IMPLEMENTING RESEARCH DATA CURATION: THE HSRC EXPERIENCE

Dr Lucia Lötter[1], Ms Anneke Jordaan[2]

[1]Human Sciences Research Council, Pretoria, South-Africa, llotter@hsrc.ac.za

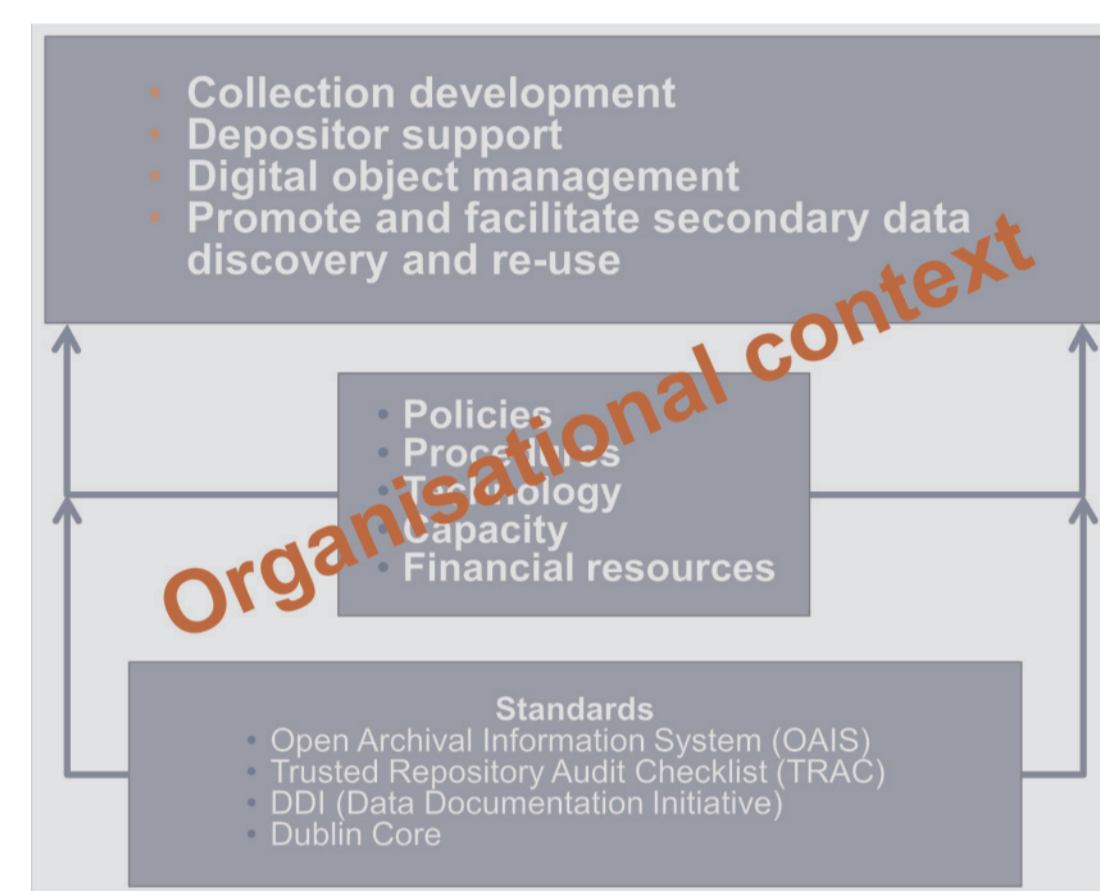[2] Human Sciences Research Council, Pretoria, South-Africa, ajordaan@hsrc.ac.za

HUMAN SCIENCES RESEARCH COUNCIL

HOW IS DATA CURATION GOVERNED IN THE HSRC?

Proper data management has a high priority in the HSRC as data is fundamental to the research undertaken. It is an organisational objective to develop and make available data sets underpinning research, policy development and public discussion of developmental issues. The HSRC intends to make its research data accessible to the maximum extent possible and to ensure its future survival and usability. The curation of RESEARCH data is one of the sentinel performance indicators which apply to the organisation as a whole, each research programme, as well as individual researchers. Performance on this indicator is audited and reported on in terms of the HSRC Strategic Plan.

THE CHALLENGE OF RESEARCH DATA CURATION IMPLEMENTATION

The implementation done at the HSRC consists of the following components: collection development, depositor support, digital object management, the promotion of use of curated data, as well as assistance with secondary data discovery and use. Standards represent the ideal, but inherent in the organisational context, are many obstacles and constraints, but also opportunities that had to be taken on board.

Factors influencing the implementation

- The nature of the social science and humanities data being generated in research programmes
- Organisational culture
- Organisational commitment
- Availability of resources (funding, capacity, technologies, and skilled staff)
  - Provide for a wide and heterogeneous data collection.
  - Embed curation in the research process.
  - Implement formalized structures and processes.

The rules of the game
  - Prevent loss, damage and data becoming unusable.
  - Digital object management has to take into account that the data is highly contextual.
  - Adhere to responsible conduct of research.
  - Cater for legal obligations to funders.

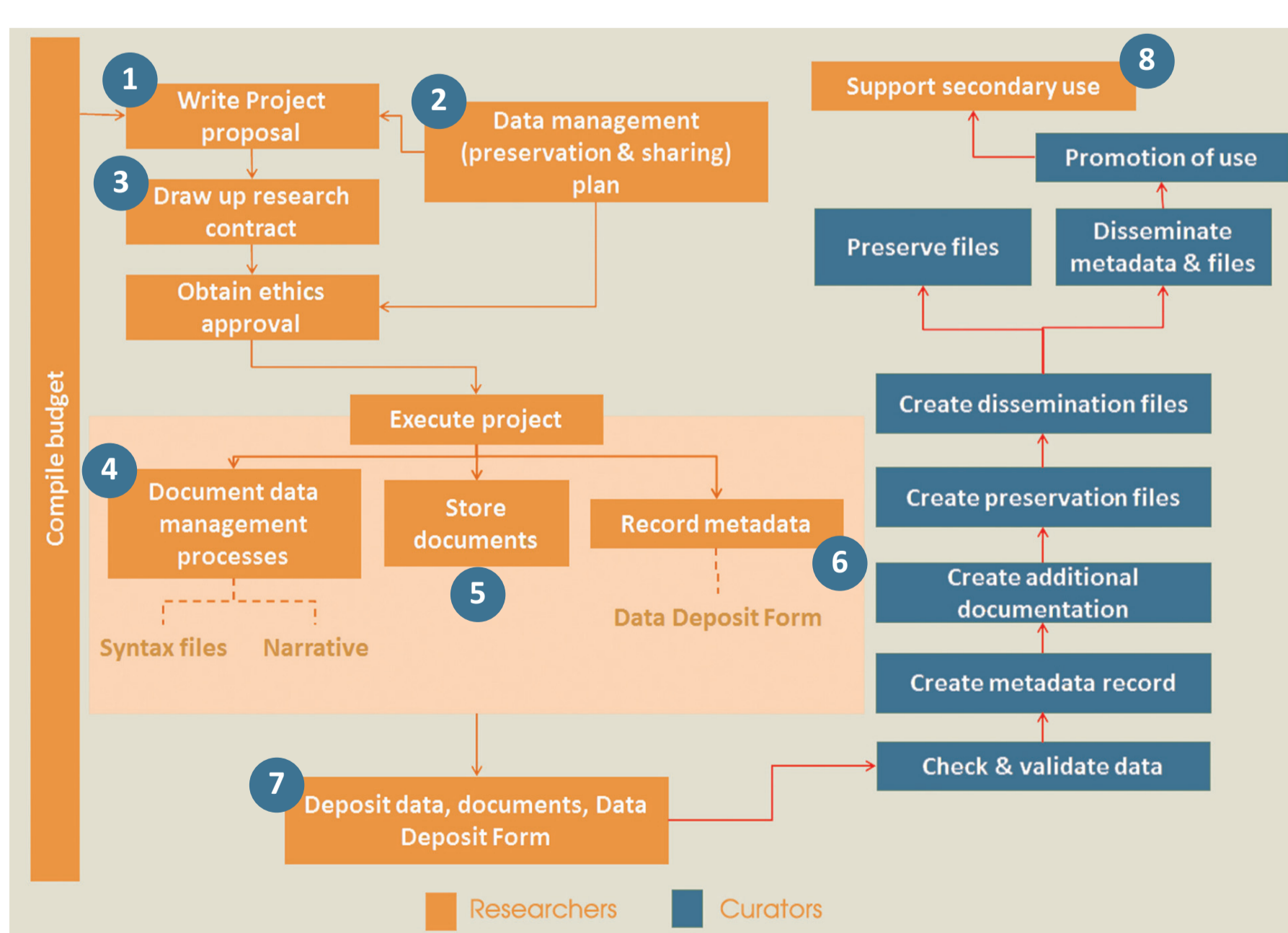### THE IMPLEMENTATION OF RESEARCH DATA CURATION AT THE HSRC

The implementation at the HSRC is an ongoing process and consists of numerous activities around data and process flows, implementing suitable technologies, developing policies and procedures, as well as capacity building and change management.

Existing 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
- An on-line dissemination interface linked to the HSRC's web portal for viewing, downloading or analysis
- Processes to monitor and audit curated data sets for performance information purposes
- Various data sets available for secondary use

PROCESS FLOW
The high level process flow within which data curation activities take place
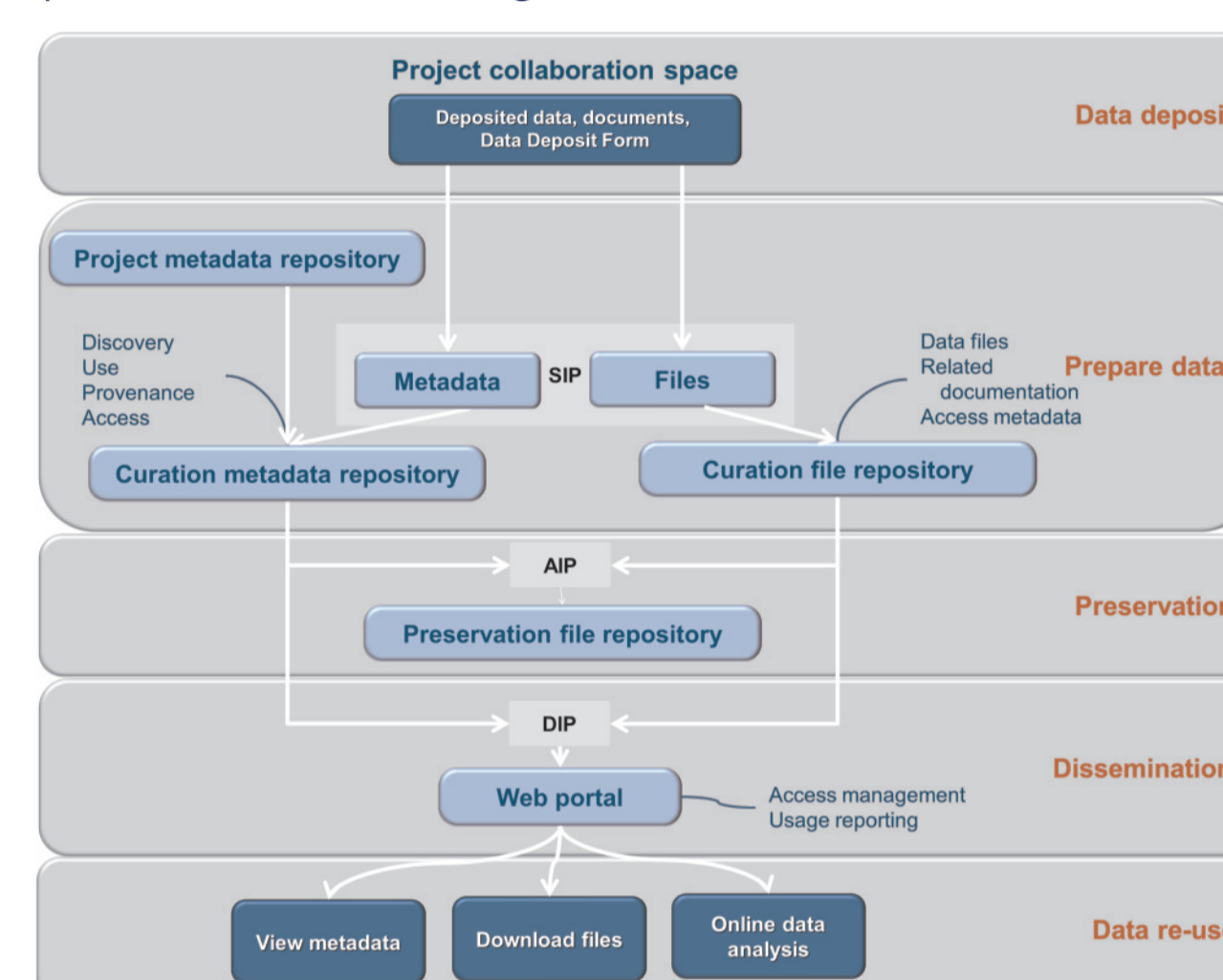
CURATION ACTIVITIES

Check and validate data	• Validation - error detection
	• Cleaning - correcting errors
	• Done by researchers
	• Document issues
	• Check anonymisation
Create data documentation	• Create catalogue record
	• Enhance internal metadata (variable and value labels)
	• Develop contextual information to facilitate re-use
	• Catalogue related documentation
	• Link research outputs to data

Disseminate data	• Create dissemination formats
	• Set access parameters
Preservation (short term)	• Preservation strategies
	• Create preservation formats
	• Storage and storage management (multiple copies, multi-media)
	• Back-up, disaster recovery
	• Security

The curation **DATA FLOW** constitutes the sourcing of files from a research collaboration storage area, capturing of information about data and documents in a metadata repository, storing all files in a curation file repository, disseminating data from the HSRC's website and moving files designated for preservation to a preservation storage area. Curation activities are performed according to Standard Operating Procedures.
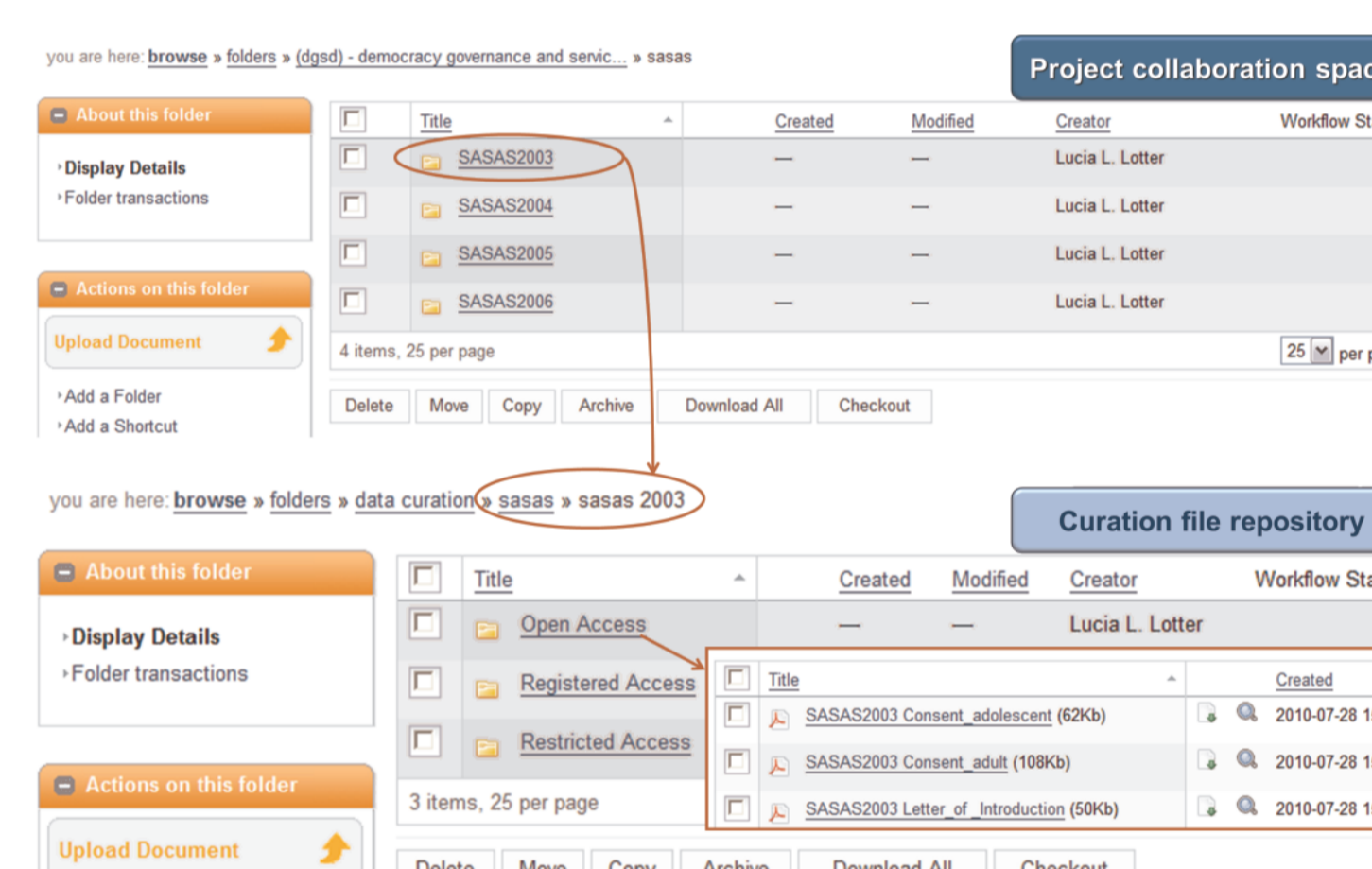
Information Package variants

- Submission Information Package (SIP)
- Archival Information Package (AIP)
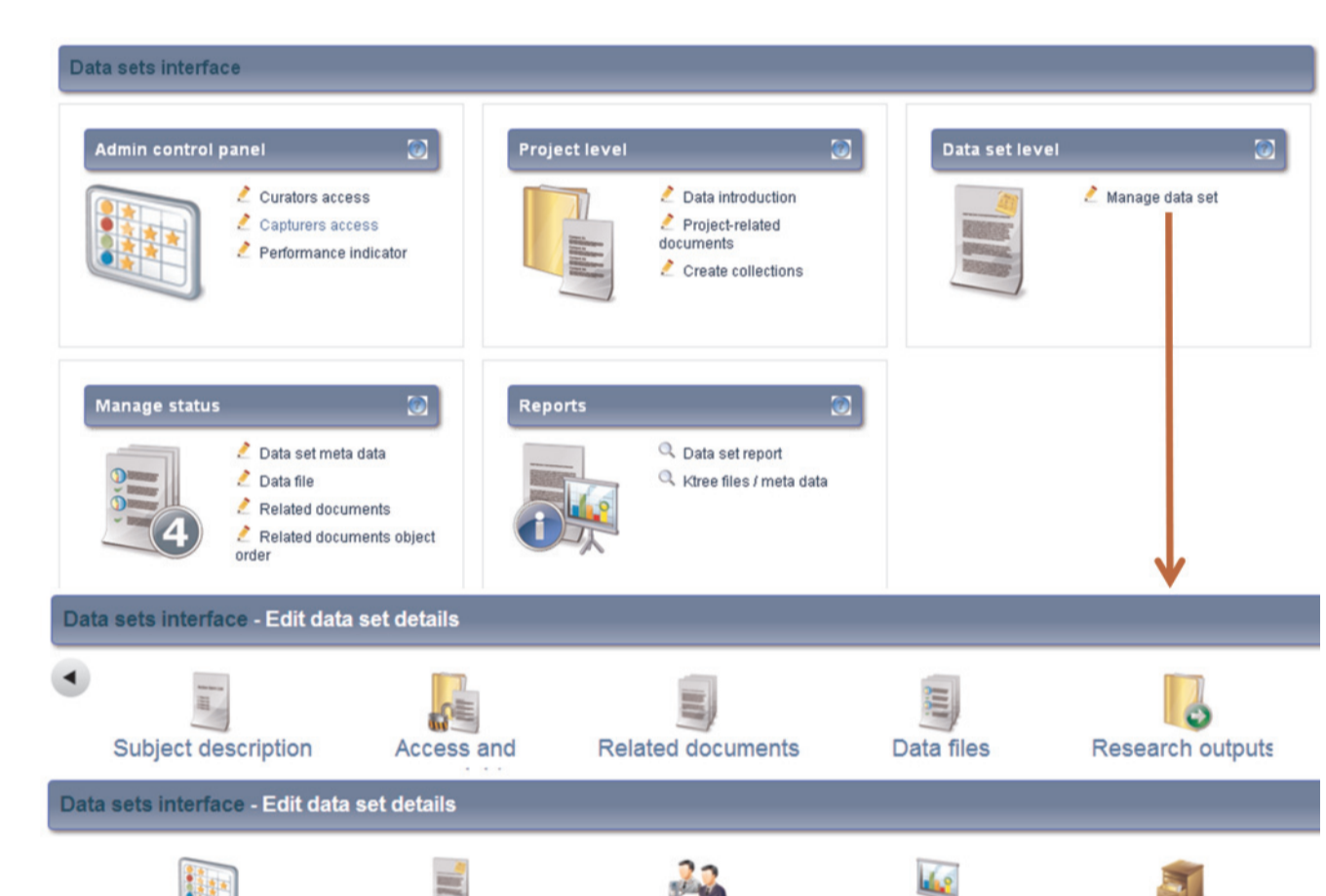- Dissemination Information Package (DIP)

The data curation process is supported by the connectivity, security, backup and disaster recovery services provided by the IT department of the HSRC.
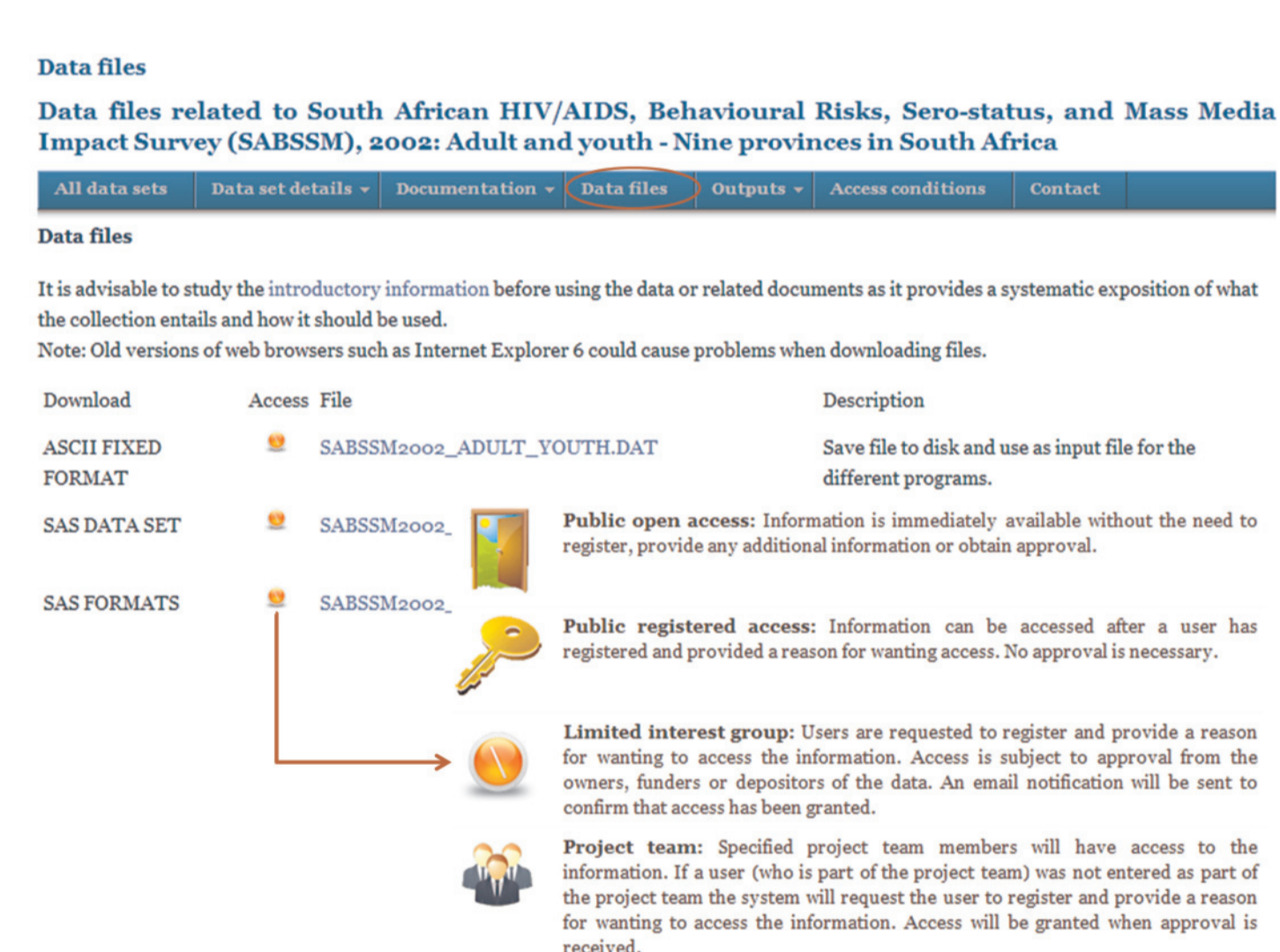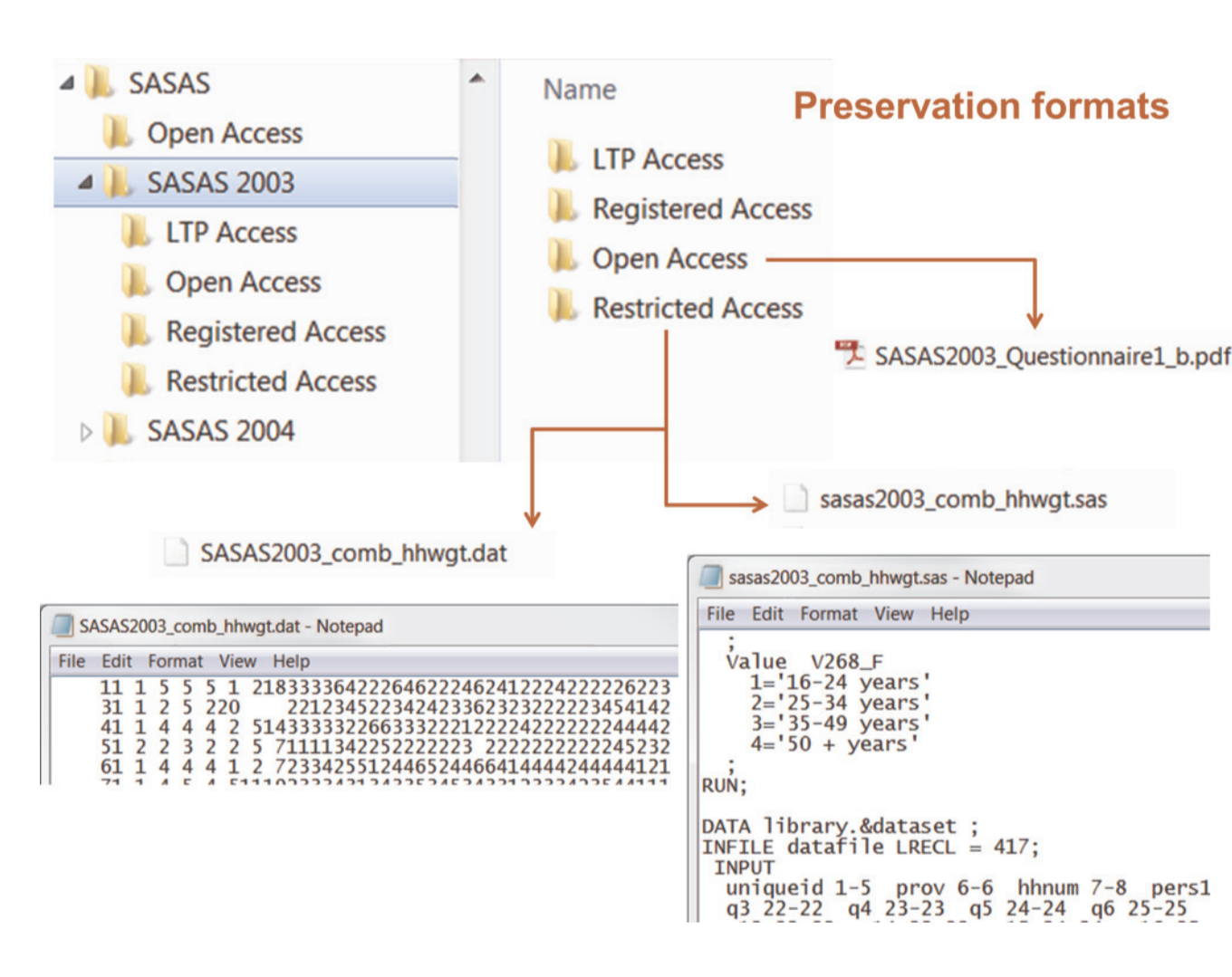
CURATION TECHNOLOGIES
FILE REPOSITORY

METADATA REPOSITORY

DISSEMINATION ON THE HSRC'S WEBSITE

PRESERVATION SERVER

CONCLUSIONS
The lessons learnt

- Successes: awareness raising and initial commitment from the organisation; making the most of an initial investment in terms of technology and curation practice; developing a competent and motivated team.
- Areas of improvement: policy development; data management and documentation training; technology development; preservation management; automation; sustainability.
- Data curation implementations will vary according to the organisational context.
- Reiterative implementation process is necessary, including continuous change management and capacity building.
- Information technology is an enabler and not an end in itself.
- Long term preservation is probably beyond the capacity of many data producers and will require investment on a large scale.

Attempting to provide on-demand, seamless access to reliable data that is usable over an extended period of time is a wide-ranging endeavor which requires continuous effort, but can realize benefits on various fronts for researchers, research organisations and ultimately society on whose behalf research is conducted.

REFERENCES

1. SOUTH AFRICA, 2008, Human Sciences Research Council Act 2008. Cape Town.
2. Human Sciences Research Council 2013. Human Sciences Research Council Strategic Plan For The Fiscal Years 2012 - 2017. Pretoria South Africa: Human Sciences Research Council.
3. Human Sciences Research Council 2013. Information Policies and Procedures, HSRC Data Sharing Policy. Pretoria South Africa: Human Sciences Research Council.
4. IEEE. The Digital Curation Centre: A vision for digital curation, 2005.
5. CCSDS.2002. 'Reference model for an Open Archival Information System (OAIS).' Blue Book CCSDS 650.0-B-1, Consultative Committee for Space Data Systems. Also published as ISO 14721:2003.
6. OCLC and CRL. Trustworthy Repositories Audit & Certification: Criteria and Checklist. Ohio, USA: CRL, The Center for Research Libraries, 2007.
7. Heidorn, P.B. 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.
8. H. Onsrud and J. Campbell. Big opportunities in access to "small science" data. Data Science Journal, 6(0), 2007.
9. Palmer, C.L. Contouring curation for disciplinary difference and the needs of small science. Sun Preservation & Archiving Special Interest Group (PASIG) meeting, November 19-21, 2008, Baltimore, Maryland.