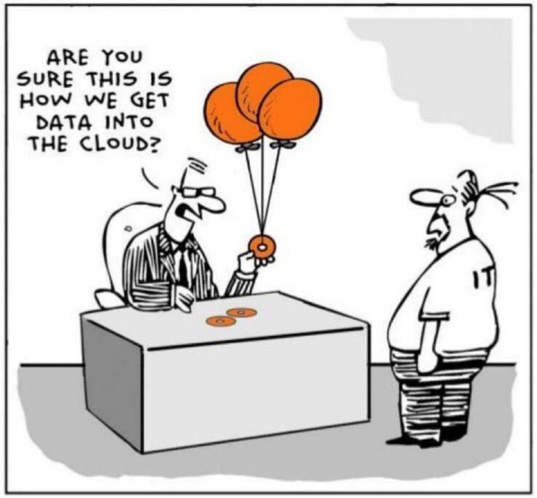


# *Digital Archives and Data Management*

# *Summer Institute on Computational Social Science*

*Hussein Suleman <hussein@cs.uct.ac.za>*
*Digital Libraries Lab, Department of Computer Science, School of IT*
*University of Cape Town*
*19 June 2019*

# Topic Outline

- Research Data Management
- Digital Repositories
- Issues in Data Management

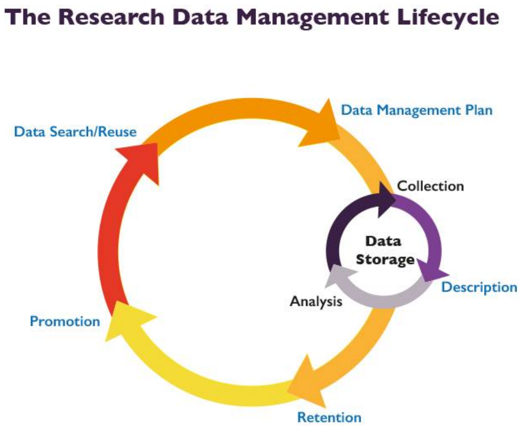
# Research Data Management

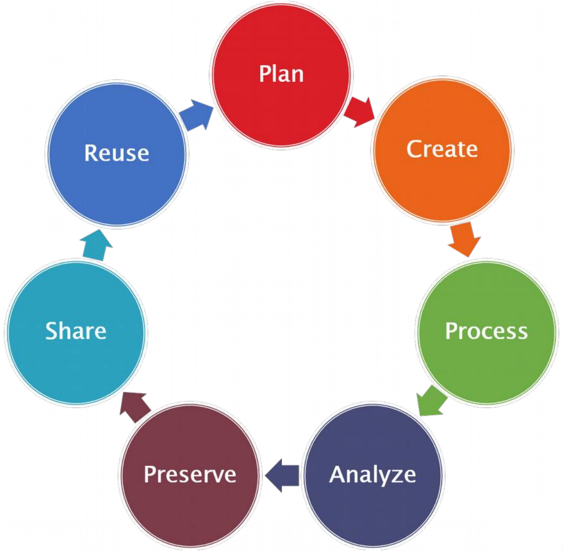
# What is Research Data Management?

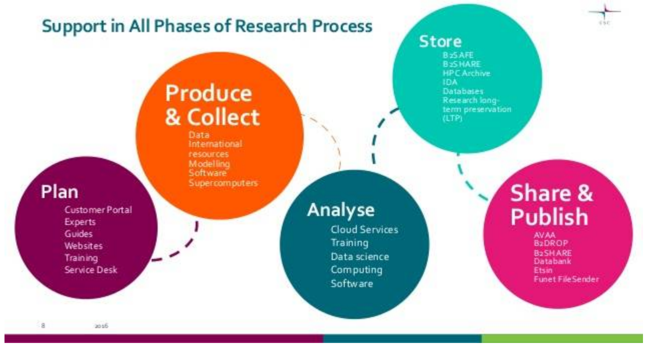
- Planning
- Organisation
- Storage
- Sharing
- Stewardship

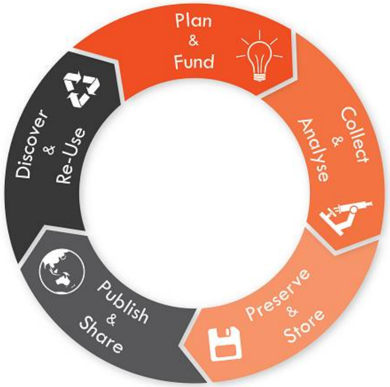
- ...of your research data.

# Research Data Management Lifecycles









# What is Research Data?

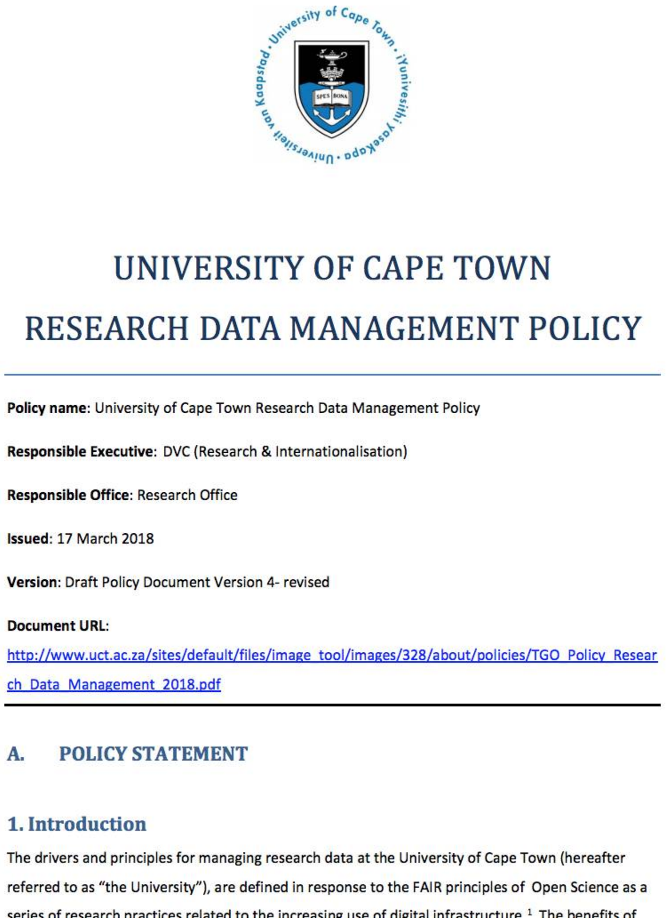
# What is Research Data?

- □ Raw survey data in CSV files?
- □ Graphs generated in Excel?
- □ Software program to produce analysis (R/Python)?
- □ Telephone interview audio files?
- □ Paper-based survey responses?
- □ Research paper/article/chapter/book?
- □ Spreadsheets containing descriptions of objects?
- □ Video recordings on cellphone?
- □ Lecture slides on a research topic?
- □ Infographic produced for a newspaper?

# Why data management?

- Plan/resource appropriately.
- Protect data from disaster/misuse/etc..
- Ensure quality in data.
- Share data.
- Allow verification of results.
- Reuse data.
    - Reduce costs.
    - Expand research.

# Why data management? Institutional Policies



University of Cape Town • iYunivesithi yaseKapa • Universiteit van Kaapstad

## UNIVERSITY OF CAPE TOWN

## RESEARCH DATA MANAGEMENT POLICY

**Policy name**: University of Cape Town Research Data Management Policy

**Responsible Executive**: DVC (Research & Internationalisation)

**Responsible Office**: Research Office

**Issued**: 17 March 2018

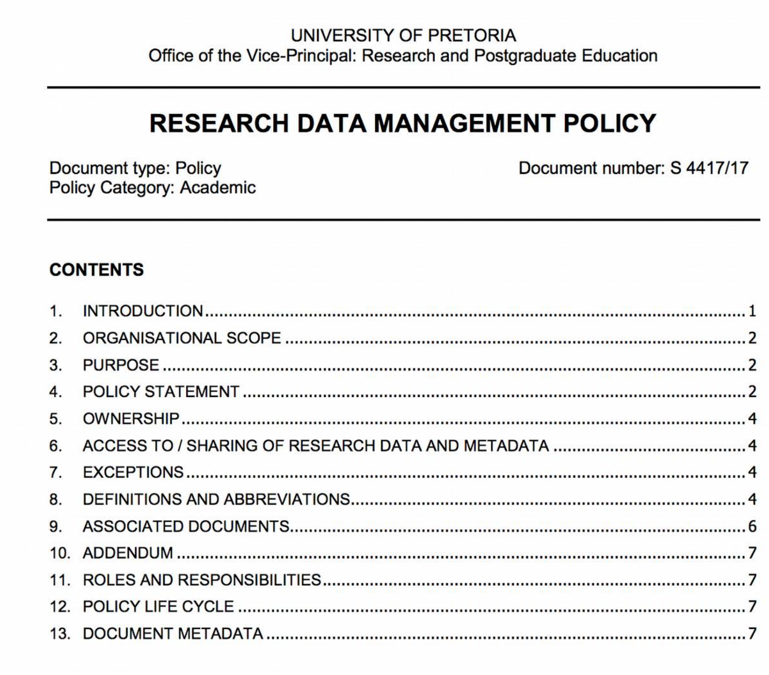
**Version**: Draft Policy Document Version 4- revised

**Document URL**:
http://www.uct.ac.za/sites/default/files/image_tool/images/328/about/policies/TGO_Policy_Research_Data_Management_2018.pdf

### A. POLICY STATEMENT

### 1. Introduction

The drivers and principles for managing research data at the University of Cape Town (hereafter referred to as "the University"), are defined in response to the FAIR principles of Open Science as a series of research practices related to the increasing use of digital infrastructure.[1] The benefits of



UNIVERSITY OF PRETORIA
Office of the Vice-Principal: Research and Postgraduate Education

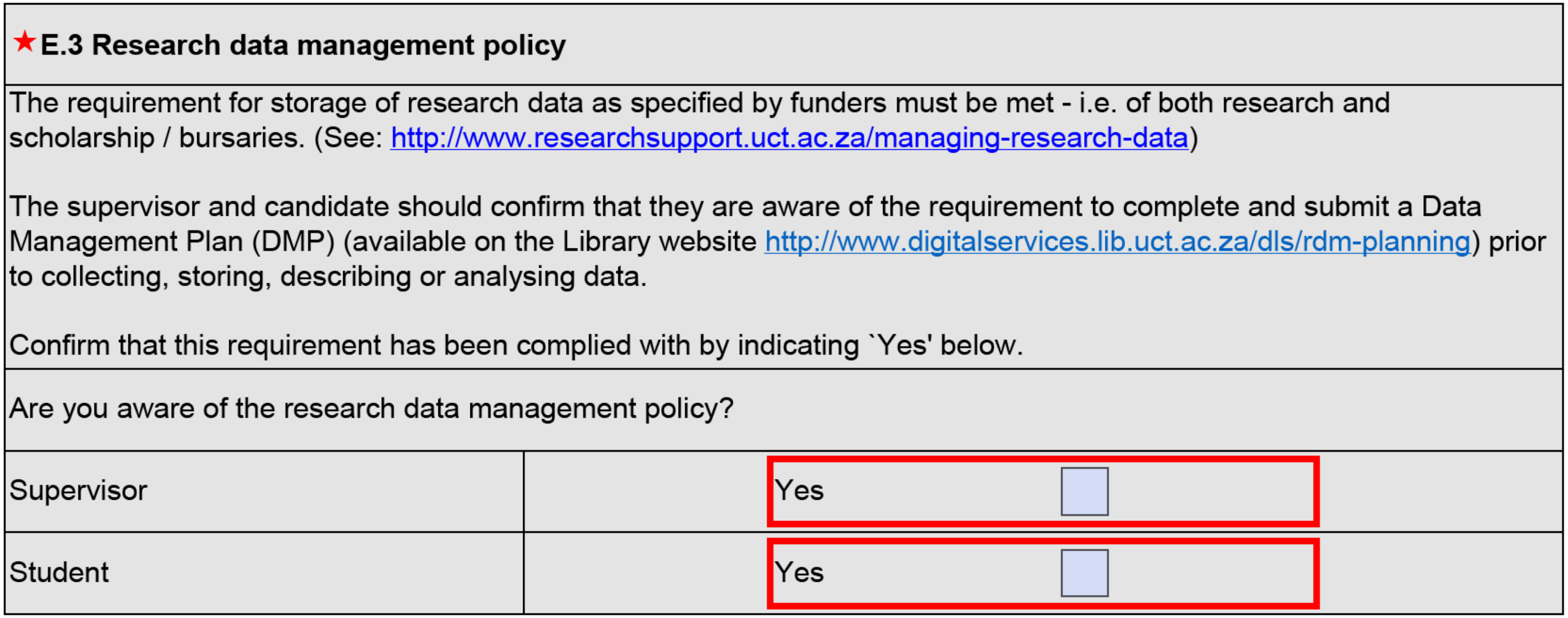
## RESEARCH DATA MANAGEMENT POLICY

Document type: Policy
Policy Category: Academic

Document number: S 4417/17

**CONTENTS**

# Why data management? Requirements



★**E.3 Research data management policy**

The requirement for storage of research data as specified by funders must be met - i.e. of both research and scholarship / bursaries. (See: http://www.researchsupport.uct.ac.za/managing-research-data)

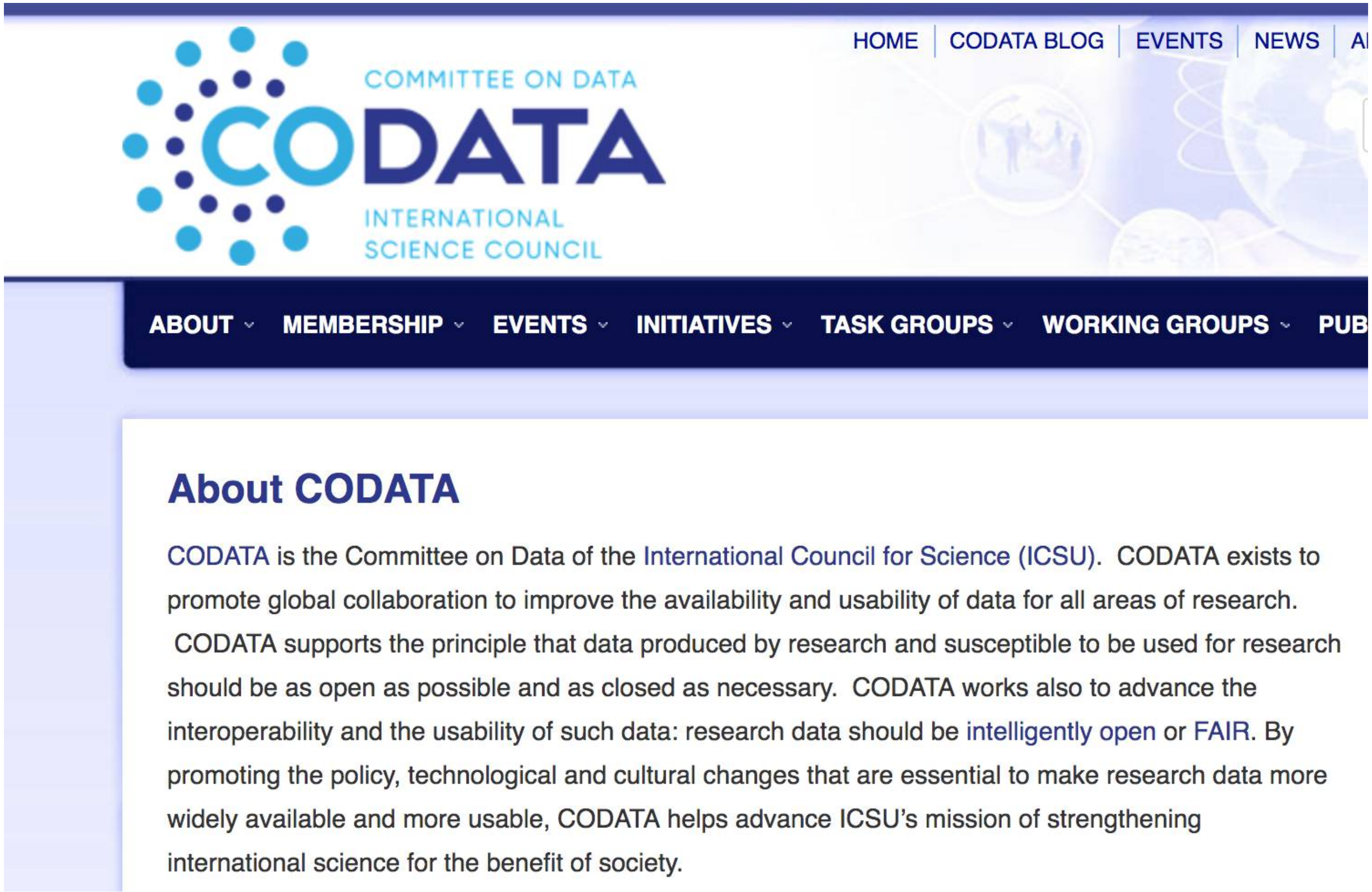
The supervisor and candidate should confirm that they are aware of the requirement to complete and submit a Data Management Plan (DMP) (available on the Library website http://www.digitalservices.lib.uct.ac.za/dls/rdm-planning) prior to collecting, storing, describing or analysing data.

Confirm that this requirement has been complied with by indicating `Yes' below.

Are you aware of the research data management policy?

| | | |
|---|---|---|
| Supervisor | | Yes □ |
| Student | | Yes □ |

# Why data management? International bodies



## About CODATA

CODATA is the Committee on Data of the International Council for Science (ICSU). CODATA exists to promote global collaboration to improve the availability and usability of data for all areas of research. CODATA supports the principle that data produced by research and susceptible to be used for research should be as open as possible and as closed as necessary. CODATA works also to advance the interoperability and the usability of such data: research data should be intelligently open or FAIR. By promoting the policy, technological and cultural changes that are essential to make research data more widely available and more usable, CODATA helps advance ICSU's mission of strengthening international science for the benefit of society.

# Why data management? Principles: FAIR

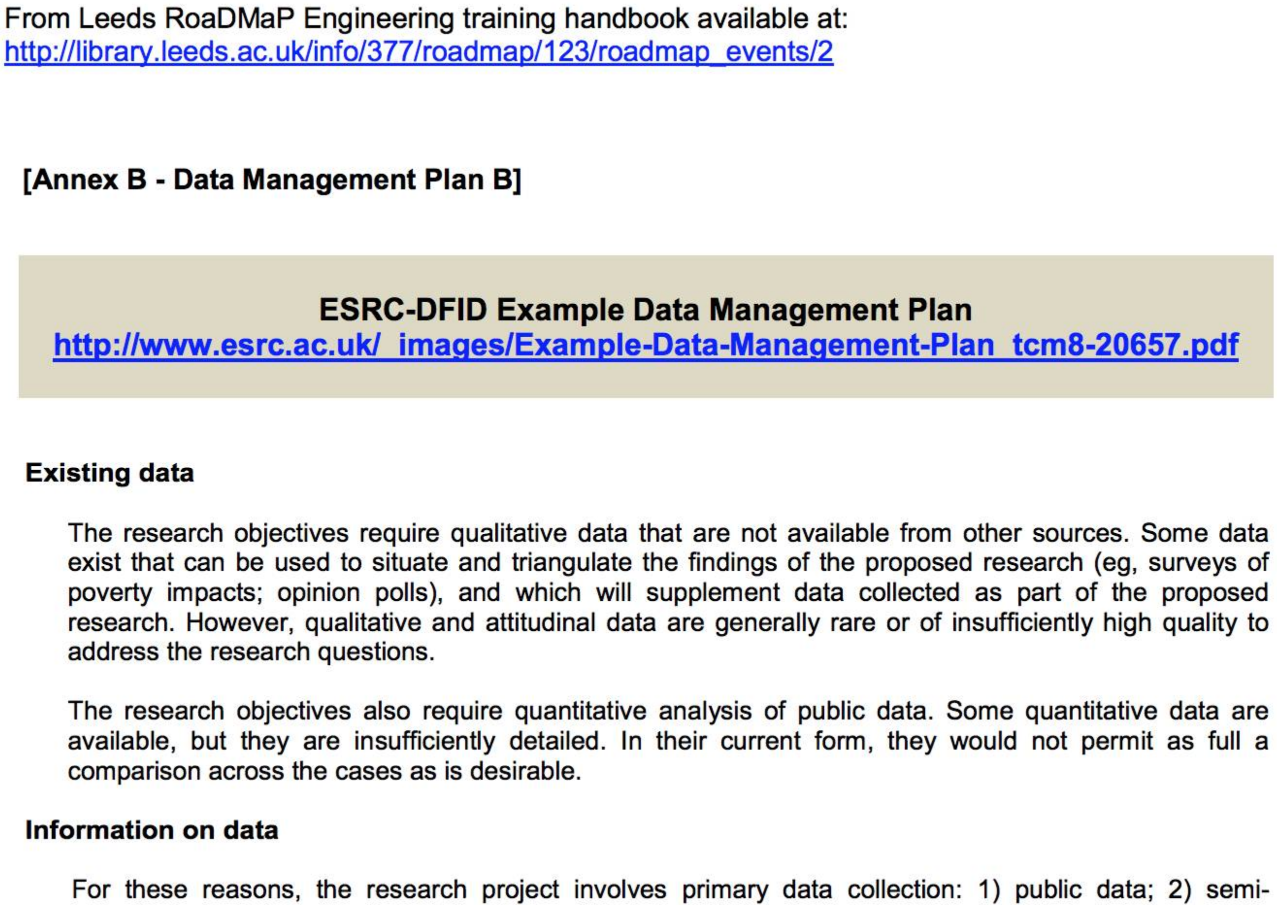
- Research data must be …
  - Findable
  - Accessible
  - Interoperable
  - Reusable

# Phase 1: Planning

- Think about the data use before a project:
  - Availability
  - Costing
  - Resourcing (store, collect, ...)
  - Analysis requirements
  - Potential problems
  - Etc.

- How methodical should you be?

# Data Management Plan



From Leeds RoaDMaP Engineering training handbook available at:
http://library.leeds.ac.uk/info/377/roadmap/123/roadmap_events/2

[Annex B - Data Management Plan B]

**ESRC-DFID Example Data Management Plan**
**http://www.esrc.ac.uk/_images/Example-Data-Management-Plan_tcm8-20657.pdf**

**Existing data**

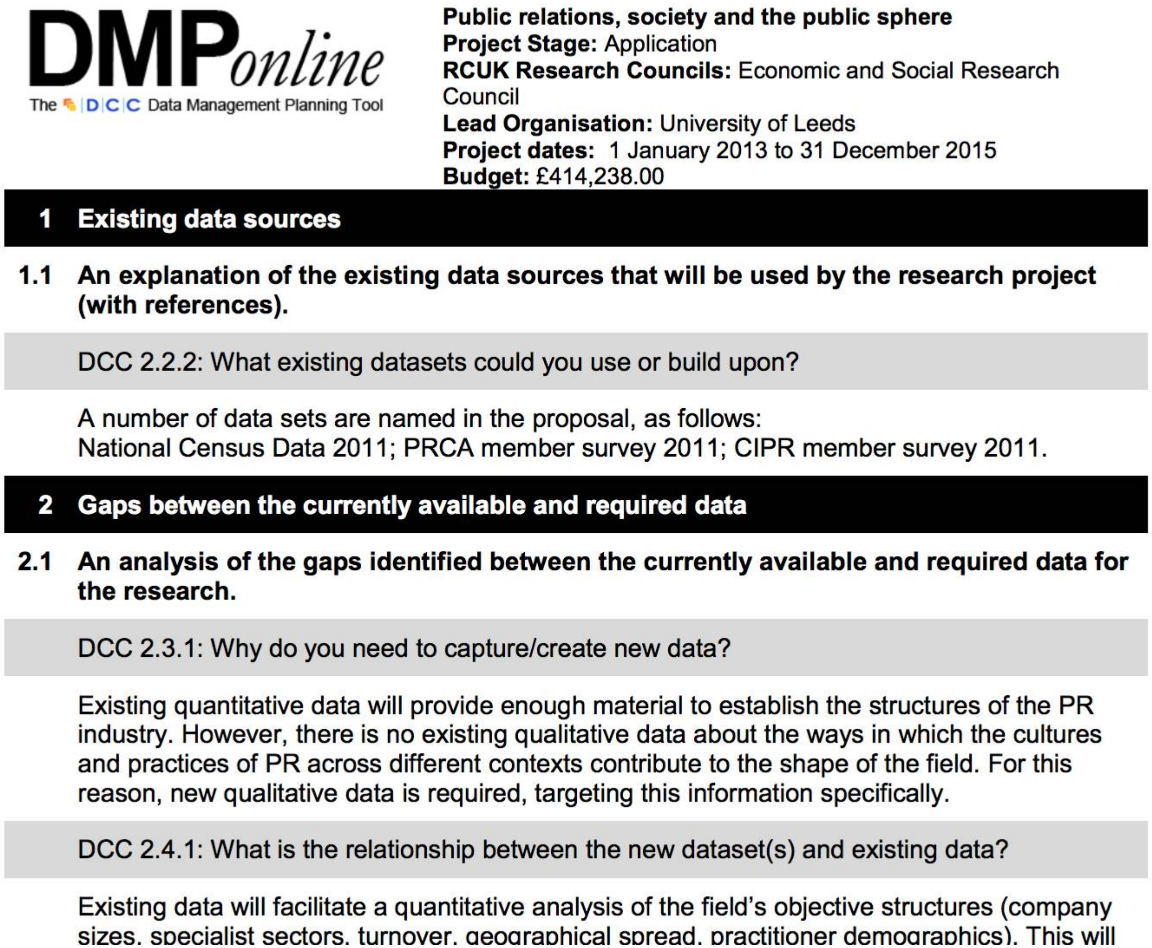
The research objectives require qualitative data that are not available from other sources. Some data exist that can be used to situate and triangulate the findings of the proposed research (eg, surveys of poverty impacts; opinion polls), and which will supplement data collected as part of the proposed research. However, qualitative and attitudinal data are generally rare or of insufficiently high quality to address the research questions.

The research objectives also require quantitative analysis of public data. Some quantitative data are available, but they are insufficiently detailed. In their current form, they would not permit as full a comparison across the cases as is desirable.

**Information on data**

For these reasons, the research project involves primary data collection: 1) public data; 2) semi-

# Data Management Plan: DMPOnline



DMPonline
The DCC Data Management Planning Tool

**Public relations, society and the public sphere**
**Project Stage:** Application
**RCUK Research Councils:** Economic and Social Research Council
**Lead Organisation:** University of Leeds
**Project dates:** 1 January 2013 to 31 December 2015
**Budget:** £414,238.00

**1 Existing data sources**

**1.1 An explanation of the existing data sources that will be used by the research project (with references).**

DCC 2.2.2: What existing datasets could you use or build upon?

A number of data sets are named in the proposal, as follows:
National Census Data 2011; PRCA member survey 2011; CIPR member survey 2011.

**2 Gaps between the currently available and required data**

**2.1 An analysis of the gaps identified between the currently available and required data for the research.**

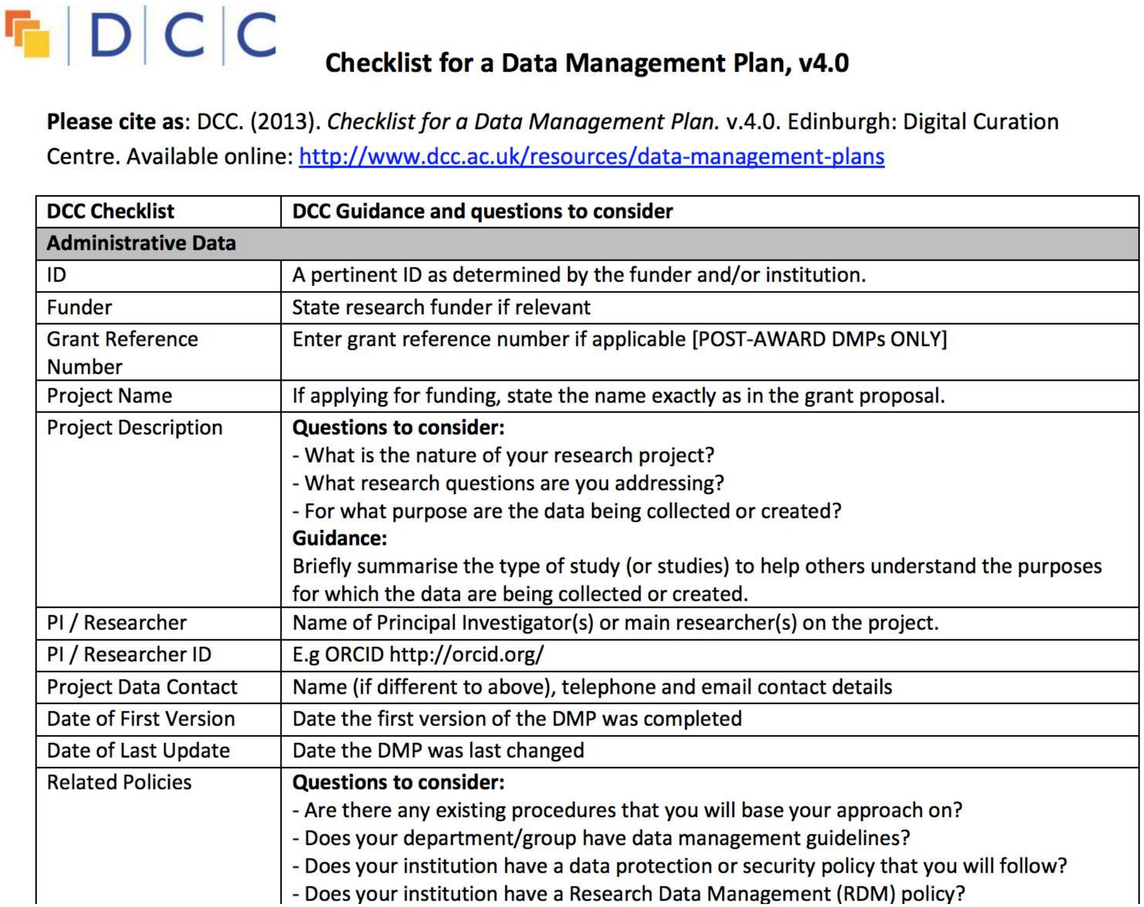
DCC 2.3.1: Why do you need to capture/create new data?

Existing quantitative data will provide enough material to establish the structures of the PR industry. However, there is no existing qualitative data about the ways in which the cultures and practices of PR across different contexts contribute to the shape of the field. For this reason, new qualitative data is required, targeting this information specifically.

DCC 2.4.1: What is the relationship between the new dataset(s) and existing data?

Existing data will facilitate a quantitative analysis of the field's objective structures (company sizes, specialist sectors, turnover, geographical spread, practitioner demographics). This will

# DMP Checklist



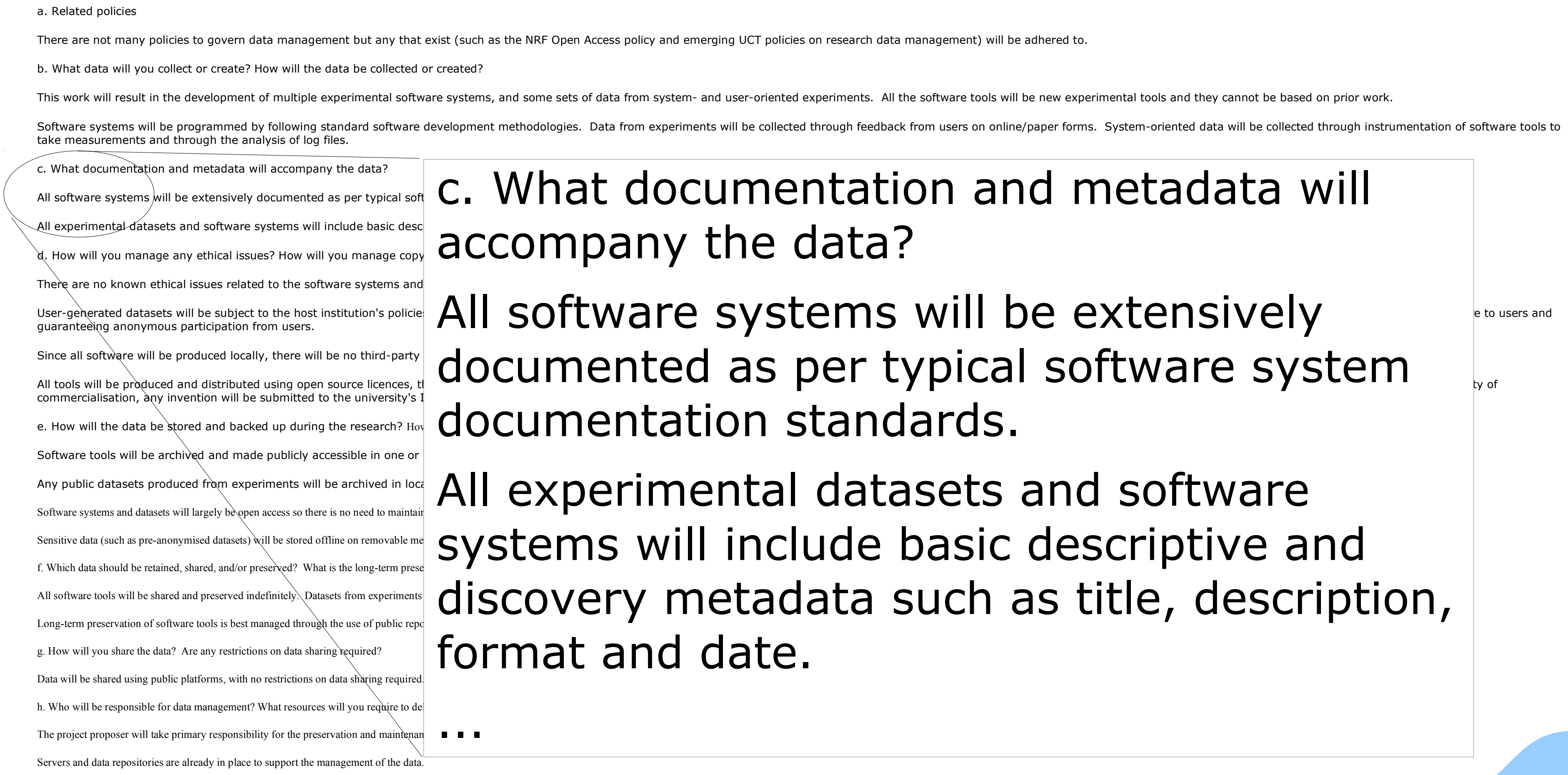
DCC

**Checklist for a Data Management Plan, v4.0**

**Please cite as**: DCC. (2013). *Checklist for a Data Management Plan.* v.4.0. Edinburgh: Digital Curation Centre. Available online: http://www.dcc.ac.uk/resources/data-management-plans

| DCC Checklist | DCC Guidance and questions to consider |
|---|---|
| **Administrative Data** | |
| ID | A pertinent ID as determined by the funder and/or institution. |
| Funder | State research funder if relevant |
| Grant Reference Number | Enter grant reference number if applicable [POST-AWARD DMPs ONLY] |
| Project Name | If applying for funding, state the name exactly as in the grant proposal. |
| Project Description | **Questions to consider:**<br>- What is the nature of your research project?<br>- What research questions are you addressing?<br>- For what purpose are the data being collected or created?<br>**Guidance:**<br>Briefly summarise the type of study (or studies) to help others understand the purposes for which the data are being collected or created. |
| PI / Researcher | Name of Principal Investigator(s) or main researcher(s) on the project. |
| PI / Researcher ID | E.g ORCID http://orcid.org/ |
| Project Data Contact | Name (if different to above), telephone and email contact details |
| Date of First Version | Date the first version of the DMP was completed |
| Date of Last Update | Date the DMP was last changed |
| Related Policies | **Questions to consider:**<br>- Are there any existing procedures that you will base your approach on?<br>- Does your department/group have data management guidelines?<br>- Does your institution have a data protection or security policy that you will follow?<br>- Does your institution have a Research Data Management (RDM) policy? |

# A DMP Checklist-driven plan



a. Related policies

There are not many policies to govern data management but any that exist (such as the NRF Open Access policy and emerging UCT policies on research data management) will be adhered to.

b. What data will you collect or create? How will the data be collected or created?

This work will result in the development of multiple experimental software systems, and some sets of data from system- and user-oriented experiments. All the software tools will be new experimental tools and they cannot be based on prior work.

Software systems will be programmed by following standard software development methodologies. Data from experiments will be collected through feedback from users on online/paper forms. System-oriented data will be collected through instrumentation of software tools to take measurements and through the analysis of log files.

c. What documentation and metadata will accompany the data?

All software systems will be extensively documented as per typical soft

All experimental datasets and software systems will include basic desc

d. How will you manage any ethical issues? How will you manage copy

There are no known ethical issues related to the software systems and

User-generated datasets will be subject to the host institution's policie e to users and guaranteeing anonymous participation from users.

Since all software will be produced locally, there will be no third-party

All tools will be produced and distributed using open source licences, tl ty of commercialisation, any invention will be submitted to the university's I

e. How will the data be stored and backed up during the research? Hov

Software tools will be archived and made publicly accessible in one or

Any public datasets produced from experiments will be archived in loca

Software systems and datasets will largely be open access so there is no need to maintair

Sensitive data (such as pre-anonymised datasets) will be stored offline on removable me

f. Which data should be retained, shared, and/or preserved? What is the long-term prese

All software tools will be shared and preserved indefinitely. Datasets from experiments

Long-term preservation of software tools is best managed through the use of public repo

g. How will you share the data? Are any restrictions on data sharing required?

Data will be shared using public platforms, with no restrictions on data sharing required.

h. Who will be responsible for data management? What resources will you require to del

The project proposer will take primary responsibility for the preservation and maintenan

Servers and data repositories are already in place to support the management of the data.

c. What documentation and metadata will accompany the data?

All software systems will be extensively documented as per typical software system documentation standards.

All experimental datasets and software systems will include basic descriptive and discovery metadata such as title, description, format and date.

...

# Phase 2: Acquisition / Organisation

- Source of Data:
  - Observations
    - Measure population movement in a region.
  - Gathered data
    - Use surveys to gather data about population movement.
  - Third party data
    - Get migration data from national census.
  - Simulated data
    - Use computer to simulate population movement.

# Data Cleaning

"This is not what I meant when I said 'we need better data cleansing!'"

- Corrections
  - Typing inconsistencies
  - Transcription errors
  - Translation errors
  - Processing errors
  - Human errors e.g., cut-and-paste
- Missing data
- Translation/transformation/standardisation

# Data Formats

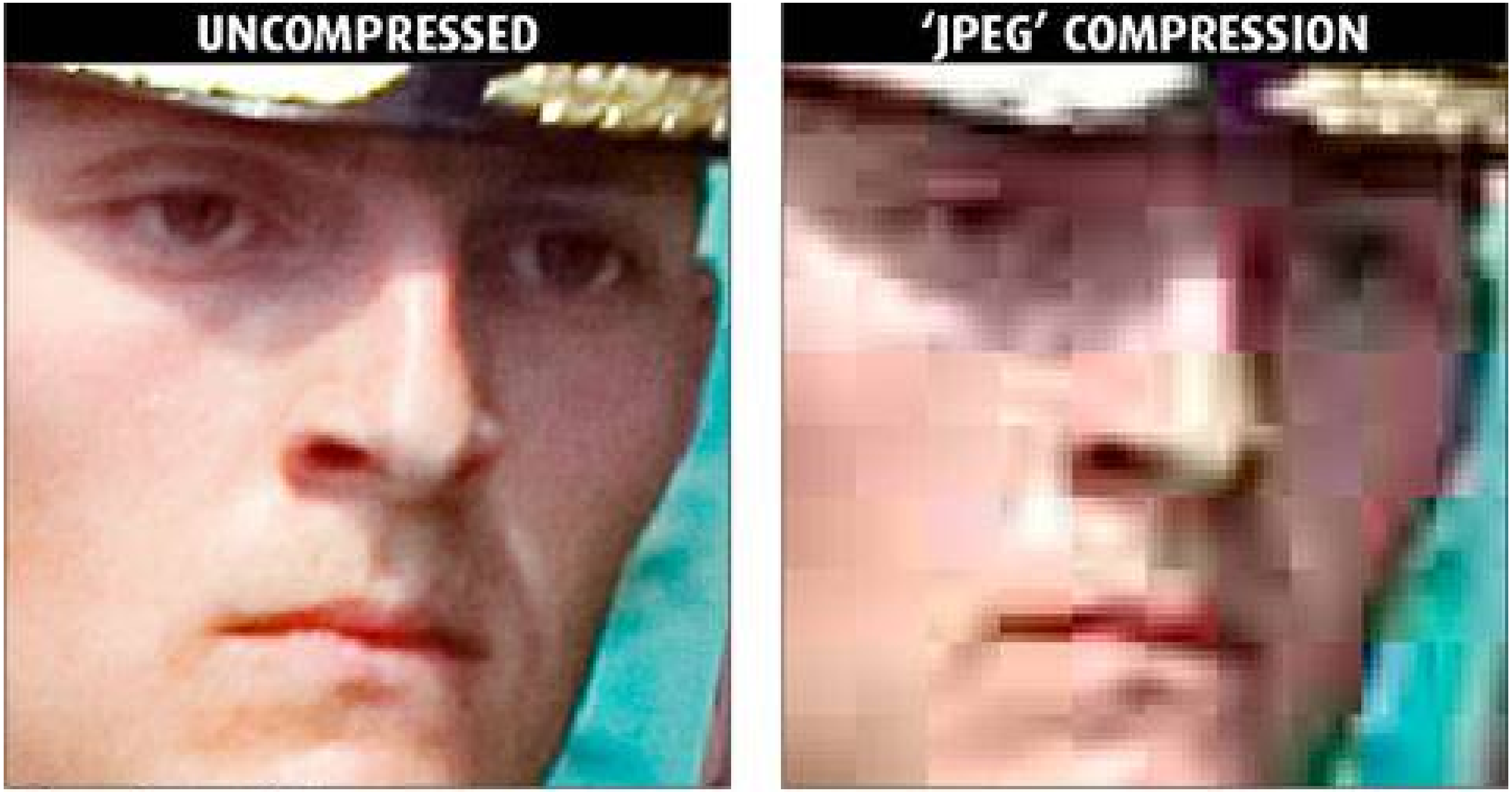
- What is a good data format to use?
  - Spreadsheet?
  - Word document?
  - Notepad text file?
  - Database?
  - XML document?
  - JPG image file?

# Data Formats - Guidelines

- Use a standard format.
- Use community guidelines.

- Syntax vs Semantics:
  - Syntax is the structure.
  - Semantics is the meaning.
  - What do we need standardisation for?

# Data Formats - Images



© Graeme Cookson / Shutha.org

# Data Formats – Levels of understanding

- Machine readable
  - Computers can load the data and decode syntax.
- Machine actionable
  - Computers can perform actions based on the data.
  - e.g., data provides enough details to determine what analysis should be performed.
- Machine executable
  - Computers can execute a dataset.
  - e.g., data includes code to perform analysis.

# Backups / Recovery (and some answers?)

- How many copies is enough?
  - 4? 8? 16?
- Where to save copies of data?
  - Some local, some online, some on removable media, some at work, some at home
- What to save?
  - Full data and changes, to reconstruct everything
  - Intermediate data and processed data

# Ethical Issues

- Anonymize all data before sharing/publishing.

- Full anonymity
  - Cannot link row of data to individual.
- K-anonymity
  - All identifiable data appears in K data items.

- Other issues? Potential for harm? Misuse?

# Phase 3: Sharing

- Why share data?
- When should data be shared?
  - Before analysis? Before publication? After publication? After project?
- How?
  - On-demand?
  - Website?
  - Attached to publication?
  - Data archive/repository?

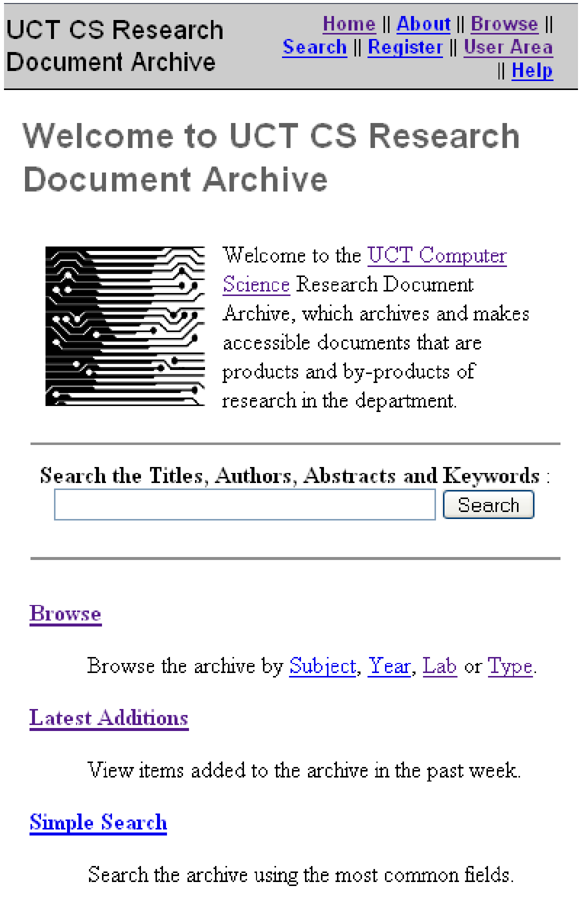
# Digital Repositories and Data Archives

# Some Definitions

- **Digital Repository** stores and provides access to digital objects.
- **Institutional Repository** provides access to data/papers/etc. Produced at a university.

- **Data Archive** is a digital repository that focuses on research data.
- **Open Access** is the principle of unrestricted access to research.

# Example Research Repository



UCT CS Research Document Archive

Welcome to UCT CS Research Document Archive

Welcome to the UCT Computer Science Research Document Archive, which archives and makes accessible documents that are products and by-products of research in the department.

Search the Titles, Authors, Abstracts and Keywords :

Browse

Browse the archive by Subject, Year, Lab or Type.

Latest Additions

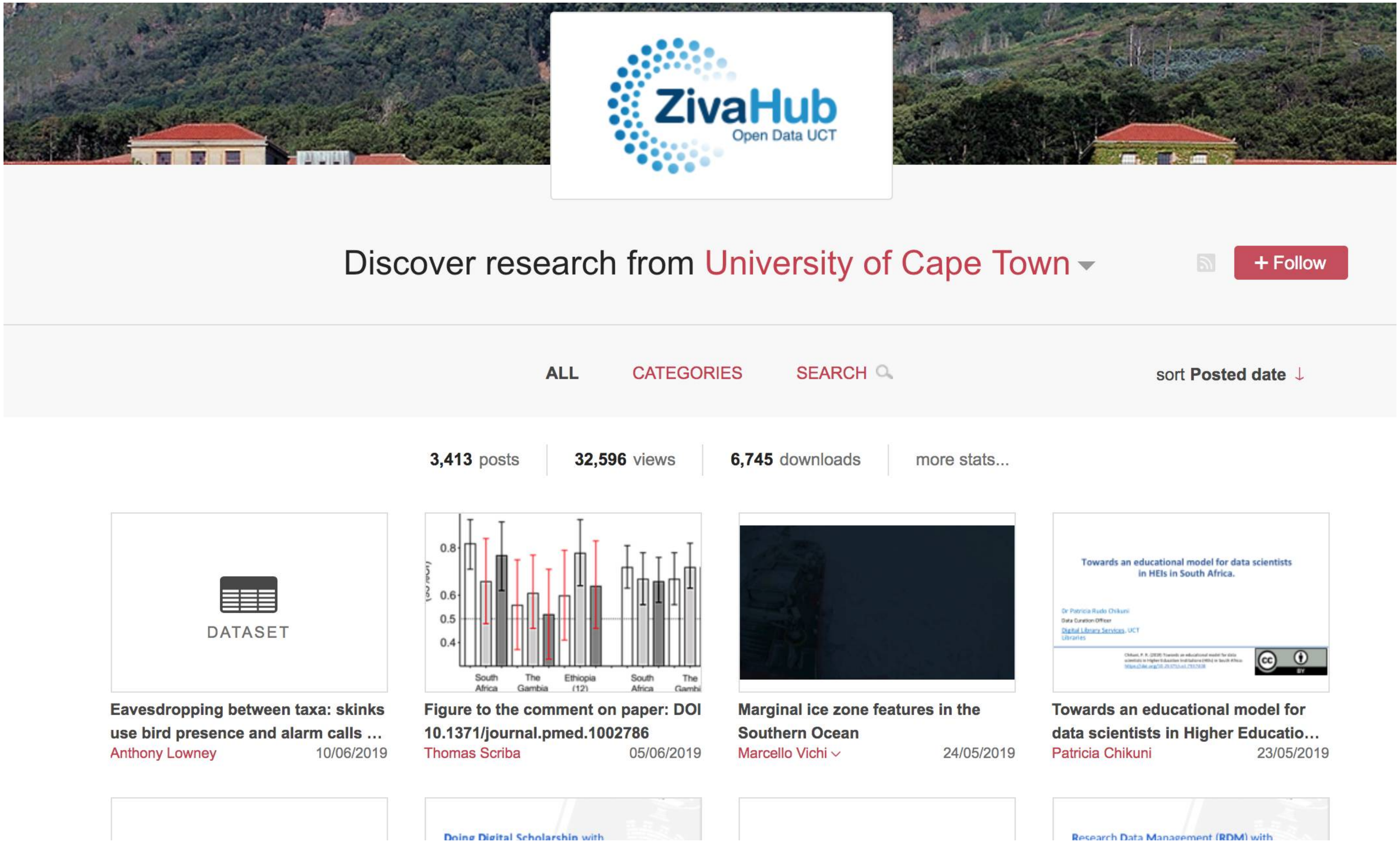
View items added to the archive in the past week.

Simple Search

Search the archive using the most common fields.

- Author self-submission
- Checking of submissions
- Archive-everything!
- UCT-CS-specific metadata and classification systems
- Hierarchical browsing
- Simple and fielded searching
- OAI-PMH compliance

# Example Data Archive



Discover research from University of Cape Town

+ Follow

ALL    CATEGORIES    SEARCH    sort Posted date ↓

**3,413** posts | **32,596** views | **6,745** downloads | more stats...

Towards an educational model for data scientists in HEIs in South Africa.

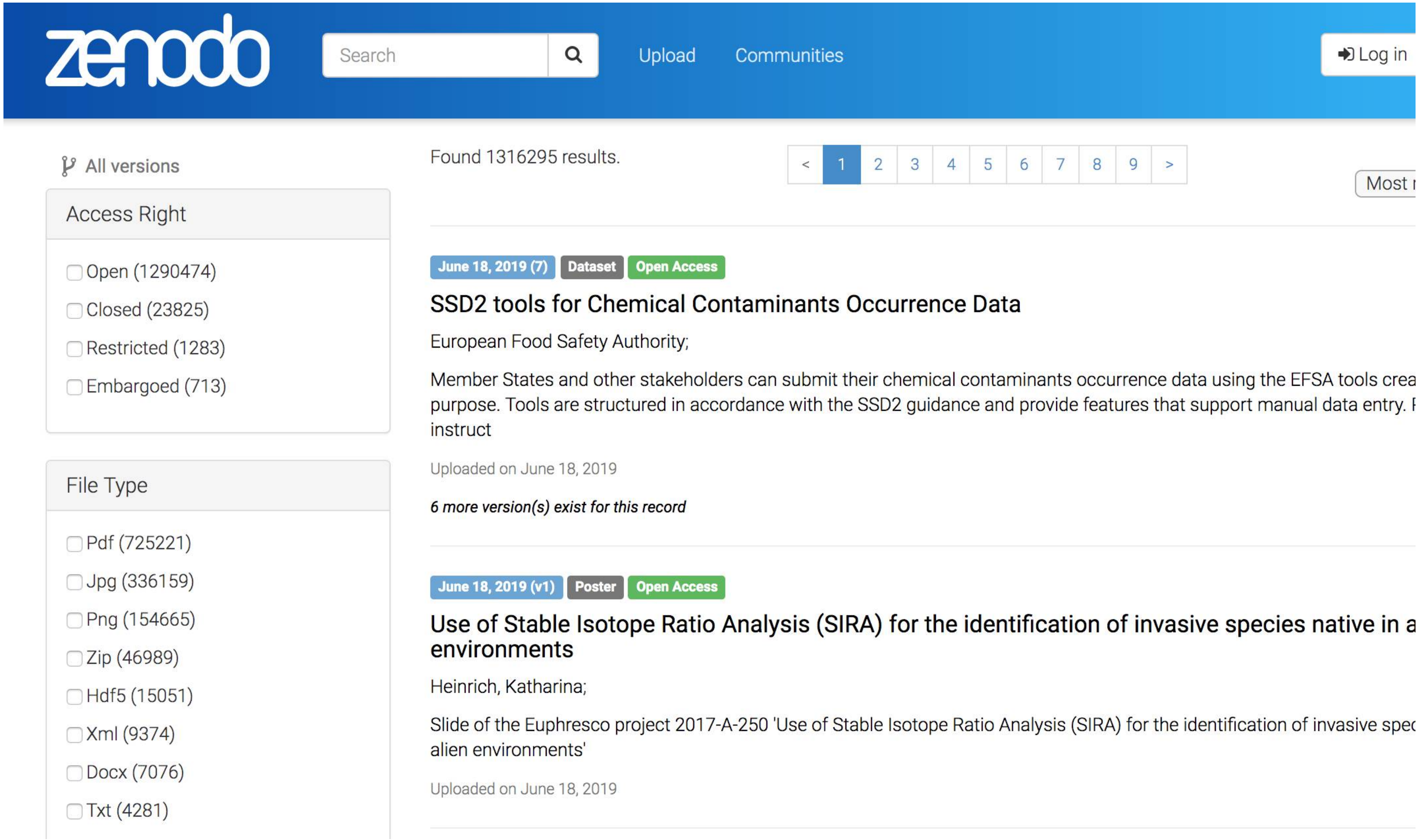
Dr Patricia Rudo Chikuni
Data Curation Officer
Digital Library Services, UCT Libraries

| | | | |
|---|---|---|---|
| **Eavesdropping between taxa: skinks use bird presence and alarm calls …** | **Figure to the comment on paper: DOI 10.1371/journal.pmed.1002786** | **Marginal ice zone features in the Southern Ocean** | **Towards an educational model for data scientists in Higher Educatio…** |
| Anthony Lowney 10/06/2019 | Thomas Scriba 05/06/2019 | Marcello Vichi 24/05/2019 | Patricia Chikuni 23/05/2019 |

Doing Digital Scholarship with

Research Data Management (RDM) with

# Example Data Archive



All versions

Found 1316295 results.

< 1 2 3 4 5 6 7 8 9 >

Most r

Access Right

□ Open (1290474)
□ Closed (23825)
□ Restricted (1283)
□ Embargoed (713)

File Type

□ Pdf (725221)
□ Jpg (336159)
□ Png (154665)
□ Zip (46989)
□ Hdf5 (15051)
□ Xml (9374)
□ Docx (7076)
□ Txt (4281)

June 18, 2019 (7) Dataset Open Access

SSD2 tools for Chemical Contaminants Occurrence Data

European Food Safety Authority;

Member States and other stakeholders can submit their chemical contaminants occurrence data using the EFSA tools crea purpose. Tools are structured in accordance with the SSD2 guidance and provide features that support manual data entry. F instruct

Uploaded on June 18, 2019

*6 more version(s) exist for this record*

June 18, 2019 (v1) Poster Open Access

Use of Stable Isotope Ratio Analysis (SIRA) for the identification of invasive species native in a environments

Heinrich, Katharina;

Slide of the Euphresco project 2017-A-250 'Use of Stable Isotope Ratio Analysis (SIRA) for the identification of invasive spec alien environments'

Uploaded on June 18, 2019

# Typical Services of Digital Repositories

- Store documents and metadata
- Search and Browse
- Submission of documents and metadata
- Moderation of content
- Access through Web interface
- Compliance with standards
- Enrichment: comments, reviews, etc.
- Linking with other systems

# Metadata

- **Metadata** refers to standardised descriptions of objects, digital or physical.

- Most digital repositories manipulate metadata records, which contain pointers to the actual data.
- The definition is fuzzy as metadata contains useful information as well and in some cases could contain all the data e.g., metadata describing a person.

# An Example of Metadata

- Metadata
  - name: Chalk
  - owner: Hussein
  - colour: white
  - size: 2.5cm
  - description: used to write on board
  - location: UCT lecture room 212
  - source: Waltons Stationers

# Types of Metadata

- Descriptive
  - title, author, type, format, …
- Structural
  - part, subpart, relation, child, …
- Administrative
  - location, identifier, submitter, …
- Preservation
  - resolution, capture device, watermark, …
- Provenance
  - source archive, previous version, source format, …

# Creating Metadata

- Follow metadata guidelines.
- Use terms from controlled vocabularies.
- Avoid duplication of information across fields.
- Use accepted standards for common elements.
  - e.g., ISO 8601 for dates
    - 2005-03-03 instead of 03/03/05

# Example: Dublin Core

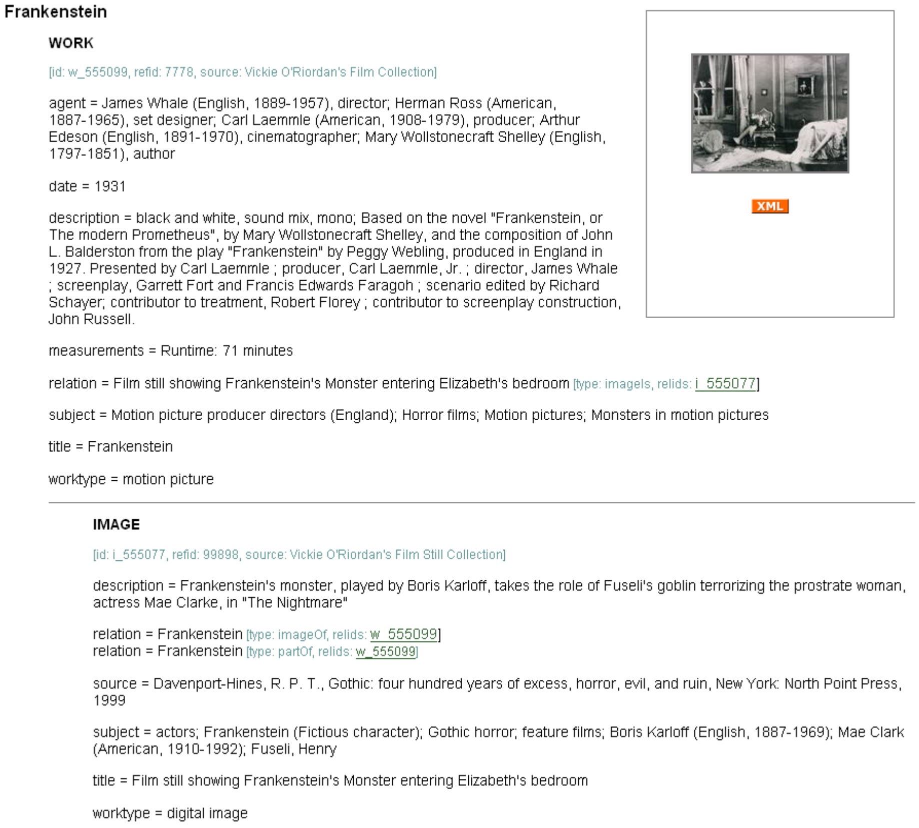
- Dublin Core is one of the most popular and simplest metadata formats.
- 15 elements with recommended semantics.
- All elements are optional and repeatable.

| | | |
|---|---|---|
| Title | Creator | Subject |
| Description | Publisher | Contributor |
| Date | Type | Format |
| Identifier | Source | Language |
| Relation | Coverage | Rights |

# DC Metadata in Valid Qualified XML

```
<oaidc:dc xmlns="http://purl.org/dc/elements/1.1/"
xmlns:oaidc="http://www.openarchives.org/OAI/2.0/oai_dc/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
   <title>02uct1</title>
   <creator>Hussein Suleman</creator>
   <subject>Visit to UCT </subject>
   <description>the view that greets you as you emerge from the tunnel under the
freeway - WOW - and, no, the mountain isnt that close - it just looks that way in
2-D</description>
   <publisher>Hussein Suleman</publisher>
   <date>2002-11-27</date>
   <type>image</type>
   <format>image/jpeg</format>
   <identifier>http://www.husseinsspace.com/pictures/200230uct/02uct1.jpg
</identifier>
   <language>en-us</language>
   <relation>http://www.husseinsspace.com</relation>
   <rights>unrestricted</rights>
</oaidc:dc>
```

# Example: VRA-Core



Frankenstein

WORK

[id: w_555099, refid: 7778, source: Vickie O'Riordan's Film Collection]

agent = James Whale (English, 1889-1957), director; Herman Ross (American, 1887-1965), set designer; Carl Laemmle (American, 1908-1979), producer; Arthur Edeson (English, 1891-1970), cinematographer; Mary Wollstonecraft Shelley (English, 1797-1851), author

date = 1931

description = black and white, sound mix, mono; Based on the novel "Frankenstein, or The modern Prometheus", by Mary Wollstonecraft Shelley, and the composition of John L. Balderston from the play "Frankenstein" by Peggy Webling, produced in England in 1927. Presented by Carl Laemmle ; producer, Carl Laemmle, Jr. ; director, James Whale ; screenplay, Garrett Fort and Francis Edwards Faragoh ; scenario edited by Richard Schayer; contributor to treatment, Robert Florey ; contributor to screenplay construction, John Russell.

measurements = Runtime: 71 minutes

relation = Film still showing Frankenstein's Monster entering Elizabeth's bedroom [type: imageIs, relids: i_555077]

subject = Motion picture producer directors (England); Horror films; Motion pictures; Monsters in motion pictures

title = Frankenstein

worktype = motion picture

IMAGE

[id: i_555077, refid: 99898, source: Vickie O'Riordan's Film Still Collection]

description = Frankenstein's monster, played by Boris Karloff, takes the role of Fuseli's goblin terrorizing the prostrate woman, actress Mae Clarke, in "The Nightmare"

relation = Frankenstein [type: imageOf, relids: w_555099]
relation = Frankenstein [type: partOf, relids: w_555099]

source = Davenport-Hines, R. P. T., Gothic: four hundred years of excess, horror, evil, and ruin, New York: North Point Press, 1999

subject = actors; Frankenstein (Fictious character); Gothic horror; feature films; Boris Karloff (English, 1887-1969); Mae Clark (American, 1910-1992); Fuseli, Henry

title = Film still showing Frankenstein's Monster entering Elizabeth's bedroom

worktype = digital image

# Some other metadata standards

- IMS Metadata Specification
  - Courseware object description.
- EAD
  - Library finding aids to locate archived items.
- METS
  - Descriptive, administrative and structural encoding for metadata of digital objects
- MODS
  - Richer than DC, subset of MARC21
- MPEG21-DIDL
  - Structural descriptions of complex multimedia objects

# Global Identifiers

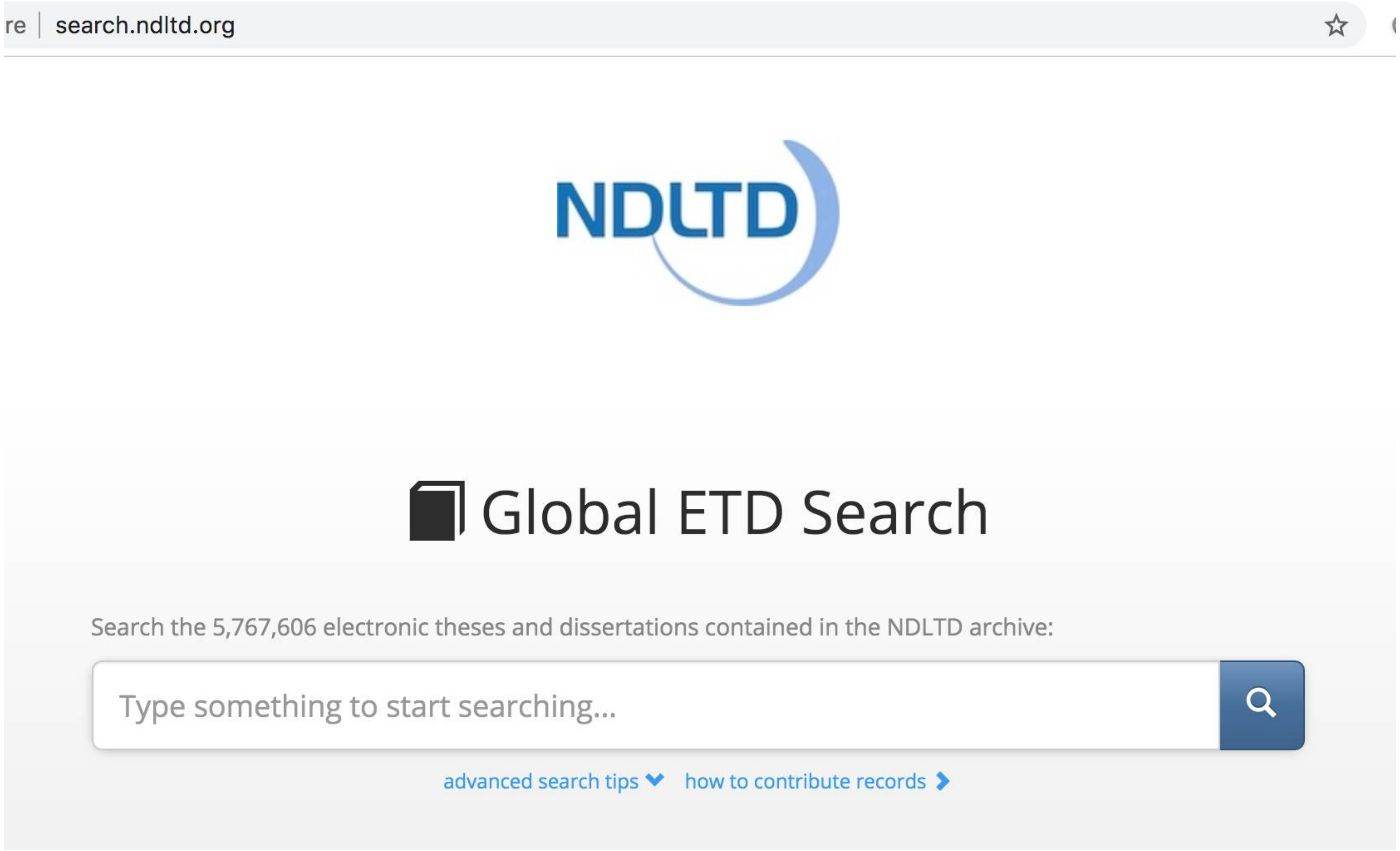
- Digital Object Identifier (DOI) is a unique reference to a dataset or publication in an archive.
  - e.g., https://doi.org/10.7927/H4Z31WKF

- Handled/generated/managed by repository.
  - You do not need to do anything!
    - Except, use the DOI and cite the DOI.

# Interoperability

- **Interoperability** is the ability of systems to work together.
- It allows connecting of repositories so researchers can find data/documents across repositories.
- 3 major aspects: standard metadata, standard network protocols and global identifiers.
- Google: no metadata, HTTP, URLs
  - Can we do better?

# Global ETD Search



re | search.ndltd.org

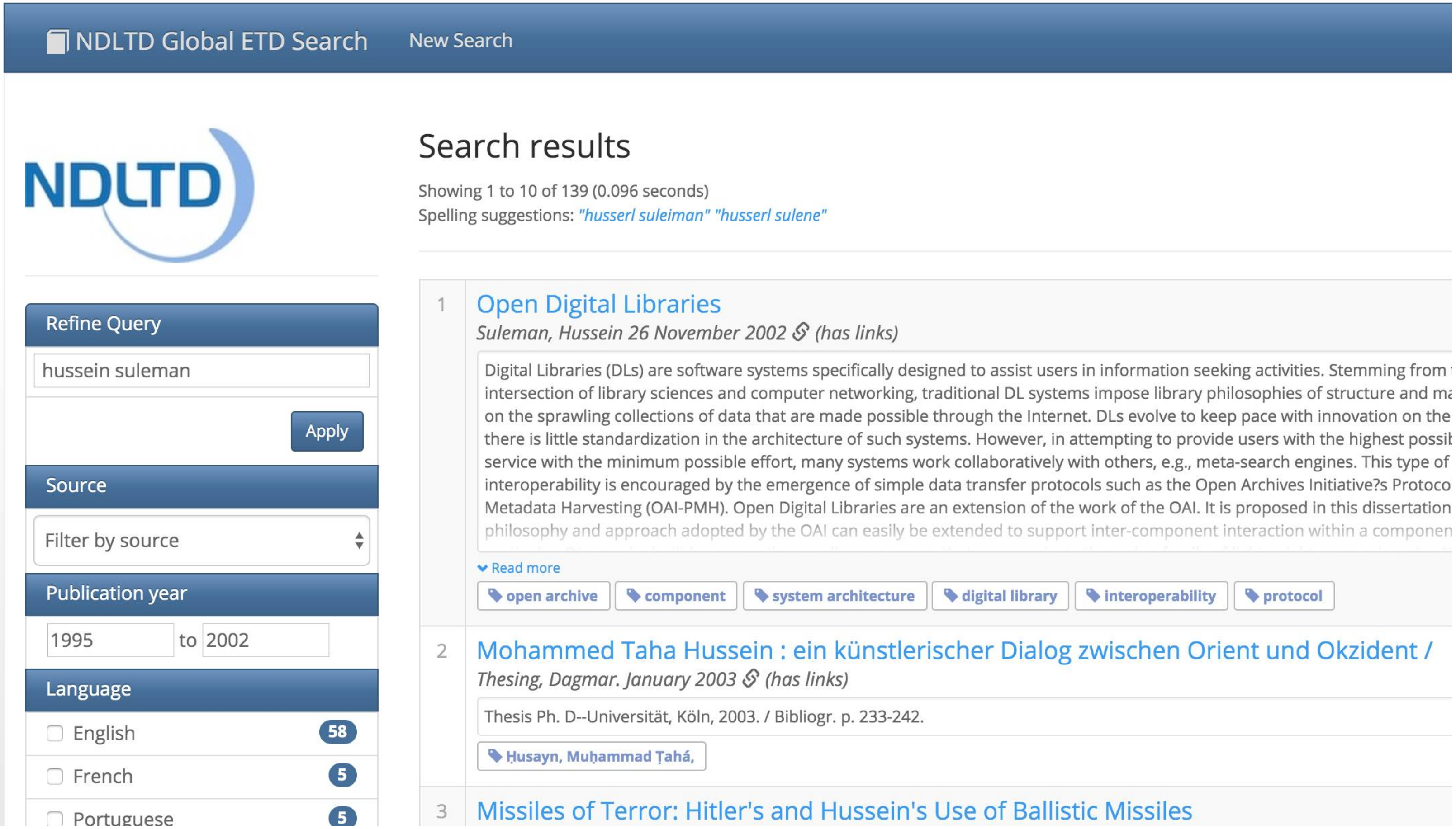
NDLTD

Global ETD Search

Search the 5,767,606 electronic theses and dissertations contained in the NDLTD archive:

Type something to start searching...
advanced search tips
how to contribute records

# Global ETD Search



NDLTD Global ETD Search    New Search

NDLTD

Refine Query

hussein suleman

Apply

Source

Filter by source

Publication year

1995 to 2002

Language

| | |
|---|---|
| □ English | 58 |
| □ French | 5 |
| □ Portuguese | 5 |

## Search results

Showing 1 to 10 of 139 (0.096 seconds)
Spelling suggestions: *"husserl suleiman" "husserl sulene"*

1 Open Digital Libraries
*Suleman, Hussein 26 November 2002 (has links)*

Digital Libraries (DLs) are software systems specifically designed to assist users in information seeking activities. Stemming from t
intersection of library sciences and computer networking, traditional DL systems impose library philosophies of structure and ma
on the sprawling collections of data that are made possible through the Internet. DLs evolve to keep pace with innovation on the
there is little standardization in the architecture of such systems. However, in attempting to provide users with the highest possib
service with the minimum possible effort, many systems work collaboratively with others, e.g., meta-search engines. This type of
interoperability is encouraged by the emergence of simple data transfer protocols such as the Open Archives Initiative?s Protoco
Metadata Harvesting (OAI-PMH). Open Digital Libraries are an extension of the work of the OAI. It is proposed in this dissertation
philosophy and approach adopted by the OAI can easily be extended to support inter-component interaction within a componen

Read more

open archive | component | system architecture | digital library | interoperability | protocol

2 Mohammed Taha Hussein : ein künstlerischer Dialog zwischen Orient und Okzident /
*Thesing, Dagmar. January 2003 (has links)*

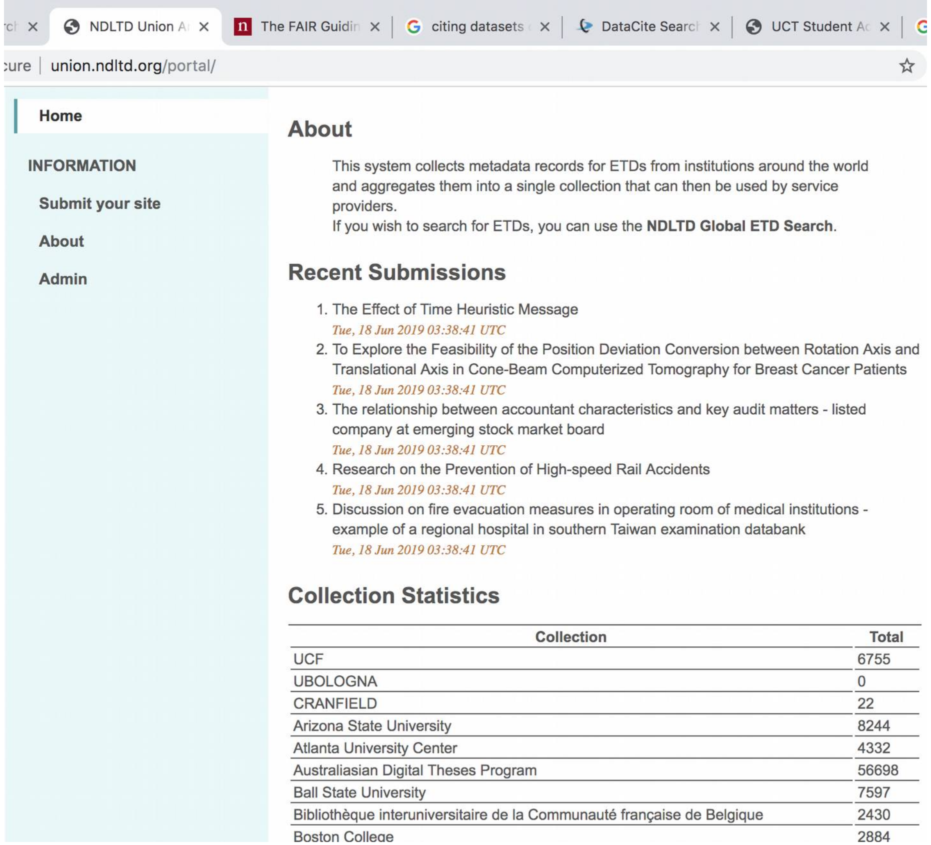
Thesis Ph. D--Universität, Köln, 2003. / Bibliogr. p. 233-242.

Ḥusayn, Muḥammad Ṭahá,

3 Missiles of Terror: Hitler's and Hussein's Use of Ballistic Missiles

# Global ETD Search: How It Works

- Standard metadata
- Efficient way to share metadata
  - OAI-PMH
- Collect updates from around the world every 12 hours
- Share through search engine



## About

This system collects metadata records for ETDs from institutions around the world and aggregates them into a single collection that can then be used by service providers.
If you wish to search for ETDs, you can use the **NDLTD Global ETD Search**.

## Recent Submissions

1. The Effect of Time Heuristic Message
   *Tue, 18 Jun 2019 03:38:41 UTC*
2. To Explore the Feasibility of the Position Deviation Conversion between Rotation Axis and Translational Axis in Cone-Beam Computerized Tomography for Breast Cancer Patients
   *Tue, 18 Jun 2019 03:38:41 UTC*
3. The relationship between accountant characteristics and key audit matters - listed company at emerging stock market board
   *Tue, 18 Jun 2019 03:38:41 UTC*
4. Research on the Prevention of High-speed Rail Accidents
   *Tue, 18 Jun 2019 03:38:41 UTC*
5. Discussion on fire evacuation measures in operating room of medical institutions - example of a regional hospital in southern Taiwan examination databank
   *Tue, 18 Jun 2019 03:38:41 UTC*

## Collection Statistics

| Collection | Total |
|---|---|
| UCF | 6755 |
| UBOLOGNA | 0 |
| CRANFIELD | 22 |
| Arizona State University | 8244 |
| Atlanta University Center | 4332 |
| Australiasian Digital Theses Program | 56698 |
| Ball State University | 7597 |
| Bibliothèque interuniversitaire de la Communauté française de Belgique | 2430 |
| Boston College | 2884 |

# Are we being FAIR? 1/2

- To be Findable:
  - F1. (meta)data are assigned a globally unique and persistent identifier
  - F2. data are described with rich metadata (defined by R1 below)
  - F3. metadata clearly and explicitly include the identifier of the data it describes
  - F4. (meta)data are registered or indexed in a searchable resource
- To be Accessible:
  - A1. (meta)data are retrievable by their identifier using a standardized communications protocol
  - A1.1 the protocol is open, free, and universally implementable
  - A1.2 the protocol allows for an authentication and authorization procedure, where necessary
  - A2. metadata are accessible, even when the data are no longer available

# Are we being FAIR? 2/2

- To be Interoperable:
  - I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
  - I2. (meta)data use vocabularies that follow FAIR principles
  - I3. (meta)data include qualified references to other (meta)data
- To be Reusable:
  - R1. meta(data) are richly described with a plurality of accurate and relevant attributes
  - R1.1. (meta)data are released with a clear and accessible data usage license
  - R1.2. (meta)data are associated with detailed provenance
  - R1.3. (meta)data meet domain-relevant community standards

# What Other Issues are There?

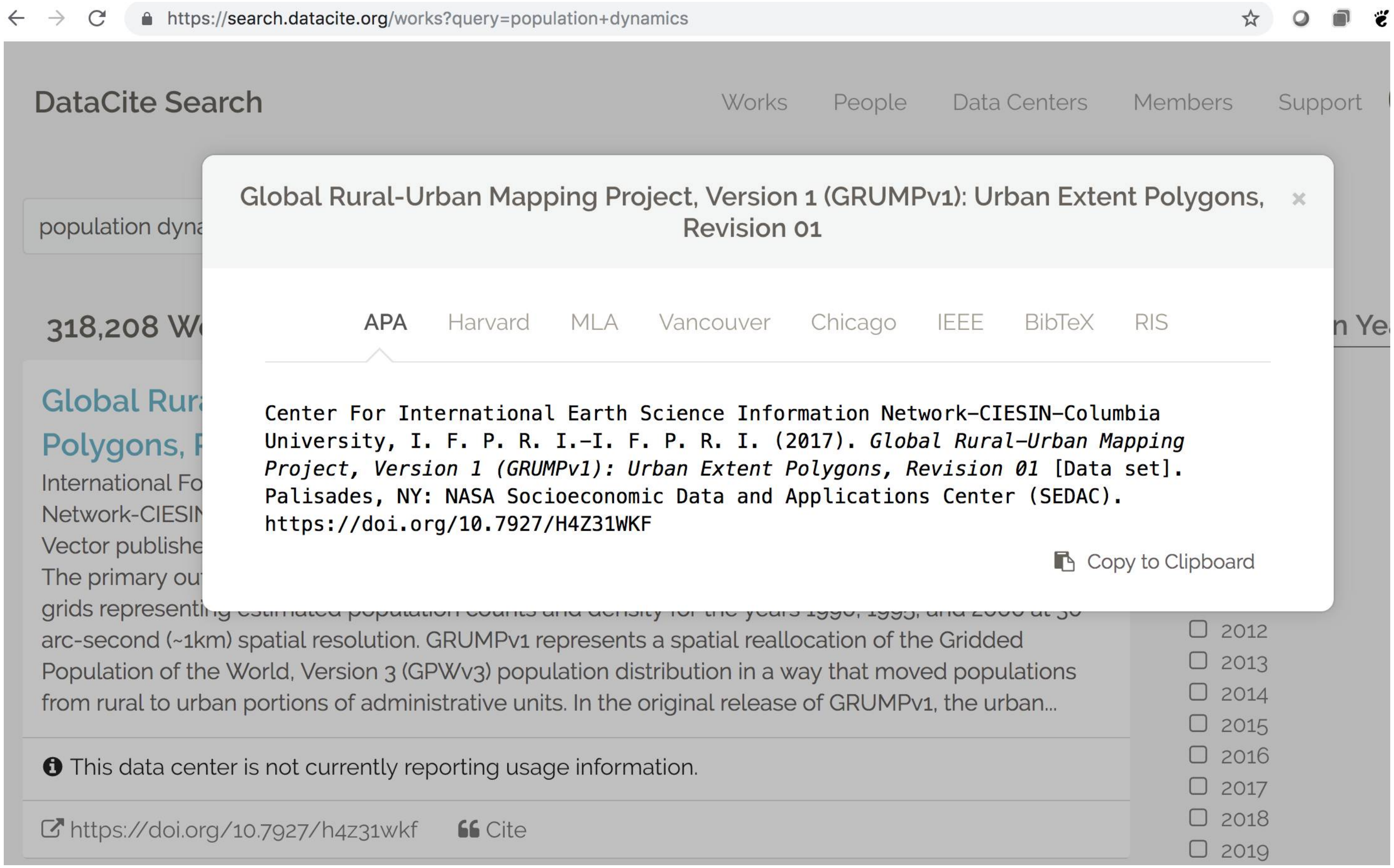
# Storing Data in the Cloud

- Cloud is a fancy name for Internet.
- Usually means data is stored in unspecified location far away and someone else will worry about it.
  - No free lunch!
  - Someone has to pay.
  - Remote access means it is easier to access for people in other countries, but maybe slower for you?

# Copyright and IP

- □ Who owns the data?
  - You?
  - University?
  - Data archive?
- □ What rules dictate how others may use it?
- □ Look at: Creative Commons licences

# Data citation



Center For International Earth Science Information Network-CIESIN-Columbia University, I. F. P. R. I.-I. F. P. R. I. (2017). *Global Rural-Urban Mapping Project, Version 1 (GRUMPv1): Urban Extent Polygons, Revision 01* [Data set]. Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC). https://doi.org/10.7927/H4Z31WKF

Global Rura
Polygons, R
International Fo
Network-CIESIN
Vector publishe
The primary ou
grids representing estimated population counts and density for the years 1990, 1995, and 2000 at 30 arc-second (~1km) spatial resolution. GRUMPv1 represents a spatial reallocation of the Gridded Population of the World, Version 3 (GPWv3) population distribution in a way that moved populations from rural to urban portions of administrative units. In the original release of GRUMPv1, the urban...

This data center is not currently reporting usage information.

# Data Stewards

- Who owns the data?
  - You?
  - University?
  - Data archive?
- Who is responsible in the long term for managing the data?

# Long Term Preservation

- Besides everything else, what more needs to be done?
  - Raise funding to manage data (remember data will only grow, never shrink).
  - Migrate data and systems.
  - Retire data.

# Costing Data Management

- How much will it cost:
  - To acquire
  - To store, backup
  - To archive in a repository
  - To migrate and preserve
- For:
  - a collection of 10MB of metadata and survey results?
  - a collection of 1TB of images?
  - a collection of 20TB of video?

# Be the cynic: Is there benefit to researchers?

- Do the established researchers do this?
- Do young researchers have the time?
  - Or should we chase publication/promotion/fame instead of following new research methods?
- Is someone counting data/publication archiving?

# Be the cynic: Decolonising Archives

- Should we as Africans share?
  - Is the move to Open Access the Western world trying to acquire our data? Who benefits more?
- Should we use the same methods/tools?
  - Does all this work for us just as easily?
- Local hosting of data?
  - Why not store the data where our access is easiest?

# that's all folks!