DATA CURATION NEEDS OF RESEARCHERS AT MZUZU UNIVERSITY

Felix MAJAWA[1] and Fiskani NGWIRA[2]

[1]University Librarian. Mzuzu University Library and Learning Resources Centre, Malawi
Email: fmajawa@gmail.com; fmajawa@yahoo.com

[2]Assistant Lecturer. Department of Library and Information Science. Mzuzu University, Malawi
Email: fiskangwira@gmail.com

Abstract

The study examined data curation aspects needed by researchers at Mzuzu University. Fifty (50) self-administered questionnaires were distributed to purposefully selected academic staff representing 28% of the total population. 57% of the respondents were not familiar with the term "data curation" however, it was revealed that they actually practiced the aspects associated with data curation. Mixed views were generated from the researchers with 34% not willing to share their data for fear that others may reuse it for other purposes contrary to the original idea. 14% of the respondents indicated that they destroyed their data after conducting research which is contrary to the whole idea of data curation. However, 80% of the respondents agreed that data would be useful for future research. The study revealed unique issues to do with the management of data, data sharing and reuse and challenges. For the fact that data curation is a new concept with a lot of concerns worldwide, there is need for more research in the field to iron out gray areas on the processes of the concept. Researchers should be endowed with knowledge about the benefits of data curation such as data preservation, data archiving, data reuse which prevents duplication.

Keywords: Malawi, Mzuzu University, Data Curation, Digital Curation, Open Access, Data Sharing, Academic staff.

1. Introduction

Researchers in Higher Education (HE) are producing ever-increasing quantities of digital data in the course of their work which needs to be managed for both immediate and potential long-term use (Borgman, 2012; Pryor, 2012). As a result of advances in technology, information processing, and new models of publication, research is now conducted, communicated, and reused in increasingly complex and rapidly changing ways. Opportunities for connecting the final output of research, in the form of a published article, with other important parts of the research process, such as raw data, analyzed data, applied algorithms for statistical analysis, metadata, and documentation, are vast and change the scale and scope of how research is disseminated and connected (Ball, 2012; National Science Foundation, 2011).

The digital revolution is transforming the way in which scientific research is conducted. The relative ease with which digital information can be created coupled with an increase in computing power has led to a proliferation of "born digital" data that is, data or information that was created and exists in a digital format and may never exist in a non-digital form (Hockx-Yu, 2006). The appropriate management, sharing, and reuse of data is becoming progressively common expectation of researchers. The study, therefore, examined data curation aspects needed by researchers at Mzuzu University.

Mzuzu University was established by an Act of Parliament in 1997 as Malawi's second national (public) university. The first batch of students was admitted in January 1999. To date the University has a total enrolment of 400 undergraduate students and 176 academic staff. It has 5 faculties and 21 departments.

2. Problem statement

Data Curation processes consolidate the activities concerned with the management and preservation of data. This is contrary to the actual processes, practiced by some researchers worldwide. For instance, some researchers believe that data sets have to be destroyed after they have been processed, results generated and conclusions drawn. Are the researchers at Mzuzu University ready to preserve, share and re-use data? These are the issues that this study attempts to address. The study examined data curation aspects needed by researchers at Mzuzu University. The preliminary investigations conducted by Majawa and Ngwira (2014) at Mzuzu University to ascertain if academic staff were aware of data curation concepts, show that majority were not aware of data curation concepts. Need therefore arises to explore the Data curations needs at the University.

3. Objectives of the study

The general objective of the study is to examine the data curation needs of researchers at Mzuzu University. The specific objectives of the study are as follows:

- To investigate awareness of the data curation concept.
- To explore the current data curation activities
- To find out challenges associated with data curation processes.

4. Research questions

The research questions of the study are as follows:

- RQ1 Are the researchers aware of Data Curation Concepts?
- RQ2 what are current data curation activities do you do?
- RQ3 what are the data curation challenge experienced?

5. Literature review

Researchers lack knowledge of tools for metadata generation, acquisition, data migration, or validation, all the way through the data life-cycle to culling tools for de-accessioning and destruction. Witt (2008) pointed out that librarians have many skills to bring to bear on research data curation. Librarians have expertise in classification and description of information through metadata services such as cataloging. Technical services and public services provide access points. Reference and instruction assist in finding and using information effectively. Collection management selects, deselects and presents information in the appropriate context (Witt, 2008). This new role for librarians is a natural continuation of already existing roles of acquiring, organizing, and making available resources needed by academic staff and students (Horwood et al., 2004). The data librarian would be responsible for advising researchers on where to locate resources, how to manage the data, and how to gain access to resources or troubleshoot data-related problems (Macdonald and Martinez-Uribe, 2008, 2010).

The source data is rarely made available with the publication of the article there is usually not enough information available in the published article to reproduce the research, a defining principle of the scientific method. Without access to the source data, another scientist must use inference and extrapolation to fill the gap between the information represented in the article and the full potential that could be derived from the raw data (Witt, 2008). Collaborative and multidisciplinary research happens in separate locations, producing and using huge amounts of digital data. Fast changing technology puts digital information at risk of being lost if not properly curated and preserved.

A record of scientific data could itself become a vibrant, useful part of the process and practices of science, whether through data-mining and reuse, data-visualization, or other techniques and methods yet to be invented. Widespread sharing of data may lead to discovery and use outside of the discipline in which the data were created, fostering interdisciplinary research and learning (Witt, 2008). There is also an institutional and community benefit to open data in terms of costs, greater accessibility and long-term preservation of research output (Macdonald and Martinez-Uribe, 2010). Apart from the value in reproducing the original results, shared data can also be used to advance the original research or another line of inquiry. In some cases, preserving and sharing existing datasets could enable them to be reused instead of incurring the expense of generating new data from scratch (Witt, 2008).

Researchers' inadequate storage capacity issues, confidentiality concerns, intellectual property rights and complexity are some of the stumbling blocks to open sharing of data identified by Macdonald and Martinez-Uribe (2010). Brandt (2007) found that researchers are unsure of how or whether to share their data, lack time to organize datasets, need help describing data so they can be found and used, want new ways of managing data, and require assistance in archiving datasets. A properly trained librarian can help researchers navigate all of these issues.

Datasets are the special collections of the digital age. Witt (2008) asked who will sift through these data, select and preserve what is valuable and make it accessible in the future? Librarians' unique skill set of organizing, classifying and preserving information make them well-suited to the task of digital data curation.

6. Methodology

This study adopted a survey design and it targeted fifty (50) academic staff who were purposefully selected. Questionnaire was the main instrument used for data collection. A questionnaire which was tagged "Data Curation Needs of Researchers at Mzuzu University" was divided into four sections. Section "A" elicited information on the background of the respondents with questions such as gander, faculty, department, and position. Section "B" solicited information on the awareness of the data curation concept, while section "C" concentrated on data curation activities. Section "D" concentrated on the data curation challenges. The forty-seven (47) representing (94%) were returned and found usable.

7. Findings and discussion

Gender	**F**	**%**
Male	38	80
Female	9	20
Total	**47**	**100**

Table 1: Gender Distribution

7.1 Gender Distribution

The Table 1, above revealed that majority of the respondents were males 38 (80%), while 9 (20%) were females.

7.2 Respondents by faculty

Out of the 47 respondents, 20 (42%) were from Education, 5 (10%) were from Health Science 3, (6%) were from Tourism and Hospitality, 7 (15%) were from Information Science and Communications, 8 (18%) Environmental Science, while 4 (9%) were from the Centres.

7.3 Awareness of Data curation

The purpose of this question was to find out from the respondents if they were aware of data curation concepts. The respondents 27 (57%) indicated that they are not aware of the data curation concept, while 20 (43%) were aware of it. However, it showed that, the respondents practiced the activities associated with data curation despite being unaware of the term "data curation" as evidenced in the responses to the subsequent question.

7.4 Awareness of Data Curation activities

When asked to identify the data curation activities which they were aware of, 13 (27%) respondents were aware of Data Sharing, 13 (27%) respondents were aware of Data Management. 11 (24%) respondents indicated that they were aware of Data Archiving, while 10 (22%) respondents indicated that they were aware of Data Preservation. The benefits of data sharing have been emphasized in a research by Borgman (2012) who found out that data sharing has been promoted for many reasons: first, without sharing data it is impossible to verify the results of research, a key principle of good science. For others it is a political issue: withholding data generated with public funds is seen as undemocratic and it would be wrong to restrict access to a public good (Arzberger et al., 2004; Murray-Rust, 2008; Vision, 2010). A less altruistic argument is made that data sharing can increase a researcher's citation rate, whether by direct citations of the data or of the associated article (Brase, 2014; Pampel et al., 2014; Piwowar and Vision, 2013).

7.5 Data Sharing

The question sought to find out if it was necessary to share data with others, majority 31 (66%) respondents indicated that it was necessary to share data with others while 16 (34%) indicated that it was not necessary to share data with others. This finding is in line with the works of Witt, (2008) that widespread sharing of data may lead to discovery and use outside of the discipline in which the data were created, fostering interdisciplinary research and learning. In some cases, preserving and sharing existing datasets could enable them to be reused instead of incurring the expense of generating new data from scratch.

7.5.1 Related Data Curation Activities Performed

The aim of the question was to solicit information from respondents on what do they do with the raw data after conducting research. Majority 19 (40%) respondents kept their raw data on PC, laptop and flash, 7 (14%) kept in envelopes, 7 (14%) destroyed, 3 (6%) archived and interpreted, 3 (6%) preserved, 2 (5%) shared with colleagues, 2 (5%) re-used, 1 (3%) stored on-line, 2 (5%) kept in cloud, 1(3%) archived. This reveals the need for designing systems for data curation. Despite attempts to keep data, there is no mechanism to locate and retrieve data sets once research for a particular study is finalized. This does not provide opportunity for data verification, sharing, re-use and leads to duplication in research.

7.5.2 The Value of Data on Future Research

An attempt was made to find out if raw data will be valuable on future research, majority 38 (80%) respondents indicated that it would be valuable while 9 (20%) respondents indicated not valuable. This finding is similar to the finding of Macdonald and Martinez-Uribe (2010) who observed that there is also an institutional and community benefit to open data in terms of costs, greater accessibility and long-term preservation of research output. Apart from the value in reproducing the original results, shared data can also be used to advance the original research or another line of inquiry. When asked to provide reasons as to why raw data will be valuable to use on future research, the respondents who indicated it will be valuable gave the following reasons:

- It could be reanalyzed to add meaning to some future research
- May be used to generate more new papers or come up with another research agenda/ programme.
- Justifying further research purposes and validation of research findings
- It is a primary source of information for future research
- Other researchers might find it useful conducting similar research interests
- Related research activities may be conducted
- For comparison with new/future research and new developments
- To make comparison of changes over time e.g. before and after or with and without
- Research in the same field will find it valuable and will not spend much time going to the field to collect data
- Research of similar nature may use same raw data as primary data or for comparison with newly acquired data. This reduces time and or increases reliability of the data collected at both occasions
- Other research might find data valuable to avoid duplication of effort

7.6 Data Curation Challenges

The major challenges identified by the respondents were categorized. The findings revealed that lack of storage facilities, equipment for raw data was topping the list of the challenges, and this is followed by:

- Safety of raw data
- Lack of knowledge on how to store and archive data electronically so as to be used in future
- Lack data bank
- There is no proper way of preserving it at a central location for access
- Lack of knowledge of period of storing such data and where
- Lack of necessary technology to handle raw data
- Lack of space for keeping such data
- Poor storage facilities because my raw data is kept in a paper folder results in exposing it to weather conditions that may cause damage
- Lack of sharing of raw data among researchers at Mzuni
- Lack of right to access and use of achieved data collected through collaborative research
- Failure to share data by colleagues who have done slightly similar research
- No reliable ICT facilities to keep raw data at Mzuni
- Lack of repository centre where such raw data can be deposited for future use

The above findings corroborate the work of Macdonald and Martinez-Uribe (2010) who identified some challenges that researchers' experience inadequate storage capacity issues, confidentiality concerns, intellectual property rights and complexity are some of the stumbling blocks to open sharing. Brandt (2007) found that researchers are unsure of how or whether to share their data, lack time to organize datasets, need help describing data so they can be found and used, want new ways of managing data, and require assistance in archiving datasets.

7.6.1 General Comments on the study

This section intended to gather comments regarding this study. From the overall comments provided below, majority of respondents are of the view that there is need for curation training and orientation. This is in line with work of Horwood et al., (2004) that the new role for librarians is a natural continuation of already existing roles of acquiring, organizing, and making available resources needed by academic staff and students. The data librarian would be responsible for advising researchers on where to locate resources, how to manage the data, and how to gain access to resources or troubleshoot data-related problems. From literature it has been observed that librarians have a major role of training faculty on data preservation, data management, data archiving and data curation. Below are some of the comments from respondents.

- It is a good study it should however give information on how raw data can be protected especially when the research outputs have not been published
- It is a strange concept for me hence need for training
- Useful study that if academic staff take this seriously there will be no duplication of efforts
- There is need to have an online repository centre where raw data can be deposited such that the same is available to subsequent users.
- Need for data curation training at mzuni
- Important study which could help identify areas for improving access to archived data
- Creation of data bank
- Need for data curation training
- Please orient us on the concepts of data curation

8. Conclusion

The study revealed unique issues to do with the management of data, data sharing and reuse and challenges. For the fact that data curation is a new concept with a lot of concerns worldwide, there is need for more research in the field to iron out gray areas on the processes of the concept. Researchers should be endowed with knowledge about the benefits of data curation such as data preservation, data archiving, data reuse which prevents duplication.

9. Recommendations

As a way forward the study recommends that the library should

- Invest part of its budget into the development to ICT facilities to preserve and archive data at Mzuni;
- Establish data bank, repository;
- Impart knowledge to researchers through workshops, training on how to store and archive data electronically so as to be used in future;
- Create space for keeping such raw data available in a paper format
- Encourage researchers to share data sets
- Improve ICT facilities to preserve and archive data sets at Mzuni

References

Arzberger, P., Schroeder, P., Beaulieu, A., Bowker, G., Casey, K., Laaksonen, L. and Moorman, D. (2004), "Promoting access to public research data for scientific, economic, and social development", Data Science Journal, Vol. 3 , November, pp. 135-152.

Ball, A. (2012), "Review of Data Management Lifecycle Models", available at: http://opus.bath.ac.uk/ 28587/ (accessed 17 December 2015).

Borgman,C.L.(2012), "The conundrum of sharing research data" ,Journal of the American Society for Information Science and Technology, Vol. 63 No. 6, pp. 1059-1078.

Brandt, D.S. (2007), "Librarians as partners in e-research", College & Research Library News, Vol. 68 No. 6, available at: http://crln.acrl.org/content/68/6/365.full.pdfþhtml.

Brase, J. (2014), "Making data citeable: datacite", in Bartling, S. and Friesike, S. (Eds), Opening Science: The Evolving Guide on how the Internet is Changing Research, Collaboration and Scholarly Publishing, Vol. 445, Springer International Publishing, Cham, pp. 327-329.

Hockx-Yu, H. (2006). "Digital preservation in the context of institutional repositories", Program: electronic library and information systems, Vol. 40 No. 3, pp. 232-43.

Horwood, L., Sullivan, S., Young, E. and Garner, J. (2004), "OAI compliant institutional repositories and the role of library staff", Library Management, Vol. 25 Nos 4/5, pp. 170-176.

Macdonald, S. and Martinez-Uribe, L. (2008). "A new role for academic librarians: data curation", El Profesional de la Informacio ´n, Vol. 17 No. 3, pp. 273-80, (original in Spanish).

Murray-Rust, P. (2008). *Open data in science*, Serials Review, Vol. 34 No. 1, pp. 52-64.

National Science Foundation (2011). "Digital Research Data Sharing and Management", No. NSB-11-79, available at: www.nsf.gov/nsb/publications/2011/nsb1124.pdf (accessed 17 December 2012).

Pampel, H. and Dallmeier-Tiessen, S. (2014), "Open research data: from vision to practice", in Bartling, S. and Friesike, S. (Eds), Opening Science: The Evolving Guide on how the Internet is Changing Research, Collaboration and Scholarly Publishing, Springer, Cham, pp. 213-224.

Piwowar, H.A. and Vision, T.J. (2013), "Data reuse and the open data citation advantage", PeerJ, Vol. 1, pp. e175, doi:10.7717/peerj.175.

Pryor,G.(2012)."Why manage research data?", In Pryor,G.(Ed.),Managing Research Data, Facet Publishing, London, pp. 1-16.

Vision, T.J. (2010), "Open data and the social contract of scientific publishing", Bioscience, Vol. 60 No. 5, pp. 330-331.

Witt, M. (2008). "Institutional repositories and research data curation in a distributed environment", Library Trends, Vol. 57 No. 2, pp. 191-201.