WP.10

ENGLISH ONLY

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

UNECE Workshop on the Common Metadata Framework
(Vienna, Austria, 4-6 July 2007)

CASE STUDY - New Zealand[1]

1. INTRODUCTION

Organisational Details: Statistics New Zealand, PO Box 2922, Wellington, New Zealand. Statistics NZ has 850 FTE. The organisational chart is in the Appendix 1.

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 tax-payer, develop a platform to support future growth. In response to these growing demands, Statistics New Zealand developed its Business model Transformation Strategy (BmTS). 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.

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.

[1] Prepared by Gary Dunnet, Craig Mitchell and Matjaz Jug

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.

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

i. 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
ii. documents **all the stages of the statistical life cycle** from conception to archiving and destruction
iii. is **centrally accessible**
iv. is **automatically populated** during the business process, where ever possible
v. is used to **drive the business process**
vi. is easily **accessible** by all potential users
vii. is **populated and maintained by data creators**
viii. is **managed centrally**

The following principles of metadata management have been established, this include:

- 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)

Lessons learned & future challenges

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.

2. THE STATISTICAL METADATA SYSTEMS AND THE STATISTICAL CYCLE

Description of metainformation systems

The Metadata Broad Logical Design [6] defined nine components of key metadata infrastructure which were needed to create the physical metadata environment. These are shown in the diagram below.

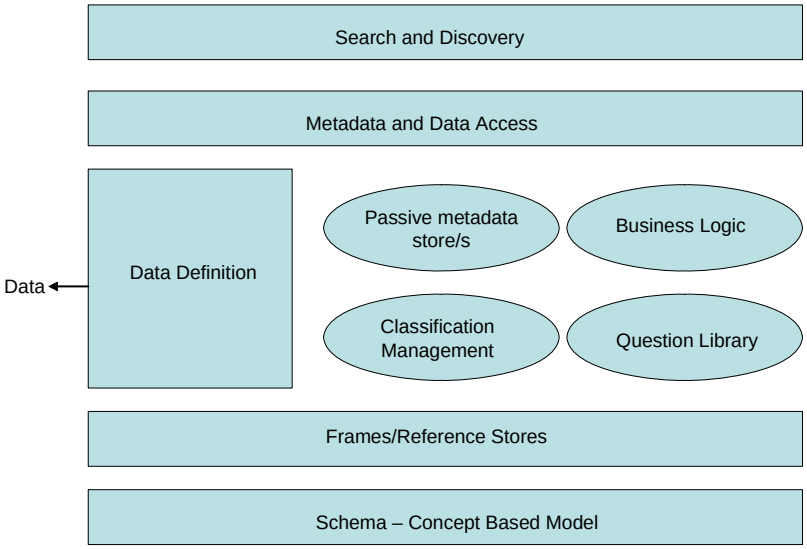


Figure 1: Key metadata infrastructure

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 it's 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.

The metadata infrastructure will be implemented within the 10 components covering the whole BmTS environment (see figure 2)

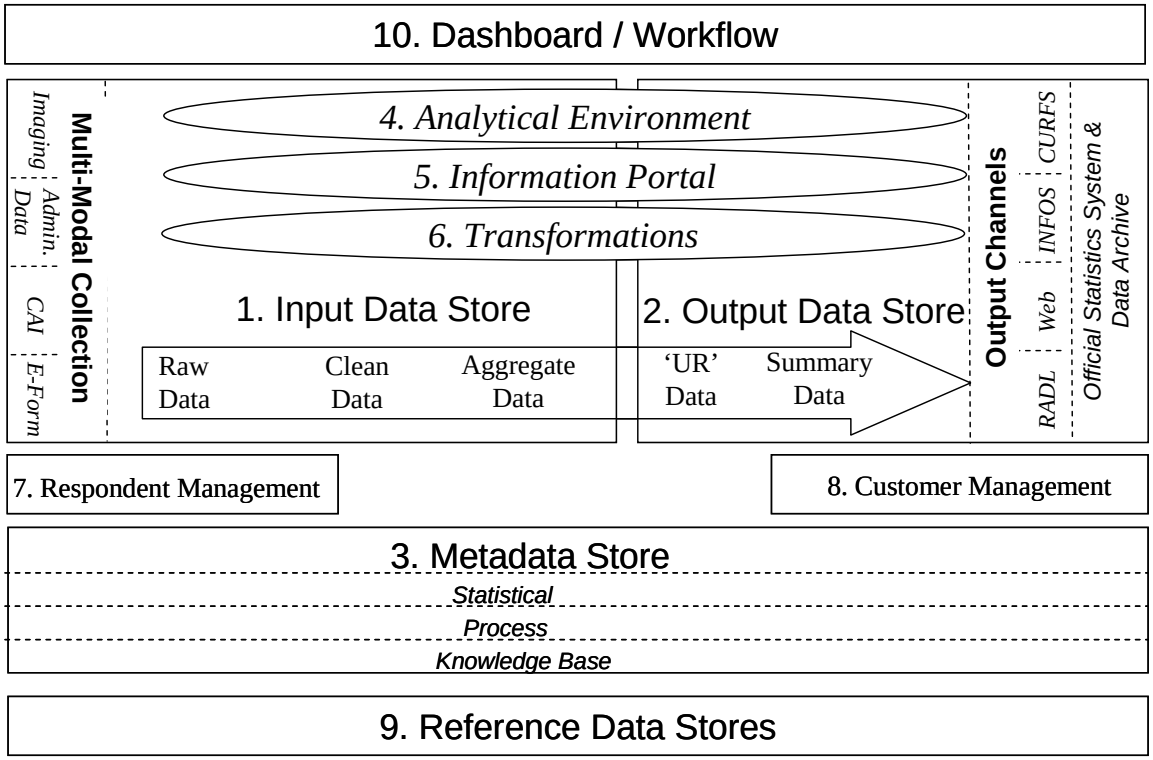


Figure 2: 10 components

Key Metadata Infrastructure relationships

The intention of the Metadata Project is to deliver an integrated environment of metadata components. Hence each of the components of the key metadata infrastructure relate to each other (and to the data) in specific ways. Figure A.1 shows the relationships between the key components.

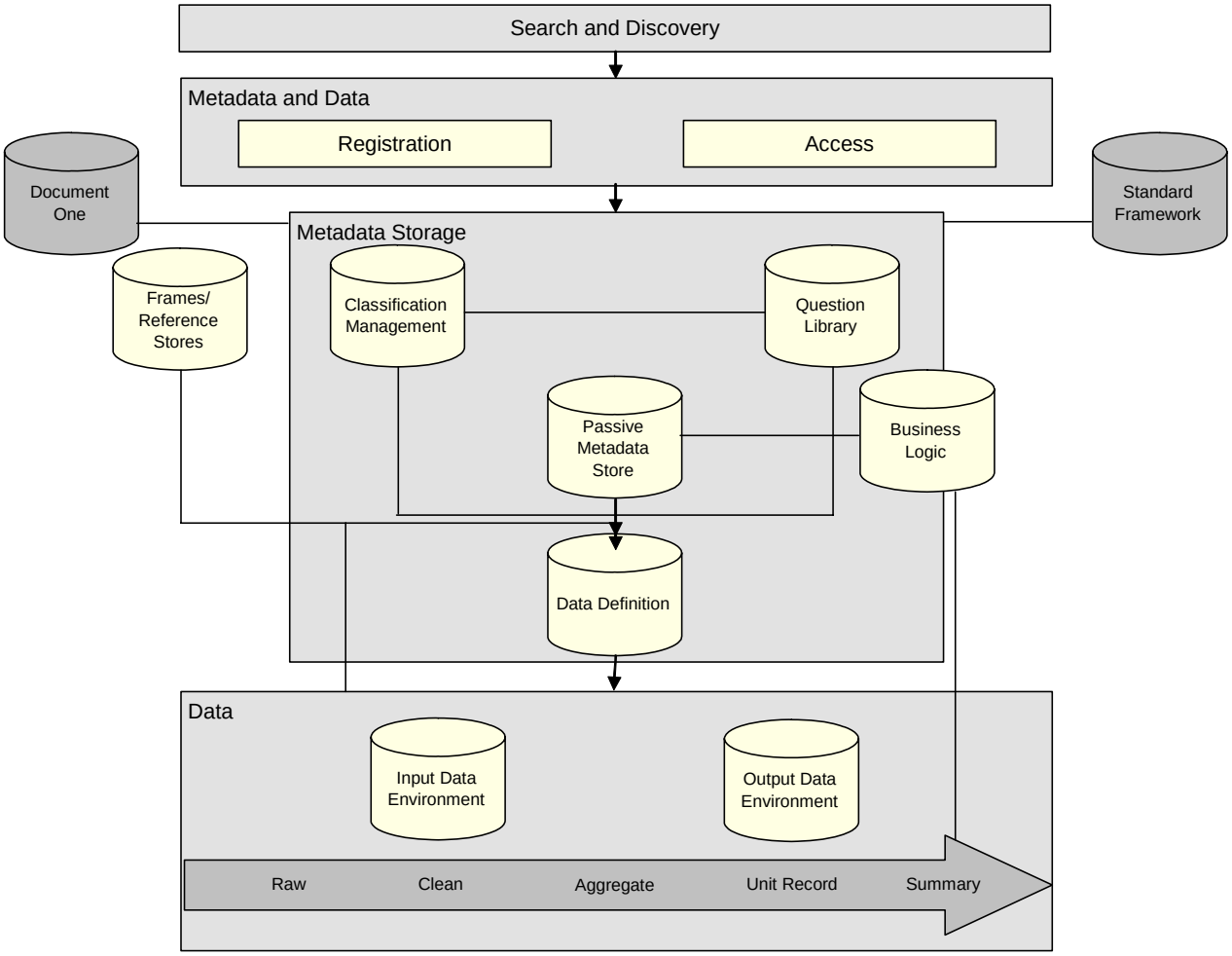


Figure 3: Key metadata infrastructure relationships

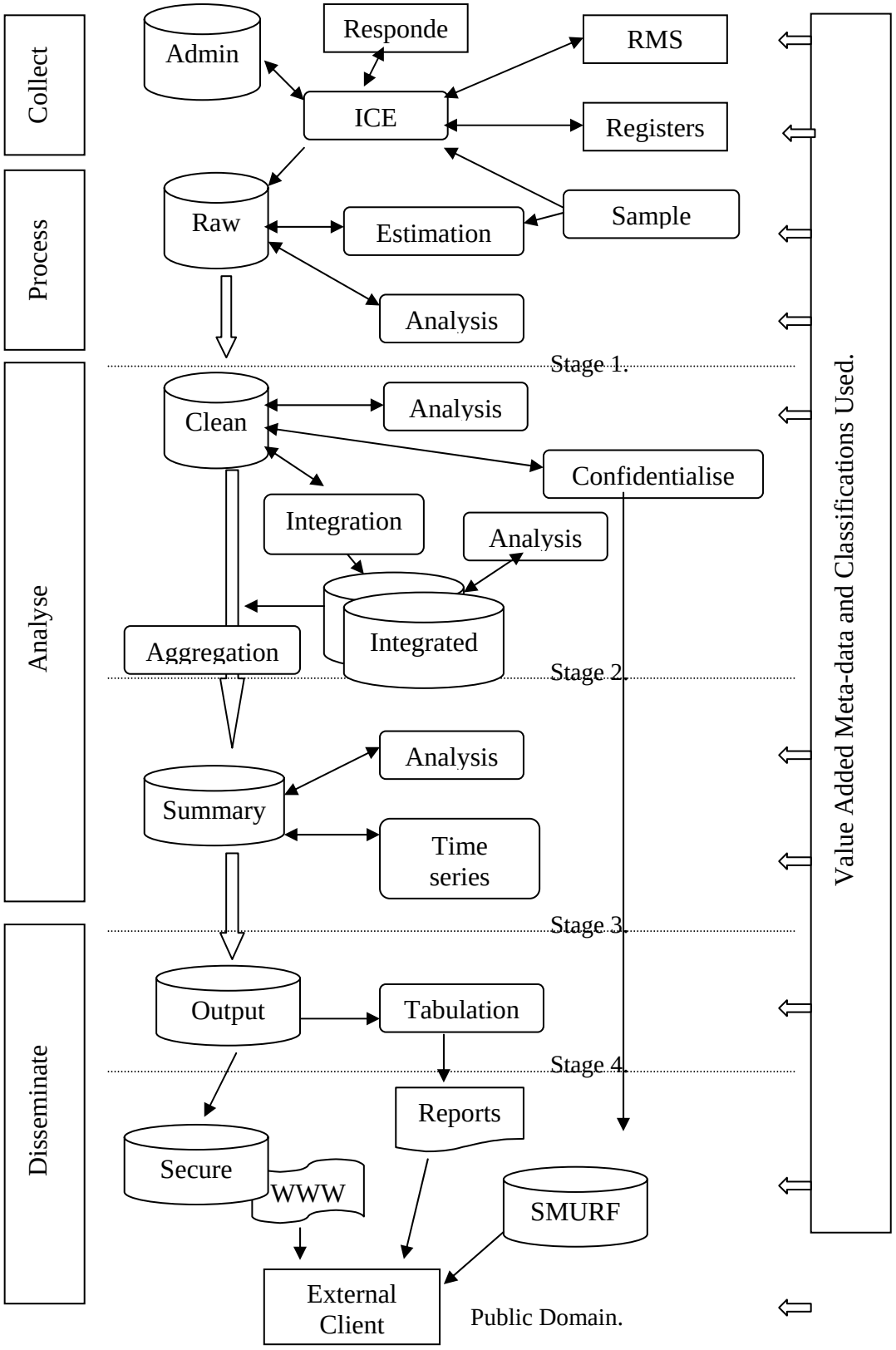


Figure 4: Data Flow through the Statistical Production Cycle

Statistics New Zealand's business process model is described in the following files. In order to get to the generic Business Process Model (gBPM) extensive consultation was undertaken across the organisation and many business process modelling sessions undertaken. Having undertaken the consultative processes it was agreed that the gBPM will fit the direction that Statistics NZ was to take going forward. We are working on some minor variations of the gBPM to better fit Data Integration and Feasibility Studies. The data flow is on the Figure 4. gBPM diagram is attached at the end of the document.

Mapping between CMF Lifecycle Model & Statistics NZ gBPM:

CMF Lifecycle Model	Statistics NZ gBPM (sub-process level)
1 - Survey planning and design	Need (sub-processes 1.1 - 1.5) + Develop & Design (sub-processes 2.1 - 2.6)
2 - Survey preparation	Build (sub-processes 3.1 - 3.7) + Collect (sub-process 4.1)
3 - Data collection	Collect (sub-processes 4.2 - 4.4)
4 - Input processing	Collect (sub-process 4.5) + Process (sub-processes 5.1 - 5.3)
5 - Derivation, Estimation, Aggregation	Process (sub-processes 5.4 - 5.7)
6 - Analysis	Analyse (sub-processes 6.1 - 6.6)
7 - Dissemination	Disseminate (sub-processes 7.1 - 7.5)
8 - Post survey evaluation	Not an explicit process, but seen as a vital feedback loop.

Metadata groups

The MetaNet Reference Model^{TM} (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.

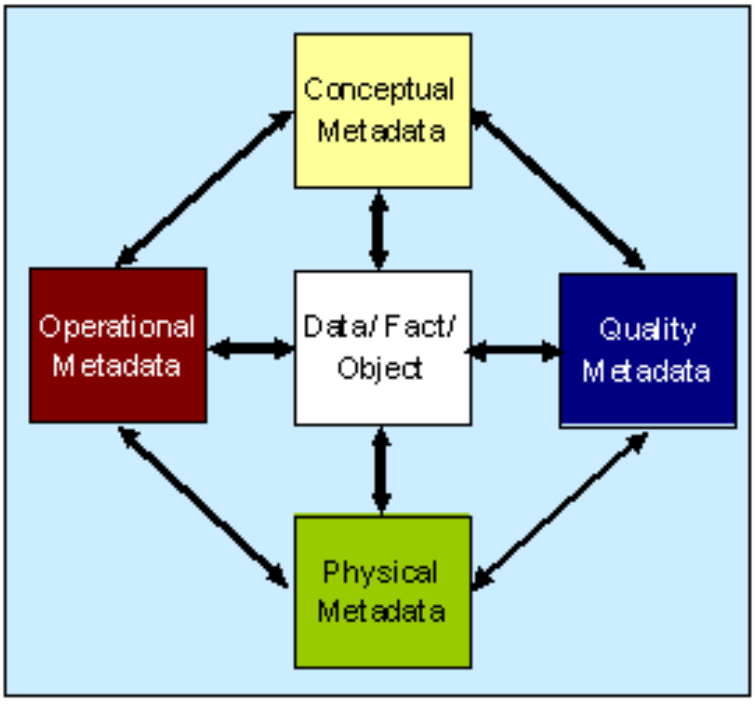


Figure 4: Metadata groups and relationship to data

3. STATISTICAL METADATA IN EACH PHASE OF THE STATISTICAL CYCLE

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
- 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)	Quality Issues	Risks and Challenges
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)	Quality Issues	Risks and Challenges
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)	Quality Issues	Risks and Challenges
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)	Quality Issues	Risks and Challenges
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)**	**Quality Issues**	**Risks and Challenges**
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)	Quality Issues	Risks and Challenges
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)	Quality Issues	Risks and Challenges
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 Issues	Risks and Challenges
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. For each phase provide:

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)	Quality Issues	Risks and Challenges
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 Issues	Risks and Challenges
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)	Quality Issues	Risks and Challenges
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 Issues	Risks and Challenges
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 Issues	Risks and Challenges
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

Group	Description	Example	Source (if reused)	Quality Issues	Risks and Challenges
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.			

Metadata Mapping to gBPM Sub process level - Breakdown by MetaNet metadata types:

Metadata Types	Need	Develop and Design	Build	Collect	Process	Analyse	Disseminate
General Comments	If no previous data is available, this stage will primarily deal with conceptual metadata to define the need.	This stage involves the refinement of conceptual metadata, and the design of operational and physical metadata.	At the end of this stage, conceptual metadata, operational metadata and physical metadata should all be created. Some aspects of quality metadata may also be defined.	During the collect phase, the primary focus will be on the physical metadata and quality metadata.	The process stage will mostly utilise the operational metadata to process the data. The outcome of this will be more quality metadata.	Analysis should utilise all types of metadata however the main focus will be how well the physical data represents the conceptual metadata.	The metadata types used in dissemination will be dependent on the audience. For instance, public audiences may mostly be concerned with conceptual metadata, while technical audiences will also have an interest in quality and operational metadata.
Conceptual Metadata - describes the basic idea (concept) behind the metadata object	Determine the concepts to be studied. At this stage, includes broad concepts such as global variables, themes , subject areas, statistical objects types	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.	May need to refine/adapt concepts due to feedback from testing or errors detected.	As data is collected it will be allocated against concepts. Some details about the relevant concepts may be used in respondent management strategies.	Conceptual metadata are used to classify and code open responses.	In the analysis stage the processed data will be used to analyse the concepts defined in the need and develop/design stages.	All audiences will have an interest in the conceptual metadata in order to understand what the data represents.
Operational Metadata - are the metadata required to view the data from an operational point of view (eg record variables, matrix operations, statistical process). This includes all processes and configuration.	May require analysis to determine what data is already available and how it was collected and processed.	This includes designing the sample methodology, collection methodology and the statistical processes. Includes concepts such as statistical process, process implementation, data collection methodology. Also need to design quality characteristics and processes.	Build and configure storage structures, collection instruments, processing requirements. Includes concepts such as record variables, record types, matrix operations.	Utilise the collection processes outlined in the 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.	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.	Operational Metadata will provide technical audiences with further information on how the data was collected and processed.

Metadata Types	Need	Develop and Design	Build	Collect	Process	Analyse	Disseminate
Quality Metadata - are the metadata for a particular instance eg response rates, status, weighting, versions.	Review quality metadata for previous collections to further understand appropriateness for meeting the need.	Quality metadata from previous collections may be used in the design of sample and collection methodologies.	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.	Collect data for each instance of the survey. Quality metadata will be populated based on the collection instance.	Further quality metadata will be populated at this stage based on the processes applied.	Quality metadata will be used at this stage to assess how well the data represents the concepts outlined in the needs stage.	Technical audiences and professional audiences may be interested in the quality metadata in order to understand the characteristics of each instance of data collection.
Physical Metadata - includes the physical, unique characteristics of the data which cannot be separated eg server locations, data base.	Locate current data through physical metadata eg datasets, server locations, etc	May begin to define/ identify where data will be stored. May also require physical metadata from previous collections to determine the best	Determine the physical location where the data will be stored ie servers, software, access packages.	Physical datasets will be populated with data at this stage.	If processed data is stored in different locations, new physical metadata will be defined.	Physical metadata will be required to locate the data, and any additional data for comparisons.	Physical metadata may be required when locating data in response to customer queries.

4. SYSTEMS AND DESIGN ISSUES

Overall description of 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.

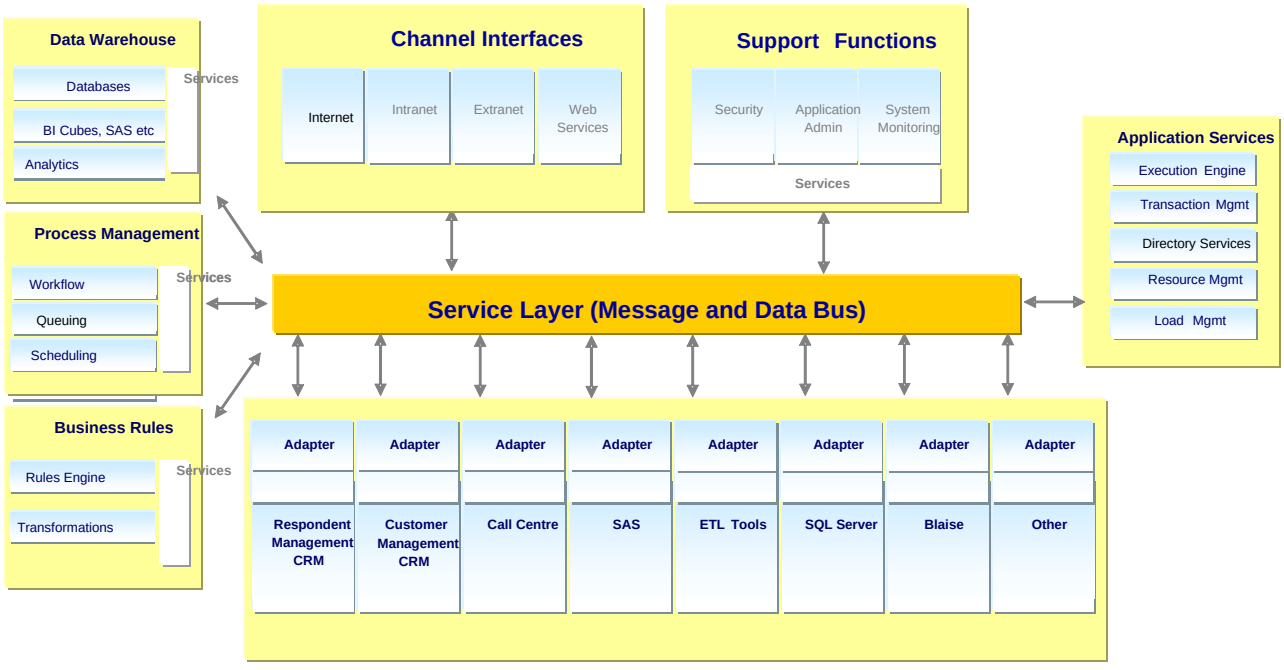
As outlined above, the BmTS has three core deliverables:

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 communization 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 center), tools (SAS, ETL, Blaise) and databases (SQL Server).



Currently within Statistics New Zealand, a project 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 Information Portal project. 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 current version of IDE is based on metadata prepared with special IDE spreadsheet tool, which will be replaced by the new metadata capturing tool.

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 developed, 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.

Question/Concept 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). Potentially the component will also be developed in alignment with questionnaire design tools to allow for the drag and drop of standard questions.

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 application for classification management. There is a plan to upgrade the platform to .NET/SQL and enhance for integration within the new SOA architecture.

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.

Document One - 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 Framework – 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 versionning, notification systems and recording of consultation. Following the population of the Standards

Framework, further investigation will be conducted regarding the integration with other components within the metadata environment.

Standards and models used

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 the 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. It is envisaged that this will form part of stage 2 of the metadata project.

5. ORGANIZATIONAL AND CULTURAL ISSUES

To ensure that metadata is relevant and useful, it is critical that the metadata project understands its audience. A full audience analysis identified several broad user groups - Public, Professional, Technical, System. The outcome of that analysis has been carried forward [12].

User Group Name	Use of the solution	Type
External		
Government	Request new statistics about NZ Use statistics about NZ	Professional
Public	Request new statistics about NZ Use statistics about NZ	Public
External Statisticans (incl. Intl Orgs)	Request new statistics about NZ Use statistics about NZ Interact with data in Statistics NZ environment	Technical
Internal Users		
Statistical Analysts	Interact with data in Statistics NZ environment Utilise workflows to manage data processing and assembly? I would have thought we need to have an operations role for this. What do you mean by assembly here? Analyse data through standard tools and processes. Maximise the re-usability of data. Work with SM, TAS, 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.)	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	Utilise workflow to manage data processing and assembly. Work with SM & TAS to design and configure workflows Monitor progress of business processes to produce statistical outputs. Manage Automated monitoring of end-to-end process Process transparency and traceability	Professional
Data Managers / Custodians / Archivists	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

User Group Name	Use of the solution	Type
Statistical Methodologists	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.)	Interact with data in Statistics NZ environment Context metadata supports data needs	Technical Professional
Architects - data, process & application	Specify standardised generic solutions. Plan for progressive architecture Isolate IT dependencies with defined integration and interoperability mechanisms	Technical Professional System
Respondent Liaison	Provide a user friendly service to external data providers.	Professional
Survey Developers	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	Propose and adopt metadata standards and business rules Manage and monitor the core corporate metadata assets	Professional Technical
Project Managers & Teams	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	Reduce development time and costs Control maintenance/support costs Define capability requirements	Professional Technical
Product Development and Publishing	Supply of Context metadata supports data needs	Professional
Information Customer Services	Maintain customer details Capture customer request for statistics Search and discover existing dataset for customer Report progress on logged customer issues	Professional

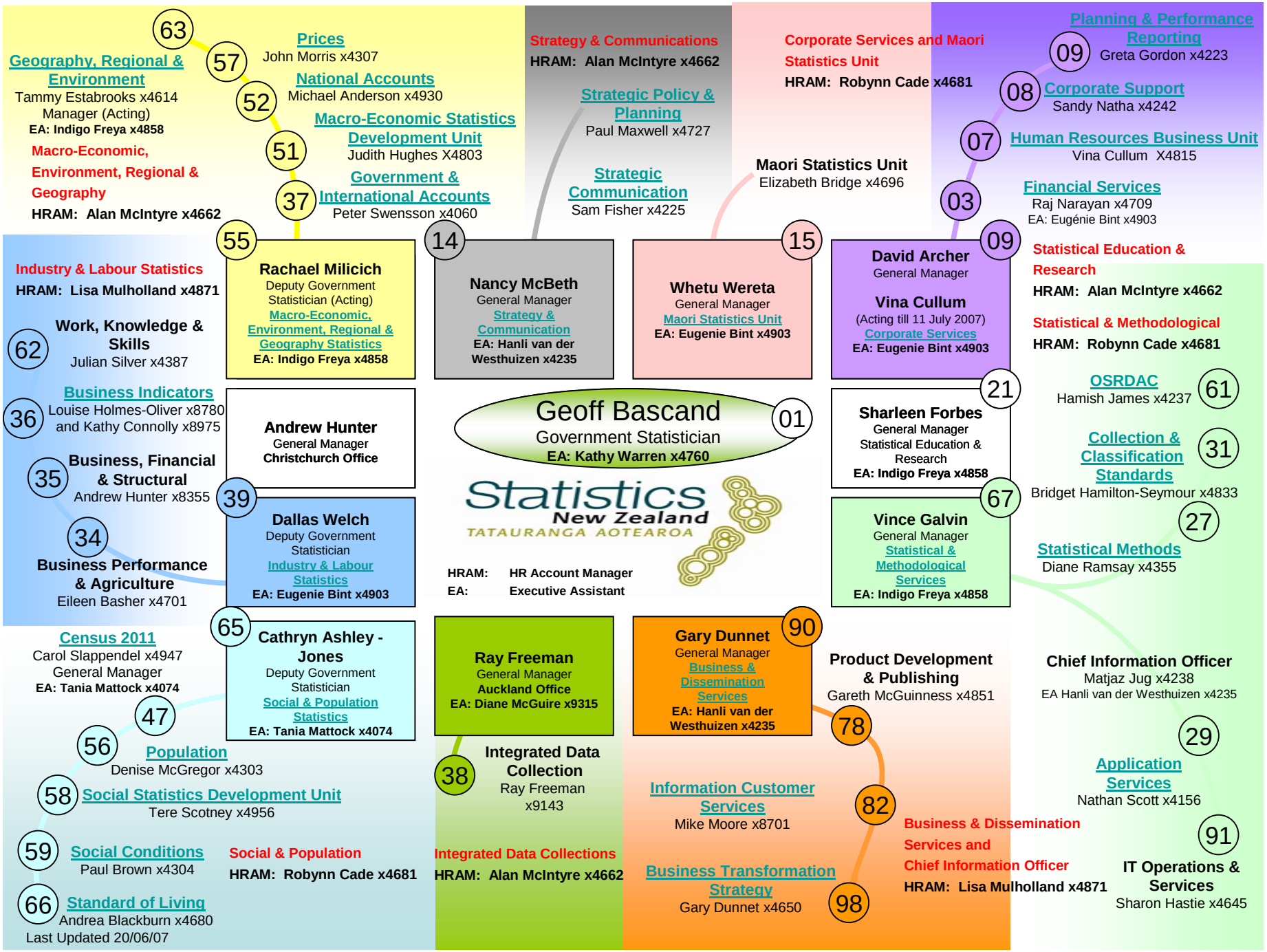
Notes: This is the list of actors, i.e. roles using the Solution.

6. ATTACHMENTS & LINKS

Details	Version	Link
BmTS Programme Charter	FINAL	attached
Metadata White Paper	v5.1 FINAL	attached
Metadata Broad Logical Design	v2.1 FINAL	available on request
Business Case for the Metadata and Information Portal Programme	v6.0 FINAL	available on request
Stage Two Metadata Vision Document	v0.2 DRAFT	available on request
Process Mapping	DRAFT	available on request
End to End Model White Paper	v3.0	available on request
Input Data Environment Whitepaper	v2.1	available on request
Output Data Environment Whitepaper	v1.1	available on request
gBPM	v1.2	available on request

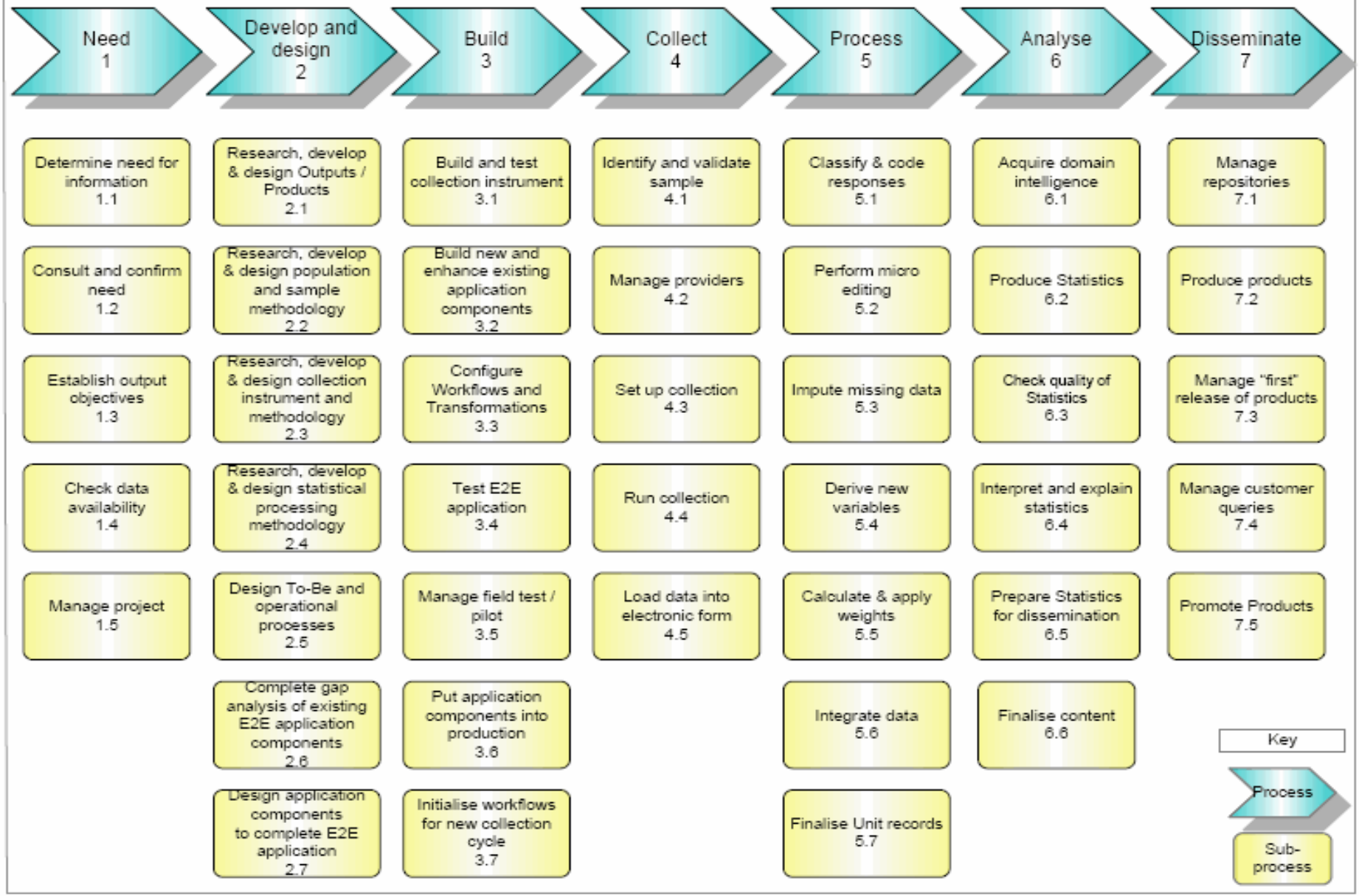
APPENDICES

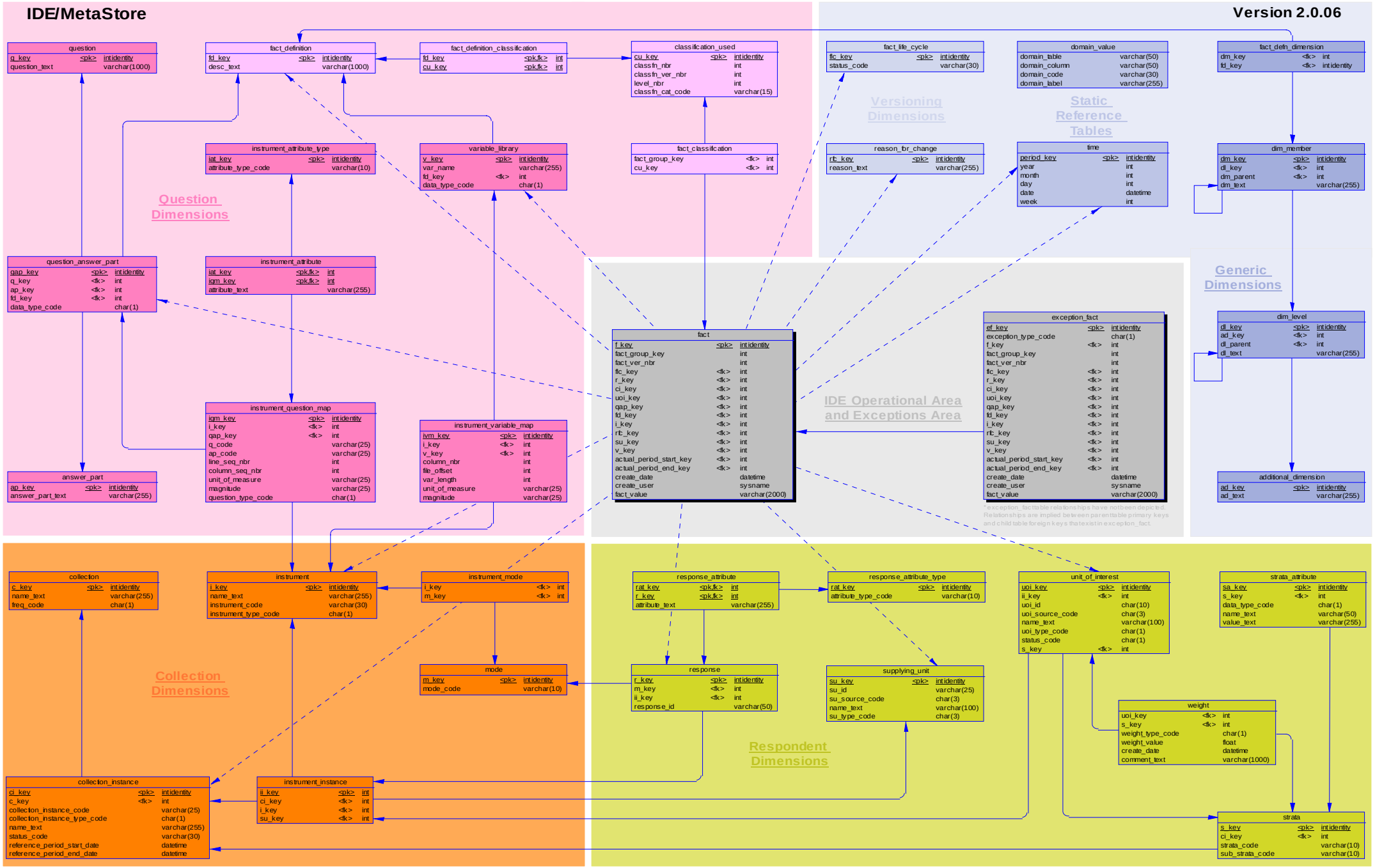
Organisational Chart
Generic Business Process Model
IDE/MetaStore data model V2.1


Geography, Regional & Environment
Tammy Estabrooks x4614
Manager (Acting)
EA: Indigo Freya x4858
Macro-Economic, Environment, Regional & Geography
HRAM: Alan McIntyre x4662
63
57
Prices
John Morris x4307
52
National Accounts
Michael Anderson x4930
51
Macro-Economic Statistics Development Unit
Judith Hughes X4803
37
Government & International Accounts
Peter Swensson x4060
Strategy & Communications
HRAM: Alan McIntyre x4662
Strategic Policy & Planning
Paul Maxwell x4727
Strategic Communication
Sam Fisher x4225
Corporate Services and Maori Statistics Unit
HRAM: Robynn Cade x4681
Maori Statistics Unit
Elizabeth Bridge x4696
Planning & Performance Reporting
Greta Gordon x4223
09
08
Corporate Support
Sandy Natha x4242
07
Human Resources Business Unit
Vina Cullum X4815
03
Financial Services
Raj Narayan x4709
EA: Eugénie Bint x4903
Industry & Labour Statistics
HRAM: Lisa Mulholland x4871
55
Rachael Milicich
Deputy Government Statistician (Acting)
Macro-Economic, Environment, Regional & Geography Statistics
EA: Indigo Freya x4858
14
Nancy McBeth
General Manager
Strategy & Communication
EA: Hanli van der Westhuizen x4235
15
Whetu Wereta
General Manager
Maori Statistics Unit
EA: Eugenie Bint x4903
09
David Archer
General Manager
Vina Cullum
(Acting till 11 July 2007)
Corporate Services
EA: Eugenie Bint x4903
Statistical Education & Research
HRAM: Alan McIntyre x4662
Statistical & Methodological
HRAM: Robynn Cade x4681
62
Work, Knowledge & Skills
Julian Silver x4387
36
Business Indicators
Louise Holmes-Oliver x8780
and Kathy Connolly x8975
35
Business, Financial & Structural
Andrew Hunter x8355
34
Business Performance & Agriculture
Eileen Basher x4701
Andrew Hunter
General Manager
Christchurch Office
39
Dallas Welch
Deputy Government Statistician
Industry & Labour Statistics
EA: Eugenie Bint x4903
Geoff Bascand
Government Statistician
EA: Kathy Warren x4760
01
Statistics New Zealand
TATAURANGA AOTEAROA
HRAM: HR Account Manager
EA: Executive Assistant
21
Sharleen Forbes
General Manager
Statistical Education & Research
EA: Indigo Freya x4858
67
Vince Galvin
General Manager
Statistical & Methodological Services
EA: Indigo Freya x4858
OSRDAC
Hamish James x4237
61
Collection & Classification Standards
Bridget Hamilton-Seymour x4833
31
27
Statistical Methods
Diane Ramsay x4355
Census 2011
Carol Slappendel x4947
General Manager
EA: Tania Mattock x4074
65
Cathryn Ashley - Jones
Deputy Government Statistician
Social & Population Statistics
EA: Tania Mattock x4074
Ray Freeman
General Manager
Auckland Office
EA: Diane McGuire x9315
90
Gary Dunnet
General Manager
Business & Dissemination Services
EA: Hanli van der Westhuizen x4235
Product Development & Publishing
Gareth McGuinness x4851
78
Chief Information Officer
Matjaz Jug x4238
EA Hanli van der Westhuizen x4235
29
Application Services
Nathan Scott x4156
47
56
Population
Denise McGregor x4303
58
Social Statistics Development Unit
Tere Scotney x4956
38
Integrated Data Collection
Ray Freeman
x9143
Information Customer Services
Mike Moore x8701
82
Business & Dissemination Services and Chief Information Officer
HRAM: Lisa Mulholland x4871
91
IT Operations & Services
Sharon Hastie x4645
59
Social Conditions
Paul Brown x4304
66
Standard of Living
Andrea Blackburn x4680
Last Updated 20/06/07
Social & Population
HRAM: Robynn Cade x4681
Integrated Data Collections
HRAM: Alan McIntyre x4662
Business Transformation Strategy
Gary Dunnet x4650
98

Statistical Business Process Model

Statistics New Zealand
TATAURANGA AOTEAROA




IDE/MetaStore
Version 2.0.06
question
q_key <pk> intidentity
question_text varchar(1000)
fact_definition
fd_key <pk> intidentity
desc_text varchar(1000)
fact_definition_classification
fd_key <pk,fk> int
cu_key <pk,fk> int
classification_used
cu_key <pk> intidentity
classfn_nbr int
classfn_ver_nbr int
level_nbr int
classfn_cat_code varchar(15)
fact_life_cycle
flc_key <pk> intidentity
status_code varchar(30)
domain_value
domain_table varchar(50)
domain_column varchar(50)
domain_code varchar(30)
domain_label varchar(255)
fact_defn_dimension
dm_key <fk> int
fd_key <fk> intidentity
Versioning Dimensions
Static Reference Tables
instrument_attribute_type
iat_key <pk> intidentity
attribute_type_code varchar(10)
variable_library
v_key <pk> intidentity
var_name varchar(255)
fd_key <fk> int
data_type_code char(1)
fact_classification
fact_group_key <fk> int
cu_key <fk> int
reason_for_change
rfc_key <pk> intidentity
reason_text varchar(255)
time
period_key <pk> intidentity
year int
month int
day int
date datetime
week int
dim_member
dm_key <pk> intidentity
dl_key <fk> int
dm_parent <fk> int
dm_text varchar(255)
Question Dimensions
question_answer_part
qap_key <pk> intidentity
q_key <fk> int
ap_key <fk> int
fd_key <fk> int
data_type_code char(1)
instrument_attribute
iat_key <pk,fk> int
iqm_key <pk,fk> int
attribute_text varchar(255)
fact
f_key <pk> intidentity
fact_group_key int
fact_ver_nbr int
flc_key <fk> int
r_key <fk> int
ci_key <fk> int
uoi_key <fk> int
qap_key <fk> int
fd_key <fk> int
i_key <fk> int
rfc_key <fk> int
su_key <fk> int
v_key <fk> int
actual_period_start_key <fk> int
actual_period_end_key <fk> int
create_date datetime
create_user sysname
fact_value varchar(2000)
IDE Operational Area and Exceptions Area
exception_fact
ef_key <pk> intidentity
exception_type_code char(1)
f_key <fk> int
fact_group_key int
fact_ver_nbr int
flc_key <fk> int
r_key <fk> int
ci_key <fk> int
uoi_key <fk> int
qap_key <fk> int
fd_key <fk> int
i_key <fk> int
rfc_key <fk> int
su_key <fk> int
v_key <fk> int
actual_period_start_key <fk> int
actual_period_end_key <fk> int
create_date datetime
create_user sysname
fact_value varchar(2000)
* exception_fact table relationships have not been depicted. Relationships are implied between parent table primary keys and child table foreign keys that exist in exception_fact.
Generic Dimensions
dim_level
dl_key <pk> intidentity
ad_key <fk> int
dl_parent <fk> int
dl_text varchar(255)
additional_dimension
ad_key <pk> intidentity
ad_text varchar(255)
answer_part
ap_key <pk> intidentity
answer_part_text varchar(255)
instrument_question_map
iqm_key <pk> intidentity
i_key <fk> int
qap_key <fk> int
q_code varchar(25)
ap_code varchar(25)
line_seq_nbr int
column_seq_nbr int
unit_of_measure varchar(25)
magnitude varchar(25)
question_type_code char(1)
instrument_variable_map
ivm_key <pk> intidentity
i_key <fk> int
v_key <fk> int
column_nbr int
file_offset int
var_length int
unit_of_measure varchar(25)
magnitude varchar(25)
collection
c_key <pk> intidentity
name_text varchar(255)
freq_code char(1)
instrument
i_key <pk> intidentity
name_text varchar(255)
instrument_code varchar(30)
instrument_type_code char(1)
instrument_mode
i_key <fk> int
m_key <fk> int
response_attribute
rat_key <pk,fk> int
r_key <pk,fk> int
attribute_text varchar(255)
response_attribute_type
rat_key <pk> intidentity
attribute_type_code varchar(10)
unit_of_interest
uoi_key <pk> intidentity
ii_key <fk> int
uoi_id char(10)
uoi_source_code char(3)
name_text varchar(100)
uoi_type_code char(1)
status_code char(1)
s_key <fk> int
strata_attribute
sa_key <pk> intidentity
s_key <fk> int
data_type_code char(1)
name_text varchar(50)
value_text varchar(255)
Collection Dimensions
mode
m_key <pk> intidentity
mode_code varchar(10)
response
r_key <pk> intidentity
m_key <fk> int
ii_key <fk> int
response_id varchar(50)
supplying_unit
su_key <pk> intidentity
su_id varchar(25)
su_source_code char(3)
name_text varchar(100)
su_type_code char(3)
weight
uoi_key <fk> int
s_key <fk> int
weight_type_code char(1)
weight_value float
create_date datetime
comment_text varchar(1000)
Respondent Dimensions
collection_instance
ci_key <pk> intidentity
c_key <fk> int
collection_instance_code varchar(25)
collection_instance_type_code char(1)
name_text varchar(255)
status_code varchar(30)
reference_period_start_date datetime
reference_period_end_date datetime
instrument_instance
ii_key <pk> int
ci_key <fk> int
i_key <fk> int
su_key <fk> int
strata
s_key <pk> intidentity
ci_key <fk> int
strata_code varchar(10)
sub_strata_code varchar(10)