Managing Access to and Use of Data Collections: A Preliminary Report

Andrew Johnson & Kristin Eschenfelder
University of Wisconsin-Madison School of Library and Information Studies
Contact: eschenfelder@wisc.edu

Purpose and Methodology

The purpose of this study was to explore to what degree data repositories employ access controls or use controls to regulate who accesses data in the repository and what uses are made of repository data.

The study included repositories from a variety of disciplines that met the following criteria:

1. Accepts data submissions from individual researchers/projects across institutions (i.e., not a repository for a single project, instrument, or institution).

2. Limits public access to some but not all data.

Due to the vast and ever-changing landscape of data repositories on the web, it is impossible to capture a complete list of every repository from every discipline. In order to have a large sample from which to draw the case studies, a list of data repositories was compiled from the recommendations of experts, funding bodies, and journals.

These sources produced a sample of 177 repositories. From this initial sample, 119 repositories met the first criterion that researchers must be able to submit data (67.2%). Eighteen of the repositories meeting the first criterion also met the additional requirement that some but not all data must have public access restrictions (15.1%).

Of the repositories that met both target criteria, four were from the biological sciences (22.2%), seven were from the earth and climate sciences (38.9%), two were from the humanities (11.1%), two were interdisciplinary (11.1%), and three were from the social sciences (16.7%).

The websites of the 18 repositories meeting the two inclusion criteria were analyzed to determine a variety of characteristics of each repository's motivations and methods for controlling access and use. Ratings for each site were based on a thorough exploration of the homepage and all relevant sections to which it linked with particular attention paid to sections concerning data access, policies, and metadata for public and restricted datasets.

Overall Results

Figure 1 shows the reasons for controlling access stated on the websites of the 18 repositories in the sample. Overall, the reasons for controlling access to restricted data are widely distributed. Six repositories had reasons for controlling access that did not fall into any of the predefined categories (33.3%). Concerns over protecting intellectual property, preventing data misuse, protecting sensitive information, and ensuring exclusivity were the next most frequent categories with three repositories falling into each (16.7%). Privacy was a reason for controlling access for two repositories (11.1%) while limiting data use to certain types of research only and ensuring attribution were motivations for one repository each (5.6%). None of the repositories in the analysis restricted access to data based on a desire to prevent commercial research, and three repositories did not state the reasons why access to data was restricted on their websites (16.7%).

Figure 1. Reasons for controlling access to data

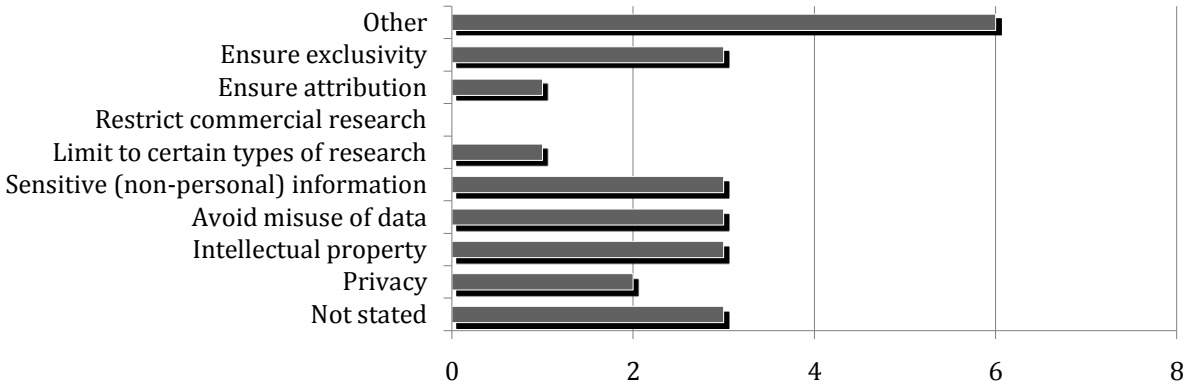


Figure 2 shows the reasons for controlling use of data as stated on the websites of the repositories in the sample. Overall, ensuring attribution was by far the most common reason for controlling use of data with 13 repositories stating it in their documentation (72.2%). Misuse of data was the next most common reason with six repositories mentioning it (33.3%) while protecting privacy and preventing commercial research were each mentioned by five repositories (27.8%). Four repositories were concerned with protecting intellectual property (22.2%). One repository each mentioned ensuring exclusivity, limiting use to certain types of research, and protecting sensitive information (5.6%). Two repositories had reasons for controlling use of data that did not fit into one of the predefined categories (11.1%), and four repositories did not state reasons for controlling use of data (22.2%).

Figure 2. Reasons for controlling use of data

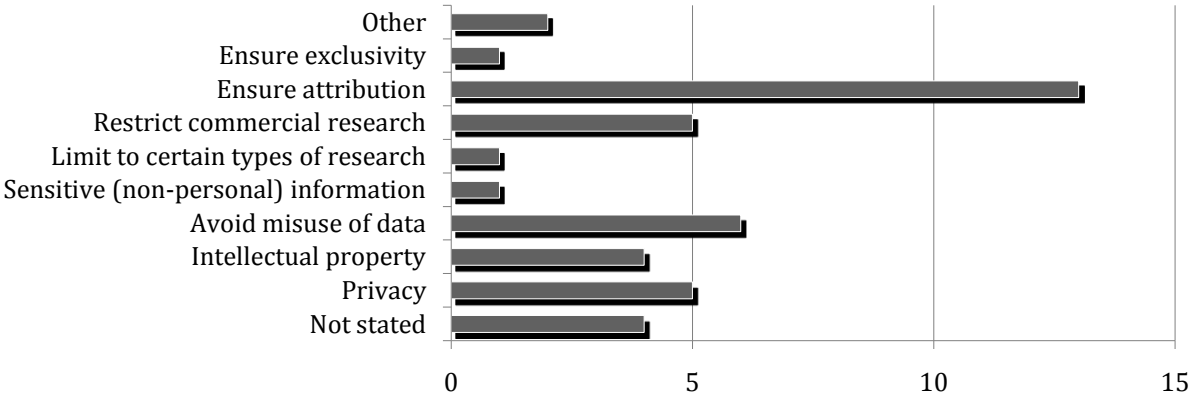


Figure 3, next page, reports the number of repositories using registration to control access to data. Only two repositories in the sample did not use some form of registration to control access to restricted data (11.1%). Five repositories required registration but did not appear to review registration requests (27.8%), nine required registration with an apparent review process for each request (50%), and five required a formal application process for obtaining restricted data (27.8%).

Figure 3. Use of registration

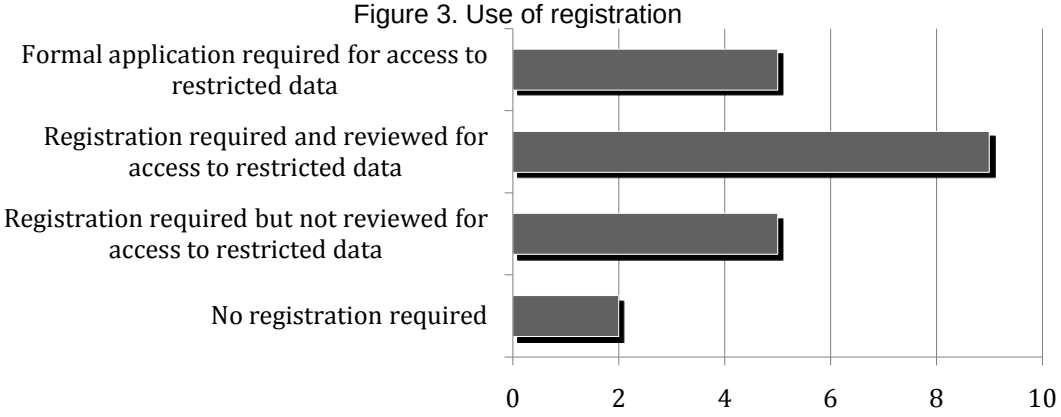


Figure 4 shows the use of IP range restrictions to control access to data. Fifteen repositories did not use IP range restrictions (83.3%) while only three used this method of controlling access to restricted data (16.7%).

Figure 4. Use of IP range restrictions

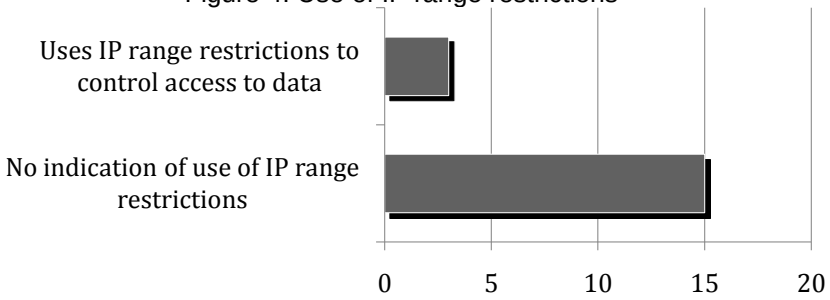


Figure 5 shows the policy methods used by repositories to control use of data. Overall, the most common policy method used was a repository-level terms of use statement with 14 repositories having this type of statement (77.8%). Ten repositories used a dataset-level terms of use statement (55.6%), and seven required users to click through a terms of use statement (38.9%). Five repositories had a repository-level copyright statement pertaining to data (27.8%) while four had dataset-level copyright statements (22.2%). It was not possible to determine whether or not dataset-level statements existed for three of the repositories (16.7%), and only one of the repositories did not have any policy documentation on its website.

Figure 5. Policy methods used

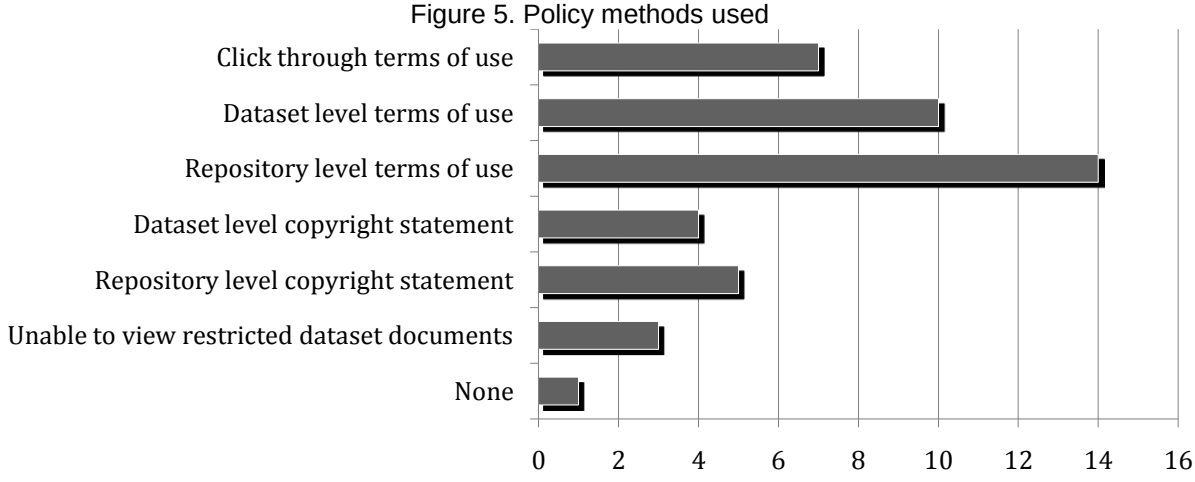


Figure 6, next page, shows other limitations repositories placed on the use of data. Of the other limitations examined, the most commonly occurring was a prohibition on further unapproved dissemination of data with eight repositories having such a restriction (44.4%). Seven repositories required data users to report how data had been used (38.9%), and five prohibited the selling of data (27.8%). Three repositories each

prohibited contacting individuals described in the data, restricted uses of data not previously approved, and required data to be kept secure (16.7%). Two repositories each prohibited uses that would lead to harm, required data to be destroyed after a set period of time, and placed a time limit on access to data (11.1%). One repository required that data users notify the repository in the event of a data security breach (5.6%).

Figure 6. Other limitations placed on data use

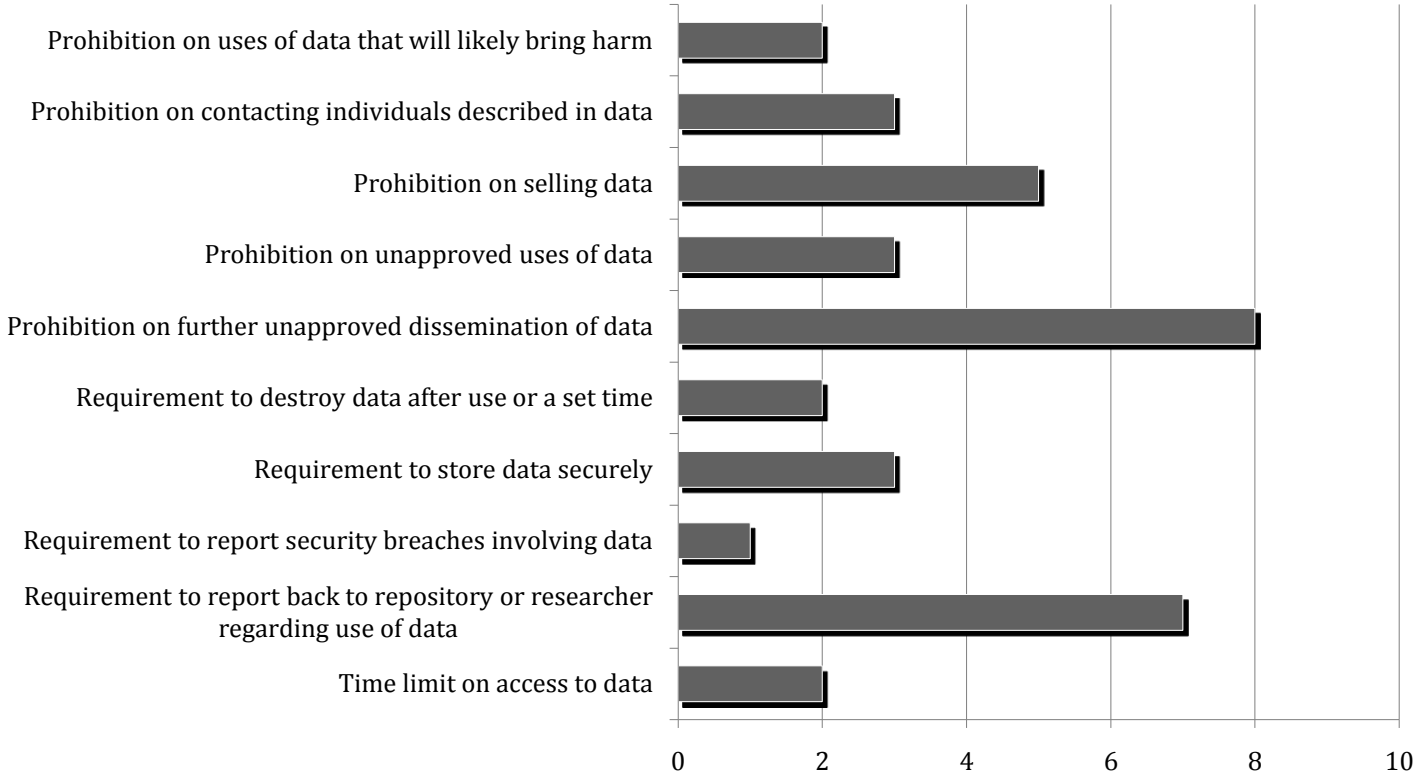
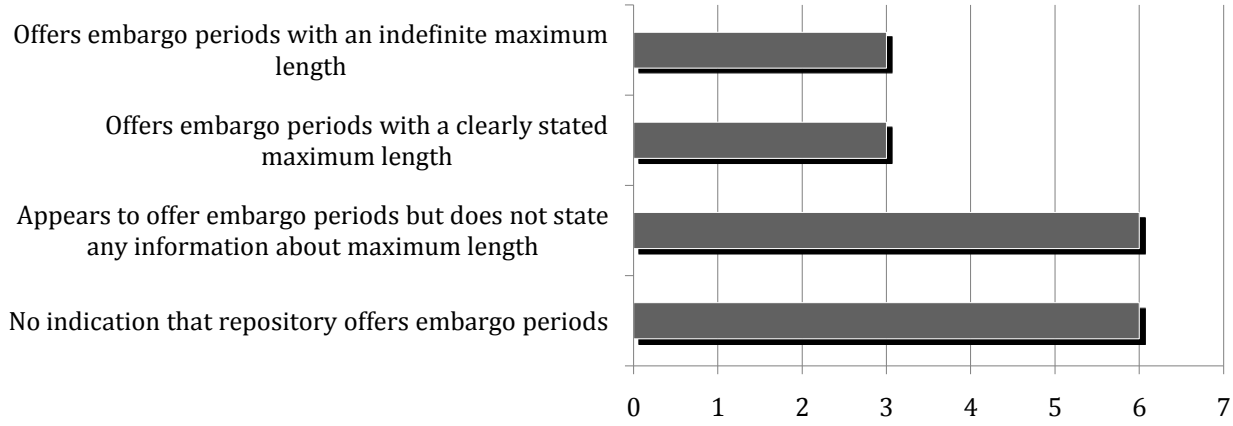


Figure 7 reports the use of embargo periods by the repositories in the sample. Six repositories (33.3%) did not appear to offer any type of embargo period, six offered embargo periods with no information about the length data could remain embargoed (33.3%), three offered embargo periods of a clearly stated length (16.7%), and three offered embargo periods of an indefinite length (16.7%).

Figure 7. Use of embargo periods



Results by Field

Because there are only a small number of repositories within each field category (e.g., four in the biological sciences, two in the humanities, three in the social sciences), it is difficult to assess differences between fields. Many of the patterns from the overall findings appear to hold up across field categories, but the results do suggest some interesting differences:

1. Not surprisingly, only repositories in fields with human subject data (biological and social sciences) cited privacy as a reason for restricting access to data; however, only one repository in each of these fields mentioned privacy as a concern. Ensuring exclusivity was a more common reason in the biological sciences while intellectual property was just as common as privacy in the social sciences.

2. Interestingly, intellectual property was not a reason for restricting access to data for either of the humanities repositories, yet it was a concern for repositories in the earth and social sciences.

3. All of the social science repositories use some form of registration that is either reviewed or requires a formal application process. Social science repositories also account for two-thirds of the repositories that use IP range restrictions to control access to data (the other is from earth sciences).

Further research is needed to see if these differences pan out in larger samples.

The data in this report is currently being confirmed with the participating data repositories. Confirmed data will be made available in a separate publication.