HSRC
Human Sciences
Research Council

What can take the dark out of the long tail?

Efforts to address the data management challenges of "small science"

Social science that makes a difference

Lucia Lötter
CHPC Conference
December 2011

Presentation overview

- Dark data
- The landscape of the dark tail
- Data curation - The light
- A data curation implementation demonstrator
- An energy source for the light – A Trusted Digital Repository system
- Suggestions for a stronger light

What is “dark” data?

“Data that has never been published or otherwise made available to the rest of the scientific community”

http://muse.jhu.edu/journals/library_trends/v057/57.2.heidorn.html

Shedding Light on the Dark Data in the Long Tail of Science

P. Bryan Heidorn

What is “dark” data?

Dark vs. Light is about **visibility**

- Shining the light
  - Usable
    - Understandable
    - Good quality
    - Readily available
    - Accessible
  - Visible = open access?

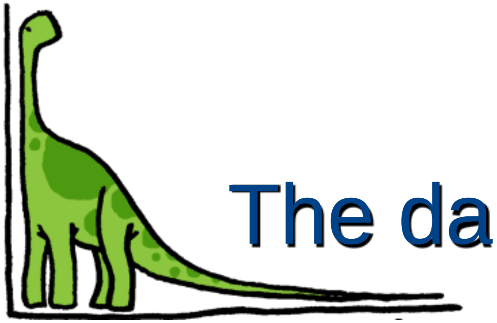
HSRC
Human Sciences
Research Council

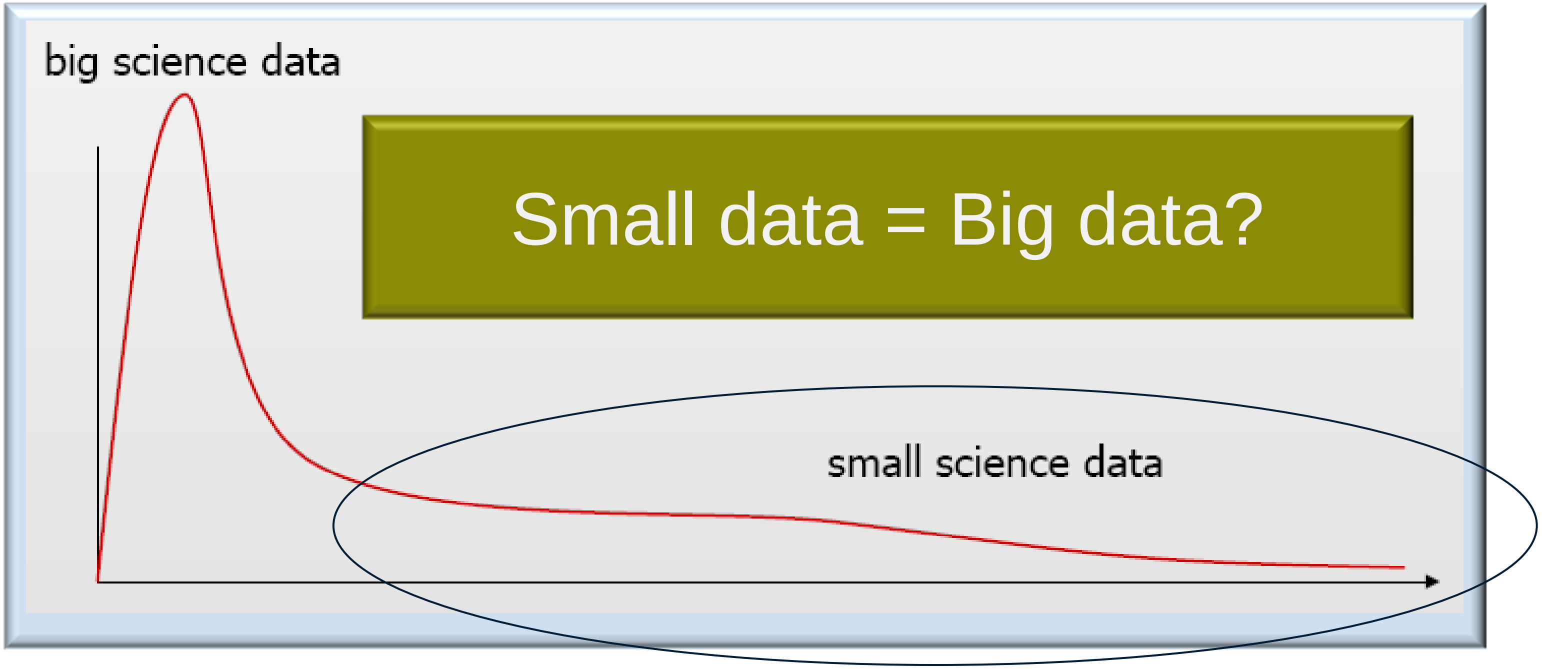
Different shades of light / dark

“Big” projects / data	“Small” projects / data
Collection automated using specialised instrumentation.	Some collected with instrumentation, but also manually-collected data.
Apply a continuum from highly structured industry-wide standards to relatively independent proprietary data standards. Can easily be submitted to structured databases.	Include in many cases user developed data structures and can contain narrative / “unstructured” data.
Highly visible in open access repositories.	Less visible or not at all. Large number not submitted to repositories.
Deep - Homogeneous collections	Wide – Heterogeneous collections
High probability of long term survival.	At risk of loss, damage, becoming unusable. Much less likely to be maintained over time.

http://muse.jhu.edu/journals/library_trends/v057/57.2.heidorn.html

The dark tail

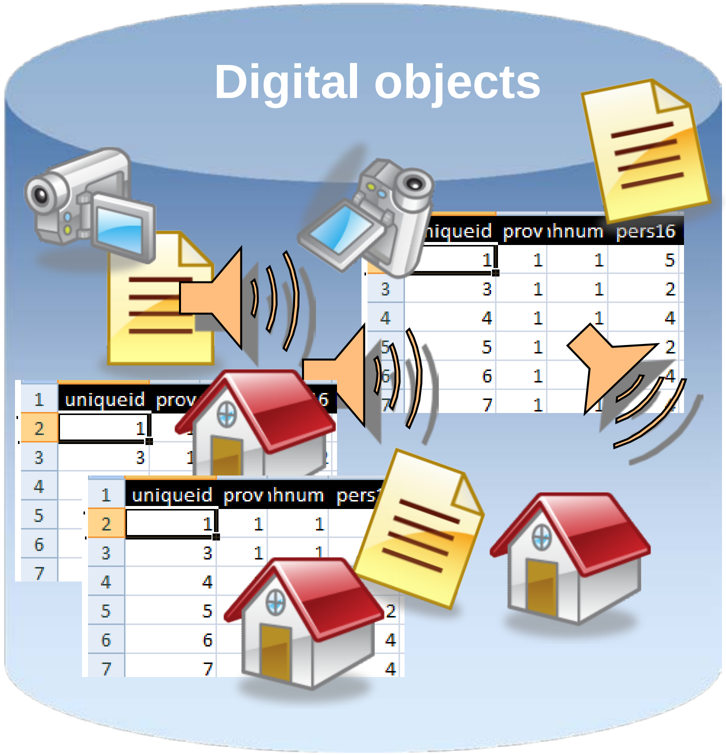




Small data = Big data?

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

Inside the dark tail



- Many data generators
- Various social science disciplines
- Nature of data
  - Qualitative
  - Quantitative
- Highly contextual

HSRC
Human Sciences
Research Council

Why a dark tail?

- Lack of focus on data as a primary scientific output
- Lack of funding
- Lack of knowledge and expertise
- Lack of infrastructure and technology

Big is better?

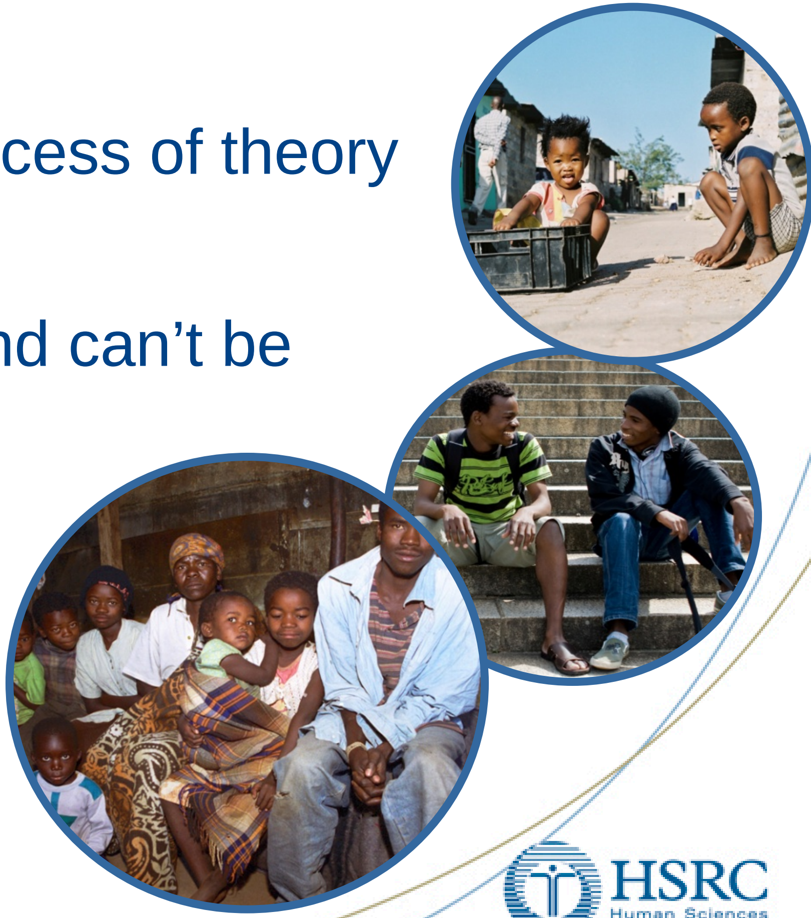
Does the dark tail matter?

Scientists are still searching for a "Unified Theory of Everything"...
They've obviously not read "The Long Tail"...
©hugh

Does the dark tail matter?

- A breeding ground for new ideas
- Addresses significant questions for improvement of society
- Essential to the scientific process of theory development and evaluation
- Data is sometimes unique and can’t be regenerated
- Cost of data collection

http://muse.jhu.edu/journals/library_trends/v057/57.2.heidorn.html



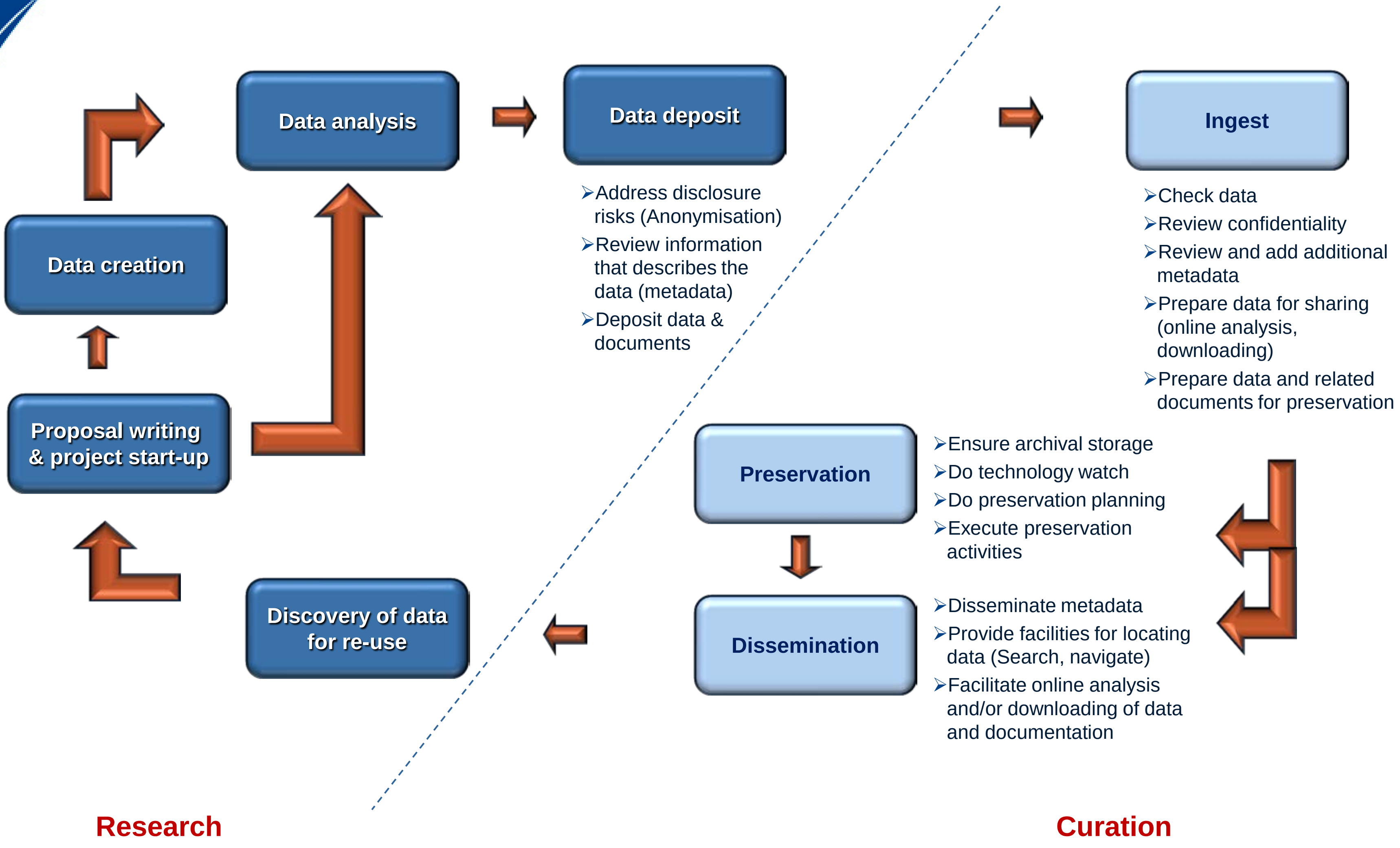
HSRC Human Sciences Research Council

Curation - The light

"[C]uration embraces and goes beyond that of enhanced present-day re-use, and of archival responsibility, to embrace stewardship that adds value through the provision of context and linkage: placing emphasis on publishing data in ways that ease re-use and promoting accountability and integration"

(http://eprints.erpanet.org/82/01/DCC_Vision.pdf)

Shining the light



Proposal writing & project start-up

Data creation

Data analysis

Data deposit

- Address disclosure risks (Anonymisation)
- Review information that describes the data (metadata)
- Deposit data & documents

Ingest

- Check data
- Review confidentiality
- Review and add additional metadata
- Prepare data for sharing (online analysis, downloading)
- Prepare data and related documents for preservation

Preservation

- Ensure archival storage
- Do technology watch
- Do preservation planning
- Execute preservation activities

Dissemination

- Disseminate metadata
- Provide facilities for locating data (Search, navigate)
- Facilitate online analysis and/or downloading of data and documentation

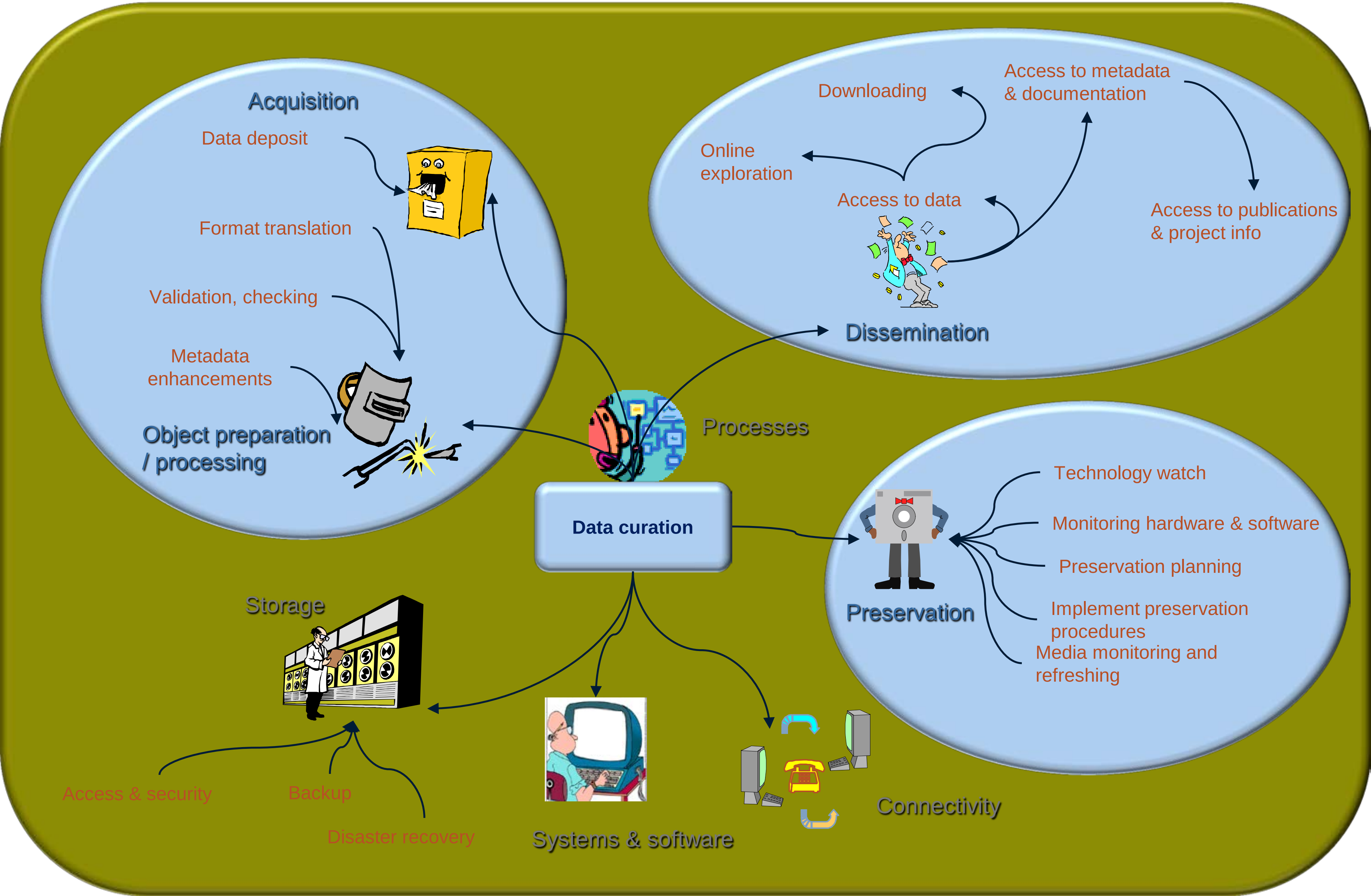
Discovery of data for re-use

Research

Curation

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

Facilitating the curation of long tail data



Acquisition

Data deposit

Format translation

Validation, checking

Metadata enhancements

Object preparation / processing

Dissemination

Downloading

Access to metadata & documentation

Online exploration

Access to data

Access to publications & project info

Processes

Data curation

Preservation

Technology watch

Monitoring hardware & software

Preservation planning

Implement preservation procedures

Media monitoring and refreshing

Storage

Access & security

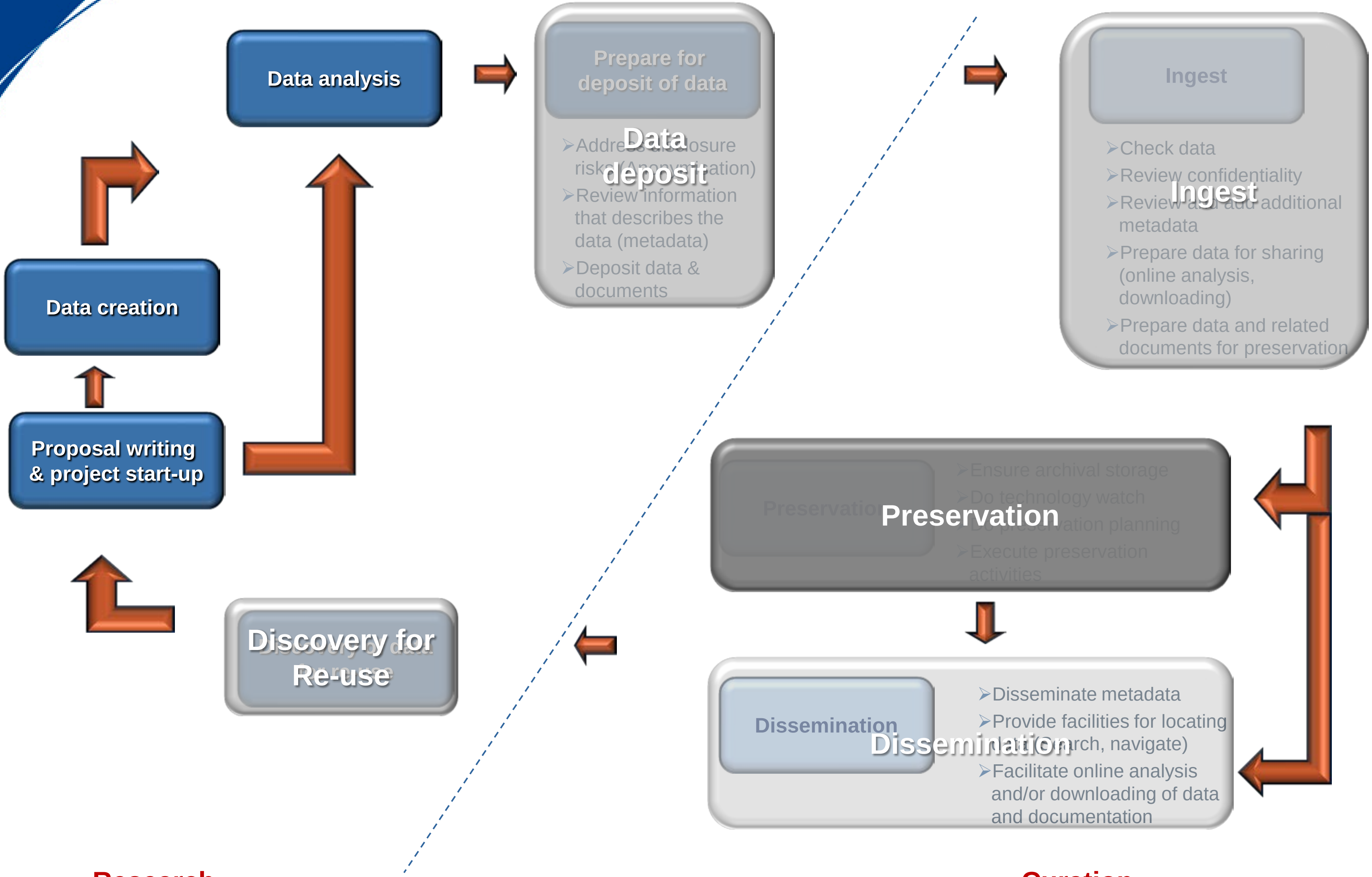
Backup

Disaster recovery

Systems & software

Connectivity

How dark is it?



Data analysis

Data creation

Proposal writing & project start-up

Discovery for Re-use

Discovery of data for re-use

Prepare for deposit of data

Data deposit

- Address disclosure risks (Anonymization)
- Review information that describes the data (metadata)
- Deposit data & documents

Ingest

Ingest

- Check data
- Review confidentiality
- Review and add additional metadata
- Prepare data for sharing (online analysis, downloading)
- Prepare data and related documents for preservation

Preservation

Preservation

- Ensure archival storage
- Do technology watch
- Do preservation planning
- Execute preservation activities

Dissemination

Dissemination

- Disseminate metadata
- Provide facilities for locating data (Search, navigate)
- Facilitate online analysis and/or downloading of data and documentation

Research

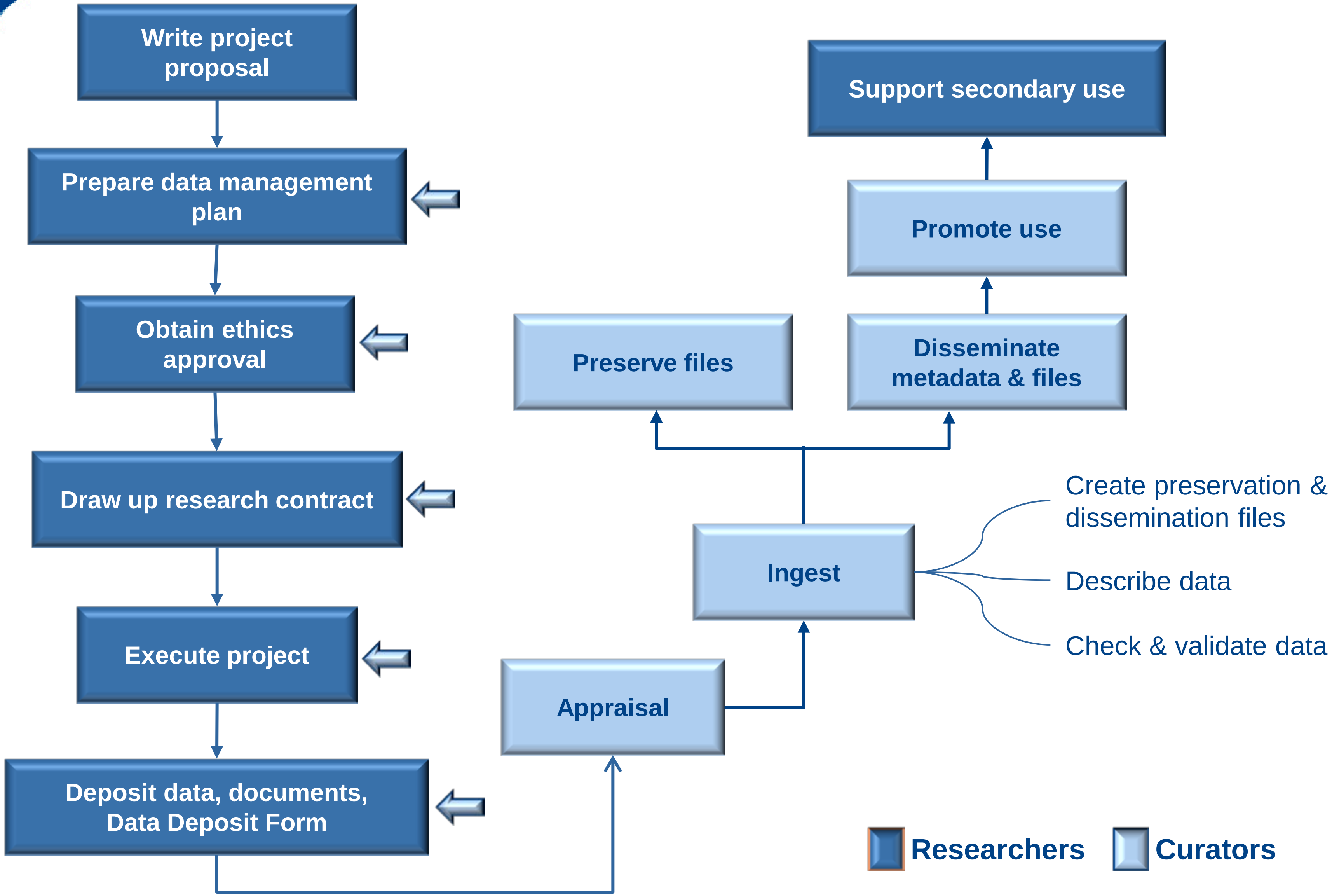
Curation

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

HSRC investment – just the beginning …

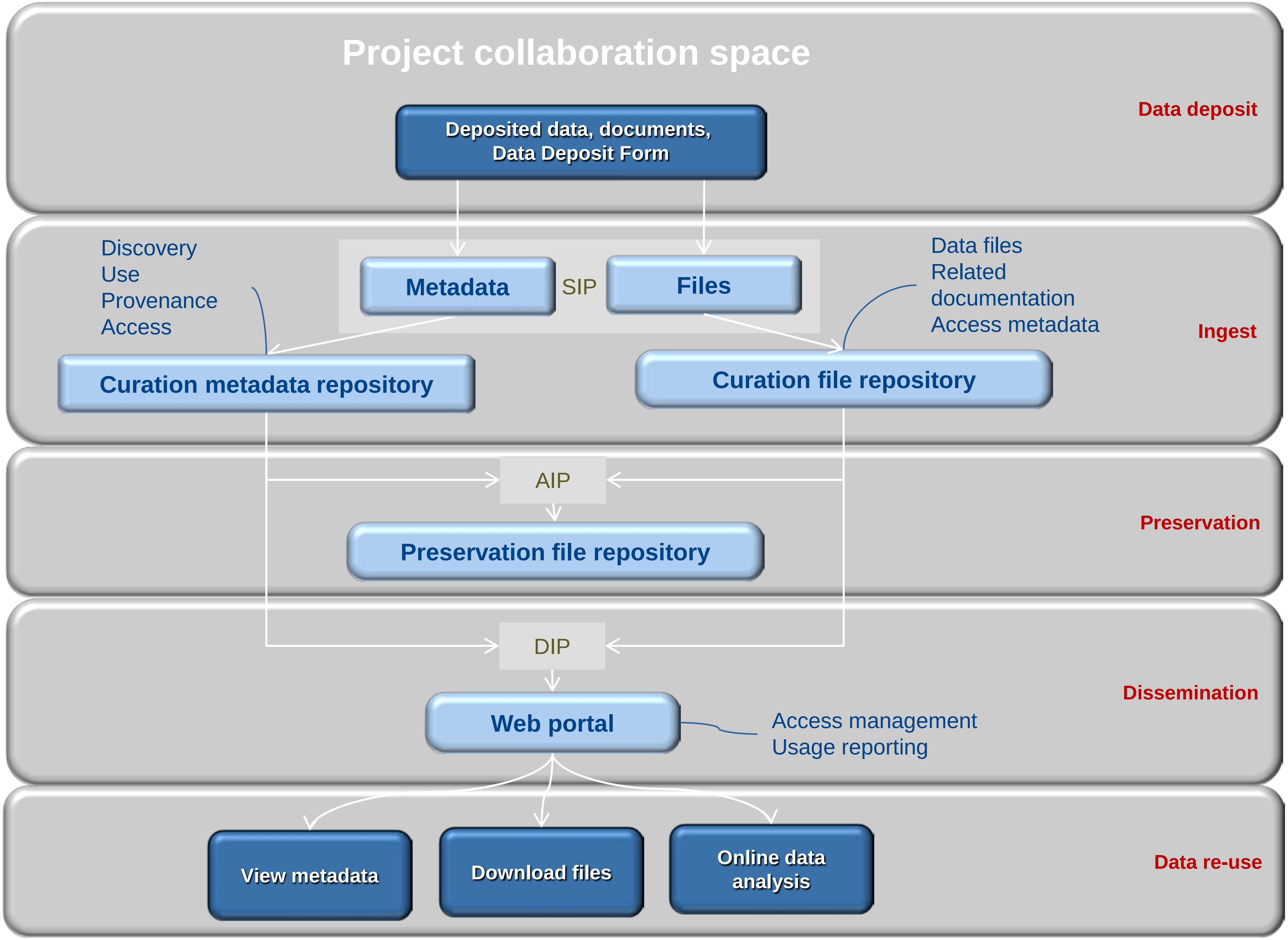
- Policies and procedures that facilitate data deposit, preparing data and related documentation
- Support for researchers in terms of data curation issues
- Training of researchers in data documentation and management
- A metadata and file repository system
- An on-line dissemination interface linked to the HSRC's Web Portal for viewing, downloading or analysis
- Operational processes that guide curation activities
- Processes to monitor and audit curated data sets for performance information purposes
- Various data sets available for secondary use

Data management process flow

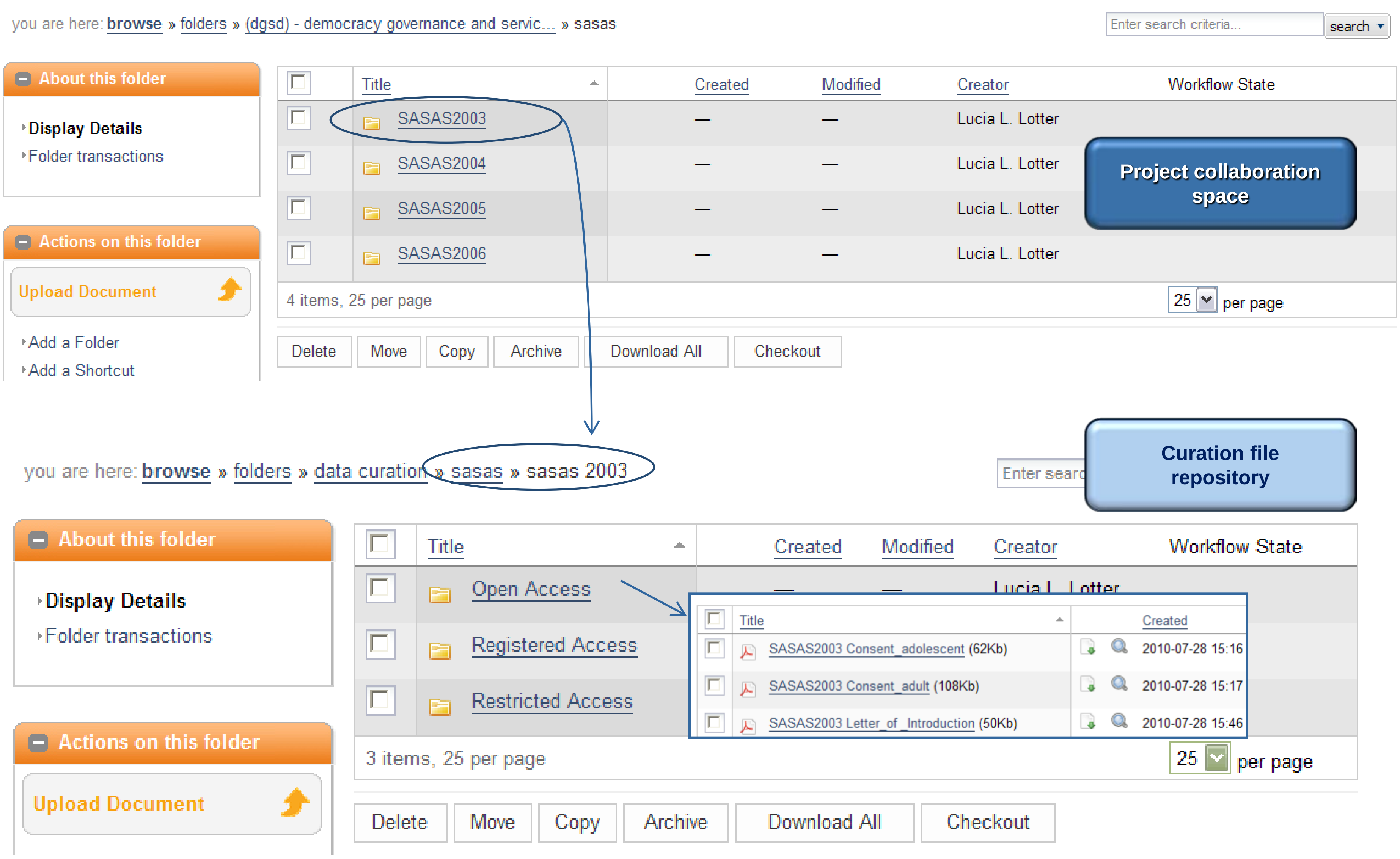

Write project proposal
Prepare data management plan
Obtain ethics approval
Draw up research contract
Execute project
Deposit data, documents, Data Deposit Form
Appraisal
Ingest
Create preservation & dissemination files
Describe data
Check & validate data
Preserve files
Disseminate metadata & files
Promote use
Support secondary use

🟦 Researchers ▢ Curators

Data flow


Project collaboration space
Data deposit
Deposited data, documents, Data Deposit Form
Discovery
Use
Provenance
Access
Metadata
SIP
Files
Data files
Related documentation
Access metadata
Ingest
Curation metadata repository
Curation file repository
AIP
Preservation file repository
Preservation
DIP
Dissemination
Web portal
Access management
Usage reporting
View metadata
Download files
Online data analysis
Data re-use

Data flow technology
Curation file repository



you are here: **browse** » folders » (dgsd) - democracy governance and servic... » sasas

Enter search criteria... search

About this folder

- **Display Details**
- Folder transactions

□	Title	Created	Modified	Creator	Workflow State
□	SASAS2003	—	—	Lucia L. Lotter	
□	SASAS2004	—	—	Lucia L. Lotter	
□	SASAS2005	—	—	Lucia L. Lotter	
□	SASAS2006	—	—	Lucia L. Lotter	

4 items, 25 per page — 25 per page

Delete | Move | Copy | Archive | Download All | Checkout

Actions on this folder

Upload Document

- Add a Folder
- Add a Shortcut

Project collaboration space

you are here: **browse** » folders » data curation » sasas » sasas 2003

Enter searc

Curation file repository

About this folder

- **Display Details**
- Folder transactions

□	Title	Created	Modified	Creator	Workflow State
□	Open Access	—	—	Lucia L. Lotter	
□	Registered Access				
□	Restricted Access				

□	Title			Created
□	SASAS2003 Consent_adolescent (62Kb)			2010-07-28 15:16
□	SASAS2003 Consent_adult (108Kb)			2010-07-28 15:17
□	SASAS2003 Letter_of_Introduction (50Kb)			2010-07-28 15:46

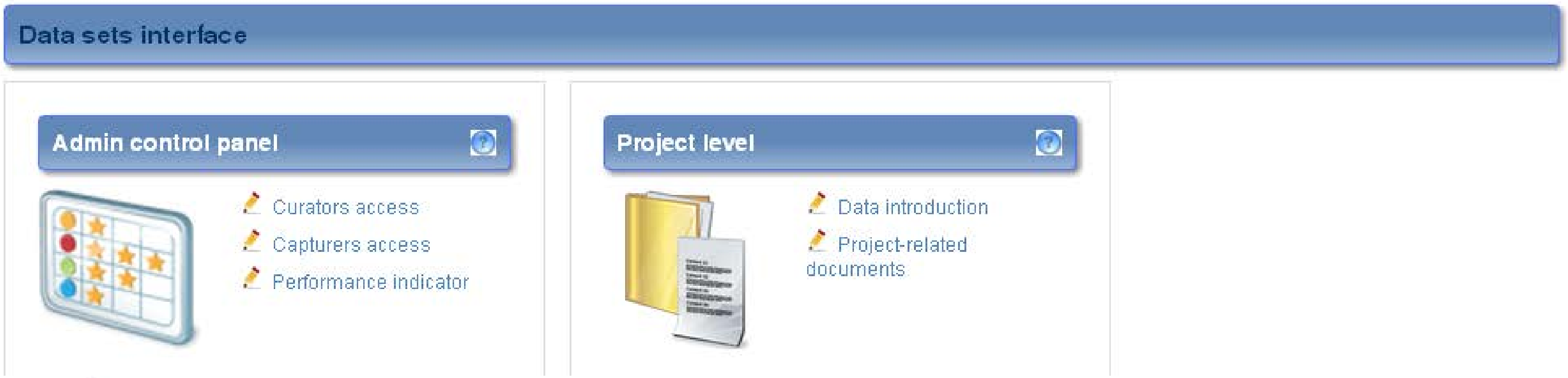
3 items, 25 per page — 25 per page

Delete | Move | Copy | Archive | Download All | Checkout

Actions on this folder

Upload Document

Data flow technology
Curation metadata repository



Data sets interface

Admin control panel

Curators access

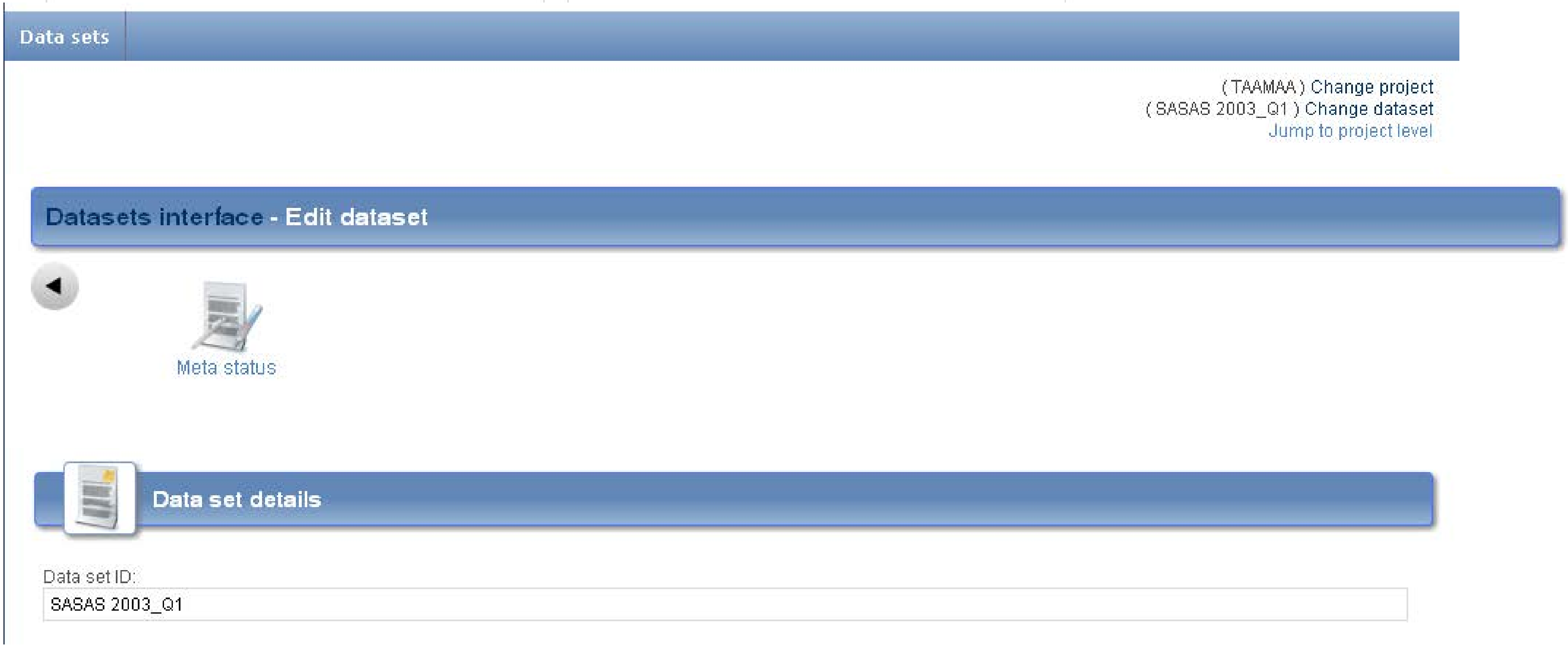
Capturers access

Performance indicator

Project level

Data introduction

Project-related documents



Data sets

(TAAMAA) Change project
(SASAS 2003_Q1) Change dataset
Jump to project level

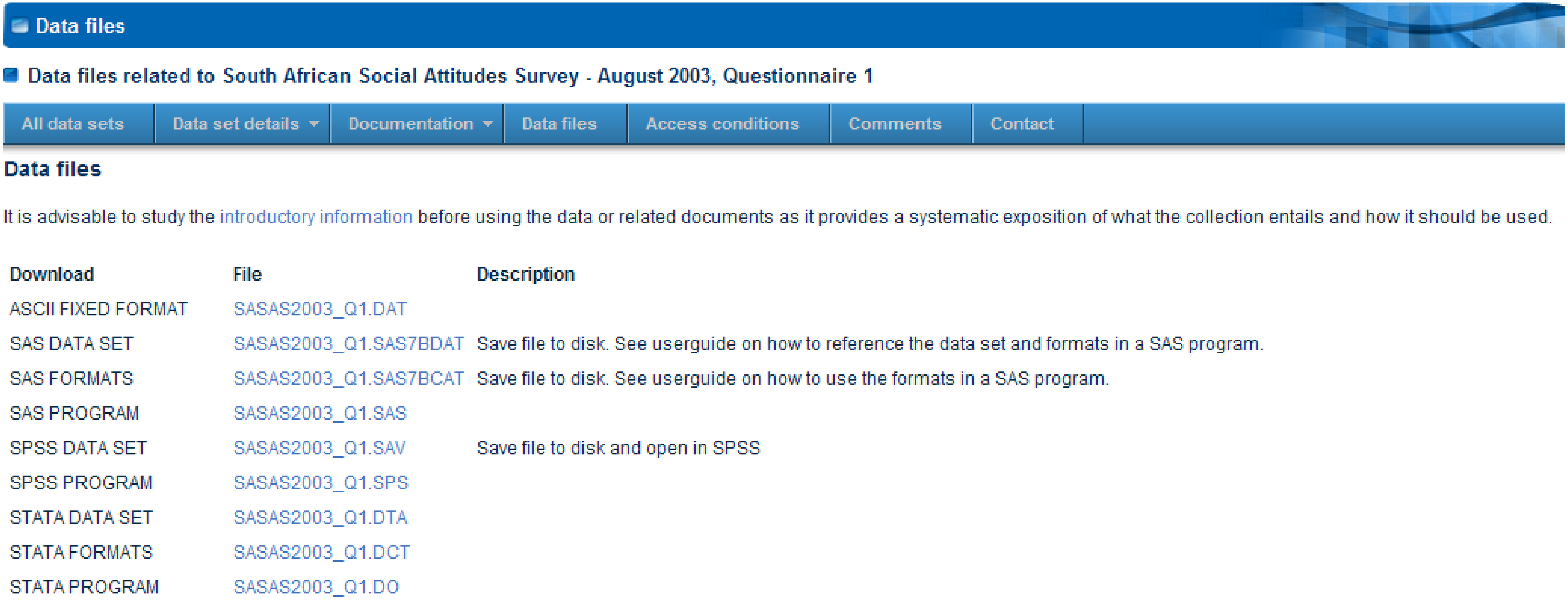
Datasets interface - Edit dataset

Meta status

Data set details

Data set ID:
SASAS 2003_Q1

Data flow technology
Web portal



Data files related to South African Social Attitudes Survey - August 2003, Questionnaire 1

Data files

It is advisable to study the introductory information before using the data or related documents as it provides a systematic exposition of what the collection entails and how it should be used.

Download	File	Description
ASCII FIXED FORMAT	SASAS2003_Q1.DAT	
SAS DATA SET	SASAS2003_Q1.SAS7BDAT	Save file to disk. See userguide on how to reference the data set and formats in a SAS program.
SAS FORMATS	SASAS2003_Q1.SAS7BCAT	Save file to disk. See userguide on how to use the formats in a SAS program.
SAS PROGRAM	SASAS2003_Q1.SAS	
SPSS DATA SET	SASAS2003_Q1.SAV	Save file to disk and open in SPSS
SPSS PROGRAM	SASAS2003_Q1.SPS	
STATA DATA SET	SASAS2003_Q1.DTA	
STATA FORMATS	SASAS2003_Q1.DCT	
STATA PROGRAM	SASAS2003_Q1.DO	

Data dissemination - examples

- **South African HIV/AIDS Behavioural Risks, Sero-Status, and Mass Media Impact Survey (SABSSM)**

  http://www.hsrc.ac.za/Datasets-PFAJLA.phtml
- **The Collaborative HIV/AIDS and Adolescent Mental Health Project (CHAMP)**

  http://www.hsrc.ac.za/Datasets-SAIAAA.phtml
- **South African Social Attitudes Survey (SASAS)**

  http://www.hsrc.ac.za/Datasets-TAAMAA.phtml
- **Trends in International Mathematics and Science Study (TIMMS)**

  http://www.hsrc.ac.za/Datasets-LAAQBA.phtml
- **National Assessment of Learner Achievement (NALA): Grade 9 Systemic Evaluation**

  http://www.hsrc.ac.za/Datasets-LFATBA.phtml

Remaining challenges

- A comprehensive policy framework, associated procedures
- Appropriately skilled staff
- Financial sustainability
- Technology improvements
- Preservation of non-textual data
- Preservation management
- Information products that will aid re-use and uptake of research evidence
- Promotion of secondary data use

The energy source for the light

"A Trusted Digital Repository (TDR) is one whose mission is to provide reliable, long-term access to managed digital resources to its designated community, now and in the future."

Trusted Digital Repositories: Attributes and Responsibilities, 2002

TDR Audit & certification criteria

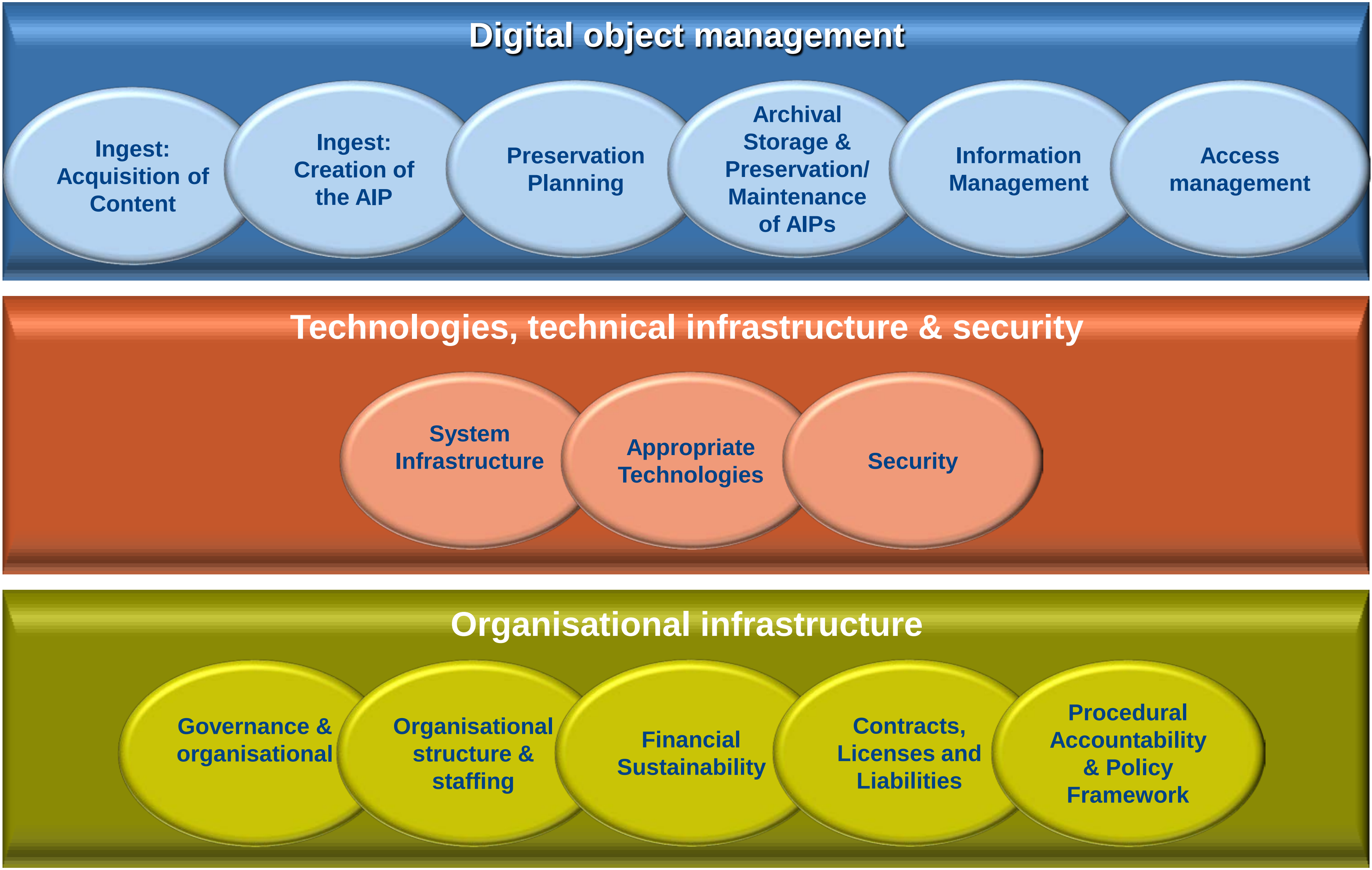
Digital object management

Technologies, technical infrastructure & security

Organisational infrastructure

Based on Trustworthy Repositories Audit & Certification: Criteria and Checklist, 2007

TDR Audit & certification criteria



Digital object management

- Ingest: Acquisition of Content
- Ingest: Creation of the AIP
- Preservation Planning
- Archival Storage & Preservation/ Maintenance of AIPs
- Information Management
- Access management

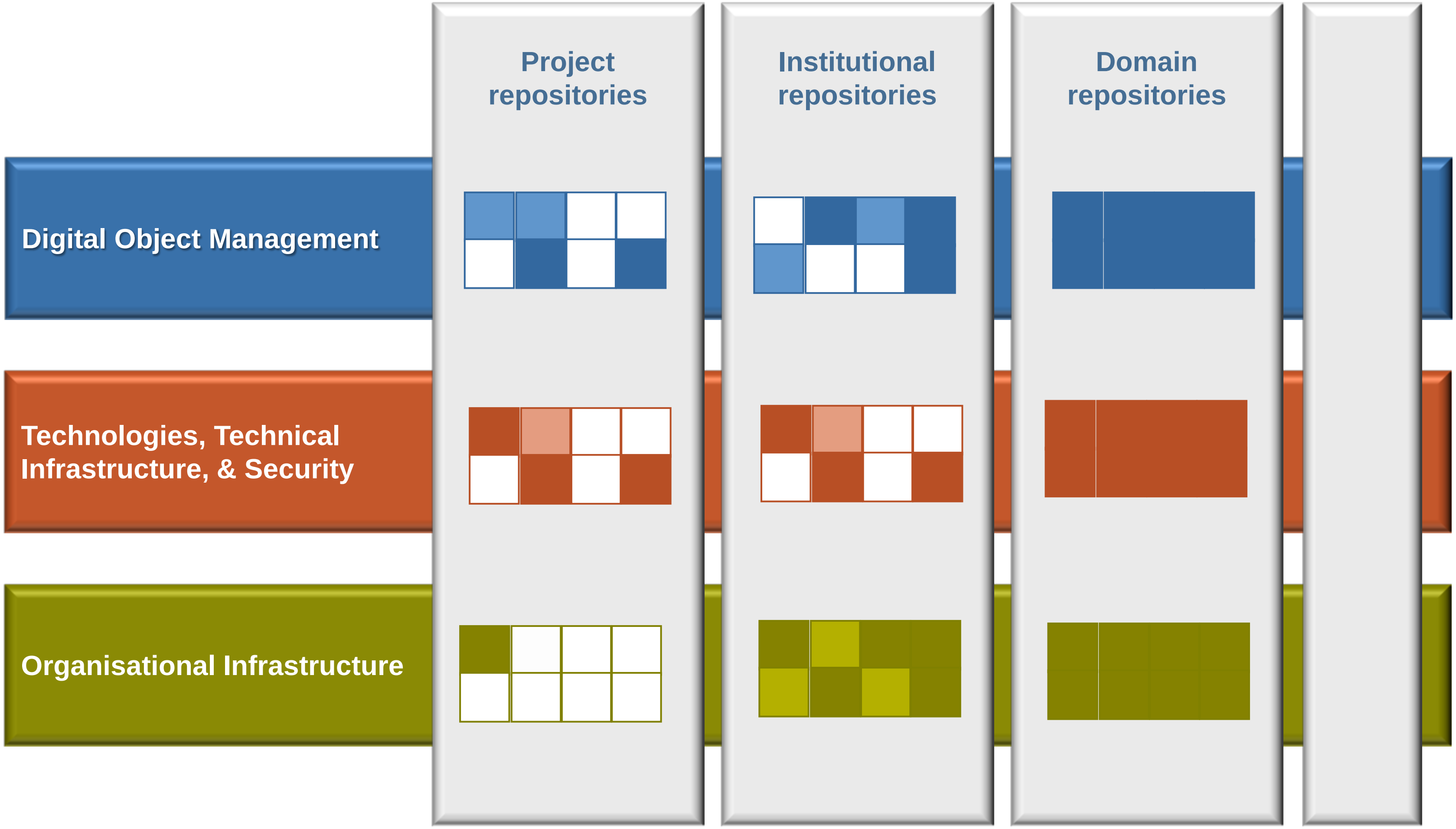
Technologies, technical infrastructure & security

- System Infrastructure
- Appropriate Technologies
- Security

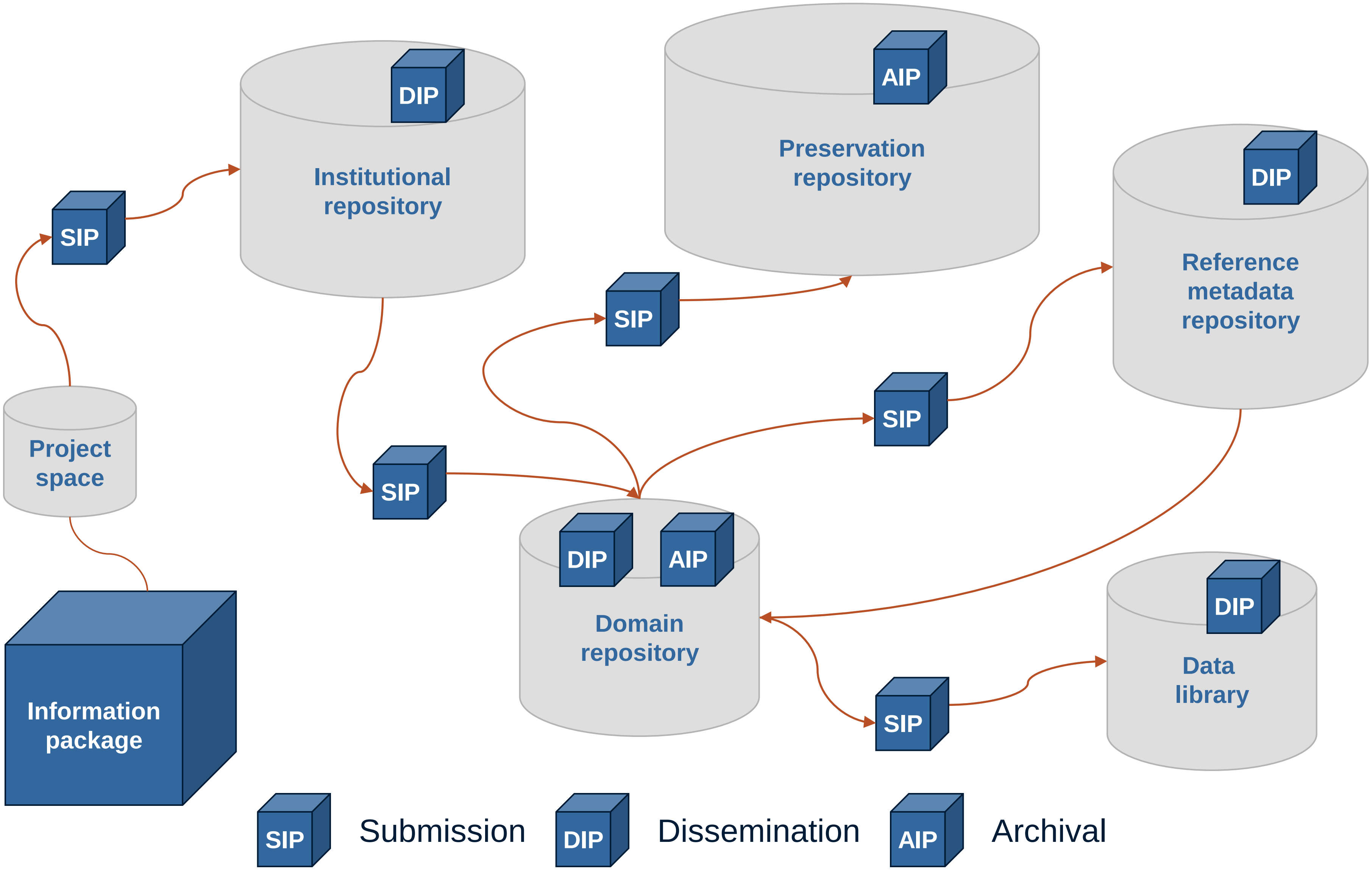
Organisational infrastructure

- Governance & organisational
- Organisational structure & staffing
- Financial Sustainability
- Contracts, Licenses and Liabilities
- Procedural Accountability & Policy Framework

Configuring audit & certification criteria throughout the TDR system


Project repositories
Institutional repositories
Domain repositories
Digital Object Management
Technologies, Technical Infrastructure, & Security
Organisational Infrastructure

Configuring a TDR System



A stronger light

- Embed the curation of social science data in an **e-Research context**
- Curation within a wider research **data management strategy** that covers data from all domains
- The process should be **inclusive**
- The process should be **comprehensive**
- **Progress** = Useful2 – Bad + New

A light tail

On-demand, seamless access to reliable data that is usable over an extended period of time.

Thank you

Building the bridge between research, policy and action

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

References

Research Libraries Group (2002). *Trusted Digital Repositories: Attributes and Responsibilities: An RLG-OCLC Report.* California, USA: RLG Inc.

Lord, P., Macdonald, A., Lyon, L., Giaretta, D. (no date). *From Data Deluge to Data Curation*. no place.: The Digital Archiving Consultancy Limited and the Digital Curation Centre.

OCLC and CRL (2007). *Trustworthy Repositories Audit & Certification: Criteria and Checklist*. Ohio, USA: CRL, The Center for Research Libraries.

Inter-university Consortium for Political and Social Research (ICPSR). (2009). *Guide to Social Science Data Preparation and Archiving: Best Practice Throughout the Data Life Cycle* (4th ed.). Ann Arbor, MI: Inter-university Consortium for Political and Social Research (ICPSR).

Heidorn, P.B. (2008) Shedding light on the dark data in the long tail of science. Library Trends, The Johns Hopkins University Press, 2009, Vol.57(2), 280-299.

IEEE. *The Digital Curation Centre: A vision for digital curation*, 2005.

E.T. Meyer. Moving from small science to big science: Social and organizational impediments to large scale data sharing. *e-Research: Transformation in Scholarly Practice. Routledge, New York*, pages 147–161, 2009.

H. Onsrud and J. Campbell. Big opportunities in access to" small science" data. *Data Science Journal*, 6(0), 2007.

C.L. Palmer. Contouring curation for disciplinary difference and the needs of small science. 2008.