Human Sciences Research Council

IMPLEMENTING RESEARCH DATA CURATION:

THE HSRC EXPERIENCE

Dr Lucia Lötter¹, Ms Anneke Jordaan²

¹Human Sciences Research Council, Pretoria, South-Africa, Ilotter@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.

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

THE CHALLENGE OF RESEARCH DATA CURATION IMPLEMENTATION

Collection development

discovery and re-use

ositor support al object management note and facilitate secondary data

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.

- 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.



CURATION TECHNOLOGIES FILE REPOSITORY

About this folder		Title	-	Created	Modified	Creator	Workflow State
Display Details		SASAS2003	7	-	-	Lucia L. Lotter	
Folder transactions		SASAS2004		-	-	Lucia L. Lotter	
		SASAS2005		—	_	Lucia L. Lotter	
Actions on this folder		SASAS2006		-	_	Lucia L. Lotter	
Upload Document 👘 🔶	4 items	s, 25 per page					25 🕶 per pag
≻Add a Folder ≻Add a Shortcut	Delete	e Move Copy Arch	ive Do	wnload All Che	ckout		

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.

METADATA REPOSITORY

Admin control panel	Project level	0	Data set level	
Curators access		Data introduction Project-related		🙎 Manage data set
Performance indicato	r Hillinger	documents <i>2</i> Create collections	Bananan	
Manage status	Reports	0		

- 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



About this folder		Title		^		Created	Modified	Creator		٧	Vorkflow State
Display Details		Dpen	Access			—	-	Lucia L. L	otter		
Folder transactions		📄 Regis	tered Access		Title				•		Created
		- Dostri	icted Access		P	SASAS2003 Con	sent_adolescent	(62Kb)		Q,	2010-07-28 15:16
Actions on this folder		<u>Kestii</u>	Cied Access		R	SASAS2003 Con	sent_adult (108k	(b)		Q	2010-07-28 15:17
	3 iten	ns, 25 per pa	ige		R	SASAS2003 Lette	er_of _Introduction	on (50Kb)		Q,	2010-07-28 15:46
Upload Document 🎐	Delete Move Copy Archive Download All Checkout										

DISSEMINATION ON THE HSRC'S WEBSITE

Data files

Data files related to South African HIV/AIDS, Behavioural Risks, Sero-status, and Mass Media Impact Survey (SABSSM), 2002: Adult and youth - Nine provinces in South Africa

All data sets	Data set de	tails 👻 Do	cumentation •	Data files	Outputs 🗸	Access conditions	Contact	
Data files								
the collection enta	ils and how it	should be us	sed.			ments as it provides a s n downloading files.	ystematic exp	position of what
Download	Access	File				Description		
ASCII FIXED FORMAT		SABSSM20	002_ADULT_Y	OUTH.DAT		Save file to disk and u different programs.	ise as input file	e for the
SAS DATA SET	9	SABSSM20	002_			nation is immediately a al information or obtain		out the need to
SAS FORMATS	9	SABSSM20	002_			: Information can be son for wanting access. I		
			→ <u>()</u>	for wanting to owners, funde	access the in	sers are requested to r formation. Access is s rs of the data. An ema granted.	ubject to app	proval from the
				information. If the project tea	f a user (who is am the system v	project team member part of the project tea vill request the user to formation. Access will	m) was not en register and p	tered as part of provide a reason



PRESERVATION SERVER

 SASAS Open Access SASAS 2003 LTP Access Open Access Registered Access Restricted Access SASAS 2004 	📕 Open A	red Access
SASAS2003_comb_hhwgt.da	at	sasas2003_comb_hhwgt.sas
SASAS2003_comb_hhwgt.dat - Notepad File Edit Format View Help 11 1 5 5 1 218333364222646222462 31 1 2 5 220 221234522342423362 41 1 4 4 2 514333332266333222122 51 2 3 2 2 7111134225222223 2 61 1 4 4 1 2 7233425512446524466 71 4 5 4 511100000040104000000000000000000000000		; Value V268_F 1='16-24 years' 2='25-34 years' 3='35-49 years' 4='50 + years' ; RUN; DATA library.&dataset ; INFILE datafile LRECL = 417; INPUT uniqueid 1-5 prov 6-6 hhnum 7-8 pers1 q3 22-22 q4 23-23 q5 24-24 q6 25-25

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.

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

- 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.

SOCIAL SCIENCE THAT MAKES A DIFFERENCE

www.hsrc.ac.za