WP. 22
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT (OECD) STATISTICS DIRECTORATE

Joint UNECE/Eurostat/OECD work session on statistical metadata (METIS)
(Luxembourg, 9-11 April 2008)

Topic 2 (iii) Metadata and the statistical cycle and Implementation

CASE STUDY: AUSTRALIAN BUREAU OF STATISTICS

Submitted by Australia[1]

[1] Prepared by Alistair Hamilton (alistair.hamilton@abs.gov.au).
.

METIS COMMON METADATA FRAMEWORK (CMF)
PART C CASE STUDY

AUSTRALIA / AUSTRALIAN BUREAU OF STATISTICS

Revision History		
Date	Section(s) updated	Comment
11/03/2008		Version 1

Organization Details

Organization Name	Australian Bureau of Statistics
Number of staff	2 925
Contact person (for Metadata)	Alistair Hamilton Director - Data Management alistair.hamilton@abs.gov.au + 61 2 6252 5416

1. INTRODUCTION

1.1 Metadata strategy

The document entitled *A Brief History of Metadata (in the ABS)* at Appendix 1 (referenced simply as *BHM* hereafter) provides information on the evolution of ABS strategies related to metadata over time.

As described in *BHM*, the ABS Metadata Strategy has evolved rapidly over the past three decades. It was formalised in an 18 month process, involving stakeholder consultation across the ABS, which culminated at the end of 2003 with the *Strategy for End-to-End Management of ABS Metadata* being reviewed, and broadly endorsed, by the ABS Executive. More information on the details of this strategy is provided below.

While the formal strategy document from 2003 hasn't yet been updated, the actual strategy employed by the ABS has evolved considerably over the past four and a half years. The 2003 document was a milestone in the evolution of corporate thinking and planning in regard to metadata management, but to some extent it only represents a "snapshot" of thinking at a particular point of time.

As described in *BHM*, the ABS has now embarked on the path towards another milestone in terms of a *2020 Vision.* This is expected to be drawn together during the second half of 2008 and then provide a platform for strategic planning over coming years.

Ultimately any ABS metadata strategy exists to support the ABS mission and objectives as set out in the organisation's corporate plan.

http://www.abs.gov.au/websitedbs/d3310114.nsf/51c9a3d36edfd0dfca256acb00118404/b1042c4ee5af9c71ca256a46008278d9!OpenDocument

In particular, the availability of appropriate metadata and the application of sound metadata management practices are critical to supporting informed use of statistics and the quality of the statistical services we deliver to the nation.

The twelve principles defined as a cornerstone of the 2003 strategy continue to be applied within the ABS

1. Manage metadata with a life-cycle focus
2. All data is well supported by accessible metadata that is of appropriate quality
3. Ensure that metadata is readily available and useable in the context of client's information need (whether client is internal or external)
4. Single, authoritative source ('registration authority') for each metadata element
5. Registration process (workflow) associated with each metadata element, so that there is a clear identification of ownership, approval status, date of operation etc.
6. Describe metadata flow with the statistical and business processes (alongside the data flow and business logic).
7. Reuse metadata where possible for statistical integration as well as efficiency reasons (no new metadata elements are created until the designer/architect has determined that no appropriate element exists and this fact has been agreed by the relevant 'standards area')
8. Capture at source and enter only once, where possible
9. Capture derivable metadata automatically, where possible
10. Cost/benefit mechanism to ensure that the cost to producers of metadata is justified by the benefit to users of metadata
11. Variations from standards are tightly managed/approved, documented and visible
12. Make metadata active to the greatest extent possible

These twelve principles are applied when planning and authorising all ABS projects that provide, and/or make use of, metadata management capabilities, even those where metadata management is a secondary rather than primary objective or requirement.

Other key points in the 2003 strategy include

- There is an agreed conceptual metadata model which is linked to processes that are part of the statistical processing cycle and this linkage is used to determine what metadata should be collected.
- The ABS metadata model takes account of and uses international standards where possible.
- The physical implementation of the metadata model is the Corporate Metadata Repository (CMR) which is used by all ABS projects. It consists of a number of shared physical databases.
- All metadata entities are managed by a 'registration authority'
- Roles and responsibilities are identified
- Data Management and Classifications Branch (DMCB) is responsible for coordination, definition and maintenance of metadata policies, procedure, systems and provides advice and consultancy to developers related to metadata matters.
- DMCB is the 'registration authority' for the CMR and ensures that other organisational units with this role for particular metadata entities understand that role, are trained and have relevant tools.
- Metadata management is part of every project and should be considered alongside resource allocations and accountabilities in the same way as business processes and data flows are considered.
- Governance of metadata management developments and the oversight of outcomes realisation is vested in line management, existing project and program boards with ABS Executive taking an ultimate corporate view.

1.2 Current situation

BHM describes how the current situation has evolved within the ABS.

Basically the majority of data collection and input processing activities for business and household surveys are moving toward implementation of a common high level metadata framework that is informed by ISO/IEC 11179. This framework was developed over the past seven years and postdates the ABS specific metadata framework which was implemented for the corporate output data warehouse which was developed during the 1990s. The creates a looming challenge for end to end metadata management within the ABS. The ABS response to this issue will be thought through further as an outcome of the 2020 Vision process.

Key elements of current metadata infrastructure include major repositories related to

- statistical activities
  - These are termed "collections" by the ABS, where these activities include surveys, censuses, statistical analysis of administrative data sources and statistical "compilation" activities such as preparing the national accounts.
- datasets
  - These are specific structured data files, data cubes and tables associated with statistical activities. Examples include various "unit record files" and aggregate outputs.
- classifications
  - This is a "legacy" system based on an ABS specific data model.
- data elements
  - This is a recent development based on the metamodel found in ISO/IEC 11179 Part 3.
- questions and question modules
  - This was developed recently for household surveys with an aim to generalise the facility in future.
- collection instruments
  - This was developed recently for household surveys with an aim to generalise the facility in future.

The more recent developments also incorporate an approach to metadata registration based on ISO/IEC 11179 Part 6. Even if some of the older repositories cannot be completely replaced in the next few years it is hoped that a common high level metadata registration framework can be implemented across the ABS for all classes of metadata. (This does not imply that all classes of metadata undergo exactly the same registration processes, but that the processes for each class of metadata are consistent with a higher level "metamodel" for registration.)

Interoperability of the current ABS metadata model, and the legacy "output" model, with third party software (eg SAS, Blaise, SuperCROSS) continues to be an issue.

A major emerging focus for the ABS is support for SDMX V2 including its extended metadata capabilities and some of the new "packages" that were not present in V1. Using SDMX V2 based structures as a common reference point is seen as a possible means of bridging differences in metadata models without "bridging" directly from structure to structure such that any change to either end (eg modernisation of one of the structures) requires a whole new bridge.

The ABS is also supporting establishment of a National Data Network (NDN) for sharing data and services from multiple content providers within Australia

http://www.nationaldatanetwork.org/ndn/ndnhome.nsf/Home/Home

This is one factor which requires development of metadata models and capabilities which are usable beyond the ABS. The NDN needs to interoperate with agencies whose data content is more "administrative", "geospatial" or "research oriented" than "statistically" oriented. This raises interesting questions about metadata modelling.

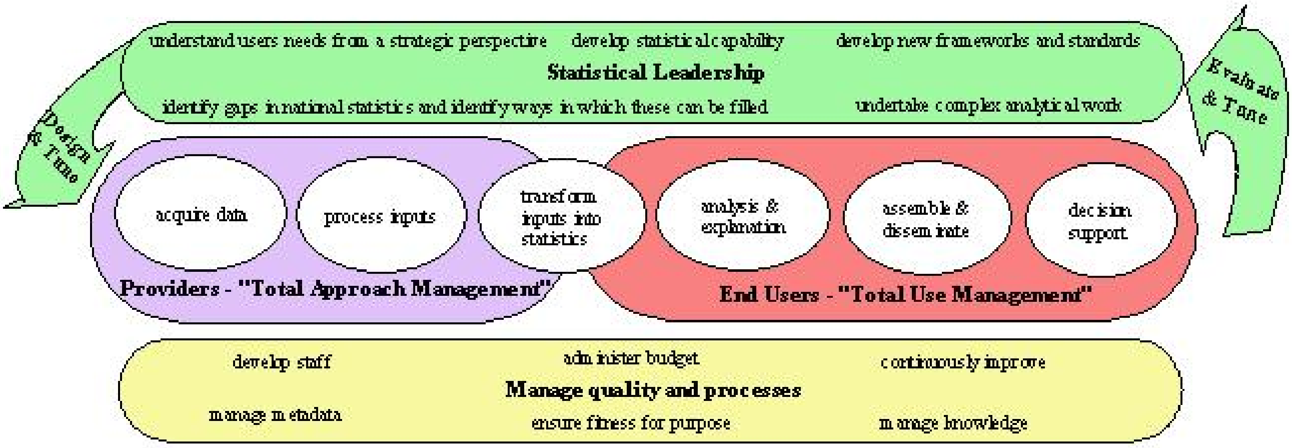
While many of those agencies are at least as passionate about metadata as the ABS - but from a different "school" - the NDN also needs to support content producers and users for whom metadata is much less of an interest and priority. This raises interesting questions about minimum metadata content and quality standards.

These challenges, and others, associated with the current situation will shape the upcoming 2020 Vision process.

2. STATISTICAL METADATA SYSTEMS AND THE STATISTICAL BUSINESS PROCESS

2.1 Statistical business process

While a few areas insist on their own variations on the following theme, the following diagram is affectionately known as "The Caterpillar" within the ABS.


understand users needs from a strategic perspective
develop statistical capability
develop new frameworks and standards
Statistical Leadership
identify gaps in national statistics and identify ways in which these can be filled
undertake complex analytical work
Design & Tune
Evaluate & Tune
acquire data
process inputs
transform inputs into statistics
analysis & explanation
assemble & disseminate
decision support
Providers - "Total Approach Management"
End Users - "Total Use Management"
develop staff
administer budget
continuously improve
Manage quality and processes
manage metadata
ensure fitness for purpose
manage knowledge

A strength of The Caterpillar is that it highlights the activities which take place throughout the cycle including six main steps within in the body of the caterpillar, and the "linking" steps at the beginning and end which open and close the cycle.

The Caterpillar was developed as part of the Business Statistics Innovation Program (BSIP) launched at the dawn of the new century as described in BHM.

It allowed a disparate range of surveys and other statistical activities whose processes were (especially prior to BSIP) very different in detail to describe what they did, why and how (eg what systems and data stores were used) in terms of a common high level reference point for the statistical life cycle. It later allowed "leading practice" to be identified in different parts of the statistical cycle. (Due to legitimate differences there is usually not just one single practice that is best for every survey. Typically a limited set of leading practice models are identified for each step in the cycle from which one can be selected depending on the specific needs and nature of the survey. This is far preferable to 50 different surveys choosing 50 completely different practices for each step.)

In terms of numbering from the CMF model

1. Needs emerge from the **Statistical Leadership** aspect of the caterpillar (possibly having been fed back from **Evaluate and Tune**). Needs can be clarified in detail during the **Design and Tune** phase (eg consulting with an external stakeholder group on design issues).
2. Develop and design very much equates to the **Design and Tune** arrow at the start of the caterpillar.
3. Build is also largely in the **Design and Tune** arrow. Within the ABS currently, partly due to a lack of metadata driven systems, "design" and "build" often tend to go hand in hand (eg for collection instruments). Also, many business surveys - which informed the original design of the caterpillar - are conducted very regularly so the actual amount of "design and tune" (and build) effort for each cycle of the survey is very limited compared with the amount of work which then occurs in subsequent steps in the caterpillar. Separating out Build from Develop and Design in the CMF model, however, might better reflect where the ABS seeks to go in the future.
4. Collect largely equates to **Acquire Data**.
5. Process largely equates to **Process Inputs** and some aspects of **Transform Inputs Into Statistics** (eg weighting).
6. Analyse in the CMF model includes some aspects of **Transform Inputs Into Statistics** in the ABS model (eg seasonal analysis, macro editing). It also covers **Analysis and Explanation** in the ABS model. The difference in the ABS model is that **Analysis and Explanation** commences once data is "finalised" (unless an anomaly is detected in the phase that causes earlier work to be redone) rather than including finalising some aspects of the data (eg producing seasonally adjusted and trended estimates).
7. Disseminate in the CMF model largely equates to **Assemble and Disseminate** and stretches into **Decision Support** in terms of helping clients make use of the content we have disseminated and answering their questions.
8. Archive in the CMF model is underdone in the Caterpillar. In reality it is covered in a low key manner as longer term aspects of **Assemble and Disseminate** and **Decision Support** as well as part of data management policy under **Manage quality and process** at the bottom of the diagram. In the context of sharing data on a sustainable long term basis within the National Data Network, however, its relevance to **Statistical Leadership** is increasingly being recognised. The explicit reference in the CMF model, therefore, makes sense in the ABS context.
9. Evaluate in the CMF model, in lieu of further detail, is assumed to correspond to the **Evaluate and Tune** arrow in the Caterpillar.

2.2 Current system(s)

There are many systems within the ABS that encompass significant metadata definition and management aspects.

- Some are fully corporate. The main examples of these are described briefly below.
- Some are "shadow systems" which extend corporate systems to supplement the standard content with attributes of local interest.

    - Making the corporate systems more readily "extensible" would help to address this issue, as would an enterprise architecture that makes it easy to marry up "local" low level system/context specific metadata with "corporate" metadata.
    - Some "shadow systems" have been designed and maintained to ensure they can be easily reintegrated with the corporate system in future. Some have not.
- Some are truly "local" systems
    - These exist for a variety of legitimate and not so legitimate reasons.
    - The best of them source relevant content from the Corporate Metadata Repository (CMR) as a properly maintained snapshot but then reformat that content to meet local needs (eg to support systems that cannot "read" the metadata directly and require it to be translated/packaged in a special way).
    - The worst of these update, evolve and create new metadata for local use independently of the CMR.
    - Others deal with classes of metadata (eg methodological parameters to drive specific processes) which are not currently managed within the CMR.

Collection Management System (CMS)

This manages high level information about "statistical activities" ("collections") undertaken by the ABS. These "statistical activities" include surveys, censuses, statistical analysis of administrative data sources and statistical "compilation" activities such as preparing the national accounts.

The basic definition of a "collection" suitable to be registered in CMS involves inputs, processing/transformation and output. Simply collating data from other collections, therefore, results in a new "product" rather than being a new "collection" in its own right.

Each collection may have many instances (cycles) - such as a monthly survey. Information can be recorded at the collection, cycle or an intermediate level called "profile". (One purpose of the "profile" level is to document small to medium "redesigns" and other changes that can occur over time within a collection.)

Many (but not all) end to end processing systems do refer to the Collection ID and Cycle ID based on the registration of the relevant activity to CMS. This provides a good starting point in terms of end to end "metadata glue" and means the corporate registry function of CMS is being used relatively actively.

As a repository for descriptive information about statistical activities undertaken by the ABS, however, it sits to one side of the processes themselves and the content is often of relatively poor quality to start with and then poorly maintained over time. This is despite the fact that managers of these activities are asked to sign off on CMS content. Much of the content visible through CMS, therefore, cannot be relied upon as an accurate, up to date description of activities in the ABS.

A subset of this content is signed off to the ABS website to become visible in the ABS Directory of Statistical Sources.

http://www.abs.gov.au/AUSSTATS/abs@.nsf/viewcontent?readform&view=DOSSbyTopic&Action=expandwithheader&Num=1

This disseminated content does tend to be better (but not perfectly) maintained.

CMS also hosts "Quality Declarations" that have started being disseminated alongside ABS data in recent months. For example, see

http://www.abs.gov.au/Ausstats/abs@.nsf/0/74BA4626F8C20DF5CA2573D20018F6F9?Open

Document

The basic design of the CMS dates back to the 1990s although it was updated to Version 5 in 2001. It's structure for describing activities doesn't correspond to the ABS Caterpillar that was developed subsequently. Also

- more of the information entered in CMS should be actively driving actual business processes rather than being "passive" independent documentation, and
- more of the content visible through CMS should be sourced from other stores of actively used metadata.

A redeveloped CMS might also, for example, be aligned with the top level modules ("Group", "Study Unit", "Data Collection") associated with DDI (Data Documentation Initiative) V3.

Redevelopment of CMS is recognised as a priority, but not imminent.

Dataset Registry

This is a widely, but not universally, used registry for defining "dataset" metadata associated with a specific unit record file, data cube etc. This metadata includes

- the set of individual "data elements" included within the dataset
- where the data is stored and how it is structured (eg field names)
- what business unit owns the dataset, when it was last updated etc
- what statistical activity (collection) produced the data

This catalogues all available "output" datasets within the ABS and assists in their management including long term retention.

Some systems working with data in specific environments have their own dataset registries, which includes structuring "dataset" metadata in somewhat different ways. Extending the corporate registry to integrate with the definition and management of "input" and "intermediate" datasets would be of value in an end to end context including being able to trace metadata usage within the ABS. (Querying the metadata model currently allows us to know, for example, which output datasets make use of a particular classification but not which input or intermediate datasets might do likewise.)

The main corporate register dates back to the 1990s and the characteristics of "data elements" recognised within its model are not fully harmonised with ISO/IEC 11179 although the differences are not monumental. This is another driver for updating the model underpinning the registry, in addition to the need to extend that model to better support definition and management of input and intermediate datasets.

While extending and updating the register is desirable it is not imminent. The issue may be "forced", however, when the ABS starts trying to "join up" the IDW and ISHS based data collection and input processing developed during recent years with output processes operating in an environment that currently dates to the 1990s. (See *BHM* for more details.)

Classification Management System (ClaMS)

This is another system that largely dates back to the 1990s. It features a "pre Neuchatel" ABS developed model for classifications. As infrastructure it is used relatively widely (although not universally) in end to end statistical processes within the ABS. For example, in addition to being used universally as part of the defining metadata for output datasets, these classifications can be linked into metadata definition for

- the Input Data Warehouse

- processing of Household Surveys
- driving aggregation, estimation and consequential confidentialisation processes
- driving the layout of publication tables
  - eg indenting labels according to the depth of the classification item in the classification hierarchy
- labelling and describing time series
  - eg based on the classification item labels associated with each dimension of the "key" for that particular time series.

While quite useful for many systematic purposes, the current system is very weak in terms of enforcing rational reuse of classifications across the ABS. For example, while a business area might define their own version of a classification and use that version more or less on an end to end basis, they are unlikely to reuse a classification defined by another area. This is because

- it is relatively hard to find existing classifications that would be structurally suitable to be reused for the area's purpose(s)
- it is relatively easy for areas to define new classifications that meet their required specifications
- areas like to exercise full control over "their" classifications rather than being dependent on other management processes

In addition, ClaMS does not properly support the following

- detailed definitions (as opposed to labels) for individual classification items
- item by item mappings from one version of a classification to another version of the same classification
- item by item mappings from one classification to another
- "special" concepts such as "cut off values" used to translate continuous variables to categorical codes

At the same time, however, the levels of sophistication and complexity of classifications which can be supported within ClaMS can make it "overpowering" for users who have very simple and basic requirements.

It should be noted, also, that ClaMS is currently sometimes used for defining lists (eg of valid values) rather than only "proper" classifications.

Redevelopment of ClaMS is recognised as a priority, but not imminent.

Data Element Registry (DER)

This is a newly developed ISO/IEC 11179 based facility which replaces a number of older "Data Item" systems.

It has been developed using a "services architecture". At the core is a repository of data elements and their building blocks (eg object classes, properties, value domains etc). There are then low level Create, Read, Update, Delete services which are in turn called by a higher level "business based" service layer. A generic user interface is supplied for the DER but it is expected that most users will be interacting with the DER as part of more general "business workflow level" metadata assembly (including reuse) tools that will work with data elements in combination with questions, question modules, collection instruments etc rather than in isolation.

The first main "take up" of the DER will be via the Questionnaire Development Tool (QDT) developed as part of the ISHS project. (See BHM for more information). The second main "take up" is expected to relate to the Input Data Warehouse associated with business statistics. This means that the first uses of DER will be at the "input" end of the statistical cycle, but full end to end utilisation, including support for dissemination requirements, is expected in future.

In addition to the "data element" repository component based on the ISO/IEC 11179 Part 3 metamodel, the DER currently comprises a more general "metadata registration" component based on ISO/IEC 11179 Part 6. The latter has been designed to be able to be separated out as register and set of services in its own right which could support registration and management of metadata "objects" that are outside the Part 3 metamodel (eg questions, "collections", "collection instruments", datasets). This separation is likely to occur (at a logical level, if not a physical level) to support the rolling out of a common high level framework for metadata registration across the ABS.

Questions, Question Modules, Collection Instruments

The ISHS project for household surveys has developed new metadata repositories and associated services related to the above, as well as making use of the new corporate Data Element Registry and the existing Collection Management System.

While the actual development work on these repositories and services to date has concentrated on household survey requirements, the high level design and IT architecture has been selected with an expectation that these repositories will be generalised and "corporatised" in future even if the higher level business services and workflow interfaces developed as part of ISHS, which currently interact with these repositories, remain specific to household survey processes.

Analysis to date suggests that some extensions to the repositories and services will be required to support business statistics and other corporate uses but this should not impact existing use by household surveys.

The infrastructure developed by ISHS is only now in the process of being "commissioned" for actual use by household surveys so it is possible there will be some further refinement to the repositories and services for that purpose prior to any thought of wider "corporatisation".

The initial use of these repositories and services will focus on survey development and input processing but full end to end utilisation, including support for dissemination requirements, is expected in future - first by household survey processes and then more generally.

Quality Infrastructure System (QIS) and Business Activity Monitoring (BAM)

Both of these systems, recently released to production, store metrics on how statistical processes are performing (eg response rates, imputation rates, edit rates etc) and support reporting and analysis related to these metrics. This data about the outcomes of processes can be termed "operational metadata" or "paradata" within the ABS. It can be useful for internal monitoring, management and tuning of processes as well as generating data quality indicators for external dissemination.

These systems rely on individual processes being "instrumented" to write relevant metrics to the QIS or BAM store. (QIS is informed more by the IDW "business statistics" data model and BAM by the household surveys approach.) This allows for progressive uptake.

At the moment the metrics recorded in QIS and BAM tend to relate to early stages in the statistical cycle but both are designed to be able to accept metrics from later in the cycle.

Process Metadata

Some early conceptual and exploratory work has been done in this area but no major design work. Seven types of "process metadata" were identified in this early work, from "configuration" metadata about the IT environment and the user running the process, through to metadata which is a formal "input" to, or "output from" the process through to metadata which describes the process itself and which describes how chains of processes fit together.

The simple SDMX package related to process definition has also been considered.

Achieving a clearer path forward in regard to structuring and managing "process" metadata is seen as an important enabler to having other metadata (eg the structural definition of data elements) actively drive statistical processes.

2.3 Costs and Benefits

Section 2.2 details infrastructure delivered as the result of diverse projects, some of which first delivered outputs more than a decade ago. Lifecycle costs and benefits are extremely difficult to even estimate meaningfully.

Costs and benefits for new developments and redevelopments are usually estimated when developing business cases. While much better than a vacuum for planning purposes, past experience suggests these cost benefit analyses are usually not borne out in practice. Often this is because decisions are made over time to diverge from the original project plan in some way rather than just because the original estimation process was flawed or based on imperfect information.

None of the major developments are currently at the "business case" stage - they are either not yet at that stage or long past it - so current "business case" estimates are not available.

2.4 Implementation strategy

This question can be viewed from several perspectives. At least in terms of metadata management, the swinging of a pendulum can be seen to some extent in the *BHM*. Developments in the 1990s tended to be on a "big bang" basis.

These were sometimes pejoratively referred to as "Cathedral Projects" for being too grand in ambition and design, and for taking much longer and much more money to complete than originally expected. Nevertheless, many of the results of these projects have proved to be of enduring value - so much so that many outputs have lived on long beyond their prime.

The strategy next became "opportunistic" and "incremental". There was a broad "master plan" of what should exist in the longer term, but individual "construction projects" were much more modest in scale.

The 2020 vision process on which the ABS has now embarked may move the balance back toward the centre.

At another level, a consistent learning has been that a well developed and managed implementation strategy (in addition to a development strategy) is essential. New capabilities are being delivered into a complex context of existing processes and infrastructure. Uptake of those new capabilities needs to be managed and promoted appropriately. (The simple "Field of Dreams" approach of "Build it and they will come!" has never yet worked for us.) Often the new capability and/or the implementation and communication strategy for it, needs to be refined based on early uptake experience. Whether it is managed by the development team or some other team, every major project requires a well planned and actively managed "Outcome Realisation" phase after it has finished delivering its major outputs.

3. STATISTICAL METADATA IN EACH PHASE OF THE STATISTICAL BUSINESS PROCESS

3.1 Metadata Classification

The ABS doesn't have a formal "taxonomy" of metadata. One was proposed early in development of the 2003 metadata strategy but it wasn't included in the final document. It was found that discussions about how to "class" particular instances of metadata (in borderline cases rather than all cases) could become very protracted without that discussion seeming to generate any real value.

The primary categorisation in use now relates to purpose/use of metadata. This means a particular "piece" of metadata may (and often should) support more than one type of use. The categories are

1. (Search and) Discovery - Help users find data (or a metadata object in its own right, such as a classification) of relevance to their needs and interests
2. Definition - Help users understand data (or a metadata object in its own right, such as the definition of a data element)
3. Quality - Help uses assess the fitness of associated data for their specific purpose
4. Process - Apply metadata to run processes, such as using a classification to drive an aggregation process or to provide a list of valid encoding values for editing purposes. It also includes defining other parameters that drive a process as metadata, such as the choice of which imputation method to use for which data element.
5. Operational - These are metrics on the results of the operation of processes such as edit rates, imputation rates etc. These can feed into internal decisions on managing and improving survey processes and into external "quality" decisions. This metadata is sometimes termed "paradata".
6. System - Low level information about files, servers etc that helps allow the physical IT environment to be updated without end user processes needing to be respecified.

The ABS also recognises "objects" in regard to which metadata can be assembled and registered. These include

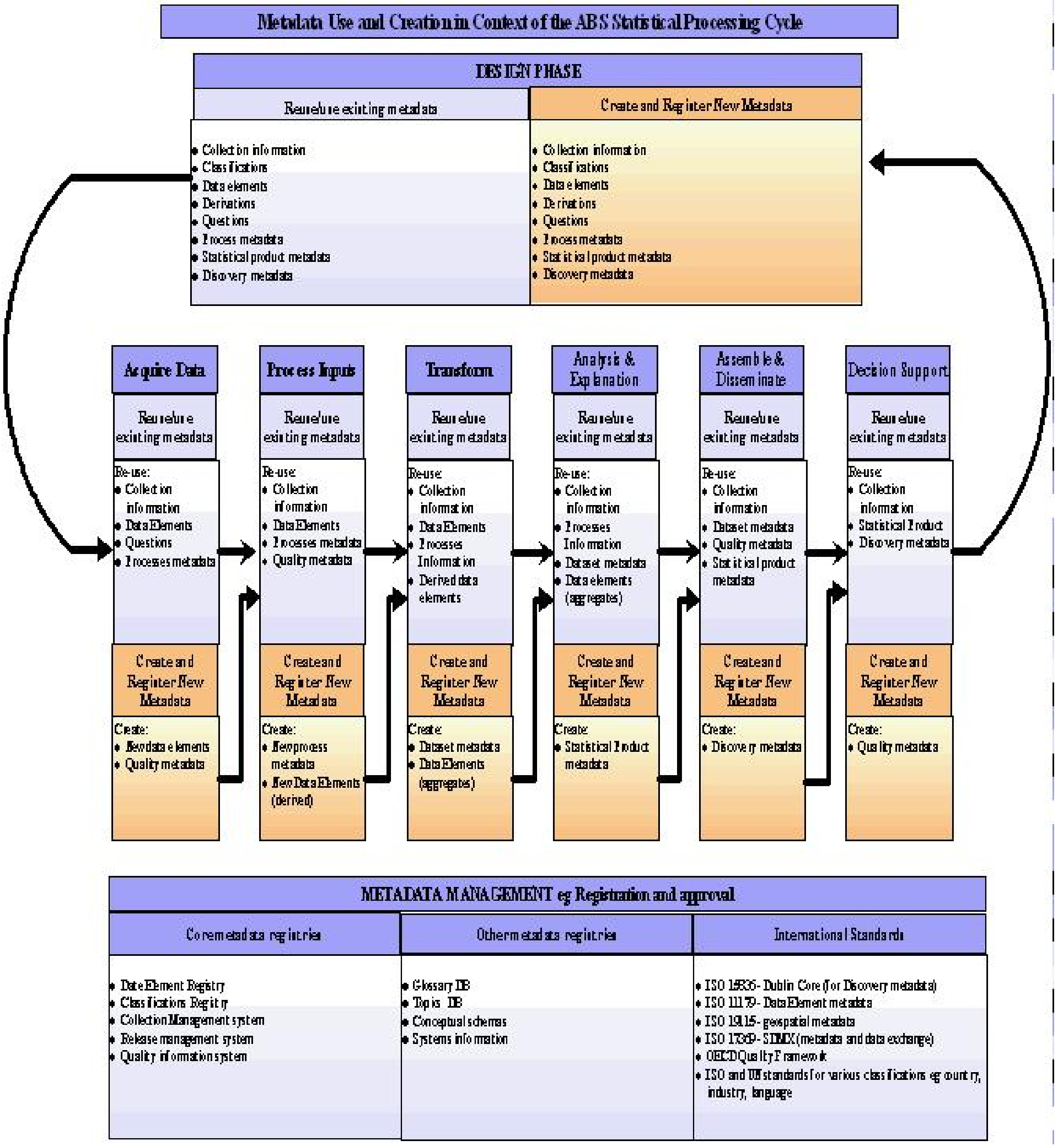
- high level end to end statistical activities ("collections")
- individual datasets
- data elements
- classifications
- individual processes
- terms
- questions
- question modules
- collection instruments

These "objects" can be further broken down (eg data elements into properties, object classes, value domains etc). While the ABS could establish a list of all the high level metadata objects we currently recognise, we wouldn't necessarily recognise a particular list as containing all of, and only, the high level objects that ever should be recognised by any statistical agency.

3.2 Metadata used/created at each phase

The ABS is an agency that aspires to achieve end to end definition, management and reuse of metadata. Section 2.2 records the extent to which we have achieved this so far in regard to our major corporate metadata systems.

While indicative rather than exhaustive, the following diagram sets out ABS aspirations in this regard as captured in a briefing paper from 2006.


Metadata Use and Creation in Context of the ABS Statistical Processing Cycle
DESIGN PHASE
Reuse/use existing metadata
• Collection information
• Classifications
• Data elements
• Derivations
• Questions
• Process metadata
• Statistical product metadata
• Discovery metadata
Create and Register New Metadata
• Collection information
• Classifications
• Data elements
• Derivations
• Questions
• Process metadata
• Statistical product metadata
• Discovery metadata
Acquire Data
Reuse/use existing metadata
Re-use:
• Collection information
• Data Elements
• Questions
• Processes metadata
Create and Register New Metadata
Create:
• New data elements
• Quality metadata
Process Inputs
Reuse/use existing metadata
Re-use:
• Collection information
• Data Elements
• Processes metadata
• Quality metadata
Create and Register New Metadata
Create:
• New process metadata
• New Data Elements (derived)
Transform
Reuse/use existing metadata
Re-use:
• Collection information
• Data Elements
• Processes Information
• Derived data elements
Create and Register New Metadata
Create:
• Dataset metadata
• Data Elements (aggregates)
Analysis & Explanation
Reuse/use existing metadata
Re-use:
• Collection information
• Processes Information
• Dataset metadata
• Data elements (aggregates)
Create and Register New Metadata
Create:
• Statistical Product metadata
Assemble & Disseminate
Reuse/use existing metadata
Re-use:
• Collection information
• Dataset metadata
• Quality metadata
• Statistical product metadata
Create and Register New Metadata
Create:
• Discovery metadata
Decision Support
Reuse/use existing metadata
Re-use:
• Collection information
• Statistical Product
• Discovery metadata
Create and Register New Metadata
Create:
• Quality metadata
METADATA MANAGEMENT eg Registration and approval
Core metadata registries
• Date Element Registry
• Classifications Registry
• Collection Management system
• Release management system
• Quality information system
Other metadata registries
• Glossary DB
• Topics DB
• Conceptual schemas
• Systems information
International Standards
• ISO 15836 - Dublin Core (for Discovery metadata)
• ISO 11179 - Data Element metadata
• ISO 19115 - geospatial metadata
• ISO 17369 - SDMX (metadata and data exchange)
• OECD Quality Framework
• ISO and UN standards for various classifications eg country, industry, language

3.3 Metadata relevant to other business processes

The 2003 metadata strategy defined its scope as relating to "statistical" metadata (rather than all the metadata potentially relevant to any aspect of ABS operations). The scope was still broad, however, because some of the metadata required in order to perform core statistical operations may not be thought of as "statistical" in nature.

Briefly exploring some of the borderline cases, the operational metadata (paradata) about statistical processes can be (and is) used for making financial planning and prioritisation decisions. For example, the financial implications of increasing sample size, increasing the length of questionnaires, accepting reduced response rates, raising the threshold for "significance" editing etc can all be gauged better, together with the likely statistical benefits/costs. This can help set priorities for expenditure, or for areas where savings can be reaped.

On the other side, "administrative" information sourced from the ABS "Corporate Directory" about individual staff members, individual positions (which might be temporarily occupied by one person while another is absent), business units, corporately defined "roles" etc is used extensively by statistical systems - including metadata systems. This may be used, for example, to determine who is currently in the set of people who have the right to edit, approve or otherwise manage a particular piece of metadata.

An intersection is early work on a proposed Statistical Content Ownership Framework (ie ownership of data and metadata). This recognises that organisational units change over time, so assigning ownership of individual content to a particular business unit can create maintenance headaches and/or responsibility headaches over time. The idea is to assign ownership/custody for particular data and metadata holdings to particular subject matter based "domains", possibly together with some additional specialised "methodological" domains related to particular concepts, methods and other artefacts. We anticipate these domains should be more stable and enduring. We would then map these domains to the current organisational structure.

While it is not currently the case, it is possible this could in turn feed into non "statistical" activities such as cost recovering the space used to store data in a particular system related to a particular domain.

4. SYSTEMS AND DESIGN ISSUES

4.1 IT Architecture

Unless otherwise noted, this section refers back to the main metadata systems as described in Section 2.2.

The newer metadata facilities are based on a Service Oriented Architecture. The older facilities tend to have monolithic coupling of the repository, the business logic and business rules (which are built into the application rather than embedded in services) and the User Interface.

Nevertheless, selected information about the collections defined in CMS is "projected" from CMS into an Oracle database. While only a small subset of the total information held in CMS, this comprises all of the core "structural" registration details about collections, cycles and profiles. Basic (read only) "collection metadata services" based on this content on Oracle are then provided for statistical processing applications to access.

A similar approach applies in the case of classifications except a much greater percentage of the total information held in regard to classifications is both "structural" and available on Oracle.

Apart from CMS and ClaMS (which include some descriptive content held only in IBM's Lotus Notes product) the other metadata holdings are all based in Oracle. There is extensive use of Oracle Stored Procedures for reusable services/functions and some use of true web services.

4.2 Metadata Management Tools

Statistical processing applications interact with metadata via services where possible although, as described in BHM, many ABS processing applications and third party vendor products are not yet amenable to this approach. Where this approach is used currently it typically involves the application "reading" relevant content from the metadata repository rather than writing back new or updated records.

As noted in 1.2, it is hoped that a simple standard reference model and set of supporting tools (eg based on SDMX V2) might assist in this regard in future.

In the meantime, as described in the introduction to 2.2, there are cases where metadata from the Corporate Metadata Repository needs to be restructured and/or repackaged relatively manually to make it suitable for use in particular processing systems.

4.3 Standards and formats

The standards and formats currently in use for the major metadata repositories, together with those we hope to use in future, are described in Section 2.2.

4.4 Version control and revisions

This tends to be a major point of debate within the ABS. As the systems have grown up at different times, their approach to version control tends to differ. The most recent major debate has been in regard to the new Data Element Registry.

In general we are now favouring the general approach to versioning set out in ISO/IEC 11179 Part 6. That standard, however, still leaves a lot of flexibility available to the relevant Registration Authority for a particular registry in terms of how versioning will be applied.

In general, where there is not a compelling case for supporting formal versioning then that complexity is avoided. Collections, for example, are not currently versioned. Many aspects of change over time for a collection, however, can be handled through descriptions of the "cycle" or the "profile" rather than edits to the main collection document itself.

The current classification system doesn't handle versioning well and could benefit from the Neuchatel approach. Currently each registered object is essentially an independent entity (ie a "new classification"). It is possible to designate one classification as being "based on" another but this can mean many different things

- The new classification is a new version of the earlier classification and is in some sense expected to supersede it (although possibly not immediately).
- The new classification is a "variant" of the earlier classification defined for a specific purpose. The earlier classification may "live on" indefinitely for the original purpose.
- Classifications are being "grouped" into a "family" without necessarily being formal variants or versions of each other.

Where versioning does need to be supported, careful attention needs to be given to defining cases that don't result in new versions ("trivial changes") and cases that must result in whole new objects (ie the change is so fundamental the new object is no longer a "version" of the old object).

Where revisions are to be made (or new versions created) as much impact analysis as possible is undertaken. This includes, for example, understanding what other metadata objects and processes refer to the object that is about to be revised (or versioned) and whether the revision will have any inappropriate impact (whether the new version should be referenced instead). The lack of fully "joined up" registries (including knowing exactly what metadata is referred to in each processing system) makes impact assessments difficult and only partially reliable in some cases.

The preceding example of impact assessment in the case of versioning illustrates the flow on impacts that versioning can have within a complex and actively used metadata registration system. If the existing metadata objects that refer to the object that just got "versioned" now need to refer to the newer version of that object, then all those existing metadata objects themselves now potentially need to get "versioned" (because they're pointing to a new version of the first object). All the objects that refer to the objects that referred to the original object now need to get impact assessed and potentially versioned themselves, and so on with a ripple effect potentially sweeping across the whole registry originating from just one object being versioned. The ABS hasn't yet resolved this issue.

4.5 Outsourcing versus in-house development

While external expert consultants have been engaged from time to time, the metadata systems described in Section 2.2 were all designed and developed "in-house". Open source and other starting points for the Data Element Registry were seriously considered. It is expected open source and other collaborative options will increasingly be selected in future, although that is different to complete outsourcing. At a minimum, interoperability between new repositories deployed within the ABS and other relevant "external" repositories, standards and vendor software solutions will be an increasingly important consideration.

4.5 Additional materials

None are supplied at this stage but it is likely that additional information can be made available on request.

5. ORGANIZATIONAL AND WORKPLACE CULTURE ISSUES

5.1 Overview of roles and responsibilities

Realisation of the objectives of the 2003 metadata management strategy, and upholding and advancing the principles set out in it, remains a responsibility shared across the ABS.

As upholding and advancing the principles was seen particularly as the responsibility of every new project within the ABS, the project planners, project managers, business analysts and IT staff associated with these projects had a particularly important role. Data Management Section (DMS) developed particular guidelines to assist such key people in understanding the practical meaning and intentions of the principles and how they might apply in the context of a specific project. DMS also provides direct interactive advice to planners, analysts and IT staff.

DMS was also assigned the lead role in terms of co-ordinating the development of specific metadata management infrastructure and ensuring this infrastructure fits together as part of a logically integrated Corporate Metadata Repository. It has the lead role in monitoring overall progress in regard to the strategy and identifying areas where refinement to the strategy, updates to policy and practice or other measures might be required.

Statistical subject matter areas are required to make appropriate use of the available facilities, adhere to the policies and follow the relevant guidelines. In particular, these areas remain responsible for the extent, accuracy and other aspects of the fitness for purpose of the metadata content related to their particular collection, classifications, data elements etc. While DMS ensures the necessary "repository infrastructure" is provided, and that the infrastructure remains "fit for purpose" in a changing organisational and technical environment, DMS does not become responsible for the quality of the content held within each repository.

In addition to documenting their metadata initially, senior subject matter staff became responsible for "signing off" that the documented content was both accurate and sufficient. Subject matter areas also became responsible for ongoing custodianship of that metadata, including ensuring it remains up to date and answering any enquiries its definition might generate from others.

At a higher level a Metadata Strategy Group comprising "Branch Heads" drawn from across the ABS was formed to elaborate upon and drive forward and "champion" the strategy. This group has direct access to the very top management with the ABS and has regularly brought critical issues and proposals before top management for input and funding approval.

5.2 Metadata management team

Data Management Section (DMS) resides within the Data Management and Classifications Branch (DMCB) of the Methodology and Data Management Division (MDMD) of the ABS. DMS consists of around a dozen staff supported by around half a dozen programmers (application developers) from the ABS Technology Services Division. In addition to looking after

- policy and strategy related to metadata
- the work program related to the Corporate Metadata Repository (CMR)
- user support and training related to the CMR

DMS also look after work program, user support and training for the output data warehouse and other aspects of data management policy and practice within the ABS.

The two other sections within DMCB are the standards areas for economic and population statistics, looking after the development, definition and promotion of key content related statistical frameworks, concepts and classifications.

Each of the three subject matter Groups within the ABS (each Group, loosely, consists of two Divisions) includes a "co-ordination section" that assists with

- requirements gathering and prioritisation for new metadata facilities and for improvements to existing ones
- targetting and co-ordinating "user acceptance testing" of, and feedback on, new/changed facilities
- co-ordinating definition of implementation programs for new processes and systems, monitoring progress of implementation and escalating the most common and most serious implementation issues to ensure they are addressed
- aiding communication between DMS and end users (translating terminology and impacts between the two)
- meshing the CMR work program with other work programs relevant to that Group

DMS also works closely with the publishing area to facilitate appropriate content from the CMR flowing through appropriately into publications or directly onto ABS web pages. The level of content flowing through in this way is increasing in terms of its scope, volume and interlinkage. Making metadata available to the public via the web raises a range of additional content, process and management issues for subject matter areas, DMS and Publishing that need to understood and addressed in an appropriate and sustainable manner.

5.3 Training and knowledge management

DMS provides a range of training. This includes an overview of concepts and systems related to metadata management. Such training is regularly made available to new starters within the ABS and other staff.

A Corporate Metadata Repository (CMR) Assistant is available from the home page of the ABS intranet. This provides a portal to overview and detailed information about the available facilities as well as related policies, guidelines and training courses. It also provides direct access to the facilities themselves by allowing users to click on the component of interest as represented in a high level diagram showing how the various facilities fit together.

As the CMR is "part of the way the ABS does business", the generic training offered by DMS is only one strand. The economic and social statistics areas provide training that includes explanations of how the CMR facilities fit within, and are used within, their business processes. The training about dissemination processes in the ABS likewise includes information about how content defined in the CMR can be drawn into the various dissemination channels and made available outside the ABS. DMS provides development assistance and input on the components of these training courses that relate to the CMR.

Similarly the corporate "Assistants" related to Business Statistics, to Household Surveys and to Publishing cross reference relevant content from the CMR Assistant where appropriate.

The strategy of presenting information about the CMR in the context of a particular wider business process, rather than trying to present everything about it exhaustively in a major CMR specific training program, appears to be working very well.

5.4 Partnerships and cooperation

The ABS is very keen to share information and experiences and to collaborate within METIS generally, as well as on a narrower (eg bilateral or "working group") basis.

A second major international opportunity for partnership and cooperation is seen to be around SDMX. ABS has provided extensive feedback on previous proposals and outputs from the consortium, and volunteered to take part in case studies. The consortium seems committed to providing National Statistical Offices with even greater opportunities to shape, rather than just respond to, the initiative in future. The ABS looks forward to that.

The ABS also contributes actively to international committees associated with other metadata standards of relevance to it, such as ISO 11179.

The ABS interest in collaborating on relevant open source software has been noted earlier.

As described in Section 1.2, the National Data Network (NDN) provides many opportunities for various collaborations. Many of these collaborations are within Australia but they also include, eg, collaborating with the US Bureau of the Census in regard to their Data Ferrett product. This included trial mappings of ABS data and metadata content into the schema associated with that tool which highlighted some useful information content which can't currently be accommodated within that schema. The NDN initiative takes the ABS beyond simply collaborating with other statistical agencies and into collaborating with the geospatial community, the research community and others.

5.5 Other issues

Over the past 15 years the term "metadata" has become common parlance within the ABS. The value and importance of metadata is widely recognised.

Some of the practical complexities of managing and actively reusing metadata throughout the statistical cycle are (not yet) so widely and well understood. This means there is a degree of disappointment and frustration expressed in some quarters that more progress hasn't been made more quickly and that we haven't yet made metadata simple to manage and maintain as well as "all powerful" in driving and describing all processes and outputs.

There is also still a tendency for projects to want to structure metadata in an exactly optimal manner for their processes and manage it directly in that form. The services layer "plumbing" to allow such an approach to be implemented efficiently, while drawing content from corporate metadata repositories that are not structured in the same manner, is far from being largely - let alone fully - in place.

6. LESSONS LEARNED

The lessons learned, and conclusions drawn, from various experiences within the ABS so far are mentioned under the headings where those experiences are discussed. These lessons learned, which are currently scattered through the other sections of the case study, will be drawn together and consolidated here as time allows.

Some key points from a separate summary of lessons learned from a year ago include the following

- While technology is a vital enabler, metadata management should be driven, governed and presented as primarily a business issue rather than a technical issue.
- Similarly, all high level organisational units need to be engaged by the metadata management program and have defined responsibilities in relation to it. Some units' primary responsibilities may simply be to contribute to corporate sign off on the objectives, strategies, policies and high level design of deliverables (systems and processes) and then to take up and apply the outputs in an agreed manner to contribute to the achievement of the corporate outcomes sought from the project. Other units, naturally, have a much more extensive role in terms of leadership, co-ordination, business analysis, design, development, implementation and ongoing management of systems and processes. If only a few specific organisational units are seen to have a direct stake in the project then it's much less likely to achieve overall success.
- It's become more and more apparent over time that applying externally recognised and supported standards, in regard to design of data models for example, has a lot of benefits - including as a means of building upon a wealth of intellectual efforts and experiences from others. At the same time, application of standards must be driven, and moderated, by the organisation's particular context and needs. The underlying effectiveness of the infrastructure should not be sacrificed in favour of complying "to the letter" with a standard, although the business case and the management arrangements for any divergence need to be defined and agreed.
- In addition to developing and deploying infrastructure, a metadata management project should be

understood, and managed, as a "cultural change" initiative for an organisation.

- Sufficient attention needs to be focused, by the project team and by other areas, on ensuring the metadata management infrastructure (systems and processes) is fully integrated with other business processes and IT infrastructure rather than being a "stand alone" development.
- In addition to allowing sufficient time and resources for the business analysis, design and development process it is crucial there is sufficient resourcing focused on
    - implementation of the new infrastructure
        - includes training, best practice advice and technical troubleshooting support for business users
    - maintaining and upgrading the infrastructure as business requirements, and as other elements of the IT environment, evolve over time
    - co-ordinating and promoting "outcome realisation" from the investment

An emerging lesson, also, is that while Service Oriented Architecture (SOA) offers a lot of opportunities and potential, it also comes with a lot of new complexities compared with earlier approaches. It requires new understandings and a new mindset from those developers who are being asked to take up, and interact with, the available services as well as requiring the same from the business analysts and programmers within the team responsible for providing the metadata repositories and services. It can make the overall environment much more complicated in some ways (eg services are calling services that call services etc and then somewhere at a low level a service is updated and everything needs to be configured appropriately to allow proper testing of that change). Implementing SOA in environments that include a lot of "legacy" processing systems that are not enabled for the new architectural directions is particularly challenging. A highly successful example of implementing an SOA based metadata management environment would be of very high value as a case study for the ABS.

Appendix 1: A Brief History of Metadata (in the ABS)

Introduction

The following document charts the historical arc of metadata management strategies and developments pursued by the ABS, particularly over the past three decades, culminating in the outlook we now face.

This history is of importance to the present because the current situation for metadata management within the ABS, and many of the challenges to be faced in the future, reflect past strategies.

While this "history" document is relatively long, it is hoped it sets out a relatively coherent context for the current situation, lessons learned and future plans described throughout the case study. Various sections of the case study refer extensively to the historical, and forward looking, perspective on ABS metadata strategy set out below. In this way the extent of context that needs to be reiterated at different places within the case study document itself is reduced significantly.

For fun, rather than representing a rigorous framework of "metadata paleontology", the historical arc described below is broken into a number of eras.

Premetazoic Era (1905-1973)

As with other statistical agencies, ABS processes and outputs involved some degree of "metadata" management even before the term was coined formally. As the term hadn't been invented, however, the ABS didn't yet have a metadata management strategy.

Protometazoic Era (1973-1990)

During the 1980s the ABS (along with many other agencies) undertook a number of major "data dictionary" projects that assembled basic definitional and structural metadata related to the individual "data elements" collected, derived and output by the ABS. Some of these initiatives created data dictionaries that spanned multiple related surveys (eg a range of "business surveys"), and allowed definitions of common data elements to be shared and reused consistently across this set of surveys. None of these initiatives were fully corporate in scope. While these initiatives did engage subject matter statisticians (usually technically oriented ones) they tended to "grow out from" new IT capabilities, rather than the specific IT capabilities being secondary to the metadata strategy and directions.

Mesometazoic Era (1991-2000)

At the start of the 1990s the ABS initiated a major focus on "data warehousing". Rather than supporting a series of different "stove pipe" survey specific output systems, many advantages were identified in establishing a "output data warehouse" as a "single version of the truth" when sourcing output data from surveys for dissemination and for secondary use within the ABS.

In 1991 the ABS was fortunate enough to have Professor Bo Sundgren undertake a five month review which resulted in an excellent paper entitled "Towards a Unified Data and Metadata System at The Australian Bureau of Statistics". This paper envisaged three components for an "ideal" ABSDB, namely "macrodata", "microdata" and "metadata".

From the outset there was a strong focus on the metadata required to support the output data. Major repositories were developed to collect and structure metadata related to the following

- statistical activities
    - These are termed "collections" by the ABS, where these activities include surveys, censuses, statistical analysis of administrative data sources and statistical "compilation" activities such as preparing the national accounts.

- datasets
    - These are specific structured data files, data cubes and tables associated with statistical activities. Examples include various "unit record files" and aggregate outputs.
- data items (data elements/variables)
    - ISO 11179 was just a glint in someone's eye at this stage. Information about "data items" was recorded using an ABS specific data model. Many similar underlying characteristics (eg a distinction between enumerated and non enumerated value domains) are recognised by the two models but the details of modelling are different
- classifications
    - Once again, an ABS specific data model was used.
- terms
    - Captured in the ABS Glossary

A lot of the metadata now being documented to support output data had already been entered in different forms elsewhere in the statistical cycle including

- internal planning and approval documentation, public consultation documentation associated with initiating a statistical activity
- all the individual processing systems associated with a statistical activity
    - either entered as metadata or "hard coded" into each system
- "Concepts, Sources, Methods" and other publications associated with that statistical activity

Assembling this "extra" metadata, which then provided little direct return for subject matter statisticians, was often regarded as an overhead. The quality of the metadata provided initially was often questionable and it wasn't then actively maintained over time.

Nevertheless, this era established a core of metadata in common corporate repositories in accordance with a common (but ABS specific) data model. It provided a platform for all that followed.

Most of the metadata repositories developed during that time are still with the ABS. They have evolved and been extended. For example, most now offer some degree of "services interface" which allows content from the repositories to be called up from within processing applications rather than needing to "jump into" a repository specific application. Nevertheless, apart from the new Data Element Registry, these applications are yet to be completely redeveloped to allow them to adopt new IT architectures and standard metadata models that have evolved since the early 1990s. The legacy from this era is now both a corporate asset and a corporate liability.

Neometazoic Era (2001-2007)

Almost from the outset there had been the notion that the metadata facilities developed during the previous era should, for many reasons, be extended to serve purposes beyond documenting output data. Major action wasn't taken on this front until around the turn of the century, however, by which time the output data warehouse was firmly established and integrated within ABS output processing and dissemination workflows.

Around 2001 the existing metadata facilities were recognised as the foundation of a Corporate Metadata Repository (CMR) which had an identity separate to the output data warehouse itself. (The latter was by now termed the ABS Information Warehouse (ABSIW)). While support of ABSIW metadata requirements remained an important purpose for the CMR, its mission now extended to supporting all aspects of the statistical cycle. (At the instant it was first formally constituted, of course, the CMR did not yet possess the capabilities required to fulfil many aspects of this extended mission.)

Around the same time the ABS commenced a massive re-engineering and consolidation of its approach to managing "input" data, from both survey and administrative sources, related to businesses. (This was the first major step in the Business Statistics Innovation Program (BSIP), a program which continues to the present day under a different name.) Similarly to the case with the ABSDB in the 1990s, metadata was of crucial importance to the emerging Input Data Warehouse (IDW). By this time, however, metadata standards such as ISO/IEC 11179 had become well established and accepted internationally. A defining characteristic of the new era was an emphasis by the ABS on applying standards which had emerged since the 1990s wherever, and to the extent, they could be applied in support of the achievement of ABS statistical and business objectives.

The IDW, and processes associated with that data store, quickly became "the second big target" for support by the CMR.

A couple of years subsequent to initiation of the IDW project, a major project was initiated to redevelop, extend and modernise the existing environment in which household surveys were developed, initiated and processed. Once again metadata was a key consideration for this Integrated Systems for Household Surveys (ISHS) project. Part of the ISHS redevelopment included alignment with ISO/IEC 11179 and with many other aspects of the new metadata framework associated with the IDW.

Where the ABSDB replaced "output" stovepipes in the 1990s, IDW and ISHS greatly reduced the number of separate input "pipes" although they did not result in "just one channel". (A range of statistical activities remain outside the scope of either IDW or ISHS currently.) The reduction in the number of "pipes" to be supported makes it easier to apply the CMR to support, in practice, to end to end statistical activities within the ABS.

The new considerations and directions evident at the dawn of this era provided impetus to the development and formalisation of the Strategy for End-to-End Management of ABS Metadata over 18 months up to November 2003.

The strategy sets out

- a model to work towards in terms of how metadata should be structured and accessed to support "end to end" purposes
- a set of metadata management principles which, if applied consistently to all new systems development work undertaken by the ABS, would lead the organisation toward that integrated metadata management environment
- processes to facilitate the application of those metadata management principles to future developments undertaken by the ABS

Major developments such as IDW and ISHS subsequently did assist in advancing these principles, and the ABS metadata management environment more generally, although not to the extent - and not with the level of coherence - originally hoped.

The ABS Data Element Registry (DER) was developed during this era. It is integrated with relevant CMR components that predated it, such as those related to classifications and to statistical activities. While its long term objective is to support definition, management and reuse of data elements through all stages of the statistical cycle, the first target was support for IDW and ISHS related requirements for data element metadata.

The ISHS project included the development of new repositories and services related to questions, question modules and collection instruments. These were designed to integrate with the metadata related to data elements from the DER. While the repositories related to questions and collection instruments have been built for ISHS specifically at this time, they were designed in consultation with "architects" for the CMR and IDW with the intention that at some stage these repositories can be enhanced and extended to meet broader corporate metadata requirements related to questions and collection instruments as an integrated part of the CMR.

The 2003 strategy of progressing ABS metadata management capabilities by "piggybacking" on other projects was at first characterised as "opportunistic".

The ABS subsequently concluded that, at a minimum, a planned "incremental" approach was required. As initial design and development of the DER neared completion, proposed programs of subsequent work related to the CMR were set out during 2006 and 2007 that stretched out across many phases over several years. These comprised a mixture of developing new capabilities and redeveloping and extending the older corporate metadata systems to integrate better with

- each other
- modern systems architecture, and
- international metadata standards.

These proposed work programs were generally agreed to be "worthy" by the ABS Executive but there were debates about priorities and interdependencies between certain developments.

Holocentric Epoch (2008)+

The strategy document from 2003, and subsequent papers, contain extensive discussion of objectives, principles, business drivers, benefits, models, proposed work programs etc related to metadata management. Earlier this year, however, the ABS executive noted these documents do not provide a clear and compelling "picture" of how the organisation aspires to be in the longer term. For example, if the ABS did achieve the end to end metadata model set out in the 2003 strategy what would it mean in practice in terms of changing and improving the way the organisation operates? What new capabilities would it deliver and are they the capabilities we need most? How certain is the ABS that "the future" the 2003 strategy targets is both achievable and the "most appropriate" future for us to be investing in working towards?

The next major step over the next twelve months is likely to be development of a "2020 Vision" encapsulating longer term ABS aspirations. Having clearly defined the state we aspire to reach longer term, the most appropriate strategy for moving forward can be determined. Not only may the target change fundamentally, but the preferred method of achieving it may (or may not) change fundamentally from the current "incremental" approach. (That said, the organisational and project management challenges and risks attendant with trying to successfully manage a "big bang" approach are also well recognised.)

Two important considerations have emerged in recent years that reinforce this need to reconsider aspirations for the future and the preferred path for progressing toward them.

Firstly, the new corporate "metadata infrastructure" delivered since 2001 has featured a "service" oriented architecture (SOA) designed to allow it to be readily "plugged into" existing processing systems. Actual take up by processing systems, however, has been extremely slow for a number of reasons. These include

- lack of funds (and business drivers etc) for updating existing processing systems
- existing processing systems being "monolithic" and not readily able to "take up" the corporate metadata services
- "special needs" of existing processing systems which are not fully met by the generic services and, at a minimum, require the added complexity of marrying up "corporate" metadata with local extensions
- lack of a standard way to present information about a specific data element, classification or statistical activity (which was originally defined from a statistician's perspective) in a way that a IT infrastructure can make systematic use of it
- inability of the processing system to provide metadata to describe "what it has done"
    - eg even if the "input" data elements have been defined, once transformation processes are undertaken within the processing system there will be no description of the resultant derived data elements

Secondly, the frame of reference for ABS metadata management requirements has become less and less defined by the boundaries of the organisation itself. Some of this breaking down of boundaries has driven by the ABS itself, seeking to develop infrastructure that could be shared with other producers of statistical data within Australia in order to assist them in contributing greater volumes of higher quality content to Australia's National Statistical Service. Other government agencies within Australia also now have their own focus on interoperability, including metadata and standards to support interoperability, which the ABS needs to recognise. Interest in software collaboration, particularly open source software, rather than purely "home grown" systems has also broadened the focus on metadata and interoperability. The focus on collaboration in regard to software development and to making data and service available means the ABS is not only increasingly working alongside other statistical agencies but also seeking to interoperate with agencies whose data content is more "administrative", "geospatial" or "research oriented" than "statistically" oriented.

These two factors might be seen to be pulling in opposite directions. Firstly, the practical complexities and difficulties of getting "the rubber to hit the road" in terms of using metadata in an "end to end" context to drive actual ABS processes is becoming clearer. Secondly, there is pressure for metadata capabilities to be more flexible, interoperable and generic - making them less ABS specific.

In combination these factors suggest that charting a successful way forward in regard to metadata management for the ABS does require a "paradigm shift", a new era.

*** END ***