Management of Research Data in ETD Systems

By

Felix N Ubogu[1]
Yasien Sayed
University of the Witwatersrand, Johannesburg
Private Bag X1
Wits 2050
South Africa
Felix.Ubogu@wits.ac.za
Yasien.Sayed@wits.ac.za

Abstract

Purpose
To establish the research data management practices in institutions which have ETD programmes. These may include the systems in place (software and hardware), and the integration of data and text.

Design/methodology/approach
Along with a review of the literature the authors carried out a brief survey of existing practices in the management of the raw data on which a thesis/dissertation is based. A questionnaire was sent to the members of the Networked Digital Library of Theses and Dissertations (NDLTD) and a number of other institutions.

Findings
Sixty percent of the respondents indicated that their institutions have research data management centres/services. Only in one institution is there a relationship between the ETD programme and the data management centre/service. In this instance, ETDs were previously housed on the Academic Computing Services servers. Most of the institutions have no policy on the stewardship of raw data of a thesis/dissertation, and raw data is only stored if provided by the author along with the full text. Other forms of data pose challenges to institutions, and they are investigating options for including other data such as artwork, datasets and recordings, that form part of the submission.

Practical implications The paper will be a source of information to institutions considering embarking on the management of research data, especially that research data on which a thesis/dissertation is based.

Originality/value
The work draws attention to the need to collect research data of long term value. Such

[1] Felix Ubogu is the University Librarian of the University of the Witwatersrand, Johannesburg and Dr Yasien Sayed is Senior Researcher, Protein Structure-Function Research Unit, School of Molecular and Cell Biology, University of the Witwatersrand.

data may for example be utilised for the transfer of knowledge and skills by using postgraduate students to train less experienced students, using the archived/stored data without duplicating costly experiments, or for multiple projects based on the same or expanded datasets.

Keywords
Data Management, Research data, Electronic Thesis and Dissertation

1.0 Background

1.1 The University of the Witwatersrand (Wits), Johannesburg

Spread over more than 400 hectares, Wits University is an urban, comprehensive university which has a distinctive capacity to contribute to the reconstruction and development of South Africa through research and the production of skilled, critical and adaptable graduates. The University is structured into five Faculties (Commerce, Law & Management, Engineering and Built Environment, Health Sciences, Humanities, and Science) comprising 37 Schools, and the student population is about 25,000. The University library system comprises two main libraries and 14 divisional libraries. Students have access to 1,138, 000 book volumes, 45,000 journal titles and a host of electronic resources.

Wits is a research-intensive University, committed to providing quality training of postgraduate students, as one means of ensuring a continuing supply of active and motivated researchers. The Electronic Theses and Dissertations (ETD) project of the University is coordinated by the University Library in conjunction with stakeholders including the University Central Records Office.

As part of the ongoing initiative to understand the research support needs of researchers, the Library's Education and Training section invited Dr. Yasien Sayed, of the School of Molecular and Cell Biology (MCB), to give a presentation on his research and “How the Library can Best Serve the Academic Research Community at Wits”. This presentation generated lively discussion and the Library identified actions which would be taken to remove some of the frustrations of the academics. The management of research data was identified as an area that needed attention.

In a formal request to the University Librarian, Dr Sayed subsequently wrote:

> The Protein Structure-Function Research Unit located in the School of Molecular and Cell Biology (MCB) of the Faculty of Science currently has about 17 post graduate students (BSc (Hons), MSc, PhD and post doctoral fellows).
>
> It has become increasingly clear over the past few years that we have a problem with regard to storage of data in our research unit. The data generated by all the post graduates in the lab have been stored, traditionally, in lab books. After a few years, however, some of these hard copy documents are lost or misplaced and as a result we have lost some very important data. Subsequently, we have impressed on our students the importance of saving

> data to a CD and computers in the lab. However, this has also proven not to be very effective because a recent MSc student had had her laptop stolen which had all her raw and refined experimental data. Unfortunately, she had not saved the data on any other PC. The raw data is now gone forever. The cost (both in terms of time and money) of redoing the experiments is not always a feasible option because some data is collected at other institutions, locally and internationally. Until recently, we (MCB) had a server called Biome, but this server has since been dismantled.
>
> In light of the above, I request that the University library with the help of CNS (Central Network Services) assist us in providing a data storage facility. In this way, raw data may be stored and archived for future reference and safe keeping. It may also be utilised for transferral of knowledge and skills by using our current post graduate students to train the less experienced students using the archived/stored data without duplicating costly experiments. One can imagine the wealth of data that is generated from experimental data by 17 students and the impact this would have on the functioning of a lab if this data were lost not only now but in the future. The storage facility may also be used to archive research articles (e.g., pdf format). Some articles are very difficult to obtain, especially the "older" papers, and such a facility may aid having access to such information very quickly. Our research unit, for example, would be responsible for depositing the papers relevant to our research.
>
> These are just a few ideas, from a researcher, that I think should be considered. This would certainly aid academics and researchers and help us to realise the University's vision of becoming one of the top 100 Research University's in the world by 2010.
>
> I trust that the University will give this matter serious consideration and take the necessary steps to assist researchers in this regard (Sayed, 2008).

In order to meet the need expressed above, the Faculty of Science has acquired a server for data storage. This request has revealed that the University needs to look at requirements for the storage of raw data/datasets on which a thesis/dissertation is based.

2.0 The Data Landscape

> *"Today's research community must also assume responsibility for building a robust data and information infrastructure for the future."* (ICSU, 2004 p.7).

In defining "data" and "information" the Assessment Panel of the Committee on Scientific Planning and Review (CSPR), in its *ICSU Report of the CSPR Assessment Panel on Scientific Data and Information*, considered data and information as a continuum ranging from raw research data through to published papers. To the Panel,

> "Data" includes, at a minimum: digital observations, scientific monitoring, data from sensors, metadata, model output and scenarios, qualitative or observed behavioral data, visualizations, and statistical data collected for administrative or commercial purposes. Data are generally viewed as input to the research process.
>
> "Information" generally refers to conclusions obtained from analysis of data and the results of research. But the distinction between them is flexible and will vary according to the situation. Increasingly, the output of research (traditionally viewed as "information") includes data and has become input to other research, rendering the output-input distinction between data and information meaningless (ICSU, 2004 p. 14).

In this paper, thesis and dissertation are considered as information derived from analysis of data and results of research.

Increasingly, users wish to gain access to digital research data alongside publications (information), and to be able to link the two, in order to assess the evidence on which reported results are based. Such linking also facilitates the granting of credit to data creators (Research Information Network, 2008 p. 11).

The report by the Research Information Network (RIN) in the UK characterises research data in terms of

- the purposes and processes through which it was created
    - scientific experiments
    - models or simulations
    - observations of specific phenomena at a specific time and location
    - derived data
    - canonical or reference data
- the different groups of people and organizations who created it
    - the research community itself in the course of research
    - a variety of bodies in the public, private and voluntary sectors
- the reasons for which it has been collected together
    - for the benefit of those engaged in a specific project, where some or all of the data may or may not retain a value beyond the life of the project
    - for the benefit of a wider group within a discipline, or across disciplines, to provide reference information, or a basis for evidence-based policy-making.

The report observes that not all the data generated or collected in the course of research are of value, and guidelines are suggested for the selection of those data which should be made accessible to others and, where appropriate, preserved for the long term.

Data and information are essential building blocks of knowledge. Advances in information and communications technologies (ICT) have heightened interest in data and information production, management, and dissemination as well as access to and reuse of research data.

The United States Committee on National Statistics lists some of the benefits to sharing research data, including

- reinforcement of open scientific inquiry;
- verification, refutation, or refinement of original results;
- promotion of new research through existing data;
- encouraging more appropriate use of empirical data in policy formulation and evaluation;
- improvements of measurement and data collection methods;

- development of theoretical knowledge and knowledge of analytic technique;
- encouragement of multiple perspectives;
- provision of resources for training in research;
- protection against faulty data; and
- science would be more efficiently advanced and more effectively applied to making decisions (Fienberg, Martin and Straf, 1985).

International and national institutions across the world are developing and putting in place policies to guide the deposit and access to data and any publications arising from research projects, especially those funded from public or philanthropic funds. The *ICSU Report* made 58 recommendations, including:

- Recommendation 6: The scientific community, through ICSU national and union members, should seek to persuade governments and private sector data providers that research data produced commercially or through public-private partnerships should be made available for free or for the cost of reproduction for purposes of research and education.
- Recommendation 13: Journal publishers should encourage authors to make the data for their articles available in electronic repositories that are stable, widely accessible and professionally managed.
- Recommendation 14: ICSU should work with its members and relevant bodies in encouraging the coordinated development of digital libraries and their integration with journal publishing and data systems.
- Recommendation 16: The panel recommends that ICSU play a major role in promoting professional data management and that it foster greater attention to consistency, quality, permanent preservation of the scientific data record, and the use of common data management standards throughout the global scientific community.
- Recommendation 17: Recognizing that scientific data and information management is undergoing rapid innovation and change, information technology specialists, librarians, research scientists, government data producers, donors, and others should be involved in a concerted effort to develop standards and curricula for professional training for scientific data managers.
- Recommendation 18: Financial support for data and information management should become a routine component in all research budgets and the evaluation criteria for assessing research funding proposals should include evaluation of data management.
- Recommendation 19: All scientists should receive training in data management as part of their graduate and postgraduate education. ICSU should encourage the development of guidelines for data management by working scientists and their institutions.
- Recommendation 20: Scientists should be recognized and given credit for the scientific contribution of the data sets that they produce as well as for the analysis of those data.
- Recommendation 21: ICSU, its members and associated bodies should raise awareness of the increasingly important role that institutional repositories play in relation to scientific information management and preservation and the need to ensure that such repositories are properly resourced, developed and maintained (ICSU, 2004).

These recommendations seem to be guiding, or are in accord with, developments of research data management in more developed countries. Many of these countries are developing national research infrastructure (systems and services) that enable data to be managed and secured. Such infrastructure has been referred to variously: “cyber-infrastructure” is used in the United States, “e-infrastructure” in the United Kingdom and

Europe, “GRID” in Canada (Anne Fitzgerald and Kylie Pappalardo, 2007), and “e-Research” in Australia.

In Australia, the Prime Minister’s Science, Engineering and Innovation Council (PMSEIC), Working Group on Data for Science in its report, *From Data to Wisdom: Pathways to Successful Data Management for Australian Science* (2006), recommended that:

> Data management expertise becomes a core skill for researchers, including graduate and postgraduate science students across all disciplines, and that they receive data management training as part of their education (Recommendation 10: p.13).

The *Report to the National Science Foundation from the ARL Workshop on New Collaborative Relationships* observes that “digital data stewardship is fundamental to the future of scientific and engineering research and the education enterprise, and hence to innovation and competitiveness. Overall, it is clear that ecology of institutional arrangements among individuals and organizations, sharing an infrastructure, will be required to address the particularities of heterogeneous digital data and diverse scholarly and professional cultures”.

Hey and Hey (2006) have observed that:

> Increasingly academics will need to collaborate in multidisciplinary teams distributed across several sites in order to address the next generation of scientific problems.
> In addition, new high-throughput devices, high-resolution surveys and sensor networks will result in an increase in scientific data collected by several orders of magnitude.
> To analyze, federate and mine this data will require collaboration between scientists and computer scientists; to organize, curate and preserve this data will require collaboration between scientists and librarians.

They further pointed out that “A vital part of the developing research infrastructure will be digital repositories containing both publications and data.”

Kaniki (2007) observes that raw data sets are very important to the scientific community around the world in terms of knowledge production but that little work has been done in the developing countries in this area.

There is a need for stewardship of the research data on which a thesis/dissertation is based. According to Harboe-Ree (2008) data should be managed for the following reasons:

- Supports individual or group research
- Optimises the investment
- Allows reuse and recombination
- Disseminates results to wider community
- Maintains the scholarly record
- Manages compliance, legal and financial risks
- Funding, IP, privacy, etc

- Facilitates verification of research results

3.0 Study Methodology

Along with a review of the literature the authors carried out a brief survey of existing practices in the management of the raw data on which a thesis/dissertation is based. A questionnaire was sent to the members of the Networked Digital Library of Theses and Dissertations (NDLTD) (listed at the NDLTD website on 15 April 2008, and a few other institutions.

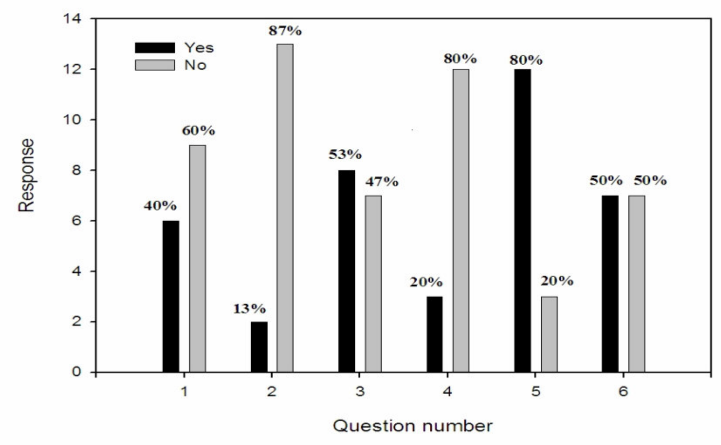
The questions covered

- Existence of institutional research centre and/or research data management centre/service.
- Relationship between institution's ETD programme and data management centre/service.
- Collection and digital storage of any raw data used in the production of a thesis/dissertation.
- Software and hardware used.
- Integration of the e-text integrated with the data.
- Inclusion of images in submitted ETDs, and management of these.
- Copyright policies that guide the use of images.
- If institution has no research and data management centre/service what plans there are to manage the raw data of theses and dissertations.

The questionnaire was sent to 65 institutions and only 15 responses (23%) have been received. Time constraints did not allow for a follow up.

4. Results of Survey

The bar chart below presents a picture of the responses to the six questions.



Question number:
1: Does your institution have a research centre and/or research data management centre/service?
2: Is there any relationship between your ETD programme and the data management centre/service?
3: Do you collect and store digitally any raw data used in the production of a thesis/dissertation?
4: Is the e-text integrated with the data?

Comment on responses

1. Existence of institutional research centre and/or research data management centre/service:

Sixty percent of the respondents indicated that their institutions have research data management centre/service.

2. Relationship between your ETD programme and the data management centre/service:

Only in one institution is there a relationship between the ETD programme and the data management centre/service. In this instance, ETDs were previously housed on the Academic Computing Services servers.

3. Collection and digital storage of any raw data used in the production of a thesis/dissertation:

Most of the institutions have no policy on the stewardship of raw data of a thesis/dissertation. In some institutions, raw data would be stored if provided by the author along with the full text. Other forms of data are posing challenges to institutions and they are looking at options for including other pieces such as artwork, datasets and recordings as part of the submission.

Software and hardware used:

The software used in the management of ETDs includes DSpace, EPrints and Oracle-based in-house systems. The EPrints software allows for supplementary files to be included in any format. The storage format is mostly PDF and MSWord.

4. Integration of the e-text integrated with the data:

Where raw data is gathered there is no integration of the e-text with the data.

5. Inclusion of images in submitted ETDs, and management of these:

In most of the institutions, images are embedded in submitted theses and dissertations.

Copyright policies that guide the use of images:

The trend is that the author is responsible for providing copyright permission and must provide written evidence. Where no evidence of permission is available, the image is removed from the digital version of the thesis but remain in the printed version.

6. If institution has no research and data management centre/service what plans there are to manage the raw data of theses and dissertations:

Some institutions are considering putting in place formal procedures for managing the raw data of theses and dissertations.

5. Observations and Recommendations

The management of research data is receiving a lot of attention in the international community. The Research Information Network in the UK suggests five principles which provide a broad framework for developing good practice for universities, research institutions, libraries and other information providers, publishers, research funders and researchers themselves. This framework includes:

- the clear definition of roles and responsibilities;
- the creation and collection of digital research data in accordance with applicable international standards;
- easy access to digital research data; protection of the rights of those who have gathered or created data; and protection of the rights of those who have legitimate interests in how data are made accessible and used;
- efficient and cost-effective models and mechanisms for managing and providing access to digital research; and
- Preservation and accessibility of digital research data of long term value for current and future generations (RIN, 2008).

The principles are similar to some of the recommendations in the ICSU Report.

Some institutions are taking bold steps to establish initiatives related to research data management. Harboe-Ree (2008) has recently indicated that the present data collection activities of Monash University in Australia are geared towards collecting the raw data for published articles. The infrastructure of their repository includes a Fedora 2.2.1 repository with VTLS software for web services (VITAL repository search and management services and VALET web submission tool) sitting on top.

The practice of having a cross-disciplinary team to deal with the implementation of data management programme is highly desirable. Such teams would include staff with expertise in data management, high-performance computing, computation, visualization, metadata specialists and staff who can advise on access and ownership issues (e.g. copyright, intellectual property and open access). Such a team would address institutional and researcher needs, as well as formulate a set of principles to guide cost modelling and sustainable funding options.

6. Conclusion

The management of the raw data on which a thesis/dissertation is based should be viewed in the broader context of the management of institutional research data. In keeping with international developments in the management of research data, it is highly desirable / essential for institutions with ETD programmes to develop the infrastructure and capacity to take advantage presented by advances in information and communications technologies.

Institutions need to establish high-level interest in the management of research data through the establishment of appropriate structures/bodies and infrastructure. Such bodies should have the mandate to establish data practices at institutional level, set policy frameworks and determine processes through which policies are implemented. The recommendations contained in the ICSU Report as well as the report of the Research Information Network provide good guidelines for implementation of research data management and the various roles of stakeholders. In conjunction with other stakeholders libraries are playing critical roles in the data management arena.

Acknowledgements

The authors are grateful to Ms Clare Walker of the University of the Witwatersrand Library for a critical reading of the manuscript.

References

Fienberg, S. E, Martin, M.E., and Miron L. Straf, Editors. *Sharing Research Data,* Committee on National Statistics, National Research Council, 1985.
Available at: http://www.nap.edu/catalog/2033.html

Fitzgerald, A. and Pappalardo, K. (2007). “Building the Infrastructure for Data Access and Reuse in Collaborative Research: An Analysis of the Legal Context”. Available at: http://eprints.qut.edu.au/archive/00008865/01/8865.pdf

Harboe-Ree, C. (2008). “eResearch support in Australia: a Monash University perspective”. Presented at the CSIR, Pretoria on 14 May 2008.

Hey, T. and Hey, J. (2006). “E-Science and its Implications for the Library Community”, pp. 1–2. Available at:
http://conference.ub.uni-bielefeld.de/2006/proceedings/heyhey_final_web.pdf

ICSU (International Council for Science). 2004. *Scientific Data and Information: Report of the CSPR Assessment Panel*. Available at: http://www.icsu.org/Gestion/img/ICSU_DOC_DOWNLOAD/551_DD_FILE_PAA_Data_and_Information.pdf (accessed 20 April 2008).

Kaniki, A. (2007). Opening address at the CODATA Workshop on Data Sources for Sustainable Development in SADC, 14-15 May 2007. p.2
Available at: http://stardata.nrf.ac.za/Codata/workshopCodataReport_2007.doc

National Science Board. *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century*. (Arlington, VA: National Science Foundation, September 2005), 19. http://www.nsf.gov/pubs/2005/nsb0540/start.htm.

Networked Digital Library of Theses and Dissertations (NDLTD). http://www.ndltd.org/

Prime Minister's Science, Engineering and Innovation Council, Working Group on Data for Science. *From Data to Wisdom: Pathways to Successful Data Management for Australian Science*, December 2006. Available at: http://www.dest.gov.au/NR/rdonlyres/D15793B2-FEB9-41EE-B7E8-C6DB2E84E8C9/15103/From_Data_to_Wisdom_Pathways_data_man_forAust_scie.pdf

Research Information Network (2008). *Stewardship of digital research data: a framework of principles and guidelines*. Available at: http://www.rin.ac.uk/files/Research%20Data%20Principles%20and%20Guidelines%20full%20version%20-%20final.pdf

Sayed, Yasien (2008). Personal communication.

To Stand the Test of Time: Long-Term Stewardship of Digital Data Sets in Science and Engineering. A report to the National Science Foundation from the ARL Workshop on New Collaborative Relationships: the role of Academic Libraries in the Digital Data Universe. Sep. 26–27, Arlington (2006). Available at: http://www.arl.org/bm~doc/digdatarpt.pdf

List of institutions that responded to survey

1. Ball State University (USA)
2. Johns Hopkins University (USA)
3. McGill University (Canada)
4. Queensland University of Technology (Australia)
5. Ohio University (USA)
6. Rhodes University (South Africa)
7. Rochester Institute of Technology (USA)

8. University of Maryland (USA)
9. University of North Texas (USA)
10. University of Pretoria (South Africa)
11. University of Tennessee (USA)
12. University of the Witwatersrand (South Africa)
13. Virginia Tech (USA)
14. West Virginia University (USA)
15. Yale University (USA)