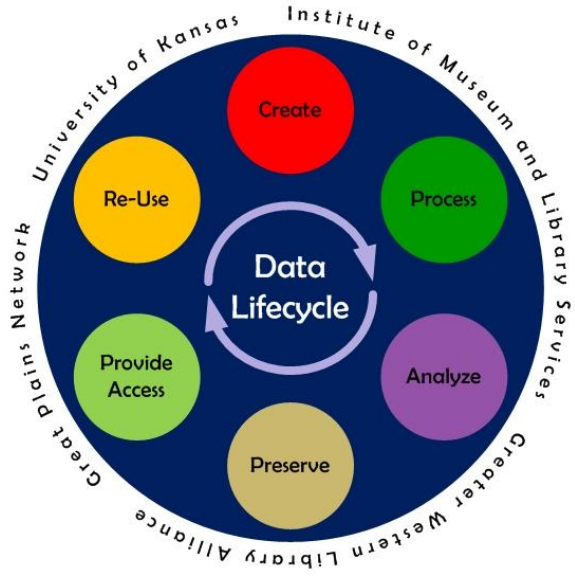
# REPORT

## PLANNING FOR THE LIFECYCLE MANAGEMENT AND LONG-TERM PRESERVATION OF RESEARCH DATA: A FEDERATED APPROACH

June 18, 2013

IMLS 51-12-0695




UNIVERSITY OF KANSAS LIBRARIES & INFORMATION TECHNOLOGY

GREAT PLAINS NETWORK

GREATER WESTERN LIBRARY ALLIANCE

INSTITUTE OF MUSEUM AND LIBRARY SERVICES

*REPORT*
*PLANNING FOR THE LIFECYCLE MANAGEMENT AND LONG-TERM PRESERVATION OF RESEARCH DATA: A FEDERATED APPROACH*

A report to the Advisory Council, IMLS Grant 51-12-0695

June 18, 2013 Version 1.1

## Grant Staff


Deborah M. Ludwig, Principal Investigator
Scott R. McEathron, Investigator
Bob Lim, Investigator
Paul K. Farran
Joni M. Blake, Investigator
Gregory E. Monaco, Investigator
Nicole A. Potter. Project Coordinator



Acknowledgements: We wish to acknowledge the efforts and contributions to this report from many individuals, including the members of the committees and research teams listed in Appendix A. We also thank Lars Hagelin of the Greater Western Library Alliance for the development of graphics and technical advice.




*This project is made possible by a grant from the U.S. Institute of Museum and Library Services*


---

# *TABLE OF CONTENTS*

# Review of Current Literature

## Introduction to the Literature Review (Scott R. McEathron)

In the process of reviewing the literature on the theme of lifecycle management and long-term preservation of research data, it became clear that many related sub-themes have emerged and continue to evolve. The literature related to this theme is like a growing tree with new branches forming and growing out from one or two main trunks. We have tried to identify the relevant branches and have grouped the literature into the following twelve topics:

1. Introduction and general works
2. Assessments of researcher behavior, attitudes and needs
3. Services, roles and responsibilities
4. Sharing, reuse, publication and citation
5. Data management planning
6. Policies and standards
7. Institutional repositories, approaches, and issues
8. Disciplinary or subject repositories, approaches, and issues
9. Federated approaches
10. Economics/costs of data curation
11. Archiving and preservation
12. Metadata and description

We have limited the literature review to works that have had a great impact (high citation rate), are excellent case studies, or show promise as examples of innovation. Thus the current literature review is selective and limited to English language works, most of which from the last 10 years. Some works could easily fit into more than one area. We have left it to the reviewer of any given section to decide which best fits that section. The selected works have been arranged thematically in APPENDIX E. An alphabetical bibliography of these works is located in APPENDIX D.

## 1. Introduction and general works (Scott R. McEathron)

As we began to review the literature relevant to lifecycle management and long-term preservation of research data, a number of bibliographies and guides surfaced. We are indebted to Charles Bailey's (2013) *Research Data Curation Bibliography.* With over 200 citations and growing, it is probably the most comprehensive bibliography on the subject. However, other distinctive bibliographies and guides are noteworthy. The Westra, et al. (2010) bibliography of *Selected Internet Resources on Digital Research Data Curation* presents a thematically organized bibliography of the more important internet based resources. Witt and Giarlo (2012) provide a description of another unique guide, *Databib: An Online Bibliography of Research Data Repositories*. *Databib* currently provides records on over 500 repositories worldwide and is an example of the growth and geographical breath in digital data repository services.

Some of the more important early works may be described as "calls to action" in response to the growth in e-sciences (Gray et al. 2002; Hey and Trefethen 2003; Hey et al. 2009). Grey (2002; Hey et al. 2009) called for tools to support the whole research cycle--and specifically the curation, archiving, and publishing of digital data. Hey and Trefethen (2003) called for the creation of new types of digital libraries to archive and curate e-science data and provide other data-specific services.

Other articles have made similar "calls to action" beyond the e-sciences (Borgman 2009; Ogburn 2010). Borgman's article is indicative of the growing interest in digital humanities. She makes comparisons of the data practices between the sciences and humanities, which she then uses to frame a series of lessons and questions. Ogburn (2009) makes a similar "call to action" in her article focused on the potential role of libraries in the area of data curation.

The digital or data curation theme has continued as a central focus for many writers (Higgins 2008; Ogburn 2010; Yakel 2007). Of note, Higgins describes the Digital Curation Centre's Curation Lifecycle Model as a tool to help plan curation and preservation activities to different levels of granularity (135). Yakel (2007) explores the evolution of digital curation as becoming "an umbrella concept that includes digital preservation, data curation, electronic records management, and digital asset management" (335).

A number of compilations have also emerged. The most noteworthy example is Graham Pryor's *Managing Research Data* (2012) with covering many of the same topics of this literature review--several of the chapters are specifically cited. What follows is a more focused literature review of each of the themes related to lifecycle management and long-term preservation of research data.

## 2. Assessments of researcher behavior, attitudes and needs (Stephanie Wright)

Publications assessing researcher needs, behaviors, and attitudes surrounding data management have proliferated over the last few years. While most assessments seem to use the same tools (surveys and/or interviews), they can vary widely in focus. Some focus on the disciplines of the researchers (Williams and Pryor 2009; life sciences), some on a particular phase of the data lifecycle (Feijin 2011; access and storage, Swan and Sheridan 2008; sharing, and Kuipers and van der Hoeven 2009, Sharpe 2006; preservation), others on specific geographic area (Sharpe 2006, Swan and Sheridan 2008; UK) or particular institutions (Marchionini, et al. 2012; Scaramozzino, Ramirez and McGaughey 2012). Not surprisingly, some publications had more than one of these foci, and there are similar findings in more than one assessment. In addition, there are publications that analyze and synthesize data from multiple assessments - attempting to provide a broader picture of data management from the researcher environment (Feijin 2011) and some included perspectives from other research-data stakeholders (publishers and data managers).

Feijin (2011) falls into the latter category and is in itself a literature study. The report focuses on researcher needs in terms of data storage and access, but covers much more - with the primary

conclusion that researchers do want and need support services for managing digital data. The authors provide a list of requirements necessary to make those support services successful, including making sure tools and services are easy to use and are "in tune with researchers' workflows". One of the publications reviewed by Feijin (2011) worth noting is the PARSE report focusing on digital preservation (Kuipers and van der Hoeven 2009). This assessment has broad geographic and disciplinary representation as well as responses from publishers and data managers, which provides a useful comparison of needs and motivations across stakeholders.

Swan and Sheridan (2008) interviewed over 100 researchers across multiple disciplines and assessed what researchers are actually doing in regards to data sharing, as well as uncovering their motivations and constraints with sharing data. In a similar vein, Scaramozzino, Ramirez and McGaughey (2012) looked at multiple data curation behaviors of California Polytechnic State University researchers and compared their actions to their expressed beliefs and attitudes.

Within the group of disciplinary studies, Williams and Pryor (2009) used information lab notebooks to supplement the more familiar assessment tools of interviews and focus groups to understand information exchange behaviors (including data sharing) of life sciences researchers within the context of their roles in a research group. This method allowed the authors to create diagrams showing the information flow within each research group and how they move through the stages of the data life cycle.

Finally, the *Data Curation Profiles Directory* out of Purdue University (Carlson and Brandt, 2013) is a publication worth exploring as it maintains a multidisciplinary collection of profiles of specific data set requirements as detailed by the researcher. In essence, the profiles are case studies identifying how researchers across institutions and disciplines deal with data management issues.

## 3. Services, roles and responsibilities (Brian Westra)

*Research Data Services* (RDS) and *Research Data Management* (RDM) services are two umbrella phrases authors have used to describe the suite of services related to data curation (Jones, Pryor, & Whyte 2013; Tenopir, Sandusky, Allard, & Birch 2013). These services may be viewed through the lens of organizational structures, degree of investment, competency requirements, or in relation to existing library services. Giarlo depicts the data curation services of academic libraries as "data quality hubs" (Giarlo 2013, 6), where curatorial practices address such factors as trust, authenticity, and usability. He then examines the implications for data curation practices. Others use the collection development metaphor to express data curation services (Choudhury 2010).

Lyon examines the implications of research data informatics on libraries and outlines how libraries can transform to meet these needs. She describes ten different data support services: surveys, planning, informatics, citation, training, licensing, appraisal, storage, access, and impact. She also outlines a framework of roles, responsibilities, requirements, and relationships for providing these services (Lyon 2012).

Lewis outlines nine areas where libraries can be active in relation to research data, ranging from developing library workforce data skills and confidence, to leading or partnering on the development of local data policies, and influencing national policy (Lewis 2010). Others have used a tier model for data management services activities that reflect increasing involvement or "embeddedness" with the researcher and the data, progressing from education to consultation to infrastructure (Reznik-Zellen, Adamick, & McGinty 2012). The data.bris project at the University of Bristol outlined four options in their business case for a pilot data management service: 'do nothing, do little, preferred, and gold-plated' (Whyte 2013).

Business cases, roadmaps, and strategic planning documents may be useful tools for defining the structures, partnerships, and organizational development required to provide new services. There are many examples including: (Beitz, Dharmawardena, & Searle 2012; Jones et al. 2013; Macdonald & Martinez-Uribe 2010; University of Edinburgh 2012; Cole 2013; Marchionini 2012; Whyte 2013; Witt 2012).

National and international initiatives such as DPN (Digital Preservation Network) and RDA (Research Data Alliance) aim to provide "an open global research infrastructure", and avenues for influencing and developing services and collaborations that can have broad impact.

Pilot cases provide an opportunity to explore, develop and apply the infrastructure that is needed to support the full lifecycle of research data. Examples include the University of California San Diego's work with five projects, in neuropsychology, archaeology, oceanography, earth science/topography, and astrophysics (Moore 2013), and the University of Manchester's work with biomedical data (Poschen et al. 2012).

The preceding examples and others exemplify successful collaborations and shared oversight between libraries and other institutional partners. For example, Purdue University Libraries partnered with the research office and information technology to develop and fund the Purdue University Research Repository. At the University of California San Diego, the Research Cyberinfrastructure Oversight Committee and Implementation Team have representatives from academic research departments, computing, the supercomputing center, the office of research, and the libraries (University of California San Diego).

Other kinds of infrastructure collaborations may include the evaluation and implementation of electronic lab notebooks (University of Wisconsin - Madison 2012), resource navigation systems (Haendel, Vasilevsky, & Wirz 2012), and image management systems (Linkert et al. 2010).

Policy development is at the higher level of Lewis' hierarchy (Lewis 2010). Some institutions are developing guiding principles that not only state policies for researchers, but institutional responsibilities as well (Flach & Price 2012). While researchers may define responsibilities for data stewardship based on a data "ownership" conceptual framework (Marchionini 2012), institutional responsibility statements can signal the commitments that the larger organization is making to data management.

## 4. Sharing, reuse, publication and citation (Scott R. McEathron)

From the White House in Washington, D.C. to an unassuming pub called the Panton Arms in Cambridge, UK, scientists, scholars, attorneys, and others are now talking about the importance of sharing data in research and more specifically, the need for open data. In a recent memorandum from The Office of Science and Technology Policy (OSTP) (2013), John Holden has directed "each Federal agency with over $100 million in annual conduct of research and development expenditures to develop a plan to support increased public access to the results of research funded by the Federal Government" (2). Similar activity can be found at the grass-roots level. In the UK, a group of scientists and others have developed a set of guidelines called the Panton Principles that aim to make open data "freely available on the public internet permitting any user to download, copy, analyze, re-process, pass them to software or use them for any other purpose without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself" (Murray-Rust, et al. 2010). The growing interests of scientists and the current and proposed future requirements of national research funders will continue to make data sharing a central theme within the literature.

Many articles used the theme of lifecycle management and long-term preservation of research data center as one of the primary reasons to preserve research data: **so it may be shared and reused**. The articles vary in their scope--from broad overviews, framing research agendas to multi-scaled or disciplinary specific data sharing and specific publisher policies.

A broad overview of the reasons for sharing (and not sharing) research data and an agenda for future research is provided by Christine Borgman (2012). A similar opinion, from the point-of–view of a scientist, on the need to share data is offered by Vision (2010). Vision concludes that instead of individual publishers or journals as the repository/archive of data, a "superior approach is a disciplinary repository that has data as its primary focus and is shared by a scientific community larger than a single journal or publisher" (330). The model he gives as an example is Dryad[1]. Vision also forwards the opinion that "journals are in the best position to promote the practice of archiving 'small-science' data upon publication" (330).

Cragin, et al. (2010) studied the data characteristics and sharing practices of "small-science" research areas and the implications for institutional repositories. They concluded that because of the high level of variation and complexity in data forms and sharing practices,…resource demands for curation services…will be high" (4036). Faniel and Zimmerman (2011) present a very thorough review of current research on data sharing and reuse and offer a research agenda in three areas: 1) how researchers manage their data; 2) questions around the re-users of data; and 3) the influences on how researchers make their data available (65-66).

An example of disciplinary concerns for data publication can be found in the field of Bioinformatics as described by Chavan and Ingwersen (2009). These authors detail specific concerns of data within the discipline such as data not being easily discoverable or accessible

[1] www.dryad.org

and the lack of recognition for publishing it. Further, they offer a "Data Publishing Framework" to incentivize the goal of sharing data.

Another way of conceptualizing the issue of data sharing is from the publishing framework. What does it mean to publish data? Lawrence et al. (2011) provide an overview of the structures that they feel are needed in order to improve a more formal system of data peer review, publication, and citation. They and others (Simons 2012) also suggest Digital Object Identifiers (DOIs) to provide a permanent identifier and locator for datasets (7). Mooney and Newton (2012) also found that the majority of articles they reviewed lacked adequate citation of data used. They concluded that full citation of data is not a normative behavior in scholarly writing (15).

Piwowar and Chapman (2008) explored the correlation between a journal's data-sharing policies and the likelihood of having accessible datasets. They reviewed the policies for data sharing of journals that publish results of research utilizing gene expression microarray data. They found that "high impact journals tended to have strong data sharing policies" (11) and also "articles published in journals with a strong data-sharing policy are more likely to have publicly available datasets" (15).

## 5. Data management planning (Brian Westra)

Although the NIH had required since 2003 that researchers address data sharing for grant awards over $500,000, the 2011 NSF requirement for data management plans (DMPs) had a more significant impact on proposal-writers and raised the importance of the services provided academic institutions. The recent Office of Science and Technology Policy (OSTP) mandate expands the basis for data management plan requirements to all federal agencies providing over $100 million in research grants (Holdren 2013).

A 2011 survey of ACRL libraries in the US and Canada (Tenopir, Birch, and Allard 2012), highlights the range of services provided by academic libraries. At the time of the ACRL survey, only 20% of library respondents were providing consultations for faculty and graduate students on data management plans (DMPs), but another 22% planned to do so within the next 2 years.

Services specific to data management plans for grant-funded research may include consultations with grant writers, DMP training and workshops, and form-based tools for creating a DMP. Some libraries have begun to review larger sets of DMPs (Parham and Doty 2012).

Libraries are also working alone and in collaboration with other campus partners to develop service models in support of the constituent elements of a plan, such as storage and backup, electronic lab notebook systems for description (University of Wisconsin-Madison 2012), and data preservation and sharing. Some aspects of these services are considered elsewhere in this document.

Understanding researcher needs, and presenting services with measurable positive impacts on those needs are critical to the success of DMP services. The data curation profile provides a framework for determining data management practices and needs of researchers (Carlson 2012). Establishing trust relationships with researchers is important, and embedded librarianship may provide one of several paths toward this end (Carlson and Kneale 2011).

The DMP Online tool developed in the United Kingdom, and its relative, the DMPTool developed by the California Digital Library and partners in the U.S., are employed by some libraries to walk grant-writers through the process of developing a data management plan for submission with a grant proposal. Sallans and Donnelly (2012) compare and contrast these two form-based web resources. The DMPTool links to data plan requirements published by the funding agency units, and local guidance materials can also be incorporated into the web pages.

Cornell and Syracuse librarians investigated funder data policies and assessed their comprehensiveness and level of detail against a rubric across a range of data policy elements (Dietrich et al. 2012). Funder data policies are usually general in scope, though a few programs have more explicit guidance and form-based resources for developing plans and reporting on data activities, such as IEDA (Integrated Earth Data Applications).

Libraries can provide training specifically about data management plans for researchers (Johnston, Lafferty, and Petsan 2012), but the data management plan structure can also be used as a rubric for broader data management training.

## 6. Policies and standards [report pending]

## 7. Institutional repositories, approaches, and issues (Sarah Potvin)

**What is an institutional repository?** Clifford Lynch's early, formative definition urged: "An institutional repository is not simply a fixed set of software and hardware" (2003, 2). Rather, it is "a set of services that a university offers to the members of its community for the management and dissemination of digital materials created by the institution and its community members. It is most essentially an organizational commitment to the stewardship of these digital materials..."(2).

**Were these institutional repository services intended to manage data over their lifecycle, and how have they evolved to do so?** Lynch's broad definition suggests the possibility; he predicted that "a mature and fully realized institutional repository ... will also house experimental and observational data" from community members (2). Subsequent research and surveying point to institutional willingness to engage with data, though financial, staffing, and technology constraints have intervened. Lynch and Lippincott (2005) report on results of a Coalition for Networked Information survey of relevant individual and consortial member institutions, aimed

at assessing the "current state of institutional repositories (IRs) in the US." In preparing the survey, they observed "two views" of IRs: "One characterizes an institutional repository as primarily addressing dissemination of various forms of e-prints for faculty work..."; "The second approach conceives of an institutional repository as broadly housing the documentation of the intellectual work—both research and teaching—of the institution, records of its intellectual and cultural life, and supporting evidence for present and future scholarship." This second variety "will include e-prints, certainly, but also datasets, video, learning objects, software, and other materials."

The 2005 CNI survey results indicated that "a significant number of institutions are committed to institutional repositories that go far beyond e-prints" (Lynch and Lippincott 2005).[2] While only four responding institutions indicated that their IRs currently held data sets, twenty-six indicated plans to do so over the next 1-3 years—the largest number of responses garnered for any type of planned content. Lynch and Lippincott conclude that institutional repositories in the US "are being positioned decisively as general-purpose infrastructure within the context of changing scholarly practice, within e-research and cyberinfrastructure, and in visions of the university in the digital age" (2005). A 2009 ARL report on repository services echoed the observation that "repositories are developing rather than developed" (Moore, et al. 2009, 8): "Just a few years ago, many libraries were acting on a vision of repositories that focused on preprints and postprints of faculty publications and theses and dissertations. ... We now understand better that institutions produce large and ever-growing quantities of data, images, multimedia works, learning objects, and digital records..." (8).

**How have particular institutional repositories engaged with data?** Many institutions have formulated data management plans, policies, and resources. Some universities have extended their "set of services" around data management while depending on repository systems beyond the university to disseminate the data.

The University of North Carolina's 2012 "Research Data Stewardship at UNC: Recommendations for Scholarly Practice and Leadership," a working report produced in response to a charge from the UNC Provost, prioritizes the placement of research data into subject/disciplinary repositories, rather than a UNC IR. It notes: "Individual researchers are best positioned to identify which repositories are most appropriate to their data and are encouraged to take advantage of public repositories whenever possible" (Marchionini, et al., 18). A central campus data registry is recommended to "provide a single place where anyone with access could get an overview of the data preservation efforts at UNC and at a minimum, a list of individual data sets that have been stored" (21). There is recognition, however, that public disciplinary repositories might not be ideal for all data. The recommendations thus extend to the development of "A repository for UNC research data where public repositories do not exist and/or data must be locally managed by contractual or sensitivity requirements" (24). This approach loosens the university's role in publishing its researchers' data, while ensuring that the data is accessible and

[2] Italics removed.

published in a repository preferred by the researcher (or funder).

Other institutions have taken a more active role in overseeing the full lifecycle management of institutional research data. Notably, the Purdue University Research Repository (PURR)[3], built on a HUBzero platform, aims to support researchers as they develop data management plans, collaborate on research, and, ultimately, publish and archive their datasets, which are issued with DOIs (PURR website, 2013). Still others, such as Indiana University Bloomington, have integrated specialized ingest processes, which prompt the collection of data-specific metadata for those contributing datasets to the institutional repository (Konkiel, 2013).

How have other institutions approached the decision by researchers to deposit data in either disciplinary or institutional repositories? Universities retain lists of potential repositories for deposit, often relying on other institutions to develop and supplement these lists (University of Oregon; University of Idaho Library).[4] They have also developed tools to facilitate data curation and, ultimately, repository deposit. Some tools are designed to interoperate with both institutional and disciplinary repositories for deposit. Cornell University's DataStaR (Data Staging Repository), "a platform and a set of services meant to facilitate data sharing" takes a destination-repository-neutral approach (Steinhart 2011, 16). The intermediate repository allows Cornell researchers to "store and share data with selected colleagues, select a repository for data publications, create high quality metadata in the formats required by external repositories and Cornell's institutional repository, and obtain help from data librarians with any of these tasks" (Steinhart, 16). Monash University had previously explored the role of intermediate or "collaboration" repositories that facilitate researchers actively working on their data prior to their appearing in "publication domain" repositories (Treloar, Groenewegen, and Harboe-Ree 2007). This approach, too, strengthens the university's set of offered services while retaining neutrality on publication site.

One issue regarding including data in institutional repository services is the question of whether data can simply be incorporated into the existing infrastructure, or whether it requires separate handling and applications. Some authors have enforced the distinction suggested by Lynch and Lippincott (2005), pointing to "publication" or "data" repositories as sites worthy of differentiation.

---

[3] PURR website. (Accessed April 26, 2013) https://purr.purdue.edu/

[4] The University of Oregon site lists repositories by discipline and includes the note: "You may want to use the University of Oregon Libraries' institutional repository, called Scholars' Bank, for your data." University of Oregon (Accessed April 26, 2013) "Data Repositories," in "Research Data Management," http://library.uoregon.edu/datamanagement/repositories.html. University of Idaho Library (Accessed April 26, 2013) "Data Management: External Links." http://www.lib.uidaho.edu/services/data/data_management/links.html. The UNC Data Stewardship report recommends that UNC "design a process whereby researchers are both informed of existing resources and encouraged to use existing resources as a first choice. ... it is unlikely the university will have the wherewithal or inclination to produce a complete listing of all data repository resources along with a full enough specification of their intended use-cases. Many universities are developing such repository lists... and UNC should collaborate with others whenever possible to ensure deep coverage." (Marchionini et al.,18).

Early efforts involved developing open-source digital repository applications such as Fedora (developed at Cornell in 1997) and DSpace (developed at MIT, with support from Hewlett-Packard, in 2002 and now managed, developed, coordinated, and supported under the umbrella of DuraSpace). Together with ePrints, these applications represent many of the institutional repository platforms currently in use in the US and form the basis of institutional repository services. Universities customize their applications of these platforms (sometimes extensively). They depend on (and ideally participate in) networks of developers and committers who improve and update the open source products. Universities might choose one application, or maintain multiple repository platforms, perhaps specifying separate instances for publication or data repositories.

Each of the three institutions participating in the DISC-UK DataShare Project (2007-2009), a JISC-funded project to investigate and "contribute to new models, workflows and tools for academic data sharing ... " in institutional repositories, relied on its own distinct repository platforms—DSpace, ePrints, Fedora—for institutional datasets. Two of the institutions involved incorporated data into their existing repository instance, while one launched a dedicated repository for datasets, though using the same application as the parallel institutional publication repository. A key conclusion of the project was that "IRs can improve impact of sharing data over the internet" (Rice 2009, 5).

As Macdonald and Martinez-Uribe (2010) summarize the evolution of institutional repositories toward data:

> Integral to the whole research base are research outputs such as publications and digital data as both evidence and the means to verify intellectual endeavor. University strategies to harvest these products have developed around the concept of digital repositories developed by academic libraries. The first realization of such information systems were publication repositories built to manage and disseminate research articles and aimed to provide open access to a significant proportion of newly published academic papers. The development of research data repositories has been seen as the next coherent step in the growth of repositories. (6)

However, this coherent step toward research data repositories has not been collectively trod. Rather, as these examples demonstrate, they have been undertaken in different ways. Currently, US research universities offer distinct and different repository services and tools. While recognizing the importance of and the need to intervene in the data lifecycle management of their researchers, universities have developed and expanded their repository services along particular lines, according to institutional priority. These distinct approaches continue to evolve.

## 8. Disciplinary or subject repositories, approaches, and issues [report pending]

## 9. Federated and collaborative approaches (Deborah M. Ludwig & Michael Bolton)

*Databib*[5] currently lists 573 repositories for research data, many of which involve partnerships or collaborative efforts. The literature about federated and collaborative approaches to managing and sharing research data looks at participative communities that build shared services and infrastructure in a variety of disciplinary and institutional contexts. This review covers literature pertaining to a few federated or collaborative examples.

**Collaborative Communities.** Beyond a place to store and access data, successful long-term curation of research data involves communities of practice, working alongside disciplinary specialists to develop tools, standards, best practices, and programs for education and training. The National Science Foundation (NSF) *DataNet*[6] program envisions, "the creation of new (virtual) organizations [integrating] library and archival sciences, cyberinfrastructure, computer and information sciences, and domain science expertise…." (Hanisch and Choudhury 2009, 2) and Treloar (2009), highlight the NSF *DataNet* call

> …to build sustainable infrastructure by creating a new type of organization that the NSF does not believe exists today. It is looking for librarians, archivists, and computer/computational/ information scientists who will work together to build excellent infrastructure for science and/or engineering, while engaging closely with intended users; domain scientists will be full partners in the process" (134-135).

Michener et al. (2011) describes the Data Observation Network for Earth (*DataONE)* initiative as a participatory endeavor, engaging scientists as the primary stakeholders in the network with numerous secondary stakeholder communities. Libraries have been prioritized as the most important secondary community network "because integrative science is data-driven and information-reliant and because libraries provide support services in each of the five [*DataONE*] science research environments" (7-8).

Many authors of articles and papers about research data emphasize the community aspects of data sharing and stewardship (Allard 2012, Michener et al. 2011, Hanisch and Choudhury 2009, Schaeffer et al. 2011, Treloar 2009, Williford and Henry 2012). Collaborative communities allow people with different skills and knowledge to work together and to contribute to successful and complex solutions. Partnerships create a foundation for success by building on the knowledge and skills of a rich community of people. Without tools and standards to make data management and data sharing practical and without the requisite best practices, education and training necessary to adhere to the established standards and best practices, data stewardship projects and programs may not be sustainable (Williford and Henry 2012). Many of the projects noted in this review of the literature offer perspectives on the essential human connections.

[5] Databib is a tool for discovering sources of online research data.
[6] http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=503141

**What do we mean by federated approaches?** Heimbigner and McLeod (1985) defined a federated database architecture as one that

> …allows a collection of database systems (components) to unite into a loosely coupled federation in order to share and exchange information. The term federation refers to the collection of constituent databases participating in the federated database" (254).

In a data repository context, we can conceptualize federations as a distributed approach to storage or repository services that manage, share and perhaps curate data from multiple sites. The point of federation is often at a registry level, which houses metadata and pointers to the actual location of the data within a larger connected network (Allard 2012; Michener et al. 2011; Treloar 2009; Warner et al. 2007).

Williford and Henry (2012), in discussing the *Digging Into Data Challenge* program[7] sponsored by ten international research funders, refer to "a digital ecology of data, algorithms, metadata, analytical and visualization tools, and new forms of scholarly expression"(2). The authors note, "the digital raw materials upon which today's humanists and social scientists rely are heterogeneous, complex, and as massive as 'big data' in the sciences"(2-3). They recommend the adoption of more models for sharing, noting that it "makes no sense to replicate resources, skills, and services at all colleges and universities"(4) and that agreement to work together offers mutual benefit.

Many articles note the challenges of cross-repository interoperability based on differences between disciplinary communities and the need for an interoperability framework to connect many heterogeneous systems housing data. Warner et al. (2007) discusses the *Pathways* project, a partnership of Cornell and Los Alamos National Laboratory to develop a lightweight interoperable data model and workflow to better enable interoperability across complex metadata and the formats represented.

**Federated Approaches to Data Management in the Sciences.** Two exemplars of federated approaches to managing scientific research data are found in *DataNet* projects the National Science Foundation funded in 2009.

1) The *Data Observation Network for Earth* (*DataONE*) with headquarters at the University of New Mexico, a member institution of the Greater Western Library Alliance.
2) The *Data Conservancy*[8] with headquarters at the Sheridan Libraries at Johns Hopkins University.

***DataONE***

[7] http://www.diggingintodata.org/
[8] http://dataconservancy.org/

Allard (2012) and Michener et al. (2012) discuss *DataONE* as a multi-institutional, multi-national, and interdisciplinary collaboration to support the full information lifecycle of biological, ecological, and environmental data and to provide tools for researchers, educators, and the public. Michener et al. (2012) explains that environmental sciences represent a challenge for discovery because of their extremely heterogeneous data. *DataONE* reaches across different scientific disciplines and institutions for sharing of data and findings, expertise and tools, and for the development and utilization of compatible data management strategies and best practices.

Allard (2012) and Michener et al. (2012) describe *DataONE's* distributed technical architecture of member nodes and coordinating nodes or repositories with accompanying infrastructure and services. The overall architecture includes repository services and data replication as well as the *Mercury* metadata catalog for data discovery across geographically distributed member node repositories. The project addresses the need for a consistent researcher interface with analysis and visualization tools spanning member repositories. The services include quality metadata, shared identity management and access control policies.

Michener, et al. (2012) and Allard (2012) both describe *DataONE's* virtual community as one that creates a strong network of people and working groups comprised of scientists, academic researchers, educators, government and industry representatives, and leading computer, information, and library scientists. To encourage a cross-disciplinary approach, the network's scientific communities are not categorized by domain . *DataONE* facilitates data preservation and re-use, interoperability solutions, and best practices for data management across its lifecycle. Michener et. al. (2012) describes scientists as *DataONE's* primary stakeholder with libraries as the most important secondary stakeholder community. Library partners include the Libraries at the University of New Mexico; the College of Communication and Information, University of Tennessee; and the UC Curation Center, California Digital Library, University of California. Allard (2012) concludes that *DataONE is* a successful partnering of librarians, information science researchers, and scientists and notes that community engagement of scientists is an important user-centric focus of the *DataONE* project.

***The Data Conservancy***

Hanisch and Choudhury (2009) describe the *DataNet Data Conservancy* project as a partnership between John Hopkins University's Sheridan Library, the US National Virtual Observatory, and a wider list of partners found on the web site[9] that includes several national research centers along with Cornell, the University of Illinois, and the University of California, Los Angeles. Use cases in astronomy, seismology, and international land use policy have informed initial development. "The Data Conservancy team is interdisciplinary and multifaceted, and rooted in the university research library at the John Hopkins University" (6). The authors explain three key terms: stewardship, sustainability, and multiple-scales.

As with *DataONE*, Hanisch and Choudhury (2009) describe the technical elements of the federation as a common user access interface layer that includes a registry of aggregated

[9] http://dataconservancy.org/community/partners/

metadata, database query and data access, and distributed storage for the actual data objects. The project envisions a black box repository environment that could be widely deployed by various organizations from which metadata is replicated out to *The Data Conservancy*.

Both *DataONE* and *The Data Conservancy* initiatives represent national and international approaches to federated management of scientific research data with metadata for discovery and tools or services providing points of federation for discovering and using distributed collections of data.

**Federated Approaches to Data Management in the Social Sciences.** Two social science data initiatives, *ICPSR* and the *IQSS Dataverse* illustrate federated approaches to social science research data management.

***The Inter-university Consortium for Political and Social Research (ICPSR)***

*ICPSR*[10] is a non-profit, membership-based, centralized data archive located at the University of Michigan and focused on social science data. Green and Gutmann (2007) from ICPSR discuss two primary repository types: institutional and discipline/domain. ICPSR represents the latter type of repository. "The current digital repository landscape is made up of a blend of repository types. Repositories can be grouped into two broad categories: Institutional digital repositories with no specific discipline focus and discipline or domain-specific data archives" (38). The authors note that many institutional repositories (e.g., asudigitalrepository[11], KU ScholarWorks[12], MOSpace[13], Scholars Bank[14]) may allow deposit of data sets, "but support services for data processing, metadata production, or analysis are not usually offered as part of the repository service….[T]hese repositories position themselves at or near the end of the scientific research life cycle. Their goal is less to partner with researchers or with domain-specific repositories throughout the research life cycle than it is to garner the value of the institution's productivity, to gather this productivity, and possibly to lower the local or community-wide cost of scholarly publications" (38-39).

The authors remind us that the social science domain reflected in discipline-specific repositories have been in existence for decades. "Rather than focusing on publication-related materials from multiple subjects areas within a single organization, domain-specific digital repositories hold collections of materials grouped by type, subject, or purpose and intrinsically support domain- or discipline-oriented research needs. Domain-specific digital repositories in the social sciences have a history of providing infrastructure for data sharing and strive to provide support throughout the data life cycle. These data archives hold the raw materials that faculty and students can reuse, repurpose, analyze, and recompile in teaching, learning, and research environments" (39).

[10] http://www.icpsr.umich.edu/
[11] http://repository.asu.edu/
[12] http://kuscholarworks.ku.edu
[13] https://mospace.umsystem.edu/xmlui/
[14] http://library.uoregon.edu/diglib/irg/SB_Role.html

Both disciplinary and institutional repositories have shared goals as well as distinctive differences. The authors note the similarities of shared metadata and common discovery platforms such as Google Scholar. "Resource discovery in the social sciences now extends far beyond consulting a stand-alone research aid or search tool. Alliances have been developed across repositories and a new set of tools allowing researchers to do complex and innovative searches to locate and explore data is emerging" (41).

***Dataverse, the Institute for Quantitative Social Science, Harvard University***

The Harvard *Dataverse* Network[15] is described by Crosas (2012) as a collection of social science research data contained in virtual data archives called "dataverses" open to all researchers worldwide to share, cite, reuse and archive research data. Institutions can also use the open source *Dataverse* software to develop their own local institutional data repositories, which through metadata harvesting and sharing can then become federated partners within the larger Dataverse network. Repositories based on other software can also share metadata with the Dataverse network using the OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) protocol.

According to Crosas (2012), a key focus for *Dataverse* is maintaining the researchers connection to the data.

> Most professional archives, although often considered the most reliable solution, do not usually facilitate control and ownership of the data by the author. Once the author submits the data, the archive becomes fully responsible for the data management, cataloging and future updates. While this can be advantageous for some researchers, many prefer to maintain control of their data and to receive increased recognition.

The author explains that *Dataverse* allows researchers to create "virtual collections" called *Dataverses*, which include custom branding for the researcher or research team – helping meet the researcher's needs for recognition, visibility and ownership.

**Federated Approaches to Data Management in the Humanities and Arts.** Researchers in the humanities and in the arts create data that may be both big and complex.

***Humanities Projects at Scale – The Digging Into Data Challenge***

Williford and Henry (2012) cover eight humanities collaborations sponsored by the previously mentioned *Digging Into Data Challenge*. All projects engage with large data corpora, apply computational analysis, require collaboration from a variety of professionals, and conduct a research process. All projects cross disciplinary and international boundaries. The authors note some of the challenges these multi-institutional, multi-disciplinary, multi-national partnerships have encountered including working as virtual teams, and the lack of effective training for students and junior scholars pursuing computationally intensive research. Long-term sustainability of the projects was an additional concern of participants in the *Challenge*. The authors make nine recommendations and expand on these in their report:

[15] http://dvn.iq.harvard.edu/dvn/

1. Expand our concept of research.
2. Expand our concept of research data and accept the challenges that digital research data represent.
3. Embrace interdisciplinary.
4. Take a more inclusive approach to collaboration.
5. Address major gaps in training.
6. Adopt models for sharing credit among collaborators.
7. Adopt models for sharing resources among institutions.
8. Re-envision scholarly publication.
9. Make greater sustained institutional investments in human infrastructure and cyberinfrastructure (2-3).

***Digging Into Image Data… (DID-ARQ)***

DID-ACQ is one of the Simeone, et al. (2011) speaks of lessons learned from the international *Digging Into Image Data to Answer Authorship-Related Questions* (*DID-ARQ*)[16]. This project looked for ways to examine an archive of images too large to examine manually. It provides a template for future collaborations involving multiple datasets in geographically distributed locations. The project includes researchers from the University of Illinois, the National Center for Supercomputing Applications, Michigan State University, and the University of Sheffield.

***TextGrid***

Another example of a federated humanities data management is the *TextGrid*[17] project involving ten institutions in Germany, which was described by Neuroth, et al. (2011). The authors describe *TextGrid* as the first large, multi-year project in Germany dealing with developing a research infrastructure and virtual research environment for the arts and humanities. The environment has two parts: an entry point to a virtual research environment and a data archive (or set of archives). Disciplinary communities served include textual philology, linguistics, art history, classical philology, and musicology. *TextGrid* is part of the scientific *D-Grid* program affiliated with various e-Science endeavors. Within the *TextGrid* context, federation refers to the shared development of tools for the virtual research environment, the shared cost model, and a future federated repository infrastructure.

**Closer to Home: Collaborative Research Data Projects of GWLA / GPN Members.** In addition to *DataONE*, several projects have been developed at Greater Western Library Alliance and/or Great Plains Network institutions that exemplify possibilities for managing research data. These projects include tDAR and MaizeGBD described below.

***tDAR***

Spielmann and Kintigh (2010) discuss development of *the Digital Archaeological Record (tDAR)*[18], an international digital repository for data access and preservation from archaeological investigations. The *tDAR* repository encompasses datasets, documents, and images from current

[16] http://isda.ncsa.illinois.edu/DID/
[17] http://www.textgrid.de/en/community/
[18] http://www.tdar.org/

archaeological research and legacy data. *tDAR* is part of the Digital Antiquity partnership, a multi-institutional group that ensures sustainability.

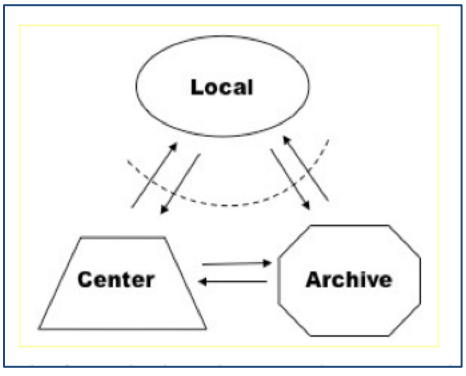
***MaizeGDB***

The *MaizeGDB* project, described by Schaeffer, et al. (2011) and Lawrence, et al. (2004), is a public database that serves the community of maize researchers by storing and curating data related to the genetics and genomics of maize (corn). The "curation focus is to facilitate data integration of very large data sets and to provide insight into development of easy-to-use interfaces and data displays. The efforts toward data integration involve gene nomenclature considerations as well as ontology development and implementation"(1). This data repository is over 20 years old and has matured from a focus on comprehensive curation of the literature, genetic maps and stocks to a current approach including the recent release of a reference maize genome sequence, multiple diverse maize genomes and sequence-based gene expression data. Originally housed at the University of Missouri, the Maize DB is the work product of a research community that provide data, establish nomenclature standards and recommends directions, priorities and strategies.

**Federation Between Different Types of Repositories.** The repository is a key element in federating long-term stewardship of data, reflecting the often-rich research partnerships built around research data. Baker and Yarmey (2009) note distinctive types of repositories that have different goals, participants, and purpose; but these repositories are also similar in terms of roles, activities and responsibilities. The authors consider repository relationships from the vantage point of connectedness to, or distance from, the researcher. The repository first created by, or for, the researcher is more connected to the active research project while a repository that provides "final" archiving of a dataset is further away from the researcher.

The authors illustrate the relationship of different types of repositories and the data flow between three different categories of repositories:

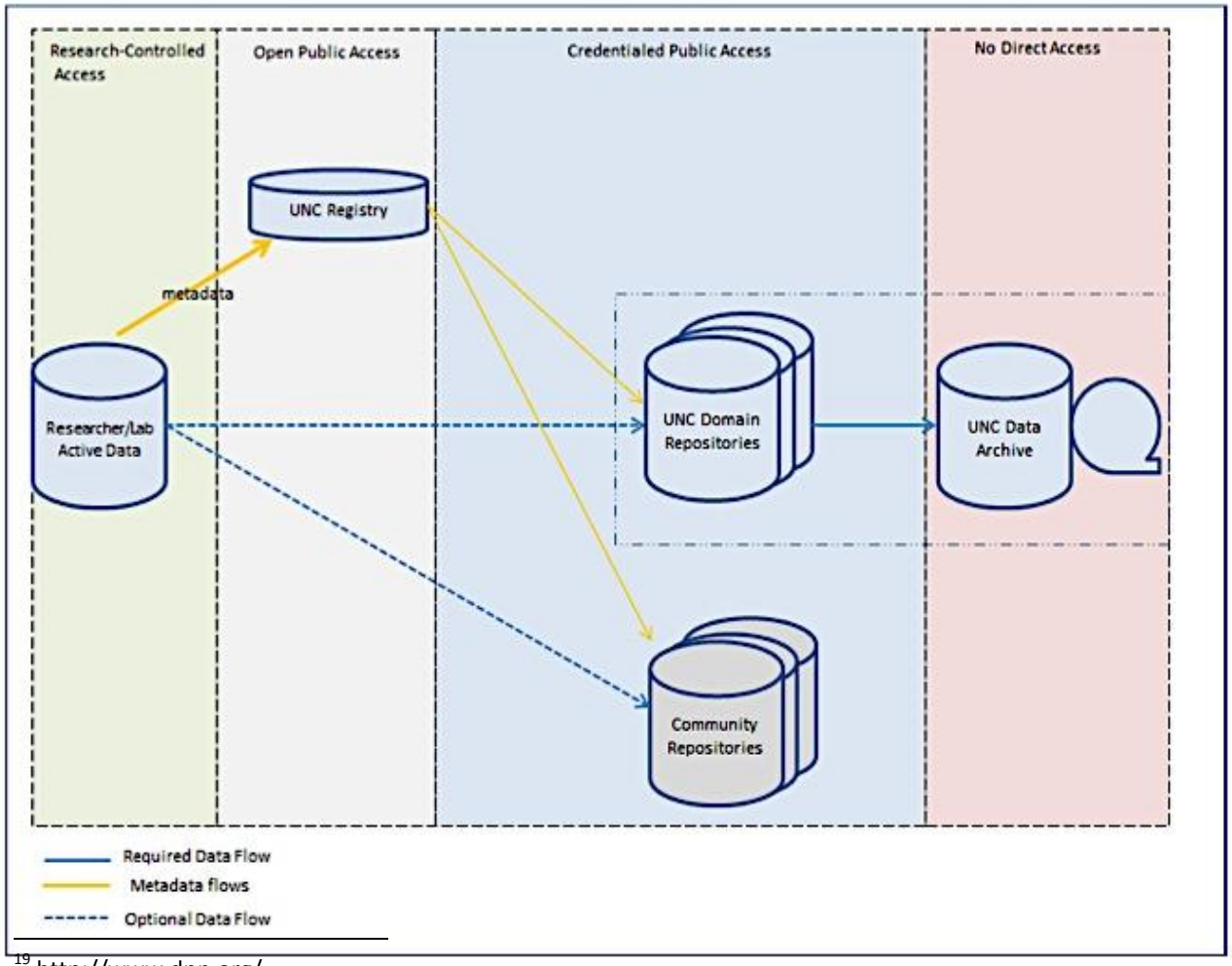
- local repositories where data lives during the active research process
- center repositories where data may be accessed during or after a project
- archival repositories concerned with long term data preservation




Data may begin the lifecycle in a repository largely controlled by the researcher(s) involved in the acquisition of research to answer a particular question. As the data matures and is ready for sharing, it may move to a "center" repository with the function of making data available or shared with a broader audience, and finally, if the data has future value in disciplinary and interdisciplinary research, the data may move to an archival repository for the purpose of ensuring future re-use.

Baker and Yarmey (2009) explain that the three types of repositories form a collective approach to research data management with each type of repository filling a different role within the federation and serving a different audience. Data can flow in both directions within the federation. For example, the archive may be used to restore data to the local researcher's repository or, in some cases, data might move directly from the local repository to an archive. Data might be synchronized between an archive and center (accessible) archive so changes and new data versions are automatically sent to the archive. An archive might periodically check to see if its version of the data matches that of a center repository and initiate appropriate events when the two versions differ. The *Digital Preservation Network* (*DPN*)[19], for example, represents a specialized type of federation to reserve research data and other digital artifacts of interest, including cultural heritage data. Baker and Yarmey (2009) would consider this a federation of dark archives used for preservation and restoration of data access when lost by a "center" archive.

**Connecting Local and Remote Repositories – the University of North Carolina View. T**he University of North Carolina (Marchionini, et al. 2012) undertook a campus study of digital research data stewardship. Working under a charge from the Provost, the report describes how




[19] http://www.dpn.org/

a task force conducted an environmental scan of research data stewardship policies and trends, discussed issues, and using interviews and a survey collected data on campus. The task force developed a set of principles and associated courses of action for the campus to consider. The UNC data flow diagram (above) shows expected interaction of UNC's repositories and with possible community/disciplinary repositories. This diagram shows metadata flowing from a researcher or lab to a university data registry for public discovery with metadata and optional data flows to university repositories or community repositories as appropriate. Data flows to university domain repositories would be coupled with deposit into a "dark" archive for presentation that excludes public access and could provide future restoration of the data to the access repository if needed.

**In Summary: Federated and Collaborative Data Management in the Literature.** Authors cited in this literature review point to the need for data to remain close to the researchers and for the researchers to play a major role in the development of any plan for the management, stewardship and curation of research data. The literature suggests developing a community to tackle complex issues of data management in a user-centric approach, broadly inclusive of partners with different skills and knowledge.

Various authors (particularly Green and Gutmann 2007) also point out that both institutions and disciplinary communities are in the business of building and maintaining repositories for access to data. Universities need to consider and articulate how they wish to balance expectations for data sharing between local campus and non-local repositories. Disciplinary communities may also need to consider how to work with other disciplinary and with local institutional repositories to federate discovery of data for researchers and perhaps to provide common sets of tools for interacting with the data.

Data is complex, originating from a variety of domains. Scientists, social scientists, and humanists use data that extends beyond numeric and encompasses a wide variety of formats. Data management planning needs to take into account the various types of data to be curated. The University of North Carolina captures data complexity in the following assumption.

> The term 'research data' is considered in its broadest form and includes the digital traces of processing and documentation associated with the research enterprise. We consider only electronic (digital) research data, yet, what this means varies enormously across disciplines and scholars. We believe that disciplinary communities (including funding agencies for those communities) are best prepared to define what constitutes 'research data' for their respective fields and that university policies and implementations should be guided by what UNC scholars and their communities define as research data. We also note that more than 40% of UNC researchers in our survey reported regular use of non-digital text, hand-written notes, and other forms of non-digital data as part of their research. Another one-third of our respondents report using biological, organic, or inorganic samples and/or specimens in their research. As UNC develops policies regarding research data – digital or otherwise –

this enormous diversity must be accommodated and embraced. UNC simply cannot seek to develop a "one-size-fits-all" policy in this environment (Marchionini, et al. 2012, 5).

## 10. Economics/costs of data curation [report pending]

## 11. Archiving and preservation (Amalia Monroe-Gulick)

The preservation of digital data and objects is a major concern in both the research and archivist communities. In a 2002 report from the Research Libraries Group (RLG), digital preservation was defined as "the managed activities necessary for ensuring both the long-term maintenance of a byte stream and continued accessibility of its contents" (Beagrie, et al. 2012, 3). The archiving and preservation of digital data has been a significant challenge to researchers and information professionals since the beginning of the "data deluge." Berman (2008) argues that it is assumed that digital data will always be available; however, it is actually fragile and can easily be lost due to many reasons including technology failure, as well as software and storage media becoming outdated.

According to Cragin, et al., "sharing is at the heart of success, as collecting, storing and making use of data can only [occur] after the means for sharing are in place" (2012, 1023). Another major concern are the lack of incentives and even disincentives for scientists to make data available for sharing. Borgman (2007) identifies disincentives for sharing which include the lack of rewards for effective data management, the amount of effort to preserve data for later use, and scientists desire to share data only after publication. In addition, Tenopir, et al. (2011) indicate scientists identify problems with funding and time as the main obstacles to sharing data. However, many funding agencies, such as the NSF, are now requiring data to be made available. Beyond funding requirements, one author argues that the scientific community needs to be fully aware of the importance of the continuum of data that is necessary for the advancement of future research (Durr 2008).

Jantz and Giarlo (2005) argue that no matter the scale of the data preservation purpose (i.e. "big data" or "small data"), methods, policies, standards, and technologies all need to be considered. In an article addressing data preservation for astronomy, they found that while many research facilities provide the archival space, a complete cycle - which includes capture, curation, preservation, and long-term access - is missing (Choudhury, et al. 2007).

One of the first steps in establishing a complete cycle or "end to end process" is risk assessment and management. Risk management of data preservation is another important issue, and challenge, when deciding how to preserve digital objects. Different types of preservation actions in terms of risk management for archivist and scientists can be clarified through the framework a Preservation Network Model (PNM) Conway, et al. (2012). The goal of PNM is to capture the interrelationships of the data, software, and other digital objects in order to accurately assess the acceptable risks. The preservation strategies identified are:

- Risk acceptance and monitoring
- Capture of software and extension through the stack
- Description
- Migration

These different strategies require that the risks with preserving each type of digital object first be identified. In addition, different categories of data have special considerations that need to be addressed during these initial assessment and planning phases. For example, McGarva (2009) identifies the specific actions with Geospatial data, including: format, systems, legal, and community actions.

Akmon (2011) indicates that another essential element for successful data preservation is developing of a preservation plan early in the research lifecycle. Because of the unique lifecycle of different types of data, this is an issue that presents challenges and cannot be completely addressed by standardized policies. Wallis, et al. (2012) discuss the importance of understanding different types of data lifecycles (including e-science and government documents) as essential, especially in interdisciplinary research. In an article discussing the preservation of remotely sensed data, Faundeen reports on the data migration systems used by the U.S. Geological Survey which address the major issue of technological obsolescence, and argues that "planning for these preservation activities is extensive and must be done before data are threatened" (2003, 162).

Akmon (2011) argues that libraries and archivists can play an important role in the preserving and archiving of data, even without specific domain knowledge. Tenopir, et al. (2012) found that most academic libraries do not develop research data policies for universities. However, because of the expertise they can offer in archiving and preservation policies, they could serve as clearinghouse for potential policies.

Jantz and Giarlo (2005) found that there are many different types of systems and methods for long-term access to research data. Institutional repositories managed by libraries, with established guidelines and procedures are one option that has been

greatly explored. Another potential method for dealing with the challenge is utilizing cloud technology. Mattmann et al (2010) report on storing and preserving NASA data with cloud technology. However, regardless of the type of system, the same fundamental questions of long-term needs must be addressed.

This review of the literature addresses some of the issues and potential solutions to the many challenges facing those attempting to preserve data. This is an evolving topic that will require ongoing exploration and discussion as the most effective, scalable, and discipline-specific processes to ensure long-term access and simplified sharing of data develop.

## 12. Metadata and description (Andrew Johnson)

Metadata is central to the lifecycle management and long-term preservation of research data. A number of metadata models and schemes have been developed to describe data from particular domains or disciplines. These include the Data Documentation Initiative for social science data and the Ecology Metadata Language for ecology data. There have also been attempts to describe scientific data more broadly. Matthews et al. (2010) developed a general model, called the Core Scientific Metadata Model (CSMD), to represent scientific study metadata for data generated in large-scale scientific facilities. The CSMD was designed to provide a core model that can be extended when necessary. Thus, the CSMD can be more specialized than general metadata models, yet broader than those that only describe a particular scientific domain (Matthews et al. 2010, 114).

In addition to describing research data in either general or domain-specific terms, metadata is also valuable for discovering and citing research data sets. The DataCite Metadata Scheme attempts to address the issues of data discovery and citation at a global level (Starr and Gastl, 2011). The DataCite schema features a small set of mandatory elements with a larger optional set that can be used for more detailed description. The core elements include general descriptive information as well as a Digital Object Identifier (DOI) to describe relationships between the cited data set and other objects.

Metadata plays a key role in the long-term preservation of research data. In order for research data to remain authentic and useable over time, it is necessary to capture contextual and administrative metadata as well as technical metadata about the digital object(s). As Wilson (2010) notes, this latter type of technical metadata, exemplified by the PREMIS standard, is not solely sufficient for long-term preservation of research data, and he argues that archivists and scientific researchers should work together to ensure that all of the necessary "recordkeeping" metadata, including contextual and administrative metadata as well as technical metadata, is captured in order to maintain data integrity, authenticity, reliability, and usability over time (215). Shaon and Woolf (2012) provide an example of one such attempt to incorporate contextual, domain-specific metadata with technical preservation metadata. In their work with

the Infrastructure for Spatial Information in the European Community (INSPIRE) Spatial Data Infrastructure, they convey that INSPIRE's application of the ISO 19115 metadata model for geospatial data captured important contextual information necessary for discovery and understanding of data; however, it failed to meet preservation metadata requirements like those described in the Open Archival Information System (OAIS) Reference Model . To solve this problem, they developed a preservation profile of ISO 19115 that includes both preservation metadata adhering to OAIS and PREMIS specifications as well as core ISO 19115 metadata that allows for accurate description and contextualization of geospatial data.

Despite the emphasis in the literature on capturing a variety of types of metadata throughout the research data lifecycle, there are numerous barriers and challenges to achieving this goal in practice. Mayernik (2010) discusses significant metadata challenges for curating and sharing data. He notes that responsibility for metadata creation is ambiguous and could include information professionals, scientists, and hardware/software tools (Mayernik 2010, 2). Another challenge lies in the discrepancy between the information professionals' formal approach to metadata creation and the informal practices of scientists. This discrepancy is further complicated by the distributed nature of metadata creation in research settings as well as the changing role of metadata through different stages of the data lifecycle (Mayernik 2010, 2). In additional work on the metadata practices of scientists, Mayernik (2011) describes how researchers who rarely create documentation beyond what is required for their own use also rarely share data with users outside of their immediate projects (248). When asked to create metadata for a widely available metadata registry, the researchers Mayernik (2011) observed still described data in ways targeted mostly to researchers with similar expertise (11). Mayernik (2011) argues that the metadata necessary for widespread data sharing is not in line with the current metadata practices of many researchers (280).

Ferguson (2012) describes additional metadata challenges for biomedical researchers attempting to understand and repurpose publicly available microarray data sets, and she notes that many of these data sets lack sufficient contextual information and metadata (51). In an effort to address this issue, a team of data curators at the Bioinformatics Core Group in the Harvard School of Public Health sought to annotate and contextualize some of this publicly available data (Ferguson 2012, 52). According to Ferguson (2012), the data curators used a suite of open source software tools to provide metadata about the experiments and data sets. However, even with tools and standards available to assist with metadata creation, she describes this process as time consuming and requiring significant subject matter expertise (55). Ferguson (2012) also notes that the data curators expressed uncertainty about how much metadata was necessary to allow for discovery and reuse (55).

Greenberg et al. (2009) also address the question of what constitutes sufficient metadata in their work developing metadata best practices for the Dryad data repository. They determined that the amount and type of metadata for data ingested into digital repositories should support both the immediate operational needs and long-term project goals of the repository (208). For example, Dryad's metadata approach addresses the immediate need of making content

available in a DSpace repository via an XML schema while responding to the long-term goal of alignment with the Semantic Web via a metadata application profile (Greenberg et al. 2009, 209).

Brownlee (2009) focused on metadata challenges particular to research data in general purpose institutional repositories (IRs), and she concludes that metadata used to describe and manage data can be dissimilar to the metadata required for submission to IRs (2). At the University of Sydney, metadata describing research data was found to be stored primarily in databases or Excel spreadsheets with variability in use of and conformance to standards (Brownlee 2009, 3). The University of Sydney Library chose to address this variability and lack of standardization by ingesting the native metadata records as digital objects along with Dublin Core (DC) metadata describing these objects (Brownlee 2009, 6). Other options the library considered were mapping native metadata to DC without ingesting the original files and creating a custom metadata schema in the IR for each new native metadata element set (Brownlee 2009, 4-6).

Variability among numerous discipline-specific metadata schemes can create barriers to interdisciplinary research. Willis et al. (2012) examined nine metadata schemes for physical, life, and social science data to identify common objectives across schemes. Their analysis indicated that despite being constrained by historical disciplinary practices and workflows many metadata-driven goals are largely independent of scientific discipline or data type (32). Willis et al. (2012) identified eleven fundamental metadata goals for documenting scientific data in order to enable sharing across disciplines, including: abstraction, extensibility, flexibility, modularity, comprehensiveness, sufficiency, simplicity, data interchange, retrieval, archiving, and publication (27). Willis et al. (2012) suggest that further research on these common goals could help to break down the metadata barriers for interdisciplinary data sharing and research (33).