Data Management

Overview: why manage data?

Legislative requirements.
Under the legislation such as the Statistics Act, data collected as part of government business is a record and must be managed throughout its life cycle. In the Pacific Islands region in most countries this means from the time it is created until it is archived or destroyed. This includes administrative data collected as part of the general business of an organisation and the data created by clients’ interaction with an agency, as well as survey data.

Demand for data.
The increased demand for evidence-based decision making has increased the demand for data. Data collected for business use can also be used to provide better client service, improve organisational efficiency and support innovation. Therefore, it needs to be managed to ensure discovery and re-use.

Transparency and security
The General Data Protection Regulation (GDPR) passed by the European Union, together with data breaches such as inappropriate release of Facebook data, has raised awareness of the need to secure data about individuals, whether collected by surveys or as part of business as usual. Being able to demonstrate that an organisation is taking care of the data it collects is important for ensuring that those providing data feel confident that their information is safe.

Use/re-use
"Metadata are critical to effective data use as they convey information that is necessary to fully exploit the analytic potential of the data. Because it is often impossible for secondary researchers to ask questions of the original data producers, metadata becomes the de facto form of communication between them." (ICPSR Guide to Social Science Data Preparation and Archiving)

Metadata and interoperability enable data to be used by different systems and users with different needs.

Data Management Plan
A data management plan puts information and context around data, enabling better discovery and re-use. It demonstrates that the data is being cared for appropriately. It provides transparency and assurance that data is secure and will not be accessed inappropriately. It identifies any legislative or contractual requirements around accessing and using the data.
What needs to be included?

**Governance and Management:**
Governance is about looking after the data in your organisation.

**Administration**
Administration ensures that records are kept of the important decisions made about the data.

Examples:
- Obtaining proof of approvals
  - to collect data from data suppliers/survey participants
  - to use collected data for an agreed purpose
  - to move data from one step in the data flow to another
  - consent to make the data available and to retain/archive it
- Ensuring legislative requirements are met
- Determining how data is kept, stored or destroyed

**Responsibilities for managing data**
Data creators need to understand the requirements for ensuring data transparency and management.

Examples:
Make sure data creators understand how:
- informed consent will be handled
- appropriate access to the data is managed throughout its life cycle
- data will meet the criteria to move to the next stage in its flow
- to ensure the dataset does not become corrupt
- long-term access will be managed at the end of the data life cycle
- decisions around disposal will be implemented.

**Access and security**
The access and security of the data is important to those who provided it. It enables transparency about the openness of the data, including reasons for limiting access.

Examples:
- Describe access and use throughout the data life cycle
- Identify any barriers to stop sharing
- Describe how others will access the data
- Identify specific privacy, security or confidentiality concerns and how these will be handled
  - ensure data is secured against unauthorised use
  - describe how privacy will be protected
- Identify how data will be shared.
Discovery and re-use
Data users do not have the opportunity to have conversations with the data creators, especially when the data is administrative data. Describing the data enables potential users to understand it and decide if they can use it. It is often the only form of communication with the researcher.

Data formats, volume and storage
Describe the formats in which the data is generated, maintained and made available
Examples
- Identify specific tools or software needed to create/process/visualize the data
- Identify storage locations and any specific backup procedures required
- Note the size of the dataset

File formats
For the time being, while the obsolescence of media is not thought to be a major issue in the Pacific with both CSPro or Stata being used as a standard format. In addition, the Pacific Data Library will always preserve both the data and the metadata in the original format. Another important consideration for data curation is what file formats to use. Careful thought about files at the beginning of a project or planned data collection can save a lot of time, money and heartache later in a project. File formats govern the ability to use and reuse data in the future, with the ongoing accessibility of data an important consideration.

Formats more likely to be accessible in the future are non-proprietary, open, documented standard commonly usage by research community, standard representation (ASCII, Unicode), unencrypted and uncompressed.

Data documentation
Documenting the dataset encourages discovery and re-use.
Examples
- Describe how the dataset was created or extracted
  - If survey data, link to the questionnaire
  - If administrative data, describe the process to extract data from the system
- Enable re-use by describing what the data items mean
  - Describe any methodology used
  - Define the variables
  - Define any classifications or codes
- Use consistent naming conventions to identify data throughout its life cycle (Raw, Processed, Final)
- Statement about the quality of the data
- List any documentation/metadata standards or formats used
Retention, preservation and disposal
Managing the retention, preservation and disposal of datasets signals to data creators and users that their data is being looked after.

Examples

- Identify relevant datasets and their location, including long-term location
- Determine how long data will be kept?
- Decide how data will be stored for long-term preservation
- Schedule reviews in data life cycle
- Determine timing of archiving and disposal activities
- Ensure the final data is read only, if it is in an archive or a data store. The final version of the data needs to be separate from the raw and processed data.

Data Management Plan
A Data Management Plan can be created in Excel or Word. Alternatively, there are templates on the web and software programs that can be used.

Although it is research data focussed, this is a good guideline for data management: https://www.cessda.eu/Research-Infrastructure/Training/Expert-Tour-Guide-on-Data-Management

Ensure that the Data Management Plan is easy to create and maintain. It should be a one-stop-shop which links out to relevant documents/shared drives where necessary.