

# 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

Gaborone, 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.

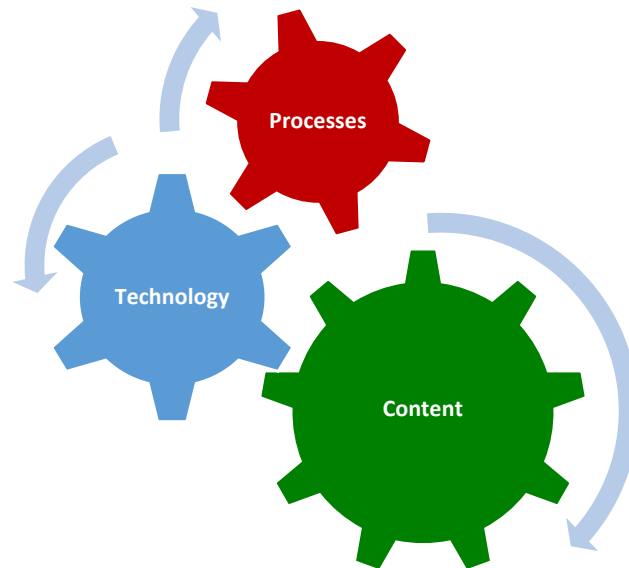
**Prerequisite: Data must be trustworthy!**



<http://datacuration.hsrc.ac.za>

# 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?

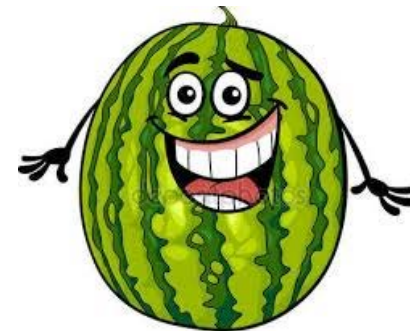
## Authenticity

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



## 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



# 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).

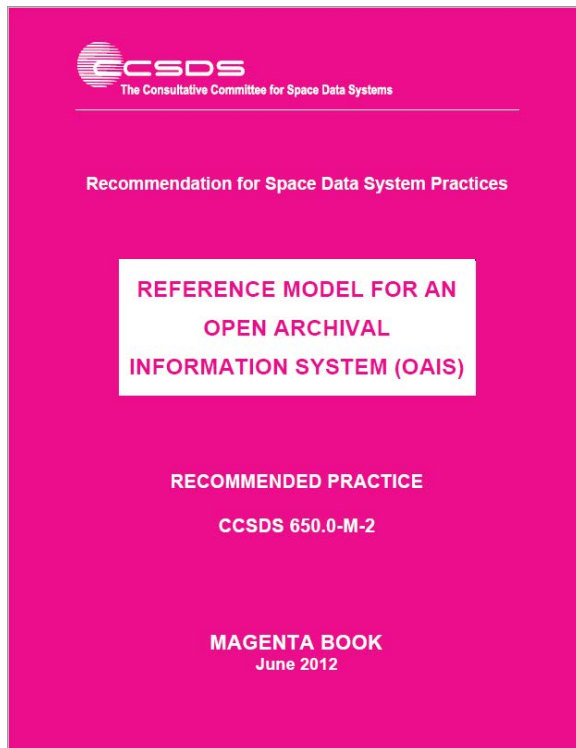
## Authenticity management

- 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



**ISO** International Organization for Standardization  
Great things happen when the world agrees

Standards | All about ISO | Taking part | **Store**

Standards catalogue | Publications and products

Home > Store > Standards catalogue > Browse by ICS > 49 > 49.140 > ISO 14721:2012

**ISO 14721:2012 (CCSDSS 650.0-P-1.1)** [Preview](#)

Space data and information transfer systems -- Open archival information system (OAIS) -- Reference model

ISO 14721:2012 defines the reference model for an open archival information system (OAIS). An OAIS is an archive, consisting of an organization, which may be part of a larger organization, of people and systems that has accepted the responsibility to preserve information and make it available for a designated community. It meets a set of such responsibilities as defined in this International Standard, and this allows an OAIS archive to be distinguished from other uses of the term "archive". The term "open" in OAIS is used to imply that ISO 14721:2012, as well as future related International Standards, are developed in open forums, and it does not imply that access to the archive is unrestricted.

ISO 14721:2012

Buy this standard

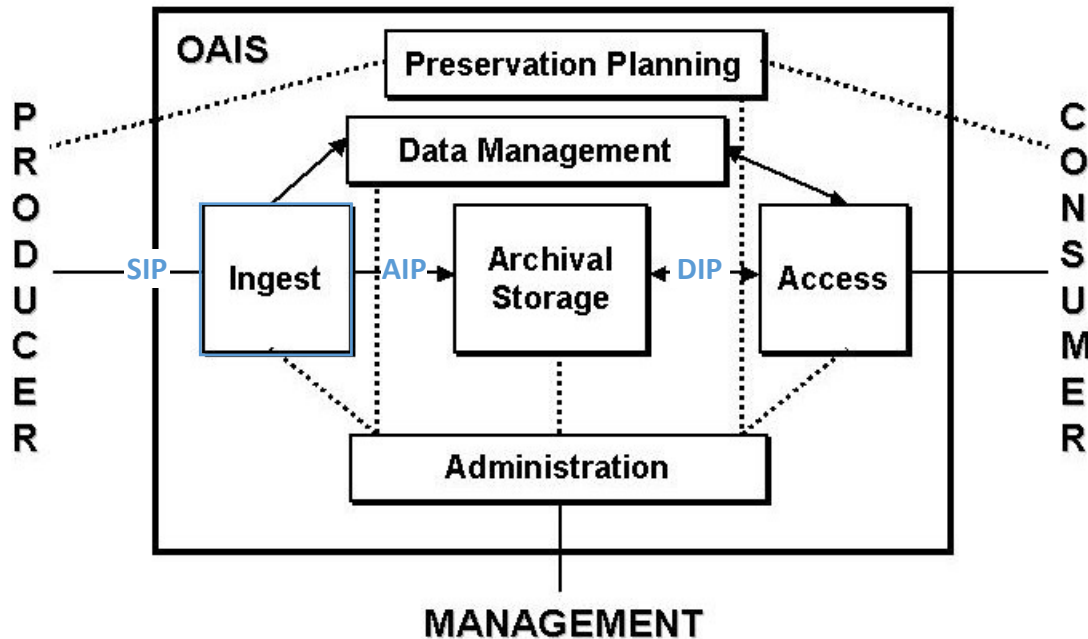
Format	Language
<input checked="" type="checkbox"/> PDF	English
<input type="checkbox"/> Paper	English

CHF 198 [Buy](#)

# 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)

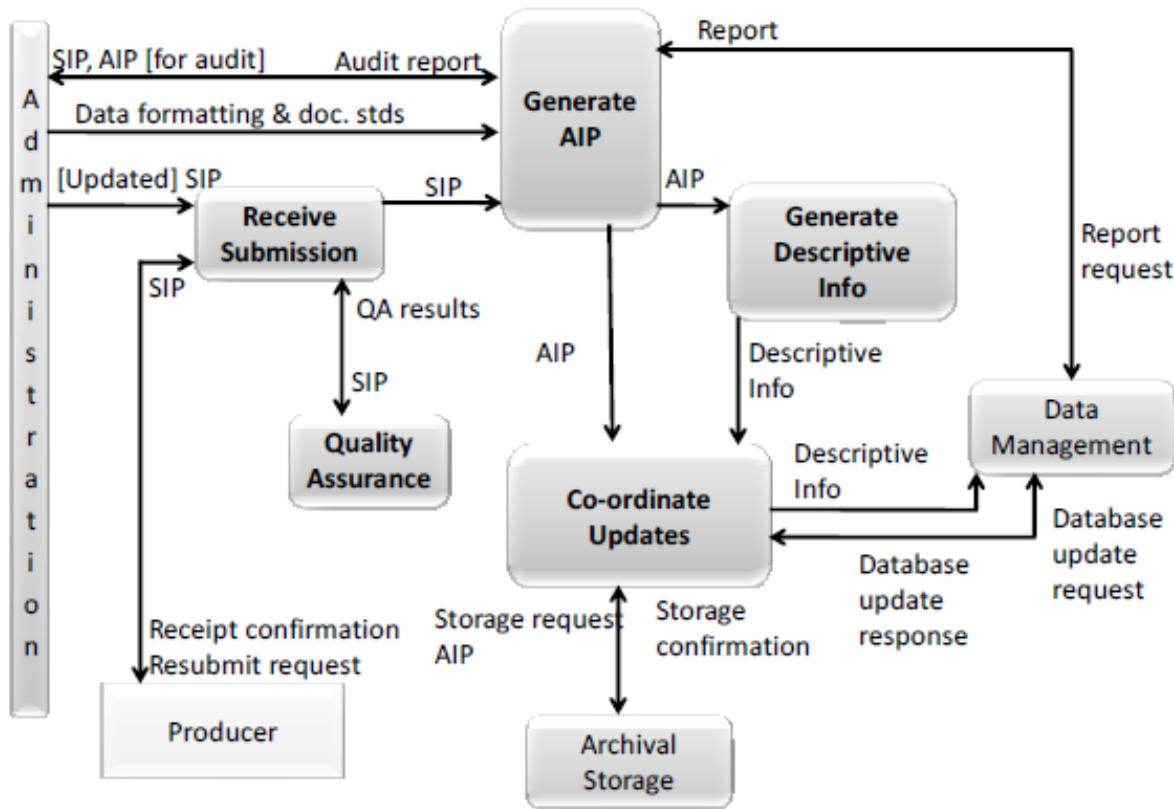


(Lavoie, 2014)

- SIP** Submission Information Package
- AIP** Archival Information Package
- DIP** Dissemination Information Package



# The “Ingest” function



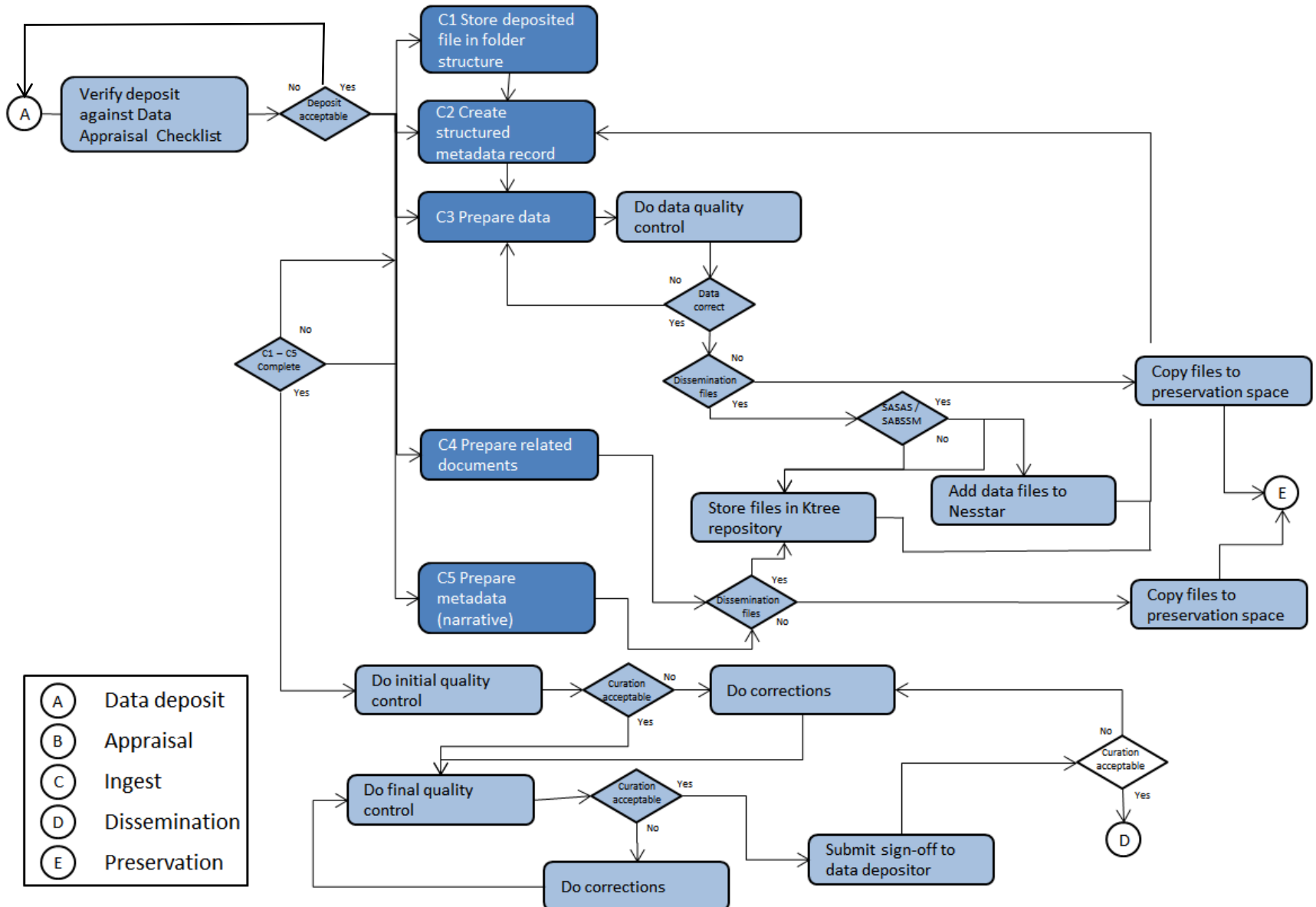
- Receipt of SIP
- Validation that the SIP is uncorrupted and complete
- Transformation of SIP into a form suitable for storage and management within the archival system
- Extraction and/or creation of descriptive metadata
- Transfer AIP and DIP and its associated metadata to the archival store

(CCSDS, 2012)

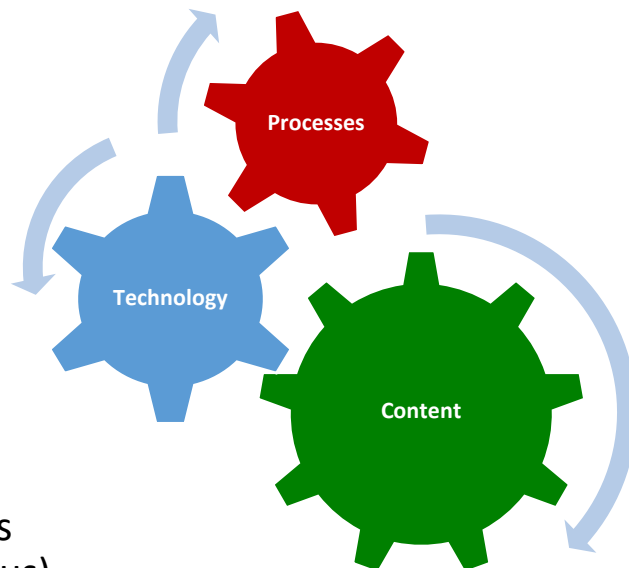
(Lavoie, 2014)



# 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?

SIZE 2010-2012 ← Data collection

1-Communication

Contract

2-Deposited

4-Ingest docs

5-Dissemination

A-SIZE 2010-12 Caregiver ← Data set

1-Communication

DDF

eMails

MCL

Sign off

2-Deposit

3-Ingest data

Data

Processing

4-Ingest docs

5-Dissemination

Data

Docs

B-SIZE 2010-12 Focal child

C-SIZE 2010-12 Household

Z-Preservation

SIZE2010_12_CG.csv	2016/10/10 04:02 PM	Microsoft Excel Com...	2 453 KB
SIZE2010_12_CG.dat	2016/10/10 04:02 PM	DAT File	3 291 KB
SIZE2010_12_CG.dct	2016/10/10 04:01 PM	DCT File	112 KB
SIZE2010_12_CG.do	2016/10/10 04:08 PM	DO File	79 KB
SIZE2010_12_CG.dta	2016/10/10 04:08 PM	DTA File	3 325 KB
SIZE2010_12_CG.sas	2016/10/10 04:04 PM	SAS System Program	187 KB
SIZE2010_12_CG.sas7bcats	2016/10/10 04:04 PM	SAS Catalog	553 KB
SIZE2010_12_CG.sas7bdat	2016/10/10 04:04 PM	SAS Data Set	6 032 KB
SIZE2010_12_CG.sav	2016/10/10 04:00 PM	SPSS Statistics Data...	2 142 KB
SIZE2010_12_CG.sps	2016/10/10 04:02 PM	SPSS Statistics Synt...	146 KB
SIZE2010_12_CG.zip	2016/10/10 04:11 PM	ZIPITF~1 file	2 020 KB
SIZE 2010-12 Caregiver Code book_b.pdf			
SIZE 2010-12 Caregiver Consent form_b.pdf			
SIZE 2010-12 Caregiver Questionnaire_b.pdf			

# How do we compile metadata?

Time recording | Reports | Data sets | Contacts | Management

( PFAJLA ) Change project  
( SABSSM 2005 Adult-youth ) Change data set  
Jump to project level

**Data sets interface - Edit data set**

Control panel | Data set details | Subject description | Scope | Data collection | Funding / Authoring | Access and copyright | Related documents | Data files

**Data set details**

Data set ID:  
SABSSM 2005 Adult-youth  
23/255

Data set title:  
South African National HIV Prevalence, HIV Incidence, Behaviour and Communication Survey (SABSSM) 2005: Adult and youth data - All provinces

Citation:  
Human Sciences Research Council. *South African National HIV Prevalence, HIV Incidence, Behaviour and Communication Survey (SABSSM) 2005: Adult and youth data - All provinces.* [Data set]. SABSSM 2005 Adult-youth. Version 1.0. Pretoria South Africa: Human Sciences Research Council[producer] 2005, [distributor] 2011.

Data set description:

This data set contains information on participants (youth and adult) aged 15 year perceptions of HIV/AIDS; marital status; sexual debut; condoms; vulnerable group perceptions about government policies; knowledge and perceptions of HIV vaccines. The data set contains 495 variables and 16398 cases.

- **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 <ul style="list-style-type: none"> <li>• Composition: Acronym of collection (upper case), Space, Year of data collection, Space, Reference to data set content.</li> <li>• Descriptives capitalised, rest lower case.</li> <li>• Linked descriptives and multiple years indicated with unspaced hyphen – not 2006/07, 2006-2007.</li> </ul>	<IDNo agency=" HSRC ">xxx</IDNo>  Examples <ul style="list-style-type: none"> <li>• SASAS 2008 Q1</li> <li>• SABSSM 2005 Adult-youth</li> <li>• EC-PSS 2008</li> <li>• SASAS 2003 Combined Household weight</li> <li>• SASAS 2004 Combined Weight</li> <li>• RnD 2006-07</li> <li>• NALA 2009 Gr9 Natural sciences assessment</li> </ul>

# 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 fields 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 ... x

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

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

1. Introduction.....	1
2. Available data sets .....	1
2.1. Program code to generate the different statistical data sets mentioned above: ....	1
3. Conversion of data.....	2
3.1. Additional notes on SPSS.....	2
3.2. Additional notes on SAS.....	2
3.3. Other formats: .....	3
4. Data and documentation notes .....	4
4.1. Deleted / recoded / derived variables.....	4
4.2. Missing values within the data sets .....	4
4.3. Variables and labels within the data set .....	4
4.4. Weighting of the data.....	4
APPENDIX A.....	5
APPENDIX B.....	8

# How do we source metadata?

## Data Deposit Form

Please provide the following information about the data<sup>1</sup> 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<sup>2</sup> (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 Legal is 20090008SAC.		
2	Data set title	Project SIZE – Caregiver (size_cg.say)			
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?

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

## Qualitative data checklist

Data set	Item	Verified	Comment
	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 <u>does not</u> comply with the ADEPTS requirement).		
	2 Data is to be <b>shared</b> , 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 <u>does not</u> comply with the ADEPTS requirement).		
	3 The data is <b>anonymized</b> . The content of the interview (e.g. place names, person names that could identify an individual) is <b>anonymized</b> . <b>This is essential</b> . See example in Appendix A. OR <b>Permission</b> was obtained from participants for the secondary use of the data OR Data can be shared within an <b>enclave</b> , i.e. only at the HSRC, with supervision and regulated in terms of a special usage agreement.		
	4 Data is <b>shared timeously</b> , i.e. before or no longer than 24 months after completion of the project.		
	5 The data is the <b>correct and final version</b> of the set. The data and documents are in an acceptable format as listed in the Data Deposit document.		
	9 <b>All variables</b> have been captured in the data set correspond with the questionnaire. Additional variables are described in terms of how they are captured. The number of variables corresponds to that specified in the Data Deposit Form.		
	10 Variables in the data set follow the <b>order</b> of the questions in the data collection instrument.		
	11 Each record has a <b>unique identifier / record number</b> . The unique identifier which links separate data files with associated information is included in all the data files.		
	12 <b>No duplicate records</b> are in the data file. (Duplicate records occur where multiple records were captured with exactly the same information in each of the variables.)		
	13 <b>Missing data for all variables</b> . Run Stata syntax to determine if data set contains records/cases without any data (missing data for all variables): Format of syntax: <code>egen test= rowmiss (varlist)</code>		
	14 <b>No duplicate record numbers</b> 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 <b>attribute</b> is represented per variable.		



- **Processing standards**

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

- SIZE 2010-2012
  - 1-Communication
    - Contract
    - 2-Deposited**
    - 4-Ingest docs
    - 5-Dissemination
  - A-SIZE 2010-12 Caregiver
    - 1-Communication
      - DDF
      - eMails
      - MCL
      - Sign off
    - 2-Deposit
    - 3-Ingest data
      - Data**
      - Master
      - Processing**

Name	Date modified	
Master	2016/11/14 0	
formats.sas7bcat	2016/10/06 0	
size2010_12_cg.sas7bdat	2016/10/06 0	
SIZE2010_12_CG_V1.sav	2016/09/16 12:09 PM	SPSS Statistics Data Document
SIZE2010_12_CG_V2.sav	2016/09/22 10:00 AM	SPSS Statistics Data Document
SIZE2010_12_CG_V3_AfterDeleteVars.dat	2016/10/05 02:41 PM	DAT File
SIZE2010_12_CG_V3_AfterDeleteVars.sas	2016/10/05 02:41 PM	SAS System Program
SIZE2010_12_CG_V3_AfterDeleteVars.sav	2016/10/05 02:13 PM	SPSS Statistics Data Document
SIZE2010_12_CG_V3_AfterLabelsCorrection.sav	2016/10/10 12:29 PM	SPSS Statistics Data Document
SIZE2010_12_CG_V3_AfterSpellcheck.sav	2016/10/03 03:00 PM	SPSS Statistics Data Document

SIZE2010-12_AlterType.sps	2016/09/21 04:27 PM	SPSS Statistics Syntax File
SIZE2010-12_CG_Processing_Multiple response question.spv	2016/10/11 10:48 AM	SPSS Statistics Output Document
SIZE2010-12_CG_ProcessingLog.spv	2016/10/10 03:57 PM	SPSS Statistics Output Document
SIZE2010-12_CG_StatTransfer_verification.do	2016/10/10 02:27 PM	DO File
SIZE2010-12_DeleteVars_Labels.sps	2016/10/11 09:56 AM	SPSS Statistics Syntax File
SIZE2010-12_Freqs.sps	2016/10/11 07:20 AM	SPSS Statistics Syntax File

```
* Encoding: UTF-8.
*Delete variables to...
DELETE VARIABLES
DATASET ACTIVATE
RECODE
EXECUTE.
FREQUENCIES
DATASET ACTIVATE
RECODE
RECODE
EXECUTE.
```

```
1 * Encoding: UTF-8.
2 *Delete variables to anonymise data.
3
4 DELETE VARIABLES
5 A_CG_7_11 A_CG_7_14 A_CG_7_17 A_CG_7_20 A_CG_7_23.
6
7 DATASET ACTIVATE DataSet1.
8 RECODE RECODE A_CG_20_3 (6=SYSMIS).
9 EXECUTE.
```

# 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

```
SIZE2010_12_CG_SASConversionLog.log - Notepad
File Edit Format View Help
3899     A_CG_22_6 3           A_CG_22_7 3           A_CG_22_8 3
3900     A_CG_22_9 3           A_CG_22_10 3          A_CG_22_11 3
3901     A_CG_22_12 3          A_CG_22_13 3          A_CG_22_14 3 ;
3902
3903
3904
3905  RUN ;

NOTE: The infile DATAFILE is:
      Filename=C:\myfolder\SIZE2010_12\SIZE2010_12_CG.dat,
      RECFM=V,LRECL=1717,File size (bytes)=3368998,
      Last Modified=10 October 2016 16:02:02,
      Create Time=10 October 2016 16:01:17

NOTE: 1961 records were read from the infile DATAFILE.
      The minimum record length was 1717.
      The maximum record length was 1717.

NOTE: The data set LIBRARY.SIZE2010_12_CG has 1961 observations and 705 variables.
NOTE: DATA statement used (Total process time):
      real time           0.49 seconds
      cpu time            0.48 seconds
```

```
* Encoding: UTF-8.
/*1. Save this progr...
/*3. Verification of...
get
WEIGHT
FREQUENCIES
get
WEIGHT
DESCRIPTIVES

1 * Encoding: UTF-8.
2 /*1. Save this program to the 3-Ingest data > Processing folder of the project.*
3 /*2. Make sure that the path where the data is, is correct. Always test the data file in the 5-Dissemination\Data-folder.
4 /*3. Verification of correctness of data when using StatTransfer procedure.
5
6
7 get file = "T:\Data Curation\Projects\SASAS 2012\C-SASAS 2012-Q3\5-Dissemination\Data\Version 3\SASAS2012_Q3_V03.sav".
8 WEIGHT OFF.
9 FREQUENCIES VARIABLES=province Q67 Q70 Q230
10 /ORDER=ANALYSIS.
11
12 get file = "T:\Data Curation\Projects\SASAS 2012\C-SASAS 2012-Q3\5-Dissemination\Data\Version 3\SASAS2012_Q3_V03.sav".
13 WEIGHT OFF.
14 DESCRIPTIVES VARIABLES=province Q67 Q70 Q230
15 /STATISTICS=SUM.
```

# Creating PDF documents

## Guidelines for creating PDF documents

### Contents

1	The PDF/A standard.....	
2	HSRC standard for long term preservation.....	
3	Creating a PDF/A-1b file.....	
3.1	Two ways to produce a PDF/A file.....	
3.2	To convert an existing PDF file to PDF/A-1b format.....	
3.2.1	Setting the properties in Adobe Acrobat.....	
3.2.2	Converting and testing the file for PDF/A1-b compliance.....	
3.3	Generating a PDF/A-1b compliant file from MsWord.....	3
3.3.1	Setting the default.....	3
3.3.2	Setting the properties in Ms Word.....	5
3.3.3	Creating headers for tagging in the PDF file.....	5
3.3.4	Converting the file to PDF.....	5
3.3.5	Setting the properties in Adobe Acrobat if the file is already in PDF/A-1b format.....	6
3.4	Generating a PDF/A-1b compliant file from MSEXcel.....	6
3.5	Setting the default of the PDF printer to PDF/A-1b compliance.....	7
3.6	Creating PDF documents from other word-processing software.....	7
3.7	Hard copy documentation.....	7
4	General guidelines.....	7
4.1	Original document formatting.....	7
4.2	Document structure.....	7
4.3	Enter metadata into the Document properties.....	7
4.4	Add bookmarks to questionnaires and user guides.....	8

## Documents

- Convert documents to PDF/A\_1b
- Add metadata
- Enhance
- Verify accuracy, error detection
- **Tools:** Conversion software



# How do we store DIPs?

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

KnowledgeTree®



Dashboard | Browse Documents | DMS Administration | Lucia L. Lotter · Preferences · About · Logout

you are here: [browse](#) » [folders](#) » [data curation](#) » size 2010-12

Enter search criteria... search

### About this folder

- Display Details
- Folder transactions

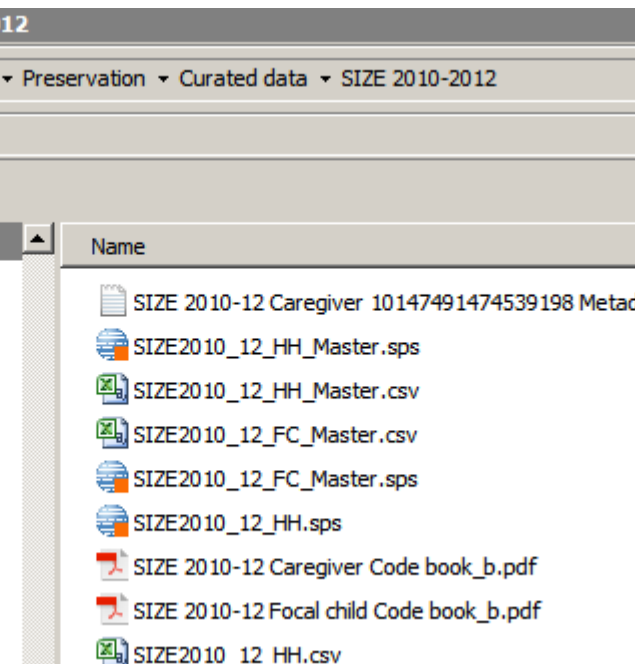
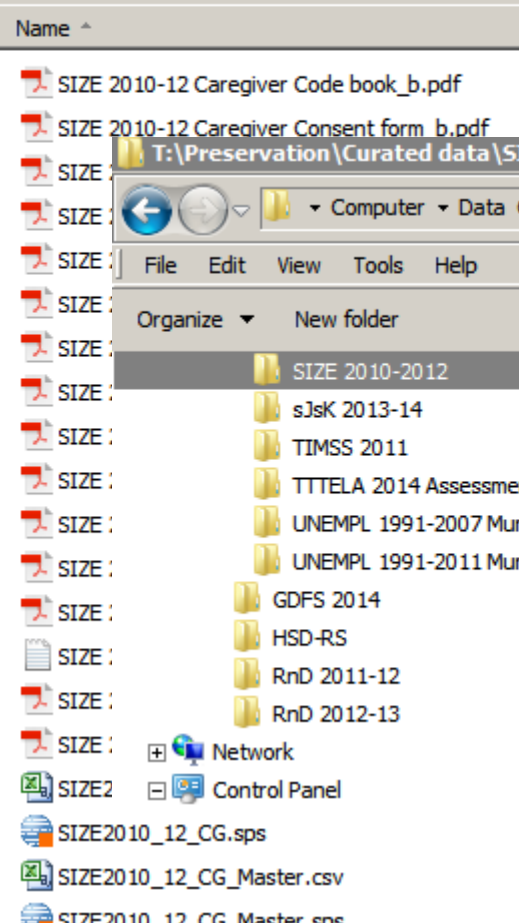
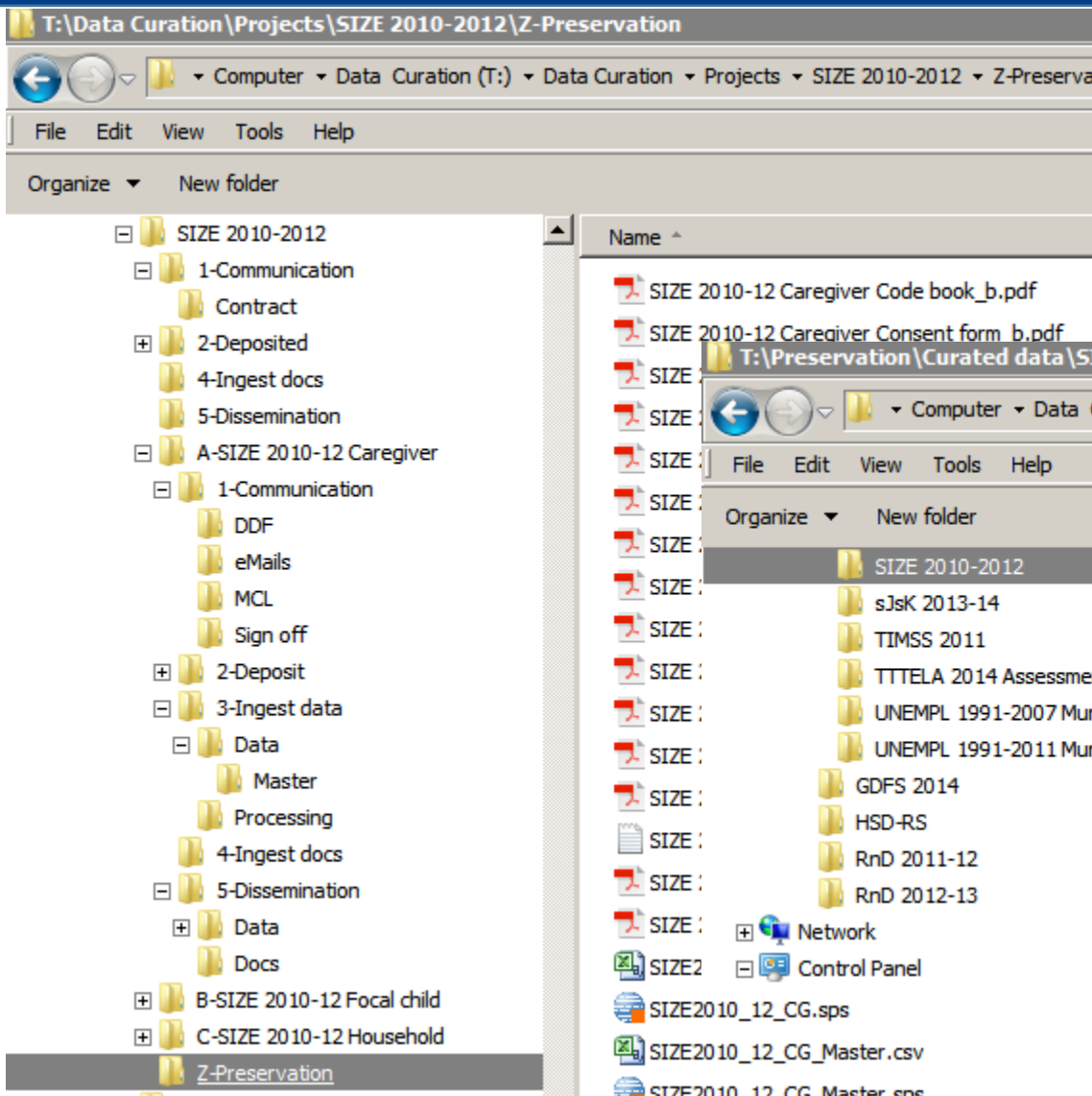
### Actions on this folder

- Upload Document
- Add a Folder
- Add a Shortcut

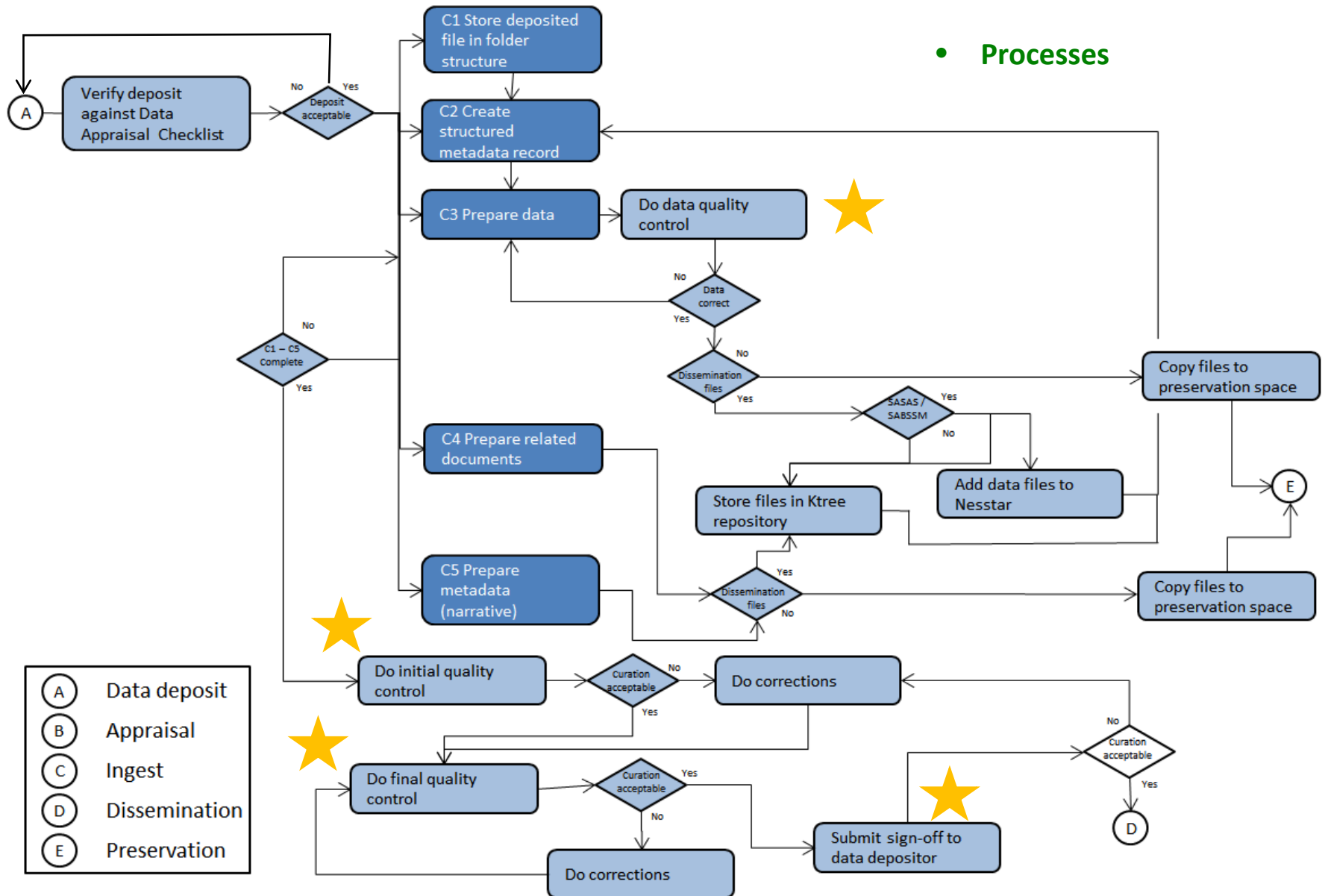
Title	Created	Modified	Creator	Workflow State
<a href="#">SIZE 2010-12 Caregiver Code book_b.pdf (3Mb)</a>	2016-10-26 10:28	2016-11-03 11:27	Annechien A. Jordaan	—
<a href="#">SIZE 2010-12 Caregiver Consent form_b.pd... (188Kb)</a>	2016-10-04 07:51	2016-10-11 08:55	Annechien A. Jordaan	—
<a href="#">SIZE 2010-12 Caregiver Questionnaire_b.p... (9Mb)</a>	2016-10-04 07:52	2016-10-11 08:55	Annechien A. Jordaan	—
<a href="#">SIZE 2010-12 Focal child Code book_b.pdf (2Mb)</a>	2016-10-26 10:54	2016-11-03 11:26	Annechien A. Jordaan	—
<a href="#">SIZE 2010-12 Focal child Consent form_b... (121Kb)</a>	2016-10-28 11:40	2016-10-28 11:40	Annechien A. Jordaan	—
<a href="#">SIZE 2010-12 Focal child Questionnaire_b... (4Mb)</a>	2016-10-26 11:07	2016-10-26 11:15	Annechien A. Jordaan	—
<a href="#">SIZE 2010-12 Household Code book_b.pdf (11Mb)</a>	2016-11-03 11:30	2016-11-03 11:30	Annechien A. Jordaan	—
<a href="#">SIZE 2010-12 Household Consent form_b.pd... (216Kb)</a>	2016-10-28 10:42	2016-10-28 10:42	Annechien A. Jordaan	—
<a href="#">SIZE 2010-12 Household Questionnaire.pdf (4Mb)</a>	2016-10-28 10:40	2016-10-28 10:40	Annechien A. Jordaan	—
<a href="#">SIZE 2010-12 Information sheet and conse... (220Kb)</a>	2016-10-04 07:50	2016-10-11 08:56	Annechien A. Jordaan	—
<a href="#">SIZE 2010-12 User guide_b.pdf (808Kb)</a>	2016-10-11 12:25	2016-11-03 11:39	Annechien A. Jordaan	—
<a href="#">SIZE2010_12_CG.csv (2Mb)</a>	2016-10-11 00:00	2016-10-11 00:00	Annechien A. Jordaan	—

# How do we store AIPs?

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



# How do we do quality assurance?



# How do we do quality assurance?

<b>1. Introduction .....</b>	<b>• Tools: Data Curation SOP</b>	
1.1 Using the Open Archival Information System (OAIS) as basis for data curation		1
1.2 Data management process flow		2
<b>2. Data deposit .....</b>		<b>2</b>
<b>3. Data appraisal.....</b>		<b>2</b>
<b>4. Ingest.....</b>		<b>3</b>
4.1 Storage of files		3
4.2 Implement the file naming convention		5
4.3 Setting access parameters		6
4.4 Creating the metadata record of a data set		6
4.5 Document programs, recordings and all other actions performed		12
4.6 Check and improve the quality of data files		13
4.7 Convert documents to PDF/A_1b		32
4.8 Create a User guide		32
4.9 Create a readme file		32
4.10 Storing dissemination files		33
4.11 Configuring team member access groups		35
4.12 Quality control		38
4.13 Sign-off		39
4.14 Registering the Digital Object Identifier (DOI) with da ra		40
4.15 Recording performance indicator information		41
4.16 Dissemination new versions of data files		42
4.17 Retracting a data set		42
<b>5. Preservation .....</b>		<b>43</b>
5.1 Introduction		43
5.2 Content to be preserved		43
5.3 Preservation strategy		44
<b>6. Disseminating data.....</b>		<b>45</b>
<b>7. Management.....</b>		<b>48</b>
<b>8. Appendix A: Reference documents .....</b>		<b>48</b>

# How do we do quality assurance?

- Follow structured process
  - Tools: Monitoring Checklist

- [-] SIZE 2010-2012
  - [-] 1-Communication
    - Contract
  - [+] 2-Deposited
    - 4-Ingest docs
    - 5-Dissemination
  - [-] A-SIZE 2010-12 Caregiver
    - [+] 1-Communication
      - DDF
      - eMails
      - MCL
      - Sign off

## Data curation – Monitoring check list for quantitative data

<b>Project number:</b>	SPADAA	<b>Start date:</b> 2016-09-19	<b>Live</b>	Yes	No	<b>AJ:</b> <b>ADEPTS</b>	Yes	No
<b>Curator:</b>	AJ	<b>End date:</b>	If not: reason		If not: reason			
<b>Collection name:</b>	SIZE2010-12							
<b>Data set ID:</b>	SIZE2010-12 Caregiver	<b>Data file name of dissemination data set:</b> SIZE2010_12_CG						
<b>Data file name:</b> <b>Master</b>	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
<b>DDF folder:</b> <ul style="list-style-type: none"> <li>• Verify that latest version of DDF was used for project deposit</li> <li>• Copyright / ownership of the data</li> <li>• Embargo date</li> <li>• Data deposit checklist</li> </ul>	AJ AJ AJ AJ	GT





# How do we do quality assurance?

- [-] SIZE 2010-2012
  - [-] 1-Communication
    - [-] Contract
    - [+] 2-Deposited
      - [-] 4-Ingest docs
      - [-] 5-Dissemination
    - [-] A-SIZE 2010-12 Caregiver
      - [-] 1-Communication
        - [-] DDF
        - [-] eMails
        - [-] MCL
        - [-] Sign off

**From:** Anneke Jordaan  
**Sent:** Monday, October 3, 2016 11:07 AM  
**To:** Alastair Van Heerden <[avanheerden@hsr.ac.za](mailto:avanheerden@hsr.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. **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**

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

## 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) <sup>16</sup>
Quantitative tabular data <sup>17</sup>	711 variables and 1961 cases (although only 1899 have data. The 62 caregivers without data have their 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_say

**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)

# How do we do quality assurance?

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

- [-] SIZE 2010-2012
  - [-] 1-Communication
    - Contract
  - [+] 2-Deposited
    - 4-Ingest docs
    - 5-Dissemination
  - [-] A-SIZE 2010-12 Caregiver
    - [+] 1-Communication
      - DDF
      - eMails
      - MCL
      - Sign off

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 refer to the footnotes for more detailed 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 when approval from the researcher / project leader is received. A record of who accesses the information, the reason for use and whether approval is granted, will be kept.
- **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 belonging to this group 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 project team, the system will request the user to register and provide a reason for wanting to access the information. Access will be provided when approval from the 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 kept of who accesses the information and the reason for use is kept.
- **Registered access:** Information can be accessed after a user has registered and provided a reason for wanting access. No approval is needed from the researcher / project leader. Record is kept 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 name	File status			Metadata status	
	Status	Audience	Status	Audience	
SIZE2010_12_CG.csv	Dissemination	Registered access	Dissemination	Open access	
SIZE2010_12_CG.dat	Dissemination	Registered access	Dissemination	Open access	

# How do we do quality assurance?

## Reports



- Data set reports
- Ktree files / meta data
- RMS versus KTREE error report
- Compare dataset citations
- Access parameters

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

## Data sets interface - Data set report

### Data set report

Please select a project.

Projects:

SPADAA - We look out for our children, South Africa (SIZE) 2010-12. Msunduzi Municipality - KwaZulu-Natal

List of all data sets assigned access to you.

### List of data sets

Data set no.	Data set ID	Data set Title	Reports
1	SIZE 2010-12 Caregiver	We look out for our children, Caregiver (SIZE) 2010-12. Msunduzi Municipality - KwaZulu-Natal	-- select report -- -- select report -- Summary Related documents Data files Metadata review
3	SIZE 2010-12 Focal child	We look out for our children, Focal child (SIZE) 2010-12. Msunduzi Municipality - KwaZulu-Natal	
2	SIZE 2010-12 Household	We look out for our children, Household (SIZE) 2010-12. Msunduzi Municipality - KwaZulu-Natal	

# The result!

Logout of Research Data Service | CONTACT US



**HSRC**  
Human Sciences  
Research Council

Research Data Service

Welcome, you are currently logged in with email address [llover@hsrc.ac.za](mailto:llover@hsrc.ac.za)

HOME ABOUT US ACCESS TO DATA AVAILABLE DATA

### Data set

#### We look out for our children, Caregiver (SIZE) 2010-12. Msunduzi Municipality - KwaZulu-Natal

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

**Data set metadata record**  
Download data set metadata record

**Data set ID** SIZE 2010-12 Caregiver

**Data set title** We look out for our children, Caregiver (SIZE) 2010-12. Msunduzi Municipality - KwaZulu-Natal

**Citation** Human Sciences Research Council. *We look out for our children, Caregiver (SIZE) 2010-12. Msunduzi Municipality - KwaZulu-Natal*. [Data set]. SIZE 2010-12 Caregiver. Version 1.0. Pretoria South Africa: Human Sciences Research Council [producer] 2016, Human Sciences Research Council [distributor] 2018.

**Data set description** 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.



<http://datacuration.hsrc.ac.za>

# Thank you

**Lucia Lötter**

Manager Digital Curation  
eResearch Knowledge Centre  
Human Sciences Research Council  
South Africa

[lloetter@hsrc.ac.za](mailto:lloetter@hsrc.ac.za)

<http://datacuration.hsrc.ac.za>



# 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>

