Digital Scholarship and your data

Thursday, 20th June

DLS
Digital Library Services

What are DATA to you?
“What (are my) research data?”

Go to www.menti.com and use the code 716700

1. Grab your phone
2. Go to www.menti.com
3. Enter the code 716700 and vote!
“What (are my) research data?”

... any information collected, stored, and processed to produce and validate original research results.
The research project (& its data) lifecycle

1. **Plan & Design**
2. **Collect & Capture**
3. **Collaborate & Analyse**
4. **Discover, Reuse & Cite**
5. **Share & Publish**
6. **Manage, Store & Preserve**
Digital Scholarship is the application and integration of digital tools and methods to discover, research and teach.
We provide Digital Scholarship services to the University of Cape Town, including the following:

- Data Curation activities towards best practices in Research Data Management (RDM);
- specialist Digitisation services towards Digital Preservation;
- expertise in Geographic Information Systems (GIS).

We advocate for Open Science, to make research done at UCT more efficient, collaborative, accessible, findable and reusable. We spearhead these practices as contributions to a more equitable and sustainable social order in the higher education landscape.

Source: DLS website: http://www digitalservices.lib.uct.ac.za/
Open Science is the movement to make scientific research (including publications, data, physical samples, and software) and its dissemination accessible to all levels of an inquiring society, amateur or professional.

Open Science is arguably simply proper science. Others are enabled to collaborate and contribute, since research data [...] and other research processes are freely available, under terms that enable reuse, redistribution and reproduction of the research and its underlying data and methods and subscribe to grounded ethical practices.

Source: Foster Open Science: [https://www.fosteropenscience.eu/foster-taxonomy/open-science-definition](https://www.fosteropenscience.eu/foster-taxonomy/open-science-definition)
Adapted from: Woolf, M.; Olliaro, P.; Todd, M. H. (2011). "Open science is a research accelerator". Nature Chemistry. 3 (10): 745–748. [https://doi.org/10.1038%2Fnchem.1149](https://doi.org/10.1038%2Fnchem.1149)
Open Science at UCT

Greater access to scientific inputs and outputs can increase scientific productivity through reducing duplication, allowing more research from the same data and multiplying opportunities for domestic and global participation in the research process.

Open science can reduce delays in the re-use of scientific research including articles and data, and promote a swifter path from research to innovation to produce new products and services.

Science, often publicly funded, should be publicly accessible to promote a greater awareness among citizens and to build public trust and support for public policies and investments in research. Open science also promotes citizen science in experiments and data collection.

Open access to scientific outputs allow for greater evaluation and scrutiny by the scientific community which means more accurate replication and validation of research results. Openness to data contributes to maintain science’s self-correction principle.

Science plays a key role in today’s knowledge economies and increased access to research results, including data, can positive impact not only scientific systems but also innovation.

Open science promotes collaborative efforts and faster knowledge transfer for a better understanding of global challenges and wicked problems.

Digital Scholarship and Research Data Management towards Open Science
• **Digital Scholarship** tools exist to help us manage our data - keep track, store, secure, cleanup and organize
• Part of management is having sufficient and clear data describing the data - or *metadata*
• We think we will always know our data - that we can store the data about our data in our heads
• In the digital and networked world, *Metadata* becomes the currency of exchange that enables data to link with other data and researchers. It contributes to making data findable, accessible and resuable.
What is Research Data Management (RDM)?

- The **organisation and documentation** of the data processes (collection, description, de-identification, curation, archiving and publication) within a research project.
- Part of an international drive towards **Open Science**, to professionalise data management practices, and make research more coherent and shareable.
- Good **Digital Scholarship** practices along every step of the research lifecycle help data management and enable Open Science.
What is closed science?

Source: NYU Health Sciences Libraries.  https://youtu.be/N2zK3sAtr-4
...to transform the way research is conducted at UCT by accelerating discovery, increasing the value of research decision-making, and catalysing changes throughout the economy and society that are of value to all citizens.

The University seeks to ensure consistent research practice related to data management principles that support effective data sharing, including open access; and the need for data to be discoverable, accessible, reusable and interoperable to specific quality standards.
The Research Data Lifecycle

*Digital Scholarship tools and methods to assist with Research Data*
The research data management (RDM) lifecycle

1. **Plan & Design**
2. **Collect & Capture**
3. **Collaborate & Analyse**
4. **Discover, Reuse & Cite**
5. **Share & Publish**
6. **Manage, Store & Preserve**
PLAN & DESIGN
A **data management plan** (DMP) is a **living**, written document explaining what you intend to do with your data during and following the conclusion of your research project.

A DMP is already a **requirement** by many **funders** (NIH, Wellcome Trust, NRF).

Even when it is not a requirement, having made such a plan can **save you time** and **effort** during your research, as it assists you with **organising your data**, preparing it for the next step in its lifecycle, and clarifying who will have access to it, how, and when.

A DMP provides **guidance for curation-specific activities**, such as file-naming, archiving, formats suitable for long-term preservation, etc.
Useful information is provided at every step.
Typical DMP questions

- **What type of data** will be generated in your research?
- How will your data be **named and referenced**?
- What **file formats** are involved?
- What data and **metadata standards** will you follow?
- Who will **have access** to your data?
- How and when will you **share** your data, if applicable?
- Will you be **digitally preserving** your data? If yes, how so?
- How will you **license** your datasets?
- How will you ensure **privacy** or **confidentiality**, if applicable?
COLLECT & CAPTURE
RedCap

https://trn-redcap.uct.ac.za/

A secure web application for building and managing online surveys and databases, useful for collecting and tracking information and data from research studies, scheduling study events and conducting surveys.

Features:

- input data from anywhere in the world
- projects can be used by researchers from multiple sites and institutions
- total control of shaping your database or survey
- data may be imported from external data sources to begin a study or to provide mid-study data uploads
- export survey results to common data analysis packages
- generate a PDF version for printing in order to collect responses offline

Adapted from: Harvard Catalyst: https://catalyst.harvard.edu/services/redcap/
Advice for the Collect & Capture Process

While collecting and capturing your data, make sure that you document it with correct, meaningful metadata:

- Describe the type of data generated:
  - The form *(What kind of data does it hold?)*
  - The stability of each dataset *(How does it change over time?)*
  - Create unique names for each of your datasets

- Document the data you are capturing, and how you are identifying it within each data set by building a data dictionary.

- Practice good file naming conventions.

- Document your process and store it together with your data (e.g. readme.txt).

Signs you might not be managing your data

1. Your only copy of your data is on a flash that you left behind in a Postnet
2. You know you saved it somewhere but the search function is not finding it and you can’t remember the file name
3. Your collaborator can’t make sense of the contents of the files: ie What does the column *love1* stand for?
4. The program that you used to collect the data doesn’t work on your updated operating system and you can’t open it anymore
5. You have four versions of the same file and can’t tell which one is the right one, is it *final.docx*, or *final_final.docx*, or *copy of final_final.docx* or *thisoneistherightone_final - Copy.docx*
COLLABORATE & ANALYSE
Open Science Framework (OSF)

- free, online platform that allows you to register your project, manage stakeholders, and centralise data that might be stored at different locations with different collaborators
- allows integrations with Google Drive, Dropbox, OneDrive, figshare, and many more
- provides unlimited, free storage
- helps with creating versions of your project at different stages (‘forking’)
- includes wiki-components for ease of documentation and description, including the development of a data dictionary

https://osf.io/institutions/uct/
Advanced digital scholarship

Data Analysis and Mining:
Tools that help you identify patterns in large volumes of data, combining statistics, AI and machine learning.
● Tools and processes for data de-identification, to safeguard privacy of patients.
● Tools and process for text analysis.

Data Visualization:
Tools that develop a graphical presentation of data and information through visual means.

Digital Humanities:
Tools, processes and critical awareness found in the intersection between digital technologies and fields of study within the humanities.

Tableau - https://www.tableau.com/
Geographic Information Systems

Everything Happens Somewhere:

- Because everything happens somewhere everything can be associated with a spatial location.
- These locations can be mapped in space, either for simple visualisation or for complex analyses.

Data Visualisation (Maps):

- Maps are an incredibly powerful visualisation tool which allow us to view and display our data in interesting and informative ways. They allow us to see patterns in our data, not just find them.
- They also allow us to communicate our findings in a clear and succinct manner.

Images sourced from UrbanObservatory.org’s App
GIS

Data Analysis (Making Information):

The full potential of GIS is realised when performing spatial analyses. Different types of analyses exist to satisfy various needs:

- **Overlay** Analysis allows us to compare different data types, e.g. Mean Annual Rainfall and Crop Type.
- **Geostatistical** Analysis allows us to perform statistical analyses of correlated spatial data, e.g. Hotspot Analysis.
- **Network** Analysis allows us to calculate travel times and service delivery areas, e.g. “Golden Hour” coverage or Clinic’s Service Area.
- **Dashboards** of real time sensor feeds for live monitoring, e.g. Resource Usage; Traffic Volumes; Fleet Management.

DLS’ GIS services assist with GIS software acquisition, project planning, troubleshooting, analysis and cartographic design.

Find us @ [www.gis.uct.ac.za](http://www.gis.uct.ac.za)

Difference in Malaria rates between 2000 and 2015.

From the [urbanobservatory.org](http://urbanobservatory.org)
DISCOVER, REUSE & CITE
Building a Culture of Data Citation

1. Dataset is stored in a publicly accessible repository. Digital Object Identifier (DOI) minted for dataset (e.g., ANDS mints DOIs for Australian repositories).

2. Researcher future funding and promotion influenced by dataset citation metrics.

3. Researcher creates a research dataset and a publication related to the dataset.

4. DOI is used in data citation.

5. Citation metrics services (e.g., Thomson Reuters Data Citation Index) accumulate citation references to the dataset and publication.

6. Funding and research groups review publication and dataset citation metrics.

7. REWARD: Grant funding and recognition for significant contributions.

USE

Research community use the DOI to access the dataset and carry out related research.

Research community generate new publications using the DOI to reference the dataset.

MEASURE

Funding and research groups review publication and dataset citation metrics.

A small overview of data catalogues, registries and repositories

- **BioLINCC** – Clinical specimen database.
- **Dataverse** – Widely used open source repository system; Example: HARVARD Dataverse
- **dataMED** – prototype biomedical data search engine to discover data sets across data repositories or aggregators.
- **Code Ocean** – Cloud-based computational platform which provides a way to share, discover and run published code.
- **ContentMine** – Uses machines to liberate 100,000,000 facts from the scientific literature.
- **DataBank** – Analysis and visualisation tool that contains collections of time series data on a variety of topics.
- **DataCite** – Establish easier access to research data by providing persistent identifiers for data.
- **Datathub** – Publish or register datasets, create and manage groups and communities
- **Dataverse Network** – Harvard-based tool to share, cite, reuse and archive research data.
- **Deveo** – Free, private Git, Mercurial, and SVN repository management platform.
- **Dryad** – Data repository system for any files associated with any published article in the sciences or medicine.
- **Figshare (.com)** – Free cloud service for managing, sharing & publishing research data.
- **GenBank** – Gene sequence database provided by the National Center for Biotechnology Information.
- **GitHub** – Online software project hosting using the Git revision control system.
- **How Can I Share It** – Information and tools to ensure your articles can be shared with your colleagues easily.
- **Open Science Framework** – Open registration, version control & collaboration software system.
- **Quip** – Combines chat, documents, spreadsheets, checklist, and more to collaborate on any device.
- **re3data** – Global registry of research data repositories.
- **Research Compendia** – Tools for researchers to connect data, code & computational methods to published research.
- **SlideShare** – Community for sharing presentations and other professional content.
- **Zenodo** – A home for the long-tail of science, enabling researchers to share and preserve any research outputs.
- **ZivaHub | Open Data UCT** – UCT’s digital repository.

Adapted from: Zimmer, Niklas; King, Thomas (2018): Data discovery and re-use. figshare. Presentation. https://doi.org/10.25375/uct.7358423.v1
Working with the **FAIR** guiding principles

- **Findable**
  - Describe your data in a data repository
  - Receive a persistent identifiers (e.g. uct doi provided by ZivaHub)

- **Accessible**
  - Consider what can be published
  - Obtain participant consent
  - Perform de-identification / anonymisation

- **Interoperable**
  - Use open formats
  - Apply consistent vocabulary
  - Use common/disciplinary metadata standards

- **Reusable**
  - Consider permitted use
  - Apply machine-readable open licenses (e.g. CC-BY etc.)

Adapted from: Zimmer, Niklas; King, Thomas (2018): *Data discovery and re-use*. figshare. Presentation. [https://doi.org/10.25375/uct.7358423.v1](https://doi.org/10.25375/uct.7358423.v1)
5 ★ Open Data [Tim Berners-Lee]

- make your stuff available on the Web (whatever format) under an open license
- make it available as structured data (e.g., Excel instead of image scan of a table)
- make it available in a non-proprietary open format (e.g., CSV instead of Excel)
- use URIs to denote things, so that people can point at your stuff
- link your data to other data to provide context

Source: https://5stardata.info/en/
‘Good RDM makes data reusable’

Source: 10 aspects of highly effective research data - Good research data management makes data reusable By Anita de Waard, Helena Cousijn, PhD, and IJsbrand Jan Aalbersberg, PhD
SHARE & PUBLISH
What Stops you from sharing data?
Why share your research data?

1. **To comply with Funding agencies and institutions** requiring results of scientific studies to be shared with the public as a condition for providing grants or awards.

2. To satisfy publishers’ requests to deposit some datasets in public platforms.

3. To **confront some of the biases in data collection and analysis** (Atici et. al, 2013).

4. To **reproduce** or to **verify** research, and to **ask new questions** of extant data.

6. To **advance the state of research** and innovation (Borgman, 2012).

7. **To increase citation rate** (Piwowar, Day and Fridsma, 2007).

8. To **increase the visibility of individual researchers and their work online** (Peters et al., 2015).
ZivaHub | Open Data UCT

https://zivahub.uct.ac.za/

- A repository to store and openly disseminate data
- Powered by Figshare for institutions
- Keeps track of views, downloads and citations
- Allows search across all Figshare platforms
ZivaHub | Open Data UCT

https://zivahub.uct.ac.za/

supports the upload of any file format, and aims to visualise all of them

embeds relevant metadata, to make data FAIR compliant

choose from a range of licensing options when publishing your data openly
share data privately with your funders, reviewers, or supervisors before publishing it later

all item records get assigned a Digital Object Identifier (DOI)
MANAGE, STORE, PRESERVE
A brief history of the ‘data deluge’

Global Information Storage Capacity in optimally compressed bytes

- **1986**
  - Analog: 2.6 exabytes
  - Digital: 0.02 exabytes

- **1993**
  - Analog storage

- **2000**
  - Digital storage

- **2007**
  - Analog: 19 exabytes
    - Paper, film, audiotape and vinyl: 6%
    - Analog videotapes (VHS, etc.): 94%
    - Portable media, flash drives: 2%
    - Portable hard disks: 2.4%
    - CDs and minidisks: 6.8%
  - Digital: 280 exabytes
    - Computer servers and mainframes: 8.9%
    - Digital tape: 11.6%
    - DVD/Blu-ray: 22.8%
    - PC hard disks: 44.5%
    - Others: < 1% (incl. chip cards, memory cards, floppy disks, mobile phones, PDAs, cameras/camcorders, videogames)

- **2002**: “beginning of the digital age” 50%

% digital: 1% 3% 25% 94%

A future problem: where is my data?

I know where it is but...

- It’s in an unsupported file format
- It’s in a legacy system
- It’s not well described so it’s irretrievable
- It’s corrupted

I don’t even know where it is...

- It was on destroyed hardware
- A third party has it
- It’s on a hard drive in a vault
- I expected it to be just where I left it

Yes, maintaining backups of your stored data is crucial! But this does not mean that they are digitally preserved. Digital preservation is an institutional endeavour to ensure that data remain accessible and usable in the long term, in view of:

- **technological change** (e.g. legacy media & formats)
- **bit-rot** (decay of digital files over time, e.g. on flash drives)
- **link-rot** (decay of identifiers over time, e.g. on websites)
- **media failure** (e.g. ‘head crash’ on hard drives, CD-Rs oxidising)

Digital preservation is generally handled by specialist staff, such as archivists and librarians, using dedicated hard- and software solutions. Researchers need to be aware that some of their data may legally require digital preservation, and ideally participate actively in the process of planning for it from the outset (see: DMP).
Closing Remarks & Future Interactions
Digital Scholarship is the application and integration of digital tools and methods discover, research and teach.

Research Data Management is the organization and documentation of research data (ideally towards making it Findable, Accessible, Interoperable and Reusable).

Open Science is a set of practices that drives all aspects of research to be more efficient, accountable, collaborative, and of good quality.

Digital Scholarship, within a research project, integrates digital technologies, works within networked environments and subscribes to Open Science. When all the above intersect, DS has the power to transform the research landscape.
Upcoming Workshops

RESEARCH DATA MANAGEMENT TRAINING
Discover how you can become a more EFFICIENT researcher in today’s digital world. Start managing your DATA and your RESEARCH process with guidance from the DLS TEAM.

RESEARCH DATA MANAGEMENT WITH DMPonline
The new Student MoI, as well as the NRF, require students to outline their data plans for their research projects in a Data Management Plan (DMP). This talk/workshop takes you through the reasons for creating a DMP, as well as guiding you through using the DMPonline website.

DOING DIGITAL SCHOLARSHIP
Doing research requires interacting with a multitude of digital spaces. This talk outlines digital processes and tools that can increase efficiencies throughout a research project. It looks at collaborative tools for managing, analyzing, mapping and visualizing research data.

SHARING AND PUBLISHING WITH ZivaHub
UCT’s open data repository is rapidly growing. Uploading your research outputs to ZivaHub makes them discoverable, citable, shareable and reusable. Learn about open data and ZivaHub which allow you to engage with researchers at UCT and the world.

Sharing and Publishing with ZivaHub:
10 July @ 10 AM
RDM at UCT

UCT Open Science Framework (OSF)

Digital preservation

OneDrive / Google Drive etc.

ZivaHub | Open Data UCT

'Open Data UCT'

UCT DMPonline

'RDM at UCT' Slack workspace

Slack: 'Searchable Log of All Conversation and Knowledge'

---

Niklas Zimmer 10:26

https://figshare.com/articles/Monash_University_s_content_migration_its_impact_on_research_outputs_Figshare/6396776

Monash University's Content Migration: A case study

This is a case study based on Monash University's experience migrating content, including their theses, into their instance of Figshare.

For more information on Monash University's content migration, including the coding required to migrate the content and work done in-house versus commissioned by the university, please reach out to Andrew: andrew.harrison@monash.edu.

References

https://monash.figshare.com/theses
Thank you!

dls@uct.ac.za

@DigitalUct

rdm-at-uct.slack.com

http://www.digitalservices.lib.uct.ac.za/