Managing the Data Commons: Controlled Sharing of Scholarly Data[1]

Kristin R. Eschenfelder
School of Library and Information Studies
University of Wisconsin-Madison
Room 4228 Helen C. White Hall
600 N. Park Street, Madison, WI 53706
eschenfelder@wisc.edu
(corresponding author)

Andrew Johnson
University of Colorado at Boulder
University Libraries
1720 Pleasant Street
Boulder, CO 80309
andrew.m.johnson@colorado.edu

ABSTRACT

This paper describes the range and variation in access and use control policies and tools used by 24 web-based data repositories across a variety of fields. It also describes rationale provided by repositories for their decisions to control data or provide means for depositors to do so. Using a purposive exploratory sample, we employed content analysis of repository website documentation, a web survey of repository managers, and selected follow up interviews to generate data. Our results describe the range and variation in access and use control policies and tools employed, identifying both commonalities and distinctions across repositories. Using concepts from commons theory as a guiding theoretical framework, in our analysis we describe five dimensions of repository rules that create and manage data commons boundaries: locus of decision making (depositor vs. repository), degree of variation in terms of use within the repository, the mission of the repository in relation to its scholarly field, what use means in relation to specific sorts of data, and types of exclusion.

KEYWORDS

Data sharing, data repositories, controlled data collections, use controls, access controls, data access polices, knowledge commons

** This is a preprint from April 2013. There is a large table (Table 2) that is saved as a separate file. The authoritative version of this article will be published in Journal of the American Society for Information Science and Technology sometime in 2014. There are no major data changes between this version and the final version, however the final version's analysis was further improved by reviewer comments. **

[1] This study was funded by the IMLS Laura Bush 21st Century Research Grant RE-04-06-0029-06. An earlier version of this paper was presented at the 2011 Annual Conference of the American Society for Information Science and Technology and the 2012 Libraries in the Digital Age Conference in Zadar, Croatia. This version contains additional interview and content analysis data and substantially different analysis. This paper has benefited from the feedback of two anonymous reviewers, Dorothea Salo, Kalpana Shankar, Gayle Nimmerguth, Puneet Kishor, study participants who kindly provided feedback on drafts, and conference attendees at ASIST 2011 and LIDA 2012.

INTRODUCTION

Web-based data repositories for “long lived” data have arisen in many disciplines to preserve data across changes in technology, accumulate data sets for data mining, and –most important to this paper - to promote wider sharing and reuse of data (National Science Board, 2005). Promotion of data reuse through “open” data has generated enthusiasm; for example, the U.S. government recently required federal agencies to provide public access to data from some federally funded research (Holdren 2013). This paper explores a less investigated aspect of data reuse -- the role of access and use controls to promote sharing and reuse. While this may seem counterintuitive, prior research suggests that providing tools to manage sharing might promote deposit of data, increasing its accessibility. For example, Pryor’s (2009) study of sharing amongst life science researchers found researchers wanted to know “who was using their data and for what purpose.” Tenopir et al. reported that many scientists believed they would share more data if they could place conditions on data access (2011).

While many repositories make their data sets accessible to any user without reuse restrictions, other repositories actively manage who uses data and control reuses. This paper explores this subset of repositories, which we call “controlled data collections” (CDC), and how CDC manage access and use of data. We define CDC as repositories where staff, or user communities, make and enforce rules to control who can access data or how data can be used.

We conceptualize CDC as “knowledge commons” as described by Hess and Ostrom (2007). They stress that knowledge commons are not synonymous with unrestricted anonymous public use; rather, knowledge commons may be bounded and their resources shared by *some* people for *some* uses (Hess and Ostrom, 2007). Research on commons governance identifies sustainable, successful commons as having clear boundaries, complex governance rules, and active management (Ostrom, 1990). This suggests that boundary setting functions of repositories may also be important for repository sustainability – a growing area of concern in digital collections (LeFurgy, 2009; Maron, Smith, Loy, 2009). For example, access and use controls may support sustainability through ensuring integrity and trustworthiness of data and business models that generate revenue. Economic downturns have increased concerns about the sustainability of digital collections.

Conceptualizing CDC as knowledge commons, this paper explores the rules that a purposeful sample of repositories have made about sharing data with *some* people for *some* uses. These “operational rules” define potential users’ interaction with the repository environment and data resources (Ostrom & Hess, 2007). For example, CDC rules may define who can access data, or what types of reuse is permitted.

From a repository best practices perspective, access and use rules are “community proxy” functions of repositories. Access and use control rules are especially important for repositories whose data have policy, legal or ethical considerations (NSB, 2005). Repository staff make rules, in conjunction with --

or on behalf of -- user communities, to ensure the integrity and trustworthiness of the repository (National Science Board, 2005). For example, repository rules might manage access to the repository as a whole, or manage access to particular records. A repository might specify different access rights for different sets of users (CCSDS, 2011).

Despite the potential importance of boundary setting, or access and use controls to deposit practices or repository sustainability, we know little about data repository access and use rules and how they vary across repositories or types of data. Increased knowledge about access and use control rules could contribute to best practices, increase data deposit, and promote critical thinking about governance of access and use control rules.

As a step toward these goals, this paper describes an exploratory study that investigated two questions:

RQ1: What technological and policy tools do repositories employ, or make available for their depositors to employ, to restrict access and use of data?

RQ2: Why do repositories control access to and use of data collections or provide means for their depositors to do so?

Our results describe the range and variation in access and use control policies and tools employed in our purposeful sample of CDC. Results also describe the rationale for restrictions provided by the repositories. Data analysis develops concepts that describe the boundary setting work of controlling access to data and use of data. The first, *locus of control* (LoC), refers to location of policy statements or decision-making about data. A repository may state policy at a repository, collection or data set level. The location of decision making may also vary: depositors may make decisions, repository managers may make decisions, or they may collaborate in decision making. LoC describes variance in whether depositors or repository managers decide (a) the terms of use for data and (b) whether to approve or deny specific access/use requests. The second concept, *repository mission*, distinguishes the degree to which managing access and use is part of the mission of the repository. The third concept, *degree of openness*, explores variance in how repositories and their managers interpreted what the terms “open” and “use” mean. The fourth concept, *terms of use (ToU) variability*, describes the degree to which terms of use vary between data sets within one repository. We also compare the arrangements repositories offer for managing very sensitive data. We then compare our findings on rationale for control with prior studies about researchers’ concerns about data sharing. The next section briefly reviews prior work on data sharing and data openness.

DATA SHARING AND RESTRICTING

The arguments for data sharing are well documented (Borgman, 2007, 2012; Piwowar et al., 2007; Tenopir et al., 2011); but, studies show that actual data sharing remains low (Blumenthal et al., 2006; Fry et al., 2009; Milia et al, 2012; Tenopir et al., 2011).

Past studies identified barriers to data sharing such as lack of time and resources, data misuse, legal issues and desire to ensure attribution (Borgman, 2007; Kuipers and van der Hoeven, 2009; Tenopir et. al., 2011). Other concerns include a desire to maintain exclusivity for publication, concern that reanalysis could lead to contrasting conclusions (Wicherts, Bakker, Molennar, 2011), large file sizes (Langille and Eisen, 2010), interference with patent opportunities (Pryor, 2009), and lack of standards (see Tenopir et al., 2011 for an overview).

Barriers to sharing are thought to vary somewhat between disciplines. For example, while privacy is a major concern in biomedical or social science research involving human subject data, intellectual property is a concern in humanities disciplines where primary source documents and publications are considered data, or in disciplines where data leads to commercial products (Borgman, 2009; Taylor, 2007; Blumenthal et al., 2006; Hilgartner, 1997). Concerns also vary within fields (e.g., biology) based on reward structures and other factors unique to sub-disciplines (Pryor, 2009). Past studies show that sharing often occurs through socially regulated informal exchanges. What is shared depends on the level of trust or “practices of trust” in the social network of researchers (Cragin and Shankar, 2006; Cragin, Palmer, Carlson and Witt, 2010; Hilgartner, 1997; Hilgartner and Brandt-Rauf, 1994; Pryor, 2009; Van House, Butler, Schiff, 1998).

Studies of sharing by institutions (i.e., digital cultural collections hosted by libraries, archives and museums rather than individual researchers or teams) found that common rationales for controlling access and use of works included the desire to control descriptions and re-representations of a work, avoiding legal risks and complexities, and ensuring social and financial credit for stewardship work (Eschenfelder and Caswell, 2010). Further, while many digital collection managers were concerned about “misuse” of their materials, what they conceived of as misuse varied. If one defines misuse broadly as a violation of some rule or norm, the rules/norms referred to by the term “misuse” included description or labeling standards, copyright law, terms of use, personal privacy expectations, cultural privacy expectations, formal promises made to participants, promises made to IRBs, and feelings of custodial responsibility (Eschenfelder and Caswell, 2010).

Degrees of open data

While the ideal of open data collections has generated enthusiasm, the question of what counts as “open” is complicated and involves issues of both access and use rules. Some argue that open data allows for unrestricted anonymous public use. For example “commons collections” typically have no access restrictions and no reuse restrictions. Other definitions permit restrictions; for example, the

Open Knowledge Foundation definition allows for acknowledgement requirements and cost recovery charges (OKF, 2012). Creative Commons offers at least seven different licenses reflecting different degrees of openness (Creative Commons, 2013).

Further, past research has shown that repository managers have widely varying personal understandings of what counts as open. For example, an archive manager whose historical images are available on the web might consider her collection open even though the terms of use of the photographs preclude any re-use without permission (Eschenfelder and Caswell, 2010).

METHODS[i]

This exploratory study of data repositories' use of access and use controls describes the range and variation in control policies and tools used across a variety of fields. We developed a purposive sample of CDC repositories in order to best describe range and variation in control policies and tools across fields. [ii] To identify CDC, we first generated a list of potential CDC from previous studies of data repositories, expert recommendations, and a review of journal and funder policies. In order to qualify as a CDC, repositories had to meet all the following criteria:

1. Accept data submissions from a broad audience (i.e., across institutions, data collection instruments or research projects).
2. Do not charge end users for access or use.
3. Control access or use of at least some data. Because the study was exploratory, we took an inclusive approach and included as many forms of control as possible. We defined controlling access or use as requiring some action or information from the end user beyond a command to access or download. Our inclusive approach means that our control restrictions range from the very onerous to things like registration requirements that some might consider inconveniences rather than restrictions. For this reason, we favor the term control over the term restriction. Many repository managers may agree that they employ controls, but may argue those controls are not restrictions.
4. Share access beyond the original depositor. We targeted what the NSB calls "intermediate" collections where the data's user community is larger than just one project (NSB, 2005).

The study sample was purposeful. We identified at least two data repositories in each field in order to ensure diversity. It was easy to find CDC in some fields and difficult in others. We eventually identified 24 CDC that fit the above criteria. We achieved diversity because our 24 CDC included repositories where almost everything was restricted and repositories where almost nothing was restricted.

We collected data about the 24 CDC through content analysis, a survey and interviews. We first conducted a structured content analysis of information available on repository public websites, producing a draft report on each repository.[iii] We sent copies of the draft reports to each repository and invited repository managers to correct and augment the reports via a web survey form.[iv] We received responses from 17 out of 24 repositories. An additional 18th respondent participated in a follow-up interview but declined to participate in the structured data correction. [v] After analyzing the web survey data, we conducted follow-up interviews with four repository managers from humanities, social sciences, health, and earth/space repositories. Interviews took place over the phone and varied in length from 20 minutes to an hour.

Analysis of the content analysis, survey and interview data raised new questions, and we re-examined documents describing data deposit policies for each CDC. We found that data deposit instructions often include information about access and use control options. To synthesize all the data from the different sources, we created case reports for each. Our case reports report on repositories within six broad fields, highlighted similarities and differences, and identified patterns and concepts. Finally, a copy of the paper was sent to each participating repository for comment.

Due to the methodological limitations imposed by purposeful sampling, the response rate to our survey and the structured nature of the interviews, our results are not representative of all CDC repositories and they are not statistically generalizable. As data from a purposeful sample, they illustrate range and variation across CDC. Our analysis generated new concepts that facilitate understanding of how CDCs manage access to data and use of data, and these theoretical concepts are transferable across a broader range of CDC (Lincoln and Guba, 1985).

RESULTS

Our analysis included the following number of repositories in each of six field-based groups:

- 6 social science,
- 4 humanities,
- 2 human health,
- 4 ecology,
- 3 chemistry or molecular data and
- 5 earth and space sciences repositories.

Our first finding was that data access and use controls were highly variable both across and within CDC repositories. Most of our CDC were “meta repositories” or repositories that hosted smaller archives managed by depositors. Terms of use (ToU) and access and use controls varied among these smaller archives. Further, the amount of restricted data in each repository varied greatly. In some, the majority of data were restricted; and while the repositories provided open metadata or open sample data, users had to create identifying accounts prior to accessing the rest. In other cases, the majority of repository data was available for public anonymous use, and only a small amount was restricted. Most repositories provided tiered service, with some data available to the public and some data requiring registration or approval. As one respondent explained, his repository has some “freely available data that anyone can access and use after agreeing to terms of use,” some “members-only data that only those at member institutions can access as a result of their membership,” and some “restricted-use data that prospective users must formally apply to use.” A small subset of our repositories required institutional-level membership for access.

In the results section, we first summarize the data in comparative data tables. Then we explore the findings in each field. We first describe the access and use controls employed by the CDC in each field, and then we summarize the CDC’s rationale for the controls. We generated this data from policy statements on websites, questions from the survey and follow-up interviews.

COMPARATIVE DATA TABLES

- Table 1 summarizes the policy documents used by repositories to document access or use controls.
- Table 2 (appendix) depicts the variation in controls employed across the repositories.
- Table 3 summarizes the rationales provided by repositories to explain their use of controls.

As indicated in Table 1, repository-level Terms of use statements (ToU) were used by all the repositories. Dataset-level ToU and copyright statements were more variable.

TABLE 1: Policy Documents by Field

Policy Documents	**SOC** ***n=6*** **# (%)**	**HUM** ***n=4*** **# (%)**	**Health** ***n=2*** **# (%)**	**Ecology** ***n=4*** **# (%)**	**Chem/Molecular** ***n=3*** **# (%)**	**Earth/Space** ***n=5*** **# (%)**
Repository-level terms of use statement	**6(100)**	**4(100)**	**2(100)**	2(50)	**3(100)**	**5(100)**
Dataset-level terms of use statement	5(83)	--	**2(100)**	2(50)	1(33)	**4(80)**
Copyright statement	3(50)	3(75)	--	1(25)	2(66)	2(40)

As shown in Table 2 (appendix), the most common access control method was embargo followed by required user registration. The most common use controls included requiring acknowledgement, prohibiting researchers from sharing data with unapproved individuals, and restricting data use to scholarly or educational use (except with special permissions). Looking across Table 2 suggests that SOC, HUM and Health CDC had more controls than repositories from other fields.

<insert Table 2 or file attached separately here>

Table 3 summarizes the rationale provided by CDC for restricting access or use of data. Our main finding is that because most repositories managed diverse data sets, they had different control rationales for the different sets. Further, we found that rationale varied across fields and that some fields referred to more rationale than other fields. We show the most common rationale in each field in bold.

Table3: Access and Use Control Rationale by Field

Rationale for controlling access or use of data	SOC n=6 # (%)	HUM n=4 # (%)	Health n=2 # (%)	Ecology n=4 # (%)	Chem/Molecular n=3 # (%)	Earth/Space n=5 # (%)
Ensure attribution	**6(100)**	2(50)	1(50)	2(50)	1(33)	3(60)
Protect sensitive/confidential information (other than personal privacy)	1(17)	**3(75)**	--	**3(75)**	**--**	--
Avoid misuse of data	4(67)	**4(100)**	**2(100)**	1(25)	1(33)	--
Ensure exclusivity	--	--	--	1(25)	**3(100)**	1(20)
Avoiding commercial research	4(67)	2(50)	--	1(25)	--	3 (60)
Intellectual property concerns	4(67)	**3(75)**	--	1(25)	1(33)	**4(80)**
Protect privacy (i.e., personal information)	**6(100)**	2(50)	**2(100)**	--	**--**	--
Use only permitted for certain types of research (e.g., only diabetes research) or research participants place limits on use of data	--	1(25)	1(50)	1(25)	--	**--**
Other	--	2 (protect physical archaeological sites)	**2(100) IRB approval**	2(50) quality control, generate use data	--	**4(80) export restrictions**

Findings from each field

In the following subsections, we describe the common access and use controls and control rationale within each field.

Social Science Data archives(SSDA)

The six SSDA included:

1. **ASSDA:** Australian Social Science Data Archive
2. **DANS:** Data Archiving and Networked Services – Discipline Social Sciences
3. **IQSS and Murray:** Institute for Quantitative Social Science Dataverse Network and the Murray Data Archive
4. **ICPSR**: Interuniversity Consortium for Political and Social Research - Main Data Archive.
5. **Odum**: Odum Institute for Research in Social Science Data Archive
6. **UKDA:** UK Data Archive Economic and Social Data Service

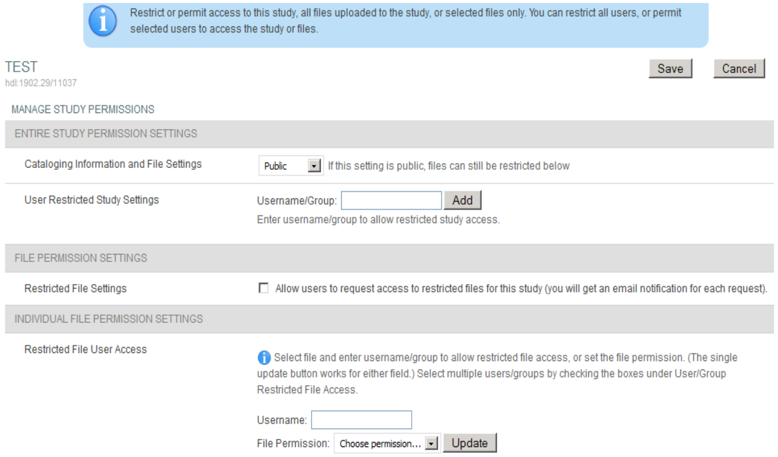
For access controls, all employed embargos, and required registration for access to some data (100%) Only two required registration to access any data (ICPSR, DANS). All held some data that required formal application. Several were membership organizations and users had to belong to member organizations in order to access most data (e.g., ICPSR, ASSDA).

In terms of use controls, all prohibited further dissemination of data beyond the approved user. All required acknowledgement. Most prohibited users from contacting individuals described in the data, limited use to scholarly use, required users to report back about data use, and prohibited further use beyond the approved study (67%). Additional controls that were not in our codebook, but which we saw in documentation, included prohibitions on re-identification of data, requirements for IRB approval and required signatures by the user's institutional research officers.

All enforced controls through repository level ToU (100%). Further, the SSDA that managed highly sensitive data (e.g., ICPSR, Murray, ASSDA) had dataset-level ToU that specified additional data set controls (83%).

In terms of Locus of control (LoC), ICPSR, ASSDA and UKDA had strong repository level LoC, and analysis suggests that they set ToU for data, sometimes in consultation with depositors. Repositories managed user applications for access to data sets. In contrast, Odum and IQSS (both of which used Dataverse software) had dataset level LoC and dataset ToU. Depositors developed their own ToU (see Figure 1), and users requested access from depositors through the Dataverse system.

FIGURE 1: Odum Dataset-Level Permissions



Restrict or permit access to this study, all files uploaded to the study, or selected files only. You can restrict all users, or permit selected users to access the study or files.

TEST
hdl:1902.29/11037

MANAGE STUDY PERMISSIONS

ENTIRE STUDY PERMISSION SETTINGS

Cataloging Information and File Settings: Public — If this setting is public, files can still be restricted below

User Restricted Study Settings: Username/Group: — Enter username/group to allow restricted study access.

FILE PERMISSION SETTINGS

Restricted File Settings: □ Allow users to request access to restricted files for this study (you will get an email notification for each request).

INDIVIDUAL FILE PERMISSION SETTINGS

Restricted File User Access: Select file and enter username/group to allow restricted file access, or set the file permission. (The single update button works for either field.) Select multiple users/groups by checking the boxes under User/Group Restricted File Access.

Username:

File Permission: Choose permission...

We also observed variation in management options for risky data. Odum, IQSS and DANS forbade deposit of high risk data. Other archive websites (e.g., ASSDA, ICPSR) reviewed deposits for sensitivity, and would provide advice to depositors about "removing, masking, or collapsing variables" to reduce risk (ICPSR 2012 Restricted Data). Some offered further protection. For example, ICPSR only permitted access to certain sensitive data available through its "secure data enclave" space. (ICPSR 2012

Restricted Data; ICPSR Enclave Data) and DANS would only send risky data to users on physical media (DANS EASY).

RATIONALE FOR CONTROL - SSDA:

The most prominent rationales (see Table 3) were privacy (of individuals, groups and organizations) and attribution (100%). Misuse was another prominent rationale (67%). One form of misuse that participants described was re-identification of study participants. Certain types of uses were prohibited; for example, ICPSR documented that its data could not be used for law enforcement, administrative or commercial purposes (ICPSR "Restricted Data"). Another prohibition was use by unauthorized users. Many of the repositories held some licensed or traded data sets subject to user restrictions (DANS, Odum, Murray). As one respondent explained, "We belong to several membership organizations and pay for access. This is only for our faculty and staff so we cannot allow non-[campus] users access to it." At least three repositories had "members-only" restrictions (e.g., ICPSR, UKDA, ASSDA).

HUMANITIES

The four humanities archives included:

1. **DANS EDNA**: Data Archiving and Networked Services - e-Depot for Dutch Archaeology (EDNA) included archaeological data.
2. **EVIA:** Ethnographic Video for Instruction and Analysis Data Archive included videos related to ethnomusicology.
3. **OTA**: Oxford Text Archive included recordings and documents of linguistic data.
4. **tDAR**: The Digital Archaeological Record included archaeological data.

Common access contols used by Humanities repositories included embargoes and registration to access some data (100%). Only one required registration to access any data (DANS). Most required formal application to access certain data (75%). Some (e.g., tDAR, OTA) offered a good deal of public anonymous data.

The two repositories with archaeology data (tDAR and DANS EDNA) restricted access to some data to official archaeologists. The tDAR system asked archaeologists to enter a "Registered Professional Archaeologist" number (tDAR Register). DANS EDNA staff explained that that they verified user identities as archaeologists by checking to see if they worked for official archaeological organizations or were enrolled students of archaeology programs.

The most common use controls (75%) included prohibiting further dissemination of data beyond the approved user, requiring acknowledgement and limiting to scholarly/educational use only.

All used repository level ToU, but none employed data set level ToU (see Table 1). The LoC for setting controls varied. Both EVIA and tDAR had strong repository LoC– both had strong repository level policies. tDAR presented itself as a largely "commons collection" repository. Its repository level ToU dictated that all deposits would be subject to an adapted Creative Commons license. The LoC at DANS and OTA was weighted toward depositors. While no formal dataset ToU were posted, depositors could choose different levels of access and use control and could approve or deny access requests.

RATIONALE FOR CONTROL - HUMANITIES:

All had concerns about misuse of data (100%), but misuse meant very different things to the different repositories. For repositories with archaeological data (tDAR, DANS EDNA), misuse referred at least in part to use of archived location data to cause damage to archaeological sites. EVIA, a cultural anthropology/musicology archive, perceived at least two major types of misuse. EVIA managers were concerned about culturally insensitive uses, such as parodies or re-interpretations of materials without accompanying cultural information. They also expressed concerns about commercial or non-educational misuse due to the "educational use only" terms of use (ToU) of the site.[vi] For OTA misuse referred to concerns about unauthorized users' use of licensed or traded data sets. Misuse may also have referred to privacy concerns - OTA's terms of use (ToU) forbade users from revealing information about individuals or organizations contained in data (OTA User Agreement).

Three indicated intellectual property concerns as a motivator (75%). EVIA videos sometimes inadvertently contained copyrighted and trademarked materials such that use needed to be restricted to educational use. OTA, as mentioned earlier, hosted licensed or traded data that carried access and use restrictions, and DANS warned users to respect the copyrights of depositors. tDAR, as a predominately open access archive, did not indicate IP concerns.

HEALTH

The two health repositories included:

1. **ANON-Bio:** (Anonymous by request of repository) contained human health study data.
2. **BioLINCC:** Biologic Specimen and Data Repository Information Coordinating Center included both study data and a means of requesting specimens.

Our two health repositories had more access and use controls than other field groups (see Table 2). In terms of access controls, both had embargos, required registration to access any data in the repository system, and required formal applications to use any data.

In terms of use controls, both prohibited further dissemination of data beyond the approved researcher and the approved study. Both required acknowledgement, required destruction of the data after the study, prohibited contacting individuals depicted in the data, and required secure data storage.

Applications for data use were extensive, involving PI credentials, descriptions of the proposed use, proof of local Institutional Review Board approvals, security plans, and signed ToU agreements. ANON-Bio required potential data users to submit signatures through a large granting agency's application review system, thereby requiring that the potential user have a credentialed account on the granting agency's system. In contrast, BioLINCC data applicants could submit scanned signed forms through the BIOLINCC system itself.

In terms of Locus of Control, both health repositories had strong repository LoC, with formal repository ToU supplemented by conditions laid out in the original informed consent for data collection and data set level ToU (Table 1). In contrast to other fields, in health, requests for data use were vetted by a committee of health research professionals, rather than being forwarded to depositors or managed solely by repository staff.

While many repositories in other fields restricted use to scholarly use, both health repositories explicitly allowed commercial use of their data (Table 2). A repository manager explained that restricting commercial use could discourage innovation due to the important work performed by commercial bio-tech companies. Not all data was available for commercial use however; for example, within BioLINCC, data from studies where participants did not explicitly permit commercial use was sequestered.

One variation we observed was whether the health repository would approve data access and use for teams of people, or whether it would require each member of a team to apply separately. BioLINCC would approve teams for some data sets, but would only approve access and use of highly sensitive genetic data to individual researchers (NHLBI, 2009 "Guidelines for NHLBI Data Set Preparation"). ANON-Bio would only approve individual researchers.

RATIONALE FOR CONTROL - HEALTH:

As depicted in Table 3, health repositories had a strong "other" rationale - both indicated ensuring home institution "institutional review board" approval as a motivator for controls. In addition, both indicated misuse as a motivator. For these repositories, guarding against misuse meant (in part) ensuring that data use would comport with limitations described in the original data collection's informed consent. Informed consent requires that researchers explain the purposes of the research and how the data will be used; but re-users of that data may wish to make use of the data for some other purposes. Therefore, repositories holding human subjects data must ensure that the downstream uses of data comport with the informed consent. As one respondent explained, "participants' informed consent must be respected."

Protecting study participants' privacy was also a prominent rationale. Like social science data repositories, the health repositories sought to ensure that deposited data had been de-identified. To further protect privacy, ANON-Bio strongly recommended that US depositors of data get a "certificate of confidentiality." These certificates shield data keys (i.e., indexes that link participant codes or pseudonyms to real contact information) from legal actions such as court orders.[vii]

Ecology

The four archives included:

1. **KNB:** Knowledge Network for Biocomplexity collects data about plants and animals including microbes and viruses.
2. **NPS-IRMA:** National Park Service Natural Resource Information Portal includes data, publications, reports, and species lists.
3. **SeaDataNet:** contains marine and ocean data on water heat, salt concentrations, sea levels, currents and ecosystems.
4. **VegBank:** a database of vegetation plots and classification and taxonomic data.

Looking at access controls, all repositories (100%) offered embargoes and registration requirements to control access to some data, but only one (SeaDataNet) required registration to access any data. IRMA and VegBank had substantial anonymous publicly accessible data. None required formal applications to access data.

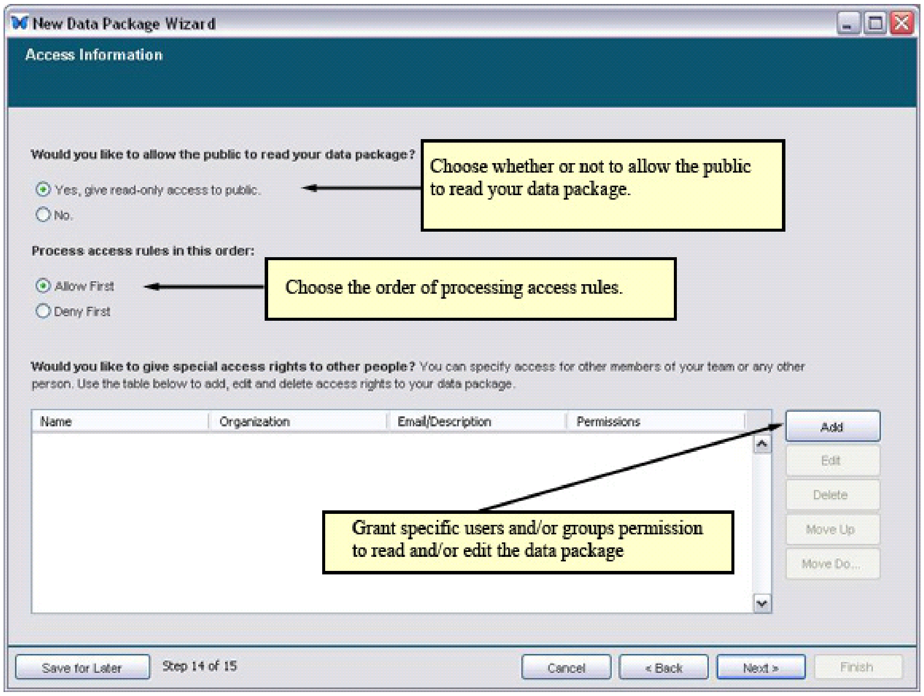
In terms of use controls, no control was employed by more than half the repositories. Fifty percent prohibited further dissemination beyond the original user, limited use to scholars, or required acknowledgement. KNB and SeaDataNet had the most limitations. Both were developed for distributed teams of researchers, so the controls may stem from assumptions about who would be using the repository data.

In general, the ecology repositories tended to have fewer use controls than the previously described fields, and several of the repositories had explicit missions to provide open data. For example, VegBank and NPS-IRMA required that depositors justify any requested use controls, suggesting a preference for unrestricted data. But VegBank provided tools for depositors to scramble longitude and latitude data to protect location data of rare or endangered materials. Similarly, NPS IRMA, as a US government database, had almost no limitations on use except for sensitive data.

The LoC for setting ToU was mixed. On one hand, NPS-IRMA and VegBank had strong repository level LoC with the emphasis being on open data. On the other hand, KNB and SeaDataNet had stronger depositor LoC, providing tools for depositors to set controls. In KNB, depositors could limit access to depositor-specified groups of registered users. Data users would negotiate permissions directly with the data depositor. KNB depositors set permissions via an access system (see

Figure 2) (KNB Morpho). KNB also provided a free text box for depositors to describe ToU at the dataset-level.

FIGURE 2: KNB Access Controls



New Data Package Wizard

Access Information

Would you like to allow the public to read your data package?

● Yes, give read-only access to public.

○ No.

Process access rules in this order:

● Allow First

○ Deny First

Would you like to give special access rights to other people? You can specify access for other members of your team or any other person. Use the table below to add, edit and delete access rights to your data package.

Name	Organization	Email/Description	Permissions

Step 14 of 15

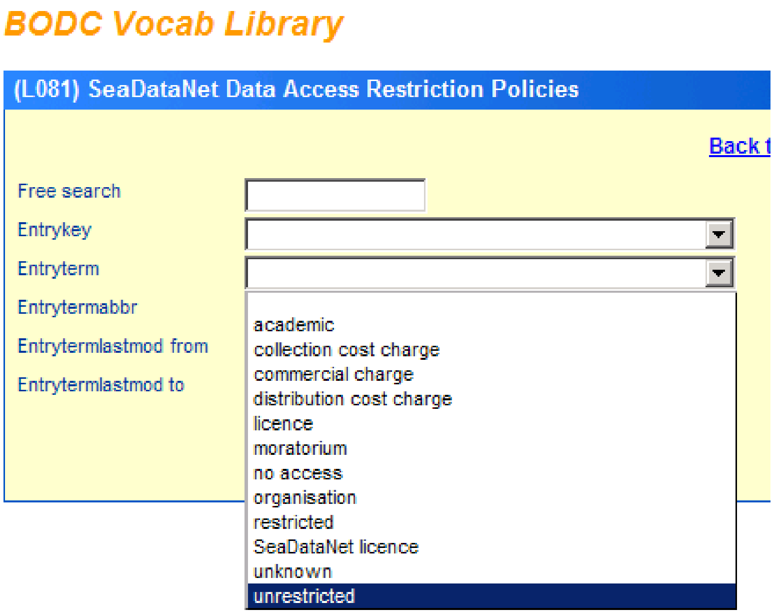
Choose whether or not to allow the public to read your data package.

Choose the order of processing access rules.

Grant specific users and/or groups permission to read and/or edit the data package

Similarly, at SeaDataNet, ToU were set by the depositor through a controlled vocabulary (see Figure 3). Data use requests however, would be reviewed by network coordinators rather that data depositors (SeaDataNet L08).

FIGURE 3: SEADATA NET Deposit Access Metadata



BODC Vocab Library

(L081) SeaDataNet Data Access Restriction Policies

Rationale for control - Ecology:

The most common rationale (75%) was protecting location data about endangered plants, animals, and (in one case) cultural resources information. For example, VegBank explained the need to "protect endangered species or rights of private land owners" whose property contained species of interest (VegBank Privacy Policy).

Gathering information on repository use and users (often via registration or attribution requirements) also emerged as a motivator. Repository managers testified that controls like registration systems provided data that depositors used to explore potential collaborations, that repositories used to gain insight into user requirements, and that funders valued as evidence of impact.

Chemistry/Molecular

The three repositories included:

- **CCDC:** The Cambridge Crystallographic Data Centre collects small molecule crystal structures or crystallographic data.
- **PRIDE:** Proteomics Identifications Database collects protein data structures.
- **Commons:** Proteome Commons Tranche Repository also collects protein data structures.

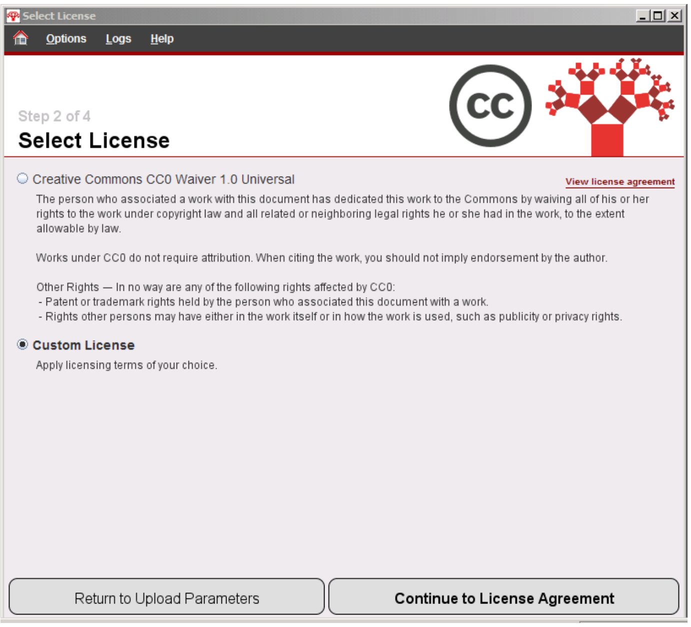
Unusual within our sample, in addition to offering controlled free access to data, the CCDC generated revenue from subscriptions for the Cambridge Structural Database System (CSD) database.

In terms of access controls, all three of the chemistry/molecular archives used embargoes and required registration to access restricted data. For example, PRIDE allowed deposit of "private data" to be "shared by a collaboration." The depositor assigned specific system-registered "collaborators" access rights to the data based on whatever terms and conditions the collaborators established outside the system (PRIDE User Manual). Only CCDC required application prior to accessing any data. Users with "bona fide" research requests could request free access to a limited number of crystal structures through a request form that included personal and institutional information. The form was reviewed by CCDC staff.[viii] Submission of data to two repositories (CCDC and PRIDE) were tied to publication in journals and offered embargo until publication.

In terms of use controls, CCDC limited use to research use only, requested acknowledgement and restricted further dissemination or copying. In contrast, Commons had no stated policy controls, and PRIDE only required acknowledgement.

The LoC within these repositories was mixed. CCDC had a strong repository level LoC. All CCDC requests involved agreeing to the repository ToU. PRIDE provided no formal way to dictate data-set level terms of use but allowed depositors to control who could access data, thereby arguably requiring potential users to seek agreements directly with depositors. The Commons allowed depositors to set ToU and potential users requested access from depositors. Uniquely, CC0 was the default license type for the Commons, but as shown below, depositors could also choose the "Custom License" option to specify other terms.

FIGURE 4: Choosing A CC0 License on Proteome Commons



Step 2 of 4

Select License

○ Creative Commons CC0 Waiver 1.0 Universal

The person who associated a work with this document has dedicated this work to the Commons by waiving all of his or her rights to the work under copyright law and all related or neighboring legal rights he or she had in the work, to the extent allowable by law.

Works under CC0 do not require attribution. When citing the work, you should not imply endorsement by the author.

Other Rights — In no way are any of the following rights affected by CC0:
- Patent or trademark rights held by the person who associated this document with a work.
- Rights other persons may have either in the work itself or in how the work is used, such as publicity or privacy rights.

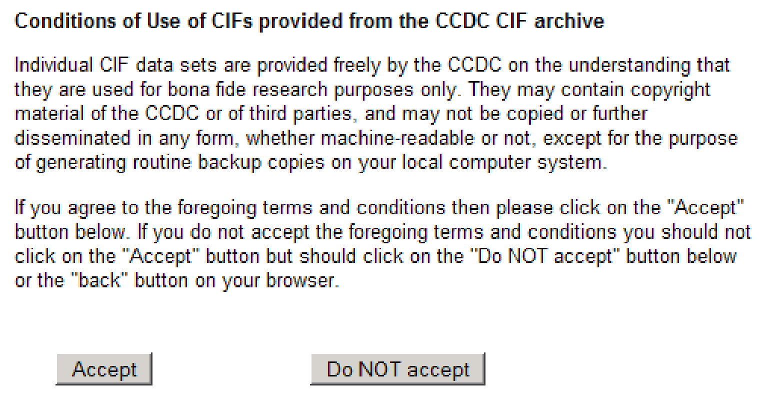
● **Custom License**

Apply licensing terms of your choice.

RATIONALE FOR CONTROL – CHEMISTRY/MOLECULAR

Submission of data to CCDC and PRIDE were closely tied to publication in journals; therefore ensuring exclusivity was a prominent rationale (100%). The partially commercial nature of the CCDC may also have created unique IP concerns. The CCDC, in addition to providing free access to post-1994 “supplementary” data, also licenses access to a commercial database of its full set of “validated” data (CCDC, 2012). Access to its full database requires institutional membership. It is not clear to what degree the free access to supplementary structures can substitute for membership access. The CCDC ToU for free data sets limits data requests to “bona fide” research (see Figure 5).

FIGURE 5: CCDC Data Request Form



Conditions of Use of CIFs provided from the CCDC CIF archive

Individual CIF data sets are provided freely by the CCDC on the understanding that they are used for bona fide research purposes only. They may contain copyright material of the CCDC or of third parties, and may not be copied or further disseminated in any form, whether machine-readable or not, except for the purpose of generating routine backup copies on your local computer system.

If you agree to the foregoing terms and conditions then please click on the "Accept" button below. If you do not accept the foregoing terms and conditions you should not click on the "Accept" button but should click on the "Do NOT accept" button below or the "back" button on your browser.

Earth and Space repositories

The five repositories included:

1. **SMOKA:** Japan Astronomical Data Archives Center Science Archive holds data from telescopes and observatories in Japan.
2. **BADC:** British Atmospheric Data Centre holds UK-based atmospheric sciences research data.

3. **NCAR CISL-RDA:** National Center for Atmospheric Research Computational and Information Systems Laboratory Research Data Archive holds meteorological and oceanographic observation data, remote sensing data and models.
4. **NSIDC:** National Snow and Ice Data Center collects data on frozen regions and their effect on climate.
5. **NEODC:** Natural Environment Research Council Earth Observation Data Centre holds UK-based earth observation data.

We found much variation, both between and within repositories. In terms of restricted data, some repositories hosted few restricted data sets, while others hosted many. Further, controls varied within repositories. Some data sets had light controls (e.g., academic use only), while others had onerous controls (e.g., requirement that the original study program manager approve all further uses) (NERC 2012 Access).

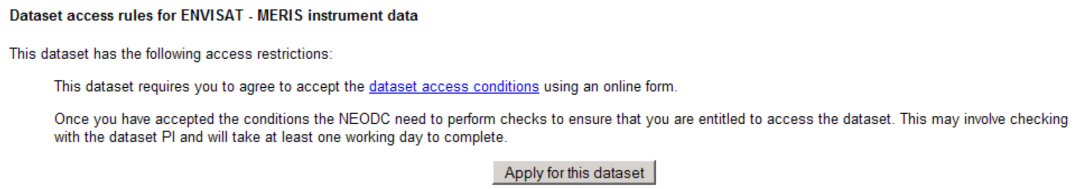
All the repositories held at least some data that required registration (100%). In addition, three repositories required registration to access any data (CISL-RDA, NEODC, SMOKA). Only one repository (NEODC) contained data sets that required formal application. Application varied by data set. Most had web-based agreements, but one data set required a signed agreement delivered via postal mail. We saw a few examples of access controls that restricted users to grant recipients from certain agencies, or required that users' institutions develop reciprocal sharing agreements (BADC, NEODC).

In terms of use controls, all repositories required acknowledgement (100%). Three (60%) limited use to scholarly use (BADC, NEODC, SMOKA). Fewer repositories (40%) prohibited further dissemination beyond the approved individual or the approved study, or required any reporting on data use.

Our data suggest that the repositories outside the US tended to have more controls than those in the US. But even within the UK repositories, which held some data with heavier controls, some data had light controls. Conversely, the US repositories, which tended to have few controls, also held some data sets with more controls.

All employed repository-level ToU, and four employed dataset-level ToU. The LoC varied; US repositories had repository-level LoC emphasizing openness (NSIDC, NCAR). Data set ToU were rare and not very visible. But, staff reported that they would sometimes restrict data at the request of the data owner, and in these cases it is not clear whether the repository or depositor made decisions about who could access and use the data. The LoC in UK repositories was more varied. They included dataset ToU customized to the needs of specific depositors, suggesting that depositors made decisions about ToU. It seemed that in most cases the repository would handle requests for use, but we found at least some NEODC datasets that required PI approval (see Figure 6).

FIGURE 6: NEODC Data Access Application



Dataset access rules for ENVISAT - MERIS instrument data

This dataset has the following access restrictions:

This dataset requires you to agree to accept the dataset access conditions using an online form.

Once you have accepted the conditions the NEODC need to perform checks to ensure that you are entitled to access the dataset. This may involve checking with the dataset PI and will take at least one working day to complete.

Rationale for Control – Earth Space

A unique rationale for control emerged from our "other" category: 80% mentioned export restrictions. For example, UK repositories (BADC, NEODC) restricted access to some data to UK researchers. In the US, the Department of State's Trade Controls "International Traffic in Arms" Regulations forbade the export of munitions to certain nations. At the time of data collection, the regulation included some satellite data. This required some US repositories to block access to some data. As one US manager explained, the regulations pressured them to ensure no users from the "unfavored nation list or terrorist lists" had access to data.

Another prominent rationale was concern over IP or informal ownership claims (80%). Analysis suggests all archives but SMOKA contained at least some licensed, traded or co-op data sets. Repositories traded access to data under ToU that made claims about data ownership. As one manager explained, trading agreements may "bear some similarity to intellectual property rights." Similarly, co-op data sets involved ownership claims. As one repository manager explained, co-op members have free access, but "everyone else should have to pay to have access." Further, some data had commercial value. For example, one BADC/NERC document justified restrictions because the data were potentially "commercially valuable" (NERC 2012 "Meteorological Office").

Discussion

This section draws on the data, past research results and information commons theorizing to generate new transferable categories and concepts (Lincoln and Guba, 1985). It is organized into three parts. The first describes five dimensions for analyzing controlled data collections (CDC). The second compares our data on motivations for control with prior findings about barriers to data sharing, highlighting new categories and distinctions. The paper concludes by outlining areas of future research.

Dimensions of Data Repository Access and Use Control Arrangements

Past work on commons led us to expect a "rich variety" of rules across our 24 CDC (Hess and Ostrom, 2007). We were surprised, however, by the diversity of rules *within* each repository. Given the within-repository diversity of rules, our analysis suggests the following five dimensions are important to understanding how CDCs manage access and use of data.

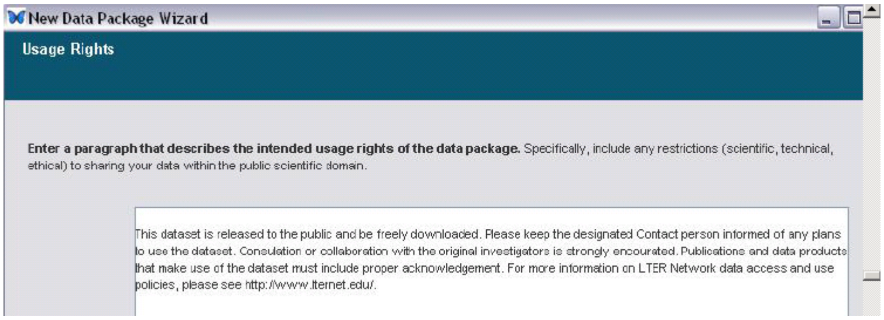
1. <u>Locus of control (LoC):</u>

One variation we saw among repositories was in Locus of Control (LoC) for (a) deciding terms of use for data, and (b) negotiating specific access/use requests. LoC refers to location of policy statements or decision-making about data. We found that repositories may state policy at repository, collection, and data set levels. We found that decisions about access and use may be made by repository managers, depositors, granting agencies, or via a collaboration of these parties.

LoC for terms of use: Most repositories had both repository and data set-level ToU. But, in some cases the ToU were largely determined by the repository while in other cases the ToU were largely determined by the depositor. For example, in

Figure 7 below, the KNB repository requires depositors to set ToU for their data sets. KNB provides default language (leaning toward open), but depositors can customize ToU for their data.

FIGURE 7: KNB Depositor TOU Entry Field



New Data Package Wizard

Usage Rights

Enter a paragraph that describes the intended usage rights of the data package. Specifically, include any restrictions (scientific, technical, ethical) to sharing your data within the public scientific domain.

This dataset is released to the public and be freely downloaded. Please keep the designated Contact person informed of any plans to use the dataset. Consulation or collaboration with the original investigators is strongly encourated. Publications and data products that make use of the dataset must include proper acknowledgement. For more information on LTER Network data access and use policies, please see http://www.lternet.edu/.

LoC for permissions decision making: The second aspect of LoC is where decisions about permissions occur. In some cases repository managers decide on requests for permission while in other cases requests are forwarded to depositors, or even granting agencies, to decide.

Proxy Archives: We label repositories with strong repository LoC for ToU and permissions as "proxy archives." These repositories create policies and make permissions decisions as a community proxy function on behalf of depositors.

Meta Archives: We label repositories with strong depositor LoC for ToU and permissions decisions as "meta archives." These organizations may see themselves as providing infrastructure through which other individuals or groups can develop and manage archives.

Negotiated Archives: We label repositories where staff negotiate with depositors about ToU and permissions as "negotiated archives."

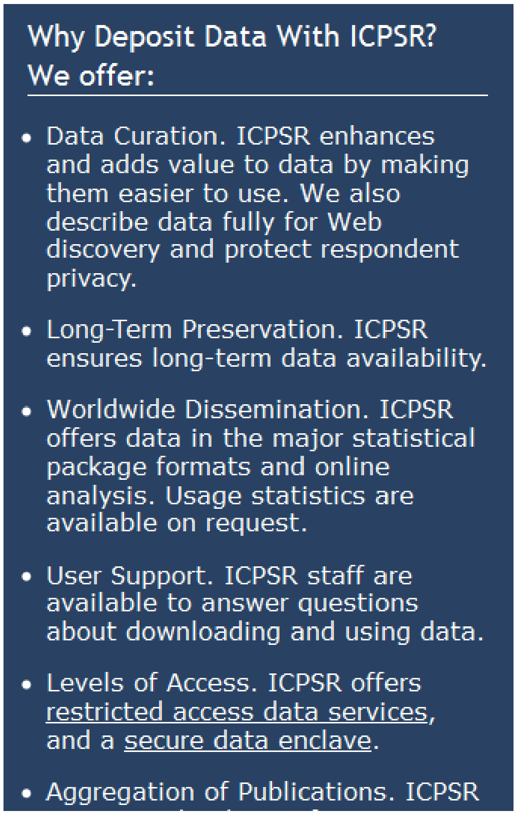
Of course given the diversity we observed within archives, many archives will not fall neatly into one of these categories; it may be that different collections within archives are managed in different ways.

2. <u>Repository mission</u>:

Another distinction we saw is the degree to which managing access and use is part of the mission of the repository. Some repositories saw themselves as primarily hosting data with no controls. They viewed data controls as an undesirable necessity to comply with regulations, license/trade terms, or to appease hesitant researchers (e.g., NSIDC, OTA, VegBank). Often in these cases, repositories did not actively advertise their ability to restrict data in deposit materials.

On the other hand, some repositories saw control as a *core responsibility* of the archive. For some repositories access and use controls are an advertised feature – a reason to entrust data to the repository. For example, ASSDA notes, "Perhaps a better way to protect the confidentiality is to set strict access conditions." (ASSDA Why Deposit?; Sharing and Access Conditions) In another example, in Figure 8, ICPSR prominently advertised different levels of access controls available to depositors (ICPSR "Deposit Data and Findings").

FIGURE 8: ICPSR Why Deposit? (Underlining Added for Emphasis)



Why Deposit Data With ICPSR?
We offer:

- Data Curation. ICPSR enhances and adds value to data by making them easier to use. We also describe data fully for Web discovery and protect respondent privacy.
- Long-Term Preservation. ICPSR ensures long-term data availability.
- Worldwide Dissemination. ICPSR offers data in the major statistical package formats and online analysis. Usage statistics are available on request.
- User Support. ICPSR staff are available to answer questions about downloading and using data.
- Levels of Access. ICPSR offers restricted access data services, and a secure data enclave.
- Aggregation of Publications. ICPSR

Some repositories may see themselves as a mix. They may feel responsibility to restrict some data but they may also seek to reduce control of other data. For future research, it would be helpful to know what percent of repository data is subject to what sort of control (e.g., licensed/traded, export restriction, funding agency requirement, insistence of researchers).

We observed the following control arrangements in our data:

- Controlling who can access and use data is a core part of the repository mission for at least some data. The ability to control is advertised to depositors/users. (e.g. EVIA, BioLINCC, AnonBIO, ICPSR).
- Repository makes no restrictions, but provides/advertises control tools that depositors may choose to use (e.g., Commons, DANS, IQSS, Odum, PRIDE, VegBank).
- Controlling access and use is only done by the repository in rare instances, at the request of depositors (e.g., NCAR-RDA, NSIDC).
- No data sets are restricted by the repository, but repository deposit rules preclude inclusion of certain types of data or require masking or removal of data (e.g., DANS, IQSS, Odum).
- Controlling access and use is actively discouraged by the repository via special charges or other burdens to the depositor (e.g., OTA).

Any one repository may have different data sets that fall under differing control conditions from the list. Future researchers could ask repository managers to estimate what percent of their data falls into the above categories.

3. Managing highly sensitive data:

We observed a variety of arrangements for handling sensitive data across CDC from different fields (beyond the basic tools of controlling who could use the data and setting terms of use).

Refuse to host: One strategy was to refuse to host sensitive data. For example, several social science repositories forbade its deposit.

Masking: Repositories in social sciences, health and ecology required depositors to remove data, mask data, or otherwise obscure sensitive data.

Contractual obligation: Archives hosting sensitive data sometimes required extensive paperwork. For example, NPS-IRMA documentation noted that any use of sensitive data required a signed confidentiality agreement. The health repositories required numerous signed agreements for access and use. ICPSR and other SSDA also required extra paperwork for access to sensitive data sets.

Non-digital transfer: Several archives stated that they would only transmit sensitive data sets via physical media to avoid network exposure.

Enclave: For certain data sets, ICPSR required researchers to travel to use data under the controlled conditions of a secure data enclave.

4. Degree of variation in ToU within a repository:

Some repositories had most data available under a basic ToU or perhaps two or three standard ToU. In other repositories, ToU varied greatly from data set to data set. While some repositories might have a set menu of control options to draw on, repositories that allowed depositors to customize ToU might produce an infinite array of control conditions. This level of variation likely has pros and cons. While highly customized control conditions might please depositors, they may make ongoing repository management more difficult.

5. What do the terms *open, restriction* and *use* mean across different repositories?

Our results illuminate complexities hidden with the term open data. Our repository managers showed varying interpretations of what counted as open, and what counted as a restriction. For example, some saw our questions about restrictions as not applying to them because their repository controlled few data sets. In their interpretation, because the vast majority of data sets were open, their repository was open. In another example, some repositories required registration but only used registration to means generate usage statistics. Sometimes these repository managers did not consider their registration requirements to be restrictions.

The variance we saw in interpretations of openness and restrictions is not surprising given the broader debate about what counts as "open data." If one assumes that open data allows anonymous public use, then user registrations qualify as restrictions. Requiring registration might be analogous to a Creative Commons license requiring attribution, as opposed to a CC:0 license which waives all rights. Other definitions of open allow for cost recovery charges (e.g., OKF). Another variance we observed was treatment of commercial use. Some definitions do not permit any restrictions on commercial use (i.e., OKF, 2012 some CC), but many of our repositories (who likely consider themselves open) did not permit commercial use. We suggest there is no simple definition of open data archives, but rather different types of openness, similar to the different Creative Commons licenses.

It is also important to clarify what *use* means in the context of openness. We defined use as the ability to download and process data locally. Some repositories, however, did not allow downloading but instead offered online analysis tools (e.g., EVIA, ASSDA). In these cases, one could still analyze data, but one could not download data. This is a distinctly different type of openness because it precludes redistribution.

Based on our data, we suggest the following menu options of data access and use control, roughly ordered from least restrictive to most restrictive:

- Public anonymous downloading.
- Sensitive data masked or removed prior to making data publicly available.
- Registration, then downloading of any data.
- Authorized users (i.e., pre-approved subset of potential users or members) get access for downloading of all data.
- Individual approval, then downloading of any data.
- Request for specific data set use requiring basic registration (with depositor or repository).
- No downloading; online analysis or streaming use only.
- Request for specific data set use with formal application (to depositor or repository).
- Formal application for specific data set to repository with IRB approval.
- Formal application for specific data set to repository with IRB approval and signed institutional agreement.

Any one repository might choose different options for different data sets, so investigating the control practices of repositories might require documenting multiple categories across different data sets.

WHY RESTRICT? OR, BARRIERS TO DATA SHARING

Table 5 summarizes our data on rationale for controlling data. The middle column notes the dominant rationale from our content analysis of repository documentation. Because our sample was purposeful, these results are not representative of the rationale of all repositories in a given field; however, they inform us about range and variation of rationale across fields. Table 3 informs us about range and variation within fields. The final column in Table 5 highlights rationale that were not included in the original deductive codebook, but that emerged inductively from our analysis of data. Both sets of rationale should be included in future research.

TABLE 5: Summary of Repository Rationale for Controlling Access and Use of Data

Repository Group	Dominant Rationale from Deductive Content Analysis of Documentation	Rationale from Inductive Analysis of Surveys, Interviews and Documentation
Social Science	Study participant privacy; Attribution	Institutional Review Board approval; Members-only/authorized user restrictions; Avoid certain uses (e.g., law enforcement)
Humanities	Avoid misuse; Protect sensitive/confidential information	Protect archaeological sites; ensure culturally sensitive uses; authorized user restrictions
Health	Avoid misuse; Protect privacy	Institutional Review Board approval; compliance with informed consent
Ecology	Protect sensitive/confidential information	Location of endangered or valuable plants and animals; privacy of landowners; ensure data quality
Chemistry and Molecular	Ensure exclusivity	--
Earth and Space	IP concerns	Export restrictions; authorized user restrictions

Our results confirm prior findings about barriers to sharing data, but also add new dimensions in the areas of privacy, attribution and misuse. Our data confirm that privacy is a strong rationale in fields involving human subject data (e.g., social sciences, human health) (Tenopir et all, 2011). Our findings however, also show that privacy concerns extend to humanities archives that contain oral histories and to ecological archives where location information might create privacy issues for property owners.
Our results also confirm that attribution (of the depositor and the repository) is an issue in data sharing (Tenopir et al., 2011). Further, social credit is important to repositories in their quest for funding (Eschenfelder and Caswell, 2010). We found that attribution requirements were common in ToU; but, repository managers did not often indicate attribution as a rationale for control. This discrepancy may

stem from respondents seeing registration systems as a means to ensure attribution/credit, but not as a control tool. Our findings also confirm prior findings that misuse is a common concern (Eschenfelder and Caswell, 2010; Pryor, 2009; Borgman, 2007). But our findings demonstrate how interpretations of misuse vary widely across and within fields. Misuse meant many different things – not all of which were forbidden by ToU. We observed the following misuse concerns among data repositories:

- Identification of study participants or sensitive locations.
- Culturally insensitive, demeaning or damaging uses that threaten participants, materials or relationships.
- Uses not conforming to limitations in informed consent or other arrangements governing data collection.
- Use by undesired users (e.g., commercial, law enforcement).
- Use by unauthorized users (license/agreement violations).
- Further distribution or re-use without permission.

Some concerns identified in prior studies did not appear in our data (i.e., lack of resources, reanalysis leading to contrasting results, disruption of patent opportunities, lack of standards). This may stem from our focus on repository rationale rather than researcher rationale. Many of our repository rationale would likely not extend to researchers; for example, repository concerns about complying with data license or agreement terms or export restrictions may not occur to individual researchers. This suggests that future studies consider two overlapping sets of barriers to sharing: barriers to individual researchers and barriers to repositories.

Another control rationale that deserves further attention is data quality control. One might assume that concerns about quality would be common; after all, repository best practices define "quality control" policies as a characteristic of trustworthy digital repositories (CCSDS, 2011). Federal agencies have data quality requirements. But quality concerns were not a common rationale in our repository content analysis data; further, it did not emerge as a concern in interviews with repository managers. That being said, embargoes were a relatively common access control option. Lack of data about quality control may stem from our methods: quality control was not listed as a rationale choice in our survey. Further, it may be that some repositories do restrict access until quality is assured, but do not bother to describe the practice in their documentation. Another possible explanation is that some repositories do not take active quality control steps. Meta-repositories like Dataverse projects may not see data stewardship services like quality control as part of their mission.

Conclusion and Next Steps

Past commons research suggests that rules governing commons resource usage are integral to sustainability (Ostrom and Hess, 2007). Using the framework of knowledge commons as a theoretical lens, this paper explored the rules that a purposeful sample of controlled data collection (CDC) repositories made about sharing data with some people for some uses. The analysis described *locus of control* (LoC), or the variance in whether depositors or repository managers decide (a) the terms of use for data and (b) whether to approve or deny specific access/use requests. It described how *repository mission*, or the degree to which managing access and use is part of the mission of the repository, varied widely across CDC and even within a repository. It pointed to how the *degree of openness,* or what "open" and "use" meant to different repository managers, varied widely. It pointed out the surprisingly high *terms of use (ToU) variability*, between data sets within one repository. It also described the arrangements repositories offered for managing very sensitive data. When we compared our findings about rationales for control with prior studies on researchers' concerns about data sharing, analysis showed areas of overlap, but also some control concerns that may be unique to repositories as institutions.

One question this study cannot answer is what is the relationship between the repository control rules and sustainability? The commons literature suggests that successful commons have complex, changing rules (Ostrom & Hess, 2007). Others have shown that repository sustainability is a concern, and that developing revenue sources is important given cutbacks by host institutions (Maron et al., 2009). Our data illustrate the many dimensions of repository control rules, but we only have anecdotal data about the relationship between rules and sustainability. Some repository managers believe that controls like required acknowledgement and user registration assist with sustainability because they provide impact data for funders. OTA's decision to not accept closed submissions without payment also points to concerns about costs and sustainability. We had several repositories with explicit revenue functions. For example CCDC had a licensed database product. Archives like ASSDA or ICPSR used membership models that required institutions to pay for access for their end users. These repositories may enjoy a sustainability advantage compared to repositories that depend solely on grants, volunteer labor or sole source institutional support (Maron et al., 2009). But we still have a poor understanding of how specific access and use rules influence sustainability.

In order to better understand the relationship between rules and sustainability, we are undertaking a complementary study of data controls and repository sustainability in one specific field -- social science data repositories. Using a small number of historically grounded case studies, the project will complement this paper by providing a context-rich understanding of how and why access and use rules developed within one sub-field and how those rules change over time with changes in technology, regulation of research and researcher expectations. This ongoing study should illuminate the complex relationships between access and use controls and sustainability. This study will also meet the National Science Board call for investigation of "how responsibility for community–proxy functions is acquired and implemented by data managers and how these activities are supported" (NSB, 2005).

Another question this study informs, but cannot answer, is whether controlling access and use promotes more sharing than might otherwise occur. We have only anecdotal evidence. Like the prior work of Pryor (2009) and Tenopir (2011) suggest, some of our repository managers *believed* that having control options would encourage deposit. For example one manager remarked, "…some investigators would not publish their data under [open] terms, so the [repository] allows them to restrict access. This gets them involved in the process of releasing data (even though restricted), and we believe that it breaks down concerns over data release, eventually leading to public release of data." As another participant explained, "[my data repository] has definitely seen data that people want to share but are unable to because of a lack of appropriate protections."

The next step is to find out what access and use control options change scholars' attitudes about sharing or change scholars' actual sharing behavior. This paper's analysis outlines an array of control options that repositories might employ. Future quasi-experimental research could test variation in researchers' attitudes toward sharing in relation to different control options. For example, within a specific sub-field, do researchers express a greater inclination to deposit data if the repository or the researcher controls the permission seeking process? Or, do researchers' attitudes about deposit change more if the repository vets potential data users? While this study cannot answer these questions, it prepares the ground for future research by describing the array of controls currently employed by repositories.

Another limitation of this study and area for future research involves looking at the actual utility of deposited data. Future work should examine the relationship between utility of data and depositor choices about what data to include, and how to structure deposited data. Past research by Hilgartner and Brandt-Rauf suggests that researchers sometimes decide to use specific formats or types of data in order to limit potential uses (1994). Given access to data and permission to use data, we should not assume that all deposited data is equally useful.

In summary, this study sheds light on a unheralded subset of web-based data repositories we call controlled data collections (CDC). From a practice standpoint, the paper provides a starting point for repositories when developing access and use control methodologies that fulfill the user proxy functions described in repository standards documents (CCSDS, 2011; NSB, 2005). For example, as U.S. federal agencies seek to comply with the call for greater public access to data (Holdren, 2013), they may determine that responsible stewardship of some data requires active management of access and use. This paper's analysis should help them decide on an appropriate menu of options to use. Repositories and other organizations managing data should also take note that our data shows that a one-size-fits-all solution does not exist within disciplines and even within individual repositories. Most repositories offer a broad array of access and use control options. Repositories are grappling with a wide variety of data types and subjects that require a custom array of policies and tools to manage access and use. As Ostrom and Hess note, managing information commons (like data repositories) require complex and ever changing rule sets (Hess and Ostrom, 2007).

Appendix

Classification codes (in parentheses after repository names)

Institution type: Government (*G*); University (*U*); Mixed (*M*).

Host nation: Australia (*AU*); Japan (*JP*); Netherlands (*NL*); United Kingdom (*UK*); United States (*US*); International (*I*).

(***contains human subject data):

Chemistry/molecular:

1. Cambridge Crystallographic Data Centre (CCDC) (*M*, *UK*) (http://www.ccdc.cam.ac.uk)
2. Proteomics Identifications Database (PRIDE) (*M*, *I*) (http://www.ebi.ac.uk/pride)
3. Proteome Commons Tranche Repository (*U*, *US*) (http://www.proteomecommons.org)

Human Biology

4. Anon-Bio (Anonymous by request of repository)***
5. Biologic Specimen and Data Repository Information Coordinating Center (BioLINCC) (*G*, *US*) (http://biolincc.nhlbi.nih.gov)***

Earth and space science (EAR) repositories:

6. Astronomical Data Archives Center SMOKA Science Archive (*G*, *JP*) (http://smoka.nao.ac.jp)
7. British Atmospheric Data Centre (BADC) (*G*, *UK*) (http://badc.nerc.ac.uk)
8. National Center for Atmospheric Research (NCAR) Computational and Information Systems Laboratory (CISL) Research Data Archive (*G*, *US*) (http://dss.ucar.edu)
9. National Snow and Ice Data Center (NSIDC) (*M*, *US*) (http://nsidc.org)
10. Natural Environment Research Council Earth Observation Data Centre (NEODC) (*G*, *UK*) (http://www.neodc.rl.ac.uk)

Ecological

11. SeaDataNet(*M, I*) (http://www.seadatanet.org)
12. VegBank(*U*, *US*) (http://www.vegbank.org)
13. National Park Service Natural Resource Information Portal (*G*, *US*) (http://nrinfo.nps.gov)
14. Knowledge Network for Biocomplexity (KNB) (*M*, *US*) (http://knb.ecoinformatics.org)

Social science (SOC) repositories (all contain human subject data):

15. Australian Social Science Data Archive (ASSDA) (*G*, *AU*) (http://www.assda.edu.au)

16. Data Archiving and Networked Services (DANS), Social Science collection, online archiving system EASY (*M*, *NL*) (http://easy.DANS.knaw.nl)
17. Institute for Quantitative Social Science (IQSS) Dataverse Network (*U*, *US*) (http://dvn.iq.harvard.edu)
18. Interuniversity Consortium for Political and Social Research (ICPSR) (*U*, *US*) (http://www.icpsr.umich.edu)
19. Odum Institute for Research in Social Science Data Archive (*U*, *US*) (http://www.irss.unc.edu/odum)
20. UK Data Archive Source-to-Output Repositories (UKDA-store) Social Sciences & Economics Collection (*M*, *UK*) (http://store.data-archive.ac.uk)

Humanities (HUM) repositories (**contains human subject data):

21. Data Archiving and Networked Services (DANS) the e-depot for Dutch Archaeology, (DANS EDNA) archiving system EASY (*M*, *NL*) (http://easy.DANS.knaw.nl)
22. Ethnographic Video for Instruction and Analysis Data Archive (EVIA)** (*U*, *US*) (http://www.eviada.org)
23. Oxford Text Archive (OTA) (*U*, *UK*) (http://ota.ahds.ac.uk)
24. The Digital Archaeological Record (tDAR) (*M*, *I*) (http://core.tdar.org)

WEBSITES REFERENCED

ASSDA "Why Deposit Data with ASSDA?" http://www.assda.edu.au/why_deposit.html Retrieved May 22, 2012.

Cambridge Crystallography Data Center (2011) "CCDC: Request a Structure" http://www.ccdc.cam.ac.uk/Community/Requestastructure/Pages/Requestastructure.aspx. Retrieved May 22, 2012.

CC Wiki (2011) "Case Study:Proteome Commons"

CISL RDA "RDA Data User Sign In/Registration" https://rda.ucar.edu/cgi-bin/login, Retrieved May 22, 2012.

DANS "DANS Conditions of Use on the reuse of deposited data" http://www.DANS.knaw.nl/en/content/DANS-conditions-use-reuse-deposited-data. Retrieved May 22, 2012

ICPSR " Restricted Data" http://www.icpsr.umich.edu/icpsrweb/ICPSR/access/restricted/index.jsp. Retrieved May 22, 2012

ICPSR “Restricted Data Use Agreement” http://www.icpsr.umich.edu/icpsrweb/ICPSR/access/restricted/agreement.jsp Retrieved May 22, 2012.

ICPSR “Deposit Data and Findings” http://www.icpsr.umich.edu/icpsrweb/ICPSR/access/deposit/index.jsp. Retrieved May 22, 2012

IQSS, Account Terms. http://thedata.org/book/account-terms, Retrieved May 22, 2012

National Heart Lung and Blood Institute (2009) “NHLBI Research Materials Distribution Agreement (RMDA) V01.1.d20090610. https://biolincc.nhlbi.nih.gov/static/RMDA.pdf

National Heart Lung and Blood Institute (2011) “Guidelines for NHLBI Data Set Preparation” http://www.nhlbi.nih.gov/funding/setpreparation.htm

National Park Service (2008) Data management guidelines for inventory and monitoring networks. Natural Resource Report NPS/NRPC/NRR—2008/035. National Park Service,Fort Collins, Colorado. Section 9.3.2

NERC “Data rules and policies” http://www.neodc.rl.ac.uk/popups/faqwindow.php?id=7, Retrieved May 22, 2012.

NERC “Access Rules” http://badc.nerc.ac.uk/data/rules.html, Retrieved May 22, 2012

NERC “User Registration and Data Application Process” http://www.neodc.rl.ac.uk/popups/faqwindow.php?id=5, Retrieved May 22, 2012

Odum Archive “Odum Dataverse Account Terms of Use” http://www.irss.unc.edu/odum/contentSubpage.jsp?nodeid=574, Retrieved June 9, 2012.

OTA “Depositing with the University of Oxford Text Archive” http://ota.ahds.ac.uk/about/deposit.xml. Retrieved May 22, 2012

Oxford Text Archive “Oxford Text Archive User Agreement” http://ota.ahds.ac.uk/documents/user_agreement.xml, Retrieved May 22, 2012

PRIDE User Manual http://www.ebi.ac.uk/pride/userManual.do, Retrieved May 22, 2012

SeaDataNet (2008) L081 Data Access Restriction Policies “Terms used to represent and classify data access policies in operation in the SeaDataNet project”

The Digital Archaeological Record “Register with the Digital Archaeological Record” http://core.tdar.org/account/new. Retrieved May 22, 2012

The Digital Archaeological Record "Terms of Use" http://www.tdar.org/support/policies/terms-of-use/. May 22, 2012

Register of Professional Archeologists "Registration Application and Renewals" http://www.rpanet.org/displaycommon.cfm?an=4. Retrieved May 22, 2012

UK Data Archive "License Agreement" http://www.esds.ac.uk/aandp/create/licence.asp). Retrieved May 22, 2012.

UK Data Archive "Economic and Social Data Service: End User License" http://www.esds.ac.uk/aandp/access/licence.asp, Retrieved June 1, 2012

VegBank Privacy Policy, http://vegbank.org/vegbank/general/privacy.html Retrieved May 22, 2012

REFERENCES

Alsheikh-Ali, A. A., Qureshi, W., Al-Mallah, M. H., & Ioannidis, J. P. A. (2011). Public Availability of Published Research Data in High-Impact Journals. PLoS ONE, 6(9), 4.

Blumenthal, D., Campbell, E.G., Gokhale, M., Yucel, R., Clarridge, B., Hilgartner, S., & Holtzman, N.A. (2006). Data withholding in genetics and the other life sciences: Prevalences and predictors. Academic Medicine*,* 81(2), 137-145.

Borgman, C. L. (2012). The conundrum of sharing research data. Journal of the American Society for Information Science and Technology, 63(6), 1059–1078.

Borgman, C.L. (2009). The digital future is now: A call to action for the humanities. *Digital Humanities Quarterly, 3*(4), http://digitalhumanities.org/dhq/vol/3/4/000077/000077.html

Borgman, C.L. (2007). *Scholarship in the digital age: Information, infrastructure, and the Internet.* Cambridge, MA: MIT Press.

Consultative Committee for Space Data Systems CCSDS (2011). *Audit and Certification of Trustworthy Digital Repositories*, CCSDS 652.0-M-1, Washington DC: NASA.

Cragin, M.H. & Shankar, K. (2006). Scientific data collections and distributed collective practice. Computer Supported Cooperative Work*, 15*(2-3), 185-204.

Cragin, M. H., Palmer, C. L., Carlson, J. R., & Witt, M. (2010). Data sharing, small science and institutional repositories. Philosophical Transactions of the Royal Society - Series A: Mathematical, Physical and Engineering Sciences, 368(1926), 4023-4038.

Creative Commons. (2013) “About the Licenses” http://creativecommons.org/licenses/ Retrieved April 1, 2013.

Eschenfelder, K.R.; Caswell, M. (2010) Digital Cultural Collections in an Age of Reuse and Remixes. *First Monday*, 15(11). Retrieved from: http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/viewArticle/3060/2640

Fry, J., Schroeder, R., & den Besten, M. (2009). Open science in e-science: Contingency or policy? Journal of Documentation*, 65*(1), 6-32.

Hess, C.; Ostrom, E. (2007) Understanding Knowledge as a Commons: From Theory to Practice. Cambridge, MA: Oxford University Press.

Hilgartner, S. (1997). Access to data and IP: Scientific exchange in genome research. In IP rights and research tools in molecular biology: Summary of a workshop held at the National Academy of Sciences, February 15-16, 1996 (pp. 28-39). Washington, D.C.: National Academy Press.

Hilgartner, S.; Brandt-Rauf, S.I. (1994) Data Access, Ownership and Control: Toward Empirical Studies of Access Practices. Knowledge: Creation, Diffusion, Utilization *15*(4), 355-372.

Holdren, John P. (2013, Feb 22) “Increasing Access to the Results of Federally Funded Scientific Research” Office of Science and Technology Policy: Executive Office of the President: Washington DC.

Kuipers, T.; van der Hoeven, J. (2009) PARSE.Insight: INSIGHT into Issues of Permanent Access to the Records of Science in Europe.

Langille, M. G. I., & Eisen, J. A. (2010). BioTorrents: A File Sharing Service for Scientific Data. (J. E. Stajich, Ed.)PLoS ONE, 5(4), 5. Retrieved from http://dx.plos.org/10.1371/journal.pone.0010071

LeFurgy, W.G. (2009). NDIIPP Partner Perspectives on Economic Sustainability. *Library Trends*, 57(3), 413-426.

Lincoln, Y.S.; Guba, E.G. (1985) *Naturalistic Inquiry.* Newbury Park: Sage.

Marcial, L.H. & Hemminger, B.M. (2010): Scientific data repositories on the Web: An initial survey. Journal of the American Society for Information Science and Technology*,* 61(10), 2029-2048.

Maron, N.L.; Smith, K.; Loy, L (2009) *Sustaining Digital Resources: An On-the-Ground View of Projects Today*, New York: Ithaka S+R.

Milia, N., Congiu, A., AnagnosToU, P., Montinaro, F., Capocasa, M., Sanna, E., & Bisol, G. D. (2012). Mine, Yours, Ours? Sharing Data on Human Genetic Variation. PLoS ONE, 7(6)

National Science Board (2005). *Long-lived digital data collections: Enabling research and education in the 21st century*. Washington DC: National Science Foundation.

Open Knowledge Foundation (2012). *Open Data Handbook.* Cambridge: OKF.

Ostrom, E.; Hess, C. "A Framework for Analyzing the Knowledge Commons" in Eds. Hess, C.; Ostrom, E. (2007) Understanding Knowledge as a Commons: From Theory to Practice. Cambridge, MA: Oxford University Press, pp 41-82.

Ostrom, E. (1990). Governing the Commons: The Evolution of Institutions for Collective Action. Cambridge: Cambridge University Press.

Kuipers, T.; van der Hoeven, J. (2009) PARSE.Insight: INSIGHT into Issues of Permanent Access to the Records of Science in Europe.

Piwowar, H. A. (2011). Who shares? Who doesn't? Factors associated with openly archiving raw research data. PloS one, 6(7), e18657. doi:10.1371/journal.pone.0018657

Piwowar, H. A., Day, R. S., & Fridsma, D. B. (2007). Sharing detailed research data is associated with increased citation rate. PLoS ONE*, 2*(3): e308. doi:10.1371/journal.pone.0000308

Pryor, G. (2009). Multi-scale data sharing in the life sciences: Some lessons for policy makers. *The* International Journal of Digital Curation*,* 4(3), 71-82.

Savage, C. J., & Vickers, A. J. (2009). Empirical study of data sharing by authors publishing in PLoS journals. PLoS ONE, 4(9), e7078.

Taylor, P. L. (2007). Research sharing, ethics and public benefit. Nature Biotechnology*, 25*(4), 398-401.

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., Manoff, M., et al. (2011). Data Sharing by Scientists: Practices and Perceptions. PLoS ONE, 6(6), 21. Retrieved from http://dx.plos.org/10.1371/journal.pone.0021101

Van House, N. A., Butler, M. H., Schm, L. R., W, S., & Berkeley, U. C. (1998). Cooperative KnowledgeWork and Practices of Trust : Sharing Environmental Planning Data Sets. CSCW '98: The ACM Conference On Computer Supported Cooperative Work. Proceedings. (pp. 335-343). Seattle, WA: ACM.

Wicherts, J. M., Bakker, M., & Molenaar, D. (2011). Willingness to Share Research Data Is Related to the Strength of the Evidence and the Quality of Reporting of Statistical Results. PLoS ONE, 6(11), e26828. Retrieved from http://dx.plos.org/10.1371/journal.pone.0026828

[i] The University of Wisconsin-Madison Social Sciences Human Subjects Institutional Review Board approved this study. Data were collected under IRB protocols SE-2009-0303 and SE-2012-0573. These protocols included a written informed consent agreement for all survey and interview participants that assured privacy and confidentiality of responses. Survey findings are only reported in the aggregate and textual responses do not include personal or organizational identifiers. The only data containing organizational identifiers are those drawn from public repository websites. The University of Wisconsin-Madison Institutional Review Board does not consider information drawn from public websites to be human subjects data.

[ii] We were not able to develop a random sample of CDC because it was not possible to develop a population list of sufficient size to draw a random sample. It took considerable effort to find our 24 CDC from the existing lists of repositories. Given this, a purposeful sample was appropriate.

[iii] We used a codebook developed from exploratory data analysis, pretests and a literature review. We pretested the codebook on a subsample of repository websites to ensure that the structured analysis captured the data of interest. We then conducted a formal structured content analysis of each of these controlled collections using the codebook and a data entry form.

[iv] We invited repository managers to participate in the survey via an email that included the draft report related to their repository and a hyperlink to the survey form. We sent out three rounds of email reminders during spring 2011 and a final two-day air letter of invitation to non-responders. In the final paper invitation, we included a paper means of providing responses as well as a hyperlink to the web-based form.

[v] How reliable is the unverified survey data? Returned surveys show that respondents only corrected approximately 13% of the data developed from the website content analysis, suggesting that our analysis was a reasonably reliable means of representing the repositories. Because we did not receive any feedback from 6 repositories, we should expect the same level of error in their data.

[vi] The educational use only terms of use stemmed from the fact that some of the videos included copyrighted or trademarked material such as songs or images of corporate logos. The repository managers perceived that the educational use only restriction provided a Fair Use justification for their repositories' activities.

[vii] In research involving human subjects, it is common to employ a data code to protect research participants' identities. Investigators assign all participants a non-identifying alphanumeric code that is connected to identifiers though a separate key. Ideally, a reader of study materials could not identify individual participants without key. Shielding the key from legal requests shields the identity of participants.

[viii] CCDC Data Deposition and Request FAQ stated that users could request for free of charge "data associated with any one paper, which can be supplied for bona fide research purposes." (May 26, 2012)