

Case study: Statistics New Zealand

Case study: Statistics New Zealand Case study: Statistics New Zealand: use of GSBPM Case study: Statistics New Zealand: use of GSIM Statistics New Zealand: use of GSBPM Statistics New Zealand: use of GSIM

Contact person*	Matjaz Jug
Job title	Chief Information Officer
Email	matjaz.jug@stats.govt.nz
Telephone	+64 4 931 4238

Summary*

Metadata strategy

Like many National Statistical Offices around the world, Statistics New Zealand faces a number of 'external' and 'internal' challenges in the years ahead. 'External' challenges include: the need to minimise respondent burden, improve timeliness of existing data releases, improve 'time to market' for new data releases, increased use of administrative data, and better access to data (incl. micro-data) by users. While 'internal' challenges include: provide a better work environment for staff, replace an ageing IT platform & application toolset, measure 'value for money' for the New Zealand taxpayer, develop a platform to support future growth.

In response to these growing demands, Statistics New Zealand developed its **Business model Transformation Strategy (BmTS)**.

Current situation

The BmTS programme was started in July 2004, three years on we have largely developed the 'core platform' and see a positive way forward. The BmTS is the main platform that provides the framework for projects related to metadata to develop to. Most metadata related projects are being undertaken within the BmTS suite of projects, but those that are not are governed by the BmTS principles.

See [6. Lessons learned \(New Zealand\)](#) for more information on our experiences developing and implementing this programme.

Delivering benefits

The BmTS is aimed at delivering a number of benefits to Statistics New Zealand, and provide a solid basis for growth and development, through:

1. Abstracting the business users and their business processes from the underlying data structures and database systems, moving our statistical staff up the analytical 'value' chain and providing an environment that would facilitate the more challenging data integration and data analysis necessary to meet the increasingly complex policy and research needs of government and the wider research community.
2. Creating the flexibility to respond to changes in user needs and demands, to make use of new data sources or methods and to provide a flexible range of information access methods; while also providing the ability to more easily match and confront data in order to increase the quality of Statistics NZ information.
3. Reducing the time to design, build and process information sources, providing more time for analytical and dissemination processes.
4. Building a professional environment that creates a more satisfying working experience.
5. Increasing the use of administrative data, reducing the number of individual collections or the need for new collections to create new statistics.
6. Providing a standard environment and uniform systems that will allow staff to quickly get up to speed with new subject matter. This will also simplify the migration of data and systems as underlying technologies change, while reducing the maintenance cost of separate subject matter systems.
7. Standardising the skills sets and professional development costs of our staff.
8. Utilising a smaller number of larger projects that are more likely to have a real rate of return through the reuse of the investment in a number of business areas.
9. Allowing Statistics NZ to provide standard information management tools and services for official statistical purposes.

Metadata Strategy

The Business Model Transformation Strategy (BmTS) is designing a metadata management strategy that ensures metadata:

1. fits into a metadata framework that can adequately describe all of Statistics New Zealand's data, and under the Official Statistics Strategy (OSS) the data of other agencies
2. documents all the stages of the statistical life cycle from conception to archiving and destruction
3. is centrally accessible
4. is automatically populated during the business process, where ever possible
5. is used to drive the business process
6. is easily accessible by all potential users
7. is populated and maintained by data creators
8. is managed centrally

Established principles of metadata management

- metadata is centrally accessible
- metadata structure should be strongly linked to data
- metadata is shared between data sets
- content structure conforms to standards
- metadata is managed from end-to-end in the data life cycle.
- there is a registration process (workflow) associated with each metadata element
- capture metadata at source, automatically (where possible)
- establish a cost/benefit mechanism to ensure that the cost to producers of metadata is justified by the benefit to users of metadata
- metadata is considered active
- metadata is managed at as a high a level as is possible - managing at the lowest level is prohibitive
- metadata is readily available and useable in the context of client's information needs (internal or external)
- tracking the use of some types of metadata (eg. classifications)

Metadata Classification

The **MetaNet Reference ModelTM (Version 2)** categorises types of metadata in the following way:

- **Conceptual Metadata** describes the basic idea (concept) behind the metadata object e.g. conceptual data elements, classifications, measure units, statistical object types. This type of metadata can be context free (eg the variable 'income' as a concept) or context-related (e.g. 'income' collected in a particular survey).
- **Operational Metadata** are the metadata required to view the data from an operational point of view (e.g. record variables, matrix operations, statistical process). This includes all processes and configuration. In other words, the operational metadata is used to explain how the data was created or transformed. Operational Metadata is one of the links between the concepts and the physical data.
- **Quality Metadata** are the metadata for a particular instance e.g. response rates, status, weighting, versions. This provides the other link between concepts and physical data. It is worth noting that the processes involved in preparing quality metadata are considered operational metadata.
- **Physical Metadata** includes the physical, unique characteristics of the data which cannot be separated eg server locations, data base.

The **MetaNet Reference ModelTM (Version 2)** categorises types of metadata in the following way:

- **Conceptual Metadata** describes the basic idea (concept) behind the metadata object e.g. conceptual data elements, classifications, measure units, statistical object types. This type of metadata can be context free (eg the variable 'income' as a concept) or context-related (e.g. 'income' collected in a particular survey).
- **Operational Metadata** are the metadata required to view the data from an operational point of view (e.g. record variables, matrix operations, statistical process). This includes all processes and configuration. In other words, the operational metadata is used to explain how the data was created or transformed. Operational Metadata is one of the links between the concepts and the physical data.
- **Quality Metadata** are the metadata for a particular instance e.g. response rates, status, weighting, versions. This provides the other link between concepts and physical data. It is worth noting that the processes involved in preparing quality metadata are considered operational metadata.
- **Physical Metadata** includes the physical, unique characteristics of the data which cannot be separated eg server locations, data base.

3.2 Metadata used/created at each phase

Phase: Need

Description:

- Need is an ongoing process to determine the statistical needs of Statistics New Zealand's stakeholders

Description of main developed functionalities:

- Online Consultation/Submission Tool (Census)?
- Documentation Storage: Lotus Notes

Metadata used (inputs): table (below) with groups of metadata (rather than individual metadata items), examples for each group

Group	Description	Example	Source (if re-used)
-------	-------------	---------	---------------------

Conceptual Metadata	Concepts of interest. Previously available information.	Global variables, themes, subject areas, statistical objects types	Disseminate (previous cycle/collection)
Operational Metadata	Analyse processes of previous collections/ available data.	Study, Study Method, Statistical Process	Process (previous cycle/collection)
Quality Metadata	Quality of previous collections, or available data	Data Quality	Process and Analyse (previous cycle /collection)
Physical Metadata	Locate available data through physical metadata	Datasets, server locations, software, access rights	Disseminate (previous cycle/collection)

Metadata produced (outputs): table (below) with groups of metadata (rather than individual metadata items), examples for each group

Group	Description	Example	Source (if re-used)
Conceptual Metadata	Defined concepts high level	Global variables, themes, subject areas, statistical objects types	
Operational Metadata	High level strategy for meeting need.	Business Case, Study Method, Statistical Process	
Quality Metadata	Consultation process	Reports	
Physical Metadata	Document storage, submission storage	Locations, References	

Phase: Develop and Design

Description:

- Develop and Design describes the research, development and design activities to define the statistical outputs, methodologies, collection instruments, sample, operational processes and end-to-end (E2E) solution

Description of main developed functionalities:

- Documentation Storage: Lotus Notes

Metadata used (inputs): table (below) with groups of metadata (rather than individual metadata items), examples for each group

Group	Description	Example	Source (if re-used)
Conceptual Metadata	Need to refine the concepts to determine the variables to be collected and the population of interest.	Includes concepts such as object variables, value domains, classifications, measure units, context data elements.	Need (High level concepts)
Operational Metadata	This includes designing the sample methodology, collection methodology and the statistical processes. Also need to design quality characteristics and processes.	Includes concepts such as statistical process, process implementation, data collection methodology.	Need (strategy) Collect/Process (previous collections)
Quality Metadata	Quality metadata from previous collections may be used in the design of sample and collection methodologies.	statistical process, process implementation, data collection methodology.	Collect/Process (previous collections)
Physical Metadata	May begin to define/ identify where data will be stored. May also require physical metadata from previous collections to determine the best	software, data sets, access package,	

Metadata produced (outputs): table (below) with groups of metadata (rather than individual metadata items), examples for each group

Group	Description	Example	Source (if re-used)
Conceptual Metadata	Defined concepts	Includes concepts such as object variables, value domains, classifications, measure units, context data elements.	
Operational Metadata	Completed end to end design and methodology	statistical process, process implementation, data collection methodology, business rules, transformations. testing plan.	
Quality Metadata	Details of the design process used. Quality introduced by proposed standards.	Design process. Quality measures of standards.	
Physical Metadata	Storage plan. Document storage	Software, Access Package, Data Set	

Phase: Build

Description:

- Build produces the components needed for the end-to-end solution, and tests that the solution works

Description of main developed functionalities:

- Dashboard - For configuring and monitoring processes
- Workflow tool - for developing statistical processes
- Questionnaire Design, CAI tools, Scanning Software - For survey collections
- CRM, Q-Master (call centre) - For running collections
- Transformation Tools - imputation, editing, coding, etc.
- Data Environments - for storage of data (in development)
- Metadata Environment Components - for storage of metadata (developing high level design)

Metadata used (inputs): table (below) with groups of metadata (rather than individual metadata items), examples for each group

Group	Description	Example	Source (if reused)
Conceptual Metadata	May need to refine/adapt concepts due to feedback from testing or errors detected.	Data Elements, Classifications, Value Domains	Develop and Design
Operational Metadata	Build and configure storage structures, collection instruments, processing requirements. Includes concepts such as record variables, record types, matrix operations.	Question, questionnaire, data collection methodology, statistical process.	Develop and Design
Quality Metadata	The final version of a questionnaire will be decided during this stage. Some quality metadata may also be used to assess pilot studies and instrument testing.	Methodology and process design. Testing plans.	Develop and Design
Physical Metadata	Determine the physical location where the data will be stored	servers, software, access packages.	Develop and Design

Metadata produced (outputs): table (below) with groups of metadata (rather than individual metadata items), examples for each group

Group	Description	Example	Source (if reused)
Conceptual Metadata	Finalised concepts	Data Elements, Classifications, Value Domains, Statistical Unit	
Operational Metadata	End to End components built and tested	Built Components, Processes. Statistical process, data collection methodology.	
Quality Metadata	Quality metrics about the build process and testing report	Test reports, validated processes.	
Physical Metadata	Complete Application Architecture	Software, Storage, Access Packages, Access Rights, Versions etc.	

Phase: Collect

Description:

- Collect acquires collection data each collection cycle and manages the providers of that data

Description of main developed functionalities:

- Questionnaire Design, CAI tools, Scanning Software - For survey collections
- CRM, Q-Master (call centre) - For running collections
- Data Environments - for storage of data (in development)
- Metadata Environment Components - for storage of metadata (developing high level design)

Metadata used (inputs): table (below) with groups of metadata (rather than individual metadata items), examples for each group

Group	Description	Example	Source (if reused)
Conceptual Metadata	As data is collected it will be allocated against concepts. Some details about the relevant concepts may be used in respondent management strategies.	Data Elements, Classifications, Value Domains, Statistical Unit	Develop and Design
Operational Metadata	Utilise the collection processes outlined in the operational metadata.	Data Collection Methodology, Questionnaire, Collection Strategy.	Develop and Design, Build
Quality Metadata	Collect data for each instance of the survey. Quality metadata will be populated based on the collection instance.	Collection strategy, Data Collection Methodology,	Develop and Design, Build
Physical Metadata	Physical datasets will be populated with data at this stage.	Software, Storage, Access Packages, Access Rights, Versions etc.	Build

Metadata produced (outputs): table (below) with groups of metadata (rather than individual metadata items), examples for each group

Group	Description	Example	Source (if reused)
-------	-------------	---------	--------------------

Quality Metadata	Operational Processes used. Quality measures of collection instance.	Collection Report, Data Quality - response rate, item non response.		
Physical Metadata	Data collected to measure collection concepts.	Software, Storage, Access Packages, Access Rights, Versions etc.		

Phase: Process

Description:

- Process describes cleaning the detailed data records and preparing them for analysis

Description of main developed functionalities:

- Dashboard - For configuring and monitoring processes
- Workflow tool - for developing statistical processes
- Transformation Tools - imputation, editing, coding, etc.
- Data Environments - for storage of data (in development)
- Metadata Environment Components - for storage of metadata (developing high level design)

Metadata used (inputs): table (below) with groups of metadata (rather than individual metadata items), examples for each group

Group	Description	Example	Source (if reused)
Conceptual Metadata	Conceptual metadata are used to classify and code open responses. Derive new concepts, aggregate data etc.	Classification, Correspondence, Data Elements, Classifications, Value Domains, Statistical Unit	Develop and Design
Operational Metadata	Further statistical processes are used in order to process the data. This may include the creation of cubes and registers, aggregating results using derivation rules, applying editing and imputation strategies, applying confidentiality rules, etc.	Matrix, Cube, Register, Statistical Process, Process Implementation, Operation Implementation, Derivation Rules, Computation Implementation.	Develop and Design, Build, Collect

Metadata produced (outputs): table (below) with groups of metadata (rather than individual metadata items), examples for each group

Group	Description	Example	Source (if reused)
Quality Metadata	Further quality metadata will be populated at this stage based on the processes applied.	Processing reports, Data Quality - imputation rates, editing rates etc.	Develop and Design, Collect
Physical Metadata	If processed data is stored in different locations, new physical metadata will be defined.	Storage, Software, Access Package etc.	Build

Phase: Analyse

Description:

- Analyse is where the statistics are produced, examined in detail, interpreted, understood and readied for dissemination

Description of main developed functionalities:

- Analytical Environment (strategy still in development)
- Information Portal (strategy still in development)
- Data Environments - for storage of data (in development)
- Metadata Environment Components - for storage of metadata (developing high level design)

Metadata used (inputs): table (below) with groups of metadata (rather than individual metadata items), examples for each group

Group	Description	Example	Source (if reused)
Conceptual Metadata	In the analysis stage the processed data will be used to analyse the concepts defined in the need and develop/design stages.	Data Elements, Classifications, Value Domains, Statistical Unit	Need, Develop and Design
Operational Metadata	Further processes may be used to generate tables for analysis. Operational metadata will also be used to prepare data for dissemination. Operational Metadata may also be used to analyse the data.	Tables, Statistical Processes, Confidentiality Rules	Develop and design
Quality Metadata	Quality metadata will be used at this stage to assess how well the data represents the concepts outlined in the needs stage.	Statistical Activity, Study, Statistical Process.	Collect, Process
Physical Metadata	Physical metadata will be required to locate the data, and any additional data for comparisons.	Storage, Software, Access Package etc.	Build, Collect, Process

Metadata produced (outputs): table (below) with groups of metadata (rather than individual metadata items), examples for each group

Group	Description	Example	Source (if reused)
Quality Metadata	Analysis of quality metadata against concept defined in need stage.	Data Elements, Classifications, Value Domains, Statistical Unit, Statistical Activity, Study, Statistical Process.	
Physical Metadata	Produce analysis reports, output products.	Data Set, Publications	

Phase: Disseminate

Description:

- Disseminate manages the release of the statistical products to the customers.

Description of main developed functionalities:

- Integrated Publishing Environment - Tool for configuring and disseminating analysis.
- CRM/ Job Tracking Systems - for recording customer usage.
- Data Environments - for storage of data (in development)
- Metadata Environment Components - for storage of metadata (developing high level design)

Metadata used (inputs): table (below) with groups of metadata (rather than individual metadata items), examples for each group

Group	Description	Example	Source (if reused)
Quality Metadata	Technical audiences and professional audiences may be interested in the quality metadata in order to understand the characteristics of each instance of data collection.	Analysis Reports, Output Releases, Quality measures/ reports, Needs definition	Need, Analysis
Physical Metadata	Physical metadata may be required when locating data in response to customer queries.	Server locations, access rights, access packages, systems and tools, online catalogues	Build, Analyse

Metadata produced (outputs): table (below) with groups of metadata (rather than individual metadata items), examples for each group

			Source (if reused)
Conceptual Metadata	Identification of new needs for collection.	Global variables, themes, subject areas, statistical objects types	
Operational Metadata	Development of new/changed standards for other collections	Study Methods, Statistical Processes	
Quality Metadata	Details of products produced, queries received and products used. Long term retention and archiving policies.	Dissemination process, Usage data	
Physical Metadata	Output products available for promotion.	Tables, Reports, Brochures etc.	

3.3 Metadata relevant to other business processes

Group	Description	Example
Conceptual Metadata	Information/ records defining organisational concepts, structures etc	Corporate directories, Corporate glossary, Organisational Structure charts
Operational Metadata	Information/ records defining how corporate processes are applied	Corporate policies, Guides, Contracts, Memorandums of understanding (MOU)
Quality Metadata	Information/ records related to specific instances	Invoices, Annual Reports, Budgets
Physical Metadata	Information relating to physical item	Catalogues, Asset Management Lists etc.

Metadata system(s)

The metadata infrastructure will be implemented within the 10 components covering the whole BmTS environment.

The Metadata Broad Logical Design 2004 defined nine components of key metadata infrastructure which were needed to create the physical metadata environment. These are shown in the diagram below.

Note: In defining the relationships, the terms period dependent and period independent are used. Period dependent refers to metadata which is linked to a specific activity/collection (includes quality metadata). Period independent metadata has meaning while held separate from data and can be applied to several collections (includes operational and conceptual metadata).

Search and discovery, Metadata and data access/ registration

These components reflect the ways the user interacts with the metadata. Ideally, searching, registration and access should be possible directly with each component, or through a central portal.

Metadata Storage

Data Definition - The data definition component is the only infrastructure linked directly to the data. This is the primary store which defines and adds meaning to the data. This is a period dependent store which compiles the relevant metadata for a single collection. All other storage components link to the data via the data definition store.

Passive Metadata Store - The passive metadata store is the next level removed from the data. It contains period dependent and period independent metadata about a collection of data (this includes survey collections and administrative data collections).

Question Library - The question library should be period independent. It contains questions and variables which have been defined independent of the data. The question library and classification management store are linked through the classifications used in questions.

Classification Management - The classification management store is another period independent store which manages the classifications used to define the data. It includes metadata linking classifications to each other (concordances) to allow more options when analysing and transforming the data.

Business Logic - The final period independent metadata store is the business logic component. While business logic is not linked directly to the data, it is applied to change the data through its various states. This contains details of the rules and processes that may be applied to the data. Business logic may also be referenced in the design and methodology content of the passive metadata store. The business logic component sits partly outside the storage environment due to the need for software to access the rules and processes (e.g. rules engines).

Other Storage

Frame and Reference Stores - While the Frame is not part of the metadata environment, it may contain information which is used to define the data. Hence there is a link between this component and the data definition component.

Document Management - Document One is a tool for the management of documentation. As several reports and documents will be created during the business process, they are considered part of the wider metadata environment.

Standards Framework - The standards framework represents a tool for the central storage of standards used in the generic business process. This includes a definition of processes and methodologies at high levels. It will also include statistical standards which define how classifications are applied. Similar to Document One, this should be considered part of the wider metadata environment.

Logical View of Metadata Infrastructure and Relationships

In 2007, further analysis was completed using the gBPM and the MetaNet Reference Model to build a more detailed understanding of the logical metadata stores and the key relationships with data (see the model below).

The reference metadata layer contains stores of metadata with similar characteristics which allow it to be managed in a consistent way. For instance, classifications are used at various stages throughout the statistical business process to define various types of data. By storing and managing classifications in a separate 'classification management' store, we are able to analyse usage and identify opportunities for further standardisation.

In order to develop a fully integrated metadata environment, each metadata object will need to be linked with objects within the store, or in other reference stores. For instance, a description of business process will be stored in the 'standards and processes' component, this will also need to link to the 'operational metadata' component containing workflows and transformations which operationalise the process. The workflows and transformations may also reference 'business rules' which utilise concepts from the 'variable library'. Linking metadata objects allows the user to consider the full usage of each object and will enhance reuse and standardisation. The 'Structural Metadata Layer' is the mechanism for linking the reference metadata with the actual data. Each data item (or fact) within the data environment should contain a profile within the data definition store which identifies all the metadata relevant to that fact. Ideally this will consist of a map identifying the location of the relevant metadata in the reference stores. However, until all the reference stores exist, this component will contain snapshots of the relevant metadata.

When translating from the logical view, to the physical design it is anticipated that the components will take a different shape to that referenced in the model above. For instance, 2008 will see the investigation of a single system to manage classifications, questions and variables.

Costs and Benefits

The high level benefits of undertaking the metadata programme were outlined in the introduction. Additional benefits to consider are as follows:

- maximising the value of metadata through reuse.
- reducing the unnecessary duplication of metadata.
- providing a more comparable, central source of metadata to allow for improvements through standardisation.

It is recognised that the development of a full metadata solution will require a large investment in the infrastructure of the organisation. There are also various levels of investment that could be applied to deliver the most practical solution (eg if the needs are a lower priority for one type of metadata, then the solution should be less complex). At this stage, the cost of delivery has not been calculated, however it is known the following considerations will need to be addressed:

- Large amounts of metadata already exist in various systems will need to be transferred into new systems where practical.

- The principle of reuse, may require more effort in creating and storing metadata at early stages of survey development in order to reduce effort at later stages (essentially shifting the effort, rather than reducing it).
- There will need to be careful management to ensure the duplication of metadata is reduced (ie if it's already entered in the reference store, it should be selected and configured for current requirements rather than duplicated).
- Detailed versioning will be required to ensure that the metadata is relevant to the instance it relates to. This may increase the storage requirements.

Implementation strategy

The work of the metadata project during 2007 focused on identifying the high level needs of an integrated environment and the adaptation of a metadata conceptual model to understand the relationships and interactions with metadata. With the bigger picture in place, the intention is to focus on developing solutions for smaller components of the wider environment. This approach allows us to focus delivery in the areas which will provide the most gain, while still progressing along the path to delivering the fully integrated solution. It also allows us to assess the strategy at each stage to determine the most practical approach and to minimise the risk of the delivering over-complicated solutions.

IT Architecture

The introduction of Service Oriented Architecture (SOA) into Statistics NZ was the culmination of researching industry trends and evaluating those trends against the new technical challenges that were arising in response to the BmTS.

The BmTS has three core deliverables:

1. A standard and generic end-to-end process or processes to collect, process, analyse and disseminate data (Value Chain).
2. An approach to data management which is disciplined and consistent (Information Architecture).
3. An agreed organisation-wide technical architecture as a framework for making system decisions.

To support the first two deliverables and to ensure that the third deliverable is achieved Statistics NZ has adopted a Service Oriented Architecture (SOA) approach. A SOA resolves the behaviour of the organisation's IT assets into a set of common behaviours or services. Services can be business services and technical services.

The SOA is a key enabler of BmTS exposing common services (business / statistical and technical) as an abstract, decoupled and consistent set of interfaces enabling the communication of as much of the process and data in Statistics NZ's core business. In addition, there are a number of benefits related to the incorporation of third party software; this includes off-the-shelf applications and providing and using services to and from other statistical agencies. Key aspects of the Statistics NZ SOA are that the consumer of the service can find and bind to services at runtime and the SOA extends to the development, deployment and management of services.

This architecture will enable the transparent exchange of metadata between different systems and tools. The current service layer is supporting some existing metadata components like process management (workflow), business rules (rules engine) and the integration with main systems (CRM-based respondent management and call centre), tools (SAS, ETL, Blaise) and databases (SQL Server).

Metadata Management Tools

Currently within Statistics New Zealand, a programme of work is underway developing the plans for the components of the key metadata infrastructure. While the plans are still unconfirmed, the summary below addresses the current thinking, likely direction and known issues for each component.

Search and discovery, Metadata and data access/ registration - These components will be developed within our wider Information Portal development. Currently investigations are underway to implement a single searching tool which will facilitate access across data environments and documentation stores. While further planning is needed to generate the high level strategy for access, search and discovery, the principles of reuse and integration are being incorporated in other components to ensure these components can be developed.

Metadata Storage

Data Definition - The data definition component will be based on metadata directly linked to the data in IDE (Input Data Environment) and ODE (Output Data Environment). The conceptual vision for the IDE and ODE is that they will contain several 'areas' which reflect the use of data. For example, in the IDE this includes a Load Area, Operational and Exceptions area (for processing), Clean area and Aggregate area (for analysis) and Data Marts (for Time Series, Longitudinal data etc). The data definition component will need to reflect the metadata needs for each of these areas and will emphasise reuse by ensuring the flow of metadata as the data flows through the environments.

Passive Metadata Store - The passive metadata store is currently implemented within the Lotus Notes Environment and is not directly linked to other metadata stores. While the strategy for developing this component are still being planned, it is recognised that there are current issues with structure and reusability which need to be addressed. Currently passive metadata is stored based on a flat structure where metadata for each output is stored. However this does not recognise the complex nature of collections where one input can be used in several outputs, an output can become an input for another collection, and inputs come in several forms (including survey collections and administrative data collections). There is also a recognised need for developing a more dynamic glossary which can be linked into multiple stores of metadata.

Classification Management - CARS (Classifications And Related Standards) is in use for classification management. The system in regular production is based on the relational model (currently implemented in Sybase) and an application for classification management (currently implemented with Centura). There is a plan to upgrade the platform to .NET/SQL and enhance for integration within the new SOA architecture.

Question/Variable Library - Current thinking in regard to this component is that a tool is required which manages variable definitions as well as question use. This is likely to take the form of a reference store where variables can be configured linking variable definitions, classifications, value domains, statistical objects and collection elements (e.g. question, questionnaire etc). At this stage planning is underway to determine the feasibility of combining the development of this component with the enhancement to our Classification Management component.

Business Logic - The Business Logic component encompasses the operational metadata for the statistical business process. This includes the processes used to change and transform data, the configuration which outlines the inputs and outputs, the business rules which set parameters for changing data, and the active metadata which is used to run the processes (e.g. variable identifiers, programme code). Business logic will also include quality metadata for a particular instance which defines the process that was run, rules applied, and audit trails including who, what and when, etc. Currently the storage of operational metadata is being developed with separate components such as workflow tools (K2) and transformation tools (CANCEIS, Logiplus etc). Investigations are currently underway to determine a way to integrate these components through generic storage schemas. A separate investigation is also planned looking at rule engine usage and the storage of business rules in a generic form.

Other Storage

Frame and Reference Stores - The business frame is in regular production and there is a link between this component and the data definition component (implemented as a reference from IDE to record in business frame). The similar component for persons & household frame is under development.

Documentation and Reports - Document One is a Lotus Notes application for the management of documentation. It is a central system for corporate document management but will be linked into the metadata environment through the information portal development.

Standards and Processes - Two aligned applications are currently in development to manage standards and processes (the Standards Framework and the gBPM Repository). The Standards Framework is a Lotus Notes application under development for the central storage of statistical and methodological standards. The tool stores standards using the generic Business Process Model as the framework so that users can access the relevant standards for any process they undertake. The application includes the development of standards within the store, through the use of versioning, notification systems and recording of consultation. The gBPM repository is a tool for storing our corporate process models including detailed process descriptions down to activities and tasks. Following the completion and population of the tools, further investigation will be conducted regarding the integration with other components within the metadata environment.

Standards and formats

When defining a concept based model for to be used as the overarching metadata framework, four concept based models were reviewed, specifically DDI, SDMX, MetaNet Reference Model v2.0 (MetaNet), and Neuchatel Terminology Model. In December 2006 a working group determined that no one model met all the needs of Statistics NZ. A blended model was recommended taking the best components of two models to create a single model - MetaNet and SDMX. Further analysis by the working group provided clarity for each model evaluated, including details of risks, impacts and gaps. They produced a second recommendation, this time a primary model - MetaNet with a secondary layer to treat any gaps - SDMX. Selection was based on simplicity, adaptability and integration and ability to support the business process. During this stage the MetaNet model was analysed and mapped against internal metadata stores to assess the usability within Statistics New Zealand. As a result of this process a series of recommendations were presented for adapting the model to better meet our needs. Work has been progressing to adapt this model with a revised version due for completion early in 2008.

Version control and revisions

Currently the broad principles for versioning have been developed, however the application will need to be part of the individual developments. Returning to the Logical View of our data and metadata stores in section 2, it is intended that versioning will be maintained within the 'reference metadata layer'. There will also need to be versions of the data definitions within the 'structural metadata layer', however these will essentially be linking structures which identify the relevant versions of reference metadata.

Outsourcing versus in-house development

The strategy for developing components of the metadata environment is still being developed, however in general a principle of enhance first, then buy before build is applied. A summary of the state of current developments is as follows:

- Currently the solution for our data definition is being developed in house as part of the wider development of the Input Data Environment (IDE).
- The decision has also been made that many of the business logic components will be dependent on the tools used to run transformations and workflows.
- Planning is underway to investigate the feasibility of enhancing our current classification management tool (CARS) to incorporate functionality for question and variable management.
- Longer term we also aim to redevelop our current survey metadata tool (SIM), however the form of this development is as yet undecided.
- Additional tools have also been adopted for managing other metadata components, eg documentation and report management is being enhanced through the implementation of Document ONE on top of our current Lotus Notes functionality.

Sharing software components of tools

Overview of roles and responsibilities

Overview of metadata audiences and use of metadata

To ensure that metadata is relevant and useful a high level analysis of the audiences of the metadata environment has been completed. The audiences have also been identified using several broad user groups - Public, Professional, Technical, System.

External Audiences

User Group Name	Use of the solution	Type
-----------------	---------------------	------

Government	<ul style="list-style-type: none"> • Request new statistics about NZ • Use statistics about NZ 	Professional
Public	<ul style="list-style-type: none"> • Request new statistics about NZ • Use statistics about NZ 	Public
External Statisticians (incl. International Organisations)	<ul style="list-style-type: none"> • Request new statistics about NZ • Use statistics about NZ • Interact with data in Statistics NZ environment 	Technical

Internal Audiences

User Group Name	Use of the solution	Type
Statistical Analysts	<ul style="list-style-type: none"> • Interact with data in Statistics NZ environment • Utilise workflows to manage data processing and assembly • Analyse data through standard tools and processes. • Maximise the re-usability of data. • Work with Statistical Methods, IT, and Management to specify standardised generic solutions • Automated task delivery • Automated activity monitoring • Process transparency and traceability • Devolve processing responsibility to known automated capability 	Technical Professional
IT Personnel (business analysts, IT designers & technical leads, developers, testers etc.)	<ul style="list-style-type: none"> • Utilise components to develop processing system of end-to-end solution • Utilise standard toolsets and methods to develop processing systems. • Focus development effort on adding value to the Statistics NZ Business Process Model • Work with Management and Statistical Analysts to determine business processes. • Work with SM, SA, and Management to design standardised generic solutions. • Build new and maintain components of standardised generic solution 	Technical Professional
Management	<ul style="list-style-type: none"> • Utilise workflow to manage data processing and assembly. • Work with Statistical Methods, IT to design and configure workflows • Monitor progress of business processes to produce statistical outputs. • Automated monitoring of end-to-end process • Process transparency and traceability 	Professional
Data Managers / Custodians / Archivists	<ul style="list-style-type: none"> • Manage Statistics NZ data & metadata holdings with an end-to-end perspective. • Inventory of Statistics NZ data and metadata • Process transparency and traceability 	Technical Professional
Statistical Methodologists	<ul style="list-style-type: none"> • Work with SA, TAS, and Management to specify standardised generic solutions, with a particular responsibility for the statistical methods applied • Maintain standard modules within the infrastructure (e.g. seasonal adjustment) • Define statistical methodologies to use • Design Transformation methods • Assist in Workflow design and configuration • Provide knowledge on statistical methodologies used by Stats NZ • Incorporation of adopted best practice into end-to-end processing model 	Technical Professional
External Statisticians (researchers etc.)	<ul style="list-style-type: none"> • Interact with data in Statistics NZ environment • Context metadata supports data needs 	Technical Professional

Architects - data, process & application	<ul style="list-style-type: none"> Specify standardised generic solutions. Plan for progressive architecture Isolate IT dependencies with defined integration and interoperability mechanisms 	Technical Professional System
Respondent Management	<ul style="list-style-type: none"> Provide a user friendly service to external data providers. 	Professional
Survey Developers	<ul style="list-style-type: none"> Utilise a standard toolset to develop and maintain collection instruments Utilise metadata around existing collection instruments to reuse in new /enhanced collection instruments. Question Library Inventory of Statistics NZ data and metadata 	Professional
Metadata and interoperability experts	<ul style="list-style-type: none"> Propose and adopt metadata standards and business rules Manage and monitor the core corporate metadata assets 	Professional Technical
Project Managers & Teams	<ul style="list-style-type: none"> Manage projects to deliver an end-to-end solution using the components Manage projects to determine feasibility of new statistical outputs. Manage field tests of new components of standardised generic solution. 	Technical
IT Management	<ul style="list-style-type: none"> Reduce development time and costs Control maintenance/support costs Define capability requirements 	Professional Technical
Product Development and Publishing	<ul style="list-style-type: none"> Supply of contextual metadata supports data needs 	Professional
Customer/ Client Management	<ul style="list-style-type: none"> Maintain customer details Capture customer request for statistics Search and discover existing dataset for customer Report progress on logged customer issues 	Professional

Notes: This is the list of actors, i.e. roles using the Solution.

Metadata management team

While currently undefined, it is recognised that additional support will be required to maintain, monitor and improve the metadata usage in the organisation. Teams within our information management group and standards, solutions and capability group are currently involved in developing our metadata processes and infrastructure.

Current contacts on metadata within the organisation include:

Name	Role	Topics	E-mail	Phone
Hamish James	Manager - Information Management	Metadata models and systems, Information management practices and systems	hamish.james@stats.govt.nz	+644 931 4237
Matjaz Jug	Chief Information Officer - IT Solutions	Metadata and statistical production systems	matjaz.jug@stats.govt.nz	+644 931 4238
Craig Mitchell	Programme Manager - Business Solutions	Metadata models and systems, statistical business process model	craig.mitchell@stats.govt.nz	+644 931 4840

These groups will also be made responsible for the ongoing training, knowledge management and support of the metadata solution. Over 2008, the metadata team completed the adaptation of MetaNet to produce the Statistics New Zealand Metadata Model. This is being consulted on internally and will be finalised in early 2009. This is being seen as the first step in building a more consistent knowledge of metadata across the organisation. Further strategies will be developed as the metadata infrastructure is completed and implemented.

Training and knowledge management

.

Partnerships and cooperation

.

Other issues

Lessons learned

1. Apart from 'basic' principles, metadata principles are quite difficult. To get a good understanding of and this makes communication of them even harder. As it is extremely important to have organisational buy-in, the communication of the organisation metadata principles and associated model is something that needs some strong consideration.
2. Every-one has a view on what metadata they need - the list of metadata requirements / elements can be endless. Given the breadth of metadata - an incremental approach to the delivery of storage facilities is fundamental.
3. Establish a metadata framework upon which discussions can be based that best fits your organisation - we have agreed on MetaNet, supplemented with SDMX. As Statisticians we love frameworks so having one makes life a lot easier. You could argue that the framework is irrelevant but its the common language you aim to use.
4. There is a need to consider the audience of the metadata. The table about users covers some of this, but there is also the model where some basic metadata is supplied (e.g. Dublin Core) that will meet one need but this will then be further extended to satisfy another need and then extended even further to meet another need.
5. To make data re-use a reality there is a need to go back to 1st principles, i.e. what is the concept behind the data item. Surprisingly it might be difficult for some subject matter areas to identify these 1st principles easily, particularly if the collection has been in existence for some time.
6. Some metadata is better than no metadata - as long as it is of good quality. Our experience around classifications is that there are non-standard classifications used and providing a centralised environment to support these is much better than having an 'black market' running counter to the organisational approach. Once you have the centralised environment with standard & non-standard metadata you are in a much better position to clean-up the non-standard material.
7. Without significant governance it is very easy to start with a generic service concept and yet still deliver a silo solution. The ongoing upgrade of all generic services is needed to avoid this.
8. Expecting delivery of generic services from input / output specific projects leads to significant tensions, particularly in relation to added scope elements within fixed resource schedules. Delivery of business services at the same time as developing and delivering the underlying architecture services adds significant complexity to implementation. The approach with the development of the core infrastructure components within the special project was selected to overcome this problem.
9. The adoption and implementation of SOA as a Statistical Information Architecture requires a significant mind shift from data processing to enabling enterprise business processes through the delivery of enterprise services.
10. Skilled resources, familiar with SOA concepts and application are very difficult to recruit, and equally difficult to grow.
11. The move from 'silo systems' to a BmTS type model is a major challenge that should not be under-estimated.
12. Having an active Standards Governance Committee, made up of senior representatives from across the organisation (ours has the 3 DGSs on it), is a very useful thing to have in place. This forum provides an environment which standards can be discussed & agreed and the Committee can take on the role of the 'authority to answer to' if need be.
13. Well defined relationship between data and metadata is very important, the approach with direct connection between data element defined as statistical fact and metadata dimensions proved to be successful because we were able to test and utilize the concept before the (costly) development of metadata management systems.
14. Be prepared for survey-specific requirements: the BPM exercise is absolutely needed to define the common processes and identify potentially required survey-specific features.
15. Do not expect to get it 100% right the very first time.

Links:

Attachments