HSRC
Human Sciences
Research Council

Research Data Management and Institutional Repositories

2014 LIS Research Symposium
UNISA

Dr Lucia Lötter

24 July 2014

Presentation overview

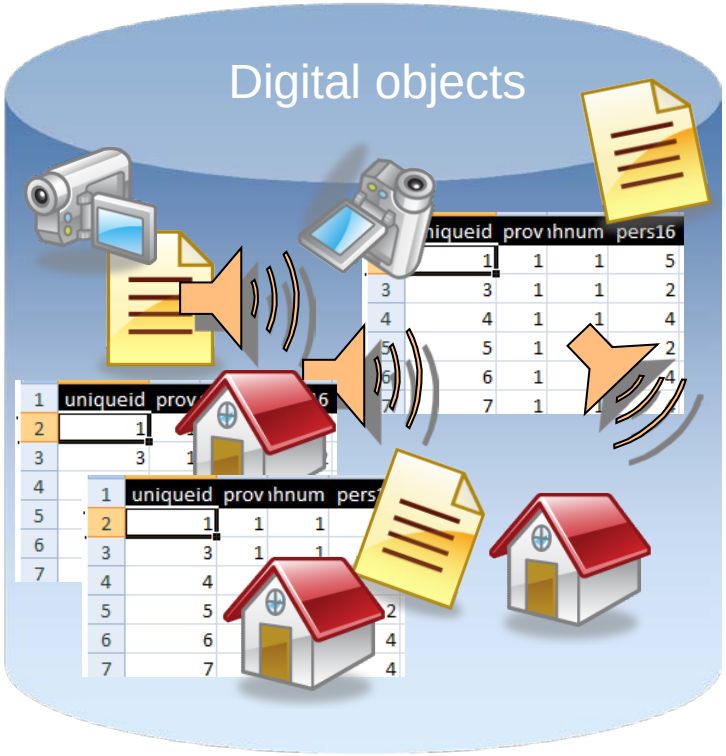
- Data and Research Data Management (RDM)
- Importance of research data and research data management
- Supporting RDM
  - Repositories in a data ecosystem
  - Institutional Repositories and RDM
    - Requirements for research data repositories
    - Nature of research data
    - The management of research data as a digital objects
- Examples
- Closing remarks

HSRC
Human Sciences
Research Council

What is research data?

"Research data, unlike other types of information, is collected, observed, or created for purposes of analysis to produce original research results."

Edinburgh University 2010



HSRC
Human Sciences Research Council

Importance of research data & RDM

"Well-managed data in digital form have great potential to be searched, accessed, mined and reused. Data may be examined to validate research results; it could be consulted by researchers of related interest and save time and resources in data re-collecting; data may even be re-purposed to answer questions unrelated to the context in which it was first generated or gathered. The value of data grows significantly as the data form more accessible collections."

"The awareness of the data deluge phenomenon, the potential impact of data reuse, and the desire to maximize the return on investment of research funding all led to increasing amounts of discussion on research data management."

Wong 2009

HSRC Human Sciences Research Council

Research Data Management

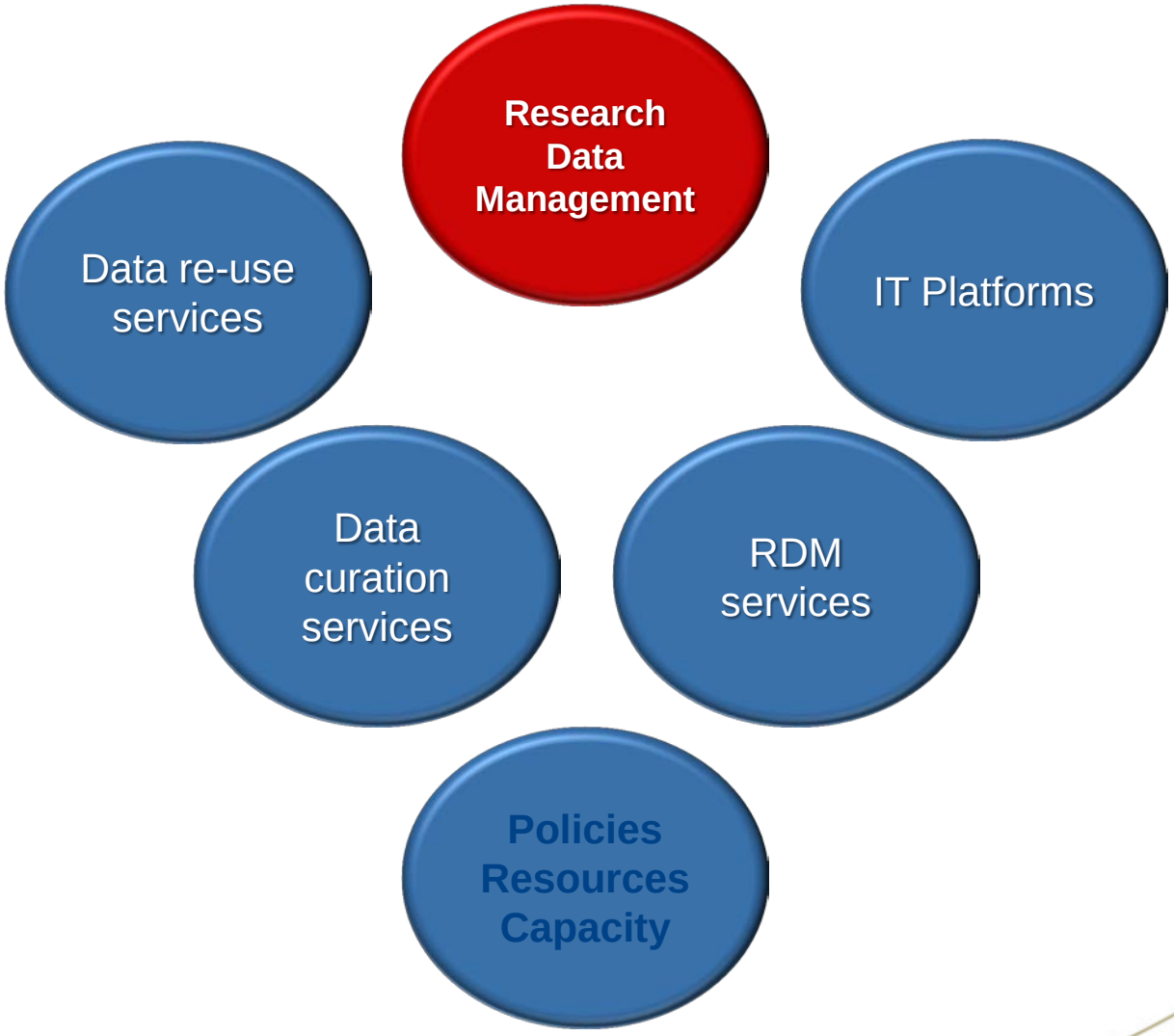
Data management is the process of controlling the information generated during a research project.

Managing data is an integral part of the research process.

How data is managed depends on the types of data involved, how data is collected and stored, and how it is used - throughout the research lifecycle.

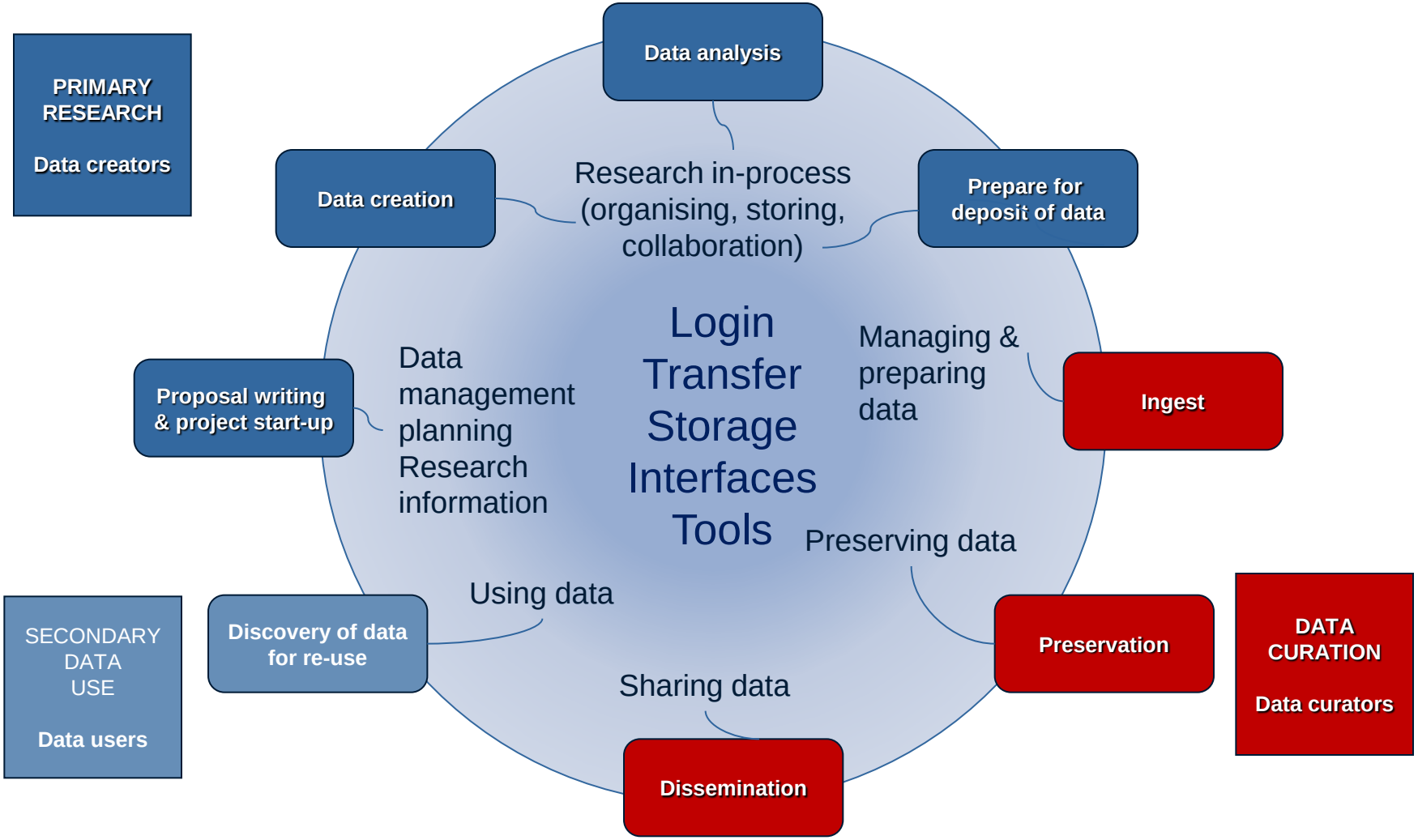
HSRC Human Sciences Research Council

Supporting RDM



HSRC
Human Sciences Research Council

Research Data Management Platforms



Repositories in a data ecosystem

- Types of repositories
  - Institutional repositories \ Discipline (subject repositories / data centres)
  - Research \ Community \ Reference data collections
  - Metadata repositories \ aggregators
- Implementation models
  - One or multiple custodians
  - Different data pathways
    - Data producer ➔ Analysis (researcher)
    - Data producer ➔ IR
    - Data producer ➔ IR ➔ Discipline repository
    - Data producer ➔ Discipline repository ➔ Metadata repository

National Science Board 2005; Beagrie, Chruszcz et al. 2008; Wilson, Martinez-Uribe et al. 2011

HSRC Human Sciences Research Council

Repositories in a data ecosystem

- The curation continuum
  - Levels of curation (Rusbridge, 2010)
    - High (high levels of expertise, where subject specialists are involved during the ingest phase of data archiving, adding and cleaning descriptive metadata)
    - Low (greater degree of automation; minimal manual intervention)
  - Information continuum

Object:	Less Metadata ⟷	More Metadata
	More Items ⟷	Fewer Items
	Larger Objects ⟷	Smaller Objects
	Objects continually updated ⟷	Objects static
Management:	Researcher Manages ⟷	Organisation Manages
	Less Preservation ⟷	More Preservation
Access:	Closed Access ⟷	Open Access
	Less Exposure ⟷	More Exposure

Wilson, Martinez-Uribe et al. 2011; Treloar, Harboe-Ree 2008

HSRC Human Sciences Research Council

Institutional Repositories

"An IR is a set of services and technologies that provide the means to collect, manage, provide access to, disseminate, and preserve digital materials produced at an institution."

Shreeves, Cragin 2008

Research outputs

Research data?

HSRC
Human Sciences
Research Council

IRs and research data

Different opinions!

HSRC
Human Sciences
Research Council

IRs and research data

role in supporting new forms of data-intensive scholarship. Scientists' incentives for changing the scholarly communication process do not relate to institutional needs, but rather the reality that data have become a new form of publication, which are critical for their research and teaching purposes. Promoting IRs as a solution to problems that may not concern faculty has been unproductive. However, presenting IRs as a mechanism for housing certain data as part of a compound object publication could be more productive. Perhaps more importantly, IRs could become an important component in a data curation strategy.

getting material into IRs first (see Kien et al., in this issue). In many cases,

Wong 2009; Shreeves, Cragin 2008, Salo 2010; Choudhury 2008

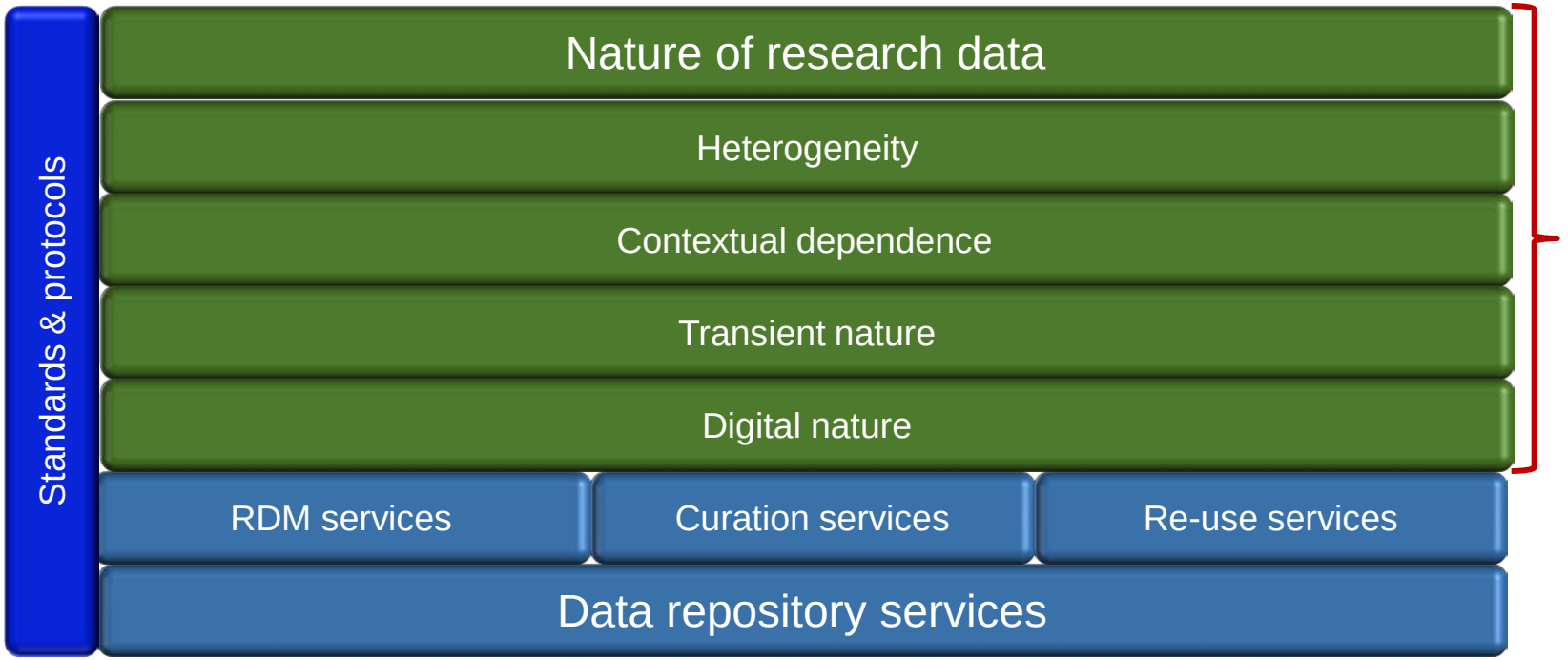
HSRC Human Sciences Research Council

Questions to answer

- Who are the users of the services and stakeholders?
- What do they need / require?
- What is the nature of the content?
  - Are data sets unique digital objects with unique requirements?
  - Do data sets require a unique set of services?
- How is the content produced and used?
- How will the platform fit into the bigger data ecosystem?
- What is achievable within the context of the organisation?
  - Other platforms and services in the organisation
  - Commitment of organisation (financial and human resources)
  - Readiness of the organisation

HSRC
Human Sciences Research Council

Data related requirements



Jacobs, Thomas et al. 2008; Wong 2009; Witt 2008; Plale, McDonald et al. 2013; Salo 2010; Taylor 2013 ; Palmer 2008; Weber 2011

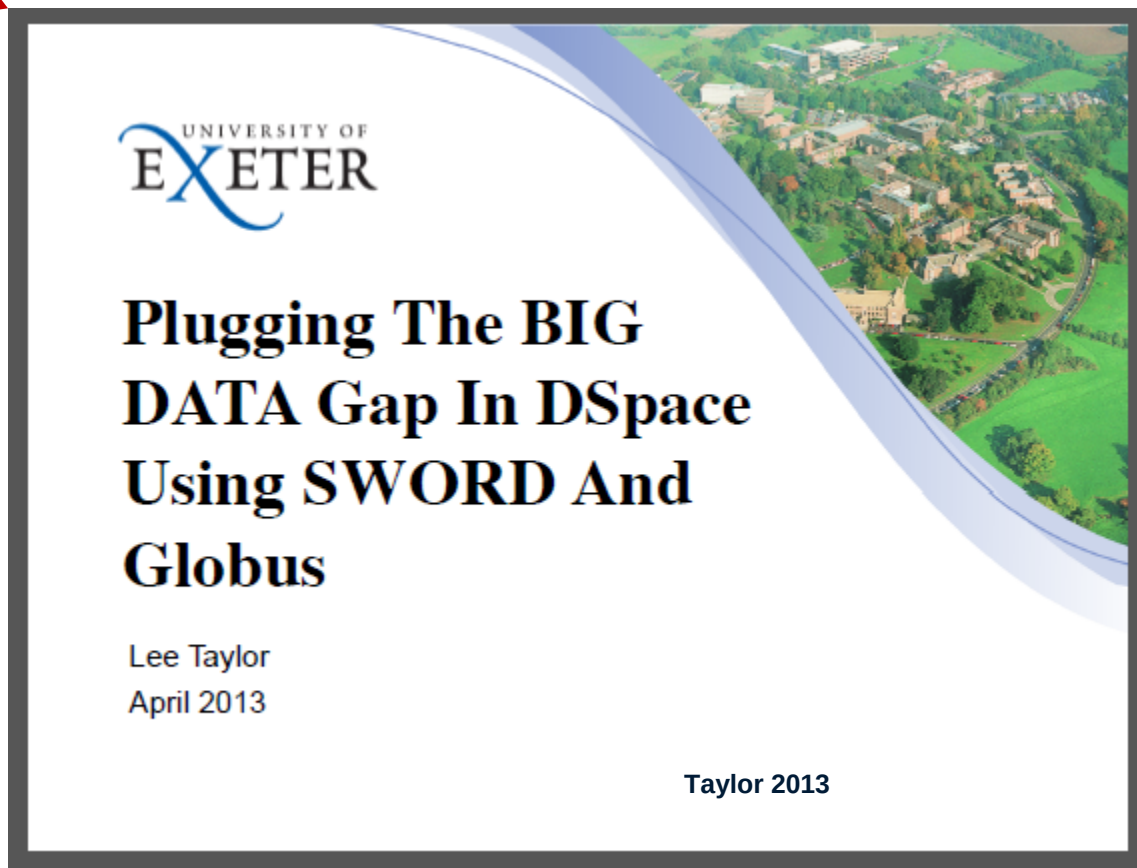
HSRC Human Sciences Research Council

Heterogeneity

- Research methodologies
- Practices of researchers
- Kinds of data, sizes, formats and composition
- Data value
- Raw, aggregated data

Flexible and highly scalable

Clarity on what the repository



Contextual dependence

- Collection composition
  - Linking items (data files, contextual documents, outputs)
  - Data with outputs or outputs with data
  - Organise into project-based collections (a ”bag””)

Data set

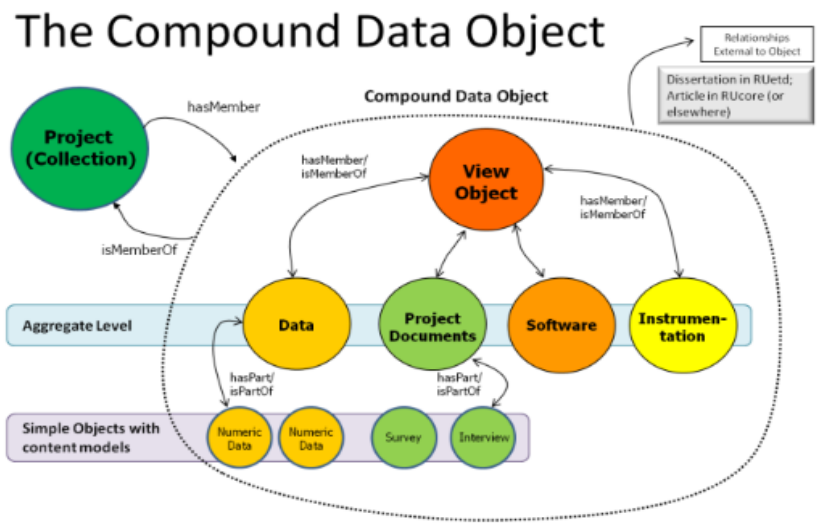
Spatial aspects of unemployment in South Africa 1991-2007: Municipalities - All provinces

All data sets	Data set details	Documentation	Data files	Outputs	Access conditions	Contact

Data set metadata record

Data set ID	UNEMPL 1991-2007 Municipalities
Title	Spatial aspects of unemployme

http://curation.hsrc.ac.za/Dataset-278.phtml



Weber 2011

Transient nature

- Snapshots vs “live” data
- Versioning of data sets

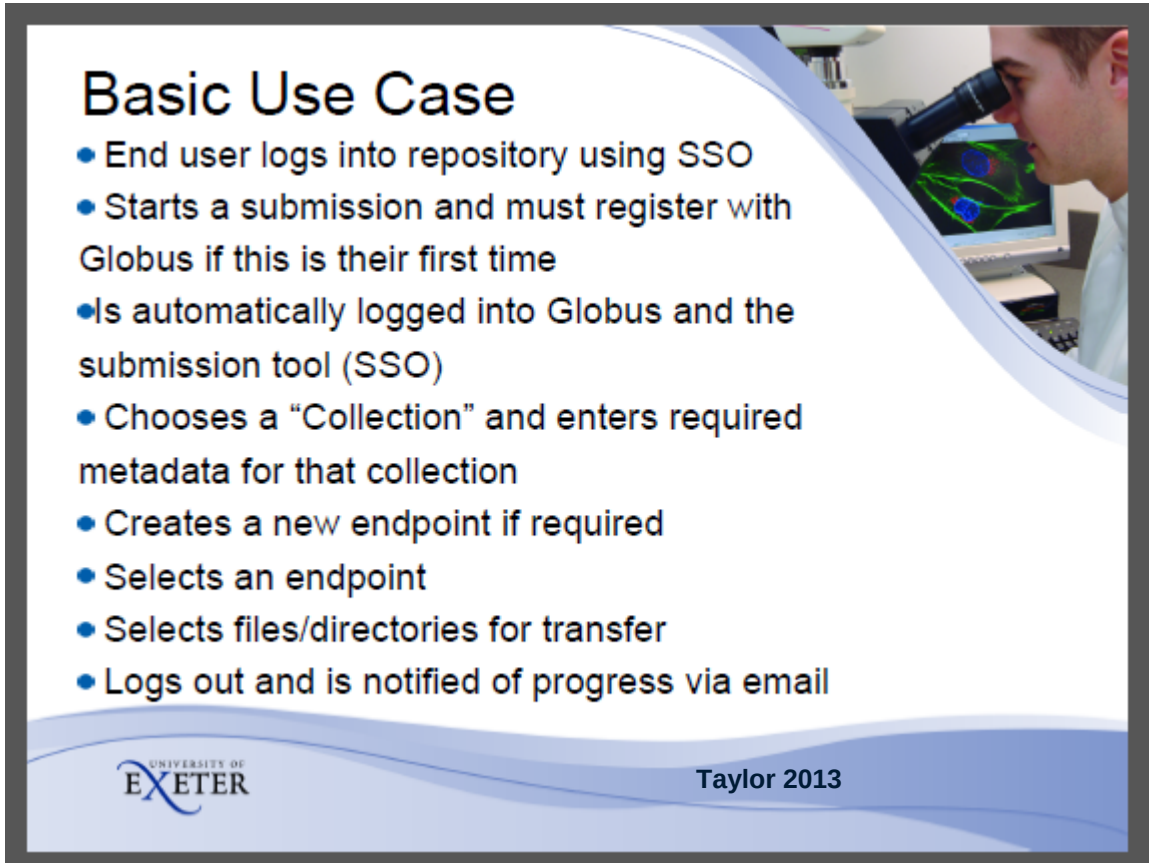
> “…digital preservation systems designed to steward only final, unchanging materials can only fail faced with real-world datasets and data-use practices.”
>
> **Salo 2010**

- Authority management (Author IDs)
- Persistent identifiers (DOIs)
- Citation standards

HSRC
Human Sciences Research Council

Digital object management

- Data submission (deposit)
  - Ethics requirements for human subjects
    - Consent to share
    - Anonymisation
  - Ownership issues
  - Self-archiving
  - Automation



Digital object management

- Access (consistent with tools and processes of the research community)
  - Discovery services (browsing, searching, OMP-MIH)
  - Metadata
    - Discovery, determine relevance, make data useable, provenance
    - Compliance with recognized standards of the community
  - Dissemination formats
  - Ways to serve and use data
  - Usage terms and conditions
  - Access management
  - Usage statistics

HSRC
Human Sciences Research Council

Digital object management

- Preservation
  - Preservation management
  - Registries
  - Retention period, de-accessioning, destruction
  - Strategies support to file formats and their long-term usability
    - Archival formats
    - Format migration
  - Storage and storage management
    - Multiple copies, multi-media, multi site
    - Back-up
    - Disaster recovery
    - Security

Standards and protocols

- Interoperability
  - Metadata (machine readable in appropriate standard)
- Various standards and protocols
  - Trusted Digital Repository (TRAC ISO 16363) (http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=56510)
  - Open archival information system Reference model (OAIS ISO 14721:2012) (http://www.iso.org/iso/catalogue_detail.htm?csnumber=57284)
  - Open Archives Initiative Object Reuse and Exchange (OAI-ORE) (http://www.openarchives.org/pmh/)
  - Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) (http://www.openarchives.org/ore/)
  - Simple Web-service Offering Repository Deposit (SWORD) (http://swordapp.org/)

HSRC Human Sciences Research Council

Examples

suggest. It is notable that both Cambridge and KCL in our case studies are developing central repositories to work with departmental facilities and discussing **federated local data repositories** for research data preservation combining services and skills from central and departmental repositories with data distributed and located at different repositories in the institution. A similar discussion and scoping project is also currently underway at the

e 2008).

Witt 2008; **Taylor 2013**

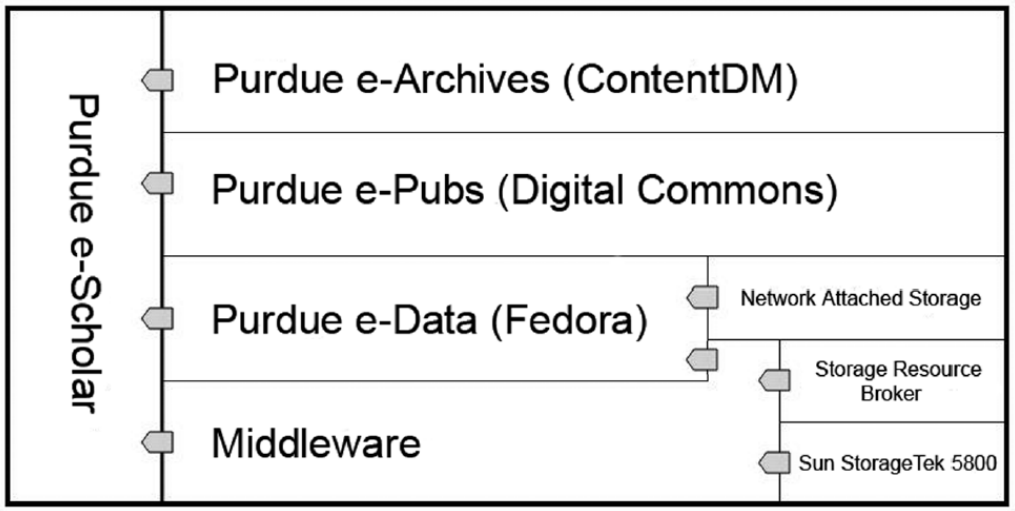
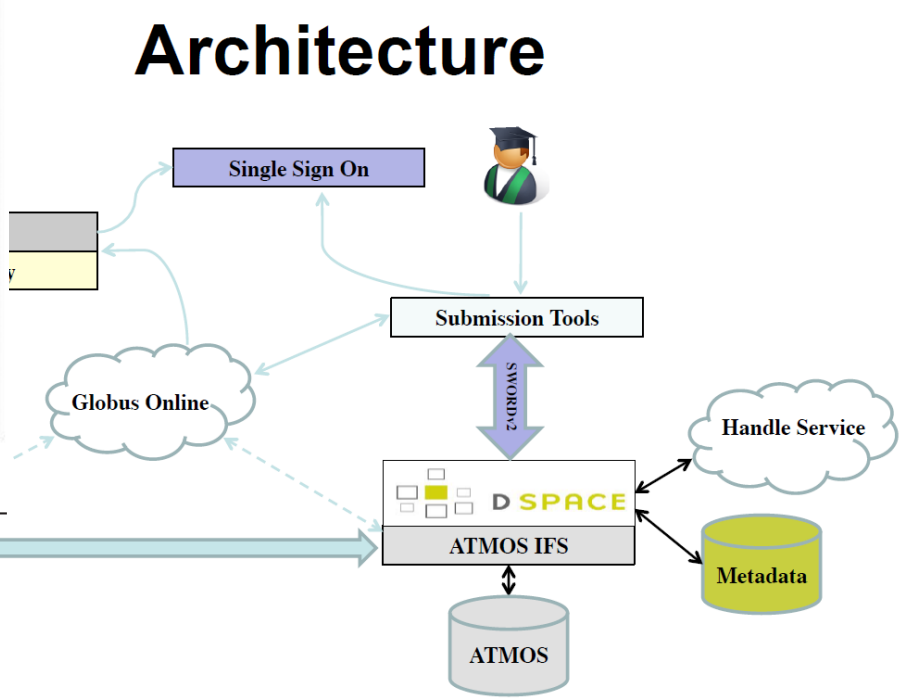


Figure 2. The Distributed Institutional Repository



- Flexible Storage and Metadata Architectures
- De-coupling Ingest, Storage and Use

Closing remarks

- Not just a one size fits all one vendor solution (install and go)
- RDM Services for small data is labour intensive (Various roles and responsibilities)
- Don’t see IRs in isolation
- Clear vision of aims
- Investigate and experiment
- Collaborate

The verdict?

HSRC
Human Sciences Research Council

References

BEAGRIE, N., CHRUSZCZ, J. and LAVOIE, B., 2008. Keeping research data safe: A cost model and guidance for UK universities.Final Report to JISC.

BUSINESS DICTIONARY, What is data management? definition and meaning . Available: http://www.businessdictionary.com/definition/data-management.html [7/23/2014, 2014].

CHOUDHURY, G.S., 2008. Case study in data curation at Johns Hopkins University. Library Trends, 57(2), pp. 211-220.

EDINBURGH UNIVERSITY, 2010. Edinburgh University Data Library Research Data Management Handbook. Edinburgh University .

JACOBS, N., THOMAS, A. and MCGREGOR, A., 2008. Institutional repositories in the UK: the JISC approach. Library Trends, 57(2), pp. 124-141.

LORD, P., MACDONALD, A., LYON, L. and GIARETTA, D., 2004. From data deluge to data curation, Proceedings of the UK e-science All Hands meeting 2004, pp. 371-357.

MIX, K. and TAYLOR JR, L., VLA Paraprofessional Forum Twentieth Annual Conference.

NATIONAL SCIENCE BOARD, 2005. Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century. National Science Foundation.

PALMER, C.L., TEFFEAU, 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.

PITTMAN, D., 2010. NSF REVAMPS DATA-SHARING POLICY Chemical & Engineering News, 88(39), pp. 46 <last_page> 47.

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.

SALO, D., 2010. Retooling libraries for the data challenge. Ariadne, 64.

SHREEVES, S.L. and CRAGIN, M.H., 2008. Introduction: Institutional repositories: Current state and future. Library Trends, 57(2), pp. 89-97.

TAYLOR, L., 2013. Plugging the big data gap in DSpace using SWORD and Globus. United Kingdom: University of Exeter.

TRELOAR, A. and HARBOE-REE, C., 2008. Data management and the curation continuum: how the Monash experience is informing repository relationships. Proceedings of VALA 2008.

UNIVERSITY OF LEICESTER, , What is research data — University of Leicester . Available: http://www2.le.ac.uk/services/research-data/rdm/what-is-rdm/research-data [7/23/2014, 2014].

WEBER, B., 2011. The RUresearch Data Portal Providing Customized Access for Specific Types of Data and Primary Users. United States of America: Rutgers University Libraries.

WILSON, J.A., MARTINEZ-URIBE, L., FRASER, M.A. and JEFFREYS, P., 2011. An institutional approach to developing research data management infrastructure. International Journal of Digital Curation, 6(2), pp. 274-287.

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.

Lucia Lötter
llotter@hsrc.ac.za
www.hsrc.ac.za