This paper describes and compares access and use control practices at twenty-two data archives in six areas: Earth and space sciences, chemistry/proteins, ecology, humanities, social studies and human health. It also compares common practices to Open Knowledge Foundation recommendations for open data and open bibliographic data. Finally, it investigates similarities and differences in motivations for controlling access and use across repository types.

Introduction

This paper examines access and use control practices at data repositories within the context of larger tensions about openness and control over cultural institution (CI) digitized content and the metadata describing that content. Linked open data initiatives that encourage network based aggregation and re-use of materials raise new questions about who should have access to collections content and what they ought to be able to do with it.

In this paper we describe twenty-two cases of current repository access and use control practices, explore what exactly “open” means in relation to those practices, and consider reasons that repositories restrict access or use. We compare those practices to recommendations made by the Open Knowledge Foundation for open data. We attempt to address questions of openness related to both content and the metadata describing that content.

Different understandings of openness for content and metadata exist. One prominent example is the Open Knowledge Foundation (OKF) definition as “free to use, reuse, and redistribute it.” (OKF, Open Definition 2012) Advocates sometimes disagree about what types of limitations may be permissible within open data. But importantly, restricting commercial use or redistribution is typically seen as “not open.” Moreover discriminating against certain users or fields of endeavor in providing access is not open. Also importantly, the OKF definition accepts requirements for attribution (OKF, Open Definition 2012)

OKF advises those who wish to host open data to employ open data licenses that make the data’s open status explicit such as CCZero. (OKF Conformant Licenses)

In an earlier part of this study, a survey of repository managers found the most common motivations for controlling access and use among these repositories. We found the most common motivations included:

- Ensuring attribution.
- Avoiding misuse/non research use of data
- Protecting sensitive information (non personal privacy)
- Prohibiting commercial use
- Concerns about intellectual property
- Concerns for privacy and privacy (Eschenfelder & Johnson 2012)

This paper examines these motivations more closely in light of the types of data contained in each repository.

Methodology

The case studies represent “controlled collections” or repositories that restrict access to data or use of data in some way. We sought to compare controlled data collections practices from across a variety of disciplines, but no list of controlled collection repositories exists. We developed a list from previous studies, expert recommendations and a review of journal and funding body policies. Repositories were limited to those with substantial English language content. We excluded repositories that did not accept data submissions from a broad audience (i.e., repositories that collect data only from a single project, instrument, or institution). We also excluded repositories that were completely open and those where data access was restricted to only the original depositor.
We were left with a list of twenty-four repositories that fit our criteria. Of those, two ceased to function over the course of the study.

Repository descriptions are primarily drawn from content analysis of data repository website documentation and instructions conducted in 2011 and 2012. To determine access and use controls for content, we examined all documentation related to searching for data, downloading data and depositing data. To determine access and use policies related to bibliographic data, we searched for each repository name with the terms “SPARQL” “linked data” or “RDF” to ascertain if the repository advertised any open bibliographic elements. We also searched repository names on the OKF Data Hub which aims to list open collections.

**Case Studies: Controlled Collections Data Repositories**

We group our cases into six categories based roughly on the scholarly discipline associated with the data in the repository. One interdisciplinary repository with sub repositories is mentioned twice, once in humanities and one in social science. A complete list of the repositories with full names and location information is included in the Appendix. Within these six categories we describe the repositories in order from fewer restrictions to more restrictions.

**Earth and Space Sciences:** We examined five repositories containing geological, atmospheric, and astrophysical data and associated computational models and documentation. The repositories took deposits primarily from research projects or funding programs with which they had relationships; unlike repositories in other areas, deposit was not marketed heavily. A few however, would consider non-affiliated deposits.

In all the cases, data could be temporarily embargoed. The most open repository, National Snow and Ice Data Center, did not require registration to view metadata or to download the vast majority of its data. Two other repositories, the National Center for Atmospheric Research/Research Data Archives and SMOKA Science Archive, allowed the public to browse metadata, but required user registration prior to downloading. A third archive, the British Atmospheric Data Center, contained some data that was open to the public, but other data required registration prior to access, and some data sets required formal application. BADC listed different terms of use for different (non public) data sets. A related UK repository, NERC Earth Observation Data Center (NEODC), could be searched but also required registration prior to accessing any data sets. Moreover, most NERC-EODC data sets required individual application for use. Some had standard license agreements that the user could agree to via a web interface prior to access. One data set required a signed license agreement delivered via postal mail. In almost all cases (except NSIDC), repository terms of use of the repository or its datasets forbade further downstream sharing.

Access and use control motivations unique to this group included the fact that some of the repositories held data from satellite or other remote sensing equipment, and export of this data was regulated to exclude certain nations. For example in the US, the Department of State’s Trade Controls “International Traffic in Arms” Regulations includes satellites and their associated systems on a list of munitions. This created the need for some US repositories (e.g., NCAR-RDA) to restrict access to some data based on the location of the data requester’s home institution. In another example, NEODC restrict some data sets to certain subsets of UK researchers that worked within a particular grant agencies funding stream or on a past or present project involving a particular satellite instrument.

**Chemistry/Proteins:** We examined three data centers containing chemical or protein data. The Cambridge Crystallography Data Center (CCDC) contained small molecule crystal structures, while the Proteome Commons (Commons) and PRIDE repositories contained data on protein structures.

In all three cases, users could search data without registering. PRIDE and the Commons held public data that did not require registration to download in addition to private data. Registration was required by CCDC to apply for and download any data. All three required registration to submit data.

Submission of data to CCDC and PRIDE were closely tied to publication in journals but the Commons was not. In both these cases, submitted data was embargoed until publication. At the CCDC, post publication data structures were moved to a related commercial licensed database. At PRIDE, data was moved to the archive and available to the general public to search. In general, deposit in CCDC only made one’s data available either by request (or subscription to the licensed product). Researchers with “bona fida” research requests could request free access to a limited number of structures by filling out a request form and agreeing to terms and conditions that precluded further dissemination of the data.

Proteome Commons and PRIDE allowed depositors to make private data accessible to preassigned user accounts (private sharing). The CCDC did not offer this option.

The Commons was the only archive to employ the CCZero license (waiving all rights) as a default deposit licensing option.

Access and use control motivations unique to this group included CCDC and PRIDE’s relationship to the journal publication process and the need to keep data exclusive prior to publication. Also, CCDC needed to manage free
access to its data in order to not undermine its commercial database product.

Ecology: We examined four ecological repositories containing data about plants, animals and ecosystems on land or in water: VegBank, the US National Park Service IRMA system, the Knowledge Network for Biocomplexity (KNB), and SeaDataNet. NPS IRMA also contained cultural history information. In all four cases, users did not need to register to browse data sets. Further, the first three contained many data sets that did not require registration to use.

Among the most liberal, VegBank did not require registration to use any data, but required registration to deposit. While much data in VegBank was public, VegBank allowed depositors to request a “confidentiality status” of full embargo for their data. To do so, depositors had to justify their embargo in their metadata. Additionally, VegBank provides tools to scramble longitude and latitude data while leaving other data accessible.

Also quite liberal, the NPS IRMA system also allowed public access to most datasets. Uploading of data to IRMA was restricted to NPS employees or contractors. Most data sets were unrestricted, and as US federal government data, had liberal use rights. But the IRMA system did offer an embargo option. Park superintendents determined what data should be protected and for how long.

The third ecology repository, the Knowledge Network for Biocomplexity, allowed the public may browse all data sets and use open data sets without registration. Getting access to restricted data required that users register and then request access. KNB also allowed depositors to limit access to predefined users with rights set at a table or data set level. Depositors could also set embargos.

SeaDataNet had the most extensive restrictions. Metadata could be searched without registration; however, registration was required for downloading any data or submitting data. Registration required users to agree to a license agreement with terms of use. Registration requests were reviewed and approved by regional data center staff.

According to the website, use requests for restricted data were evaluated based on combinations of “user type qualification” and access metadata for the data. The user type qualification was set by the regional data center as part of the user registration process. Access metadata for each data set is set by the data contributor and employed a controlled vocabulary of limitations. Based on the combination of user type qualification and metadata, some requests were granted automatically, while other requests were reviewed by data centers.

Access and use control motivations unique to this group included protecting location data about endangered plants, animals and (in one case) at risk cultural resources information (VegBank, NPS IRMA, KNB), protecting privacy rights of private land owners whose land contained described resources, shielding data that had not been officially inspected for quality (NPS IRMA), and national legal requirements to protect at risk materials or locations. (NPS IRMA)

Humanities: We examined four humanities repositories. Two contained archeological materials (tDAR and DANs), one contained video ethnomusicology files (EVIA), and one contained text and audio data for linguistics (OTA). DANs also included oral histories.

On the most liberal end, anyone could search the Oxford Text Archive (OTA), and many data sets do not require registration to use. Other data sets were restricted and required application for use.

Depositors to OTA were encouraged to de-identify all datasets that contain speakers or writers. Depositors were given four access choices when depositing: open, restricted to educational community, requiring permission of the depositor, closed archive. Interestingly, OTA reserved the right to charge a fee for deposits requiring permission and closed options; further, the website stated it would no longer take restricted data sets except for under exceptional circumstances.

Two repositories contain archeological data, the Digital Archeological Record (tDAR) and a subrepository of DANs called EDNA. While both sites were open to the public to browse, both sites required users to register to download data.

tDAR’s registration process included display of a terms of use disallowing any use that could “damage to the archaeological record.” tDAR’s registration form also asks users to identify if they are professional archeologists by entering their “Registered Professional Archeologist” number.

EDNA similarly sought to identify professional archeologists although it did not ask archeologists to identify themselves during the registration process. For depositors, EDNA offered four restriction options: open, restricted, embargoed and an option of only making data accessible to archaeologists registered within the DANs EDNA system. It is not clear how archeologists become registered with the system.

In comparison, DANs History data archive only offered three deposit options (open, restricted and embargoed) On the other hand, the history deposit instructions required that depositors uploading interview data submit a document describing the terms of access agreed upon between the interviewee and interviewer.

At the Ethnographic Video for Instruction and Analysis (EVIA) anyone may search the metadata, but like EDNA, only registered users could view most materials. Further, users were typically restricted to affiliated educational
institutions although others may request access. The archive was restricted to educational use only.

Users cannot download data from EVIA. All video was shared through streaming and the EULA forbade capture and reuse of archive materials.

Due to the costs of archive services, depositors apply through a competitive process to have their work taken in by the repository. The application process does not privilege openly accessible work.

Access and use control motivations unique to this group included the need to prevent misuse of politically, culturally, or personally sensitive materials (EVIA). Further, like the ecology archives, the archeology archives needed to protect endangered cultural resources/locations (tDAR, DANs). Like the social science archives, history archives needed to ensure that identifiable interview data was only used in conjunction with original terms of informed consent and the linguistics archives needed to ensure that data was either de-identified or that identities of participants were protected.

Social Sciences: We examined six social science data repositories: Odum, IQSS/Murray, DANs, ASSDA and ICPSR. The social science repositories were the most uniform of the archives. All allowed searching by the public but most required registration prior to use of any data. All hosted data that were “open” in that they didn’t require further permissions beyond registration. All also hosted data licensed to institutional affiliates as well as restricted data sets requiring more extensive permissions. They differed in the degree to which they were willing to accept risky data and the protection measures in place for their most high risk data categories.

Both The Odum Institute and IQSS/Murray Archive (discussed next) employed the open source Dataverse software created at IQSS to manage downloading and deposit. The software dictated many access and use control features common to both repositories.

The Odum Institute hosted some open data sets that did not require registration to access. Downloading any data set (even open) or registering for an account, triggers a terms of use agreement. Odum required registration to deposit data and to access restricted data sets. The Dataverse software ties access permissions to particular user accounts.

Odum depositors could restrict access at the file and study level. Users can also check a box to trigger an email alert for every request for access. Depositors can set custom use terms. Users can also set metadata to be public or private.

Odum is one of a subset of data archives whose policies preclude users from depositing data that contains “high risk information” (e.g. social security numbers of health records), and requires that data not constitute an “unwarranted” invasion of privacy.

The Murray Data Archives at the Institute for Quantitative Social Sciences at Harvard University also employed the Dataverse software and correspondingly had many similarities to Odum in terms of levels of data restrictions, typing access to user accounts, custom terms of use for their data and email notification of access requests.

IQSS required users wanting to use publicly accessible data to agree to terms of use. Similarly, to deposit or use restricted data IQSS required users to register and agree to further conditions that included notifying the archive of any issues related to their data, not depositing high risk data or data that constitute a breach of privacy.

IQSS/Murray required that depositors of data with privacy or confidentiality concerns employ the Dataverse restricted access tools. IQSS further included a liability statement warning depositors that depositors could be held liable for damages resulting from deposits that violate the terms of use. Murray also held restricted access data outside of Dataverse and required a formal application and agreement to use terms and conditions for access and use.

The remaining data repositories all required registration for access and use of any data set. For example, the Data Archiving and Network Services (DANS) offered three levels of access: open access, restricted access and embargoed. But users had to register to download any data, (even open data) or to deposit data. Registration involved a click through terms of use; further, during registration, DANs asked users to have their use tracked and displayed in the DANs system. If depositors chose restricted access, then users had to request access and depositors could choose to grant or deny permission.

Like other social science data archives, DANs asks its users to de-identify data. But unlike Odum, IQSS (and the remaining data archives) would accept data with privacy issues through restricted access categories.

The third social science repository, ICPSR also required users to create an account to download any data. Open data was accessible to all registered users. ICPSR had two categories of restricted access data. To use “restricted access” data, ICPSR required users to submit an application, sign a “Restricted Data Use Agreement” and develop a data protection plan. To use “secure enclave” data, ICPSR also required users to access the data at ICPSR offices, and develop a confidentiality agreement between ICPSR, the PI and the PI’s institutional representative.

ICPSR encouraged depositors to remove or mask data that can identify individuals. But, if this negatively affected the value of the data, ICPSR would take the original data and restrict access to it. For each deposit, ICPSR conducted a risk review of each data set.
The fourth repository, ASSDA distinguished between 3 types of use: browse, analyze and download. Anyone could browse metadata. Registration was required to run analysis online through ASSDA interface tools, and registration was required to download data sets. Further, downloading was limited to researchers; students were required to have a researcher sponsor to download data.

For depositors, ASSDA offered three categories of access; each requires application. In the unrestricted category, an application form was required but the depositor was not informed of the requested use. In the restricted category, application was required and the depositor would be informed of requests and may give or withhold permissions. In the special category, application was required and the depositor had included additional special access conditions.

Like the other archives, ASSDA encouraged depositors not to deposit risky data, and to mask data that identified individuals. Like ICPSR, ASSDA stated that staff will review data for risks.

Human health: We examined two repositories containing health data from human subjects. These two repositories have no open data. Both only took data deposits from funded research projects associated with their institutions.

Both repositories featured rigorous access controls and numerous use restrictions due to the sensitive nature of personally identifiable health data. Both repositories needed to account for the myriad ethical and legal issues associated with privacy, adherence with the terms of consent, laws like HIPAA and requirements from Institutional Review Boards.

The first repository, BioLINCC, hosted datasets from studies in both the "open" and "proprietary" periods. BioLINCC controlled the application process for data from "open studies," whereas the parent study approved access to data in the proprietary period. "Proprietary" refers to an embargo period of time before the data become sharable via BioLINCCs rules and policies.

BioLINCC had a substantial application process in which applicants needed to submit a research plan, curriculum vitae, research materials distribution agreement, and proof of IRB approval. The application process limited access to vetted, institutionally-affiliated researcher.

The other repository in this category requested to remain anonymous and will be referred to hereafter as ANON-BIO. ANON-BIO housed genetic data from its funded research. Like BioLINCC, ANON-BIO required a substantial application process. Only researchers with accounts managed by the funding agency that operates ANON-BIO are able to apply for restricted data.

Upon acceptance of a request for access to data, researchers must agree to a "Code of Conduct" and other terms of use. The requirements outlined in the “Code of Conduct” were among the most detailed of any repository in this study (a close second would be ICPSR’s enclave data requirements). ANON-BIO provided detailed guidelines for securely storing and working with restricted data including a requirement to destroy data after it has been used for the purpose described in the application.

The human health data group had similar access and use control motivations to the social science data archives and those humanities archives that contained human subjects data (e.g., oral histories, linguistic audio files). These motivations included protecting the privacy of research participants (either through direct identification or combining datasets to bootstrap identities), ensuring data reuse complied with conditions established as part of the original study informed consent process, and forbidding contact with research participants.

Open Metadata: We found evidence open bibliographic data activity at NSIDC, NCAR RDA, PRIDE, VegBank, SeaDataNet, DANs. This data however is extremely unreliable given the difficult we experienced ascertaining whether repositories provided open bibliographic data. WebHub is not a complete resource. Repositories do not advertise their open bibliographic data activity on their websites. We most evidence through Web searching.

The evidence of activity we found typically suggested a narrow scope that did not necessarily apply to all metadata in the repository. For example, we found evidence that one project within NSIDC made an ontology open for others to share. It is unclear if or when these limited experiments will spread to the other projects or data sets included in the repository. In sum, open bibliographic efforts appeared as nascent, experimental, not well advertised and not clearly licensed for open use.

Analysis and Conclusion

What does open mean? Few of the reviewed repositories fit these stringent OKF requirements for open. First, the OKF requires that there not be “technical barriers” to reuse. It is unclear if the common registration requirement counts as a barrier. We found only three repositories that offered anonymous public access to most datasets: NSIDC, VegBank and NPS-IRMA. Even in these cases however, the data was not clearly licensed as open and some data within each repository was embargoed because of sensitivity (VegBank, IRMA) or licensing limitations (NSIDC). One partial exception to the lack of clarity in licensing was Proteome Commons which encouraged, but did not require, depositors to use of the CCZero license.

Repository documents and responses to our survey provide ample evidence that most of our repositories saw their organizations as contributing to the broad goals of
data sharing. Given this, the data suggest several different definitions of “open” seen in these practicing repositories.

First, for many repositories, “open” meant making data available for free for scholars and educators to use for non-commercial purposes without further redistribution. Open repositories might employ a registration process in order to generate usage data to support the repository and improve services for users.

Second, our participants were also pragmatic about offering options in order to get researchers to deposit data so as to increase its sharing potential. From this perspective, being open meant offering a range of access options including, but not limited to anonymous public use in order to encourage sharing. Access options were a means to an end to promote sharing.

But in other cases, repositories saw controlling access as a responsibility as data custodians. For example the mission of the human biology repositories required them to ensure that data were only used in concordance with original research protocols and consent terms. For these repositories, ensuring responsible use by qualified researchers was a requirement to support sharing. So finally, in some instances, advertising the availability of data to qualified users and making the data more conveniently accessible was being open.

On the other hand, in some communities closed data sets may receive little use and thus not be cost effective in an expenditure to impact calculation. Repositories like OTA stated a desire to not accept any more closed data sets without extra custodial fees. Other repositories required that depositors justify their request for closed access (VegBank). Some field repositories may discourage closed data sets due to a need to prioritize resources to data sets that will show impact for funders.

These alternative views of openness complicate our understanding of what “open” means in open data and what options repositories might offer when seeking to encourage deposit and sharing of data among scholars.

Reasons for controlling access and use

In our previous work we compared motivations for controlling access and use across biological, physical, social sciences and humanities repositories (Eschenfelder and Johnson, 2012). Our failure to find strong patterns is explained by the diversity of practices found within each group by this paper’s analysis. This paper also divides the repositories into six groups that better reflect commonalities among their data types. Our descriptions of repositories practices within each of the six groups illustrate the range and variation of within each group as well as similarities across groups. We found the Humanities group requires further subdividing as archeology repositories tend to look similar to ecology repositories (both concerned about protecting locations) and history and linguistic repositories look more like some social science repositories. We found the human biology repositories had similar control arrangements as social science repositories that contained “risky” data.

Our micro analysis also illuminated a number of motivations that were not adequately represented in the survey categories including:

- Protecting the privacy of private landowners whose property contains protected plants animals or cultural resources.
- Protecting the location of things such as archeological sites or rare plants or animals.
- National regulation: Trade restrictions may limit the distribution of data stemming from satellites. National level laws may require the protect species/cultural resource location information.
- Limiting use to researchers approved by funder or with certain security status.
- Ensuring compliance with terms of data use specified when consent was given.
- Shielding data of uncertain quality
- Keeping data exclusive prior to publication
- Protecting the commercial value of data
- Compliance with contract terms for licensed data sets

ACKNOWLEDGMENTS

This research was funded in part Institute of Museum and Library Services Laura Bush 21st Century Research Grant RE-04-06-0029-06, and by funding from the Wisconsin Alumni Research Foundation.

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).

Earth and Space Sciences:

Astronomical Data Archives Center SMOKA Science Archive (G,JP) (http://smoka.nao.ac.jp)

British Atmospheric Data Centre (BADC) (G,UK) (http://badc.nerc.ac.uk)

National Center for Atmospheric Research (NCAR) Computational and Information Systems Laboratory (CISL) Research Data Archive (G,US) (http://dss.ucar.edu)

National Snow and Ice Data Center (NSIDC) (M,US) (http://nsidc.org)

Natural Environment Research Council Earth Observation Data Centre (NEODC) (G,UK) (http://www.neodc.rl.ac.uk)

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

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

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

Social Sciences:
Australian Social Science Data Archive (ASSDA) (G, AU) (http://www.assda.edu.au)
Institute for Quantitative Social Science (IQSS) Dataverse Network (U, US) (http://dvn.iq.harvard.edu)
Interuniversity Consortium for Political and Social Research (ICPSR) (U, US) (http://www.icpsr.umich.edu)
Odum Institute for Research in Social Science Data Archive (U, US) (http://www.irss.unc.edu/odum)
Data Archiving and Networked Services Electronic Archiving System (DANS EASY) (M, NL) (http://easy.dans.knaw.nl)

Humanities:
Data Archiving and Networked Services Electronic Archiving System (DANS EASY) (M, NL) (http://easy.dans.knaw.nl)
Ethnographic Video for Instruction and Analysis Data Archive (EVIADA) (U, US) (http://www.eviada.org)
Oxford Text Archive (OTA) (U, UK) (http://ota.ahds.ac.uk)
The Digital Archaeological Record (tDAR) (M, I) (http://core.tdar.org)

REFERENCES