

Research data – The new e-Resource

**Dr Martie van Deventer (CSIR)
&
Dr Lucia Lötter (HSRC)**

Presentation overview

In terms of research data, this presentation will address what could be regarded as

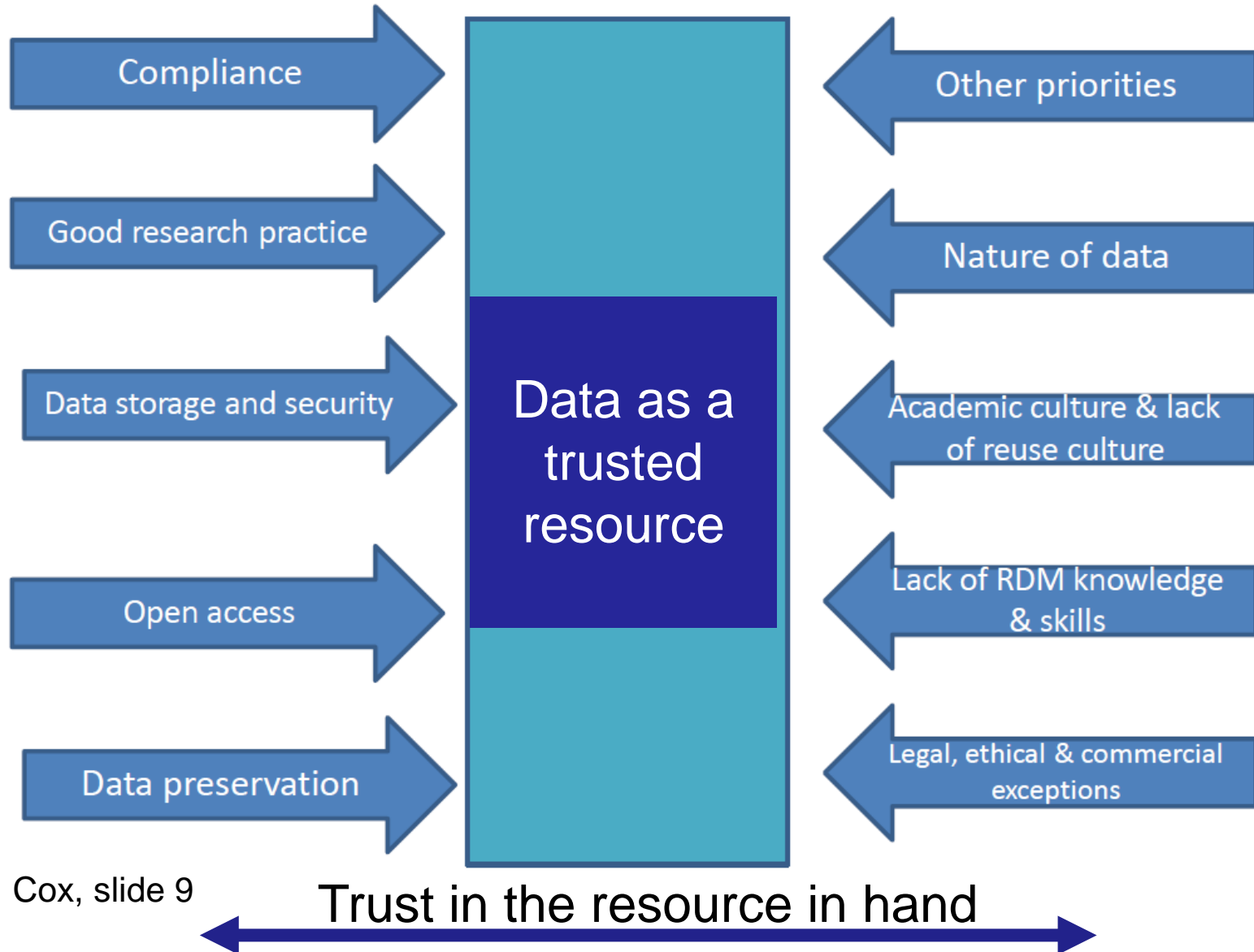
- the focus area for libraries,
- the skills required to manage data and then focus on the
- process required to ensure that the data resource is a product with a standard of quality that is clearly understood by all.



Why is it so difficult to manage research data as a resource?



Force field analysis of RDM

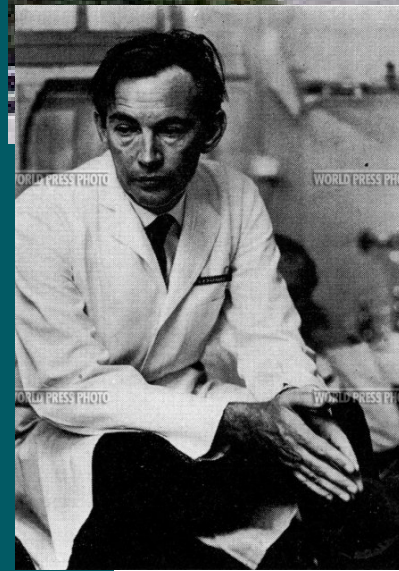
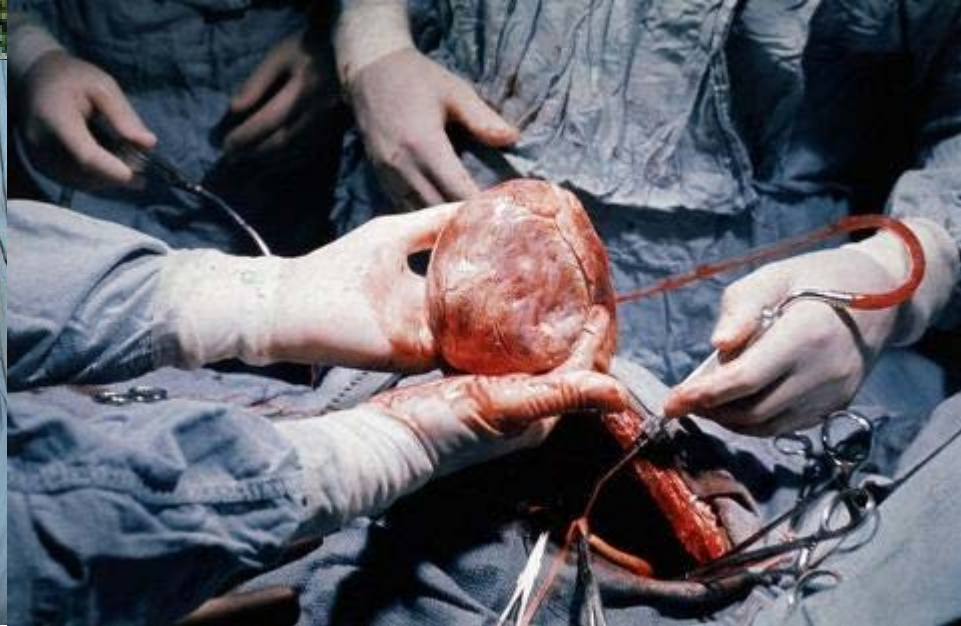
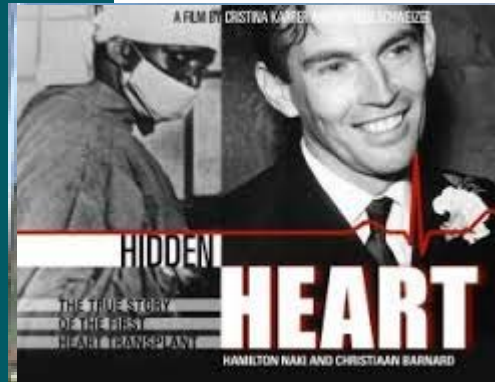


Source: Cox, slide 9

What is the library/ librarian's role in establishing a trusted eResource?

(1) Understand what it is you need to do!

Where is this data? How is it being re-used?



This is where the action is!

Closed environment

Highly qualified team – collaborating & well prepared

Practiced and focussed procedures

Sophisticated instruments

Emergency management

Record keeping

Proper 'cleaning up' procedures

... but it is often the aftercare that determines how the patient will live ...

Open (but controlled) environment

Qualified **teams** – varying levels & expertise

Practiced and focussed procedures

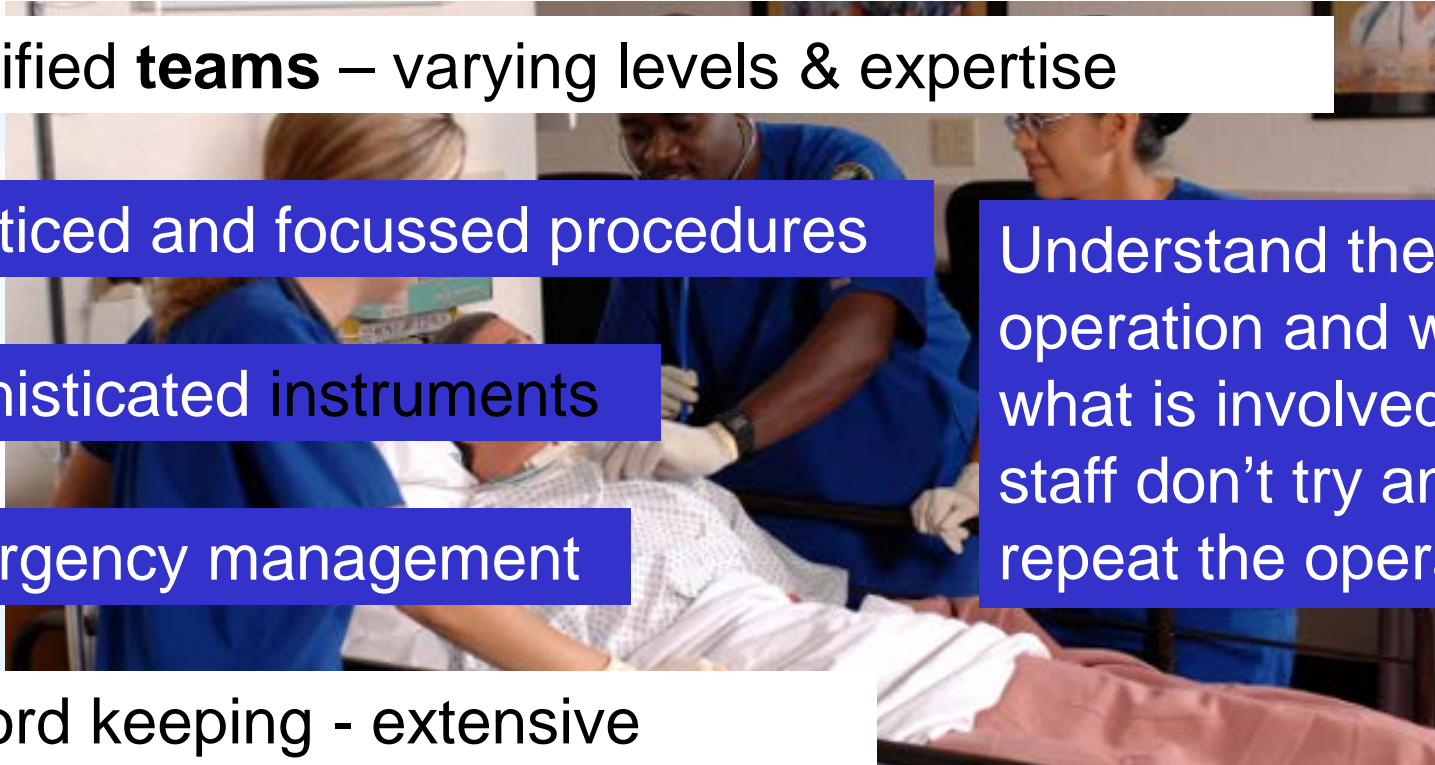
Sophisticated instruments

Emergency management

Record keeping - extensive

Proper procedures for everything

Understand the operation and who/ what is involved but staff don't try and repeat the operation!





Excellent
infrastructure



Very qualified doctors



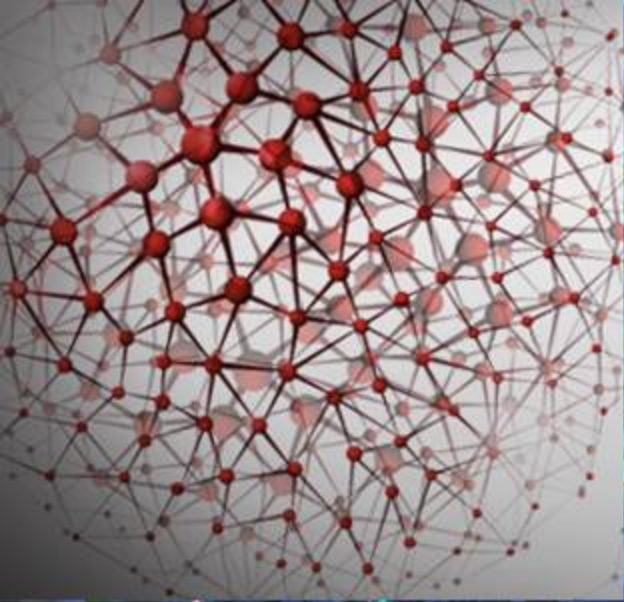


Post
operative
care



Bottom line ... we (librarians) are in the 'post operative' business! We take care of 'data patients' so that they can have a 'long and useful research life'!

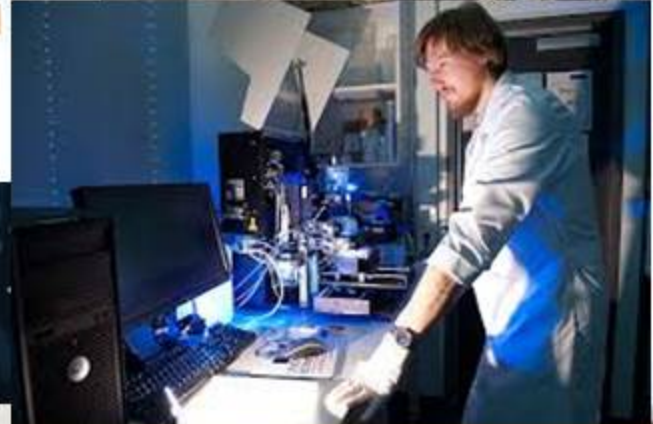




ICT
Infrastructure



Continuous learner
Conscientious
Results driven
Associative thinker
Love of complex problems
Intellectual fearlessness
Critical thinker
Love of data
Creative
Empathetic
Making research actionable
Pragmatic
Genuine interest in people
Good listener
Attention to details
Ability to see all sides
Intuitive feel for data
Broad knowledge
Curiosity
Lives client/customer service
Fore sight
Good communicator
Flexibility



Researchers



Data librarians



**The most important task is to give
dependable access to a reliable
eResource!**

The data life cycle

Re-using data

- follow-up research
- new research
- undertake research reviews
- scrutinise findings
- teach and learn

Creating data

- design research
- plan data management (formats, storage etc)
- plan consent for sharing
- locate existing data
- collect data (experiment, observe, measure, simulate)
- capture and create metadata

Processing data

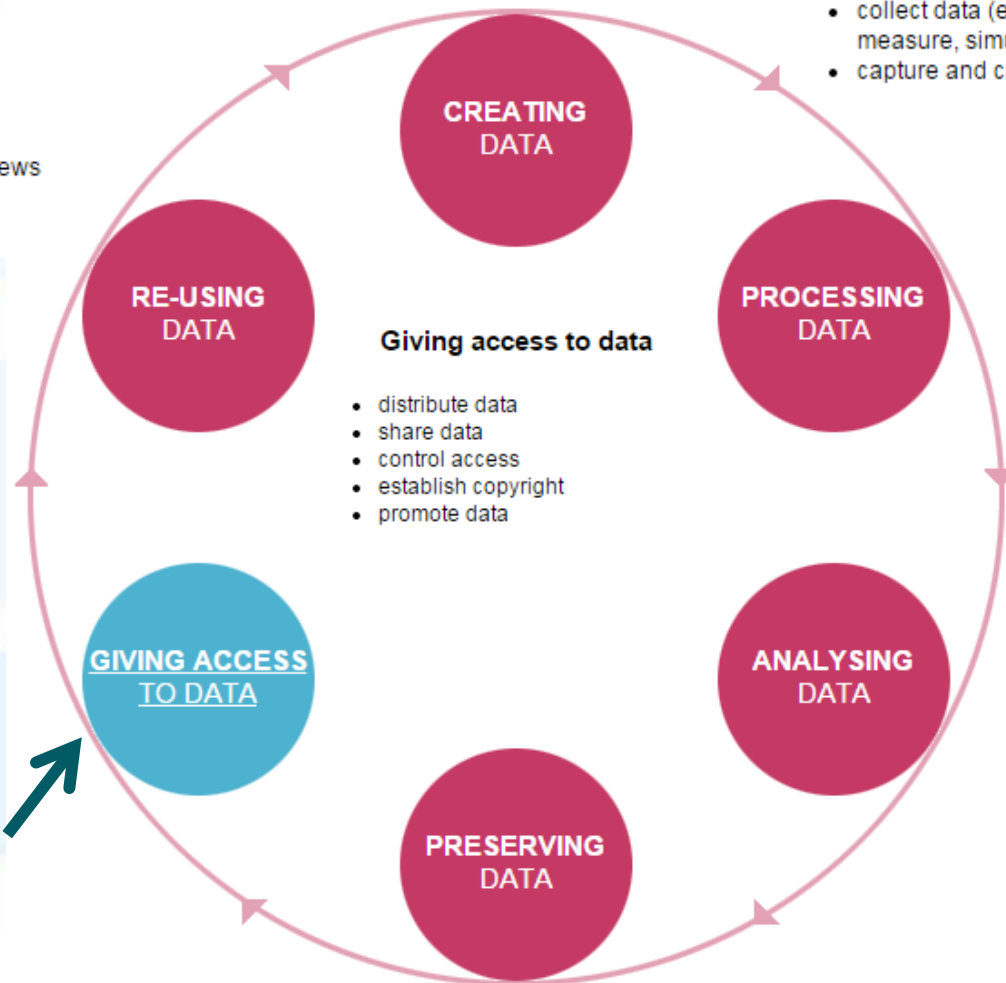
- enter data, digitise, transcribe, translate
- check, validate, clean data
- anonymise data where necessary
- describe data
- manage and store data

Analysing data

- interpret data
- derive data
- produce research outputs
- author publications
- prepare data for preservation

Giving access to data

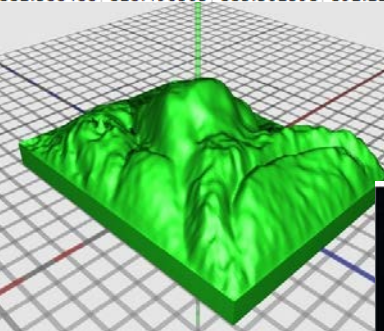
- distribute data
- share data
- control access
- establish copyright
- promote data



E-Resource

These are all data ...

10/29/04	124.761606	129.27356	122.260995	209.66
11/1/04	119.977679	129.534731	124.739135	176.31
11/2/04	130.46875	135.839924	130.84732	168.28
11/3/04	135.895502	149.510531	140.795689	120.68
11/4/04	134.127052	140.495868	132.823819	206.13
11/5/04	129.851598	137.880438	124.888856	189.67
11/8/04	123.797241	131.84633	126.146789	202.49
11/9/04	118.435374	130.691651	112.877008	140.36
11/10/04	112.401212	121.561443	114.237637	125.29
11/11/04	112.388488	128.496503	113.302591	192.22
11/12/04				
11/15/04				
11/16/04				
11/17/04				
11/18/04				
11/19/04				
11/22/04				



- Instrument measurements
- Experimental observations
- Still images, video and audio
- Text documents, spreadsheets, databases
- Quantitative data (e.g. survey data)
- Qualitative data (interview transcripts)
- Simulation data, models & software
- Slides, artefacts, specimens, samples
- Sketches, diaries, lab notebooks ...

Data landscape

The underlying attributes and requirements are very different!

Extreme Data

- Global, massive, well-typed, homogeneous volumes
- LHC & SKA

Research Big Data

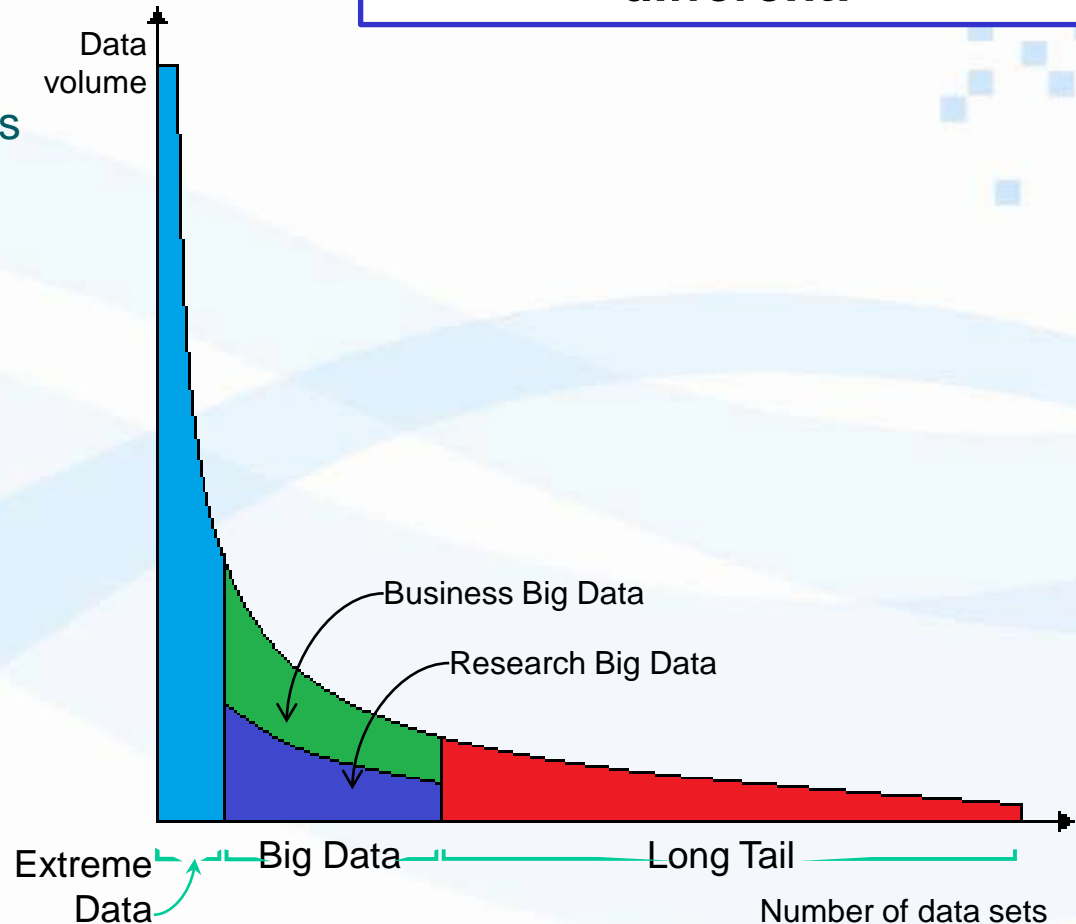
- Large, mixed-typed volumes
- Imagery, text, audio, etc

Business Big Data

- Lots of (closed) transactional, serialised data
- Sentiment data (Facebook, Twitter, etc)

Long Tail Data

- Lots of relatively small and diverse data sets



Source: Vahed, 2014

Criteria for the assessment of data repositories as an eResource



or



- Accessibility
- Usage
- Quality
- Trustworthiness

Criteria for the assessment of data repositories as an eResource

- **Accessibility**

- Research for the public good, commercial value
- Confidentiality
- Access continuum

Data enclave

Restricted access

Open access

- Visible

- Discoverable
- Accessible
- Usable

} Data repositories

Criteria for the assessment of data repositories as an e-Resource

- **Usage**

- Acceptance of data sharing as part of the responsible conduct of research
- Promotion of the secondary use of research data
- Incentives for data producers (data as a “research output”)

Criteria for the assessment of data repositories as an eResource

- Usage

- Usage evidence

- Metrics

- Data use indexes (Download, browse, search events)
 - Citation prevalence

Human Sciences Research Council. *South African National Innovation Survey (INNOV) 2008: All provinces*. [Data set]. INNOV 2008. Version 1.0. Pretoria South Africa: Human Sciences Research Council, Department of Science and Technology [producers] 2009. Human Sciences Research Council [distributor] 2013. <http://dx.doi.org/doi:10.14749/1400837297>.

- Traces of scientific activity on the social web (Altmetrics)

- Impact stories

Criteria for the assessment of data repositories as an eResource

- **Quality**
 - Quality dimensions
 - Relevance
 - Accuracy
 - Timeliness
 - Accessibility
 - Interpretability
 - Comparability and coherence
 - Methodological soundness
 - Integrity
 - Peer review
 - Important:
 - Responsible conduct of research
 - Best practice curation

Criteria for the assessment of data repositories as an e-Resource

- Trustworthiness
 - Registration of repositories
 - Re3data/ DataBib/ DataCite
 - Self-assessment
 - DRAMBORA (Digital Repository Audit Method Based On Risk Assessment)
 - Nestor (Network of Expertise in Long-Term Storage of Digital Resources) Catalogue of Criteria for Trusted Digital Repositories
 - CARDIO (Collaborative Assessment of Research Data Infrastructure and Objectives)



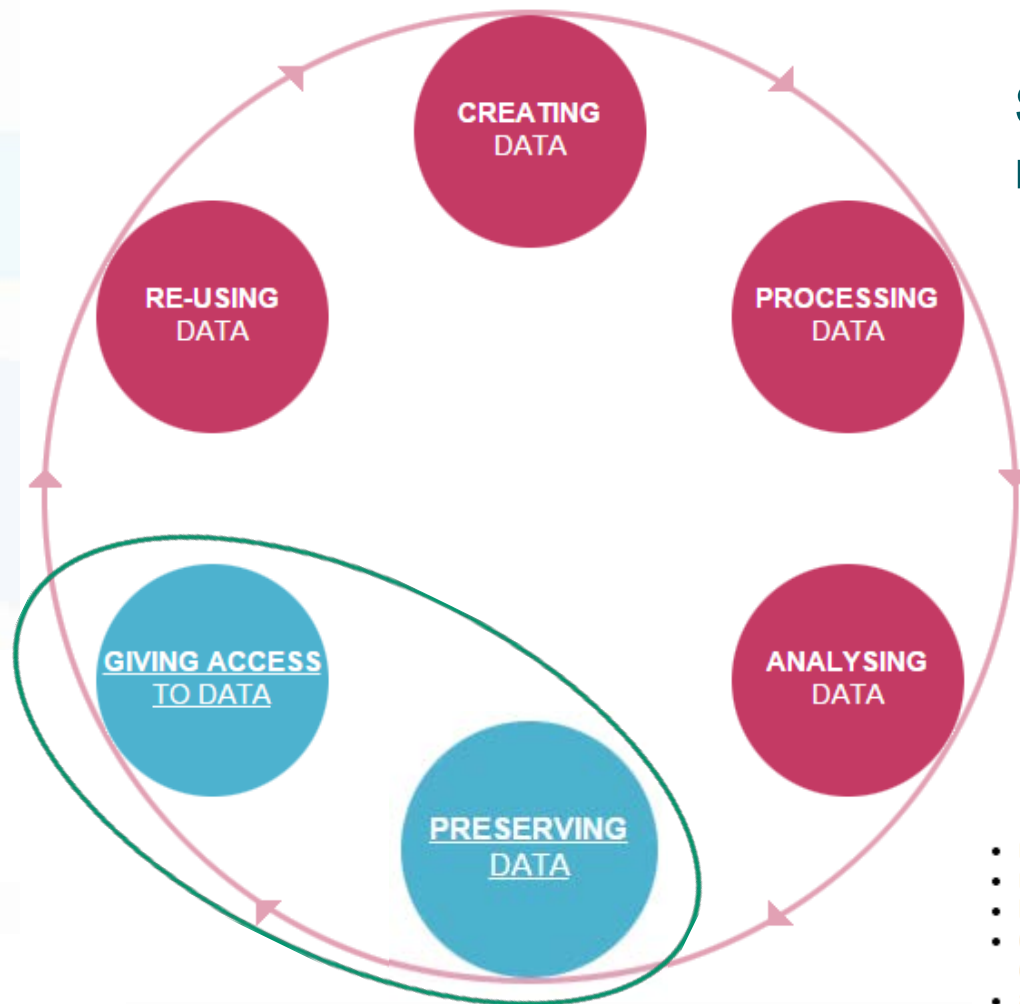
Criteria for the assessment of data repositories as an e-Resource

- **Trustworthiness**
 - Digital repository certification
 - Standards of trust
 - ISO 14721:2012 (OAIS – a reference model for what is required for an archive to provide long-term preservation of digital information)
 - ISO 16363:2013 (Audit and certification of trustworthy digital repositories – sets out comprehensive metrics for what an archive must do, based on OAIS)
 - DIN 31644 (Criteria for Trustworthy Digital Archives)
 - Accreditation
 - ICSU World Data System certification
 - Data Seal of Approval



Competence required to create/ manage/ maintain data as a trusted eResource

The data life cycle



Support data management

Curate data

Giving access to data

- distribute data
- share data
- control access
- establish copyright
- promote data

Preserving data

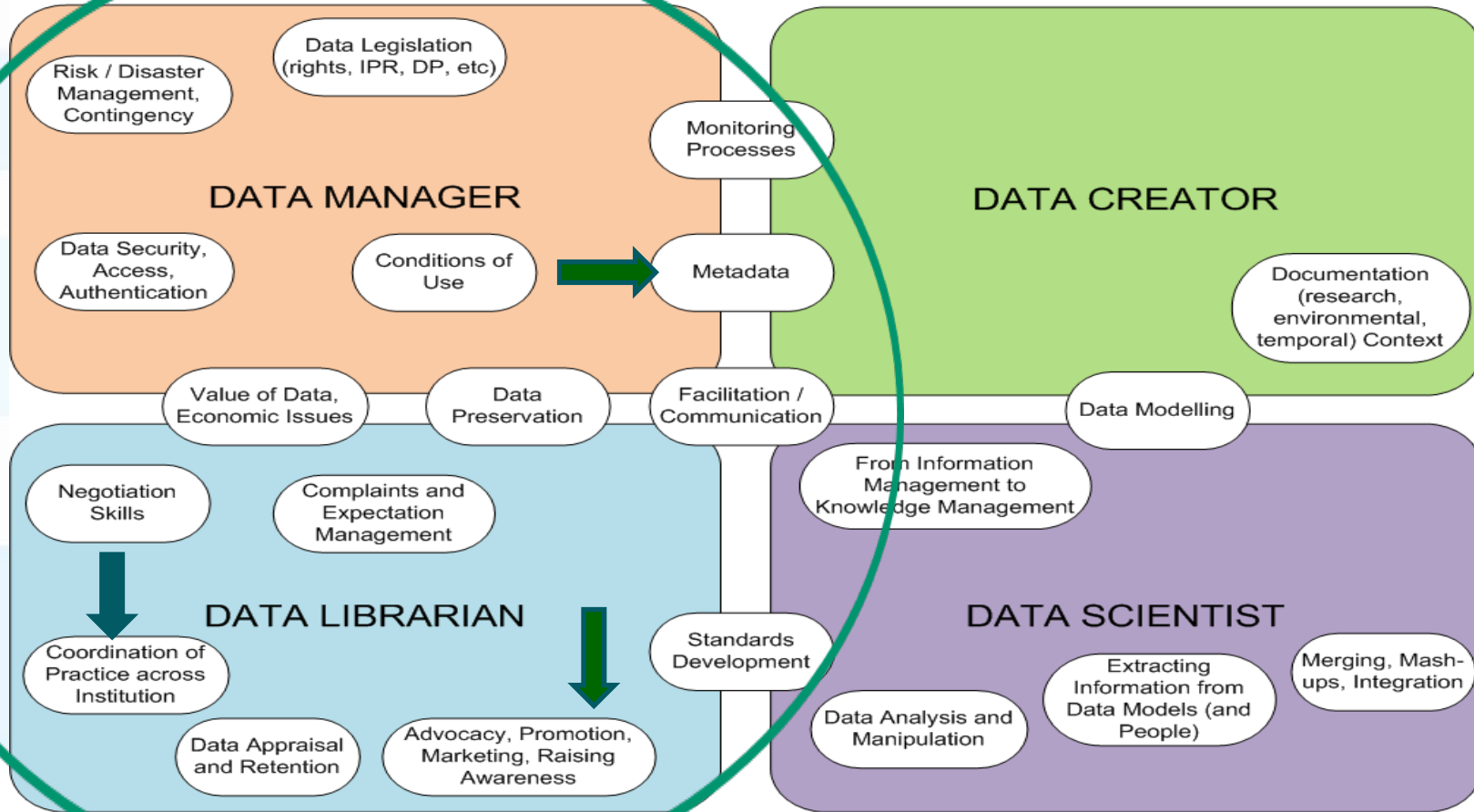
- migrate data to best format
- migrate data to suitable medium
- back-up and store data
- create metadata and documentation
- archive data

What competence?

(training, skills & expertise)

CORE SKILLS FOR DATA MANAGEMENT

A follow-up from the second DCC Research Data Management Forum (November 2008)



What should data librarians be doing?

- (1) Support the entire research data lifecycle but focus on access to and preservation of the resource.
- (2) Participate in institutional planning for RDM.
- (3) Promote best practice RDM (awareness & training).
- (4) Facilitate access to storage (infrastructure).
- (5) Provide & maintain the discovery/ access infrastructure (repositories).
- (6) Advise where & how to deposit data (journal/ funder requirements).
- (7) Know enough so that you could advise on trusted data resources.
- (9) Promote proper data citation.

Helping yourself!

- (1) Develop your own understanding of the discipline (the methodologies, nature of the data).
- (2) Liaise/ collaborate with researchers early on ... get your hands dirty!
- (3) Be totally comfortable with metadata.
- (4) Develop your own skills/ expertise in taking care of data.
- (5) Create your own research data ... to develop skills and expertise!
- (6) Build trust in the eResource.
- (7) Build trust in your professional ability to manage this eResource.
- (8) Learn from each other!

Conclusions

- It is essential that we understand when data could be regarded as a reliable eResource.
- We need to spread the word:
 - Help users understand how to create reliable/ reusable data.
 - Also where to find reliable data.
- It is essential that we encourage users to cite the data source – as they do with other reliable resources.
- There is no turning back – the NRF statement on open access is the encouragement that was required to get researchers to deposit their research data.
- Deposit does not automatically mean reliable eResource!

Conclusions

- It is urgent that more people get involved in data management.
- The competencies to manage data do still need to be identified and developed but a good foundation already exists in cataloguers, indexers, quality controllers, and systems librarians.
- Data as an eResource is an important component of science as well as society. It requires expertise and dedication to manage. It is resource intensive and requires a multi-disciplinary approach.
- Data as a reliable eResource is in its infancy but it is growing.
- Establishing data as a trusted eResource will NOT go away!

References

- DCC. 2008. Core skills for data management. Available: <http://www.dcc.ac.uk/sites/default/files/documents/RDMF/RDMF2/coreSkillsDiagram.gif> Accessed on 24 April 2015.
- Cox, A.M. 2014. CILIP Executive briefing of research data management. Welcome and introduction. Available: <http://www.cilip.org.uk/cilip/events/cilip-research-data-management-executive-briefing>
- Jacobs, N., Thomas, A. and McGregor, A., 2008. *Institutional repositories in the UK: the JISC approach*. Library Trends, 57(2), pp. 124-141.
- Plale, B., McDonald, R.H., Chandrasekar, K., Kouper, I., Konkiel, S., Hedstrom, M.L., Myers, J. and Kumar, P., 2013. *SEAD Virtual Archive: Building a Federation of Institutional Repositories for Long-Term Data Preservation in Sustainability Science*. International Journal of Digital Curation, 8(2), pp. 172-180.
- Palmer, C.L., Tefteau, L.C. and Newton, M.P., 2008. *Strategies for institutional repository development: a case study of three evolving initiatives*. Library Trends, 57(2), pp. 142-167.
- Salo, D., 2010. Retooling libraries for the data challenge. Ariadne, 64.
- Taylor, L., 2013. *Plugging the big data gap in DSpace using SWORD and Globus*. United Kingdom: University of Exeter.
- Vahed, A. 2014. Data Intensive Research Initiative for South Africa (DIRISA). eRresearch Africa conference. Cape Town: November.
- Witt, M., 2008. *Institutional repositories and research data curation in a distributed environment*. Library Trends, 57(2), pp. 191-201.
- Wong, G.K., 2009. *Exploring research data hosting at the HKUST institutional repository*. Serials Review, 35(3), pp. 125-132.
- Weber, B., 2011. *The RU Research Data Portal Providing Customized Access for Specific Types of Data and Primary Users*. United States of America: Rutgers University Libraries.
- Weber, N., Thomer, A., Mayernik, M., Dattore, R., Zaihua, J., & Worley, S. (2013, January). *The product and system specificities of measuring impact: Indicators of use in research data archives*. 8th International Digital Curation Conference, Amsterdam, Netherlands.