Going Full Circle: Research Data Management @ University of Pretoria

Presentation at eResearch Africa 2014 Conference, held at University of Cape Town, Cape Town, South Africa, 23-27 November 2014

By Johann van Wyk and Isak van der Walt
Introduction

Internationally research data is increasingly recognised as a vital resource whose value needs to be preserved for future research. This places a huge responsibility on Higher Education Institutions to ensure that their research data is managed in such a manner that they are protected from substantial reputational, financial and legal risks in the future. This presentation will focus on the Research Data Lifecycle, with an overview of RDM at the University of Pretoria, and a demonstration of pilot projects implemented at the institution.
Research Data Management: A (Brave) Complex New World

Data Sharing, Messy, Complex, Open Data, Complex New World, Small Data, Various formats, Various devices, Various Versions, Copyright License, Sensitive Data, Data Archiving, DOI, Data Repository, Data政策, Data Publishing, Data Citation, DMP, Data Preservation, Linked Data, Data Journals, Visualisation, Data Provenance, Big Data, Data Anonymisation, RDM, Data Analysis.
Why Manage Research Data?

By managing research data you will:

• Meet funding body grant requirements, e.g. NSF, NIH;
• Meet publisher 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 your institution from reputational, financial and legal risk.
Research Data Lifecycle

Based on UK Data Archive Lifecycle
Research Data Lifecycle in the context of the Research Life Cycle

- Identification of Research Area
- Creating Data
- Re-using Data
- Processing Data
- Analysing Data
- Preserving Data
- Experimenting and Analysis (Empirical Research)
- Writing up results
- Dissemination of Findings
- Identification of Collaborators
- Proposal Writing
- Identification of Funding Resources
- Literature Review and Indexing
Various stakeholders in RDM

- Executive Management
- Deans & Dept Heads
- IT Services
- Research Office
- Library
- Principal Investigator/Researcher
- Funders
- Government
- Publishers
- External (disciplinary) data repositories

(De Waard, Rotman and Lauruhn, 2014)
Chronological Development of RDM at University of Pretoria

- **2007** - Policy for the preservation and retention of research data
- **2010** - Survey of RDM practices at UP (October 2009 – March 2010)
- **April 2013** – Meeting - Director of Institute for Cellular and Molecular Medicine (ICMM) on possibility of pilot project with students, which was subsequently implemented
- **August-November 2013** - Interviews with Deputy Deans Research Faculties to determine the “Essential Research Data that the University should manage”
- **December 2013** – 2nd Pilot project – Neuro-Physiotherapy
Chronological Development of RDM at University of Pretoria

- **April 2014** – Visit by Deputy Director Innovation and Technology and Library IT Specialist (UP Library Services) to Purdue University in USA: investigate Purdue’s Research Data Repository (PURR) and long-term preservation processes as possibility for replication at UP

- **June 2014** - Assistant Director RDM attended CODATA International Training Workshop in Big Data for Science for Researchers from Emerging and Developing Countries, in Beijing, China.
Chronological Development of RDM at University of Pretoria

- **July 2014** - High Level Report on RDM sent through to University Executive for review
- **August 2014** – Proposed new University policy on RDM sent through to University Executive for review
- **Jan 2015** – Task Team to investigate infrastructure needed for RDM across the University
Survey of RDM practices at University of Pretoria, October 2009 – March 2010

- 52 interviews conducted by 15 information specialists from relevant Faculty Libraries
- At least 3 lecturers and one Postgraduate student from each Faculty were interviewed
- The information specialists received formal training in interview techniques.
- Interviews were conducted according to a semi-structured interview framework.
Findings of Survey - October 2009 – March 2010

• Funding: In most cases no need for data management or data sharing plans, depending on funding agency requirements

• Data Collection: Wide variety of data collection methods used. Both primary and secondary data. Data sets are often small.

• Data Storage: Ad hoc storage of data, both on paper and electronically

• Publishing: In general raw data not published

• Support: Lack of support with regard to storage of data (physical and electronic).

• RDM does not exist in any formal manner (with the exception of one or two departments) at the University of Pretoria
Recommendations: Survey - Oct 2009 – Mar 2010

• Investigate Very Large Database initiative from the Department of Science and Technology as possibility to support UP’s RDM needs

• Central UP server or repository

• Address the need for physical storage space.

• Create a formal staff position of ‘research data manager’ to drive RDM at UP
UP Survey of RDM at UP – August – October 2013

• Interviews with Deputy Deans Research of each of the Faculties

• Focus: 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
Trends

- **Essential Data**
  
  Interview data, Questionnaires, Spread sheet data, Lab books, Experiment /laboratory data, Images (e.g. graphs, models, Sketches, X-Rays, scans etc.), Literature reviews, Sequencing data, Computer-generated data.

- **Level of Data**

  Some see only raw data as essential, some see processed /analysed data as essential, while others see both raw and processed/analysed data as essential

- **Volume of data**

  A small number work only with small data sets.

  Majority work with small and big data sets, with exponential increase in big data sets
Trends

- **Data Formats used**
  Excel, Pdf, MS Word, Text Format, images in various formats, video, sound, various computer generated formats, SPSS, SAS, AMOS, Qualtrics data, SurveyMonkey data, simulation data formats, and even data from social media. Some in paper format.

- **RDM Plans**
  None have Research Data Management Plans in place.

- **Uploading capacity**
  No capacity to upload these data sets to a repository

- **Willingness to share data**
  Majority willing to share their data under certain conditions.
  Health Sciences not willing to share their data
Recommendations

• A new Research Data Management Policy for UP
• Establishment of a central research data management office
• Establishment of a RDM presence in each Faculty
• Consider impact of RDM on workload and time of researchers and students
• Establishment of data repository for UP
• Investigate necessary IT infrastructure for RDM – (handle small and big data sets, and HPC)
• Determine a time frame for the roll-out of a RDM system for UP
Pilot Projects at University of Pretoria

- Two data management pilot projects in 2013-2014:
  Institute for Cellular and Molecular Medicine (ICMM) and the Neuro-Physio-Group.
- Currently also implementing more pilot projects: Potato Pathology Programme, Powdery Scab, and Psychiatry Dissociation
- An Open Source Document Management System Alfresco was customised for this purpose

**Why Alfresco?**
Open Source
Captured provenance of data
Has a versioning function
Good metadata function
Easy to integrate with other software
Workflow function gave supervisor overview of progress of students
Sync function with Dropbox and Google Drive
Drag and Drop function
File Sharing function
Mobile App
Process followed to implement Pilot Projects

Step 1 - Determine Researchers’ Needs

Step 2 - Select Software

Step 3 - Customise System and demonstrate to researchers

Step 4 - Train Researchers & Subject Librarian/Information Specialist

Step 5 - Identify Champion

Continuous Software Evaluation

- Re-select software where necessary
- Monitor continuously
- Have regular meetings with Champions
- Make adjustments where necessary
- Evaluate system

Re-select software where necessary
Research in Process part of RDM
The Next steps in our RDM pilot studies
Overview

1. ECM Approach
2. Supporting the Research Data Lifecycle
3. Research in Process
4. Dissemination
5. Preservation
6. Further Development and Hurdles
“Enterprise Content Management (ECM) is a formalized means of organizing and storing an organization's documents, and other content, that relate to the organization's processes. The term encompasses strategies, methods, and tools used throughout the lifecycle of the content”

*(What Is Enterprise Content Management (ECM)?*. AllM. Association for Information and Image Management)

To Note:

- Enterprise content management is not a closed-system solution or a distinct product category.
- Focus should be focused on your environment.
Supporting the Research Data Lifecycle

- Creating Data
- Processing Data
- Analyzing Data
- Preserving Data
- Re-using Data
- Giving Access to Data
Supporting the Research Cycle Cont.

BagIt - Specification

Alfresco CMS

Dspace (UPSpace)

Preservation

Dissemination

Research in Process
NASA Langley Research Center

The Challenge
The NASA Langley Research Center conducts hundreds of tests each year designed to make aircraft and spacecraft safer and more efficient. These tests – which are performed by different teams of engineers, researchers, technicians, managers and customers – are highly collaborative, as teams need to be able to share ideas and view each other’s test documentation.

NASA built a homegrown collaborative portal, aeroCOMPASS, which allowed Langley to create individual team sites for sharing and commenting on documents, notes and other research files. But after 10 years of use, the aeroCOMPASS software had become outdated and could no longer meet NASA’s strict security guidelines.

With over **800 sites in use**, NASA needed to move aeroCOMPASS to a new collaboration and document management environment with a similar look and feel, but using a **more secure, modern architecture**.

“We have users coming into the system from all over including researchers, Lockheed Martin engineers, and researchers wanting to build teams. When they come back to the system, all of their documents and research is there waiting for them.” — David Cordner, IT Architect, Research Directorate
Alfresco – Case Study 2

KLM

A key objective of the project was to keep document management simple for users, while meeting the company’s technical and budget requirements. Solutions evaluated by KLM needed to provide users with the following framework:

- **Personal documents** – a personal document management space accessible to all employees on the Web, anywhere at any time, with Web Attached Secure Storage Anywhere (WASSA);

- **Project/team documents** – a place for project, departmental or team documents to be shared, stored and collaborated on;

- **Company shared documents** – a company-wide document repository that included search capabilities without a complex taxonomy; and

- **Operational documents** – a structured repository with restricted authorization and only the last version of operational documents that included taxonomy.

“Alfresco was easy to implement on our standard infrastructure and was rolled out to all 30,000 employees within six months. Any preconceived ideas we had about working with an open source vendor were quickly dismissed. In terms of technical support, working with Alfresco was just like engaging in any commercial service level agreement, but with the added benefits of the open source architecture and cost structure.”

Pieter Janssen, Chief Architect at KLM
DSpace – Dissemination
Preservation – Bagit Specification
What is BagIt?

- "BagIt is a hierarchical file packaging format designed to support disk-based storage and network transfer of arbitrary digital content."

- "A "bag" consists of a "payload" (the arbitrary content) and "tags", which are metadata files intended to document the storage and transfer of the bag."

- "They are also well-suited to the export, for archival purposes, of content normally kept in database structures that receiving parties are unlikely to support."
Contents of a Bag

1. Data Directory
   - Contains the data payload
   - Can be single or multiple files and directories

2. Manifest File
   - Text file with listed items in the payload with their checksums

3. BagIt File
   - Contains info on BagIt version and encoding

4. BagIt Info File
   - Contains the Metadata for the bag

5. Tag Manifest File
   - Contains checksums to verify the above mentioned txt files
Example
Example
Example
<table>
<thead>
<tr>
<th>Name</th>
<th>Date modified</th>
<th>Type</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>EasyNetMonitor</td>
<td>11/5/2013 5:25 AM</td>
<td>File folder</td>
<td></td>
</tr>
<tr>
<td>recover</td>
<td>11/21/2012 1:34 PM</td>
<td>File folder</td>
<td></td>
</tr>
<tr>
<td>isak11072014.zip</td>
<td>7/16/2014 8:34 AM</td>
<td>Compressed (zipped) Folder</td>
<td>5,661 KB</td>
</tr>
<tr>
<td>Default.rdp</td>
<td>12/20/2012 8:09 AM</td>
<td>Remote Desktop Connection</td>
<td>0 KB</td>
</tr>
<tr>
<td>pgadmin.log</td>
<td>4/11/2013 10:15 AM</td>
<td>Text Document</td>
<td>2 KB</td>
</tr>
<tr>
<td>data</td>
<td></td>
<td>File folder</td>
<td></td>
</tr>
<tr>
<td>bag-info.txt</td>
<td></td>
<td>Text Document</td>
<td>1 KB  No</td>
</tr>
<tr>
<td>bagit.txt</td>
<td></td>
<td>Text Document</td>
<td>1 KB  No</td>
</tr>
<tr>
<td>manifest-md5.txt</td>
<td></td>
<td>Text Document</td>
<td>1 KB  No</td>
</tr>
<tr>
<td>taumvmanifest-md5.txt</td>
<td></td>
<td>Text Document</td>
<td>1 KB  No</td>
</tr>
</tbody>
</table>
Metadata Captured During The Data Lifecycle

Data Retention

Descriptive Metadata

Structural Metadata

Research In Process - Phase

Administrative Metadata

Dissemination Phase

Preservation Phase
Descriptive Metadata describes a resource for purposes such as discovery and identification. It can include elements such as title, abstract, author, and keywords.

Structural Metadata describes the physical and/or logical structure of digital resources.

Administrative Metadata often captures the context necessary to understand information resources, such as creation or acquisition of the data, rights management, and disposition.
Levels of digital preservation

<table>
<thead>
<tr>
<th>Level</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Protect Your Data</td>
</tr>
<tr>
<td>2</td>
<td>Know Your Data</td>
</tr>
<tr>
<td>3</td>
<td>Monitor Your Data</td>
</tr>
<tr>
<td>4</td>
<td>Repair Your Data</td>
</tr>
</tbody>
</table>

**Table 1: Version 1 of the Levels of Digital Preservation**

<table>
<thead>
<tr>
<th>Level 1 (Protect your data)</th>
<th>Level 2 (Know your data)</th>
<th>Level 3 (Monitor your data)</th>
<th>Level 4 (Repair your data)</th>
</tr>
</thead>
</table>
| Storage and Geographic Location | - Two complete copies that are not collocated  
- For data on heterogeneous media (optical discs, hard drives, etc.) get the content off the medium and into your storage system | - At least three complete copies  
- At least one copy in a different geographic location  
- Document your storage system(s) and storage media and what you need to use them | - At least three copies in geographic locations with different disaster threats  
- Have a comprehensive plan in place that will keep files and metadata on currently accessible media or systems |
| File Fixity and Data Integrity | - Check fixity on ingest if it has been provided with the content  
- Create fixity info if it wasn’t provided with the content | - Check fixity on all ingests  
- Use write-blockers when working with original media  
- Virus-check high risk content | - Check fixity of content at fixed intervals  
- Maintain logs of fixity info; supply audit on demand  
- Ability to detect corrupt data  
- Virus-check all content |
| Information Security | - Identify who has read, write, move and delete authorization to individual files  
- Restrict who has those authorizations to individual files | - Document access restrictions for content  
- Maintain logs of who performed what actions on files, including deletions and preservation actions | - Perform audit of logs |
| Metadata | - Inventory of content and its storage location  
- Ensure backup and non-collocation of inventory | - Store administrative metadata  
- Store transformative metadata and log events | - Store standard technical and descriptive metadata |
| File Formats | - When you can give input into the creation of digital files encourage use of a limited set of known open formats and codecs | - Inventory of file formats in use  
- Monitor file format obsolescence issues | - Perform format migrations, emulation and similar activities as needed |
Further Development

Identify a Campus-wide database /repository for the publishing of open access data sets

e.g. Dspace, Fedora, Purr
Further Development

• Investigation and development of Data Publishing Platform
  • Fedora (DuraSpace Organization)

• Creation of DMP (data management planning) tool

• Automatization of certain processes

Hurdles

• IT infrastructure support

• Understanding of the concept and processes
References


References


• **Table 1: Version of the levels of preservation** NDSA Levels of Preservation. [Online] available at http://www.digitalpreservation.gov/ndsa/activities/levels.html (Accessed 19 November 2014)
Thank You