A Pilot Study to Locate Historic Scientific Data in a University Archive

Shannon L. Farrell
Research Data Services Librarian
University of Minnesota
Minneapolis, MN
sfarrell@umn.edu

Julia A. Kelly
Librarian
University of Minnesota
Minneapolis, MN
jKelly@umn.edu

Lois G. Hendrickson
Curator, Wangensteen Historical Library of Biology and Medicine
University of Minnesota
Minneapolis, MN
l-hend@umn.edu

Kristen L. Mastel
Outreach and Instruction Librarian
University of Minnesota
Minneapolis, MN
meye0539@umn.edu

Abstract

Historic data in analog (or print) format is a valuable resource that is utilized by scientists in many fields. This type of data may be found in various locations on university campuses including offices, labs, storage facilities, and archives. This study investigates whether biological data held in one institutional university archives could be identified, described, and thus made potentially useful for contemporary life scientists. Scientific data was located and approximately half of it was deemed to be of some value to current researchers and about 20% included enough information for the study to be repeated. Locating individual data sets in the collections at the University Archives at the University of Minnesota proved challenging. This preliminary work
points to possible ways to move forward to make raw data in university archives collections more discoverable and likely to be reused. It raises questions that can help inform future work in this area.

**Keywords:** University archives, Analog data, Historic data, Research data

**Recommended Citation:**


**Introduction**

Data management services currently being offered by academic libraries focus almost exclusively on machine-readable data that has been produced as part of a current research project. They support funder mandates that require that data be made freely available, but data management services mostly overlook non-digital data that has long existed in labs, offices, and archives on college and university campuses. This data, in analog (or print) format, includes numeric or descriptive data (either handwritten, typed, or printed out) as well as photographs and other items.

While much of this data is widely scattered across campuses and is under the control of either individual researchers or their departments or centers, some portion of it has been deposited and is held in the archives of the institution. In most cases, any data in university archives is under the control of the library.

This data is an institutional asset and like the currently produced machine-readable data in data repositories, it has the potential to be reused to inform future research, to be incorporated into new projects, or to be compared with current observations.

The purpose of this initial study was to determine whether potentially usable data exists in the University Archives at the University of Minnesota (University Archives) and to sample the data that was discoverable. We wanted to gain a better understanding of terms used to describe analog data as well as how processing guidelines and content standards affect users’ ability to discover these records. In order to recommend whether a project like this one would be both feasible to repeat at other institutions and scalable to cover a larger segment of the archives, we kept track of the time spent assessing collections and individual boxes for the presence of data. We also wanted to consider possible avenues to make the existence of this data discoverable.

**Literature Review**

Historians are the primary users of academic archives although a few librarians and archivists have speculated about growing use by science researchers. Tulane researchers suggest that their historical archive might have material of interest for fisheries, transportation, and energy researchers and pose questions about which scientists might be the most interested and how those potential users would find the material ([Kearney](#))
Researchers at the University of Maryland did a pilot with three case studies to assess a framework for assessing the potential reusability of data in archival collections at the National Agricultural Library (Hoffman et al., 2020). They are also conducting a survey of scientists “investigating the landscape of current data recovery efforts and identifying differences and commonalities among the priorities, processes, and practices of data recovery and reuse in different research contexts” (Shiue et al., 2021, Abstract).

Historians and other regular users of archives are accustomed to searching by the name of the primary researcher (Duff & Johnson, 2002). The same is not the case for scientists and social scientists. In a study of the large data repository at the Inter-university Consortium for Political and Social Research (ICPSR), the authors found that subject searches were by far the most common. Of the top 500 searches in 2014, 73% (365) were of the “keyword or phrase” type (Table 2). Just over 25% of the top searches in 2014 were for a specific study or serial collection. Only 1% referenced an author or primary investigator of a study. (Pienta et al., 2017, p. 366).

Additionally, Grivell (2006) of the European Molecular Biology Organization discusses the use of subject keywords in a variety of search systems and Krämer et al (2021) found that for social scientists, subject-based literature searches were a common way to initiate a search for research data. Assante et al (2016) investigated 5 scientific data depositories and noted critical features including rich subject description in the metadata.

Although a few archivists are involved in data repositories, only in recent years has their literature begun to explore data in particular (Niu, 2016; Noonan & Chute, 2014). In the work “Appraising the records of modern science and technology: A guide,” Haas et al. (1985) assume that there are data centers to collect raw data in many fields, focus their attention on machine-generated data from the hard sciences, and make distinctions between observational and experimental data. It has been suggested that since archives cannot afford to collect and document the work of all scientists, they might focus on those who are renowned, those who are connected to the history of the institution, and perhaps those outside the institution who have contributed to an area of particular strength of the institution (Janzen, 1980). Bias and subjectiveness in acquiring records as they relate to the kinds of data available for future researchers are the subjects of discussion by others (Schoenebeck & Conway, 2020; Carter, 2006). At least two studies have highlighted the importance of learning about the practices of scientists in order to better deal with their materials (Akmon et al., 2011; Lauriault et al., 2007).

Field notebooks are one category of material containing raw data that have consistently found their way to archives. They have sometimes been the object of digitization projects such as the Smithsonian Institution Archives’ Field Book Project (https://siarchives.si.edu/about/field-book-project) and the subsequent Biodiversity Heritage Library’s Field Notes Project (https://www.biodiversitylibrary.org/collection/FieldNotesProject) that currently involves eight institutions.

A recent literature review provides an overview of the use of older analog data including discussions in some scientific subdisciplines about the challenges of
preserving older data (Kelly et al., 2022). The use of historical data is not uncommon in research, especially in the life sciences (Farrell & Kelly, 2018; Griffin, 2015; Kelly, 2016; McClenachan et al., 2012; Wiebe & Allison, 2015). However, researchers seldom make note of using an archive at an academic institution as a source of historic data (Kelly et al., 2022). This means it is not clear how much they are utilized, although this may change as the use of data citations becomes more common, and the possibility to add citations to data sets in finding aids is explored, thus allowing papers to more easily cite archival sources and link to finding aids. As discussed by Gracy (2015), linked data might offer a partial solution. One example of the use of historic data is Pagnotta et al (2009) who reexamined astronomical plates found in the archives at Harvard in order to discover nova eruptions. Another example of reuse of material from an archive is from the University of Minnesota. Buma utilized the records of William S. Cooper to study plant succession in Glacier Bay, Alaska (Buma et al., 2017; Buma et al., 2019). Although the older data is not cited in the references, this statement does appear in the acknowledgements: “All historical data is available from the University of Minnesota Libraries’ Archives” (Buma et al., 2019, p. 11). He also went on to note the importance of analog data in general in a separate publication (Buma, 2018).


### Methods

For the purpose of this study, raw data was defined as numeric data, descriptive data such as field notes, and other materials including photographs, maps, and drawings. A data set is raw data within a single study or set of observations.

In order to locate collections in University Archives at the University of Minnesota that potentially contained raw data in the agricultural and life sciences, we began with a very simple search of the online finding aids. From initial investigations it was clear that the description used in the finding aids was mainly free text, due in part to processing guidelines and standards that do not contain consistent descriptive terms to note the presence of data. That means there was no consistent way to do a comprehensive search. Since the objective of this study was to locate raw data, but not necessarily to be exhaustive, we simply searched for the word “data” anywhere in the finding aid, knowing that this would not be comprehensive. We then assessed each collection that was retrieved to limit to agricultural and life sciences topics. We also looked to see where the word “data” appeared and from there made a judgment about whether the collection might actually contain raw data or not.

Collections were screened both in print and online. Print collections were searched on site in the University Archives’ reading room using their protocols which specify that only one folder or bound item may be removed from a box at a time and only one page
may be removed from the folder at a time. Any folder or item that appeared to have potential for containing data was investigated. This was judged by the title or folder name as well as the type of material that it contained. Four of the 14 of the collections that were assessed had at least some material that had been digitized and in those cases the items were usually, but not always, viewed online.

Although we knew that we eventually wanted to provide a robust description of any data that we located, we were not sure where that description might reside in the future or how it would be used to make the data more discoverable. In order to try to make the description compliant with systems that currently exist, we looked at several metadata schemas including Data Documentation Initiative and Darwin Core that are used for science data in electronic data repositories and designed our inventory to include the fields that would be necessary to utilize them. These included format, creator, dates, topic, and location of study. We also noted the amount of data, whether it was numeric or descriptive, if a description of the methods was included, whether variables could be identified, an assessment of whether the study could be repeated based on the methods provided, and if the data was potentially reusable or of use to current researchers as background. We tested the inventory with data from two collections from the University Archives before beginning the project in earnest. All collections that needed to be assessed in person were conducted by members of the research team. Student employees, all of whom were majoring in the biological sciences, were trained on how to review digital collections and assess them for raw data.

As the collections were assessed, any terms that might be used in the future to search the finding aids for raw scientific data were recorded. Examples include field notes, counts, studies, observations, photographs, and experiments. We recorded a total of 75 terms (https://hdl.handle.net/11299/253806).

Results

Our initial search resulted in 32 agricultural or life sciences related collections that potentially included raw scientific data. After studying the finding aids to judge the likelihood of actually finding raw data, we selected 14 for further review. Eight of the collections represented the work of an individual researcher, three were collections from academic departments, and there was one each from an off-campus research center, a biological field station, and a natural history museum. See Table 1 for details.

Of the 14 collections, 13 yielded raw data. In total, 1,374 separate instances of data were identified and the total number of pages of data was 51,911. Of the 1,374, 304 or 22% of the instances included a description of the methods that was judged to be sufficient to be able to repeat the study. In all, 627 instances of data or 46% were deemed to contain data that might be of use to current researchers. Topics represented a wide range across the agricultural and life sciences from bird counts in Wisconsin to peach harvests in Minnesota to tree species observed in Asia.
Table 1. Archival collections assessed for data in the University Archives

<table>
<thead>
<tr>
<th>Name of Collection</th>
<th>Date Range</th>
<th>Total Boxes</th>
<th>Boxes Pulled</th>
<th>Boxes with Data</th>
<th>Data Items Found</th>
<th>Type of Data</th>
<th>Geographic Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Botany researcher</td>
<td>1939-1964</td>
<td>9</td>
<td>8</td>
<td>8</td>
<td>40</td>
<td>X</td>
<td>MN, outside US</td>
</tr>
<tr>
<td>Wildlife management</td>
<td>1939-1969</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>53</td>
<td>X</td>
<td>MN, other US states</td>
</tr>
<tr>
<td>Plant pathology</td>
<td>1948-1991</td>
<td>51</td>
<td>8</td>
<td>5</td>
<td>49</td>
<td>X</td>
<td>MN, other US states, outside US</td>
</tr>
<tr>
<td>Agronomy researcher</td>
<td>1961-1962</td>
<td>8</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>X</td>
<td>MN, outside US</td>
</tr>
<tr>
<td>Fisheries researcher</td>
<td>1947-1978</td>
<td>21</td>
<td>9</td>
<td>9</td>
<td>137</td>
<td>X</td>
<td>MN, outside US</td>
</tr>
<tr>
<td>Ornithology researcher</td>
<td>1953-1976</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>Ecology researcher</td>
<td>1956-1998</td>
<td>11</td>
<td>3</td>
<td>2</td>
<td>20</td>
<td>X</td>
<td>Other US states</td>
</tr>
<tr>
<td>Plant pathology</td>
<td>1921-1952</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>15</td>
<td>X</td>
<td>MN, other US states, outside US</td>
</tr>
<tr>
<td>Dept of Agricultural</td>
<td>1933-1950</td>
<td>9</td>
<td>3</td>
<td>2</td>
<td>7</td>
<td>X</td>
<td>MN, other US states, outside US</td>
</tr>
<tr>
<td>Biochemistry</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Unknown</td>
</tr>
<tr>
<td>Dept of Botany</td>
<td>1879-1996</td>
<td>186</td>
<td>22</td>
<td>16</td>
<td>184</td>
<td>X</td>
<td>MN, other US states, outside US</td>
</tr>
<tr>
<td>Dept of Horticultural</td>
<td>1890-1953</td>
<td>35</td>
<td>3</td>
<td>3</td>
<td>97</td>
<td>X</td>
<td>MN</td>
</tr>
<tr>
<td>Science</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Biological field station</td>
<td>1960-1961</td>
<td>10</td>
<td>1</td>
<td>1</td>
<td>31</td>
<td>X</td>
<td>MN</td>
</tr>
<tr>
<td>Forestry center</td>
<td>1911-1956</td>
<td>12</td>
<td>4</td>
<td>4</td>
<td>30</td>
<td>X</td>
<td>MN, other US states</td>
</tr>
<tr>
<td>Natural history museum</td>
<td>1863-1958</td>
<td>180</td>
<td>44</td>
<td>44</td>
<td>710</td>
<td>X</td>
<td>MN, other US states, outside US</td>
</tr>
<tr>
<td>Totals</td>
<td></td>
<td>537</td>
<td>112</td>
<td>99</td>
<td>1,374</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Data that was identified was collected from 1863-1998. Figure 1 shows when the data began to be collected by decade for each data set. Table 2 shows the numbers of data items that were collected locally in Minnesota, in the United States outside of Minnesota, and outside the United States when noted in the data set. The majority represents research conducted in Minnesota. The research conducted outside the United
States included every continent except for Antarctica. Table 3 lists the various physical formats of the data.

![Graph showing data sets identified in the University Archives by decade](image)

**Figure 1. Data sets identified in the University Archives by decade**

<table>
<thead>
<tr>
<th>Location of Research</th>
<th>Number of Data Sets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minnesota</td>
<td>1,069</td>
</tr>
<tr>
<td>Other US states</td>
<td>119</td>
</tr>
<tr>
<td>Outside the US</td>
<td>106</td>
</tr>
</tbody>
</table>

Table 2. Location of data collection noted in the data set

<table>
<thead>
<tr>
<th>Physical Format</th>
<th>Number of Data Sets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Loose leaf paper</td>
<td>669</td>
</tr>
<tr>
<td>Notebook</td>
<td>462</td>
</tr>
<tr>
<td>Folder</td>
<td>217</td>
</tr>
<tr>
<td>Binder</td>
<td>19</td>
</tr>
<tr>
<td>Notecard</td>
<td>4</td>
</tr>
<tr>
<td>Bound checklist</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 3. Physical format of data sets

In all, 112 boxes from the 14 collections were inventoried. Time spent per box was recorded for 75% of the boxes and it varied widely, from less than 15 minutes to over 50 hours. The latter represents three boxes from the natural history museum that each contained many field notebooks. More than half of the boxes took two hours or less to complete and the average when excluding the three outliers was just over four hours per box.

**Discussion**

Collections in the University Archives do contain raw data and the majority of what we found in this initial sampling was deemed to be of some potential use to current scientists. However, the range of ways that it might be used varies greatly. Only 22% of
the data sets were complete, well labeled, and had a detailed enough description of the methods for the study to be repeated. In many cases, if a full description of the methods was found, it was part of a draft manuscript that accompanied the data. The number of studies that could be repeated would likely increase if we had gone a step further and identified any publications, including theses and dissertations, that were generated from the data that was located in the University Archives. Those papers would likely have a detailed “methods” section and spell out some abbreviations that were sometimes not fully identified in the archival material. We undertook a small pilot to locate the papers in which particular data sets were used, but it proved to be very labor-intensive, so we did not pursue it further.

The data sets that were deemed useful but did not include explanations of the methods would often fall into the category of background material that could inform current or future research. Examples include lists of species observed in a particular geographic area, studies of animal behavior, descriptions of shape, color, and flavor of horticultural crops, or notes on bird migrations. Some lacked details about exact locations or dates and others were simply descriptive in nature. Some of the data that was located was completely unusable and included pages full of hash marks, tabular data with no column headings, and columns or rows that were labeled with undefined abbreviations.

Given that life sciences researchers at large universities do not necessarily carry out their work in their own localities, it was not surprising to see that the data found in the University Archives was collected around the globe. The fact that the natural history museum was included in the study probably skewed the results toward Minnesota locations. Their 710 data sets, which represent 52% of the 1,374 in this study, included 603 that were collected in Minnesota locations.

Discovering these data sets was not simple. The biggest challenges were related to the description and organization of the collections. Life scientists searching for raw data are not currently among the primary audience for materials in a university’s archives, and the descriptions in the finding aids reflect that. When collections are being processed and the finding aids created, there are, as of yet, few guidelines or standards for data sets, especially raw data. Information about data could reside in multiple places in a finding aid, and there are no controlled vocabulary terms to apply to highlight it. A variety of terms hinting at data may be contained in the scope and content notes for entire collections, some in free text description about individual boxes or folders, or none may be present, so all boxes in a collection would have to be examined. The naming scheme for the folders that was used by the individual researchers or departments is typically retained and reflected in the finding aids, so each one we discovered was unique, and dependent on the archivist inputting that information somewhere in the finding aid. Folder names were often vague; some folders with promising names contained nothing of interest and others had rich contents but a name with no indication of that. This led to many dead ends.

If librarians and archivists are interested in having scientists discover and reuse data currently residing in archives, the routes of discovery should be enhanced. Scope and content notes could be written with a broader audience in mind and as noted above, and an additional scope or content note should be added at the folder level so users
looking for this information would be pointed directly to where it is found in the collections. This would make finding the data a lot easier especially if it is a large collection. Archivists might also consider creating a note with a specialized label, or labeling data sets in a multipart note so that all data sets in all collections could be easily located and discoverable. It is our observation and experience that historians may follow the trail of a single researcher and discover what institution’s archives hold their papers, but scientists are in the habit of looking for journal articles by using a global database and searching by topic and they will probably expect to do the same when searching for historic data. In an ideal world, the data in an archive would also be linked to the journal article or other report of the project so that details about the methods would be readily available.

If data in university archives are to be more easily discovered, an effort could be made to gather more detail concerning the data when the collection is initially accepted into the archives. If the creators of the data are not available, tapping the expertise of researchers that know the field can be useful. In a recent project to scan and add older fruit breeding data to the University’s institutional repository, the authors worked with a current fruit breeder to jointly write readme files for the data to accurately describe the methods and allow reuse of the material (https://hdl.handle.net/11299/201897).

Some of the collections were available online which initially allowed work to be done outside of hours of the University Archives reading room and later allowed the project to continue during the pandemic when all work was remote. Although those are positive aspects, all seven of the individuals involved in the searching and screening preferred to examine the materials in print. Everyone involved in the project has a background in life sciences that was very helpful, especially when deciding if and how the data might be reused.

Examining the materials in print under the rules in place in the University Archives reading room could provide a challenge for science researchers. Only one folder can be out of the box at a time and pages are to be viewed one at a time, precluding most comparisons. The rules are in place to protect fragile materials but this is not the way that scientists currently operate based on the authors’ experience. Workarounds include taking photographs, making arrangements for photocopies, or printing copies of materials that are available electronically, but all of these add time and effort to the process.

This initial investigation into potentially reusable data existing in one university archive yielded some answers but also questions. The authors believe that a first step toward addressing those questions would be to broaden this study by investigating the holdings at other university archives.

We now know that there is potentially useful data in at least one university archive and probably many more. If the overall goal is to foster the connection between that data and current and future life science researchers, how do we as librarians and archivists see that realistically unfolding? If we conduct similar surveys in a few archives across the world and publicize it widely to scientists, will that catalyze use of data from archives? Should one or two subject areas be selected for more intensive study across
archives? Is it feasible to set a long-term goal of locating and describing as much of this data as possible and making those descriptions easily findable and searchable? Given what we found about the time it took to assess collections for data, how can it be reasonably resourced? Do we proceed and to what extent?

Whether a small amount of data ends up being located and more thoroughly described or if it is a larger corpus, how would researchers find those descriptions? Data registries could be developed but would they be utilized by scientists if registries are not in their regular workflow? Could descriptions be included in existing data repositories and point users to the location of the analog data, similar to the citations to journal articles in an index? Only a few data repositories currently include anything but the machine-readable data itself. An example is AgData Commons (https://data.nal.usda.gov/) that accepts scanned but not necessarily machine-readable data. Another option would be to write data papers covering particular data sets in archives and publish those papers in the growing number of journals that accept that type of publication. Those papers would be covered in indexes that researchers already use on a regular basis, an option that is much closer to their current workflow rather than introducing a new tool such as a data registry.

**Conclusion**

This investigation into potentially reusable data in one university archive found that these collections did indeed contain raw data. The majority of the data that was found was deemed to be of some potential use to current scientists. However, the data ranged from studies that could be replicated to those that could be used as background information to inform current or future research to those that were completely unusable. Finding the data was not easy in any case. If data in university archives are to be more easily discovered, an effort should be made to gather more detail concerning the data when the collection is initially accepted into the archives.

If librarians and archivists are interested in having scientists discover and reuse data currently residing in archives, the routes of discovery should be enhanced. Some of the questions that could be investigated include: How can researchers find descriptions of reusable data - whether that is a small amount of data that is found and ends up being more thoroughly described or if it is a larger corpus of data? If data registries were developed, would they be utilized by scientists outside of their regular workflow? Could descriptions be included in existing data repositories and point users to the location of the analog data, similar to the citations to journal articles in an index? How would this be possible, since few data repositories currently include anything but the machine-readable data itself? Despite not having answers to these questions, the authors’ first steps might include working with archivists and scientists to enhance the descriptions of data in finding aids and other places where data is located.

Working with historical analog data could open new avenues for academic librarians and archivists. Much of this archival data is already considered to be in the purview of librarians and archivists and could lead to deeper interactions with researchers about their data and its description and management.
Acknowledgements

The authors would like to thank Kate Dietrick, Processing Archivist and Archivist for the Nathan and Theresa Berman Upper Midwest Jewish Archives at the University of Minnesota for providing insight and expertise.

References


This work is licensed under a Creative Commons Attribution 4.0 International License.