**UNITED NATIONS ECONOMIC COMMISSION
FOR EUROPE (UNECE)
CONFERENCE OF EUROPEAN STATISTICIANS**

**EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN
UNION (EUROSTAT)**

**ORGANISATION FOR ECONOMIC COOPERATION
AND DEVELOPMENT (OECD)
STATISTICS DIRECTORATE**

Meeting on the Management of Statistical Information Systems (MSIS 2011)
(Luxembourg, 23-25 May 2011)

Topic (i): Architectures, models and standards

Statistics 2020 and Platform Approach
Te Käpehu Whetü

Invited Paper

Prepared by Les Cochran and Matjaz Jug, Statistics New Zealand

I.   Introduction

1.   Statistics NZ has embarked on a new 10-year strategic roadmap which involves transforming delivery of statistics, based on improved relevance, increased use of administrative data, standardisation and business improvement. This programme is supported by the new enterprise architecture with standardised application platforms.

2.   This paper provides information on the current risks with the legacy technologies and architecture in Statistics New Zealand. It describes new architectural approach, based on generic business process model and statistical architecture (clusters) and implementation of this approach within Statistics 2020 Programme. Current state of platforms implemented to date is also presented and at the end strategic benefits are discussed including the benefits to the government, official statistical system and wider statistical community.

II.   BACKGROUND

A.   Current Legacy Risks and Generic Business Process Model

3.   Statistics New Zealand has the high risks associated with the legacy technologies and architecture. External review in 2010 has confirmed that the situation will become urgent and put Outputs at risk in approximately five years.

4.   Generic Business Process Model describes Statistics NZ's end-to-end statistical business process. Model has been developed within BMTS programme and has been later used as a model for international Generic Statistical Business Process Model (GBPMS). Processes are generic, across Statistics NZ, down to the sub-process level for all statistical collections and outputs. There are seven processes in the Statistical Value Chain. These are:

     (a) Need: is an ongoing process to determine the statistical needs of Statistics New Zealand's stakeholders.

     (b) 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

     (c) Build: produces the components needed for the end-to-end solution, and tests that the solution works

     (d) Collect: acquires collection data each collection cycle and manages the providers of that data

     (e) Process: describes cleaning the detailed data records and preparing them for analysis

     (f) Analyse: is where the statistics are produced, examined in detail, interpreted, understood and readied for dissemination

     (g) Disseminate: manages the release of the statistical products to the customers.

B.   Statistics New Zealand IT Strategy 2009 - 2012

3.   One of the main outcomes defined in IT Strategy is "Statistics NZ's IT infrastructure is sustainable and robust with standardised platforms based on the generic Business Process Model (gBPM) and aligned with statistical architecture (clusters). Stove-pipe, legacy IT applications will be a thing of the past."

4. There have been a number of significant activities undertaken across the organisation in the last couple of years which support enablement of the required business capability. These include:

- the development of a Statistical Architecture vision for:
  - micro-economic statistics with a focus on the increased utilisation of administrative data sources rather than direct surveying.
  - social statistics with a focus on growing needs for social indicators, reduced collection vehicles and cross-government integration needs.
- the development of a Dissemination Strategy with a focus on standardising on a dissemination data warehouse which will provide static and dynamic data to the varying dissemination tools to improve access to statistics (static web pages, tables, mapping, data visualisation).
- the development of a Statistical Metadata Model with a focus on providing a common language by standardising the application of concepts to the statistical process, and recording metadata once at the point of capture.
- the development of a Collection Strategy.

5.   The combination of these activities and viewing the gBPM from a business capabilities perspective has resulted in an Enterprise Architecture approach which we have termed the 'platform approach'. Five platforms have been identified to date, and are in various stages of implementation; the POSS Platform for household/social statistics, the BEST Platform for micro-economic statistics, the National Accounts Platform, the Collection Platform and the Dissemination Platform. This approach is enabling the mitigation of a number of the risks associated with implementation and is beginning to provide Statistics NZ with a sound basis on which to progress Enterprise Architecture and sustainability initiatives.

C.   Te Käpehu Whetü – Statistics 2020

6.   Te Kapehu Whetu is derived from Kapehu (compass) and Whetu (the stars). A waka is navigated by aligning the points of the craft with the points on the horizon where the sun and particular stars rise and set a metaphor for the organisation's journey towards its strategic objectives – we travel with certainty and an eye on the ultimate destination.

7.   The Strategic Plan 2010-2020 sets out what the organisation is seeking to achieve over the next 10 years and how it plans to get there. The Plan contains four key Strategic Priorities:
      (a)   Leading the Official Statistics System so that it efficiently meets the country's needs for relevant, trustworthy, accessible information (Strategic priority 1)
      (b)   Obtain more value from official statistics (Strategic priority 2)
      (c)   Transforming the way we deliver our statistics (Strategic priority 3)
      (d)   Creating a responsive, customer-focused, influential, sustainable organisation (Strategic priority 4).

8.   Programmes and projects implementing new enterprise architecture will happen within 6 portfolios:
      (a)   Official Statistical System Portfolio (SP1)
      (b)   Access and Use Portfolio (SP2)
      (c)   Foundation Portfolio (SP3)
      (d)   Economic Portfolio (SP3)
      (e)   Social Portfolio (SP3)
      (f)   Capability Portfolio (SP3)

9.   The Enterprise Architecture direction outlined below has been confirmed as a critical enabler of Strategic Priority 3 (Transforming Delivery) and provides a key foundation to all four components of the Delivery Transformation Plan; Relevance, Standardisation, Administrative Data and Business Improvement.

III.   Platform Approach

A.   Statistical Architecture Framework

10.   Platforms support (primarily) the Collect through Disseminate 'processes' of the gBPM. These are:
      (a)   A Collection cluster – to enable the generic Collect process
      (b)   A Dissemination cluster – to enable the Disseminate process.
      (c)   Process and Analyse are enabled by:
          •   A Micro Economic cluster – for Micro economic collections.
          •   A Macro Economic cluster – for Macro economic collections.
          •   and a Social and Population cluster – Social and Population collections.
These clusters are all supported by Infrastructure - Statistical Infrastructure (Statistical services, tools, and data) which are in turn supported by IT Infrastructure.
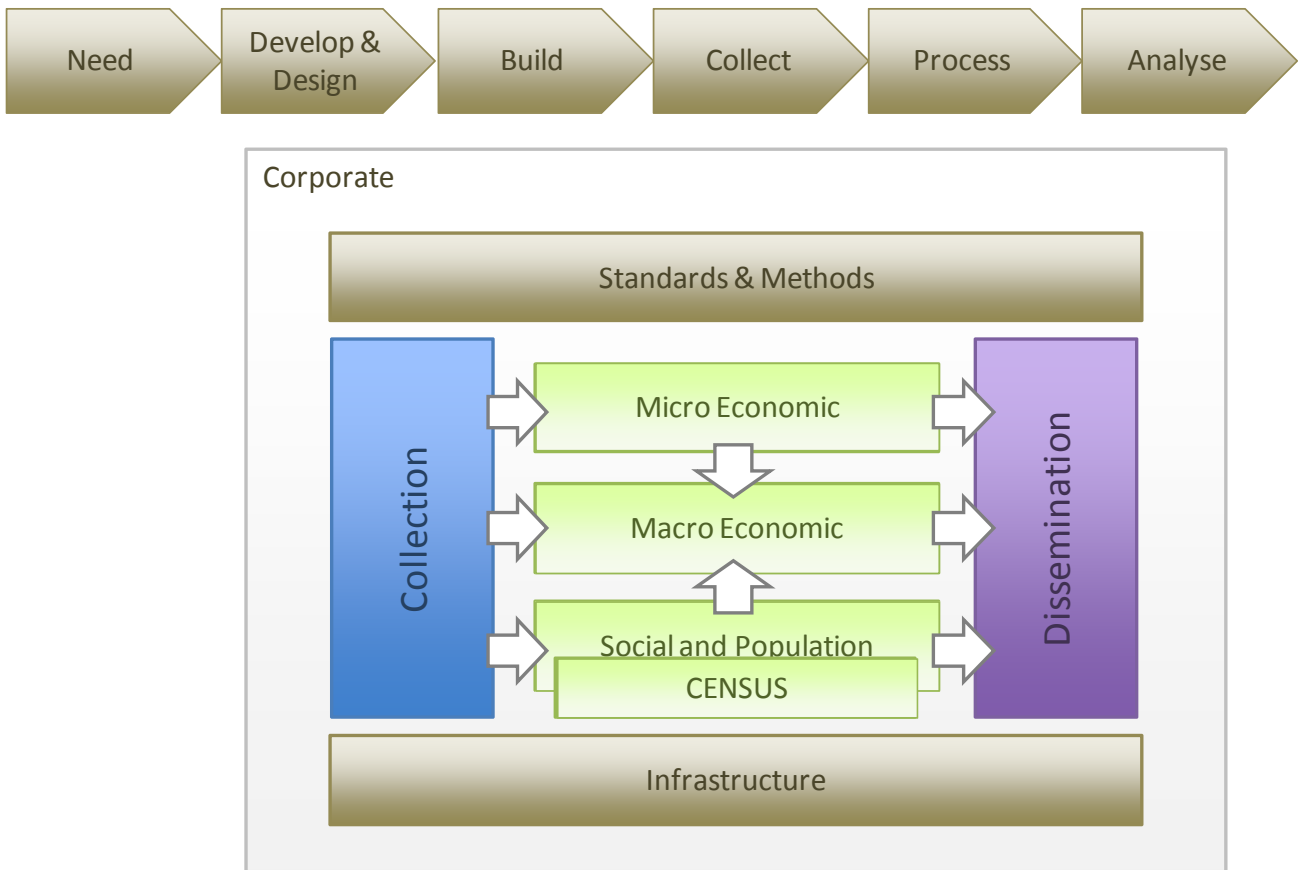
Figure 1: Statistical Architecture Framework.

B. Enterprise Architecture Framework

11. The IT Architecture with standardised platforms to support the business process model and align with the statistical architecture. Each layer is supported by the layer below and all layers supported by IT Infrastructure (see Figure 2).

12. The IT Infrastructure supports all of the other layers. This layer includes database servers, lotus notes servers, SAS servers, Backup and disaster recovery systems, phone systems, networking systems and corporate desktop and standard software (Lotus Notes client, Office, SAS EG, etc).

13. The Data layer includes the Frames, all collected survey, census, and administrative data, classifications and metadata, all documents stored either within notes or on the file servers, HR and Finance data. It is this data that is created, collected, managed, modified, shared, disseminated, and archived via the tools and services.

14. The Tools & Services perform most of the tasks on the underlying data. Whether it is a standard statistical tool performing imputation, an IT service capturing progress information, or a corporate tool capturing time and costs information. These tools typically perform some process on the underlying data but may also be acting with one another in order to provide the required service.
The three main areas or groups of tools and services are:
  (a) Statistical Tools & Services – includes statistical toolbox with tools for editing, imputation, estimation, confidentialisation, etc

(b) Non Statistical Tools & Services – mostly IT tools and services to enable the platforms and the Statistical tools and services – these include Data loading services, workflow tools, configuration and logging tools

(c) Corporate tools – e.g. those used for HR and Finance, Project Management, etc.
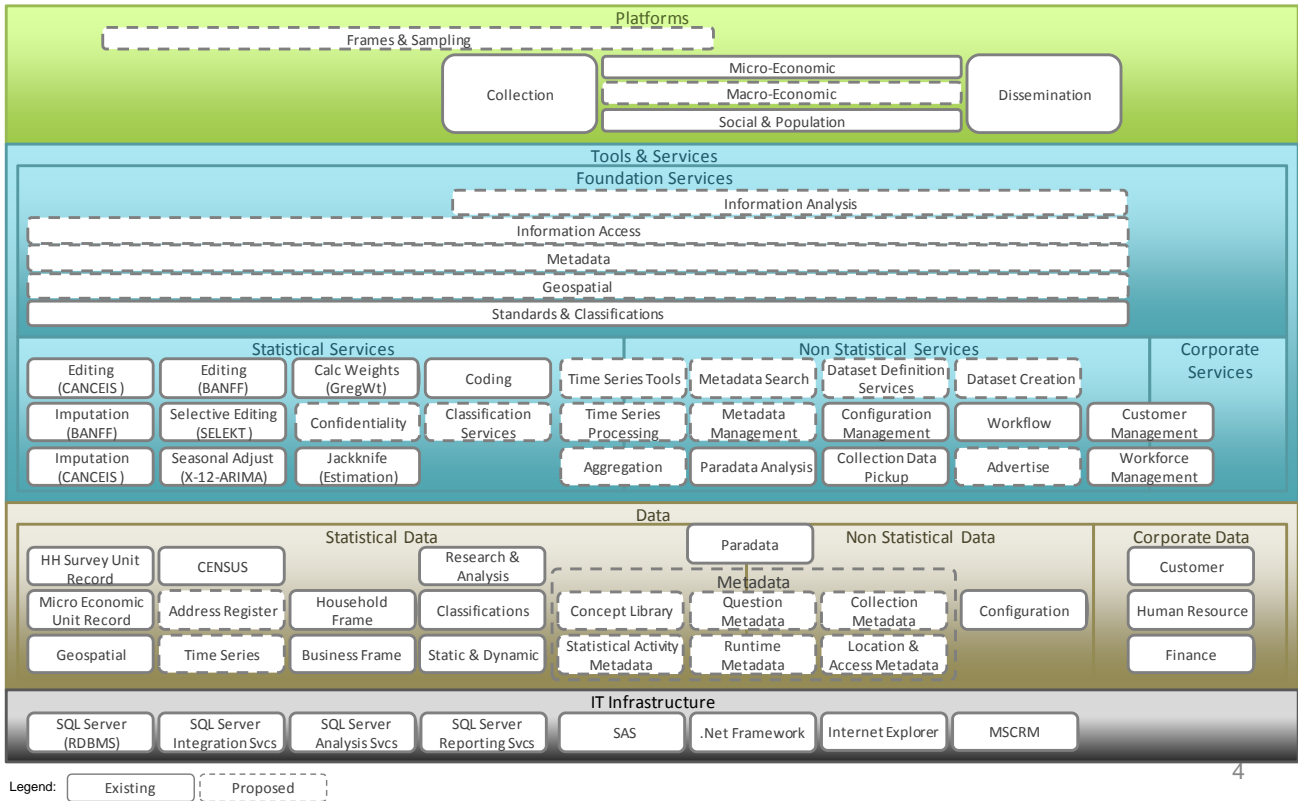
Figure 2: Enterprise Architecture Framework

C. Reuse of Tools and Services

15.   Current Platforms are aligned to Statistical Architecture and bring together standard tools & services to support each statistical area. The platforms also include some custom platform components – These are either used to coordinate and manage the tools that provide the functions available through the platform, or are components which may actually become standard tools themselves and will be pushed down to the standard Tools & Services layer so to enable them to support other areas (see Figure 3), e.g. Micro Economic platform has produced configuration tools which will become standard a standard IT tool and will support many platforms as well as other tools and services. Platforms Support Platforms – in fact the tools & services that support one platform may actually be used to support other platforms as well. The display of the platforms in layers is intentional and indicates the level of potential reuse between the tools that support one layer and the tools to support the layers above it, i.e. platforms in the higher layers are more likely to so use the tools or services from lower layers. Also apparent in the diagram is the similarity of the Collection, Micro-Economic, Macro-Economic, Social & Population, and Dissemination platforms to the Statistical Architecture Framework.
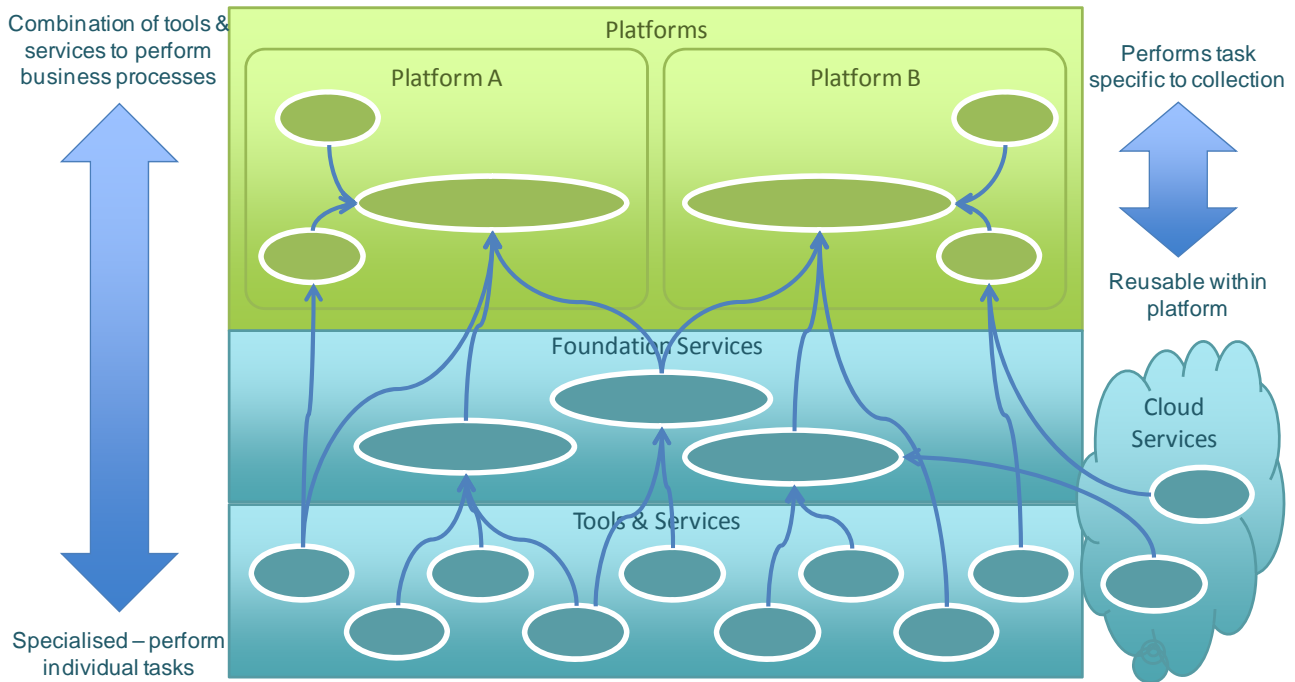
Figure 3: Reuse of Tools and Services

D.   Platform Integration and Transition from Legacy Environment

16.   One of the most difficult challenges will be transition from current stove-pipe to the standardised, platform-based environment. While we don't anticipate difficulties with the IT infrastructure which is already consolidated on common HW and server/client SW stacks, it is the data and application layers where non-standardised and unique processes are hardcoded within old legacy systems. This requires a significant change in processes, applications and culture for transition to be completed successfully.

17.   Five initial platforms have been identified, Household Platform (POSS), Micro-Economic Platform(BESt) , Macro-Economic Platform, (DNA), Collection and Dissemination platforms. These have been initially developed for the new statistical collections and services, for example HH platform for new Programme of Social Surveys (GSS, Time-Use etc.), microeconomic platform for tax data and dissemination platform for Business Toolbox web service. We are now at the end of this phase.

18.   In the second phase the emphasis will be on functional extension of existing platforms and integration between them with the ultimate goal to create e2e process required to move some key collections from legacy environment to the new infrastructure. Dissemination platform will steadily replace current multiple concurrent dissemination systems. Collection platform will provide collection function for surveys to be migrated first (sub annual micro-economic surveys and HLFS) as well as enable true multi-modal collection and respondent management. We also plan to redevelop statistical infrastructure, including classifications and metadata management, business and geo frames etc. To be able to achieve smooth transition and coexistence of both old and new, we are standardising processes and data/metadata structures on the boundaries of collect/process/analyse/disseminate phases. Our aim is to allow migration of outputs before e2e process is finished. While continuing production from legacy environment, Legacy Mitigation Programme is responsible to maintain the risk to output production at an acceptable level.

19.   The final phase will be migration of majority of surveys to the new platform-based environment. Through that phase of transition programme additional platforms may be required, for example population/demography and data integrated however it is our aim to standardise on minimal number of platforms. Majority of statistical infrastructure will be redeveloped so additional consolidation will occur at

the Tools and Services level with reuse of statistical (confidentiality, seasonal adjustment etc.) and non-statistical (configuration, workflow management) services by several platforms. Cultural change is key to the success: many owners of the stovepipe systems believe that their systems and processes are unique but as more areas move onto the new platforms, the thinking is slowly changing to be more acceptable to the standardised systems (and business processes). Also where an area does have a something that is unique, the aim is to extend the platform (or the underlying services) to accommodate the unique feature.

IV. Current State

A. Collection Platform

20.   Provides "Collection" capability for all collections including surveys and administrative data supplies. Systems will include respondent contact and management, data capture, coding and management, work-flow and work-force management, paradata analysis and instrument build. The following section outlines the existing key components of the Collection Platform:

   (a)   Contact CRM system, the system provides respondent management, forms post-out and return, sample load and reporting about response rates.
   (b)   Contact Computer Aided Interview (CAI) system, the system provides the following functions: field interviewer management, CAPI collection using BLAISE and CATI collection using BLAISE and Zeacom telephony systems.
   (c)   Mode Allocator, assign case to different mode (CAPI or CATI), currently only available for Disability Survey.
   (d)   Data Collection Load, load collection data to processing data store.
   (e)   Imaging system, it allows forms to be scanned, and repair captured data.
   (f)   Business/Social Administrative data, currently SNZ receives administrative data in three channel; Post (Cartridge Tape, CD-Rom, Paper), Electronic File Transfer and Email (Excel spreadsheet, CSV and encrypted file).

B. Household Survey Platform

21.   The Programme of Official Social Statistics (POSS) program has delivered the Household Survey Platform to provide data processing capability for Social Survey data – data about New Zealanders. This platform provides standardised data storage and standard processing for social survey data. Household Survey platform services and components:

   (a)   Data Management: Storage of data and processes to manage versioning, audit trail, and loading of data.
   (b)   Metadata Management: Loading metadata to set up a new collection.
   (c)   Process Management: Configure (Create and configure processing jobs and tasks), Run (Execute jobs), and Review (Examine processing results) all processing. This employs the Process Workflow component to perform tasks for:
      (i)   Classify and Code: Employs the Coding Services to assign standard categories to uncategorised survey response information
      (ii)   Impute: Employs the CANCEIS tool – see Statistical Infrastructure
      (iii)   Derive: Employs SAS to execute SAS routines to calculate derived values.
      (iv)   Calculate and Apply Weights: Employs SAS to execute a combination of Social Survey SAS routines and routines – see Statistical Infrastructure (GregWt & Jackknife)
      (v)   Quality Assurance: Produces SAS datasets based on criteria specified by users in a Web based user interface
      (vi)   Finalise   : Produces SAS datasets with all data for the selected collection. (in addition to locking down the system to prevent any further data modification for the selected collection)
      (vii)   Micro Edit: Performed manually using a metadata driven Web based user interface

    (d)   Process Workflow: Custom Data Transformation packages with workflow capabilities to perform all tasks for a specified job as determined by configuration values set under Process Management

    (e)   Questionnaire Performance Analysis OLAP tools to enable the analysis of paradata collected as part of the survey process. The paradata includes information about the time to complete questionnaires, sections of a questionnaire, and questions within each section.

22.   Statistical Infrastructure tools used by the platform

    (a)   Coding Services: Services used to search for categories within a specified classification for a given string (or uncategorised value)

    (b)   CANCEIS: CANCEIS (CANadian Census Edit & Imputation System) is an edit and imputation system developed at Statistics Canada

    (c)   GregWt: A SAS macro developed by the ABS (Australian Bureau of Statistics) for Generalised Regression and Weighting of sample survey results

    (d)   Jackknife: A SAS macro to implement a method of variance estimation using systematic samples from the original sample. Developed by SNZ.

    (e)   Configuration Management: An SNZ developed component that provides storage, retrieval, management (including versioning) of configuration data and metadata. Data is stored in the Configuration Data store

23.   Corporate Data sources used by the platform

    (a)   Classification Data: CARS (Classifications and Related Standards) data referenced by many parts of the platform including the Data and Metadata Management components, data-mart management, and SNZ Cube Tools.

    (b)   Household Frame

    (c)   Configuration Data: The repository for Configuration Management

    (d)   Process Logging: Repository for all messages and logs produced by all processes within the platform. All data items are linked back to the process that produced them and the corresponding logs.

    (e)   Questionnaire Performance Data: Paradata produced as part of the collection process with detailed timings the performance of the collection instrument, its contents, and the interviewer:

C.   Micro-Economic Platform

24.   The Better Economic Statistics (BESt) program is delivering the Micro-Economic Platform to provide data processing capability for Micro-Economic data – data about New Zealand businesses. This platform provides standardised data storage for micro-economic data, standard processing for micro-economic data and the ability to trial new methodological processes on micro-economic data. Micro-Economic platform services and components:

    (a)   Data Management: Storage of data and processes to manage versioning, audit trail, and loading of data.

    (b)   Metadata Management: Loading metadata to set up a new collection.

    (c)   Process Management: Configure (Create and configure processing jobs and tasks), Run (Execute jobs), and Review (Examine processing results) all processing. This employs the Process Workflow component to perform tasks for:

           a.   Derive: Employs BANFF tool – see Statistical Infrastructure

           b.   Edit: Employs SELEKT tool – see Statistical Infrastructure

           c.   Impute: Employs BANFF tool – see Statistical Infrastructure:

    (d)   Dataset Linking Management: Tools to enable simple linking of data between collections with common concepts.

    (e)   Data-mart management: Maintenance (Create/Delete) of processing data and OLAP cubes

    (f)   Process Workflow: Custom Data Transformation packages with workflow capabilities to perform all tasks for a specified job as determined by configuration values set under Process Management.

    (g)   Quality Assurance: Employs OLAP technology to review processing changes

(h) Process Analysis: Employs the SNZ Cube Tools to provide cubes to enable the analysis of all processes and their results.

(i) SNZ Cube Tools: A set of tools developed by SNZ that enable rapid development of OLAP cubes linked to standard classifications (input and output).

25. Statistical Infrastructure tools used by the platform

(a) BANFF: BANFF is an editing and imputation system developed by Statistics Canada. It is made up of a collection of specialised SAS procedures, each of which can be used independently, or put together, in order to satisfy the edit and imputation requirements of a specific collection. BANFF is designed to edit and impute continuous numeric data, so is most useful for processing economic or financial data.

(b) SELEKT: a standard tool for selective editing developed by Statistics Sweden.

(c) Configuration Management An SNZ developed component that provides storage, retrieval, management (including versioning) of configuration data and metadata. Data is stored in the Configuration Data store

26. Corporate Data sources used by the platform

(a) Classification Data: CARS (Classifications and Related Standards) data referenced by many parts of the platform including the Data and Metadata Management components, data-mart management, and SNZ Cube Tools.

(b) Business Frame/ Longitudinal Business Frame - The register of all New Zealand business (as deemed by Stats NZ).

(c) LEED (Linked Employer –Employee Database): reference to current employment counts, etc

(d) IRD Prod: input storage data for all data received from IRD (tax data). Forms the primary source for the 'Admin' collections.

(e) Configuration Data: The repository for Configuration Management

D. National Accounts Platform

27. National Accounts Platform is replacing the current suite of programmes used to compile quarterly gross domestic product (GDP) statistics. The project will upgrade the current legacy technology, implement separate development, test and production environments, and provide a clear and consistent approach to the development and maintenance of methodologies and outputs through increased standardisation.

28. DNAs platform services and components:

(a) Data Store: Storage of data for the platform holding various versions and states during the process.

(b) Management Tool: A UI to assist in the management of data and manipulation of data during process.

(c) Import: Services used to bring data into the platform from external sources.

(d) Processing: Stored Procedures and Macros that manipulate the data and manage versions of datasets.

29. Statistical Infrastructure tools used by the platform: SAS EG.

30. Corporate Data sources used by the platform: Classification system CARS.

E. Dissemination Platform

31. The SDDM (Standardised Data Dissemination Management) program will deliver a standardised system to provide publication & release capability. This will provide an end-to-end production and quality assurance process from data delivery to official release. It will accept data from multiple, disparate

processing systems in multiple formats. Benefits will include better management of data, efficiency gains by creating one output and minimise checking of data, centralising rules and confidentiality and reducing the costs involved in the maintenance and licensing of the many existing tools.

32.   Dissemination platform services and components:
   (a)   Translation & Configuration Application: Used to describe and transform legacy data formats into richly defined datasets that meet the publishing standard.
   (b)   Provider Management: A set of services around an index that enlists authorised data providers; this list exclusively controls the endpoints that can contribute data for dissemination.
   (c)   Data Identification: Manage the availability of data into the dissemination process.
      (i) Data Register: An index of datasets that are available to use within the creation of a statistical information product; defines their structure and creation details.
      (ii) Data Locator Service: Polls systems to identify data of the correct status or allows systems to advertise the existence of data.
   (d)   Product Management: Control the creation of a statistical information product and manage the life cycle.
      (i) Product Library: A single consolidated index of all statistical information products along with their release calendar and processing metadata.
      (ii) Product Manager: A UI component to manage the life cycle of the product from creation to release and then retraction.
      (iii) Product Management Services: Service components that orchestrate the life cycle stages and retrieve transform and load the actual data ready for consumption.
   (e)   Product Delivery: Make the statistical information product available for consumption.
      (i) Data Warehouse: A data warehouse that hosts the final data structures and their associated metadata.
      (ii) Embargo Mechanism: Controls the visibility of complete or partial datasets based on pre-defined criteria.
      (iii) Data Access: A collection of services that sit in front of the data warehouse to control the view of data; they receive queries and respond with standards based results (such as SDMX).
      (iv) Data Visualisation Suite: A suite of applications that consume data from the data services and provide the interface to build queries, view data in various manners and extract to pre-defined formats.

33.   Statistical Infrastructure tools used by the platform: Configuration Management (not currently but will be used eventually)

34.   Corporate Data sources used by the platform:
   (a)   CARS
   (b)   Output Classification Manager
   (c)   Subject Area Hierarchy: A hierarchy of subject area groups held and managed by the Product Development & Publishing team.

V.   Benefits

29.   New platform-based enterprise architecture is a key enabler for strategic benefits Statistics 2020 programme is going to deliver (see Figure 4):
   (a)   Continued supply of important and trusted statistics (B1) includes maintaining the highest levels of credibility and accuracy of statistical information that is produced as well as maintaining the scope of statistics that are produced. The programme of work to stabilise and standardise current systems and processes will reduce the risk to the continuing supply of quality statistics. It will also ensure continued national and international trust in the scope and quality of New Zealand's official statistics.

(b)   An agile and responsive NSO able to respond to changing needs (B2): this benefit includes maintaining relevance over time (producing the statistics that users want), as well as prioritising between competing demands to ensure that focus goes on the statistics that are of highest value to New Zealanders. New Enterprise Architecture will help enable this using much more flexible and configurable architecture which will require much less need for IT development.

(c)   Costs to government, businesses and households minimised (B3): new architecture is expected to reduce the long-term operating costs of producing statistics and to substantially reduce the risks inherent in the fragmented and dated systems currently used by Statistics NZ. Reducing costs also includes reducing the compliance burden that falls on respondents by smarter use of technology to collect data.

(d)   Increased use of government data (B4) has both an efficiency dimension (making better use of what is already available and reducing reliance on costly surveys to generate information) and a value dimension by making more statistics available to use and facilitating the linking of data into new datasets. New enterprise architecture will support data integration and analysis, as well as more effective and customer-oriented dissemination of existing data.

(e)   The Government has confidence that its investment in official statistics is value for money (B5): this aspect of benefit looks beyond the production role of Statistics NZ to include its leadership role in relation to the OSS. Nedw approach with reusable services will make these services available to areas that are not on the platforms. Which could be used externally by other producers of official statistics. There is a scope in Statistics 2020 programme to determine which are the areas where OSS infrastructure could be developed and used to provide common capability across OSS. Further more – there is also an opportunity to share development of services and tools with other NSIs
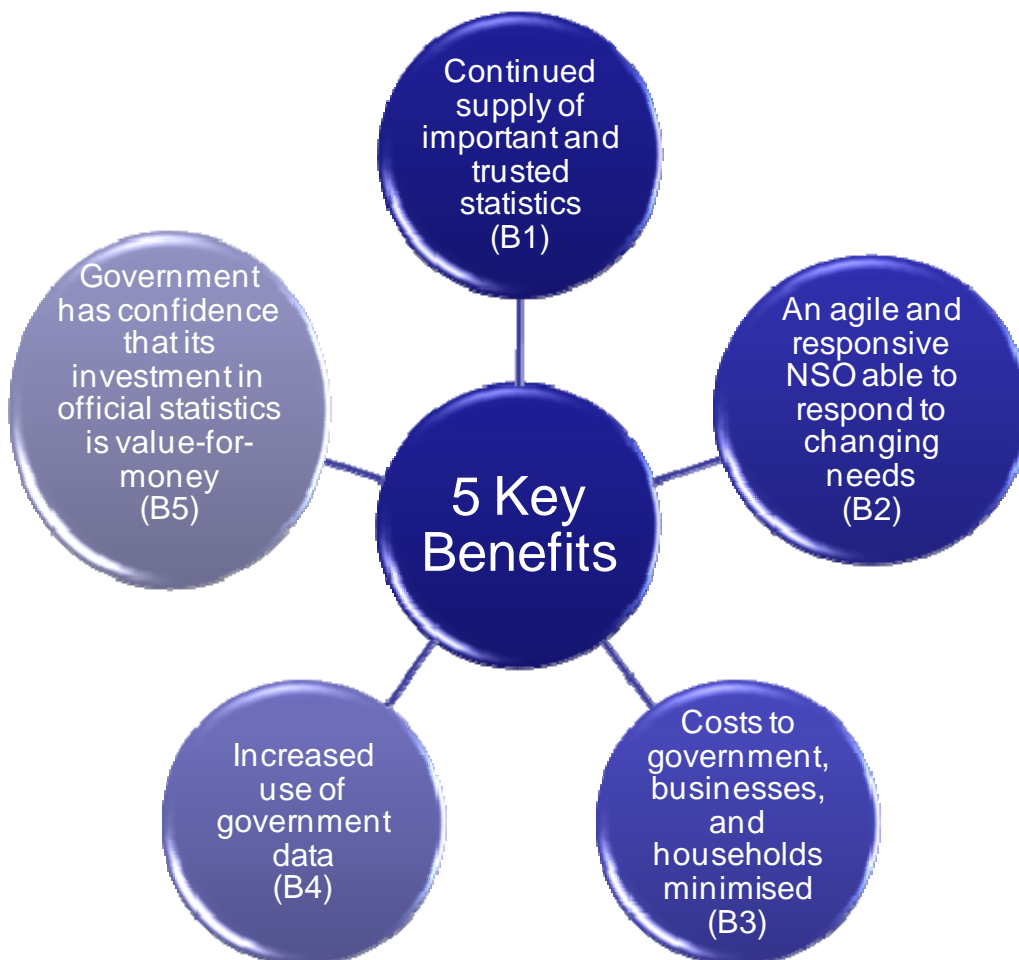
.

Figure 6: Strategic Benefits

REFERENCES

Statistics New Zealand Strategic Plan 2010 – 20,
http://www.stats.govt.nz/about_us/corporate-publications/snz-strategic-plan-2010-20.aspx

Statistics New Zealand IT Strategy 2009 – 2012, internal document

SNZ Enterprise Architecture Poster, internal document