

ETD 2011 Cape Town

Data Curation Workshop

Focus of this presentation

- **The role of data curation**
 - What is research data?
 - Research data – related concepts
 - What is data curation?
 - The curation process
 - Drivers for data curation
 - Why data curation?
 - Roles and responsibilities
- **Data curation implementation**
 - Barriers and challenges
 - What should be in place?
 - Where to start?

Focus of this presentation (cont.)

- What to do?
 - Engage with data producers
 - Develop facilitating workflows
 - Implement a suitable technology platform
 - Develop data curation policies
 - Create and pilot service models
 - Do change management

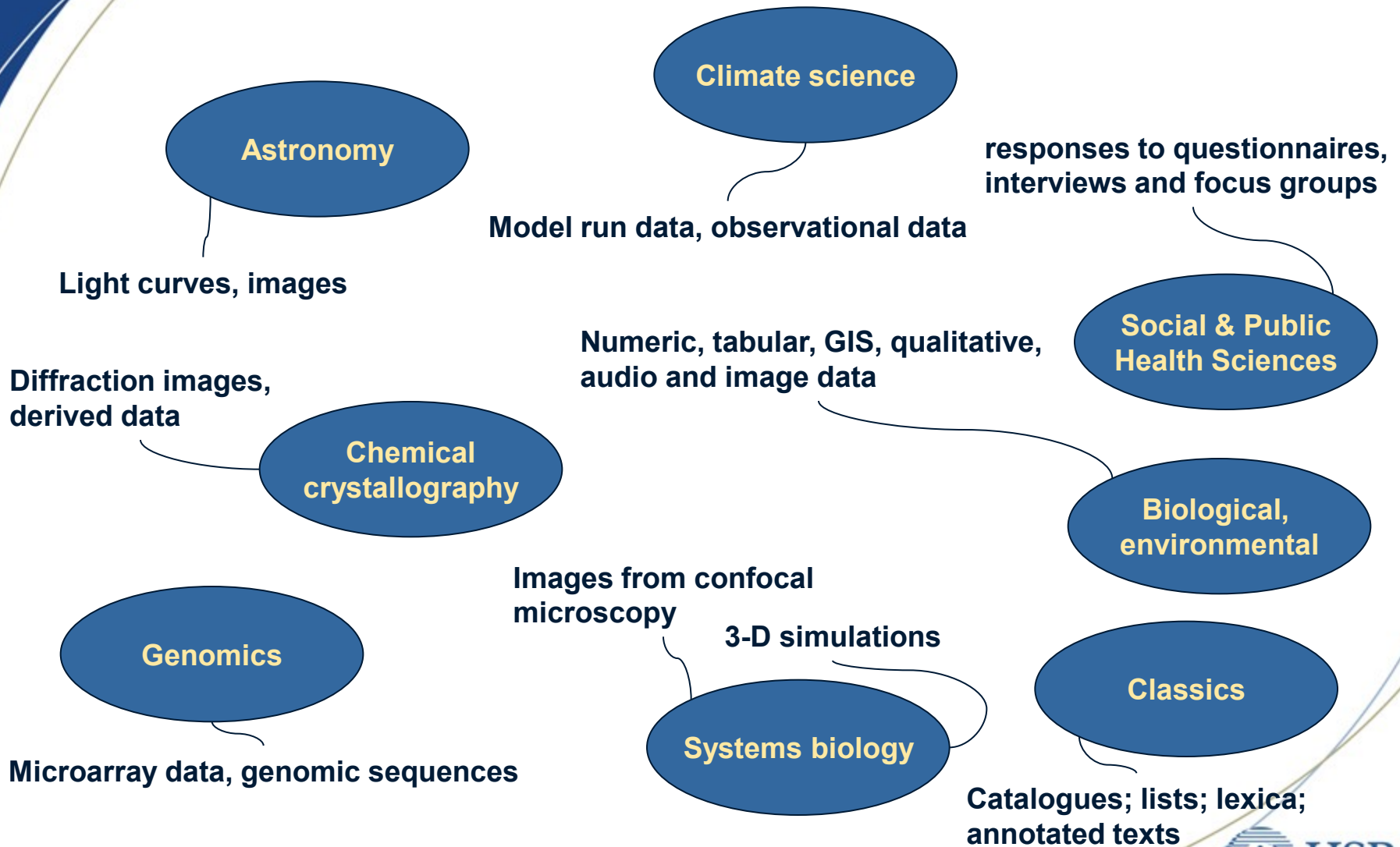
What is research data?

Collections of **records or measurements** used by researchers to undertake their research or provide an evidential record of their research

Attributes

- Digital
- Heterogeneous
- Contextual
- Valuable

Research data – Examples



Research data – Related concepts

- **Raw / micro** data vs. **aggregated / summarized / derived** data (Micro data are data in which every record is at the unit of analysis level and all records must be added up to get the totals for each data item)

1	uniqueid	prov	hnum	pers16	hhper
2	1	1	1	5	5
3	3	1	1	2	5
4	4	1	1	4	4
5	5	1	2	2	3
6	6	1	1	4	4
7	7	1	1	4	5

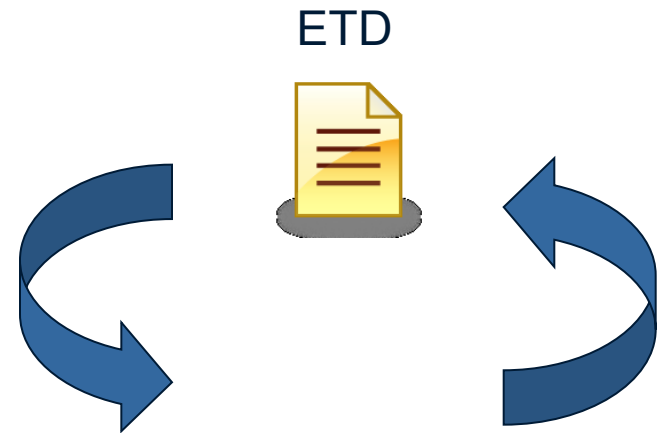


	Business	Government	Higher education
OCCUPATION			
Researchers	12 626	253	2 214
Technicians directly	827	62	1 335
Other personnel directly	2 314	123	2 325

- **Primary** data (created for the first time and there is no previous source available, did not exist before) vs. **secondary** data (readily available data)

ETDs and Data

- Digital objects
 - Metadata
 - Descriptive
 - Provenance
 - Access
 - Use
 - Preservation
- ETD and Data relationship
 - Context, provenance
 - Evidence



1	uniqueid	prov	ihnum	pers16	hhper
2	1	1	1	5	5
3	3	1	1	2	5
4	4	1	1	4	4
5	5	1	2	2	3
6	6	1	1	4	4
7	7	1	1	4	5

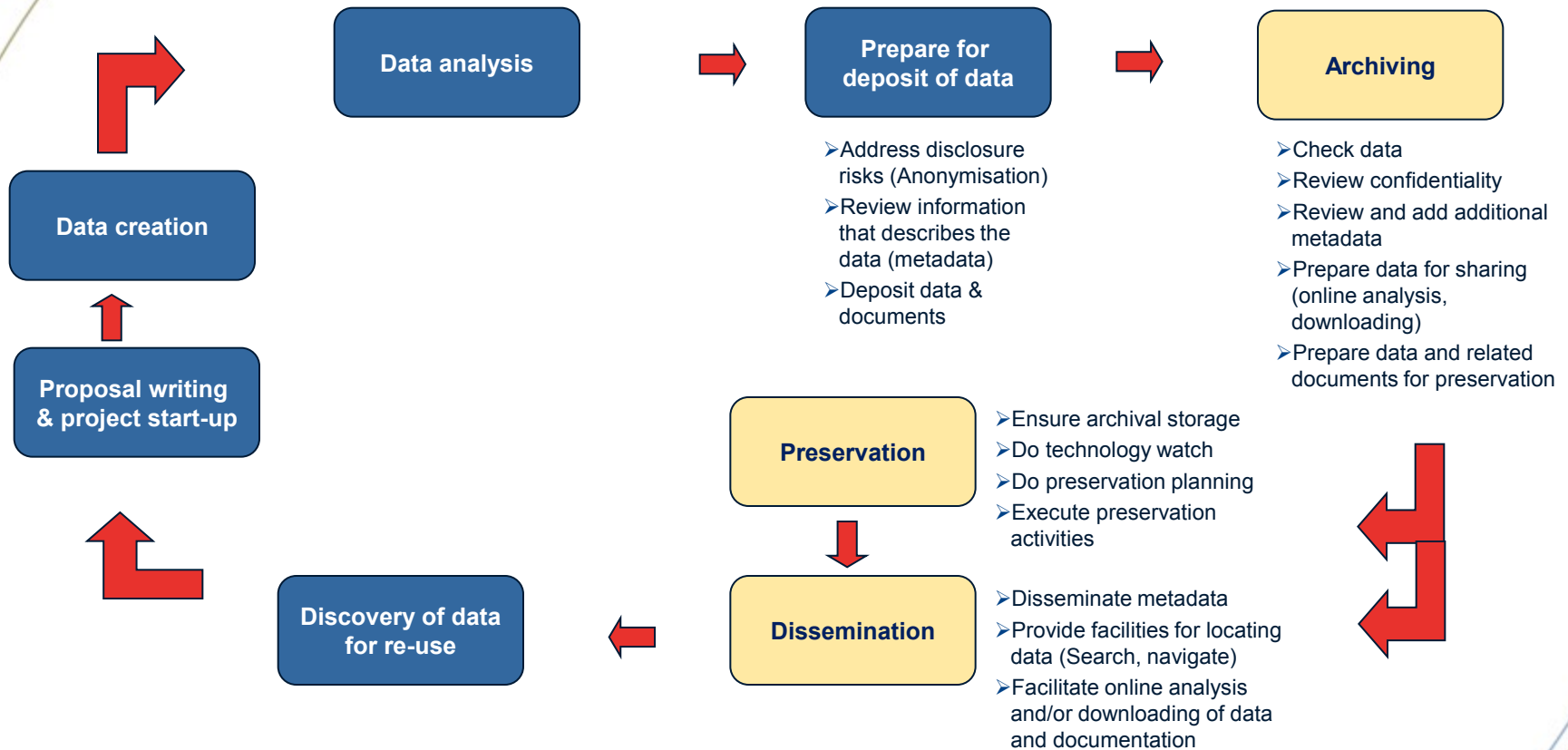
Primary data

Defining data curation

The activity of, **managing** and **promoting the use** of data from its point of creation, to ensure it is fit for contemporary purpose, and available for discovery and re-use

Lord, P. & Macdonald, A. (2003). *Data curation for e-Science in the UK: an audit to establish requirements for future curation and provision*. e-Science Curation Report prepared for: The JISC Committee for the Support of Research (JCSR)
http://www.jisc.ac.uk/uploaded_documents/e-ScienceReportFinal.pdf

Data curation – A process



(Based on ICPSR Guide to Social Science Data Preparation and Archiving, 2009:5)

Drivers for data curation

- Technology obsolescence of digital objects, media, etc.
- Open access to data from public funded research
- Best practices
 - Accountability - Future requirement for access to data when publishing in accredited journals/publications
 - Potential for creating 'new' knowledge from existing data is recognised
 - Institutional asset management
- Shift towards e-research and participation in the global research arena
- Promoting the institution, research group or individual

Why data curation?

- Contribute to advance science
- Contribute to data quality improvement
- Ethical management of research data
- Cost effective
- Data is intellectual currency – valuable, competitive edge
- Minimise data-at-risk

Roles and responsibilities

- Scientist
- Institution
- Data centre
- User
- Funder
- Publisher
- Government

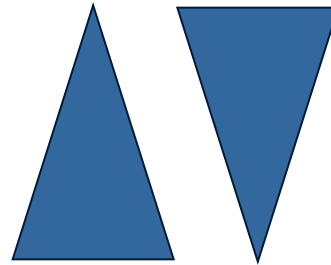
Lyon, L. (2007). *Dealing with Data: Roles, Rights, Responsibilities and Relationships Consultancy Report*. Paper delivered at the JISC Digital Repositories Conference, Manchester, June.

http://www.jisc.ac.uk/media/documents/programmes/digitalrepositories/dealing_with_data_report.pdf:

Digital repository types

- Three levels of collections

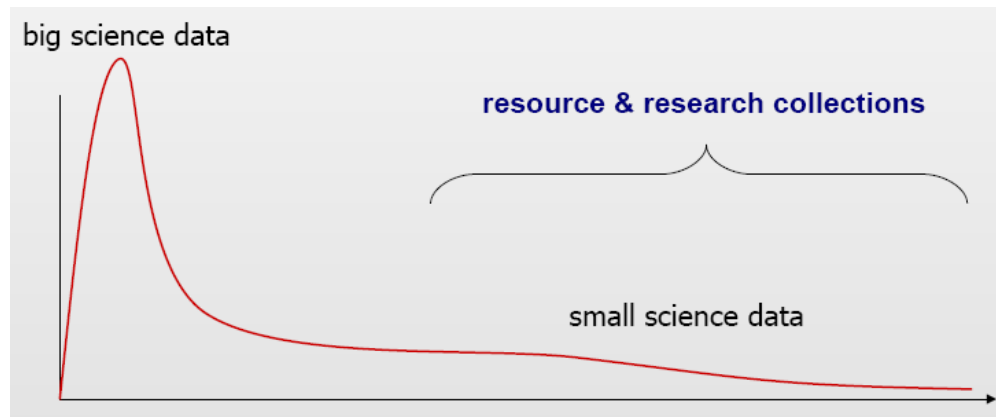
- Research
- Resource
- Reference



ETD Collections?

National Science Board (2005). *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century*. National Science Board, USA.

<http://www.nsf.gov/pubs/2005/nsb0540/>



Palmer, C.L. (2008). *Contouring Curation for Disciplinary Difference and the Needs of Small Science*. Sun PASIG Fall 2008 Meeting. 26 October.

Barriers and challenges

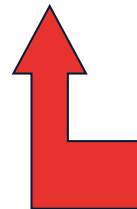
- Existence of data
- Nature of the data
- Technical
- Policies
- Research culture
- Cost
- Confidentiality

What should be in place?

- Collection development
- Depositor support
- Digital object management
- Promotion of secondary data discovery and use



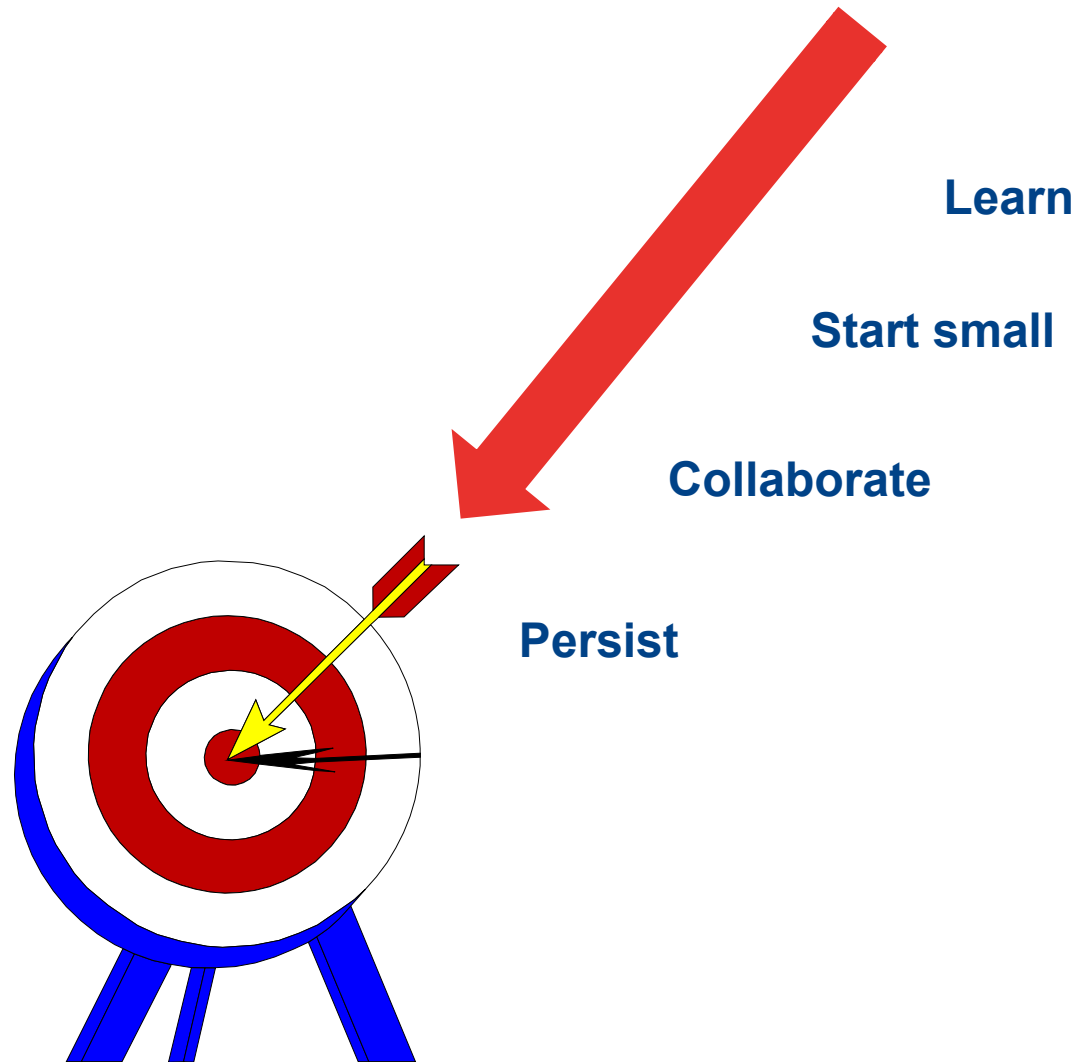
- Policies
- Procedures
- Technology
- Capacity
- Financial resources



Models, standards

- OAIS
- TRAC

Can this become a reality?



Where to start?

- **Bottom-up**
 - What is research and data all about?
 - What are researcher expectations?
 - How can researchers be supported?
 - How do researchers manage and document data?
 - What are the domain specific ontology, thesaurus, or metadata scheme?
 - Which data repository services do the researchers use?
- **Top-down**
 - What is the organisations priorities?
 - What is aim of the repository?
 - What is organisational commitment, policies, procedures, resources?

Where to start? (cont.)

- **Outside-in**
 - What are the policies of the top journals in the domain?
 - What are legislative, regulatory requirements?
 - What is the research culture like?
 - What are the policies of funders of research in the domain?
- **Inside-out**
 - What resources, technology, capacity can be used?



Data Repository Prospectus

What to do?

- Engage with data producers
- Develop facilitating workflows
- Implement a suitable technology platform
- Develop data curation policies
- Create and pilot service models
- Do change management

Walters, T.O. (2009). Data Curation Program Development in U.S. Universities:
The Georgia Institute of Technology Example. *The International Journal of Digital Curation* 3(4):83-92.
www.ijdc.net/index.php/ijdc/article/view/136/153

Engage with data producers

Data interview



Data management plan (DMP)

A DMP describes the data that will be authored as well as how the data will be **managed** and made **accessible** throughout its lifetime.

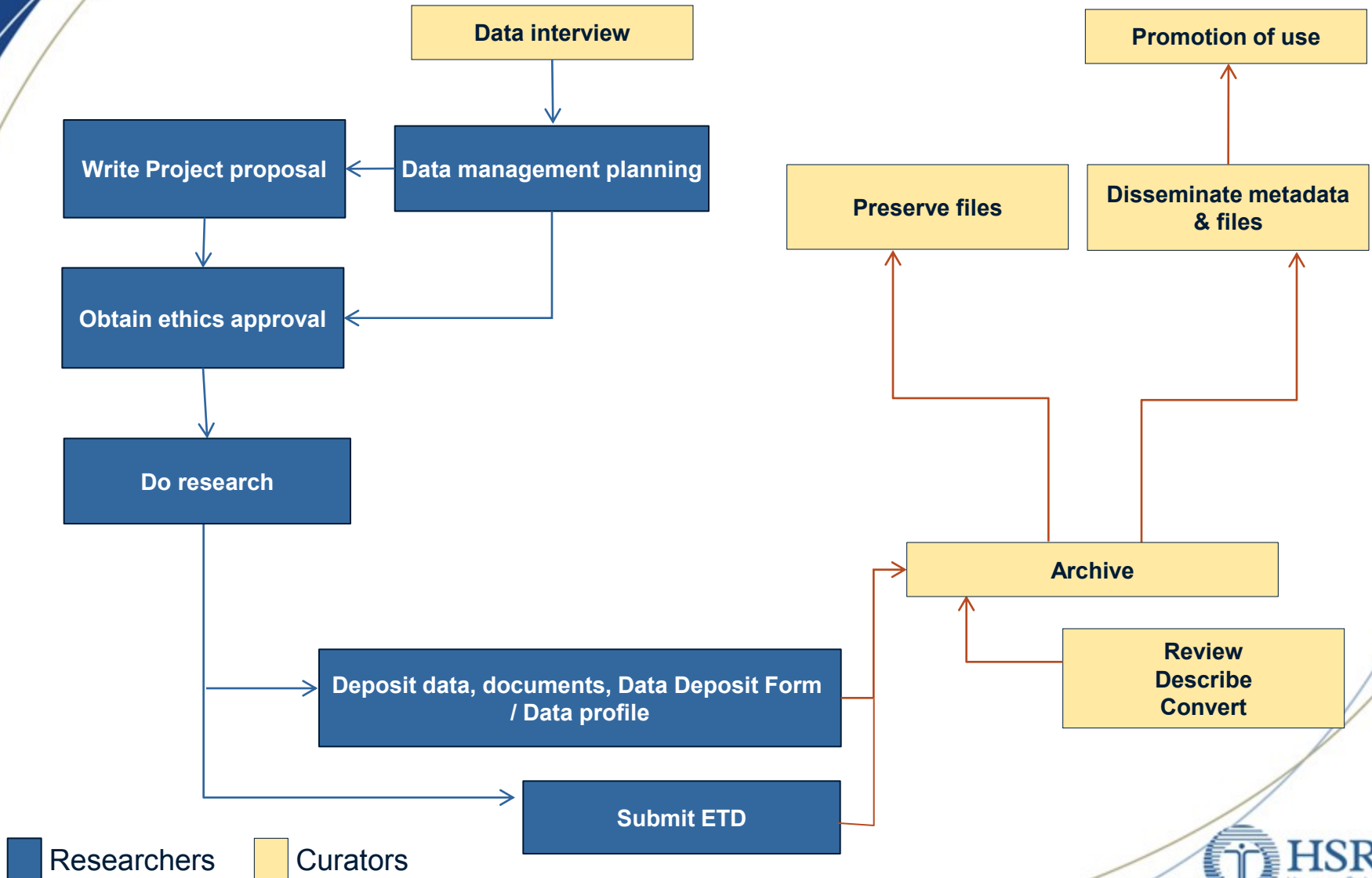
Data management planning

- The contents of a data management plan (DMP)
 - The types of data to be authored
 - The standards that would be applied for format, metadata content, etc.
 - Provisions for archiving and preservation
 - Access requirements and provisions and
 - Plans for eventual transition or termination of the data collection in the long term future

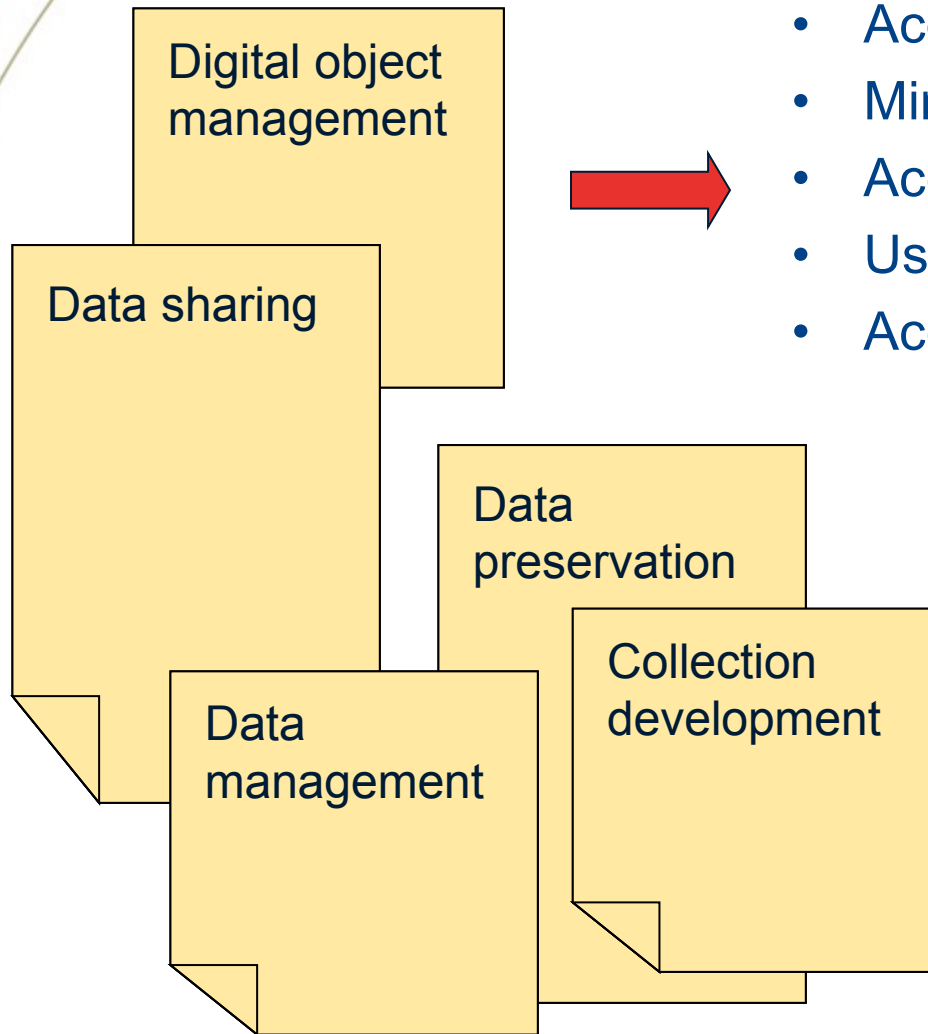
Develop workflows

- Part of research / data life cycle / Embedded within the workflows of particular research communities
 - Data management planning
 - Data documentation
 - Ethics approval
- Cater for lifecycle steps that are
 - essential in terms of digital object management
 - most critical to an institution's scientists
- Tools
 - Consider collaboration environments as an implementation mechanism

Process flow - Example



Develop data curation policies



- Acquisition
- Minimally required metadata
- Acceptable digital formats
- Use and re-use parameters
- Access regulation

Walters, T.O. (2009). Data Curation Program Development in U.S. Universities: The Georgia Institute of Technology Example. *The International Journal of Digital Curation* 3(4):83-92.
www.ijdc.net/index.php/ijdc/article/view/136/153

Create and pilot service models

- For depositors
 - Guidance, support
- For digital objects
 - Storage, description, access management, preservation
- For data users:
 - Data discovery (promotion)
 - Data use (training)

Implement a technology platform

- Functionality
 - Ingest, describe, store, access, share, reuse, preserve
- Metadata
- Access (including monitoring of data use)
- Storage
- Connectivity
- Security and disaster recovery
- Preservation
- Persistence

Technology platform - Example

Institutional repository of the University of Illinois at Urbana-Champaign

IDEALS

Illinois Digital
Environment for
Access to Learning
and Scholarship



Login | Non-Illinois Login

Search



Search IDEALS



This Community

Advanced Search

Browse

IDEALS

Titles

Authors

Subjects

Date

Communities

This Community

Titles

Authors

Subjects

Date

My Account

[IDEALS Home](#) ▶ [College of Agricultural, Consumer and Environmental Sciences \(ACES\)](#) ▶ [Dept. of Crop Sciences](#)

▶ [Illinois long-term selection experiment for oil and protein in corn](#)

Illinois long-term selection experiment for oil and protein in corn

The history of the Illinois long-term selection experiment for oil and protein in corn was reviewed extensively in the proceedings of a symposium on long term selection (Dudley and Lambert, 2004). Details of selection intensity, breeding procedure, and methods of chemical analysis are found in those proceedings. Briefly, the experiment was started by C.G. Hopkins in 1896 (Hopkins, 1899). The open-pollinated corn cultivar Burr's White was the founder population. Four selected strains were established: Illinois High Oil (IHO), Illinois Low Oil (ILO), Illinois High Protein (IHP), and Illinois Low Protein (ILP). Selection in each strain was in the direction indicated by the name of the strain (i.e., IHO was selected for high oil concentration in the kernel). Mass selection was used with a selection intensity of approximately 1 out of 5 for most of the strains.

Browse Collections of Items

strai  [Data and Publications from the Illinois long-term selection experiment for oil and protein in corn](#)

Browse by

- [Titles](#)
- [Authors](#)
- [Subjects](#)
- [Date](#)

Data and Publications from the Illinois long-term selection experiment for oil and protein in corn

This collection contains the data files and, where copyright allows, the published research for the Illinois long-term selection experiment for oil and protein in corn. Also included is a read me file and a list of all refereed publications resulting from this research.

The data files included here are:

- **Means by year and generation of the oil and protein concentration measured each year (1896-2004)** : This file contains the means by year and generation of the oil and protein concentration measured each year during the experiment. Also included are the generation numbers for each strain.
- **Raw data from each ear analyzed each year of the Illinois long-term selection experiment for oil and protein in corn 1896-2004)** : This file contains the data from each ear analyzed each year of the experiment. These are the raw data from the experiment.
- **Number of ears analyzed, the number of ears saved, and the selection differentials for the forward selection strains (1896-2004)** : This file contains the number of ears analyzed, the number of ears saved, and the selection differentials for the forward selection strains (IHP, ILP, IHO, and ILO).
- **Number of ears analyzed, the number of ears saved, and the selection differentials for the reverse strains (1947-2004)** : This file contains the number of ears analyzed, the number of ears saved, and the selection differentials for the reverse strains (RHP, RLP, RLP2, RHO, RLO, and SHO).
- **Values obtained for protein in the strains selected for oil and the values for oil obtained for the strains selected for protein each generation (1896-2004)** : This file contains values obtained for protein in the strains selected for oil (IHO, ILO, etc.) and the values for oil obtained for the strains selected for protein (IHP, ILP, etc.) each generation.

Browse by

- **Titles**
- **Authors**
- **Subjects**
- **Date**

Data File : Means by year and generation of the oil and protein concentration measured each year (1896-2004)

Bookmark or cite this item: <http://hdl.handle.net/2142/3526>

Files in this item



File	Description	Format
 GENMNS&GEN.SAS (16KB)	SAS file	Unknown
 GENMNS&GEN.txt (16KB)	Plain text	Text file

Title:	Data File : Means by year and generation of the oil and protein concentration measured each year (1896-2004)
Alternative Title:	GENMNS&GEN (file name)
Subject(s):	Corn Corn Composition
Abstract:	This file, part of the Illinois long-term selection experiment for oil and protein in corn (https://hdl.handle.net/2142/3524), contains the means by year and generation of the oil and protein concentration measured each year during the experiment (1896-2004). Also included are the generation numbers for each
	Rights Information: If data is used, the University of Illinois at Urbana-Champaign and the Illinois long-term selection experiment for oil and protein in corn must be acknowledged.
	Date Available in IDEALS: 2008-02-01
Issue Date:	200
Publisher:	Dej
	Relation: See the Read Me file (http://hdl.handle.net/2142/3527) for more information.
Genre:	Data

Read me file for data files for the Illinois long-term selection experiment for oil and protein in corn

Bookmark or cite this item: <http://hdl.handle.net/2142/3527>

Files in this item

File	Description	Format
 READ ME FILE.doc (26KB)	Read Me file for data files	Microsoft Word
Other Available Formats		
 READ ME FILE.doc.pdf (70KB)	Automatically converted using OpenOffice.org	PDF

Title:	Read me file for data files for the Illinois long-term selection experiment for oil and protein in corn
Author(s):	Dudley, John
Subject(s):	Corn Corn Composition Read me file
Abstract:	Read me file for the data set associated with the Illinois long-term selection experiment for oil and protein in corn. This file contains historical information and description of the data files available.
Issue Date:	2007
Publisher:	University of Illinois at Urbana-Champaign
Genre:	Data Other
Type:	Dataset / Spreadsheet Text
Language:	English
URI:	http://hdl.handle.net/2142/3527
Date Available in IDEALS:	2008-02-01
Relation:	References http://hdl.handle.net/2142/3526

Do change management

- Do advocacy
- Demonstrate success and value incrementally
- Obtain executive custody
- Do training and provide support
- Build relationships

The background of the slide is a black and white photograph of four hands of different skin tones clasped together in a circle, symbolizing unity and collaboration. The hands are positioned in the center of the frame, with the fingers interlaced. A red and orange beaded bracelet is visible on the left hand. The image is framed by a blue arc on the top left and a green arc on the bottom right.

Thank you

**Building the bridge between
research, policy and action**

Lucia Lötter
lloetter@hsrc.ac.za
HSRC Research Data Centre
www.hsrc.ac.za