Shepherding data on the way to long term stewardship:
Research Data Management at the University of Pretoria: a case study

By Mr Johann van Wyk (UPLS, UP); Prof Michael S. Pepper (Institute for Cellular and Molecular Medicine, UP), Mr Isak van der Walt (UPLS, UP); Ms Candice Honing (Institute for Cellular and Molecular Medicine, UP)

Presented (+ edited) by Dr Heila Pienaar and Isak van der Walt, University of Pretoria Library Services, 5 June 2014, 12thSAOIM

Wordle of the data management policy of the Office of Digital Humanities, National Endowment for the Humanities
Definitions

- **Research data**: Research data, unlike other types of information, is collected, observed, or created, for purposes of analysis to produce original research results. [http://www.ed.ac.uk/is/data-management](http://www.ed.ac.uk/is/data-management)

- **Research data management**: “is the process of controlling the information generated during a research project. Any research will require some level of data management, and funding agencies are increasingly requiring scholars to plan and execute good data management practices”. Managing data is an integral part of the research process. How data is managed depends on the types of data involved, how data is collected and stored, and how it is used - throughout the research lifecycle. [http://www.libraries.psu.edu/psul/pubcur/what_is_dm.html](http://www.libraries.psu.edu/psul/pubcur/what_is_dm.html)
Why manage research data?

By managing research data you will:

• Meet funding body grant requirements;
• Ensure research integrity and replication;
• Ensure research data and records are accurate, complete, authentic and reliable;
• Increase your research efficiency;
• Save time and resources in the long run;
• Enhance data security and minimise the risk of data loss;
• Prevent duplication of effort by enabling others to use your data;
• Comply with practices conducted in industry and commerce; and
• Protect the University from reputational, financial and legal risk.
Research Data Life Cycle

**Creating data**
- design research
- plan data management (formats, storage etc.)
- plan consent for sharing
- locate existing data
- collect data (experiment, observe, measure, simulate)
- capture and create metadata

**Processing data**
- enter data, digitise, transcribe, translate
- check, validate, clean data
- anonymise data where necessary
- describe data
- manage and store data

**Analysing data**
- interpret data
- derive data
- produce research outputs
- author publications
- prepare data for preservation

**Giving access to data**
- distribute data
- share data
- control access
- establish copyright
- promote data

**Preserving data**
- migrate data to best format
- migrate data to suitable medium
- back-up and store data
- create metadata and documentation
- archive data

**Re-using data**
- follow-up research
- new research
- undertake research reviews
- scrutinise findings
- teach and learn

Based on UK Data Archive documentation:
http://www.hsrc.ac.za/Page-156.phtml
http://www.data-archive.ac.uk/media/2894/managingsharing.pdf
http://www.data-archive.ac.uk/
Team

Dr Heila Pienaar, Deputy Director: Innovation & Technology

Johann van Wyk, Assistant Director: UP Research Data Management

Isak van der Walt: IT Specialist in the Library
Chronological Development of RDM initiatives at UP

- **2007** - Policy for the preservation and retention of research data (Rt 306/07)
- **2010** - Survey of RDM practices at UP (October 2009 – March 2010)
- **11 April 2013** – Meeting with Prof Michael Pepper – Director of Institute for Cellular and Molecular Medicine (ICMM) to discuss the possibility of a pilot project with his students
- **August-November 2013** - Interviews with Deputy Deans Research of all the Faculties to determine “Essential Research Data that the University should manage” as requested by UP Vice Principal: Research – Johann van Wyk
- **December 2013** – 2nd Pilot project – Neuro-Physiotherapy
- **17 March 2014** – High Level Report compiled on RDM internationally, nationally and at UP for UP Vice Principal Research – Johann van Wyk
UP Survey of RDM: what is seen as essential data that should be managed?

- Interviews with Deputy Deans Research of each of the Faculties
- Focus: to determine what is seen as the **essential data** of the faculty that must be managed
- Conducted **eleven interviews**, August – October 2013
- **Trends** were then identified

Wordle of the data management policy of the Office of Digital Humanities, National Endowment for the Humanities
UP Survey of RDM: recommendations

- A Research Data Management Policy for UP
- Establishment of a central research data management office
- Establishment of a research data management presence in each Faculty
- Take into account the impact of RDM on workload and time of researchers and students
- Establishment of data repository for UP
- Investigate necessary IT infrastructure that can accommodate small and big data sets, as well as high performance computing
- Determine a time frame for the roll-out of a research data management system for UP
Pilot Studies

• Current Projects:
  – Institute for Cellular & Molecular Medicine project
  – Neuro-Physiotherapy project

• We plan to expand this so that we can have a pilot project in each Faculty as a learning experience

• Open invitation for information specialists to be involved in this
Example of a doctoral student’s data

Creating data

Analysing data
- interpret data
- derive data
- produce research outputs
- author publications
- prepare data for preservation

Processing data

Giving access to data

Preserving data

Re-using data

Categories
- Articles
- CFU pictures
- Flow cytometry data
- Lab book
- PhD Thesis Final 2013
- Raw data
Pilot Projects: ICMM
Next phase

- A campus-wide database / repository for open access data sets with DOI’s
Long-term preservation

Archival information package (AIP)

• **Bagit** “bag” contains: (bag it and tag it)
  • bag declaration file, manifest file, data files
    – Metadata file (XML):
  • METS wrapper
  • Dublin Core and MODS (descriptive metadata)
  • PREMIS (preservation metadata)

• **MetaArchive**: LOCKSS replication network (7 copies)-USA example