**Version circulated for user consultation as at 17/04/07**


# Metadata - White Paper

**Business model Transformation Strategy**
**Business Solutions Team**

**"Creating a New Business Model for the National Statistical Office of the 21st Century"**

## VERSION SUMMARY

| VERSION | SUMMARY / PURPOSE OF VERSION |
|---|---|
| v1 (5 October 2004) | |
| v 2 (13 October 2004) | Incorporated comments from C&S, OSS, Brian Pink and others |
| v 3 (November 2004) | |
| v 4 (2 December 2004) | Final comments incorporated |
| v 5 (April 2007) | Revised to incorporate definitions of Metadata Types from MetaNet Reference Model |

**Update By** : Craig Mitchell
**Updated** : April 2007
**Issued By** : Gary Dunnet
**Prepared By** : Rebecca Merrington
**Prepared** : October 2004
**Acceptance Authority** : Gary Dunnet

**Project Code** **:** 90010.00

## 00. Purpose of the paper

**Update April 2007**: Following further thinking and discussion in the development of the Metadata Environment for Statistics New Zealand, it was felt that the definitions of Metadata Types needed to be reviewed to remove potential confusion and improve their application. Current thinking in the Metadata Project is to align definitions to those proposed in the MetaNet Reference Model™. These definitions are discussed in the revised section '2. What is Metadata?' (paragraphs 16 - 20). Changes have also been made to the tables in the sections named '11. Which processes (within the Business Process Model) are 'touched' by Metadata Management?' (paragraph 48) and '13. Proposed Metadata Framework' (paragraph 53). Appendix A has also been included to show examples of how various types of metadata would be defined. Appendix B shows the differences between old and new defintions.

Comments regarding these changes can be made to Craig Mitchell, Gary Dunnet or any of the Business Solutions Team (BST).

This paper has been prepared for the metadata strategy owner, the General Manager BuDS (Business and Dissemination Services). The General Manager SMS (Statsitical and Methodological Services) is the key stakeholder.

This paper is designed to provide the reader with an understanding of the scope, issues and benefits of metadata management, as well as providing detail of some of the potential metadata infrastructure required. This paper is designed for staff at the SMPM level and above, however, the paper is open and available to anyone who has an interest in metadata management.

While you are reading this paper you are encouraged to think about metadata management and how it will impact on your work area and how your work area will interact with it. You are also encouraged to challenge the contents and ask questions of the author. Comments can be made to Craig Mitchell, Gary Dunnet or any other member of the Business Solutions Team (BST). Your direct feedback will be requested later this year when requirement collection begins.
**Note:** The *"ABS Metadata Strategy papers 1- 4"* , have been referred to and used in areas where applicable metadata issues and principles exist.

## 0. Executive summary

1. Statistics New Zealand has always recognised the benefits of robust metadata management, including its role in data integration and managing business processes. In fact, during the last large capability enhancement exercise, the IT Strategy, in the mid 1990's, Statistics New Zealand commissioned Bo Sundgren to provide a strategic direction paper on metadata management. From this work and within the IT Strategy project, Statistics New Zealand created metadata infrastructure like CARS and SIM. This infrastructure has performed exceptionally well in some areas (classifications management) and adequately (at best) in others. However, where there are issues with the existing infrastructure, the issue is not about the design of the infrastructure, but that this infrastructure is in part, poorly maintained, unstructured, difficult to access, unintegrated, passive and not carefully managed.

**What is the Business Model Transformation Strategy (BmTS), going to add to Statistics New Zealand's existing metadata management strategy?**

2. Essentially, 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**

3. From this list, it is obvious that one of the most exciting aspects is that in this environment metadata will be created, used and stored throughout the end-to-end business process. Metadata becomes the heart of "How we do things around here" as professional statisticians.

4. The use of metadata to drive business processes is not a new idea, but is yet to be fully realised in other statistical agencies. This is because the ideal requires an extensive understanding of the way an organisation operates, how it should operate, and the relationship between these processes and the enterprise's architecture. Under the BmTS, Statistics New Zealand is currently addressing both its business processes and the infrastructure needed to support these processes. As a result, end-to-end data/metadata management will become a reality.

5. Even though this work is pioneering, the Business Solutions Team (BST) are convinced that end-to-end metadata management is the only way to fully realise Statistics New Zealand's leadership of the official statistics system and need for standardisation, simplification and integration of statistical information.

**What is the Business Model Transformation Strategy (BmTS) going to add to Statistics New Zealand's existing metadata infrastructure?**

6. From investigation of international best practice, and an understanding of the Statistics NZ's existing metadata infrastructure, it is likely that the following will be developed:

1. **metadata requirements** inventory (collected from metadata users)
2. **roles and responsibilities charter** establishing metadata owners, populators, maintainers and regulators
3. a **policy framework** that provides the right incentives
4. comprehensive **metadata model** (to define the content of metadata)
5. standard **data item definition** framework/template
6. **thesaurus of statistical terminologies** which define and organise terms into a structure
7. **taxonomy of relationships** between the terminology structure and other metadata and data
8. **metadata storage plan** (this may require the establishment of separate repository/ies)
9. **Extract Transfer and Load (ETL) plan** establishing methods of metadata population
10. **metadata registry** (to organise the metadata defined in the model)
11. **metadata 'viewer'** that ensures that data users have easy access to relevant

metadata in a timely manner

12. **classifications repository** (CARS already exists)
13. **questions Library**
14. **metadata store**
15. **training and implementation plan**

**How does metadata management fit in with the other components of the BmTS?**

7. Metadata is the cornerstone of the other initiatives within the BmTS, including:

- establishing robust metadata management is essential to the structure of **Input** and **Output Data Stores**;
- understanding what metadata is required to be used, created and potentially stored throughout the end-to-end business process model and then ensuring systems deliver the required outcome (e.g. Project Contact - **Respondent Management**);
- as the driver of the statistical and operational **Transformations** and processes;
- providing access to information (**Information Portal**, **Analytical Environment** and Dissemination channels); and
- ensuring metadata is reused and updated from different systems/stages in the business process life cycle.

**Risks of developing an integrated and dynamic metadata management system**

8. While developing an integrated and dynamic metadata management system has many benefits, some of the risks and issues to be managed include, the size of the initiative and that from initial investigation there does not seem to be a one stop shop solution to metadata management.

**The Next Step**

9. The next step in introducing the new metadata management strategy is developing a broad logical design paper. This paper will articulate the broad design of the component, including the processes, collection points, design of repository and consider agreed standards .

## 1. Background

### 1.1 The Business model Transformations Strategy (BmTS)

10. The Business model Transformations Strategy (BmTS) is an initiative that aims introduce a new business model that will:

- Standardise
- Simplify
- Integrate

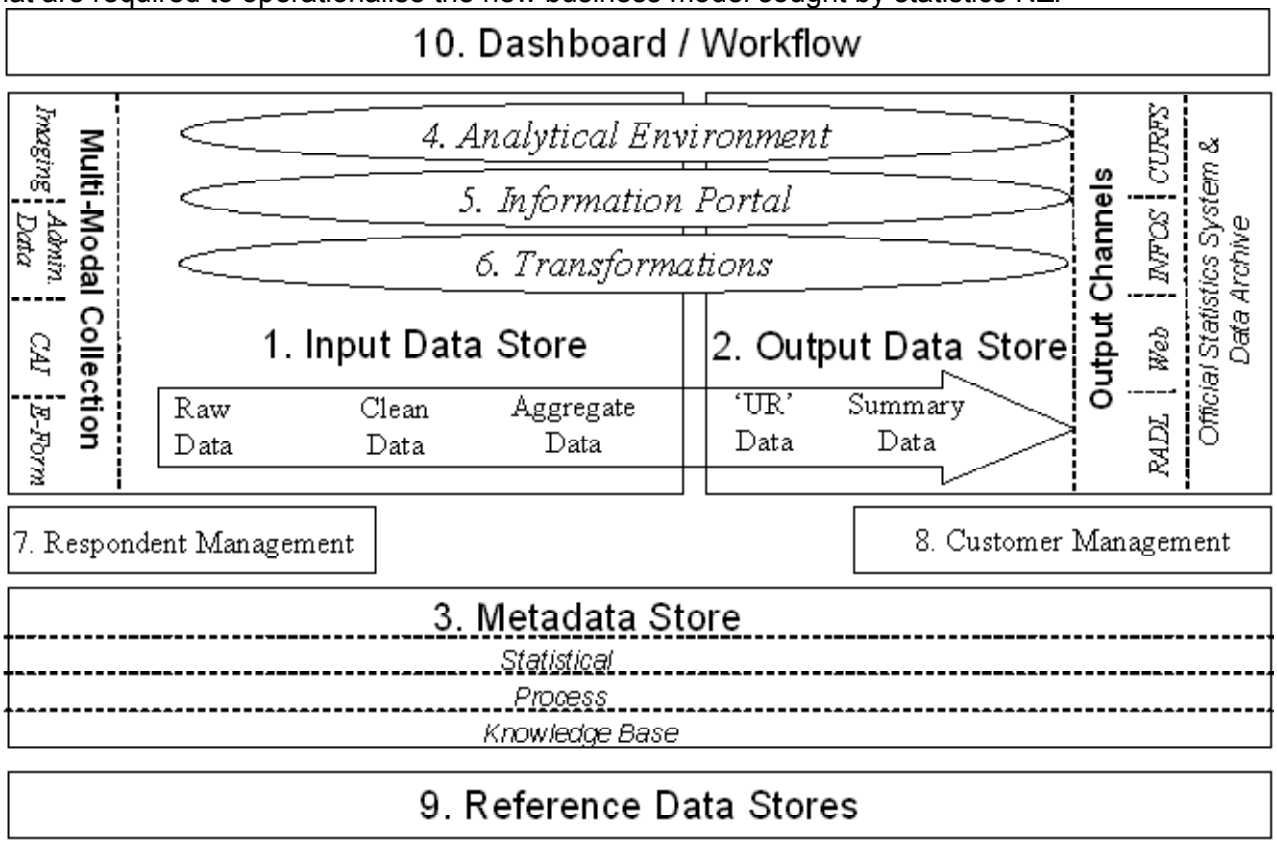
Statistics New Zealand's business processes and systems, in order to:

- improve quality
- reduce development lead times
- improve data access
- minimise respondent burden
- ensure confidentiality, security, privacy
- reduce costs

11. The BmTS is achieving this by:

- developing an end-to-end strategy that clarifies and standardises the way Statistics New Zealand's business processes and systems operate and integrate
- developing specific strategies for the development of 9 key business components that are required to operationalise the new business model
- supporting the introduction of these new business components.

12. The following diagram presents a conceptual view of the key business components that are required to operationalise the new business model sought by statistics NZ.



13. Although this diagram displays a Metadata Store, this paper focuses on metadata management. This is because the strategy for the Metadata Store (or more correctly Metadata Environment) is inherently the same as that for Metadata Management, the environment is just the tool to bring the metadata management strategy to fruition.

### 1.2 Metadata Vision

14. To enable Statistics New Zealand to meet the specific objectives of the BmTS, and its overarching sector outcome of

*"Governments, businesses, communities and citizens use official statistics on New Zealand's economy, society and environment to inform debate, research and decision making."*

By, implementing a robust metadata management strategy that:

- supports Statistics New Zealand created data in its transformation from data to information(end-to-end)
- allows Statistics New Zealand's internal and external audiences to have ease of access to relevant and timely metadata
- is recognised as the backbone of Statistics New Zealand's operation (and the success of the BmTS)
- supports New Zealand's Official Statistics Strategy

Thus:

- maximising the value of Statistics New Zealand's and other agencies' statistical information
- improving Statistics New Zealand's efficiency
- managing data as a corporate resource and in a life cycle. context

## 2. What is Metadata?

15. The Joint UNECE/EuroStat Work Session on Statistical Metadata, March 2002 define metadata in the following way:

"Metadata is information which is needed for the production and usage of statistical data. It provides information on definitional content; on the processes for the collection, processing, storage and dissemination of data; measures of quality; and information of IT and related issues." They also add that ... "metadata can be actively linked to data and to IT systems, or simply prepared as stand-alone descriptive files, although that consideration concerns the use rather than the definition of metadata." *Source: "Developing a Common Understanding of Standard Metadata Components - A Statistical Glossary"*

16. The ABS take this definition further:

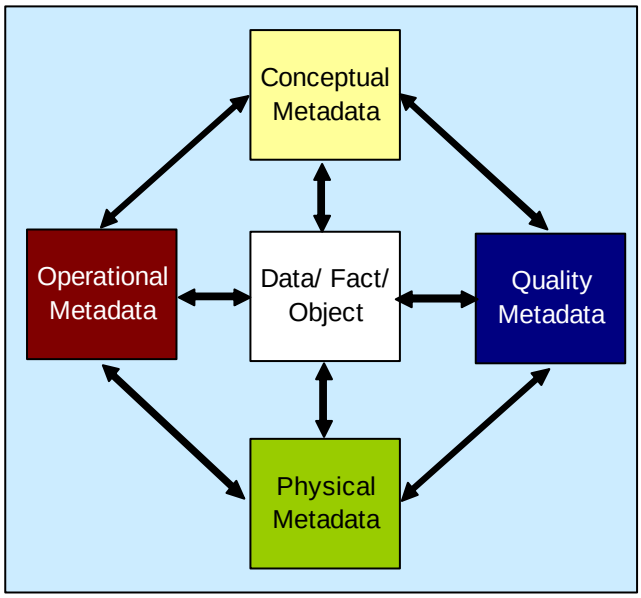
"Metadata simply means data about data, and refers to the definitions, descriptions of procedures, system parameters, and operational results which characterise and summarise statistical programs. Metadata may be passive (descriptive), i.e. the form of documentation which is used by agency staff, or may be active (prescriptive), i.e. determining the actions of automated surveys processes."

17. The MetaNet Reference Model™ (Version 2) categorises types of metadata in the

following way:

Conceptual Metadata describes the basic idea (concept) behind the metadata object eg 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 (eg 'income' collected in a particular survey).
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. 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 eg 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.

18. Each type of metadata describes a different way of viewing the metadata. The types of metadata are linked as shown in the following diagram:




The MetaNet Model itself covers various levels of detail. This diagram represents the highest level. At the lowest level, the model includes over 100 concepts each with their own characteristics. The model can easily become complex and covers multiple levels and dimensions, however it provides a flexible mechanism for viewing information at the level of detail required.

19. In the diagram above, the arrows represent access. Hence, from the facts or data, you can view any of the related metadata aspects depending on the need. However, this structure also allows for viewing information via the metadata. For instance, you can view the data item in a specific database (physical) through its quality metadata (eg non-response) and operational metadata (eg a non-response calculation), which are all based on conceptual definitions of what the data represents (eg organisational income). The relationships between metadata types are usually 1 to many (1:N), from top to bottom. That is, one concept may have several instances, and several means of operationalising, resulting in several physical outputs.

20. In the broadest sense, metadata are not exclusive to statistical data or statistical fact. Metadata can be used to describe and explain a range of statistical objects (eg a survey, a questionnaire, etc). The term 'object' is, therefore, included in the middle box to emphasise that the MetaNet Model can be applied to most statistical objects of interest, besides statistical data/facts. For example, the conceptual metadata of a questionnaire would identify the theme, the questions and the topics of interest; the operational metadata could be represented by questionnaire id numbers, line codes or routing instructions; quality metadata for a questionnaire could be the version used, the mode, the completeness of a response; and the physical metadata could be the medium used (eg electronic, paper), software, fonts and colours etc.

Appendix A contains further examples of how metadata may be defined using these definitions.

**3. What is Metadata Management?**

21. "Metadata Management refers to the **content, structure, and designs necessary to manage the vocabulary** and other metadata that **describes statistical data, designs and processes**. This includes the **development of metadata models** to define the content of metadata within some context, building **metadata registries** to organise the metadata defined in the model, developing **statistical terminologies** which define and organise terms into a structure within relationships (eg thesaurus) and **identifying the relationships between the terminology structure and other metadata and data**." (Bargmeyer and Gillman, METIS 2000)

22. In addition to this definition, in the Statistics New Zealand context, metadata management also includes:

- understanding what metadata is needed
- establishing metadata owners, populators, maintainers and regulators
- establishing methods of metadata population
- ensuring that data users have easy access to relevant metadata in a timely manner (this may mean managing potentially disparate metadata and providing access that disguises the complexity of data stores)
- determining appropriate technical storage environments

**4. What are the key benefits of Metadata Management?**

23. In his 1996 paper, Data and Metadata Management at Statistics New Zealand, Bo Sundgren outlined the key benefits of metadata management, including ensuring:

- internal and external users to be able to correctly interpret and analyse statistical data;
- producers to perform production processes in an efficient and consistent way;
- producers can continue to improve the performance of production and dissemination processes by learning from well documented experiences.

24. In addition to these benefits, robust metadata management provides a coordinated and consistent framework for managing data which:

- facilitates the life cycle management of data from end-to-end (from conception to archiving or destruction)

- facilitates statistical and data integration, allowing data confrontation
- improves collection efficiency
- facilitates reuse of data and the data structure
- simplifies data management
- drives computer systems
- ensures consistency
- provides access to consistent documentation of standard concepts, definitions, questions, techniques, business rules etc.
- widens access to corporate knowledge and ensures business continuity

**5. Common metadata standards**

25. Over recent years there has been a significant amount of work invested locally and internationally in developing metadata standards for statistical and non-statistical data. Essentially, these models are trying to:

- describe data in a generic manner
- ensure potential integration
- ensure international and cross agency cohesion

26. The predominant standards used throughout the world for statistical metadata are:

- the α-β-γ-τ model of Professor Bo Sundgren
- ISO / IEC 11179 Specification and Standardisation of Data Elements

27. Other standards that are relevant for Statistics New Zealand include:

- METANET Reference Model™
- eGIF Standard - a sub-set of Dublin Core e.g. NZGLS

28. These models are all highly relevant for Official Statistical Offices, the world over. An example is that the Australian Bureau of Statistics (ABS) has incorporated ISO 11179 in their Metadata Framework, to be used within the design of their Input Data Warehouse.

29. In terms of metadata management, it is likely that Statistics New Zealand will use one or a combination of these models to describe its data. The decision on the model to be used will be made at the logical design stage.

NOTE: As at January 2007, the MetaNet Reference Model has been selected as the base model for the development of Statistics New Zealands Conceptual Metadata Model.

30. The following documents discusses some of these models further. (Database: "Transition Programmes"; Subject: "Metadata 101 - The Theory"; Author: Gary Dunnet; Date Created: 08/05/2004; Date Modified: 08/07/2004)

**6. Who uses Metadata?**

31. In order to ensure that metadata is relevant and useful, it is critical that Statistics New Zealand understands its audience. In 2002, Gareth McGuinness undertook an Statistics New Zealand audience analysis project ("Full Audience Analysis Paper") and identified several broad user groups (Public, Professional, Technical). It is considered necessary to include a fourth group (System) to facilitate a move to a more 'active' metadata environment.

1. Public - the 'person on the street'
2. Professional - analysts and researchers
3. Technical - academics and statisticians
4. System - computers driving statistical or business processes

**The Public Audience -** comprises all members of society wanting information in a personal capacity or on behalf of a community organisation. Users might range from students completing school projects to interest groups wanting information and therefore metadata about their communities. This group also includes respondents to Statistics New Zealand's direct collections.

**The Professional Audience-** is composed of professionals in the public, private and educational sectors who use and analyse Statistics New Zealand statistics for the purposes of policy-making and other high-level endeavours.

**The Technical Audience** - comprises professionals, subject matter experts, academic and statisticians working in the area of statistical methodology or expert data.

**Systems Audience** - comprise automatic processes that use metadata to operate systems.

The Technical Audience is further divided into **Technical(Operational)** and **Technical(External)**.

**Technical (External)** - comprises non-Statistics New Zealand professionals, subject matter experts, academic and statisticians working in the area of statistical methodology or expert data.

**Technical (Operational)** - comprises mainly internal Statistics New Zealand staff and includes: **survey designers** evaluating existing data sources, investigating success and results from previous collections, assessing the impact of statistical methodologies, **respondent managers** assessing burden or response patterns by mode, **subject matter statisticians** evaluating results pre and post imputation, comparing results of different collections, **disseminators** assembling data for publication, **developers** writing applications that use metadata to drive processes.
This audience model is used in the metadata framework described later in this paper.

**7. What are the boundaries of Metadata Management**

32. As will be discussed later, metadata is created and used in every component of Statistics New Zealand's business process as well as in the compliance, staff and management processes.

33. In addition to the metadata created within Statistics New Zealand, in the future, other suppliers of Official Statistics will be required to provide metadata for their statistics in a way that will allow comparability between data sources.

34. However, even though it is obvious that the breadth of metadata needed and created is vast there are some boundaries to the scope of Statistics New Zealand's metadata management.

**Inclusions:**

- statistical processes and statistical metadata - "statistical metadata" means metadata about data collected or compiled for publication as statistics
- the registration of survey activity by other government departments
- metadata from other agencies (whether administrative by-product or directly collected survey data) should be considered within the scope of the strategy - an Statistics New Zealand objective is to encourage the use of standards to facilitate greater usefulness of government information
- a metadata stores (to encourage and maximise re-use of metadata)
- the Business Frame and similar infrastructure systems and repositories (these are included as these stores currently hold metadata and these holdings should be made transparent as part of a metadata system)
- management of metadata for any data warehouse
- new infrastructure including question library
- access control and security

**Exclusions:**

- The Metadata Strategy and Metadata Store does not support management processes e.g. Financial and Human Resources management. However, some infrastructure like the Corporate Directory is likely to be included within metadata management.

35. As you can see, the scope of metadata management is extensive. And the exclusion identified can easily be argued to be included if Statistics New Zealand is wishing to for example measure the financial and human resource implications of the BmTS.

## 8. Current Metadata Practice

### 8.1 Statistics New Zealand's current Metadata Environment

36. Over the past decade Statistics New Zealand has identified the benefits of good quality metadata for the organisation, outlined what is required and developed systems designed to provide that information. (See documents like the 1996 Data Management Strategy and the 1996 Information Management Project Plan).

37. However, in the late 1990's the emphasis was on creating unstructured, survey specific metadata stored within central, passive systems. This focus was perhaps a function of the technology available to the department and the prevalent thinking about metadata in the mid 1990's when the policy was developed. As a result, the systems that were created, were developed with specific functions in mind and operate independently of one another.

38. In an evaluation of metadata management in 2001, Elspeth Maclean found that although the department had developed a range of packages for storing metadata there were a number of areas where existing systems were failing to meet their objectives.

39. In 2004, much of the metadata infrastructure that was created in the late 1990's and evaluated in 2001 still exists.

### 8.2 Issues with and potential of the existing infrastructure

40. As a result of technology and prevalent thinking in the mid-1990's, Statistics New Zealand faces the following issues with its existing metadata infrastructure:

- metadata is not kept up to date
- metadata maintenance is considered a low priority
- metadata is not held in a consistent way
- relevant information is unavailable
- there is confusion about what metadata needs to be stored
- the existing metadata infrastructure is being under utilised
- there is a failure to meet the metadata needs of advanced data users
- it is difficult to find information unless you have some expertise or know it exists
- there is inconsistent use of classifications/terminology
- in some instances there is little information about data, where it came from, processes it has been under or even the question to which it relates

41. These issues have resulted in the corporate realisation that the significance of metadata to the effective performance of a National Statistics Office has been overlooked with the result that the current metadata infrastructure is not dynamic enough nor suitably integrated.

42. However, although some of the existing metadata infrastructure remains passive and unintegrated, the value of the existing metadata infrastructure should not be discounted (this existing infrastructure is critiqued in the attached document). Therefore, any metadata management strategy will, while addressing the issues highlighted, wherever appropriate re-use either the existing infrastructure or the principles/structure behind it.

43. CARS remains an integrated and useful piece of infrastructure. SIM although it has the potential to hold useful information remains unintegrated and passive.

### 8.3 External Metadata Environment

44. Many statistical and related agencies around the world are developing or have developed metadata management strategies. In some instances, these agencies have developed or are developing Metadata Stores.The Business Solutions Team (BST) is watching these developments carefully and has begun evaluating a few Metadata Stores (incl. ABS, ONS and OECD systems). A full evaluation of potential solutions will be available in the logical design paper. One of the key assessment criteria will be establishing the link between the solutions and appropriate statistical frameworks. (Database: "Transition Programmes"; Subject: "International Metadata Developments"; Author: Rebecca Merrington; Date Created: 23/09/2004)

## 9. Principles of Robust Metadata Management

45. From an assessment of internal and external best practice. The following principles of metadata management have been established. These principles 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)

These principles and the benefits they generate are included in this document .

**10. Metadata management - Inventory**

46. What metadata infrastructure is needed to create a successful metadata management strategy:

1. **metadata requirements** inventory (collected from metadata users)
2. **roles and responsibilities charter** establishing metadata owners, populators, maintainers and regulators
3. a **policy framework** that provides the right incentives
4. comprehensive **metadata model** (to define the content of metadata)
5. standard **data item definition** framework/template
6. **thesaurus of statistical terminologies** which define and organise terms into a structure
7. **taxonomy of relationships** between the terminology structure and other metadata and data
8. **metadata storage plan** (this may require the establishment of separate repository/ies)
9. **Extract Transfer and Load (ETL) plan** establishing methods of metadata population
10. **metadata registry** (to organise the metadata defined in the model)
11. **metadata 'viewer'** that ensures that data users have easy access to relevant metadata in a timely manner
12. **classifications repository** (CARS already exists)
13. **questions Library**
14. **metadata storage solution**
15. **training and implementation plan**

47. The Business Solutions Team and Project Services are responsible for creating all items, with the physical infrastructure (items 9-15) developed by aligning to 'lead' projects. These 'lead' projects include the OSRDAC, Social Statistics and Business Performance projects  . The range of 'lead' projects reflect the different stages of the end-to-end business process these projects represent.

**11. Which processes (within the Business Process Model) are 'touched' by Metadata Management?**

48. Metadata management and the Metadata Store supports all phases of Statistics New Zealand's agreed 'As is' business process model from end-to-end. A more specific discussion about the metadata produced and used at each stage of the business process is outlined below.

| 1. Need | 2. Design | 3. Build | 4. Collect | 5. Process | 6. Analyse | 7. Disseminate | 8.User Support |
|---|---|---|---|---|---|---|---|
| • Identify key trends nationally or internationally that may impact on the range of statistics we produce or the way we produce them<br>• capture requirements eg usage of data, quality requirements<br>• access existing data element concept definitions to clarify requirements | • capture constraints, basic dissemination plans eg products<br>• access information about similar collections<br>• capture design parameters that could be used to drive automated processes eg stratification<br>• capture conceptual and quality metadata about the collection<br>• reuse required data definitions, questions, classifications<br>• access operational metadata about similar collections to assist in design | • store operational metadata about selection process eg number in each stratum<br>• create operational metadata to drive selection process<br>• create required data definitions, questions, classifications | • capture metadata about the collection process<br>• use operational metadata about rules<br>• capture quality metadata eg quality metrics<br>• use operational metadata to drive collection processes | • capture metadata about the process<br>• use operational metadata, eg edit parameters, derivation definitions and imputation parameters<br>• capture quality metadata related to the processing instance | • use quality metadata eg quality measures<br>• use design parameters to drive estimation processes<br>• create information about quality assurance and sign-off of products<br>• use conceptual metadata to create products | • use operational metadata about customers to disseminate to relevant audiences<br>• use conceptual and quality metadata in response to customer queries<br>• use operational metadata to determine storage, archiving or disposal | • needed to support Search, Acquire, Analyse, Report<br>• capture re-use requirements<br>• detail on length of data life cycle.<br>• importance of data - fitness for purpose |
| Life Cycle Management Model (running end-to-end, from conception to archiving or destruction) | | | | | | | |

**How does metadata management fit in with the other components of the BmTS?**

49. Metadata is the cornerstone of the other initiatives within the BmTS, including:

- establishing robust metadata management is essential to the structure of **Input** and **Output Data Stores**;
- understanding what metadata is required to be used, created and potentially stored throughout the end-to-end business process model and then ensuring systems deliver the required outcome (e.g. Project Contact - **Respondent Management**);
- as the driver of the statistical and operational **Transformations** and processes;
- providing access to information (**Information Portal**, **Analytical Environment** and Dissemination channels); and
- ensuring metadata is reused and updated from different systems/stages in the business process life cycle.

**12. Metadata Custodians**

50. The following areas have specific responsibilities in terms of metadata population and key metadata repositories

- Input data providers
- Production Statisticians (subject matter statisticians, questionnaire and process designers)
- Software tools
- Statistical Designers (methodologists)
- Statistics New Zealand corporate

51. A key challenge of the strategy will be to promote the importance of metadata management to the metadata 'owners' and then to ensure that they are confident in their role as metadata populators and in some instances metadata store holders.

52. However, although metadata population and even storage may be the responsibility of these 'owners', metadata management and metadata registration will be coordinated centrally. It is recommended that Metadata management and metadata registration should be coordinated between the Statistical and Methodological Services (SMS) and the IMTS areas in general and the Classifications and Standards section and the Data Management unit specifically.

## 13. Proposed Metadata Framework

53. The following is the proposed metadata framework. This framework takes the definition of metadata established above and then discusses different attributes of the definition.

| Attributes of Metadata | Categories | Types of Metadata | | | |
|---|---|---|---|---|---|
| | | **Conceptual** | **Operational** | **Quality** | **Physical** |
| | | • statistical objects/population<br>• scope and coverage definition<br>• frameworks (eg SNA, POSS)<br>• concepts<br>• classifications, concodances<br>• value domains<br>• statistical terminology, glossary, thesaurus | • definition of collection instruments, question modules, instructions, CATI scripts, IFU scripts<br>• methodologies (incl imputation and estimation methods, sample method)<br>• business process information, eg approval process for dissemination<br>• business rules eg edit rules, derivation rules. coding rules<br>• definition of structure for variables and records when stored | • versions and status<br>• response rates, sample size and distribution<br>• sample frame information<br>• edit failures<br>• coding quality information<br>• quality metrics<br>• release and various approval sign-offs<br>• statistical release information | • datasets - structure, footnotes, titles etc<br>• metadata to describe micro and macro output datasets<br>• dissemination products<br>• software/tools used for access<br>• server locations<br>• data items |
| **Example** | Population Selection | | Y | Y | |
| **What is it used for?** | 1. Assess data quality (all the quality dimensions).<br>2. Understand how data is created.<br>3. Identify and locate data<br>4. Remember production steps.<br>5. Train new staff.<br>6. Tune and improve processes.<br>7. Run software tools.<br>8. Understand software tools. | 1, 5 | 2, 3, 5, 6, 7, 8 | 1, 2, 6 | 3, 7, 8 |
| **Where in the business process is this used?** | 1. Need<br>2. Design<br>3. Build<br>4. Collect<br>5. Process<br>6. Analyse<br>7. Disseminate<br>8. Archive and Destroy | All | 2, 3, 4, 5, 6, 7, 8 | 2, 4, 5, 6, 7, 8 | 2, 3, 4, 5, 6, 7, 8 |
| **Who defines it?** | 1. Statistical Designers (Technical Architecture, Methodologists, Questionnaire Design)<br>2. Input data providers<br>3. Production statisticians<br>4. Software tools | 1 | 1, 2, 3 | 2, 3 | 4 |
| **Who uses it? (in general, not related to access rights)** | 1. Public<br>2. Professional<br>3. system<br>4. Technical - External<br>5. Technical - Operational | All | 2, 3, 4, 5 | 2, 4 | 3, 5 |
| **How is it used?** | 1. Active<br>2. Passive | 2 | 1 | 1 | 1 |
| **How is it populated?** | 1. Automatic<br>2. Manual | 2 | 1, 2 | 1, 2 | 1 |
| **What format is used?** | 1. Structured<br>2. Unstructured | 1, 2 | 1 | 1 | 1 |

This framework is being used to develop the metadata logical design.

## 14. Risks of developing an integrated and dynamic metadata management system

54. While developing an integrated and dynamic metadata management system has many benefits, some of the risks and issues to be managed include:

| Risk | Comment/Mitigation |
|---|---|
| The potential scope of this initiative is large | The Business Solutions team manager to work closely with senior management to ensure appropriate staging, resourcing, ownership and budget |
| The ideal metadata/BmTS solution requires an extensive understanding of the way an organisation operates, how it should operate, and the relationship between these processes and the enterprise's architecture | Develop sound understanding of Statistics New Zealand's processes, data flows and stores |
| The use of dynamic metadata driven business processes may dilute corporate understanding of its business | Metadata will be accessible to all, rather than hidden in applications. There will be more eyes on the processes to ensure they are working as designed |
| In order to achieve all of the benefits identified, much of the BmTS changes need to have been made as well | Although the BmTS initiatives need to work together, there are gains to be made along the way e.g. establishing a standard data definition could realise some of the gains of the full data warehouse solution |
| There will be an obligation to manage metadata in a structured and central way. This will increase internal metadata compliance in some instances | One of the main principles of metadata management is to have metadata updated by the system as often as possible. There will need to be sound processes and clear ownership around the centralised management of metadata |
| Insufficient internal capability | This initiative may require external capability |
| Organisation resistant to the proposed changes -issue with metadata population is that it is often managed and populated by uninterested parties (metadata not seen as important when you are getting on with the job) | Ensure organisation recognises the overall benefit of metadata management |
| From initial investigation there does not seem to be a one stop shop solution to metadata management | Identify off the shelf infrastructure to be used as much as possible when the solution is being designed |

**Note:** this is not a comprehensive risk/mitigation plan, further work will be completed at the logical design stage.

## 15. Conclusion

55. Robust metadata management is the key to the success of the BmTS. Collecting comprehensive requirements and learning from past and international experience will ensure that Statistics New Zealand develops a robust metadata management system.

## 16. What's next

56. The next phase of the metadata management strategy is:

- collect and articulating high-level user needs
- deciding on a metadata taxonomy to be used within Statistics New Zealand
- evaluate of several existing metadata stores
- outline a sensible, usable, deliverable solution
- detail how metadata will interact with business processes and other components of the BmTS
- developing a proposed time line that fits with Statistics New Zealand's business requirements and their interdependencies
- confirming the content requirements of the broad logical design paper
- presenting the broad logical design paper

Note: the Broad Logical Design Paper was delivered in January 2004  , however amendments will also be made to ensure that the new definitions carry through both papers.

## 17. Contact Information

57. If you want to know more you can contact Craig Mitchell on 03 964-8831 or Gary Dunnet on - 04 931-4650.

## Bibliography

(Database: "Corporate Documents"; Subject: "96/059 - Data Management Strategy"; Author: Dallas Welch; Date Created: 20/06/1996; Date Modified: 27/08/1999)

(Database: "Transition Programmes"; Subject: "ABS Metadata Strategy - Part 1,2,3 & 4"; Author: Gary Dunnet; Date Created: 13/05/2004)

Australian Bureau of Statistics, Using ISO/IEC 11179 to help with Metadata Management Problems, Joint UNECE/Eurostat/OECD work session on statistical metadata (METIS), 2004

(Database: "Transition Programmes"; Subject: "Business model Transformation Strategy - Project Plan - DRAFT v4"; Author: Gary Dunnet; Date Created: 14/07/2004; Date Modified: 04/08/2004)

Developing a Common Understanding of Standard Metadata Components - A Statistical Glossary, Joint UNECE / EuroStat Work Session on Statistical Metadata, March 2002.

Gillman, Dan, Metadata Registries, Data Harmonization, and Maximizing Use of a Warehouse, US Bureau of Labor Statistics, 2004

(Database: "Corporate Documents"; Subject: "IRP - IRB Papers, IRB 1996/54 - Information Management Project"; Author: Margaret O'Sullivan; Date Created: 26/07/1996; Date Modified: 03/09/1999)

(Database: "Corporate Documents"; Subject: "DTM - Data Management, Data and metadata management at"; Author: Sandra McDonald; Date Created: 05/07/1996)

(Database: "Transition Programmes"; Subject: "Existing metadata infrastructure benefits, issues, conclusions"; Author: Rebecca Merrington; Date Created: 28/09/2004; Date Modified: 01/10/2004)

(Database: "Transition Programmes"; Subject: "International Metadata Developments - Draft"; Author: Rebecca Merrington; Date Created: 23/09/2004; Date Modified: 08/10/2004)

(Database: "Transition Programmes"; Subject: "Metadata 101 - The Theory"; Author: Gary Dunnet; Date Created: 08/05/2004; Date Modified: 08/07/2004)

(Database: "Classifications and Standards"; Subject: "Metadata Management Review- by Elspeth"; Author: Chris Toohey; Last Editor: Catherine Flanagan; Date Created: 29/10/2001; Date Modified: 28/04/2003)

McGuinness, Gareth, Full Audience Analysis Paper, Statistics New Zealand, 2002

(Database: "Classifications and Standards"; Subject: "New Framework for Metadata Development and Dissemination"; Author: Stephanie Prosser; Last Editor: Christine Stephenson; Date Created: 16/10/2002; Date Modified: 13/04/2004)

Sundgren, Bo, Statistical Metadata - a tutorial, R&D Report, Statistics Sweden, 1995:5

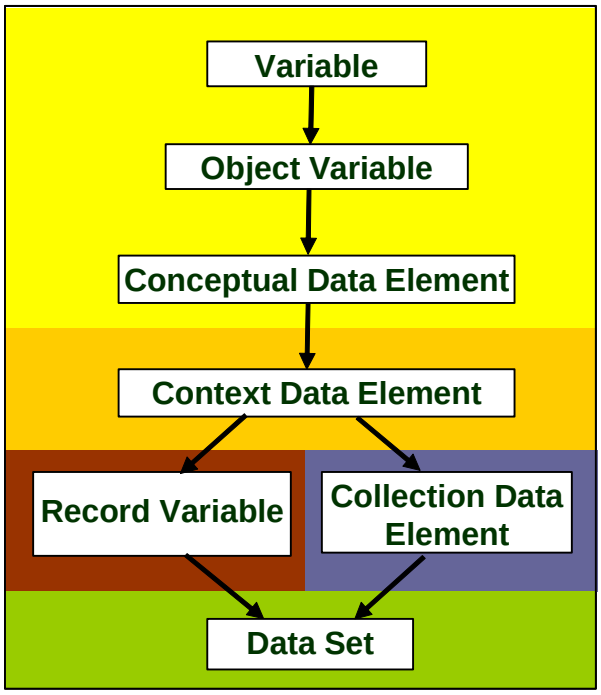
Karge, Reinhard, MetaNet Reference Model - Version 2, 2003

Word document circulated externally - 20th October 2004

Word copy of Metadata White Paper.doc

## Appendix A - Two examples of defining metadata using MetaNet Reference Model™ definitions

### Example 1 - From Variable to Data Set




In the diagram above, the coloured boxes represent the following metadata types:

- yellow represents conceptual metadata (context free)
- orange represents conceptual metadata (context related)
- red represents operational metadata
- blue represents quality metadata
- green represents physical metadata

| Concept | Description and Example |
|---|---|
| **Conceptual Metadata** | |
| Variable | The variable describes a single topic of interest eg ***income, age, activity, occupation*** . This concept is also called the global variable. |
| Object Variable | When the variable is combined with a statistical object type (ie a person, household, enterprise) it becomes an object variable.<br><br>Examples of object variables would be ***age of a person, income of an enterprise, etc*** . |
| Conceptual Data Element | In order to measure the object variable, different measurement criteria are usually defined. This concept is considered to be defined in the value domain. This will include linkages between other concepts such as classifications, time periods and measure units. When combined with the object variable, this becomes our conceptual data element.<br><br>An example of a conceptual data element would be the ***income of an enterprise in the manufacturing industry (as defined by ANZSIC06) in NZ dollars for the financial year ended 2006*** . |

| | |
|---|---|
| Context Data Element | To define the context data element a conceptual data element is given context by associating it with a study (ie Programme of Social Statistics (POSS), Annual Enterprise Survey (AES), Screen Industry Survey).<br><br>Hence an example of a context data element is ***income of an enterprise in the manufacturing industry (as defined by ANZSIC06) in NZ dollars for the financial year ended 2006 as collected in AES*** . |
| **Operational Metadata** | |
| Record Variable | Once the concepts behind the data element are defined, the detail of how it will be recorded needs to be identified/ designed. The record variable outlines the format and structure of the data element when it is stored (this can be likened to a column in a table or matrix).<br><br>The record variable includes the identification of short names, presentation titles (eg column headings), size (maximum length), etc. For example, our record variable for the above data element may identify the following: ***short name = ent_inc, size = 15, numeric, etc.***<br><br>Record variables are usually combined with record types to create a matrix. Operations and processes may also be applied to the record variable to transform the data. All these concepts are considered operational metadata. |
| **Quality Metadata** | |
| Data Collection Element | Upon collection of the data, there will be metadata specific to each instance. The data collection element defines the metadata for a particular instance of collection which is related to our data element of interest.<br><br>The data collection element may include the response rates and other quality measures, questionnaire versions, coverage etc. For example, our data collection element may identify the following: ***response rate = 85%, questionnaire version = AES 05/06 AF/MW/01, Coverage = Enterprises on the Business Frame*** .<br><br>The outcome of operations and processes being applied are also considered quality metadata as they will be specific to the instance of a collection. |
| **Physical Metadata** | |
| Dataset | All the information about a dataset will be defined by the links to conceptual, operational and quality metadata. However, there will always be some information which is specific to the dataset itself. This is the physical metadata.<br><br>This will include information such as server locations, database names, size of the dataset, the software used to access the dataset etc. For example, our dataset metadata may identify the following: ***server = syb03, size= 3MB, access type = SAS via the BmTS Dashboard*** . |

**Example 2 Types of Metadata used in generic business processes for the design of a new study.**

This example is based on the development of a new survey. It details how different types of metadata may be used in different stages of a survey collection. Please note: This is only meant as a guideline and is not a complete list of the metadata used in each circumstance.

| 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** | 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** | May require analysis to determine what data is already available and how it was collected and | This includes designing the sample methodology, collection methodology and the | Build and configure storage structures, collection instruments, processing | Utilise the collection processes outlined in the operational metadata. | Further statistical processes are used in order to process the data. This | Further processes may be used to generate tables for analysis. | Operational Metadata will provide technical audiences with further information |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | processed. | statistical processes. Includes concepts such as statistical process, process implementation, data collection methodology.<br><br>Also need to design quality characteristics and processes. | requirements. Includes concepts such as record variables, record types, matrix operations. | | may include the creation of cubes and registers, aggregating results using derivation rules, applying editing and imputation strategies, applying confidentiality rules, etc. | Operational metadata will also be used to prepare data for dissemination.<br><br>Operational Metadata may also be used to analyse the data. | on how the data was collected and processed. |
| **Quality Metadata** | 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.<br><br>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** | Locate current data through physical metadata eg datasets, server locations, etc | May begin to define/ identify where data will be stored.<br><br>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. |

**Appendix B - New Definitions of Metadata Types compared to Previous Definitions**

The definitions of Metadata used in the original versions of the white paper were as follows:

Definitional metadata relate to statistical units and populations, topics, classifications, data item definitions, questions and question modules, and statistical terms.
Operational metadata arise from and summarise the results of implementing the procedures. They include measures of respondent burden, response rates, edit failure rates, costs, and other quality and performance indicators.
Systems metadata are active metadata used to drive automated operations, including, for example, file layouts and access paths, business processes.
Dataset metadata comprise the minimal systems metadata used to describe, access and update datasets, data structure.
Procedural/Methodological metadata relate to the procedures by which data are collected and processed.

Definitional, Dataset and Procedural/Methodological metadata are examples of Statistical Metadata while Operational and Systems metadata are examples of Process Metadata. Together they combine to form a complete picture of Statistics New Zealand's Knowledge.

These definitions are compared to the new definitions in the following diagram (based on the examples given):