Keeping your data in good shape

For statistical data to achieve maximum usage and accuracy, users need to know what was being measured, how it was measured, and how effectively it was measured. Metadata provides this information and so is vital to effective data management.

About metadata

Metadata is simply information about data. It defines and describes data so we can find it, understand it, use it, and share it. Without it we can't use data to shed light on economic, social, and environmental issues, reach better conclusions and make evidence-based decisions. Without it, statistical data is just numbers.

Metadata comes in many forms, and includes things as diverse as tags for web pages, survey questions, table structures and data quality statements.

Metadata is not a new subject. Librarians have been managing metadata about books for hundreds of years. More recently, changing needs and expectations have made new forms of metadata management relevant across governments, economies, and society. This is due to the continuing "explosion" in information which is dramatically increasing the volume of data, and to an increasing expectation that people should be able to:

- find and access data electronically, and interact with it (such as in a spreadsheet) rather than use it in a static form (such as a printed page), and
- draw together data from different sources rather than considering data from each source in isolation

To use and share data, we need to improve, manage and share the metadata associated with it.

Why should we manage metadata?

Managing metadata ensures better business outcomes for both producers and users of data.

Without metadata, or with metadata that is poorly described or inaccessible, it is difficult to:

- coordinate statistical processes for responsive outcomes
- determine what is being measured, its quality and comparability
- analyse data and draw conclusions
- match data and metadata across sources

However, good metadata practice underpins:

- the coordination and responsiveness of statistical processes
- the maximum exploitation of data across users
- data exchange between organisations

Types of metadata

There are so many different types of metadata it can be useful to group similar types together. Unfortunately, there are also many different perspectives on this. Statisticians, geographers, librarians, web publishers, records managers and other communities of interest, all group metadata in different but related ways.

A subject-independent way of grouping metadata is to consider the aspect of the data that the metadata is describing:

- **Conceptual metadata** – *what are we measuring?* Examples include data item labels, definitions, and value ranges, as well as classifications
- **Operational metadata** – *how did we measure it?* Examples include question wording, coding indexes and derivation definitions

- **Quality metadata** – *how well did we measure it?* Examples include response rates and standard errors for sample data, or known reporting issues for administrative data
- **Structural metadata** – *where is it* and *what forms does it take?* Examples include dataset file structures, relationships between record types, and dataset creation programs
- **Administrative metadata** – *who created the data, who can access it and when?* Examples include creation dates and access control lists

The way any given type of metadata is used varies between data producers and data users. For example, data producers use conceptual metadata to describe their data, while data users use it to find data or compare it to other data.

Principles for managing metadata

Some simple principles can guide metadata management to ensure it delivers benefits to both producers and users:

Transparency

Metadata management relies on recording metadata and making it explicit, so it can be used to manage, find, understand, use and share data. You also need to make your general approach and standards clear so stakeholders can understand and use them. Some tips for optimal transparency are:

Develop and communicate an underlying framework

Develop a framework showing the structure that underpins your metadata, and share it with others. Show how metadata flows with your statistical or business processes.

Why? Communicating this model within your organisation or externally promotes understanding and consistency. It can inform both human interpretation and computer-based processes.

Use data and metadata standards

Use agreed authoritative standards. If these do not exist or are unsuitable, then develop and use local standards.

Why? Standards are an agreed ways of doing things. Using authoritative standards, which have been agreed across jurisdictions, encourages real interaction and comparability between jurisdictions and agencies.

There are standards that relate to the structure and content of both data and metadata:

- **Data standards** include vocabularies and classifications like the Australia and New Zealand Standard Classification of Occupations (ANZSCO); and
- **Metadata standards** include standards for geographic information, metadata registries, web resources and statistical data exchange.

You may also have *preferred use* or *local* standards, which allow consistent description of data within a jurisdiction. Which standards you use will depend on the context and type of metadata.

Make metadata accessible

Record metadata and make it visible to others in a consistent manner.

Why? To be useful, the metadata must be readily available in form that is meaningful to those using it. As more use is made of metadata, its quality tends to improve because more people have reviewed it, which leads to it becoming even more useful.

Whenever you provide data to someone, give them the metadata too. Data without metadata is just numbers.

Provide information about your metadata

Provide information about your metadata so it is clear how authoritative, final and current it is.

Why? A data producer or user should be able to determine who created a given metadata element, what authority they had, when it was created, its current status (e.g. whether draft, final, or retired) and who is currently responsible for it.

Ideally this information should be recorded in a metadata registry. Registering metadata emphasises the authority of the metadata. It promotes use of standard metadata, the re-use of metadata and saves confusion.

Automation

Metadata should be seen as an integral and active part of any business process not just passive documentation. Embed metadata into your systems or business processes to ensure consistency, accuracy, coverage and speed; and use it to drive processes wherever possible. Some key tips to achieve automation are:

Capture metadata at source

Capture or develop metadata at the time you are undertaking the process which creates it, rather than re-creating it later.

Why? The more time that passes, the greater the chance the metadata either won't be recorded or will be recorded inaccurately.

Produce metadata automatically

Generate metadata as a by product of automated business systems to ensure it is accurate and available for all possible uses of the data. Remember to also automate production of information about the metadata

Why? Data is not always used immediately. If metadata is not generated when data is produced, the data may be of little value to those using it later when no one remembers what it is or how it is structured.

Capturing metadata automatically minimises human effort, promoting accuracy and completeness because metadata is always created whenever a process occurs.

Control automation with metadata

Some metadata is just passive documentation. Active metadata can be used to control automated processes. Make metadata an active part of the business process.

Why? Making metadata active whenever possible will support more consistent processes as well as ensuring the metadata itself is current and consistent.

Uniqueness

Control the creation and documentation of metadata so your current active metadata items are obvious and unique. Avoid proliferation – have as few items as possible that can do the job. And store the key metadata in one place! Having a range of different places where the same item is stored can create confusion and discrepancies. Some key tips to achieve uniqueness are:

Keep each metadata item in only one place

There should only be one place for storing each type of metadata. Only create and update metadata in this place.

Why? Multiples stores for metadata create confusion, discrepancy and complexity. Single metadata stores promote re-use and consistency

Re-use existing metadata

Don't create new metadata items until you've checked to see whether an appropriate metadata item already exists. Re-use metadata if it is fit for purpose.

Why? Different but similar metadata creates confusion and complexity. Re-use promotes efficiency and consistency.

Re-purpose existing metadata

Use existing metadata for as many purposes as practical rather than defining something slightly different for differing purposes. For example, metadata produced for end users should be consistent with metadata used throughout business processes.

Why? Re-purposing metadata avoids apparent inconsistency in metadata as well as inefficiencies in its creation. It supports the re-use of metadata throughout the life of the data, and avoids the need to develop slightly different metadata for reporting or output later.

Manage variations from standards

Use relevant standards where possible. Where they are not used make this obvious to users.

Why? The use of relevant standards enables comparisons across time as well as data sources. Where data is not comparable it is important that users know.

Preservation

Keep records of metadata used in the past, and information about it, to enable historical data analysis. Some key tips to achieve optimal preservation are:

Preserve all versions of your metadata

Systems and process that deal with metadata should preserve the history of each version of your metadata.

Why? Users need to know about the metadata that went with the data at the time it was created, not just the latest version of the metadata. This allows the accurate comparison of data over time even though methods may have changed.

Preserve superseded metadata

Develop processes for dealing with draft, retired or superseded metadata as well as metadata in current use.

Why? Metadata has a lifecycle, from creation to active use and eventual retirement. Maintaining retired metadata allows current metadata to be identified more easily, making it less likely outdated metadata will be used inadvertently. Also preserve drafts that show development stages.

A word of caution... Your metadata strategy must be sustainable. It should create value for data users and be targeted to user needs – no more, no less.