Suggesting an Institutional Data Repository (IDR) for UCT

eResearch Africa Conference

4 May 2017

Kayleigh Roos & Erika Mias
Digital Curation Officers
Suggesting an IDR for UCT

Overview

- Introduction
- Scope & mission
- Precedents
- Considerations
- Potential solutions
- Eliminations & Evaluation
- Scenario planning
- Recommendations
- Current solutions

http://datablog.is.ed.ac.uk/2013/12/10/new-research-data-storage/
Introduction

Institutional repositories
Drivers

Suggesting an IDR for UCT
4 May 2017
Different types of repositories

Repositories preserve, manage, and provide access to many types of digital materials in a variety of formats. Materials in online repositories are curated to enable search, discovery, and reuse. There must be sufficient control for the digital material to be authentic, reliable, accessible and usable on a continuing basis.

- Different types of repositories:
  - Institutional repositories
    - OpenUCT (scholarly outputs)
    - UCT Digital Collections (digital projects)
  - Discipline-specific repositories
    - Ocean Biogeographic Information System (OBIS)
    - National Institutes of Health (NIH)
  - Government
  - Community
Drivers for OA repositories

• **Bottom-up:**
  - Data deluge
  - Publicity
  - Re-use
  - Validation

• **Top-down:**
  - Community and journal standards
  - Funding agency mandates
  - Innovation regulations
Local context and drivers

- **NRF open access statement:**

  “From 01 March 2015, authors of research papers generated from research either fully or partially funded by NRF, when submitting and publishing in academic journals, should deposit their final peer-reviewed manuscripts that have been accepted by the journals, to the administering Institution Repository with an embargo period of no more than 12 months.

  ....

  In addition, the data supporting the publication should be deposited in an accredited Open Access repository, with the provision of a Digital Object Identifier for future citation and referencing.”
Institutional Research Data Repository (IDR) for UCT

UNIVERSITY OF CAPE TOWN

Report was created from the perspective of UCT eResearch, based on projected, multi-disciplinary requirements for an IDR at UCT

- UCT-wide
- Data that directly supports published research outputs (scholarly publications)
- All data-types across all research disciplines
- Focusing on the ‘long-tail’
The UCT Institutional Data Repository (IDR) will serve principal investigators at the University of Cape Town in need of a repository to preserve and openly disseminate data that support their published research findings. The repository service is provided in terms of the UCT Research Data Management Policy (in draft). It will preserve curated datasets and make them accessible through persistent identifiers linking to their respective scholarly publications accessible on OpenUCT or elsewhere. UCT’s IDR is intended to become an accredited trusted digital repository (TDR).

The UCT IDR aims to accommodate a breadth of data types across all research disciplines. It will focus primarily on supporting all researchers without access to an existing domain-specific repository. Very large datasets associated with published work will be supported where possible when they cannot be accommodated elsewhere.

The IDR will provide open access to research data files, making them available for download in standard formats. In order to ensure the findability, accessibility, interoperability and reusability of the data, high quality documentation and metadata will be required at the time of collection. Abiding by the institutional guidelines, quality standards and curation practises of local discipline communities at UCT will also be necessary for publication in the IDR.
Considerations

Essential capabilities of software and needs of research community
Considerations

Potential software solutions were to be evaluated based on several criteria:

1. Storage
   - on-site / cloud
   - cost recovery

2. Operation & maintenance
   - installation, customisation and upgrades
   - community support

3. Publication workflow
   - curation layers
   - user roles
   - privacy and embargo
   - validation and feedback procedures

4. Dissemination and sharing
   - harvesting (DOIs) and linking data (OpenUCT)
   - OA licensing
   - integration capabilities (OSF)
Potential software solutions were to be evaluated based on several criteria:

5. Reporting
   ● usage statistics, altmetrics
   ● citation, ORCID

6. Archiving and preservation
   ● software integration
   ● versioning
   ● hardware upgrades, back-up, migration

7. Ingest
   ● UI design
   ● upload limits
   ● enforced metadata

8. Visualisation and analysis
   ● processing of file types from within UI
   ● interactivity
   ● tables, graphs, geographic analysis
Potential software solutions
Suggesting an IDR for UCT

Precedents - who’s using what?

- Investigation of software used at leading research Universities
- Institutional data repositories registered on Re3Data

<table>
<thead>
<tr>
<th>Institution</th>
<th>Research data repository</th>
<th>Research publication repository</th>
</tr>
</thead>
<tbody>
<tr>
<td>University of British Columbia</td>
<td>Datacure(Abacus)</td>
<td>Dspace(cifRole)</td>
</tr>
<tr>
<td>Edinburgh University</td>
<td>Dspace(DataShare)</td>
<td>Dspace(ERA)</td>
</tr>
<tr>
<td>Harvard University</td>
<td>Datacure</td>
<td>Dspace(DASH)</td>
</tr>
<tr>
<td>Oxford University</td>
<td>Fedora(CRA-Data)</td>
<td>Fedora(CRA-Data)</td>
</tr>
<tr>
<td>University of Essex</td>
<td>Eprints</td>
<td>Eprints</td>
</tr>
<tr>
<td>University of Liverpool</td>
<td>Eprints</td>
<td>Repository under construction</td>
</tr>
<tr>
<td>University of Munich</td>
<td>Eprints</td>
<td>Eprints</td>
</tr>
<tr>
<td>MIT</td>
<td>Dspace or Datacure</td>
<td>Dspace</td>
</tr>
<tr>
<td>TU Berlin</td>
<td>Dspace</td>
<td>Dspace</td>
</tr>
<tr>
<td>University of Wisconsin</td>
<td>Dspace(MINDS)</td>
<td>Dspace(MINDS)</td>
</tr>
<tr>
<td>Cambridge University</td>
<td>Dspace</td>
<td>Dspace</td>
</tr>
<tr>
<td>University of Pretoria</td>
<td>Dspace</td>
<td>Dspace</td>
</tr>
<tr>
<td>University of Toronto</td>
<td>Dspace (Ontario Council of University Libraries)</td>
<td>Dspace(Tspace)</td>
</tr>
<tr>
<td>York University</td>
<td>Dspace(Yorkspace)</td>
<td>Dspace(Workplace)</td>
</tr>
<tr>
<td>Johns Hopkins</td>
<td>Datacure</td>
<td>Dspace</td>
</tr>
<tr>
<td>University of Queensland</td>
<td>Fedora(UQ eSpace)</td>
<td>Fedora(UQ eSpace)</td>
</tr>
<tr>
<td>University of Copenhagen</td>
<td>Dspace(DataBox)</td>
<td>Subject specific only</td>
</tr>
<tr>
<td>Peking University</td>
<td>Datacure</td>
<td>Eprints</td>
</tr>
<tr>
<td>University of Western Australia</td>
<td>DSpace(RDO)</td>
<td>DSpace(UWA research data online)</td>
</tr>
<tr>
<td>University of Auckland</td>
<td>Figshare</td>
<td>Figshare</td>
</tr>
<tr>
<td>Brunel University</td>
<td>Figshare</td>
<td>Figshare</td>
</tr>
<tr>
<td>Melbourne University</td>
<td>Figshare</td>
<td>Figshare</td>
</tr>
<tr>
<td>University of Amsterdam</td>
<td>Figshare</td>
<td>Fedora(UvA-DARE)</td>
</tr>
<tr>
<td>Stockholm University</td>
<td>Figshare</td>
<td>DIVA Swedish universities pubs repository - software unknown</td>
</tr>
</tbody>
</table>

Research data repository software filtered from re3data.org

DSpace 29%
Dataverse 23%
Eprints 15%
Fedora 14%
CKAN 5%
There is no unanimous repository software solution, but…

- **American** universities favour **Dataverse** for research data
- **European** universities favour **Dspace** for research data
- **Australian** universities favour **Figshare** for research data
- **Fedora** is popular across these continents, but is used as a specialist repository solution where funding, staff and resources are prevalent and sustainable.
- **DSpace** is the most popular software choice for institutional research publication repositories (theses & dissertation)
Suggesting an IDR for UCT

Potential software solutions

Dscape, Fedora, Dataverse and Figshare are the most popular research data repository solutions

We also considered DRYAD, Zenodo, TIND, Globus, CKAN & EPrints
Eliminations & Evaluation
Eliminations

What was disregarded during the initial evaluation and on what basis?

1. **Budget/cost**
   - Commercial software as a service (SaaS) licence costs?
   - Free SaaS?

2. **Suitability of software**
   - Mission
   - Consideration criteria
### Commercial Software as a Service (SaaS)

#### Figshare for Institutions
- **Cost:** R800 000 + p/a
- **Includes:**
  - Training
  - 10TB AWS cloud storage
  - Basic customisations
  - Ongoing maintenance, upgrades and 24/7 support
- **Excludes:**
  - Local storage space
  - Preservation

#### Globus Data Publication
- **Cost:** R286 000 p/a
- **Includes:**
  - Standard subscription benefits, see Globus subscriptions
- **Excludes:**
  - Local storage costs
  - Preservation

#### TIND Institutional Repository
- **Cost:** R500 000 p.a.
- **Includes:**
  - Training
  - 15TB AWS cloud storage
  - Basic customisations
  - Ongoing maintenance, upgrades and 24/7 support
- **Excludes:**
  - Local storage space
  - Preservation

---

**Suggesting an IDR for UCT**

4 May 2017
Eliminations

Commercial Software as a Service (SaaS)

Pros:
• Speed of implementation
• Visualisation of many file types
• No need to employ a repository developer (staff expertise and time)
• 24/7 support
• Basic customisations included
• Basic training for staff included

Cons:
• Prohibitive start up costs
• Sustainability models for annual license fees would require cost recovery for annual license as well as storage costs
• Lost opportunity to develop skills in-house
Eliminations

**Free Software as a Service (SaaS)**

- **Cannot track institutional outputs**: repository services used to self-publish individual user data.

- **Cannot function as an institutional repository**:
  - Lack of control over deposit (no SSO)
  - No agreement to meet institutional RDM policies, workflows, etc.
  - Little or no institutional branding
  - Platform DOIs vs institutional DOIs
Other eliminations

- **CKAN** openly declares that the software is **not an institutional repository**. It requires vast customisation and is mainly used as government data repositories.

- **DRYAD** is a customised version of DSpace that currently only hosts data linked to **commercial research publications**. The university must have a commercial academic press in order to use Dryad for Organizations as an institutional data repository, otherwise it is **extremely expensive**.

- **EPrints** is used as an institutional data repository only in cases where the university’s research publication repository was already built on EPrints.
What was left to evaluate?

- Open Source Institutional Data Repository Platforms

<table>
<thead>
<tr>
<th>Category</th>
<th>Feature</th>
<th>DSpace 5.x</th>
<th>Fedora 4.x with Islandora CLAW</th>
<th>Fedora 4.x with Hydra</th>
<th>Dataverse 4.x</th>
</tr>
</thead>
<tbody>
<tr>
<td>VISION</td>
<td>The DSpace Project will produce the world’s choice for repository software providing the means for making information openly available and easy to manage.</td>
<td>The Fedora Project’s mission is to lead the advancement of Open and Open Source Software and content as a collaborative community.</td>
<td>One body, many heads - Tailored applications and workflows for different content types, contexts and user interactions. Building a framework, not an application.</td>
<td>The Dataverse is an open source web application to share, preserve, cite, explore, and analyze research data.</td>
<td></td>
</tr>
<tr>
<td>INFRASTRUCTURE</td>
<td>Server resources</td>
<td>Linux, Windows, postgres, maven, and Java Server (Tomcat etc.)</td>
<td>Linux</td>
<td>Linux (Red Hat), Windows, TomCat, soir etc. Deployment hardware information and server structure is complex and variable see <a href="https://goo.gl/bHxgf">https://goo.gl/bHxgf</a></td>
<td>Linux, glassfish, postgresQL, soir</td>
</tr>
<tr>
<td></td>
<td>Local / Hosted solutions</td>
<td>Local and hosted Hosting through Discovery Gateway @ $12000 p.a.</td>
<td></td>
<td>Local</td>
<td>Local and hosted</td>
</tr>
<tr>
<td></td>
<td>Hardware required</td>
<td>8G RAM, 1TB storage (dependant on content)</td>
<td>VMs to run docker containers</td>
<td>vms for: Fedora &amp; soir and hydra &amp; rails and oracle database. Deployment hardware information and server structure is complex and variable see <a href="https://goo.gl/bHxgf">https://goo.gl/bHxgf</a></td>
<td>min: 1 VM, 2 2.8 GHz, 8 GB mem, 50 GB disk max: 6 servers, 64 GB ram each. TBs storage from NAS using NFS, hardware load balancer</td>
</tr>
<tr>
<td></td>
<td>Build scripts</td>
<td>mvn</td>
<td>Docker, Ansible</td>
<td>ansible, vagrant</td>
<td>vagrant, puppet, ansible</td>
</tr>
<tr>
<td></td>
<td>Open source</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Platform customization</td>
<td>Yes, modularity through Manakin, xmlui much easier than jspui which required java development to customize</td>
<td>Through community. Ruby on Rails development</td>
<td>Possible with development in Java</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Support community</td>
<td>Yes - thousands of institutions</td>
<td>30 official partner institutions (and growing annually) and many working collaborators</td>
<td>17+ institutions</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Administrator configurations</td>
<td>Yes, with restrictions</td>
<td>Highly configurable</td>
<td>Highly configurable</td>
<td>Yes, per dataverse</td>
</tr>
<tr>
<td></td>
<td>User roles</td>
<td>Yes, with restrictions</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes, per dataverse and per user and per file. Pre-defined and custom roles</td>
</tr>
<tr>
<td>USER INTERFACE/FRONT-END</td>
<td>Integrated front-end</td>
<td>Yes, JSPUI(java) or Manakin (cmi)</td>
<td>Islandora</td>
<td>Hydra (see also Sufia <a href="http://sufia.io/">http://sufia.io/</a> hydra powered repository front end)</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Suggesting an IDR for UCT

4 May 2017
Scenario planning

Suggesting an IDR for UCT

4 May 2017
### Possible options...

<table>
<thead>
<tr>
<th>Commercial SaaS</th>
<th>Local Open Source</th>
<th>Free online platform</th>
<th>Hybrid (local SaaS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eg. Figshare / TIND with cloud storage</td>
<td>Eg. DSpace / Dataverse / Fedora</td>
<td>Eg. Zenodo</td>
<td>Eg. Figshare / TIND installed on top of local storage</td>
</tr>
<tr>
<td>Annual subscription fee</td>
<td>No license fees</td>
<td>No license fees</td>
<td>Annual subscription fee</td>
</tr>
<tr>
<td>Storage included in provider plan</td>
<td>Local storage &amp; compute infrastructure</td>
<td>Free cloud storage</td>
<td>Local storage &amp; compute infrastructure</td>
</tr>
<tr>
<td>No local IT staff</td>
<td>Dedicated IT staff</td>
<td>No IT staff</td>
<td>Local IT staff with support from provider’s IT staff</td>
</tr>
<tr>
<td>Customisation &amp; integration by provider as part of implementation</td>
<td>Customisation &amp; integration by institutional staff - ICTS developers</td>
<td>No customisation. No (or minimal) institutional branding. No local DOIs.</td>
<td>Customisation &amp; integration by provider as part of implementation</td>
</tr>
<tr>
<td>Upgraded &amp; maintained by provider</td>
<td>Upgraded &amp; maintained by institutional IT staff</td>
<td>Upgraded &amp; maintained according to platform’s schedule</td>
<td>Upgraded &amp; maintained by provider</td>
</tr>
<tr>
<td>Training &amp; support provided</td>
<td>Training &amp; support must be provided by institution</td>
<td>No training or support necessary</td>
<td>Training &amp; support provided</td>
</tr>
<tr>
<td>Repository management &amp; curation staff employed by institution</td>
<td>Repository management &amp; curation staff employed by institution</td>
<td>No repository managers, but curation staff are necessary</td>
<td>Repository management &amp; curation staff employed by institution</td>
</tr>
</tbody>
</table>
Scenarios

Scenario 1: National or institutional funding is available for operating the repository

- Commercial Saas
  eg. Figshare / TIND
- Locally hosted Open Source
  eg. DSpace / Fedora / Dataverse
Scenarios

**Scenario 1: National or institutional funding** is available for operating the repository

- **Commercial Saas**
  - eg. Figshare / TIND
- **Locally hosted Open Source**
  - eg. DSpace / Fedora / Dataverse

**Scenario 2: No funding** is available, but **costs can/will be recovered** from repository users

- **Locally hosted Open source**
  - eg. DSpace / Fedora / Dataverse
Scenarios

**Scenario 1:** National or institutional funding is available for operating the repository

- Commercial SaaS
  eg. Figshare / TIND
- Locally hosted Open Source
  eg. DSpace / Fedora / Dataverse

**Scenario 2:** No funding is available, but costs can/will be recovered from repository users

- Locally hosted Open source
  eg. DSpace / Fedora / Dataverse

**Scenario 3:** No funding is available and no costs can/will be recovered

- Free online platform
  eg. Zenodo
Recommendations
and current status of implementation

Suggesting an IDR for UCT
4 May 2017
Recommendations and status of implementation

1. Draft a sustainable business model for IDR
   - IDR costing evaluation
   - Pending tier 2 funding...

2. Comparative installation & solution testing
   - Re-evaluated due to pending tier 2 consortium

3. Create UCT community within free SaaS
   - Successfully implemented UCT Zenodo Community

4. Engage with select research communities for solution testing
   - Zenodo harvesting with ROER4D
   - Double-blind review process tested with researcher using Zenodo

5. Install Open Science Framework (OSF) alongside repository
   - Installation completed (but pending SSO) - UCT OSF
## Goals shifted...

<table>
<thead>
<tr>
<th>Who?</th>
<th>Institutional</th>
<th>Regional</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>• Principal investigators (PIs)</td>
<td>• PIs at consortium institutes</td>
</tr>
<tr>
<td></td>
<td>• No access to discipline repositories</td>
<td>• Mirrors?</td>
</tr>
</tbody>
</table>

| What?      | • Curated datasets linked to scholarly output (OpenUCT)                       | • Big data?                     |
|            | • All data types and disciplines                                              |                                 |

| How?       | • Open access to data                                                        | • Curators?                     |
|            | • Standard formats                                                           |                                 |
|            | • Curation by communities?                                                   |                                 |

| Function?  | • Preserve & disseminate published research data for validation & reuse      | • Raw data?                     |
|            |                                                                               |                                 |
Current Work

UCT OSF and UCT Zenodo
Open Science Framework (OSF) for institutions offers a free online service concerned with ‘filling the gaps’ between research data management tools and services to enable reproducible science.

Suggesting an IDR for UCT

4 May 2017
Suggesting an IDR for UCT

UCT (Open Science Framework) OSF

The OSF connects services you already use

View the Badges:
- Open Data
- Open Materials
- Preregistered

Services connected by the OSF:
- Dropbox
- Amazon Web Services
- S3
- Dataverse Project
- Box
- Figshare
- Zooniverse
- GitHub
Suggesting an IDR for UCT

University of Cape Town (UCT)

Recent uploads

- **Suggesting an Institutional Data Repository for the University of Cape Town**
  - Mias, Erika; Roos, Kayleigh; van Rooyen, Jason;
  - The UCT Institutional Data Repository (IDR) will serve principal investigators at the University of Cape Town in need of a repository to preserve and openly disseminate data that support their published research findings. The repository service is provided in terms of the UCT Research Data Management Policy.
  - Uploaded on January 30, 2017

- **The OER Adoption Pyramid**
  - Trotter, Henry; Cox, Glenda;
  - This Pyramid was developed in the course of a research paper focusing on why South African academics adopt OER or not. We understood that numerous factors shaped their choices, but it became apparent that some factors were "essential" to OER activity while others were merely "influential". To clarify
  - Uploaded on May 18, 2016

- **ROER4D Sub-project 1 - South and Southeast Asia OER review**
  - Dhamarajan, Gajaraj;
  - The ROER4D project endeavours to cover at least three regions in the Global South, namely, South America, Sub-Saharan Africa, and South and South East Asia. However, background information on OER projects, policies and research and information on infrastructural, legal, socio-cultural and/or economy
  - Uploaded on March 10, 2016
Suggesting an IDR for UCT

UCT Zenodo Community

Pros:
• Free
• User-friendly UI
• Covers basic data sharing and review needs

Cons:
• No institutional DOIs
• Limited institutional identity
• No view/download statistics
• Curator must approve every submission manually (no SSO)
Thank You! Questions?

Read the report
View the full presentation

Contact us:
Erika Mias: erika.mias@uct.ac.za
Kayleigh Lino: kayleigh.roos@uct.ac.za

Follow us:
Twitter: Digital@UCT
Facebook: DigitalLibraryServices