Digital Curation and Preservation at DataFirst

1. Introduction

DataFirst is a research data service dedicated to giving open access to African data for quantitative social research. For this purpose, DataFirst must undertake long-term preservation of data and ensure the data is accessible and trustworthy. This document covers our principles, policies, and procedures for each stage of the digital data lifecycle at DataFirst. It is a framework for how we prepare, preserve and release datasets to ensure the long-term availability of high-quality African data from our repository. The aim of the digital preservation policies and procedures recorded in this document is to make sure that data curation at DataFirst is undertaken within a clear data governance framework. Active governance of data ensures that our preservation activities are principles-based and uniform across datasets and support the durability and usability of the digital objects that make up the datasets in our collection. Adherence to clear policies that are based on principles ensures our datasets are:

- *Discoverable* - through keyword indexing and the creation of other discoverability metadata
- *Trustworthy* - through checks and validations at the deposit (ingest) stage and throughout the data life cycle at DataFirst to ensure data integrity and reliability
- *Usable*, that is, prepared as research-ready datasets accompanied by informative metadata to assist data access and data analysis
- *Sustainable* through being preserved in a state suitable for long-term storage and access

Recording and regularly assessing principles, policies and processes for digital preservation is important to help us to maintain our standing as a Trusted Digital Repository. It guarantees that DataFirst's practices continue to satisfy the needs of our data users and comply with data curation standards.

1.1 Digital Curation Reference Model

Our framework for digital curation at DataFirst is informed by the Open Archival Information System (OAIS) reference model. The OAIS is an international standard for digital preservation systems and system components (ISO 14721, 2018). The OAIS reference model depicts the stages of digital preservation of a dataset. The first stage is where the repository accepts data deposits. In the OAIS model this is called the Ingest stage and the dataset at this stage is called the Submission Information Package (SIP). The subsequent stages include Preparation of digital objects that make up the datasets and then Archival Storage of the dataset, referred to at the Archival Information Package (AIP) at this stage. The final stage of the OAIS is distributing of the dataset in its most usable form. At this stage the

dataset is called the Dissemination Information Package (DIP). Figure 1 shows the OAIS reference model.

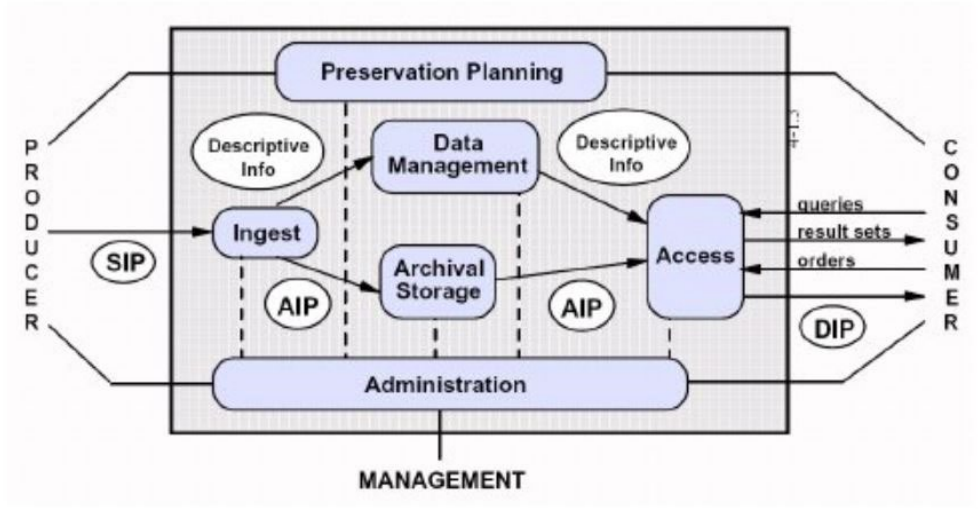

Figure 1. The Open Archival Information System model (ISO, 2018)

Our Digital Curation Reference Model aligns with the OAIS but uses our terminology and is adapted for our curation environment. The numbering in the model reflects each stage of the digital data lifecycle at DataFirst. Our digital curation principles, policies and procedures are discussed in the next sections with reference to our OAIS-based Digital Curation Reference Model depicted as Figure 2.

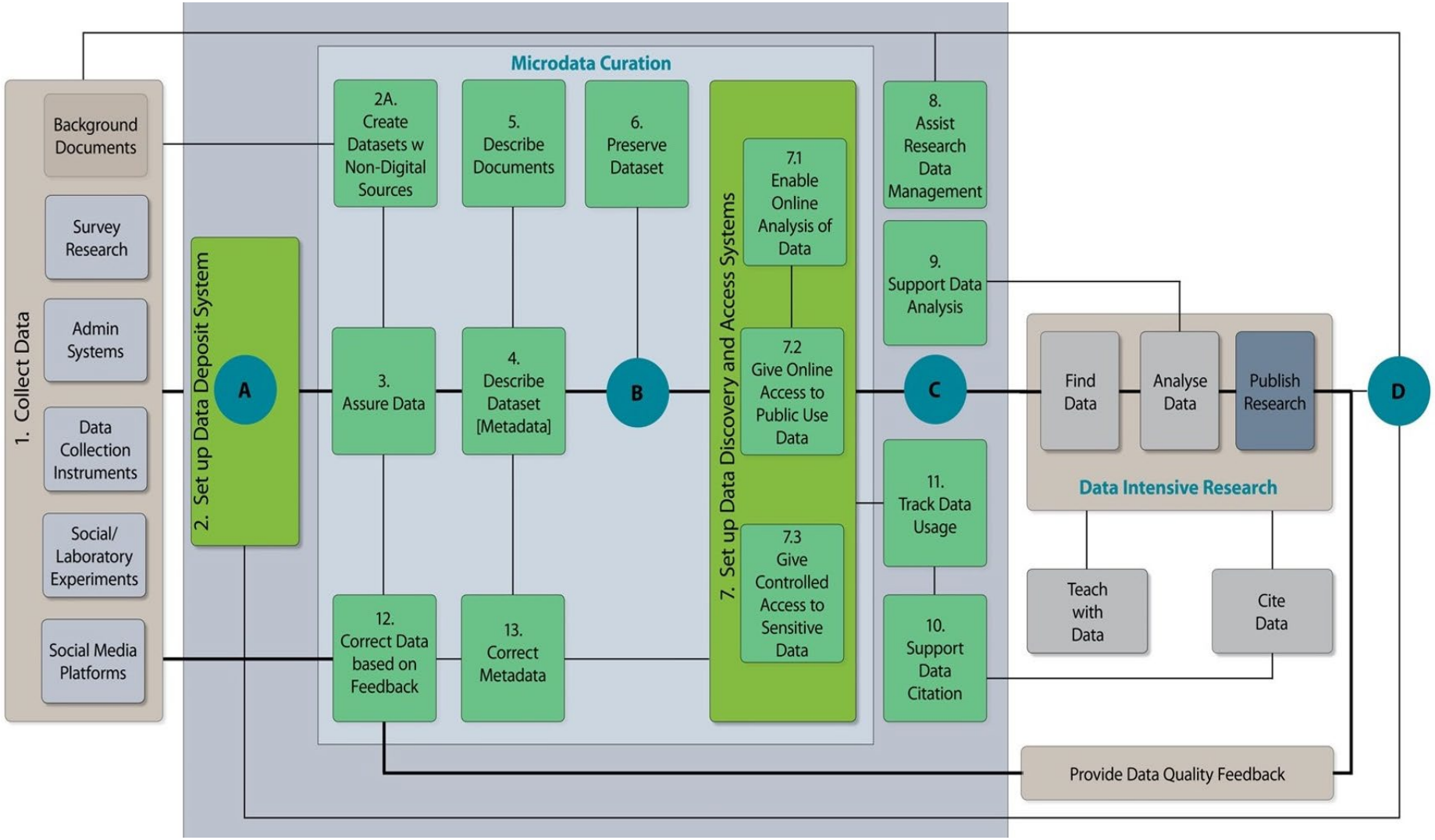

Figure 2. Digital Curation Reference Model at DataFirst (Woolfrey, 2015)

1.2 Summary of Stages in our Reference Model for Digital Curation and Preservation at DataFirst

As depicted in our model, Stage 1 (Data Collection) is outside our workflows, as we do not collect data. We source data and documents from ongoing institutional partnerships and via the academic and policy research community. Digital Preservation activities at DataFirst begin with Stage 2 which involves accepting data deposits (Ingest Stage in the OAIS). We refer to the dataset at this stage as the *Deposit Dataset* (SIP in OAIS) and depicted as A in our model. Stages 3 to 6 involve digital preservation activities in which we make changes to convert the deposited dataset for preservation and reuse. These activities contribute value to datasets, by protecting the integrity of their content and by increasing their usability. We call the archived dataset the *Preservation Dataset* (AIP), shown as B in our model. Stage 7 is where we make datasets available for reuse. The dataset at this stage is referred to as the *Dissemination Dataset* (DIP) and is C in our model. Stages 8 to 11 are those in which we support researchers to discover, download, analyse, and cite datasets. Stages 12 and 13 depict how we utilise the "virtuous cycle of open data" to advance the quality of the data and metadata. In these stages we draw on user feedback and consult with depositors to fix data errors or document data issues. Researchers may produce secondary datasets based on their analysis of our datasets and this type of dataset is depicted as D in our reference model. The next sections discuss our digital preservation policies and procedures according to the stages in our model.

Stage 2: Managing Data Deposits

2.1 Deposit System

Accepting data from our Depositors is depicted as Stage 2 in our Digital Curation Model and aligns with activities in the Ingest stage in the OAIS. At this stage the dataset is referred to as the *Deposit Dataset* and is depicted as digital object A in our Curation Model. The ingest stage is important for dealing with ethical as well as practical issues around data deposits. Ethical concerns include ownership, consent, and privacy. Practical issues include locating data collection documents and managing secure data transfer, as well as agreeing with Depositors on timeframes for disseminating data. The ideal data transfer method is secure self-deposit by depositors using an application that captures basic metadata. We are working with software developers to upgrade our dissemination software to accept online deposits of data and metadata. Currently however, we arrange methods with depositors depending on the level of security required. Depositors are given a short Data Description form to complete to provide basic metadata on the Deposit Dataset. At this stage sign a Memorandum of Agreement with each depositor which codifies the roles and responsibilities of the Depositor and DataFirst and data sharing arrangements. Information for Depositors can be found on our Deposit Data page on our website.

DataFirst does not collect data. However, we undertake projects to find and digitise historical African data that is at risk of obsolescence because the data is in deteriorating paper formats. We convert these historical records and relevant documents to digital formats and disseminate them openly for further research. Information on Data Rescue Projects undertaken by DataFirst is on our website.

2.2 Open Data Principles

We have built our digital curation activities around open data principles embodied in the Open Data Charter. These principles state that data must be "as open as possible and as closed as necessary" (EC, 2016, p. 4). Open Data principles inform both our Data Collection policies (what data we accept for deposit) and our Data Dissemination policies (how we share data). Open Data Principles state that data must be:

- Accessible - free and online, universally available, and licensed as open
- Complete - entire datasets must be made available, not just data subsets
- Primary - data must be made available in its most granular form, referred to as microdata
- Interoperable - both technologically interoperable (machine-readable) and semantically operable, which depends on adherence to data standards
- Interpretable - shared with metadata that helps researchers to access and use the data
- Sustainable - which refers to both persistence of data access and longevity of data usage support

Our curation activities are also underpinned by the FAIR data principles, a more recent Open Data framework that requires that data and supporting metadata be Findable, Accessible, Interoperable, and Reusable (FAIR) (GO FAIR Initiative, 2022). Table 1 depicts how our repository practices comply with Open Data and FAIR data principles.

Compliance with FAIR and Open Data Principles at DataFirst			
Principle		**Definition**	**Compliance Indicators**
Accessibility	Online access	Data and metadata must be easy to find and download	Data files and metadata are online so data can be discovered and downloaded with ease
	No usage costs	Data must be free	We do not charge the end-user for data access
	Use of open licenses	Data must be clearly licensed as in the public domain	We adopt the Creative Commons CC-BY (Attribution only) license for most of the data we share. Some data are shared under a CC-BY-NC (Non-Commercial Use) license
	Non-discrimination	Anyone should be able to access data for any purpose	Public access data is available to anyone for any purpose. Researchers must register on our data site once. This information is not used for the administration of access rights. Registrants' information is only used in anonymised form to report data usage statistics to our Depositors
	Use of open standards	Data and metadata must be in open formats and not dependent on proprietary software for its analysis	Data is shared in software-agnostic formats such as .csv as well as commonly used statistical analysis programmes. Metadata records can be downloaded in xml.
	Machine-readability	Data and metadata must be stored in formats that can be computer-processed	We adopt international standards for accessible and processable file formats and standards for data and metadata
	Interpretability	Data must be well-documented to support analysis	All data is accompanied by data collection documentation, as well as DDI-compliant metadata to help researchers to access and analyse the data
Primacy		Data must be released at a primary, unit-record level	Data is shared as unit-record data (microdata) and not aggregated
Completeness		Data must be released in its entirety and not as data sub-sets	Entire datasets are released, including data series and all waves of panel survey data
Timeliness		Data must be release as soon after collection as possible	We work with data Depositors for them to deposit data with us in a timely manner
Sustainability		Data and metadata must be available in the long-term, online and versioned	DataFirst's infrastructure supports long-term data preservation and dissemination and ongoing support to data users

Table 1. Repository compliance with Open Data Principles

2.3 Fair Information Practice Principles

Our data ethics approach is informed by the Fair Information Practice Principles (DHEW, 1973, p. xxii). These principles are also known as the Fair Information Principles. Not to be confused with the FAIR framework for Open Data, this core set of principles relates to the personal rights of data subjects. These rights include the right to privacy but also the right to agency (right of consent over what is done with the data they provide) (Kitchin, 2014, p. 212). The Fair Information Principles have a long history and underlie most national privacy laws (Borgesius, 2015, pp. 2076, 2102). Table 2 lists these principles.

Principle	Description
Consent	Data is only generated and disclosed with data subject's consent (the agency aspect of privacy)
Notice	Data subjects are informed that data are being generated and the purpose to which it will be put
Limitation	Use of the data is limited to this purpose or compatible purposes
Choice	Data subjects have an opt-in/opt-out choice on the disclosure or use of their data
Access	Data subjects can access and verify data on themselves
Integrity	Data is accurate, complete, current and reliable
Security	Data is protected from alteration, loss, misuse and unauthorised access
Accountability	The data holder is accountable for ensuring compliance with the above principles

Table 2. Summary of the Fair Information Practice Principles (OECD, 2007, pp. 19-21).

2.4 Data Quality Principles

Data accessibility is a key quality attribute, but we also consider other dimensions of quality in our preservation activities. Our reputation as a trusted digital repository relies on confirming, extending, and documenting the quality of data deposited with us. DataFirst adopts the influential Statistics Canada Quality Guidelines as a framework to determine data quality. Statistics Canada defines quality as "fitness for use" and lists 6 quality attributes or principles for data: Accessibility, accuracy, coherence, interpretability, relevance, and timeliness. Figure 3 illustrates these *Data quality principles,* which are addressed at every stage of digital curation at DataFirst.

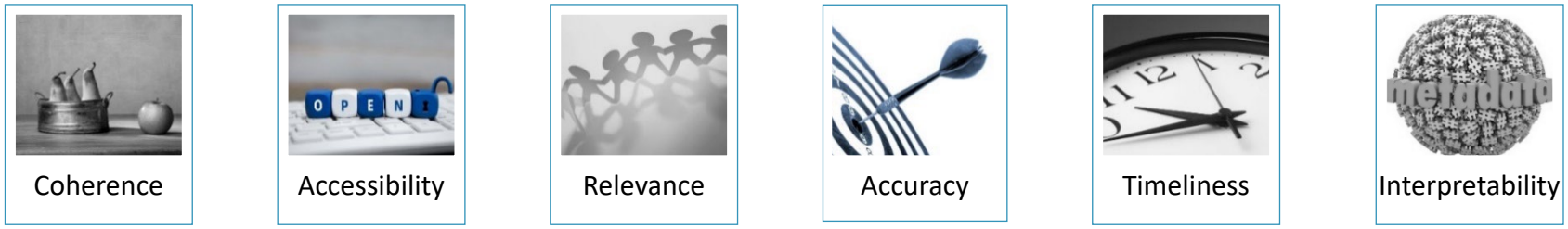

Figure 3. Data Quality Principles (Adapted from Statistics Canada, 2002)

Managing data quality at DataFirst includes setting up controls for data validation, such as adherence to data curation standards and established quality control procedures. Our procedures enable us to monitor and report on incidents that threaten data quality and make recommendations for ways to mitigate such incidences.

2.5 Standards

We maintain and seek to improve the quality of the datasets deposited with us by complying with international data standards and best practices. We consider these benchmarks at every stage of the digital curation process. At the Deposit Stage we must consider safe file transfer protocols as well as compliance with data ethics and data legislation. When preparing datasets and their digital objects we address data privacy and data quality standards, and disclosure control rules, as well as file conversion and file versioning best practices. Metadata records are created according to international metadata schema and citation standards. Archival storage must be based on repository standards and preservation best practice, as well as IT security specifications. Finally, we disseminate data according to open and FAIR data principles, considering data access and data privacy and agency rights, standard licenses, and service quality considerations. Figure 4 depicts standards and best practices relevant at each stage of the digital curation workflow at DataFirst.

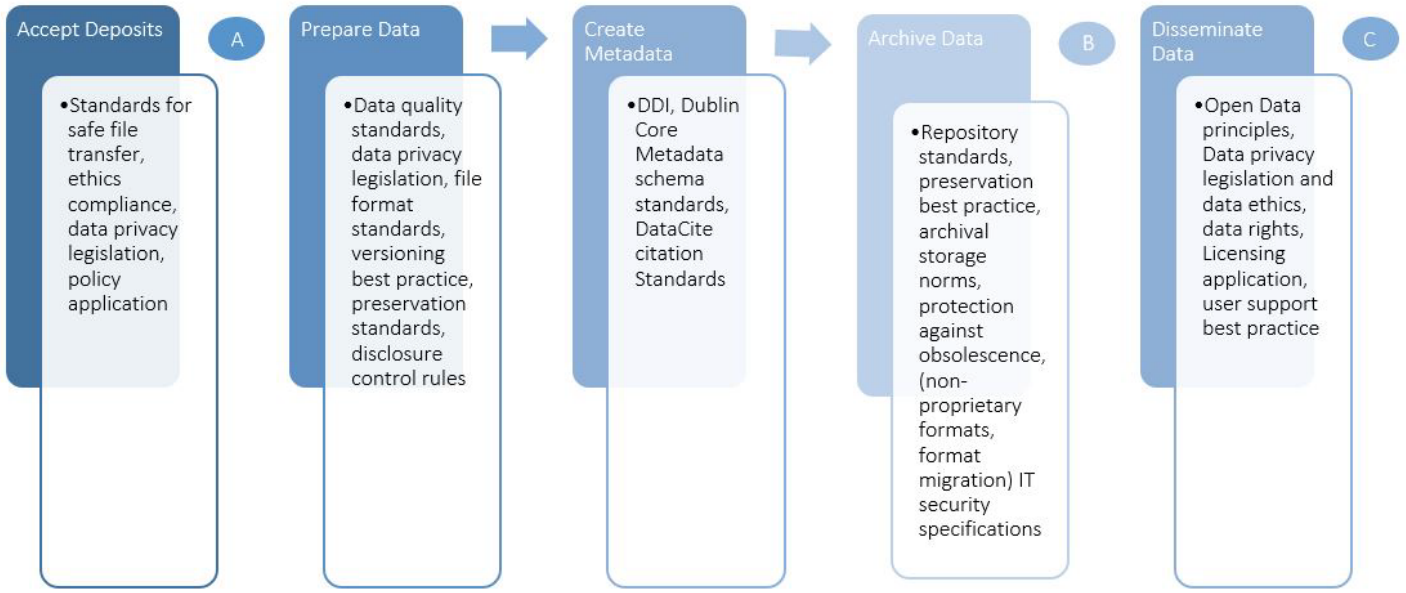

Figure 4. Standards and best practices for digital curation workflows at DataFirst

2.6 Legislative Framework

At the Deposit Stage we are cognisant of and work within national legal and regulatory frameworks as well as within the policies and regulations of our parent institution. The University of Cape Town is the legal entity under which DataFirst operates as a University Department as DataFirst's repository has no legal status. National legislation also informs our curation practices at the Dissemination Stage. Relevant SA legislation includes the Copyright Act 1978 and its Amendments, the Statistics Act of 1999 and the Protection of Personal Information Act (POPIA) of 2013.

2.7 Depositor Agreements

Our relationship with Depositors is governed by legally binding Agreements. These can be Contracts, Memoranda of Agreement (MOA), or Service-Level-Agreements (SLA). For example, we enter legally binding data-sharing contracts with national and local government agencies. All contracts with Depositors are signed by the University's Research Contracts and Innovation Office as the University is the legal entity under which DataFirst functions. We sign SLAs with large-scale University Research Projects that undertake regular data collect activities. MOA's and Service-Level Agreements with Depositors are signed by

DataFirst's Data Services Manager acting for the Director of DataFirst. The choice of contract is determined by the Depositor. Contract and SLAs are created for specific Depositors, but we use a standard MOA. Our MOA fully documents the relationship between Depositors and DataFirst, namely:

- The rights and responsibilities of the Depositor and DataFirst, as parties to the agreement
- Copyright conditions or waivers, including the moral right under copyright law to be acknowledged as the data originator
- Agreement on the access conditions under which datasets are shared with third parties (Access and Use Licenses)
- Depositor statement that they are the data owners or have permission from the data owners to deposit data with DataFirst
- Confirmation that the Depositor has the legal right or ethics clearance for the original data collection activity along with consent from data subjects
- Acknowledgement that ownership of datasets does not pass to DataFirst a deposit but rather curatorship and dissemination rights

2.8 Collections Policy

2.8.1 Data Sharing Mission

Mission Statement:
DataFirst is a research data service dedicated to giving open access to data from South Africa and other African countries. We promote high quality research by providing the essential Open Research Data infrastructure for discovering and accessing data and by developing skills among prospective users, particularly in South Africa. We undertake research on the quality and usability of national data and encourage data usage and data sharing.

Our Collections Policy must be aligned to our mission and principles and must consider national legislative and regulatory frameworks. It must also consider relevant University of Cape Town policies, namely, the Research Data Management Policy 2018, and Open Access Policy 2020. We do not accept data we cannot share in some manner. We also require Depositors to confirm in the MOA we sign with them that they are the data owners or have permission from the data owners to deposit data for sharing. DataFirst will not accept deposits where ownership or permissions to share are not clear. Finally, we request that depositors provide evidence that they obtained legal or ethics approval when they collected the data to be deposited with us.

2.8.2 Scope

Our collections policy is to only collect data on African countries. However, some datasets may be from multi-country projects, and we do then accept both the African and non-African components if these make up one dataset linked to a project. For example, we host a dataset on aging in Brazil and South Africa.
Our subject focus is social science, humanities, and health data. This focus aligns with that of social science data repositories in other countries, such as repositories in the Consortium of European Social Science Data Archives (CESSDA) and the International Federation of Data Organisations (IFDO).

We do not collect aggregated data but focus on making primary data (microdata) available because this type of data has high value for policy researchers for giving an accurate picture of local situations. Like the CESSA and IFDO repositories we are actively engaged in sourcing, curating, and sharing this type of data and assisting data users to analyse the data. We source data from African government agencies, donor organisations, and research institutions, amongst others. We curate mainly quantitative data but are building up a collection of data from qualitative research undertaken in African countries. The scale of our datasets ranges from data from large, multi-site research projects e.g. Afrobarometer and multi-year panel surveys e.g. South Africa's National Income Dynamics Study to small data subsets deposited to support research replication and reproducibility. The latter are shared from our African Research Replication and Reproducibility Data Collection page.

Our Collections Policy aims for long-term retention of datasets as data may have ongoing value for future reuse. In some cases, we are the only distributor of a dataset, and have a responsibility to sustain the sharing of this data. Our policy considers the scientific or historical value of datasets. For example, our Data Rescue Projects have increased access to data that can provide valuable insight into South Africa's recent history. We also consider the uniqueness of data and its non-replicability when we source data or encourage data holders to deposit their data for reuse.

Stage 3: Data Preparation and Preservation

Preparing datasets for preservation and sharing is a component of Ingest in the OAIS model as well as part of Preservation and Planning in the OAIS system. Data Preparation and Preservation is depicted in Stages 3-5 in our Digital Curation Reference Model.

3.1 Preparation and Preservation Policies

DataFirst's policy when preparing data and documents for reuse is that preparation activities must be designed to preserve the integrity of the data while also ensuring the data is as usable as possible. Data integrity is considered at each stage of data preparation. Ethics policies also apply at the data preparation stage, as personally identifiable data must be handled securely and disclosive information removed in public use files. Our policy is also to share as much data as possible, within ethics parameters. We may therefore seek permission to share an anonymised release of a dataset originally provided to us as a restricted-access copy and will prepare these datasets in accordance with their access levels.

3.2 Preparation and Preservation Procedures

3.2.1 Branding

One value-added service DataFirst offers Depositors is the creation of a webpage that brings their datasets together under their branding to highlight their work. Depositors supply content and logos for their branded pages. We also work with Depositors who do not already have a name for their dataset to come up with names and acronyms that are memorable. Open Data means accessible data, but

accessibility has many dimensions. Datasets names can influence their use and selecting memorable dataset titles can be a way to highlight datasets to increase their accessibility and can thus be seen as good dissemination practice.

3.2.2 Dataset Digital Objects

Dataset entities include data files and all materials relevant to the collection and ongoing use of the data. Digital objects that are not data files are referred to as External Resources at DataFirst, in line with standard Nesstar DDI-standard compliant metadata terminology. They are generally documents, such as data collection instruments (administrative forms or questionnaires) and reference documents such as codebooks, as well as analytic documents and technical reports. However, these digital objects may be podcasts, videos or other entities. Like data files, external resources are prepared, versioned, and stored in accordance with our digital curation policy and agreements with Depositors. Figure 5 models the digital objects that are prepared for public access and restricted (data enclave) access datasets.

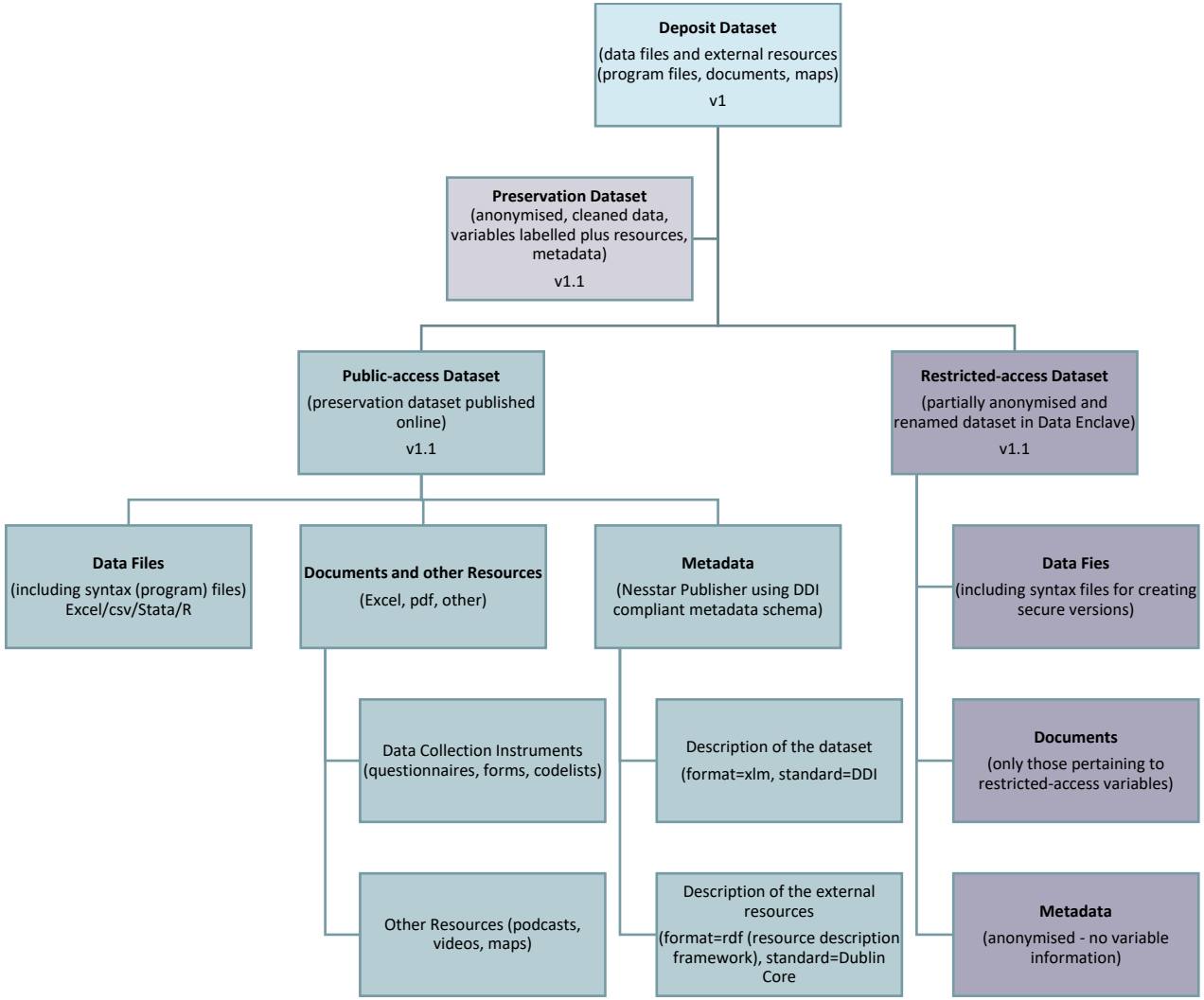

Figure 5 Digital objects prepared for public access and data enclave access dataset releases

3.2.3 Data Cleaning and Quality Checks

The Deposit Dataset (SIP in the OAIS) is the integral copy and is never changed by DataFirst. The Preservation Dataset (Archival Information Package (AIP) in the OAIS) will undergo changes in the data preparation stage which is part of the Ingest stage in the OAIS model. Preservation copies include all iterations of the dataset as we keep all previous copies as a change record and for reference purposes. For example, outdated versions may need to be consulted for tombstone pages on a removed record or citations to an earlier version of the dataset. The Dissemination Dataset (DIP in OAIS) will be the latest and most usable version of the dataset and the one shared online. The Preservation Dataset may be preserved with additional resources such as administrative documents from the data collection stage or created during negotiations with Depositors (MOA, branding information, logos). These are not made publicly available in Dissemination Datasets.

The Preservation Dataset (AIP) and Dissemination Dataset (DIP) may undergo changes and these changes are documented in dataset- and file-level metadata. Change metadata provides an audit trail from the Deposit Dataset to other versions to ensure the authenticity of the data. Data quality checks are undertaken by our Data Analysts but we also crowd-source for quality input from data users who communicate with us through our online support site. Quality assessments include checks for accuracy and consistency. Consistency checks determine whether the data is coherent within the same dataset e.g., whether variables and values are consistently represented and whether separate data files within a dataset can be merged. We also check for data comparability across a dataset series which may be compromised by changes in data collection methods, for example, a change in the sample frame or survey question between survey rounds. Where possible, DataFirst strives to correct data errors. Error correction is carried out in consultation with Depositors. Where error correction is not possible, data errors and other quality issues are documented in the metadata we publish with each dataset.

3.2.4 File Formats

Our Deposit Policy is to accept data files in any format because our Data Analysts have the necessary skills and tools to convert files to our preferred formats. However, our preference is for data files to be provided in commonly used statistical analysis packages such as SPSS or Stata. For preservation and dissemination copies, data files are converted to Stata as this is the statistical analysis software programme used at DataFirst and among many researchers in the quantitative social sciences and humanities. It is also possible to import Stata files into other data analysis software programmes. We also prepare dissemination data files as csv files because this is a software agnostic format which enables us to comply with the Open Data principles of interoperability and non- proprietariness of formats. Storing data files in non-proprietary formats can also prevent data files becoming unusable due to software obsolescence.

Preferred formats for document files are MS Word or Excel. We try to source and share all relevant documents used in the data collection process. These documents also inform the metadata records we create. MS Word documents are converted to pdf and files in both formats are retained in the Preservation copy. Dissemination Datasets include only the pdf copies as well as any document files in Excel. We do not convert spreadsheets to pdf as this reduces their functionality.

3.2.5 File Compression

Compression reduces the size of data files without information loss. Reducing file sizes is important in environments where researchers may struggle to download large files because of limited bandwidth. When we prepare dissemination data files, we compress them so that the files are smaller for researchers to download. Two types of compression are available to us when preparing data files. First, we can use the Windows compression option which allows the compression of almost any file type (right-click > send to > compressed folder). We use this option for all large document or data files. Second, large data files are compressed using the Stata compress command.

3.2.6 File Truncation

At DataFirst we may truncate large data files when we upload them to our metadata creation software, Nesstar Publisher, for the application to store variable and value labels. We do not truncate the data files in the dissemination dataset. Truncation means that we delete (or drop) observations in the dataset when the Stata compression option is not sufficient to reduce the size of the files for uploading to Nesstar. We also use the truncation option to upload variable information for restricted-access data, to ensure the metadata does not show even summary statistics. Truncation involves either (i) keeping 9999 observations (which will show some observations in summary statistics which may be useful for data discovery) or (ii) keeping 0 observations (which will still show variable and value labels). Figure 6 shows our process diagram for compressing and truncating data files during dataset preparation (created by Data Analyst Bruce McDougall, 2020).

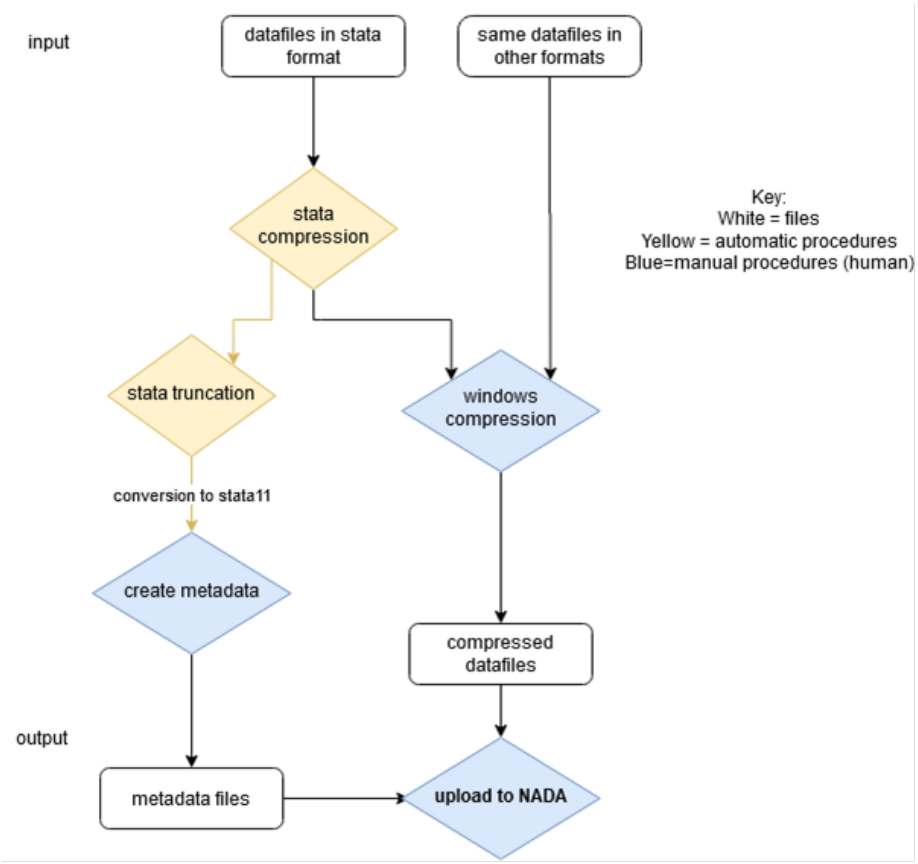

Figure 6. File compression and truncation processes (McDougall, 2020)

3.2.7 File Labelling Conventions

We have developed standard labelling conventions for different file types. Standardised labelling ensures a consistent approach by all staff when preparing data files and other digital objects and supports version control. Like version control, file labelling conventions prevent storing of duplicate files. The file labelling convention adhered to at DataFirst for each type of digital object that make up a dataset are:

- Data files = abbreviated dataset name plus year plus version number. Example: ghs-2002-hhold-v1 for the first version of the data file on households from the General Household Survey conducted in 2002.
- Document files = abbreviated dataset name plus year plus document type. Example: ghs-2002-report for the report from the General Household Survey for 2002. Questionnaires include a q prefix. Example: q-ghs-2002-hhold for the questionnaire used to collect information on households in the General Household Survey 2002.
- Metadata (Nesstar Publisher) files = ISO 3-letter country code plus producer acronym plus abbreviated dataset name plus date plus version number.
  Example: zaf-statssa-ghs-2002-v1 for the metadata record for the first version of Statistics South Africa's General Household Survey of 2002. This labelling convention for metadata files is a standard of the Data Documentation Initiative (DDI) metadata schema.

3.2.8 Version Control/Change Procedures

Some changes may need to be made to digital objects that make up the datasets for the data to be accurate and optimally usable. We use versioning or version control to manage multiple variations of datasets and their digital objects. The benefits for data management of version control are:

- It creates a consistent approach to be applied by all staff collaborating in data preparation
- It creates an audit trail of the change processes that each digital object has undergone
- Along with file labelling conventions, it prevents storing of duplicate copies of digital files
- It clearly identifies the latest version of the data files which should be the only one disseminated for reuse.

Corrections and updates to any of the dataset components will be denoted by a new version number. In the case of a Dataset deposited without version information, we follow a policy of labelling the Deposit Dataset and its files as version 1. Our versioning policy distinguishes between a *new version* and a *new release*. That is, changes we make to deposited data files or documents will be indicated by a new version number denoted by a minor version update and indicated by increments to the decimal place, e.g., v1 becomes v1.1. However, if a Depositor re-issues a dataset the new release will be labelled as the next major version indicated by a whole number e.g., If v1 is recalled the replacement dataset will become v2. Re-releases may occur when errors are discovered in data entities, either by Depositors or by DataFirst. In these cases, the depositor recalls the data and reissues a corrected version.

We version at *file level* and change the version of only the corrected file, leaving other data files with the old version number. We align the dataset version with the version of the most recently changed data file. Researchers have confirmed that they prefer this option as it allows them to download only

the changed files rather than all digital objects that make up a dataset, including unchanged files. This choice is relevant in places where low bandwidth may be an issue when downloading files.

All changes to all digital objects that make up datasets must be documented in the descriptive and structural metadata. Version information is recorded in file labels and file-level metadata as well as in the metadata record for the whole dataset that we post online. Version information can be found in the "Version Notes" field in the online metadata. The metadata records are also versioned, in compliance with the DDI metadata schema. This allows researchers to see immediately that change information is available.

3.2.10 Dataset Deletion

Dataset components may be deleted during the data preparation and preservation stages. In these stages references to dataset content may be removed from access, but the content is retained, although only accessible to repository staff. The UK Data Archive calls this activity "soft deletion" to distinguish it from "hard deletion" where both content and references to content are deleted (UKDA, 2022, p. 14). Soft deletion methods are preferred at DataFirst because there can be no certainty that data content may not be of future value. Hard deletion is available as an option for depositors and written into our MOA with depositors. With soft deletion of a dataset, the change and change purpose are indicated in the administrative metadata as well as the online metadata record of more recent versions of datasets.

3.3 Ethics Compliance Procedures

Ethics compliance practices at DataFirst are concerned with balancing the privacy and agency rights of data subjects with the information access rights of data users within an Open Data framework. Ethics-related data preparation procedures are to establish that ethics clearance was received for data collection as well as data subjects' consent. These procedures also remove potentially disclosive data to respect the privacy of data subjects.

3.3.1 Ethics Clearance and Consent

In line with the Fair Information Practice Principles and our agreements with Depositors, we request documents from Depositors that show ethics clearance for the collection of the data to be deposited with our repository. We also request deposit of any consent forms agreed by data subjects. Consent forms are uploaded with the data and other supporting documentation. We do not share ethics approval documents online but have added an ethics compliance field to our metadata, in which we document details of ethics permissions obtained for data collection, including the institutional entity responsible for ethics approval and the date approval was obtained.

3.3.2 Disclosure Control

We undertake disclosure control on data deposited with us, to ensure we do not publish personally identifiable information. Most deposits come to us already anonymised. But where we do receive personally identifiable data our Data Analysts undertake statistical disclosure control on the data to

create safe usable versions. Our statistical disclosure control processes comply with ethics norms and data privacy legislation and are framed by our Disclosure Control Flowchart The flowchart is depicted in Figure 7.

Disclosure control at DataFirst includes dealing with the following types of identifiers in the data files:

- *Direct identifiers*, that is, variables that directly identify a data subject or entity, such as names and identity numbers. They represent the greatest disclosure risk but are usually easy to deal with and at DataFirst these are removed from the Preservation and Dissemination data files.

- *Indirect Identifiers* include values at either extreme of a distribution (outliers) which may indirectly identify a data subject. Outliers can occur in variables such as age and income. DataFirst deals with outliers in continuous variables by top- and bottom-coding or grouping.

- *Geographic variables* in the data deposited with DataFirst are usually non-disclosive, generally at the level of Province or sometimes District Municipality. However, some data deposited with us has more specific geography which poses disclosure risk. In these cases, we may remove variables at fine geographic levels (such as GPS coordinates or postal codes) in the datasets prepared for public use. Depositors may permit detailed geographic variables to be shared in our Secure Research Data Centre at the University. Low-level geographic variables pose more of a disclosure risk in combination with indirect identifiers e.g., elderly HIV-positive individuals in small villages may be identified by linking the age cohort and village-level geography. In such a case our approach would be to remove the Village variable and top-code or group the age variable.

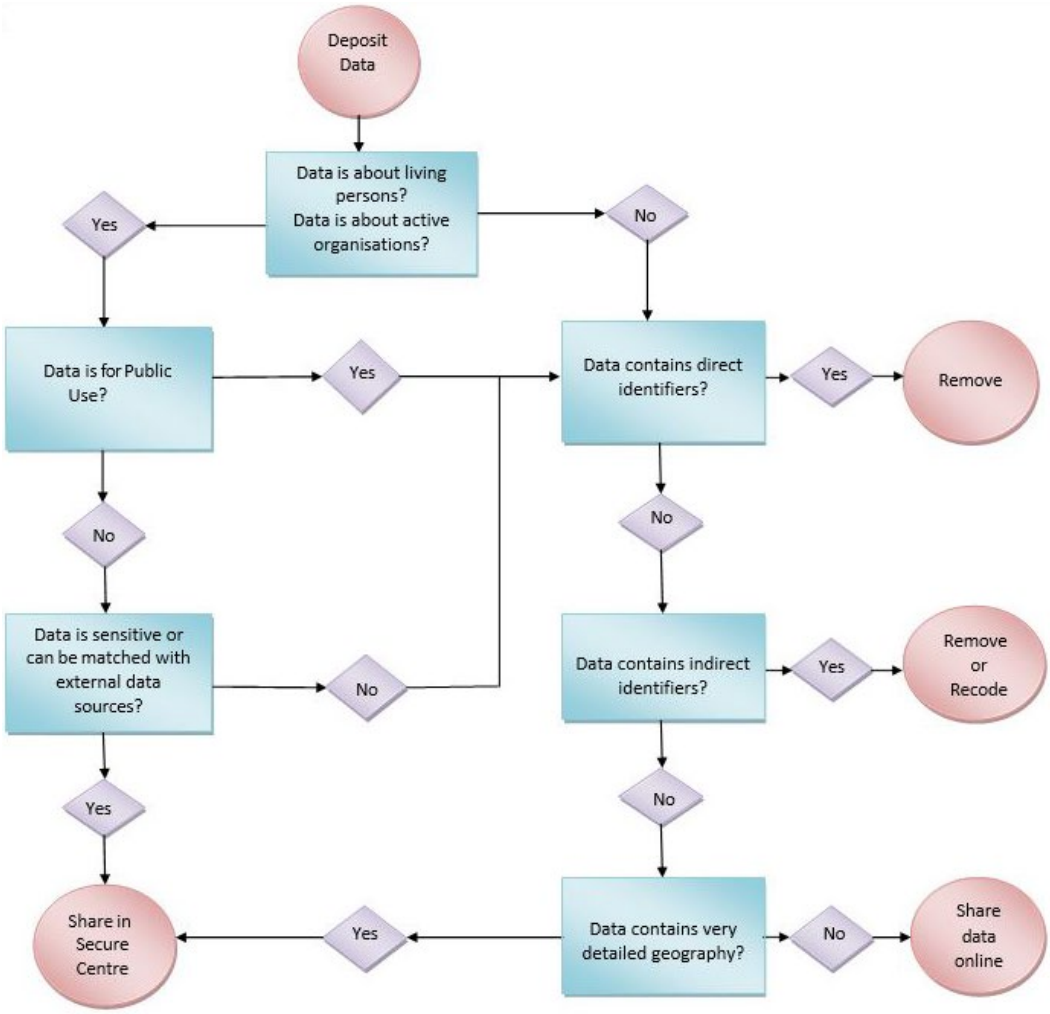

Figure 7. DataFirst's Disclosure Control Flowchart (based on O'Rourke, J. 2012. *ICPSR Disclosure Risk Decision Tree* and *Five essential steps for risk reduction*)

3.4 Dataset Upload Process

Datasets are uploaded to our web-based dissemination application from the Site Administration menu. Data files (in Stata and .csv) documents (in pdf or Excel) and other resources are uploaded to the application. Dataset-level metadata records (in xml) and metadata records for documents and other resources (in rdf) are also uploaded. Data and document files and then linked to their metadata descriptions using the Site Administration options. The Administration page allows for the setting of data access levels and creation of a pdf copy of the full metadata record for download. Files can be deleted and replaced on this page as changes are made. There are also options to bring datasets together under a "Collections" page, for example all datasets from the same Depositor

Stages 4 and 5: Metadata Management

Stages 4 and 5 in our Digital Curation Model is where we create standardised descriptions of the datasets and their entities to enable researchers to find, interpret, and analyse data. When presented in standardised schema these descriptions are known as metadata. Descriptive metadata includes metadata elements that allow for identification of datasets, for example, title, author, and abstract. In addition, administrative metadata includes management and use information, such as version history and access permissions. Some metadata elements that our researchers find of value are information on the scope (subject area) of the data and lowest level of geographic aggregation of variables. Our metadata records also help researchers to cite datasets in their research publications. We include a Recommended Citation field in our metadata record to assist researchers to cite data according to the DataCite standard. The citation includes a persistent identifier(Direct Object Identifier (DOI)) which we generate for each of our datasets on DataCite's Fabrica DOI registration and management platform which is a service available to us as a DataCite member organisation.

Our repository has adopted the Data Documentation Initiative (DDI) metadata schema for describing socioeconomic datasets. Records created according to the DDI schema are expressed in XML (eXtensible Markup Language) which allows the markup of the content of metadata records so that they are standardised and thus machine-readable and interoperable. We use the Nesstar Publisher free data markup software for the creation of our xml-compliant metadata. In Publisher we can create dataset-level metadata using templates customised for our Dataset Collection and harvest variable-level metadata from data files for inclusion in the final metadata record. Variable-level metadata includes variable and value names and labels and summary statistics.

Using the Nesstar Publisher application we also create metadata records for documents and other resources to be shared with data files. In Nesstar these elements are referred to as External Resources. Publisher uses the simpler Dublin Core metadata schema for creating metadata records for External Resources which are exported in the Resource Description Format (rdf) standard for web data. The dataset and variable-level metadata record (in xml) and the external resource-level metadata record (in rdf) are web-compliant which enables them to be posted online with the data and document files. The Nesstar Publisher markup software is preferred by us as it is freeware and so can be shared with Data Managers in under-resourced institutions. It is bundled with the freeware data dissemination software application we use for disseminating data from our open data site.

At DataFirst our Data Analysts are responsible for creating metadata for the datasets they prepare. However, metadata records are regularly reviewed by our Data Services and Operations Manager to ensure the information they provide is accurate and that staff adhere to our standard metadata schema and are using the latest metadata templates customised for our dataset Collection.

Stage 6: Archival Storage and Sustained Preservation

6.1 Archival Storage and Long-Term Preservation

Retention of datasets includes archival storage but also managing of archival copies, backing up and long-term maintenance. Archiving and long-term preservation of datasets is shown as Stage 6 in our Digital Curation Reference Model as *Preserve Dataset* and the Preservation Dataset(s) are shown as B in the model. This stage aligns with the "Archival storage" function in the OAIS model.

At DataFirst the aim of archiving or storage management is to preserve the accession, storage, and use copies of a dataset according to repository standards to ensure their integrity and continued accessibility. Our server infrastructure includes first, one virtual server, our Preservation Server, which houses our datasets and is hosted by the eResearch Department at the University. Second, our infrastructure includes one physical server with our dissemination software application and dissemination copies of datasets. Third, we have a physical server which is a secure server for hosting our restricted-access datasets. The physical servers are housed in a secure facility in our IT department on campus. Our Preservation Server has a standardised directory system, with folders for data, documents, and metadata files. The Secure Server has folders for user accounts and for each of the secure datasets which have the standard three folders for data, documents and metadata. Figure 9 depicts the one virtual server and two physical servers that make up our server infrastructure

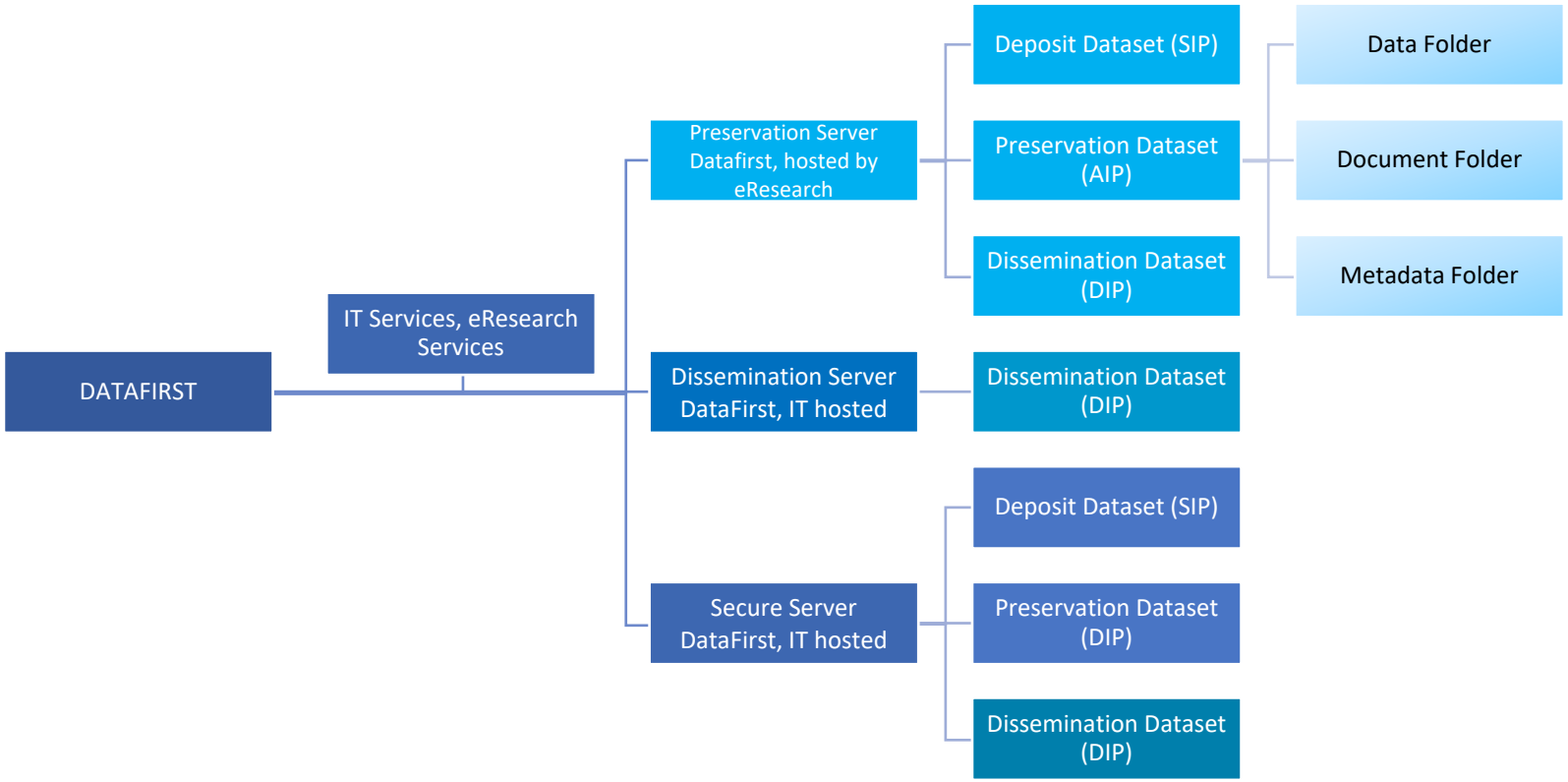

Figure 8. DataFirst's storage infrastructure and server directory system

6.1.1 Our Preservation Server

We archive multiple copies of datasets on our Preservation Server, namely:

- The *Deposit Dataset* (Submission Information Package (SIP), in OAIS terminology) which is the original accession copy deposited with DataFirst.
- The *Preservation Dataset* (Archival Information Package (AIP)) which is the copy converted for storage. The AIP will seldom be one copy, as versions of the DIP are replaced and retained. Changes may be to correct dataset entities, or to migrate digital objects to new formats to ensure that the data and document contents remain accessible.
- The *Dissemination Dataset* (Dissemination Information Package) which is the research-ready copy made available online

Our Preservation Server's standardised directory system includes a folder for each dataset within which there are sub-folders for data files, document files, and metadata files. File labelling is standardised: Data and documents are labelled according to our in-house standard and metadata files are labelled according to the DDI metadata schema. The principle behind our file labelling convention is that the parent dataset of an individual file and the file content should be immediately identifiable from the file label. This is important in an environment where researchers may download individual files from a dataset for their research purpose rather than all the files in a dataset. A consistent file structure also speeds up server searches and allows for cross-folder changes to be easily automated. Figure 9 shows the directory system of our Preservation Server.

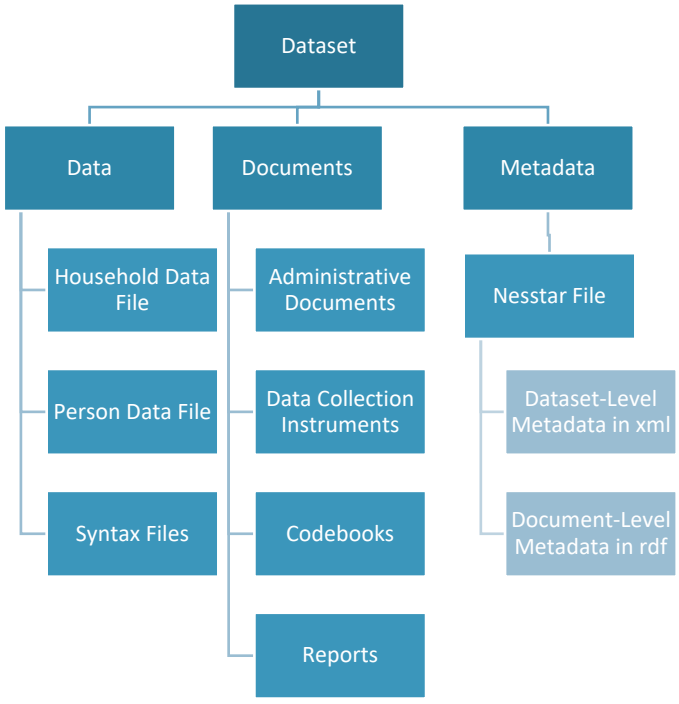

Figure 9. Directory System of DataFirst's Preservation Server

Digital objects may also be migrated to new formats ahead of changes in technologies that may render some formats obsolete. These digital preservation activities ensure the data and metadata keep their

integrity and remain usable over time. File conversions to avoid software obsolescence must be documented in metadata at the dataset level and data file level e.g., in variable labels, to show that a format change has not altered the data content.

6.1.2 Preserving Access through Preventing File Obsolescence

A Preservation Dataset includes the original (deposit) copy which we keep unchanged as a record of the deposit and in case we need to reference the original copy. All digital objects that make up the dataset are converted to preservation formats relevant to the data type. Long-term preservation of digital objects must consider loss prevention. Preservation strategies must consider the risk of file obsolescence when hardware and software environments become outdated. This may result in data or document files being unreadable. At DataFirst we adopt the following strategies to protect against the risk of digital objects in our collection becoming inaccessible through software or hardware obsolescence:

- Intentional Redundancy - we store multiple copies of AIP and DIP files on different media
- Multiple Formats – we convert data files to a number of software analysis programme formats. Files must also be stored in software independent formats so they can be usable beyond the life of the software programme in which they were created. At DataFirst data files are stored in ASCII format or more recently as .csv files. Open formats that have a high degree of interoperability and assist our FAIR data compliance. Document and programme files are stored in text format as well as converted to pdf for the DIP.
- Data Migration – We undertake two types of data migration. First, we convert data files from older versions of data analysis software programmes to the latest iterations, when these become available. Second, we change file formats of both data and document files when existing formats are in danger of obsolescence e.g., we converted older questionnaires in WordPerfect to Rich Text and MS Word before that word processing package went off the market. These strategies allow for ongoing access for data types in the event of format changes.
- Software Updates - We ensure we use the latest versions of software programmes, to protect our collection against software obsolescence
- Detailed metadata records also play a role in supporting the fidelity of digital objects by describing elements of the data which may be changed during migration. For this reason we keep copies of metadata records on different media to the AIP and DIP copies.
- Hardware obsolescence is protected against at DataFirst by hardware monitoring, technical support for server problems, and regular server upgrades, as well as off-site backups to different hardware systems, managed by our IT support department.

Obsolescence checks are a component of our annual review of our preservation server and server and PC software. Figure 7 shows the obsolescence decision-tree used at DataFirst to ensure files remain accessible in their original and reworked forms in the long term.

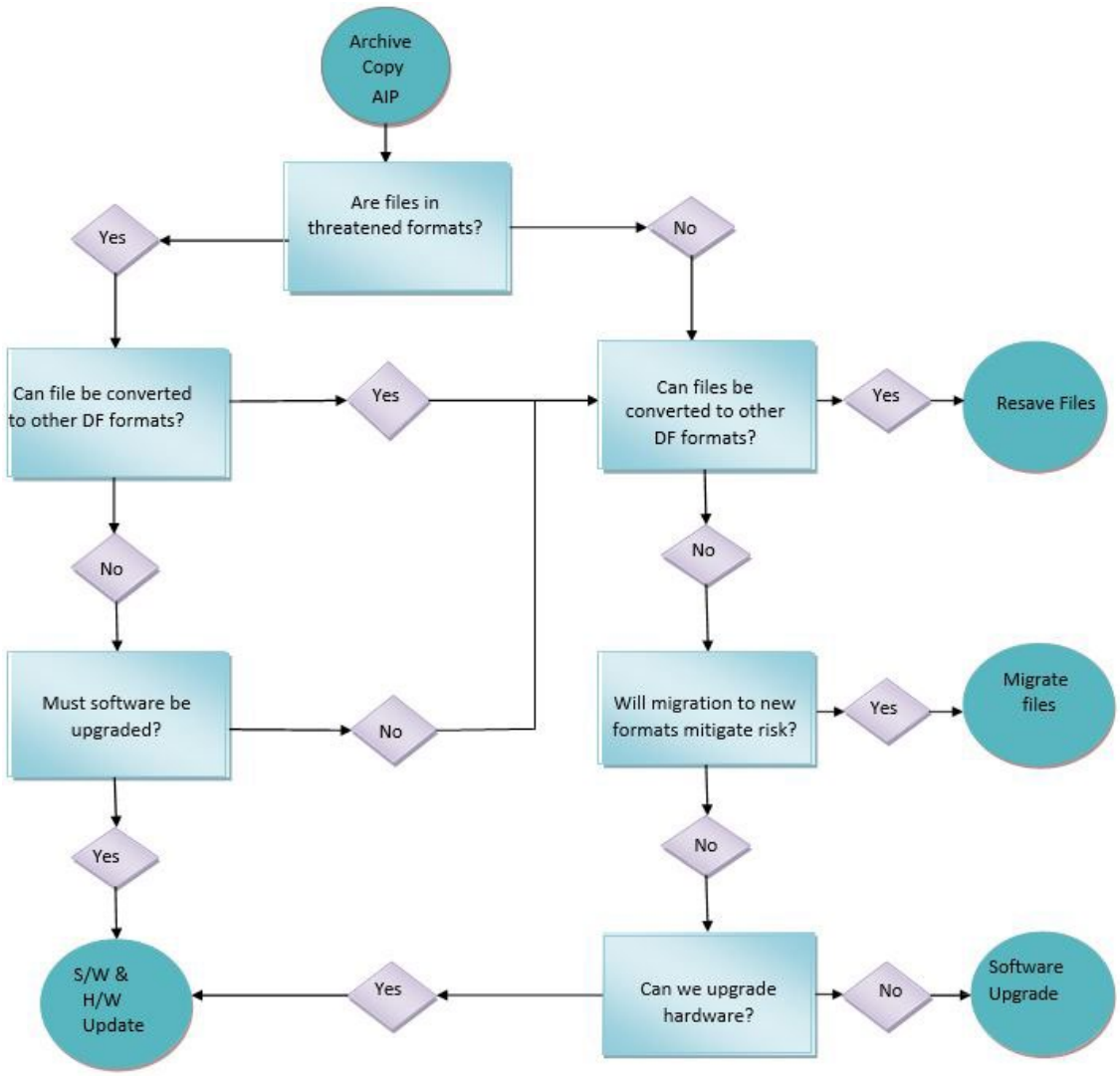

Figure 10. File obsolescence decision-tree

Changes to digital objects may be necessary, such as when errors are corrected, or variables and values labelled or relabelled. Inaccurate documentation will also need to be replaced. Any such changes will be indicated by new file versions. Each version of a digital object is stored with its preservation metadata. Metadata records are also versioned - in the metadata file label but also in the content of the metadata record. Versioning of metadata records allows researchers to see immediately whether additions have been made to the metadata we post online. Superseded Nesstar metadata files are retained on the Preservation Server as well as the Dissemination Server with only the latest version published.

6.1.3 Software

Our data curation operations and services utilise data analysis and data curation software, namely Excel, R and Stata for data analysis and data cleaning, the free Nesstar Publisher metadata editing software for creating web-compliant metadata records. Our Dissemination Server is installed with the National Data Archive software web application which is Freeware developed and maintained by the World Bank

for hosting and disseminating micro-datafiles. Finally, we subscribe to the Freshworks web-based customer support system for our online helpdesk which is a ticketing system that allows us to respond to and document user queries, as well as generate query statistics for reporting purposes.

6.2 Archival and Network Security

Physical and logical security are an integral aspect of the data preservation or archiving stage. However, at DataFirst security procedures are considered at all stages of digital curation. Security policies and procedures aim to protect data confidentiality, prevent unauthorised data access and unintended changes to data, and loss prevention strategies to protect datasets from damage or destruction. Security procedures are also relevant to ensure appropriate soft or hard deletion of records. Our preservation server holds multiple copies of datasets and, to prevent loss of data and ensure data integrity, we manage access to this server by this server by assigning access rights to authorised DataFirst and IT personnel according to clear rules.

6.2.1 Security During Data Preparation and Preservation

At DataFirst stricter security procedures apply where we receive potentially disclosive data and prepare anonymised versions for public access and when we prepare the data to be shared in our on-site Secure Research Data Centre. Data security is considered from the deposit stage. We do not use general file-sharing services to receive data deposits if the depositors cannot attest to their anonymity. Receipt of data is through encrypted access to temporary secure accounts. Secure storage involves storing sensitive data separately and restricting access to servers and spaces to those with authorisation. We remove disclosive data content and store this content as data subsets on our Secure Server. Disclosive data removed from public access versions may include variables that are direct identifiers, as well as other identifying elements, such as found in family pedigrees, on questionnaire flaps, and completed consent forms. The data files stored on our secure server have unique IDs for linking to parent data to enable relinking of data by authorised staff for possible future data checks.

Loss of data on our servers is prevented by multiple copy resilience. The preservation system is duplicated three times, first on our Preservation Server, second on a backup server in the server room in the IT Department which is in another location at the University. The third copy is held on our dissemination server from which copies are downloaded. This means that there is system redundance so that data can be regenerated if non-recoverable disk errors occur. Monitoring and updates of servers and peripherals as well as server software is undertaken by IT staff who regularly check our servers for wear and tear. They also check for out-of-warranty states which require us to either replace our servers or purchase additional maintenance licenses to carry us through to replacement dates. Servers are repaired when issues are identified by the IT Department and are generally replaced every five years to ensure we have enough storage capacity, and that they remain compatible with server software.

6.2.2 Physical and Network Security

The security of the servers housing our datasets is the responsibility of the University's IT Department. The University's Health, Safety and Environment staff and Campus Security Services are responsible for

fire protection, environmental control, and intruder prevention for UCT buildings. All doors in our building, including the IT offices, are protected by card-reader systems and fitted with alarms. Virtual server access is password protected. The University of Cape Town's IT department has set up stringent firewall protection and provides upgrades and patches to server software to protect against computer viruses and malicious codes. Physical servers are protected against power-surge and power outage damage with uninterruptible power supply (UPS) systems. The building where our physical servers are housed also has a generator. These backup power systems are important as South Africa is currently experiencing regular power outages. Security procedures must account for the risks of cloud-based storage, which is not necessarily secure or hosted in-country. Our preservation server is a virtual server hosted on a University of Cape Town cloud and maintained by the UCT's e-Research Services Department. This option is not appropriate for high-risk data like the datasets accessed in our Secure Centre.

6.2.3 Archival and Network Security for Restricted-Access Datasets

Security procedures are more stringent for sensitive data such as personally identifiable information, in line with our ethics principles and national data protection legislation. Our secure (restricted-access) datasets are housed on an institution-administered physical server in a boundary-controlled on-site server room in our IT Department. Our Secure Research Data Centre has approved protected access repository status under the Open Science Framework. Network security requires that the Secure Server is not connected to any external networks or internet services. Backup is done with a temporary link to another unconnected server. Our on-site Secure Centre is a wi-fi free space which has boundary protection with biometric readers. The Centre's workstations have no active ports and user accounts are password-secured. Researchers who are accredited by DataFirst to access data in the Centre must sign non-disclosure agreements in their application to access the Centre. They must also participate in an information session on disclosure risk run by our Data Services and Operations Manager. Dissemination datasets are copied to secure user accounts once users have permissions to access the data in the Centre and user accounts are deactivated at project close-out. Procedures for our Secure Centre are covered in our *Secure Research Data Centre procedures manual.* Our security compliance document for the Centre can be downloaded from the Secure Research Data Centre webpage.

Stage 7: Data Dissemination

Stage 7 in our Digital Curation Model is Data Dissemination which aligns with the "Access" function in the OAIS reference model. The Dissemination Dataset equates to the Dissemination Information Package (DIP) in the OAIS model and is depicted as C in our model. This stage is where the repository user locates, requests, and downloads data. The role of DataFirst at this stage is to have systems in place for data discovery and for access to data files and documents and other resources. Dissemination systems must include protocols that support data ethics and data security.

7.1 Dissemination Infrastructure

Our application infrastructure for disseminating data is built on the National Data Archive (NADA) data cataloguing and dissemination software application which is freeware developed by the World Bank's

International Household Survey Network. The NADA software is bundled with the free Nesstar Publisher metadata software to support presentation of online metadata as well as data downloads. We adopted this free and readily available software because we can install it at resource-constrained African institutions when we train data and metadata management teams. The software is maintained and updated by our Data Services and Operations staff in collaboration with UCT's IT department and software developers at the World Bank. DataFirst has given input to the development of the application since 2010 including suggesting new functionalities such as the Citations function which allows for the linking of datasets to citations to research publications based on our data. The DataFirst Data Operations Team have also partnered with donor agencies to install the software and train African Data Managers to prepare and distribute data using the NADA application.

Our Open Data site based on the NADA software allows researchers to search across datasets by keyword at dataset and variable level. The Collection can also be searched by Country or Date/Date Range. Researchers must register once (Login-Register-submit form) on the site and activate their account in an email link to be able to download data files. Registration information does not influence data access and registrants' information is only used in anonymised form to report data usage statistics to Depositors and for service improvements. Documentation can be downloaded from the "Downloads" tab without registration. The metadata record can be downloaded in pdf or xml/rdf. Downloading data files involves logging in, selecting a dataset, and the "Get Microdata" tab on the dataset landing-page and completing and submitting an online data request form. Data files are then immediately available to download. Non-commercial access requests follow the same process but are actioned through an email link sent to registrants.

We also create metadata records on African datasets that are openly available from other repositories. We obtain permission from data holders to create this discovery metadata and to link to their sites to help researchers to discover and use other open African data sources.

7.2 Data Access and Use Licenses

Intellectual Property Rights may exist on data, in the form of copyright on datasets, or usage restrictions may exist for some data. It is therefore important that repositories clearly license their data to explain how the data can be accessed and used. We have adopted *Open Copyright licenses* that allow data producers to legally share their work under open access conditions. The Open Copyright license regime we use is that of the Creative Commons (CC) which is an open knowledge non-profit. We have adopted 3 of the 6 CC licenses for the datasets we disseminate, namely:

- CC-BY Attribution-only Licence which makes data open on condition data producers receive attribution. We also ask that DataFirst be acknowledged as the data distributor
- CC-BY-SA Attribution plus Share-Alike License which makes data available for with an attribution requirement but also requires that any works based on the data be shared under the same license
- CC-BY-NC License for data that can be used for Non-Commercial purposes only. Datasets shared under this regime are called "Licensed" data on our site. This license is used only where data may be sensitive, or where we would otherwise not be able to share the data.

Our dataset collection can be searched by License Type by selecting the "License" option on the left sidebar. An additional license is for *Data Enclave* access and use. This license is used for restricted-access data available for use in our on-site Secure Research Data Centre at the University of Cape Town. Centre use is reserved for sensitive or potentially disclosive data such as data with household GPS coordinate variables, or detailed firm-level data. The Centre is open to all researchers based at research institutions who present research projects that can be undertaken with the data. The metadata for datasets in the Centre (Data Enclave) is available online and researchers can apply for access to the Centre by submitting an application form to our Helpdesk.

Researcher compliance with license conditions is obtained through various compliance measures, namely:

- We include an up-front agreement in our online data request form. Registrants must agree to comply with the relevant license conditions before they can access data files. The data request form can be viewed from the *Get Microdata* tab on the landing-pages of datasets shared under a CC-BY license.
- We commit to making data as open as possible and only as closed as necessary, as a FAIR data principle. From more than 20 years of sharing research data we have found that if researcher trust us to share data as openly as possible they will be trustworthy in their use of data.
- We check that researchers cite data producers in their research publications when we update our citation records. Where possible, we follow up with researchers who fail to cite data sources.
- In the case of CC-BY-NC licensed data there is an additional step to access data, where we check the intended use information in the data request form before giving online access. The application form for non-commercial use data can be viewed from the "Get Microdata" tab on the landing-pages of the data shared under a CC-BY-NC license (called "Licensed" data on the application form).
- For restricted-access data shared in our Secure Centre, compliance is obtained by a data-use agreement included in the application form to access the Centre which is signed by the researcher and a representative of their institution

Stages 8-10: Support to Researchers

Researchers and other users of our data are referred to as Clients in our Digital Curation Model and are called "Customers" in the OAIS Model. Client support takes place in Stages 8-10 of our Digital Curation Reference Model. Stage 8 depicts the role we play in research data management at the University and for the wider academic community. We run Research Data Management workshops to assist Research Projects, early career researchers, and postgraduates to manage their data during the life of their research and to share their data at project close-off. In this way we enable researchers to fulfil any data sharing mandates from research funding agencies, and we also encourage voluntary data sharing.

As shown in Stage 9, we assist researchers to analyse the data they download from our site. First, we assist with data access and data analysis queries from researchers via our online helpdesk. Researchers can contact our support site via email or post a query directly to the Helpdesk. We also offer regular data analysis and data curation training workshops.

Once researchers have downloaded data, we assist them to cite the data according to international citation standards. This support is shown as Stage 10 in our model. Citation information is given in a citation field in the metadata for each dataset. We also include information on our website on how to cite datasets in research publications.