libraries.* Washington, DC: Association of Research Libraries. http://www.arl.org/bm~doc/ARL_EScience_final.pdf (Accessed 12/28/2007).

Karasti, Helena, Karen S. Baker, and Eija Halkola. 2006. Enriching the notion of data curation in e-science: Data managing and information infrastructuring in the long term ecological research (LTER) network. *Computer Supported Cooperative Work: CSCW: An International Journal* 15, (4): 321-58.

Kenney, Anne. 2007. CUL All Staff Meeting, November 2007. http://www.library.cornell.edu/staffweb/2007Nov_AllStaff.pdf (Accessed 04/01/2008).

Library of Congress. 2008. Digital Preservation Partners. http://www.digitalpreservation.gov/partners/. (Accessed 04/18/2008).

Lord, P., and A. Macdonald. 2003. *e-science curation report data curation for e-science in the UK: An audit to establish requirements for future curation and provision.* JISC Committee for the Support of Research. http://www.jisc.ac.uk/uploaded_documents/e-ScienceReportFinal.pdf (Accessed 10/30/2006).

Lynch, Clifford. 2006. Open computation: Beyond human-reader-centric views of scholarly literature. In *Open access : Key strategic, technical and economic aspects.*, ed. Neil Jacobs, 185-193. Oxford: Chandos.

———. 2007. The Shape of the Scientific Article in The Developing Cyberinfrastructure. CTWatch Quarterly 3(3). http://www.ctwatch.org/quarterly/articles/2007/08/the-shape-of-the-scientific-article-in-the-developing-cyberinfrastructure/. (Accessed 04/18/2008).

Lyon, L. 2007. *Dealing with data: Roles, rights, responsibilities and relationships (consultancy report).* JISC. http://www.jisc.ac.uk/media/documents/programmes/digital_repositories/dealing_with_data_report-final.pdf. (Accessed 1/31/2008).

Madrigal, Alexis. 2008. Google to host terabytes of open-source science data. http://blog.wired.com/wiredscience/2008/01/google-to-provi.html (Accessed 04/01/2008).

Martin, Carolyn. 2008. Cornell university library academic assembly: The Cornell Plan. http://www.library.cornell.edu/staffweb/AcadAssem/AAmin.html (Accessed 04/01/2008).

Messerschmitt, David G. 2003. Opportunities for research libraries in the NSF cyberinfrastructure program. *ARL*(229) (Aug): 1-7.

Michener, W., J. Beach, S. Bowers, L. Downey, M. Jones, B. Ludascher, D. Pennington, et al. 2005. Data integration and workflow solutions for ecology. Paper presented at Data Integration in the Life Sciences. Second International Workshop, DILS 2005. Proceedings, 20-22 July 2005, .

NASA. 2007. Earth system science data centers. http://science.hq.nasa.gov/research/daac/ (Accessed 04/01/2008).

National Academy of Sciences. 2008. U.S. national committee for CODATA. http://www7.nationalacademies.org/usnc-codata/ (Accessed 04/01/2008).

National Geophysical Data Center. n.d. USA home, world data center system. http://www.ngdc.noaa.gov/wdc/ (Accessed 04/01/2008).

National Institutes of Health. 2007. NIH data sharing information. http://grants.nih.gov/grants/policy/data_sharing/ (Accessed 04/01/2008).

National Science Board and National Science Foundation. 2005. *Long-lived digital data collections*. Washington, D.C.: National Science Foundation.

National Science Foundation. 2007. US NSF - funding. http://nsf.gov/funding/ (Accessed 04/01/2008).

National Science Foundation Office of Cyberinfrastructure. 2007. Sustainable digital data preservation and access network partners (DataNet). Arlington, VA. Available from http://www.nsf.gov/pubs/2007/nsf07601/nsf07601.htm (Accessed 01/23/2008).

OCLC. 2008. Panel to address economic sustainability of digital preservation. http://www.oclc.org/news/releases/200673.htm (Accessed 04/01/2008).

Open Data Commons. 2008. Public Domain Dedication and Licence. http://www.opendatacommons.org/odc-public-domain-dedication-and-licence/. (Accessed 04/11/ 2008).

Parr, Cynthia Sims. April 2007. Open sourcing ecological data. *Bioscience* 57:309,310(2).

Purdue University Libraries. 2007. D2C2 - Purdue University. http://d2c2.lib.purdue.edu/ (Accessed 04/01/2008).

Science Commons. 2007. Protocol for Implementing Open Access Data. http://sciencecommons.org/projects/publishing/open-access-data-protocol/. (Accessed 04/11/2008).

Smith, Robin S. 2007. Geospatial Data-sharing in UK Higher Education: informal repositories and users' perspectives, http://edina.ac.uk/projects/grade/Grade_reportRSSv2.pdf. (Accessed 3/24/2008).

Swan, Alma. 2008. The business of digital repositories. In *A DRIVER's guide to European research repositories.*, eds. Kasja Weenink, Leo Waaijers and Karen van Godtsenhoven, 15-47. Amsterdam: Amsterdam University Press, http://dare.uva.nl/document/93898 (Accessed 01/21/2008).

Thibodeau, Kenneth. 2007. Critical competencies for digital curation: Perspectives from 30 years in the trenches and on the mountain top. Chapel Hill, NC, http://www.ils.unc.edu/digccurr2007/papers/thibodeau_paper_7.pdf (Accessed 07/17/2007).

Treloar, Andrew, David Groenewegen, and Cathrine Harboe-Ree. 2007. The data curation continuum: Managing data objects in institutional repositories. D-Lib Magazine 13(9). University of California San Diego. 2008. Chronopolis Project Launched Under Library of Congress Partnership to Preserve At-Risk Digital Information. *http://ucsdnews.ucsd.edu/newsrel/supercomputer/04-08Chronopolis.asp*. (Accessed 04/18/2008)

Wilbank, John. 2007. Open Access Data: Boring, but Important. John Wilbanks' blog on Nature Network. http://network.nature.com/blogs/user/wilbanks/2007/12/17/open-access-data-boring-but-important. (Accessed 04/11/2008).

Working Group on Data for Science, PMSEIC. 2006. *From data to wisdom: Pathways to successful data management for Australian science.* , http://www.dest.gov.au/sectors/science_innovation/publications_resources/profiles/documents/Data_for_Science_pdf.htm (Accessed 01/21/2008).