Trust is in the detail! Curating data to ensure integrity and authenticity

A case study

Dr Lucia Lötter

SciDataCon Session: Improving Data Repository Trustworthiness

International Data Week 2018

Gabarone, Botswana

6 November 2018

HSRC Human Sciences Research Council | eRKC eResearch Knowledge Centre

The need for ensuring data integrity and authenticity

The Human Sciences Research Council (HSRC)

- develops and makes research data sets available which underpin research, policy development and public discussion of developmental issues.
- shares data resulting from research undertaken for research and educational purposes with researchers, academics and students.

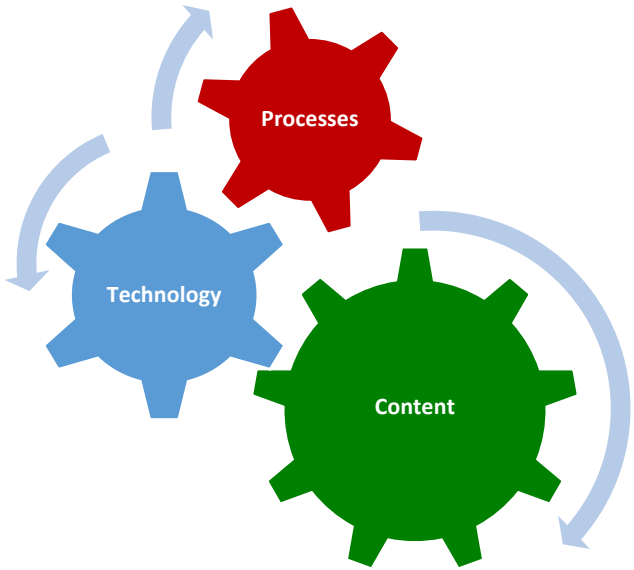
Prerequisite: Data must be trustworthy!

http://datacuration.hsrc.ac.za

HSRC Human Sciences Research Council | eRKC eResearch Knowledge Centre

The need for ensuring data integrity and authenticity

- Secondary data users must be assured that the data which they will be using is accurate, complete, reliable and consistent
- Measures to mitigate risks emanating from people, processes, technology and information



What is integrity and authenticity?

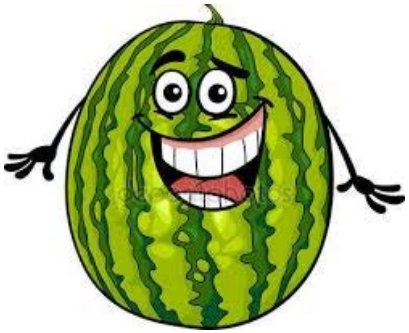
Authenticity

The degree to which a person (or system) regards an object as what it is purported to be.

Integrity

Internal consistency or lack of corruption of digital objects. Integrity can be compromised by hardware errors even when digital objects are not touched, or by software or human errors when they are transferred or processed.

(DSA & ICSU, 2016)



Integrity and authenticity

- Span the entire research and curation process (data life cycle)
- It is an essential function of a Trusted Digital Repository

Data Seal of Approval

ICSU WORLD DATA SYSTEM

Digital Object Management

VII. Data integrity and authenticity

R7. The repository guarantees the integrity and authenticity of the data.

- Essential to digital object management – specifically when information submitted by producers are accepted and prepared for inclusion in the repository

HSRC Human Sciences Research Council | eRKC eResearch Knowledge Centre

Evaluation in terms of RDR criteria

Integrity management

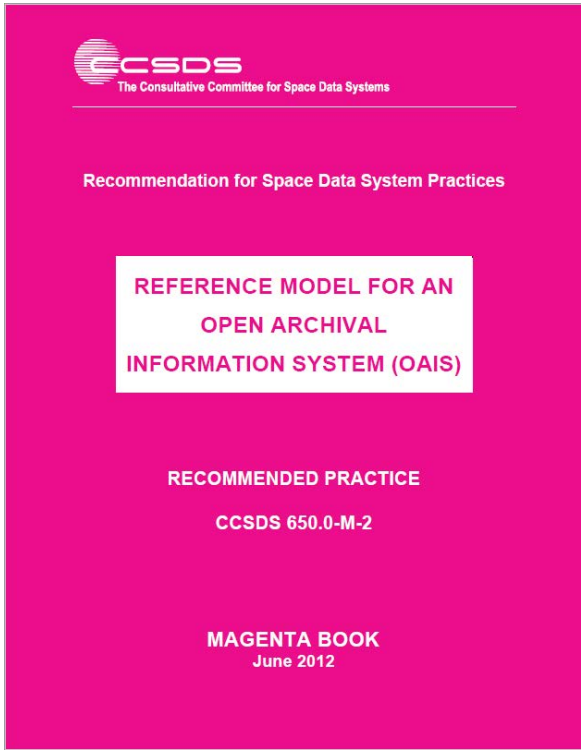
- Description of checks to verify that a digital object has not been altered or corrupted.
- Documentation of the completeness of the data and metadata.
- Details of how all changes to the data and metadata are logged.
- Description of version control strategy.
- Usage of appropriate international standards and conventions (which should be specified).

Authenticity management

- Does the repository have a strategy for data changes?
- Does the repository maintain provenance data and related audit trails?
- Does the repository maintain links to metadata and to other datasets?
- Does the repository compare the essential properties of different versions of the same file?
- Does the repository check the identities of depositors?

What is “Ingest”?

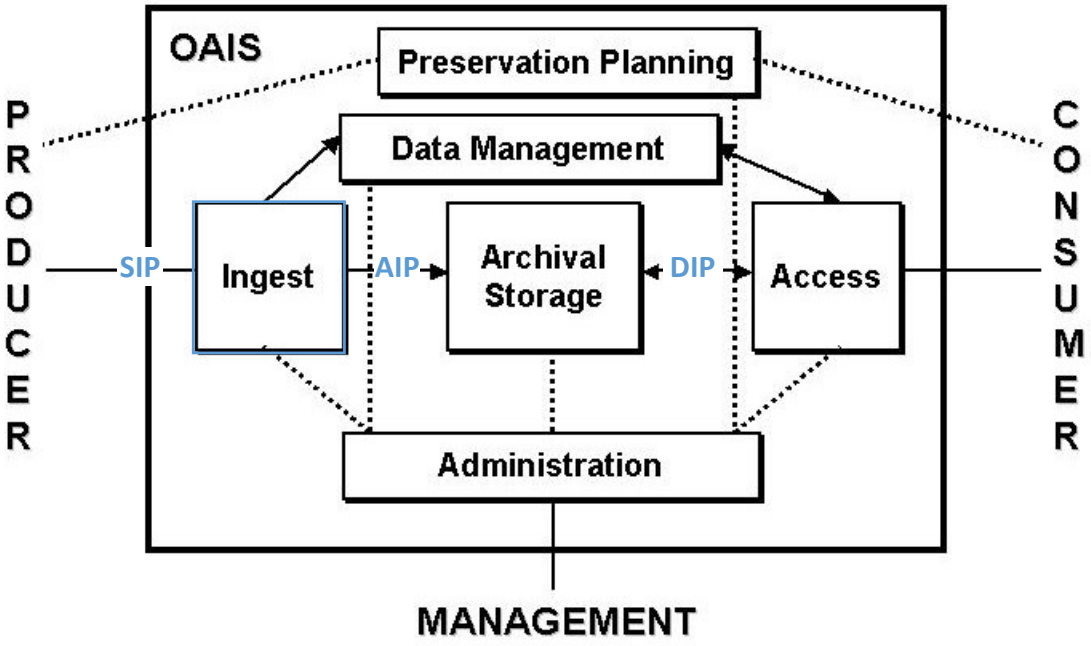
- Digital object management function described in the OAIS



What is “Ingest”?

“Ingest is the set of processes responsible for accepting information submitted by Producers and preparing it for inclusion in the archival store. ”

(Lavoie, 2014)



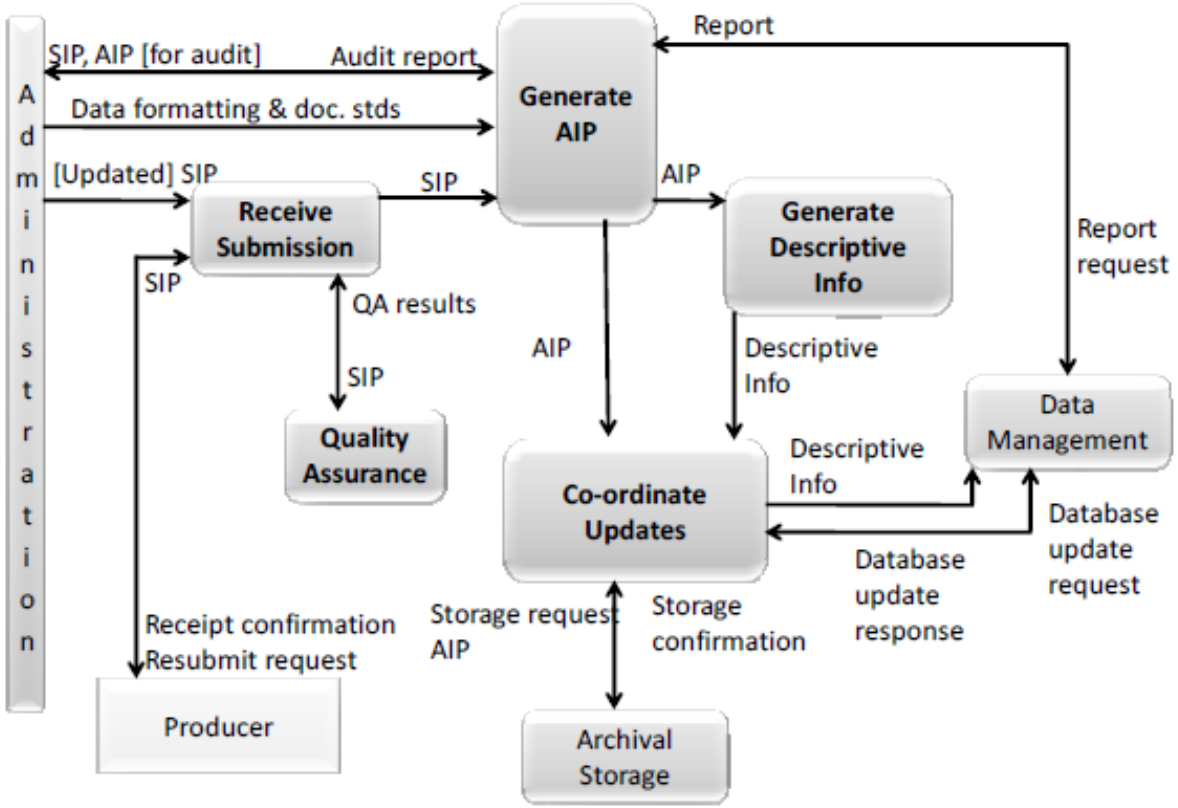
(Lavoie, 2014)

SIP **Submission Information Package**

AIP **Archival Information Package**

DIP **Dissemination Information Package**

The “Ingest” function

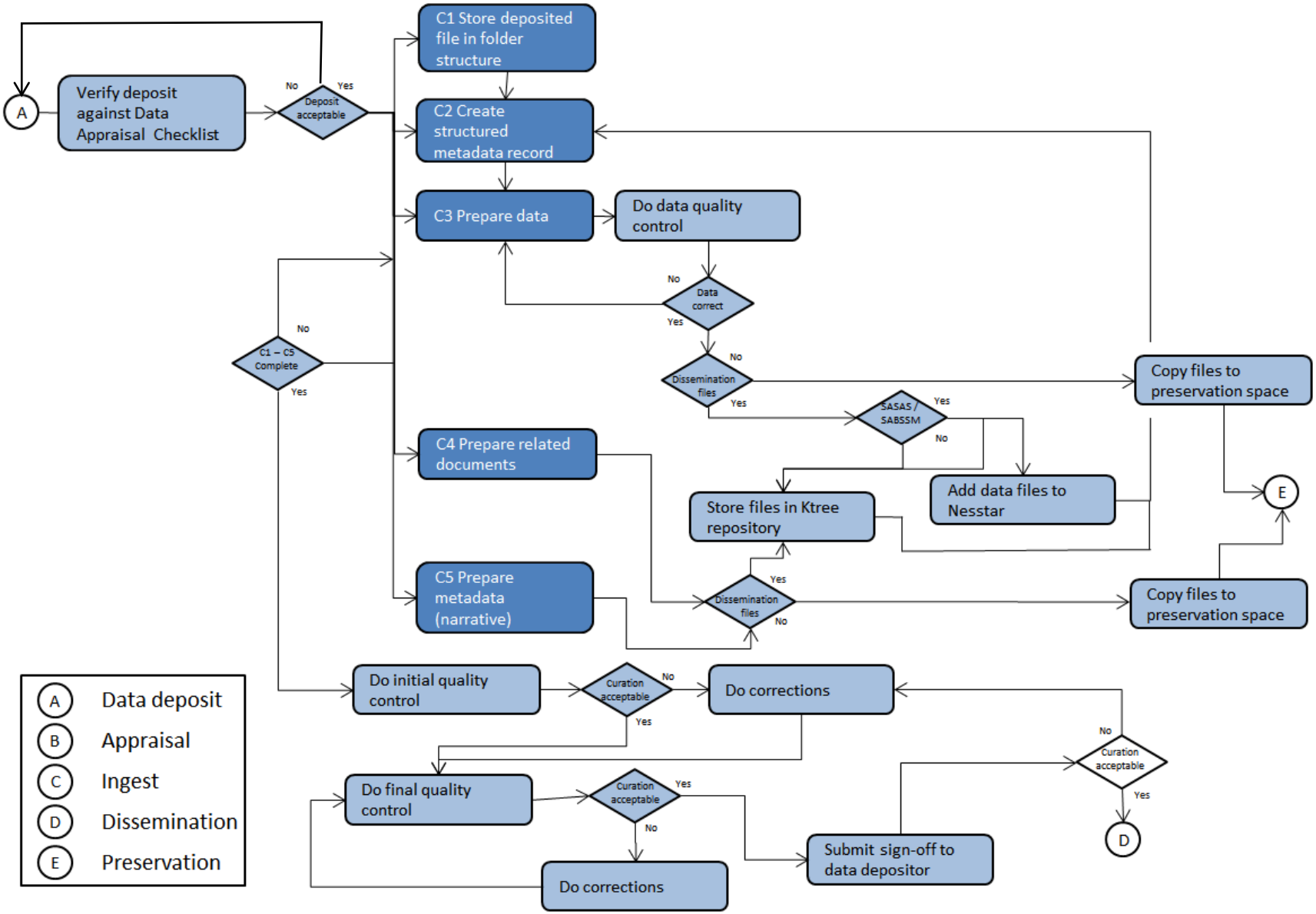


(CCSDS, 2012)

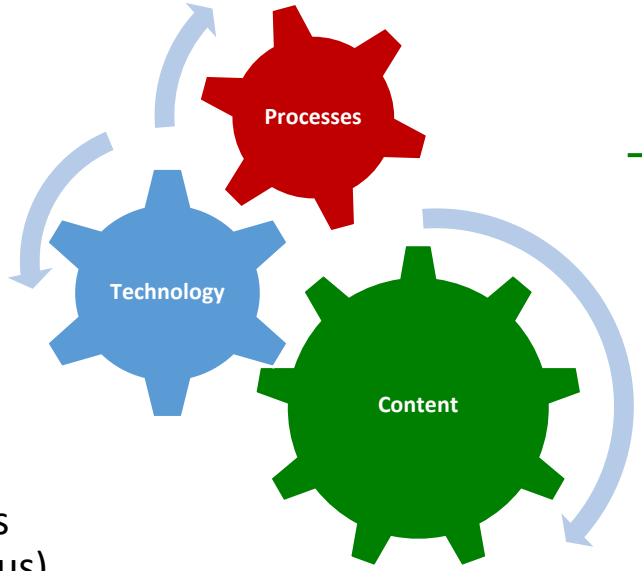
- Receipt of SIP
- Validation that the SIP is uncorrupted and complete
- Transformation of SIP into a form suitable for storage and management within the archival system
- Extraction and/or creation of descriptive metadata
- Transfer AIP and DIP and its associated metadata to the archival store

(Lavoie, 2014)

The HSRC's "Ingest" workflow



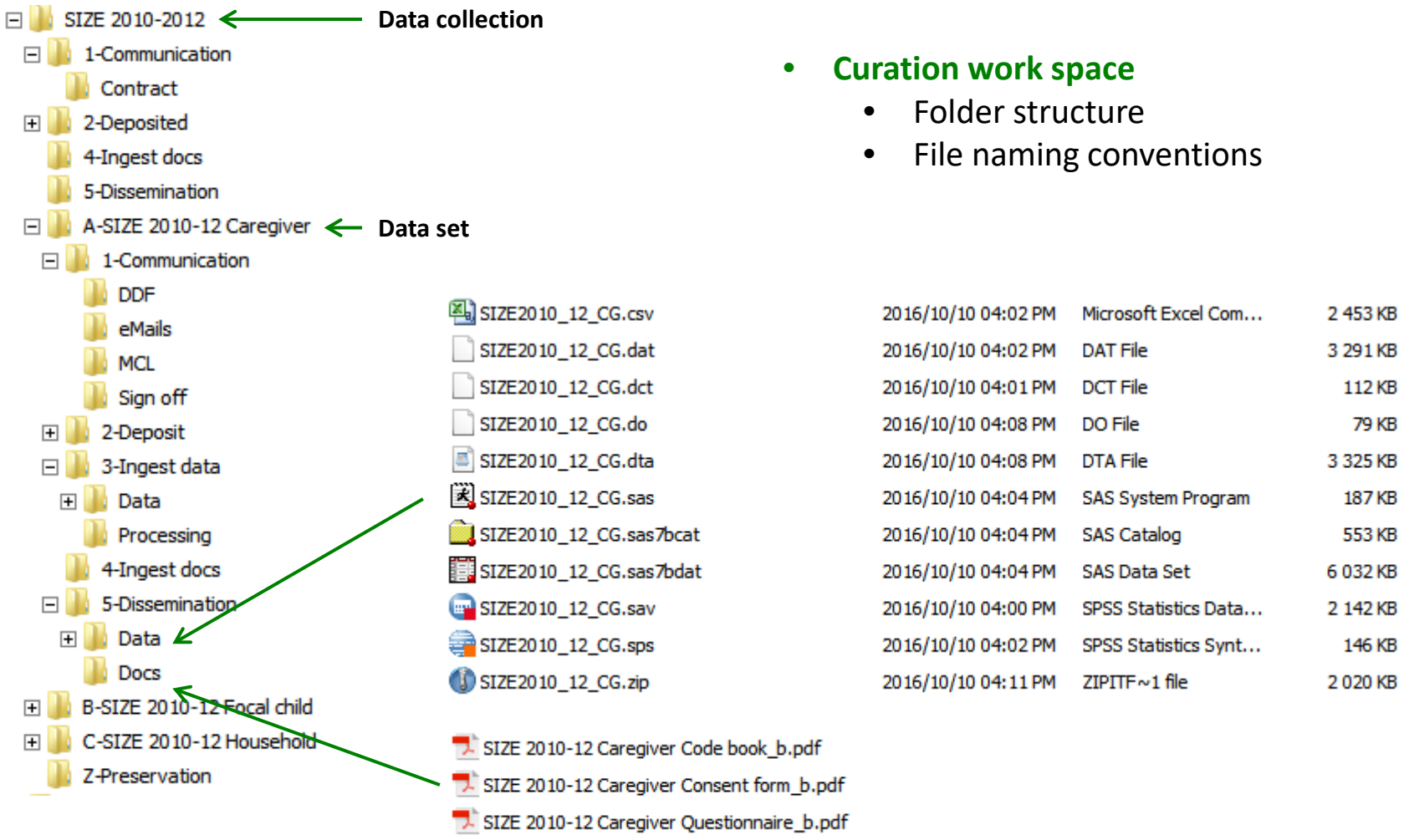
Mitigation of risk



- Risks
  - Transfer of custody
  - Corruption, loss
  - Unauthorised changes (Accidental or malicious)

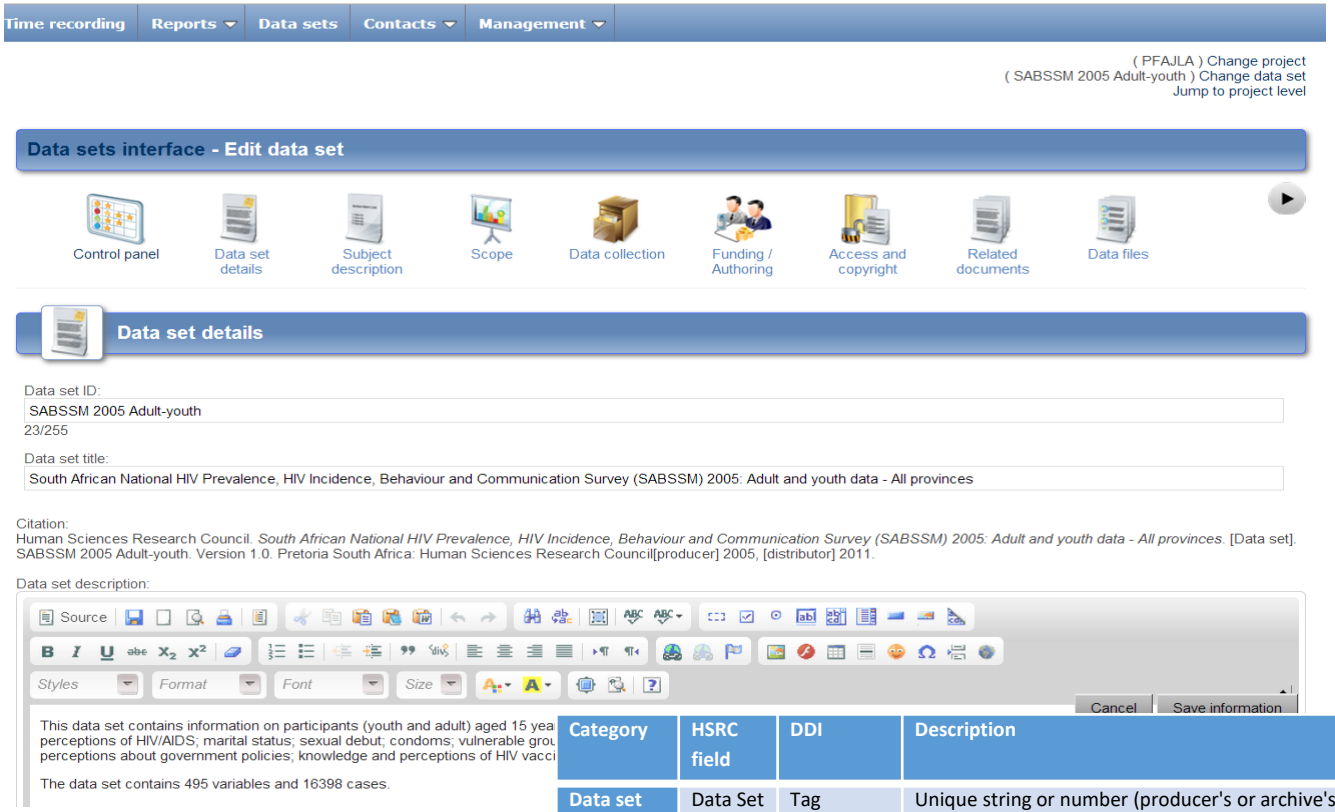
- Mechanisms
  - Checking mechanisms
  - Checksum
  - Quality assurance processes
  - Record keeping
  - Security
  - Audit trail
  - Tools (Software, SOP, Checklists, templates)

How do we process files?



- **Curation work space**
  - Folder structure
  - File naming conventions

How do we compile metadata?



- **Structured metadata record**
  - Standards (DDI, Dublin Core)
  - Management parameters
- Aid understanding, management
- Tools: Metadata repository, Conventions, Style guide

Category	HSRC field	DDI	Description	Examples
Data set details	Data Set ID	Tag < IDNo > Qualifiers Agency	Unique string or number (producer's or archive's number). Input guidelines • Composition: Acronym of collection (upper case), Space, Year of data collection, Space, Reference to data set content. • Descriptives capitalised, rest lower case. • Linked descriptives and multiple years indicated with unspaced hyphen – not 2006/07, 2006-2007.	<IDNo agency=" HSRC ">xxx</IDNo> Examples • SASAS 2008 Q1 • SABSSM 2005 Adult-youth • EC-PSS 2008 • SASAS 2003 Combined Household weight • SASAS 2004 Combined Weight • RnD 2006-07 • NALA 2009 Gr9 Natural sciences assessment

How do we compile metadata (narrative)?

SIZE 2010-12 Readme.txt - Notepad

File Edit Format View Help

```
HUMAN SCIENCES RESEARCH COUNCIL:  STUDY  INFORMATION


Title: We look out for our children (SIZE) 2010-12. Msunduzi Municipality - KwaZulu-Natal


HSRC Processing standards
-------------------------
All survey responses were recorded electronically on mobile phones. The commercially available Mobenzi Researcher m
survey software and data management portal were used (www.clyral.com). Mobenzi Researcher is a Java 2 Micro Edition
application and provides full survey functionality, including the ability to create various question types, mark fi
mandatory and intelligently manage survey branching.

Conversion of documentation
---------------------------
All electronic documentation is supplied with this study on the HSRC web site in PDF format.
The conversion programmes used are the latest versions of Adobe PDF Professional for electronic documentation.

The following data sets are part of this data collection
-------------------------------------------------------
SIZE2010_12_CG (Caregiver)
SIZE2010_12_FC (Focal child)
SIZE2010_12_HH (Household)

Conversion of data
------------------
Ingest format(s) of the data =  SPSS  (.sav)
Data sets are provided in the available version of the software at the time of

All data files created are subject to visual checking by a data processor.
The data sets that are provided for further analysis, contain micro data and a
Stata and ASCII. For additional information on these data, please refer to the


CONTEXTUAL INFORMATION
----------------------
This information is provided in the following documents:

USER GUIDE
```

- **Metadata in narrative**
  - Readme, User guide
  - Conventions, Style guide
  - Aid understanding, facilitate use
  - Accuracy, Completeness

SIZE 2010-12 User guide_b.pdf - Adobe Acrobat Reader DC

File Edit View Window Help

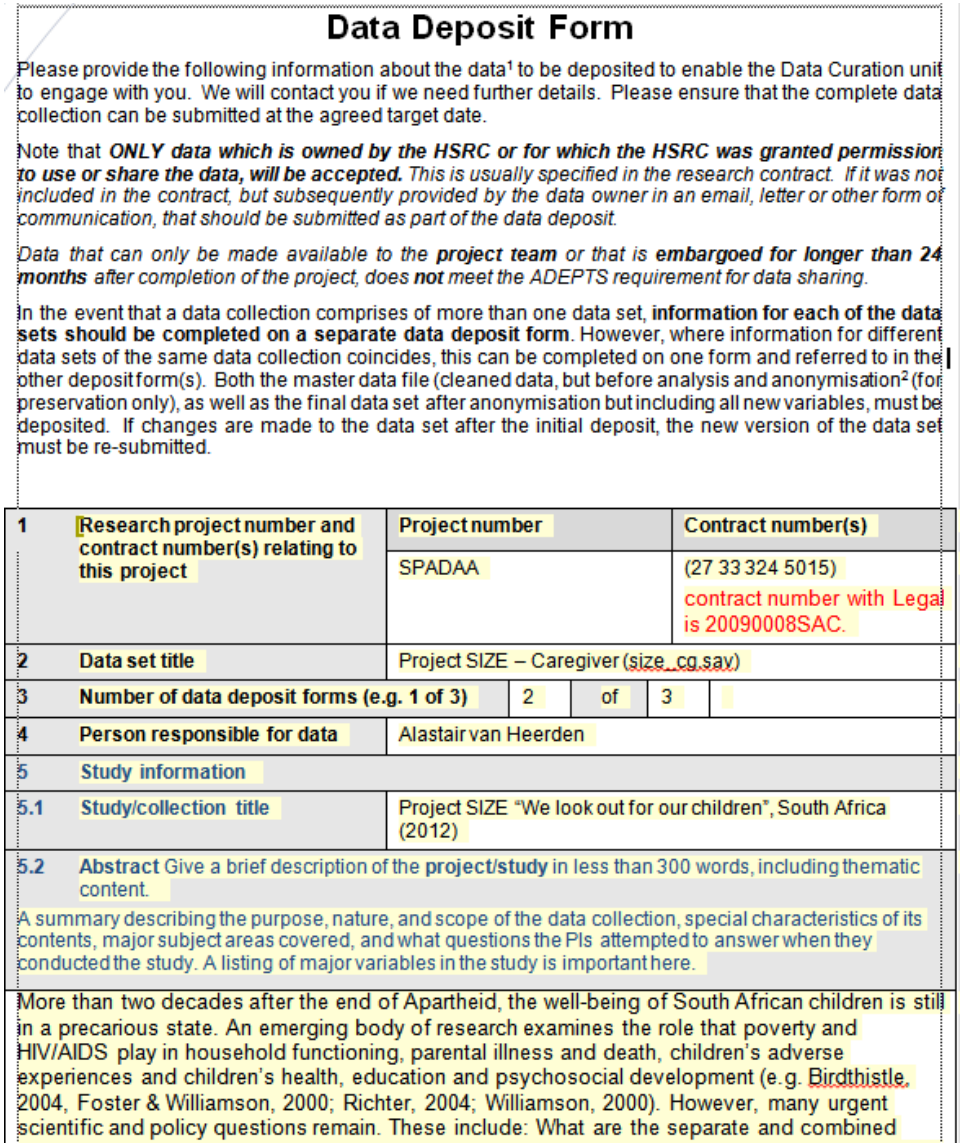
Home Tools SIZE 2010-12 User ... ×

2 / 10 100%

This file claims compliance with the PDF/A standard and has been opened read-only to prevent modification. Enable Editing

Table of Contents

How do we source metadata?



Data Deposit Form

Please provide the following information about the data[1] to be deposited to enable the Data Curation unit to engage with you. We will contact you if we need further details. Please ensure that the complete data collection can be submitted at the agreed target date.

Note that ***ONLY data which is owned by the HSRC or for which the HSRC was granted permission to use or share the data, will be accepted.*** *This is usually specified in the research contract. If it was not included in the contract, but subsequently provided by the data owner in an email, letter or other form of communication, that should be submitted as part of the data deposit.*

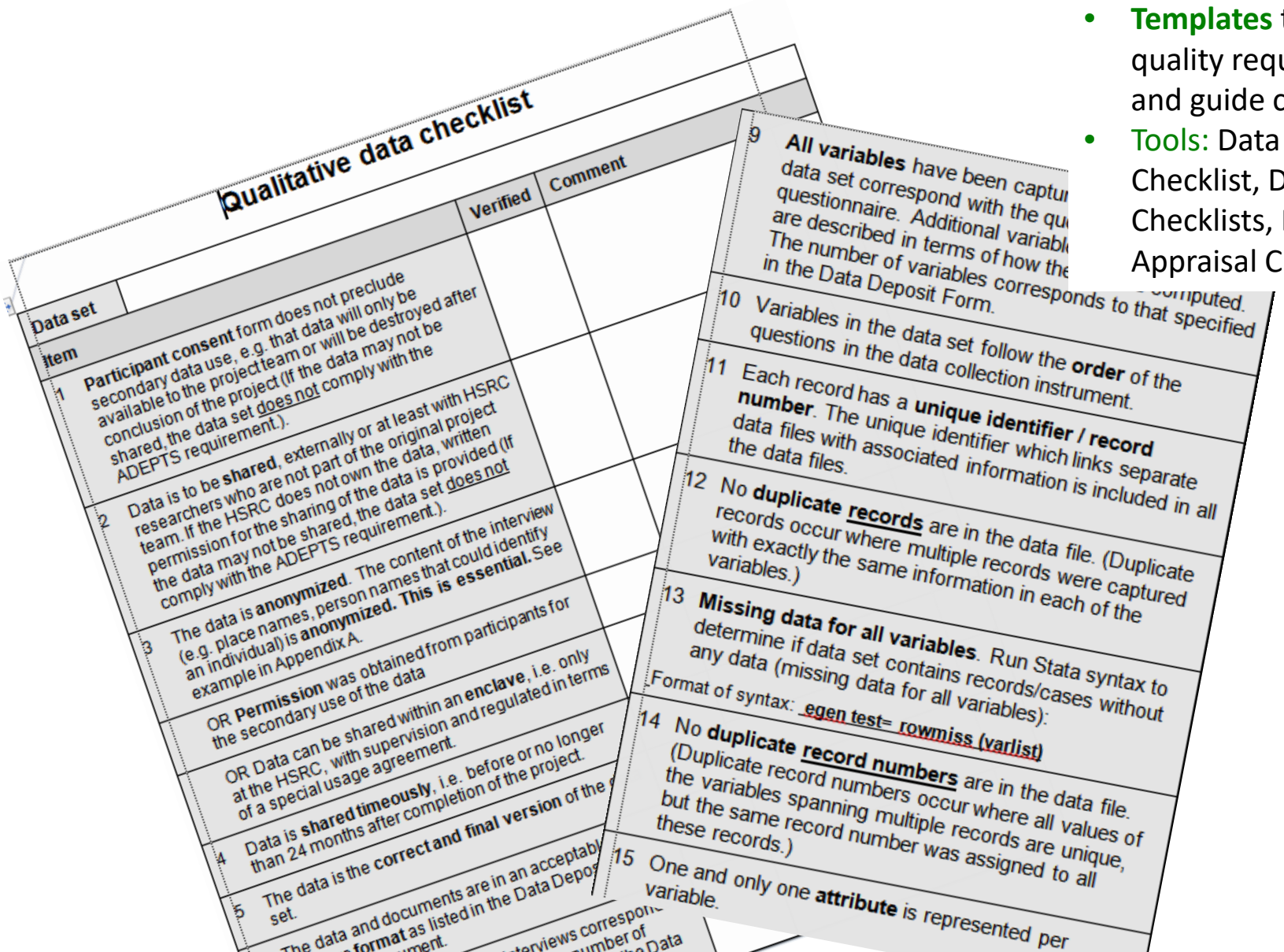
Data that can only be made available to the ***project team*** *or that is* ***embargoed for longer than 24 months*** *after completion of the project, does* ***not*** *meet the ADEPTS requirement for data sharing.*

In the event that a data collection comprises of more than one data set, **information for each of the data sets should be completed on a separate data deposit form**. However, where information for different data sets of the same data collection coincides, this can be completed on one form and referred to in the other deposit form(s). Both the master data file (cleaned data, but before analysis and anonymisation[2] (for preservation only), as well as the final data set after anonymisation but including all new variables, must be deposited. If changes are made to the data set after the initial deposit, the new version of the data set must be re-submitted.

1	**Research project number and contract number(s) relating to this project**	**Project number**			**Contract number(s)**	
		SPADAA			(27 33 324 5015) contract number with Lega is 20090008SAC.	
2	**Data set title**	Project SIZE – Caregiver (size_cg.sav)				
3	**Number of data deposit forms (e.g. 1 of 3)**	2	of	3		
4	**Person responsible for data**	Alastair van Heerden				
5	**Study information**					
5.1	**Study/collection title**	Project SIZE "We look out for our children", South Africa (2012)				
5.2	**Abstract** Give a brief description of the **project/study** in less than 300 words, including thematic content.					
A summary describing the purpose, nature, and scope of the data collection, special characteristics of its contents, major subject areas covered, and what questions the PIs attempted to answer when they conducted the study. A listing of major variables in the study is important here.						
More than two decades after the end of Apartheid, the well-being of South African children is still in a precarious state. An emerging body of research examines the role that poverty and HIV/AIDS play in household functioning, parental illness and death, children's adverse experiences and children's health, education and psychosocial development (e.g. Birdthistle, 2004, Foster & Williamson, 2000; Richter, 2004; Williamson, 2000). However, many urgent scientific and policy questions remain. These include: What are the separate and combined						

- Tools: **Data Deposit Form**
- Identification & administrative information
- Study Information
- Data set for preservation
- Data set for dissemination
- Related documentation
- Research outputs
- Promotion of the data for re-use
- Data deposit sign-off

How do we ensure data quality?



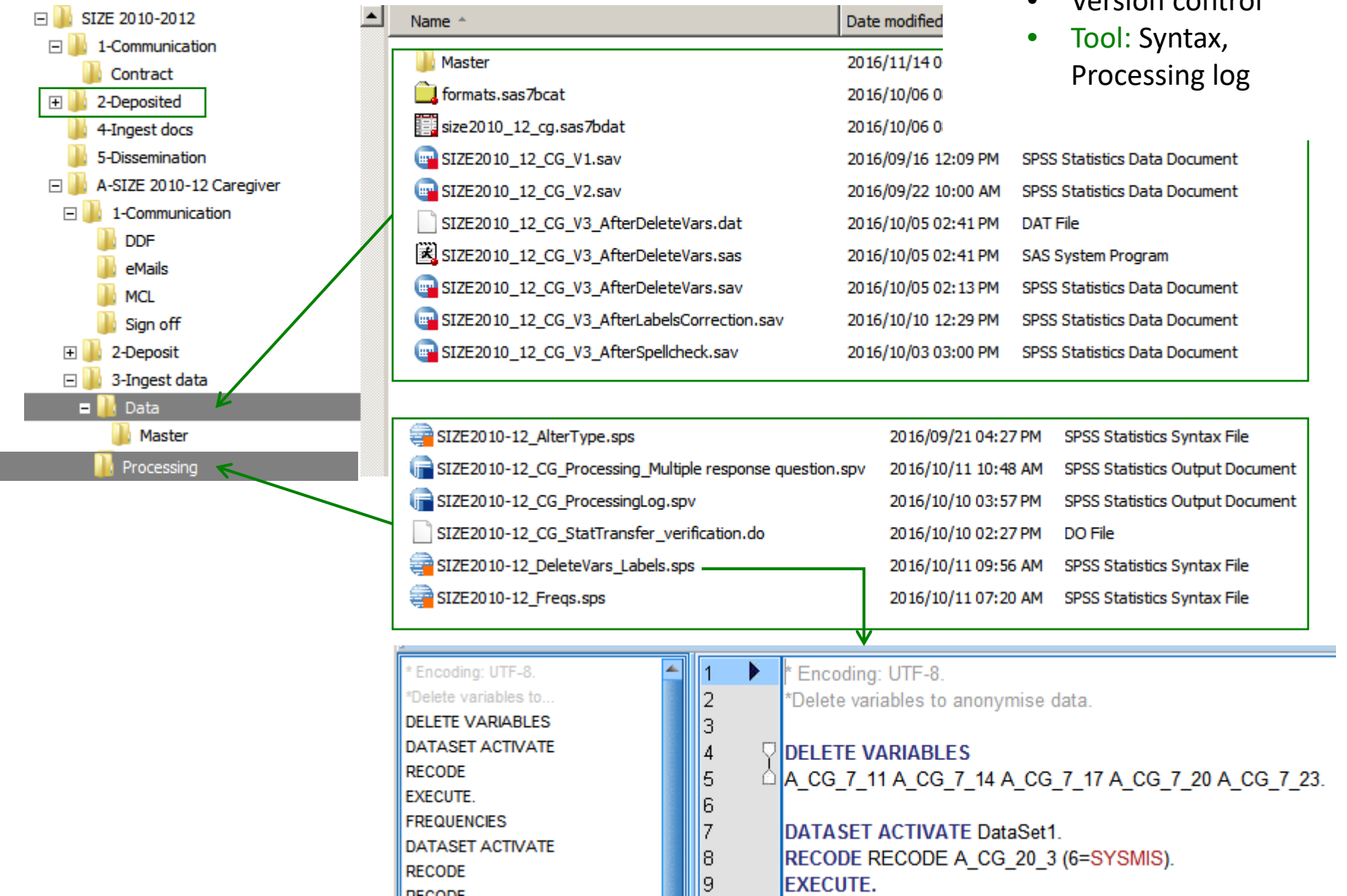
Qualitative data checklist

Item		Verified	Comment
Data set			
1	**Participant consent** form does not preclude secondary data use, e.g. that data will only be available to the project team or will be destroyed after conclusion of the project (If the data may not be shared, the data set does not comply with the ADEPTS requirement.).		
2	Data is to be **shared**, externally or at least with HSRC researchers who are not part of the original project team. If the HSRC does not own the data, written permission for the sharing of the data is provided (If the data may not be shared, the data set does not comply with the ADEPTS requirement.).		
3	The data is **anonymized**. The content of the interview (e.g. place names, person names that could identify an individual) is **anonymized. This is essential.** See example in Appendix A.		
	OR **Permission** was obtained from participants for the secondary use of the data		
	OR Data can be shared within an **enclave**, i.e. only at the HSRC, with supervision and regulated in terms of a special usage agreement.		
4	Data is **shared timeously**, i.e. before or no longer than 24 months after completion of the project.		
5	The data is the **correct and final version** of the set.		
	The data and documents are in an acceptab **format** as listed in the Data Depos		
	terviews correspon umber of e Data		

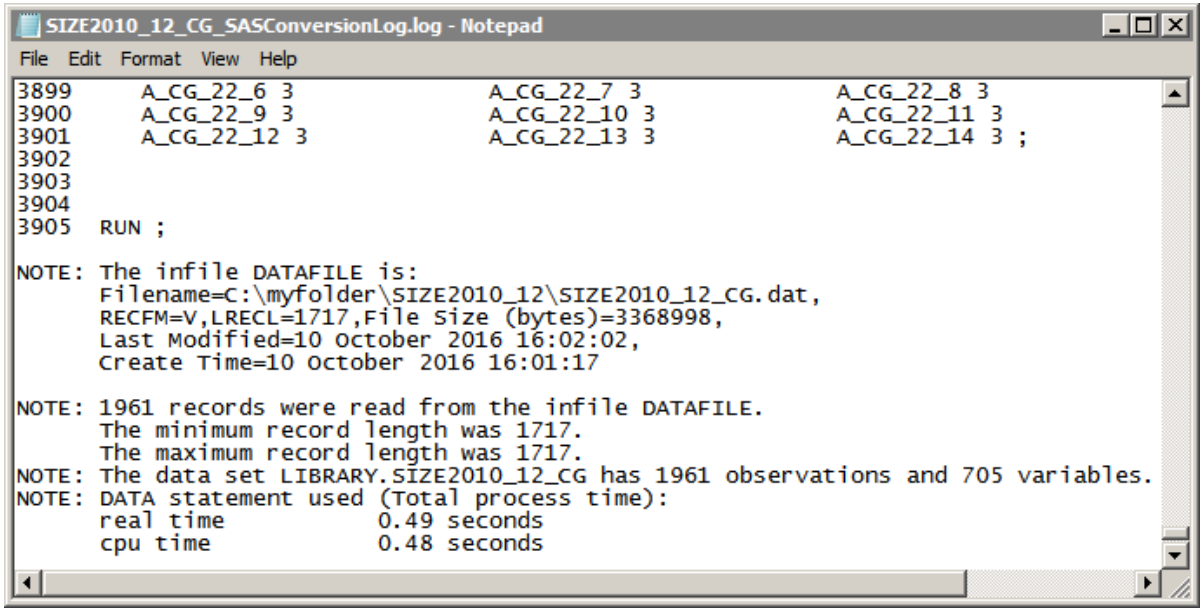
9	**All variables** have been captu data set correspond with the qu questionnaire. Additional variabl are described in terms of how the The number of variables corresponds to that specified in the Data Deposit Form.
10	Variables in the data set follow the **order** of the questions in the data collection instrument.
11	Each record has a **unique identifier / record number**. The unique identifier which links separate data files with associated information is included in all the data files.
12	No **duplicate records** are in the data file. (Duplicate records occur where multiple records were captured with exactly the same information in each of the variables.)
13	**Missing data for all variables**. Run Stata syntax to determine if data set contains records/cases without any data (missing data for all variables): Format of syntax: **egen test= rowmiss (varlist)**
14	No **duplicate record numbers** are in the data file. (Duplicate record numbers occur where all values of the variables spanning multiple records are unique, but the same record number was assigned to all these records.)
15	One and only one **attribute** is represented per variable.

- **Templates** to clarify quality requirements and guide compliance
- Tools: Data Deposit Checklist, Data Checklists, Data Appraisal Checklists

- **Processing standards**
  - Retain original deposit
  - Keep record
  - Version control
  - Tool: Syntax, Processing log

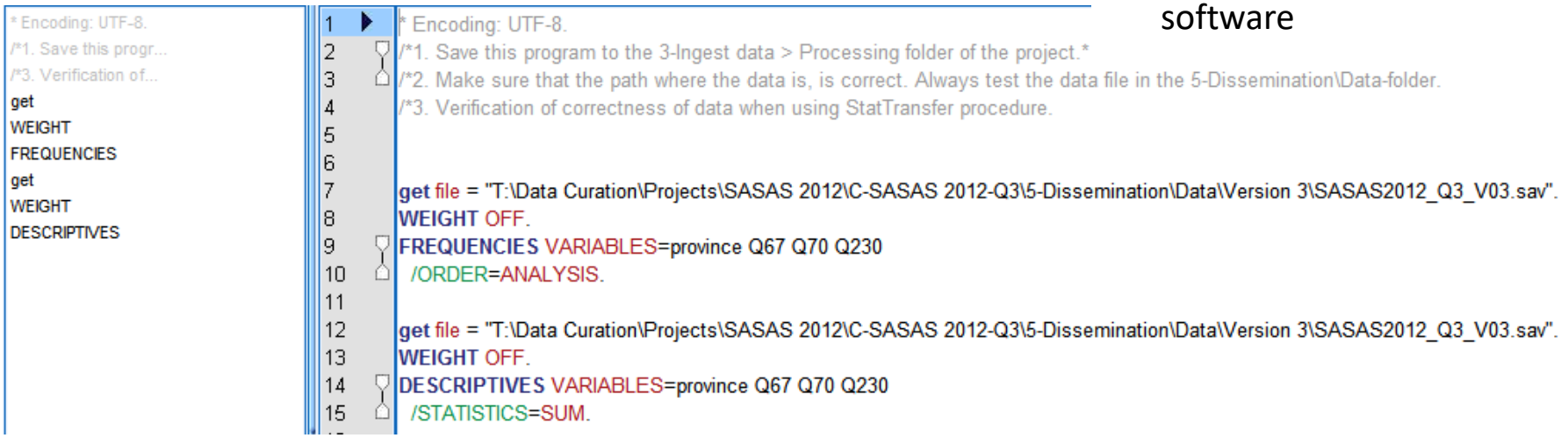


How do we create DIPs and AIPs?



Data

- Convert qualitative text data to RTF
- Convert quantitative data to SAS, STATA, SPSS, ASCII csv
- Verify accuracy, error detection
  - Program + ASCII Data File
  - Conversion logs, visual inspection, verification statistics
- Tools: Conversion, statistical software



Creating PDF documents

Guidelines for creating PDF documents

Contents

Documents

- Convert documents to PDF/A_1b
- Add metadata
- Enhance
- Verify accuracy, error detection
- Tools: Conversion software

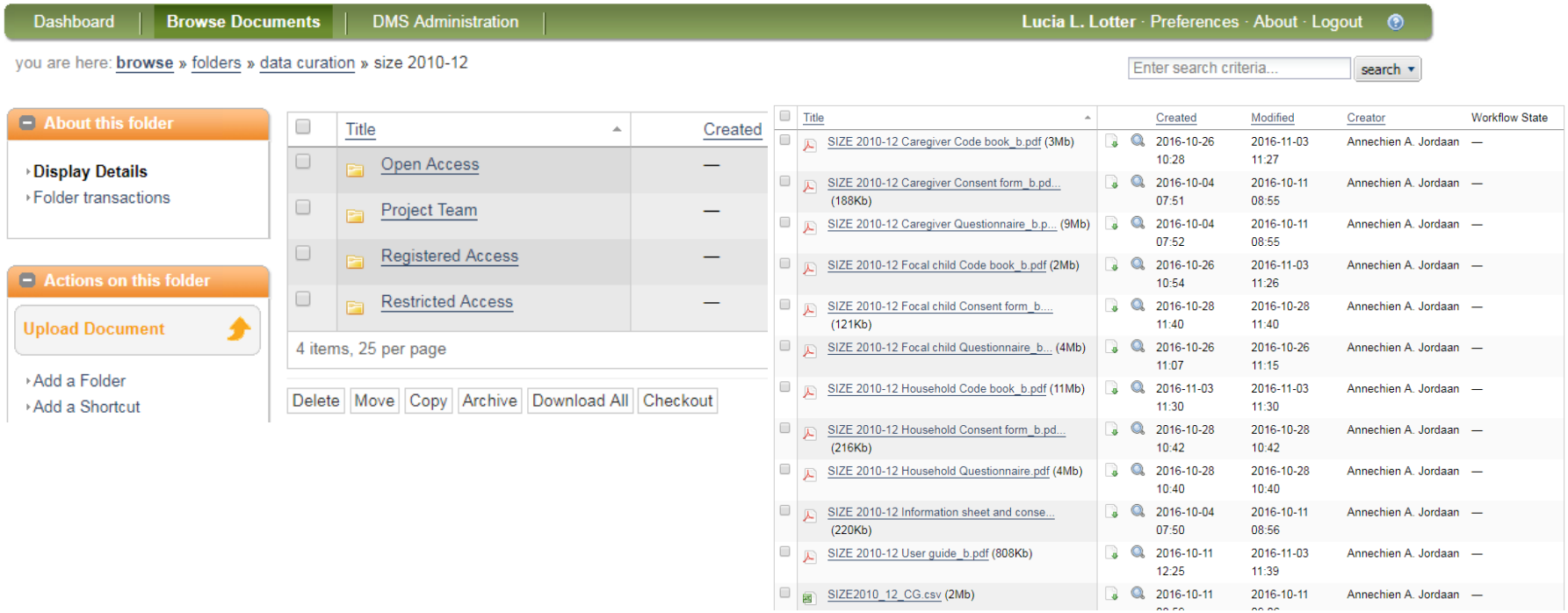
HSRC Human Sciences Research Council | eRKC eResearch Knowledge Centre

How do we store DIPs?

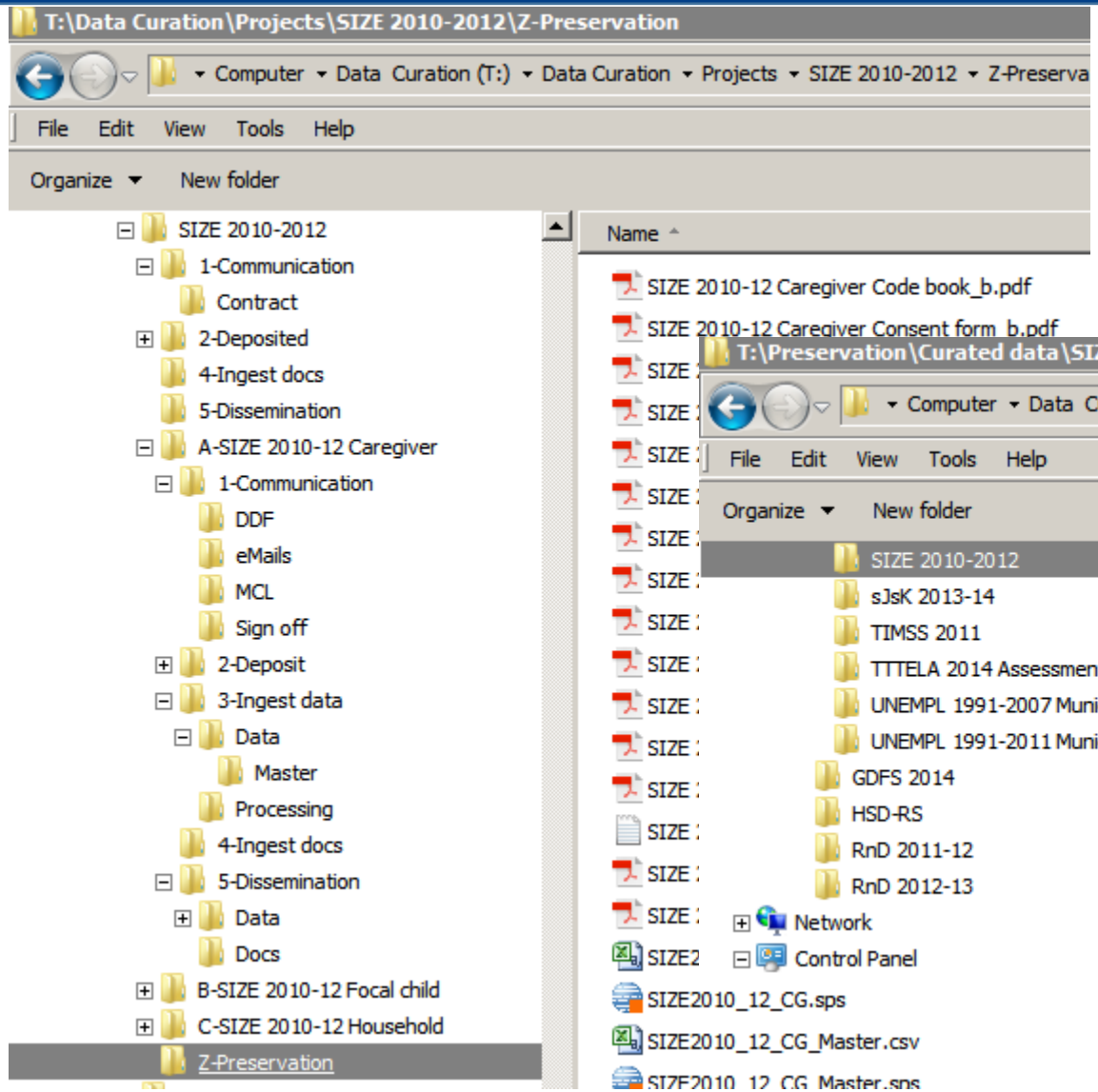
- **File repository**
  - Permissions
  - Link to metadata record
  - Tool: File repository

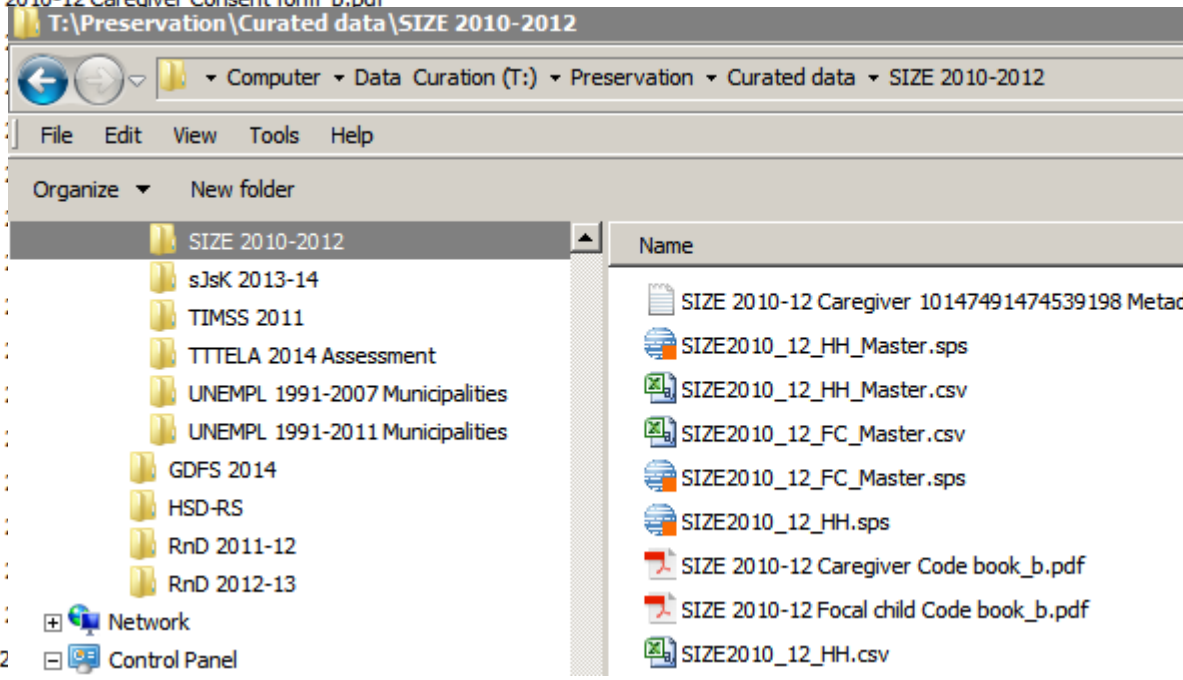
KnowledgeTree®

HSRC Human Sciences Research Council



How do we store AIPs?

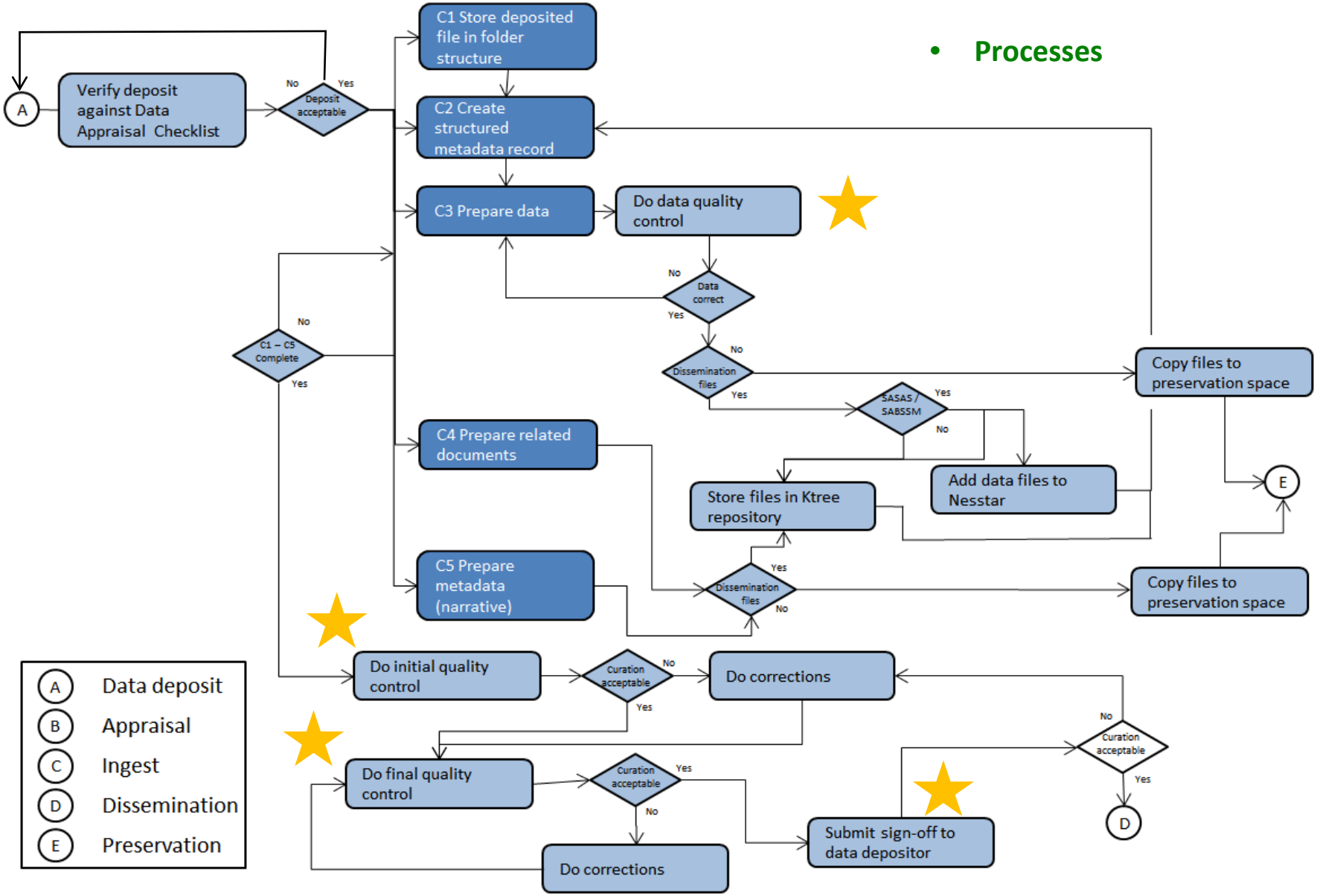




- Preservation store
  - Data, documentation, metadata
- Tools: File server

How do we do quality assurance?

- **Processes**

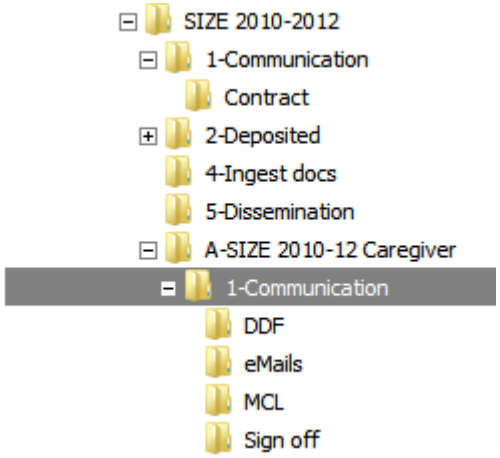


How do we do quality assurance?

- **Tools: Data Curation SOP**

How do we do quality assurance?

- **Follow structured process**
  - Tools: Monitoring Checklist



Data curation – Monitoring check list for quantitative data

Project number:	**SPADAA**	**Start date: 2016-09-19**	**Live**	Yes	No	**AJ: ADEPTS**	Yes	No
Curator:	AJ	**End date:**	If not: reason			If not: reason		
Collection name:	SIZE2010-12							
Data set ID:	SIZE2010-12 Caregiver	**Data file name of dissemination data set:** SIZE2010_12_CG						
Data file name: Master	SIZE2010_12_CG_Master							

1. Appraisal of deposited data

	Curator	QC
Appraisal should be done within a week of the date the data was deposited. Appraisal checklist completed by curator should be saved in the MCL folder.	AJ	GT

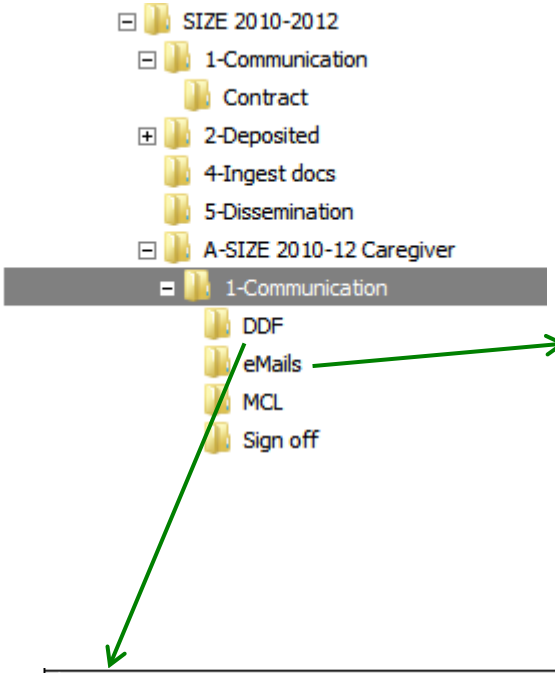
2. Sharing of data:

	Curator	QC
Verified copyright; permission to share	AJ	GT
Verified permission from respondents to reuse data – consent form Most important to do this in the data appraisal stage	AJ	

3. Communication folder:

	Curator	QC
DDF folder: • Verify that latest version of DDF was used for project deposit • Copyright / ownership of the data • Embargo date • Data deposit checklist	 AJ AJ AJ AJ	GT

How do we do quality assurance?



From: Anneke Jordaan
Sent: Monday, October 3, 2016 11:07 AM
To: Alastair Van Heerden <avanheerden@hsrc.ac.za>
Subject: RE: SIZE Data Set for Curation

- 3.9.6. What year was the grant first received? – Added 9999 as a discreet missing value. Please indicate if you agree.
  agree
- What does the response mean that appears in several of the string variables data: (e.g. A_CG_7_26) "***Missing - recoded into numeric variable***"?
  They entered a string response to other that was actually one of the formal options – it should not have been other. We could take this text out as it is now coded in the correct place. I will remove the text, no problem!
- 3.7.43. How much does it cost per week for FC's transport to school? Code 9999 – can I specify as discreet missing value? Done - please indicate if you agree. agree

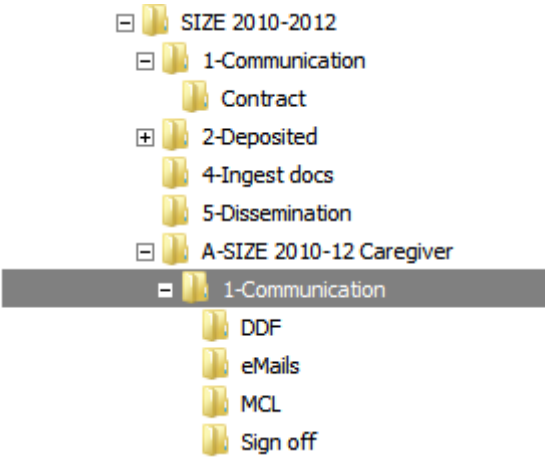
12.3 Details of the data set for dissemination If anonymisation was performed, describe the contents nature and size of the data set for dissemination.

	Number of variables and cases	Give a short description of the contents of the data files, e.g. this data set contains the responses of the adult participants for the pre and post-test of the study.	File name(s)[16]
Quantitative tabular data[17]	711 variables and 1961 cases (although only 1899 have data. The 62 caregivers without data have there PIDs captured for easier linking across files)	This dataset contains information on 1899 caregivers from the 1961 households that participated in the household survey. However, to facilitate linking, the HHIDs of households completing the household survey were added to this file with no data.	size_cg.sav

Comment [AJ3]: Six variables were deleted – 5 school names to anonymise data and A_CG_17_2 (QOL prompt). Please approve – approved (see email in T:\Data Curation\Projects\SIZE 2010-2012\A-SIZE 2010-12 Caregiver\1-Communication\eMails)

- **Record decisions and changes**
  - Tools: Data Deposit Form, Emails records

How do we do quality assurance?



- **Sign-off**
  - Tools: View dissemination on test site, Sign-off form

Sign-off document for the data curation of *We look out for our children, South Africa (SIZE) 2010-12. Msunduzi KwaZulu-Natal* - Quantitative data

Status of the data sets on 14 November 2016

The following tables provide a summary of how the data sets and related documentation will be disseminated on the HSRC web portal. Please explanations of the meta status and file status in the footnotes. The interpretation of audience is explained below:

- **Restricted access:** Users are requested to register and provide a reason for wanting to access the information. Access will be provided w the researcher / project leader is received. A record of who accesses the information, the reason for use and whether approval is granted,
- **Project team:** If a list of the project team is given to Research Use, the names will be entered into a Web portal group. The users belongi then have access to the information for viewing and downloading. If a user (who is part of the project team) was not entered as part of the system will request the user to register and provide a reason for wanting to access the information. Access will be provided when approva researcher / project leader is received.
- **Open access:** Information is made available without the need to register, provide any additional information or obtain approval. Record is that the information has been accessed.
- **Registered access:** Information can be accessed after a user has registered and provided a reason for wanting access. No approval is ne of who accesses the information and the reason for use is kept.

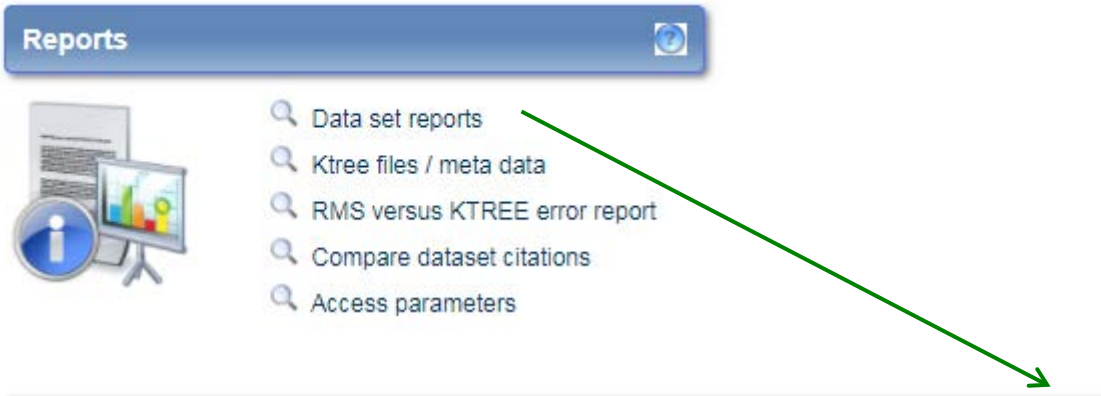
Table 1 Meta status of data

Data set ID	Project no.	Status	Date	Audience	Data file permission
SIZE 2010-12 Caregiver	SPADAA	Dissemination	2018-12-31	Open access	Registered access
SIZE 2010-12 Focal child	SPADAA	Dissemination	2018-12-31	Open access	Registered access
SIZE 2010-12 Household	SPADAA	Dissemination	2018-12-31	Open access	Registered access

Table 2a File status of data sets: Caregiver

	File status		Metadata status	
File name	Status	Audience	Status	Audie
SIZE2010_12_CG.csv	Dissemination	Registered access	Dissemination	Open
SIZE2010_12_CG.dat	Dissemination	Registered access	Dissemination	Open

How do we do quality assurance?



- **Check metadata for accuracy and completeness**
  - Tools: Metadata repository (Review and error reports)

The result!

http://datacuration.hsrc.ac.za

HSRC Human Sciences Research Council | eRKC eResearch Knowledge Centre

Thank you

Lucia Lötter
Manager Digital Curation
eResearch Knowledge Centre
Human Sciences Research Council
South Africa
llotter@hsrc.ac.za
http://datacuration.hsrc.ac.za

HSRC Human Sciences Research Council | eRKC eResearch Knowledge Centre

References

Lavoie, B. (2014). The Open Archival Information System (OAIS) Reference Model: Introductory Guide (2nd Edition). DPC Technology Watch Report. Retrieved from http://www.dpconline.org/docs/technology-watch-reports/1359-dpctw14-02/file

Consultative Committee for Space Data Systems (CCSDS). (2012). Reference Model for an Open archival Information System (OAIS). Recommended practice CCSDS 650.0-M-2 Magenta Book. DDSDS: June 2012. Retrieved from https://public.ccsds.org/pubs/650x0m2.pdf

Space data and information transfer systems -- Open archival information system (OAIS) -- Reference model ISO 14721:2012 (CCSDSS 650.0-P-1.1). Available from https://www.iso.org/standard/57284.html