RESEARCH ARTICLE / ARTICLE DE RECHERCHE

What Not to Keep: Not All Data Have Future Research Value[1]

Janice Yu Chen Kung[2] and Sandy Campbell

Abstract: The rise of academic library involvement in research data management has presented numerous challenges for academic libraries. Although libraries and archives have always had collection development policies that defined what they would or would not collect, policies for selecting research data for preservation are in their infancy. This study surveyed and interviewed health sciences academic researchers. From this research an initial list of eight types of health research data were identified as data that should not be preserved and made public. These include research data that are: sensitive or confidential; proprietary; easily replicable; do not have good metadata; test, pilot, or intermediate data; bad or junk data; data that cannot be used by others for a variety of reasons; and older data that are not used and have no obvious cultural or historical value. Conclusions drawn from the study will help librarians and archivists make informed decisions about which types of research data are worth keeping.

Introduction

Data curation and data preservation go hand in hand in that they both manage data through its lifecycle to ensure that datasets are retrievable for validation purposes or future use. The rise of research data management in health sciences has created new challenges for academic libraries and archives. Due to increasing pressures from government agencies and regulatory bodies to adopt open data policies, information professionals and researchers face new challenges related to how data should be managed during and after research projects and what types of data should be preserved. Based on the Canadian Tri-Agency (Canadian Institutes of Health Research (CIHR), the Natural Sciences and Engineering Research Council of Canada (NSERC), and the Social Sciences and Humanities Research Council of Canada (SSHRC)) statement on digital data management, the federal granting agencies are fostering "open science" whereby future researchers can access publicly funded research data and results for reuse [1]. When libraries consider which data should and should not be kept, it is usually with the intent that the preserved data will be made open for use by future researchers.

The Tri-Agency principles recommend that in deciding what data to share and preserve, researchers consider "the data needed to validate research findings and results, support replication and reuse and consider the potential benefit to their own fields of research, fields other than their own and society at large" [1]. They also recognize that "data must be managed with all commercial, legal and ethical obligations" [1] and that "not all data may need to be shared or preserved" [1]. Various scholarly studies have broadly reviewed aspects of academic research data including the kinds of data created, how researchers manage data, barriers to data curation, and how libraries and archives can support researchers in managing their data [2, 3]. However, they do not address criteria for data that should not be kept. Further, these studies were not specific to academic health research environments.

Several organizations have published guidelines for the retention of data. The United States National Oceanic Atmospheric Administration (NOAA), which preserves atmospheric and oceanic data, developed guidelines for preservation of data. The guidelines are informed by several factors including: the evaluation of societal benefits, the uniqueness of the data, and consultation with external groups such as the broader community and other agencies [4]. NOAA cites several kinds of data that should not be preserved: obsolete or redundant data, data for which storage costs exceed the cost of reproducing or regenerating the data, data that have little value once the project ends, and multiple versions including raw data and manipulated data [4]. Further, Tjalsma and Rombouts [5] describe pre-conditions that must be met for data to be preserved. These include usability of the data formats;

Janice Yu Chen Kung. John W. Scott Health Sciences Library, 2K3.26 Walter C. Mackenzie Health Sciences Centre, University of Alberta, Edmonton, AB.
Sandy Campbell. John W. Scott Health Sciences Library, 2K3.26 Walter C. Mackenzie Health Sciences Centre, University of Alberta, Edmonton, AB.
[1]This article has been peer-reviewed.
[2]Corresponding author (email: janice.kung@ualberta.ca)

adequacy of metadata; whether the data is raw, intermediate, or published; clarity on intellectual property rights including copyrights, patent, and privacy; and availability of appropriate infrastructure, and preservation costs [5]. Data not meeting these criteria would not be preserved. Although these guidelines and recommendations are helpful to libraries and archives for the development of policy, they are presented from an institutional perspective, and again, are not specific to health research data. Our study was designed to investigate researcher attitudes and elicit information about what kinds of data academic health researchers think should not be kept by libraries and archives for the purpose of reuse.

Methods

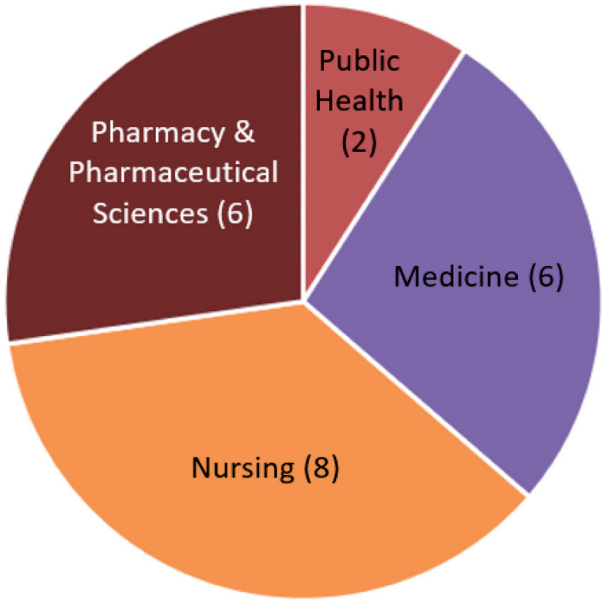
The researchers conducted a qualitative study in two phases, rooted in grounded theory [6]. Ethics approval was granted by the University of Alberta Research Ethics Board 1 for both phases. In the first phase, an anonymous survey was sent electronically to the University of Alberta health sciences community through faculty and department listservs including the Faculty of Nursing, Faculty of Pharmacy and Pharmaceutical Sciences, School of Public Health, Departments of Dentistry and Dental Hygiene, and Department of Medicine. The survey, accompanied by an information letter describing the project, was administered from February to March 2015. The information letter can be seen in Supplementary Appendix A and the complete list of the questions in Supplementary Appendix B. Responses garnered from the survey helped inform the questions for the next phase of the project.

In the second phase, between April and June 2015, the researchers conducted semi-structured interviews with survey respondents who self-identified and expressed interest in providing additional information. Because the survey responses were anonymous, participants were asked to provide contact information to inform the project team that they were amenable to being interviewed. Using convenience sampling, the researchers also approached faculty members and health sciences researchers with whom they had had previous working relationships to recruit more participants. At the beginning of each interview, the participants were given an information letter describing the project, and the purposes of the project were reviewed verbally with them. They were also asked to review and sign permission forms to allow the interviews to be audio recorded. Samples of the information letter and permission forms are in Supplementary Appendix A. For a complete list of the interview questions, see Supplementary Appendix C. Qualitative content analysis was used to conceptualize the data by identifying major themes. With the application of inductive coding, the researchers concurrently collected and analyzed data during both phases.

Results

There were 22 survey respondents: 15 faculty members, 2 clinical instructors, 1 graduate student, 1 undergraduate student, 1 research fellow, 1 research assistant, and 1 adjunct/clinician. Figure 1 outlines the faculties with which the respondents were affiliated.

Fig. 1. Survey respondents by discipline.



Findings from the survey indicate that eight of researchers had permanently preserved data in institutional repositories, on personal servers, or as supplementary material in publications. Almost half of the researchers (9 respondents) were in possession of data that could not be published due to confidentiality or proprietary concerns. The majority, 15 respondents, think that some data should not be preserved permanently, for example pilot data and un-validated data. Of the 22 respondents, 10 indicated that they were aware of risks or problems that might be inherent in permanently preserving research datasets. Some examples included liability issues when working with patient data, security, and confidentiality challenges.

Eight researchers participated in the interviews affiliated with each of the following disciplines: Public Health (1), Cell Biology (1), Medicine (4), and Nursing (2). In addition to being researchers, data creators, and users in their own right, three of the participants had further responsibilities with research data. Two were departmental data repository administrators and one managed research in a department.

In addition to recording the interviews, interviewers also took notes to confirm the conversation captured in the audio recording. Interview transcripts and notes were reviewed to identify references to types of data that should not be preserved by the libraries and archives. Two researchers (SC and JK) reviewed the transcripts and interviewer notes to identify themes related to data that should not be kept. Related comments were then grouped into eight categories of health research data that researchers thought should not be preserved and made public. These categories are presented here in alphabetical order: bad or junk data; data that are easily replicable; data that cannot be used by others; data without good metadata; older data that are not used and have no obvious cultural

or historical value; pilot, test, or intermediate data; proprietary data not owned by the researcher; and sensitive or confidential data.

A ninth theme that arose from the data concerns the importance of involving data creators in the data management lifecycle.

Bad or junk data

Bad or junk data implies that the data are not usable or reusable by researchers. McCallum [2] describes bad data as data that have missing values, have malformed records, and are stored in problematic file formats. One of the interview subjects, a cell biologist, considered bad or junk data as data collected without rigorous methodology or a scientific approach. For instance, experiments can be contaminated due to factors including temperature, equipment failure, or human error, thus compromising the data. Researchers would record such instances in their lab notebooks, but the data itself would have no research value.

Data that cannot be used by others

There are several reasons that prevent data from being used by researchers, either by the data generators themselves or by secondary data users. When datasets are too specific to be combined with other datasets, it prevents researchers from manipulating them in a meaningful way. Some data require proprietary software that might not be available to future researchers. A researcher from the School of Public Health provided insight into the challenges with using NVivo, a proprietary software used for qualitative studies, especially with the upgrade from Version 9 to 10. She stated:

> NVivo, they change their format and as soon as they change their format you don't have access to your analysis in their other platform unless you keep a copy of that platform on a computer. So people can actually lose access to their own analysis to that level of data because, five years from now, NVivo's going to have a different format.

There may be work arounds and alternative solutions to accessing analyzed data hidden behind proprietary software, but such barriers to access would pose challenges to future researchers such as additional costs required to migrate files to the latest version. If the library cannot afford to maintain older versions of proprietary software that are required to read old data files, then the data files should not be preserved.

Other data that cannot be used by other researchers are data that require knowledge of the context in which it was generated to be fully understood and appreciated. A researcher working with Indigenous youth described how only being in the environment, listening to the youth over time, and understanding their body language as they spoke, made the stories that they told meaningful. Another researcher looking at the data without that contextual knowledge could not fully understand it. Qualitative research and, to a certain extent, quantitative research are context specific. Without the proper documentation and background knowledge, there would be little value in permanently preserving this data collection and making it available to other researchers.

Data that are easily replicable

There are instances where the cost effectiveness of regenerating data on demand makes data preservation impractical. One of the interview participants indicated that data collected through citation analysis projects are easy to regenerate so there is no need to keep the information. Similarly, with systematic reviews, researchers provide replicable search strategies for databases, describe the datasets they use, and any manipulations they do to the data, but they do not keep all intermediate datasets. In cases such as these, as the NOAA criteria point out [4], it makes more economic sense to recollect data at the time of need rather than expending resources to preserve the dataset.

Data without good metadata

Savage and Vickers [7] argue that sometimes the ability to reuse datasets is hindered in part by suboptimal metadata. Descriptive metadata must accompany research data to ensure that future researchers will be able to understand and interpret the dataset. Therefore, datasets are not worth preserving if the metadata are incomplete, not standardized, inaccurate, or inconsistently applied. This echoes one of Tjalsma's and Rombouts' pre-conditions [5]. One of the interview participants claimed that only 2% of the collected data from his research would ever be published but he felt that the remaining 98% would still be useful if someone applied metadata to it. However, limited staffing resources preclude this, making use of the data by secondary users very difficult. He pointed out that there are no rewards at the annual faculty evaluation review for the application of metadata to unanalyzed data.

Older data that are not used and have no obvious cultural or historical value

The concept of finite space for storage and preservation is not unique to physical libraries; it applies to digital collections as well. According to the guidelines from NOAA, obsolete or redundant data should not be archived [4]. Since server space and administrative costs are not infinite, NOAA also recommended reducing access to older or less commonly used data rather than removing data from the archive [4]. Not all data are valued equally so it is necessary to evaluate the current and potential future research value of datasets to assess the feasibility of archiving and access requirements to those that are less well used. Data that cover short periods of time, small samples, and have no cultural or historical content would have less future use than longitudinal data, large studies, and culturally based studies. Although some data are not used regularly, caution must be observed when weeding and additional criteria need to be applied. The literature considers any data with historical value as "heritage" [5], including data that support the history of science or cultural heritage.

Pilot, test, or intermediate data

Data derived from instrument testing or trial runs have little future research value since they are used for calibrating lab equipment and testing the data collection methods to ensure that the results will answer the research questions appropriately. Sometimes there are so many iterations of data generated while developing a method that they are

only retained during the test phase, and they may not be documented as thoroughly as the data collected during the full-scale project. Another researcher was adamant that only raw data should be kept, along with a very detailed description of what manipulations had been done to achieve the final research outcomes. In his opinion, all intermediate data should be discarded if not required for validation. This aligns with Tjalsma and Rombouts' [5] view to use primary data over secondary data for verification purposes, when there is a need to recreate the environment from which the analyses were initially performed.

Proprietary data

Proprietary data appeared in both the survey responses and the researcher interviews. This is a category that is well-understood in academic environments. For example, the University of Exeter in its guidelines confirms that data do not have to be released to the public if there are commercial factors to consider [8]. The Tri-Agencies also recognize the importance of recognizing "commercial, legal, and ethical obligations" [1]. Often researchers do not have ownership rights to the data with which they are working but, rather, are working with data that have been released under contract by companies or organizations for a particular research project. Sometimes these data are supplied with the understanding that they will be used by one individual or one research team only. A notable example is drug information that comes from pharmaceutical companies released to academic scholars for research purposes. If the researchers are unable to ever make these data public, then there is no purpose for academic libraries and archives to preserve them once the researcher has finished working with them and the period for validation of results is over.

Sensitive or confidential data

The issue of confidential data was raised both by respondents to the survey and by interview participants. When research involving human subjects is being conducted, ethics agreements often define when data must be destroyed (for example, five years after collection). Researchers must abide by these restrictions. Participant consent forms may also assure participants of the data destruction date. In the age of digital curation, it is critical to ensure "conformance to funder requirements and managing institutional risk and liability" [9]. Funding agencies also recognize the importance of maintaining the privacy of certain data. The Tri-Agency Open Access Policy stipulates that there are some types of data that CIHR-supported researchers do not have to archive, including personal or sensitive data and administrative, clinical, and longitudinal data [10]. The Open Access and Data Curation Team from the University of Exeter affirms this practice by acknowledging funders' requirements in the United Kingdom as well as the need to adhere to the Data Protection Act, which protects individuals from being identified from those data and other pertinent information [8]. Dryad, an international data repository, further supports the privacy of confidential and sensitive data by not accepting data submissions in which human subject data have not been anonymized [11]. Data preservation policies by government and funding agencies must be acknowledged, such as the Tri-Agency Open Access Policy [10] and guidelines developed by Research Data Canada that explore similar issues on metadata, privacy, confidentiality, and version control [12].

A ninth theme: community involvement in data management

In the analysis, an important ninth theme emerged from the interviews with researchers that was not a category of data, but rather related to the need to involve users who create and deposit datasets in the decision-making process. This is one of NOAA's recommendations, as well [4]. Data creators must be involved in data preparation, such as the creation of metadata. This is also true when depositing and weeding datasets in the data management lifecycle described by NOAA (see Figure 2). The two red stars represent two decision points in the data management lifecycle whereby decisions particularly require the user community's input.

Discussion

Although we recognize that the list of eight types of data is probably not exhaustive, it does represent the kinds of data that academic health researchers in the study identified as data that should not be preserved and made open for reuse. All libraries and archives have guidelines defining the kinds of materials that they will and will not collect. This list will aid in the development of libraries and archives collections policies with regard to which research data will and will not be kept. The ninth theme that emerged, the importance of the involvement of the original creators and owners of the data is a reminder that no matter what inclusion and exclusion policies are established, library and archival data repositories need to work closely with their communities to ensure the viability and continued usefulness of the data that they collect.

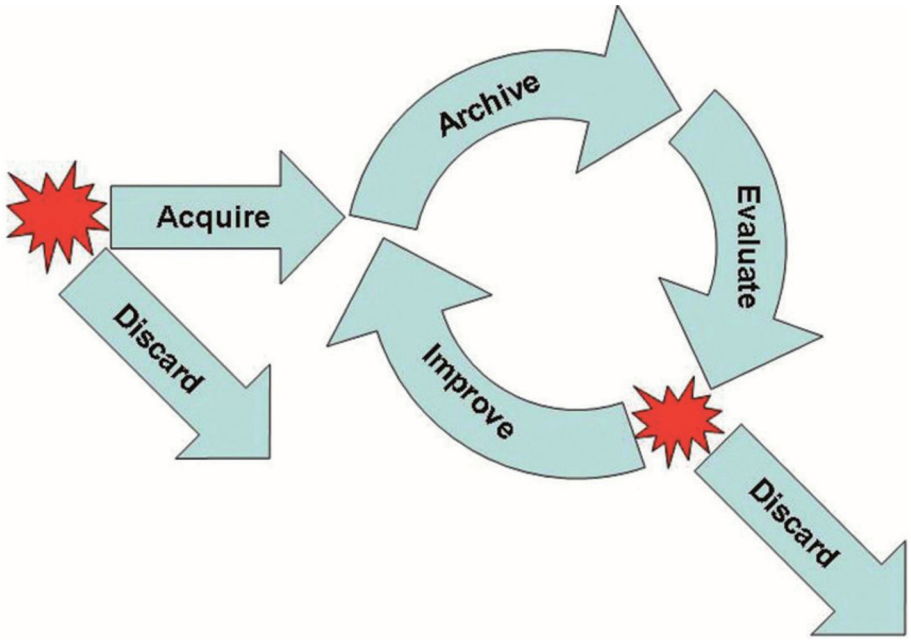
The limitations of the study include the survey's low response rate and the potential bias from the interview respondents, as more than half of the interview participants were contacted directly by the researchers based on previous working relationships with them. As a result, the participants may not be representative of all health sciences researchers at the University of Alberta or elsewhere. The categories of data types that should not be archived grew out of the two authors' analysis of the interview and survey results and were finalized through consensus. Although some of the categories exist in the examples provided by the literature, it is possible that other researchers might group commentaries into other distinct categories.

Future research arising from this study would include a study of the applicability of these guidelines to library and archival data preservation and storage for data generated in disciplines other than the health sciences.

Conclusions

This study further defines, from a sample of health researchers' points of view, which data should not or cannot be maintained in libraries and archives for the purpose of being made open for reuse. From the survey we

Fig. 2. NOAA's data management lifecycle [4].



learned that researchers are aware of the need to preserve data, but are also aware of data that should not be preserved. From the interviews we learned in more detail about the characteristics of data that should not be permanently preserved.

To date, a comprehensive preservation policy does not exist for curating datasets in the health sciences domain. This study is a contribution to the establishment of more detailed library and archival best practices, policies, and procedures for the preservation of health research data, specifically by identifying which data should not be preserved.

References

1. Government of Canada. Draft tri-agency statement of principles on digital data management [Internet]. Ottawa (Canada): Science.gc.ca; 2015 [cited 17 Oct 2015]. Available from: http://www.science.gc.ca/default.asp?lang=En&n=83F7624E-1
2. McCallum QE. *Bad data handbook*. Sebastopol (CA): O'Reilly; 2012. 245 p.
3. McLure M, Level AV, Cranston CL, et al. Data curation: a study of researcher practices and needs. *Libr Acad*. 2014;14(2): 139–164. doi: 10.1353/pla.2014.0009.
4. *Environmental data management at NOAA: archiving, stewardship, and access*. Washington (DC): National Academies Press; 2007 [cited 25 Nov 2015]. Available from: http://www.nap.edu/catalog/12017.html
5. Tjalsma H, Rombouts J. *Selection of research data: guidelines for appraising and selecting research data*. Den Haag: Stichting SURF; 2011.
6. Corbin J, Strauss A. Grounded theory research: procedures, canons, and evaluative criteria. *Qual Sociol*. 1990;13(1):3. doi: 10.1007/BF00988593.
7. Savage CJ, Vickers AJ. Empirical study of data sharing by authors publishing in PLoS journals. *PLoS One*. 2009;4(9): e7078 [cited 25 Nov 2015]. Available from: http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0007078
8. Cole G, Lloyd-Jones H, Evans J. *What to keep/delete: how to appraise your data (RDP)*. PPT presented at Exeter, University of Exeter; 2013 [cited 25 Nov 2015]. Available from: http://hdl.handle.net/10871/8241
9. Lynch C. The next generation of challenges in the curation of scholarly data. In: *Research data management: practical strategies for information professionals*. West Lafayette (IN): Purdue University Press; 2013. p. 395–408.
10. Government of Canada. *Frequently asked questions* [Internet]. Ottawa, Canada: Science.gc.ca.; 2015 [cited 2 Nov 2015]. Available from: http://www.science.gc.ca/default.asp?lang=En&n=A30EBB24-1
11. Dryad. Templates for correspondence [Internet]. Potentially inappropriate files: human subject data; [revised 24 Sept 2015; cited 25 Nov 2015]. Available from: http://wiki.datadryad.org/Templates_for_Correspondence#Potentially_inappropriate_files:_Human_subject_data
12. Research Data Canada. Strategic documents [Internet]. In: *Guidelines for the deposit and preservation of research data in Canada*. Ottawa (Canada); 2016; [cited 7 Feb 2016]. Available from: http://www.rdc-drc.ca/resources/