WP. 15
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Oslo, Norway, 24–26 September 2012)

Topic (ii): Global solutions to editing;

ON TAP: DEVELOPMENTS IN STATISTICAL DATA EDITING AT STATISTICS NEW ZEALAND

Prepared by Allyson Seyb, Felipa Zabala, Les Cochran, Chris Seymour, Statistics New Zealand

I. Introduction

1. Statistics New Zealand (Statistics NZ) is the major producer of official statistics in New Zealand and the leader of the New Zealand Official Statistics System (OSS). The organisation tries to maintain the quality of its outputs in a tight fiscal environment that requires the organisation to minimise costs and maximise effectiveness. Statistics NZ's aim is to produce statistics that are fit for purpose in a cost effective and sustainable way.

2. Traditionally at Statistics NZ, decisions about statistical redevelopment of outputs were made independently of prioritisation and quality frameworks and were driven by the need to replace outdated production systems. Because they were reactive rather than planned for, IT systems developments were expensive and, because they happened at widely spaced intervals, resulted in production 'silos', unique production systems for each survey where any further changes to systems (eg updating classifications) could only be made by IT teams. Different production systems for each survey meant it was difficult for people to move between surveys without a significant amount of training, and subject matter specialists became experts in processing data and troubleshooting IT systems rather than in data analysis (Seyb et al, unpubl.).

3. Over recent years, the New Zealand government has invested significantly in developing modern processing systems, and Statistics NZ has actively pursued a strategy of standardisation of business and information concepts, methods, processes, and technology. Modern processing platforms use standard tools for common processes. These platforms make it possible to process data quickly and easily. The Micro-economic platform is also highly configurable, which allows subject matter specialists to create their own production systems by choosing from a selection of standard tools that carry out the usual processes associated with statistical products. For example: editing data, imputing for erroneous data or non-response, seasonal adjustment, and applying confidentiality rules. A new governance model ensures decisions about spending are made at the right level, and across the whole of Statistics NZ's products and services.

4. This paper follows an invited paper presented at UNECE 2011 (Bentley & Zabala, 2011) and describes the latest developments in Statistics NZ's economic and household processing platforms. Section II focuses on the strategic framework that is helping Statistics NZ to achieve

the desired changes to its systems and direction. The household and economic platforms are described in section III. Section IV focuses on challenges and lessons learnt, and section V contains concluding comments. Future directions are indicated throughout the paper.

II. Strategic Developments

A. Strategic Framework

5. In 2011, Statistics NZ embarked on a 10 year programme of change called *Statistics 2020 Te Kāpehu Whetū: Achieving the statistics system of the future* (Stats 2020). The Stats 2020 programme of change is extensive: transforming how our statistics are delivered, leading the New Zealand OSS, obtaining more value from official statistics, and creating a responsive and sustainable organisation (Statistics New Zealand, 2010). Standardisation of processes, methods, tools, and systems is a key component. It also includes improving the quality of the statistics produced, by ensuring they are relevant and remain so over time.

6. Since 2008, the organisation has focussed on developing new statistical architecture designs for both economic and social statistics. By 2020 the organisation's aim is that administrative data will be the primary source of information, supplemented where necessary by direct collection. The new household and economic platforms are designed to enable this aim to be achieved, and also to enable the organisation to overcome challenges in its internal operating environment. These challenges include a number of aging IT systems that are expensive to maintain; processes which move data physically between departmental teams, with specialist job roles limiting efficiency gains and creating a barrier to innovation; and unnecessary diversity and duplication conceptually - in terms of business and information concepts - and practically - in terms of methods and technology.

7. In the last 12 months, many more of the organisation's outdated IT systems have been replaced with fewer systems that are more flexible and standardised. The common processing platforms for clusters of social and business collections have been extended to include more outputs and a wider range of standard tools. The organisation is documenting current methods and tools to use as a benchmark to allow us to understand how much progress we are making in terms of standardisation. The use of standard tools for common processes has resulted in a reduction in the number of methods and tools in use in the organisation.

8. Metadata is fundamental to an end-to-end solution and integral to the new platforms. A metadata tool, 'Colectica' has been implemented. Colectica is off-the-shelf software, and stores all the information about Statistics NZ's outputs in one place. It is possible to browse for information by statistical output, concept, or business unit. The information is available internally at present, but in future it will be available through the organisation's website.

9. Processes to research and introduce new standard methods and tools have been established, and the use of standard statistical tools is mandatory for outputs using the new platforms. A review of our generic Business Process Model is planned for 2012, and our Quality Management programme has just been reviewed and aligned with the recently released UN standard. Diagnostic reports, created automatically during processing on the new platforms, indicate the quality of processes and are used in the continuous improvement of the process and process components. Production systems are no longer 'black boxes' where business rules are hidden

inside production systems and inaccessible to subject matter specialists. Generic reusable system components mean faster migration of existing production systems and a greater ability to respond to changing user needs in an efficient way.

B. Collaboration

10. In 2009, the Australian Bureau of Statistics proposed an initiative that focuses on stronger collaboration on statistical information management systems among national statistical offices (NSOs). Statistics NZ, along with statistical offices in Canada, Norway, the Netherlands, Sweden, and Australia are part of this collaboration effort. NSO's recognise the need to work together in a climate of limited funding, aging infrastructure, and increased user demand. Two types of collaboration were identified as worth pursuing: longer term efforts that would result in the largest gains but would require adopting enabling information management and architecture management standards, including standards to facilitate the exchange of data; and shorter term efforts that would help build trust between NSO's. Statistics NZ is involved in collaboration opportunities in confidentiality and editing and imputation.

11. The Statistical Network on Industrialisation of Editing operated for almost two years between June 2010 and April 2012. The network identified four short-term tasks to focus on: a glossary of key concepts, a general editing workflow, the identification of a minimum set of standard editing and imputation methods, and the establishment of some overall objectives and principles for the industrialisation of editing. Some progress has been made on each of the tasks. The closure report also contains reflections on challenges the network encountered in terms of establishing a shared vision: the management of the project and clarity around roles; the governance of a project that involves staff at many NSOs, particularly securing the required resources; and the importance of continued support for the initiative from the highest levels of each organisation. In 2013, Statistics NZ, along with the other members of the network, plans to focus on exploring the use of the selective editing tool SELEKT, developed by Statistics Sweden, with the aim of including it in the library of tools available on the processing platforms.

12. Similarly, the Statistical Network on Confidentiality was created in June 2010 and closed in April 2012. The network focused on sharing information and practices by developing a list of methods and framework for methods for statistical disclosure control. This network also noted as a challenge the problems securing resources as individual NSO's priorities shifted. Statistics NZ is considering evaluating SAS2ARGUS in 2013. This tool, developed by Statistics Sweden, is a version of the Tau Argus software for statistical disclosure control, implemented as a service.

C. Enterprise Architecture

13. Statistics NZ promotes a 'shared services' enterprise architecture model: systems are designed and built by connecting standard business capability, where the process elements are mostly implemented as common services. The services are reusable software which may be developed in-house, or sourced from other NSOs or private companies. Business logic is extracted from applications and formalised as configuration rules which chain together processes and services into meaningful business workflows. Data and metadata are defined and managed using standards-based formats aligned with the generic statistical information model reference framework which is currently being developed. Processes and services are implemented in a standard way to collect performance and quality metrics to allow continuous improvement (Clarke, 2010).

14. Currently, five platforms cover all aspects of statistical production: data collection, economic outputs processing, social and household outputs processing, National Accounts processing and data dissemination. The platforms are at various stages of development, ranging from the first generation collection platform to the Micro-economic platform which will have eight regular outputs by the end of 2013. Good progress has been made on a standard data dissemination platform, which includes a web browser that provides access to data stored on the dissemination platform and an SDMX gateway that provides a machine-to-machine data exchange service. A key area for development in the near future is expanding the collection platform to include electronic collection, and moving away from expensive paper-based collection of information. No limit has been set on the number of platforms that will ultimately be developed, but it is expected to be in the tens and will certainly be far fewer than the hundreds of survey processing systems used by the organisation just a few years ago. In the future, it is likely there will be population statistics platforms that will include processing of the Population Census and population estimates.

15. A key element of the platforms is the use of SAS for statistical processing and analysis. The SAS runtime environment is highly scalable and easily able to process large volumes of data in batch mode. Also, standard tools from the statistical toolbox Banff and CANCEIS are available on the platforms for editing economic and household data, respectively. A simple selective editing solution has been implemented on the economic platform. The selective editing score is an impact score rather than an impact and suspicion score (vis-à-vis SELEKT) and is useful for outputs that have very little editing overall where the cost of using a more sophisticated tool for selective editing outweighs the potential savings. Use of the score resulted in a 66 percent reduction in the number of records reviewed in one output.

III. Statistical Production Systems

A. Household Platform

16. The Household Survey platform in current development is the second generation platform, which will provide the capability to process and analyse social survey data. This platform was previously referred to as the Programme of Official Social Statistics (POSS) platform (Bentley & Zabala, 2011).

17. The design of the new platform was informed by an evaluation of the interim platform, and the direction outlined in the Social Statistics Architecture (Bycroft, 2009). The platform will accommodate three social surveys and their supplements. The three surveys act as vehicles for a wide range of supplementary modular topics. These three surveys are the Household Labour Force Survey, which will be a vehicle for work-related topics; the General Social Survey, a vehicle for a range of topics in the social domain; and the Household Economic Survey, which will be a vehicle for income, expenditure and wealth-related topics. The new platform will also accommodate a redesigned Household Labour Force Survey, and will provide capability that can be used for other systems, such as a potential population platform or a future census system (Cochran, 2011).

18. The platform currently supports only the ‘Process’ phase of the gBPM. It utilises standard tools to load, code, micro-edit, and finalise a unit record dataset. Processing on the platform ends with the creation of a clean unit record dataset, from which the subject-matter specialist creates statistics. The platform can process a Blaise-based survey, providing the platform with the ability to automatically load response data from the collection platform interface as well as from file sources. Future work required to handle new Blaise-based surveys will be limited to building the the Blaise instrument, loading metadata, configuration, and the creation of specific business rules

(eg derived variables and edits). Work in the next two years will include the addition of new functions, or extension of existing ones to handle the unique features of new surveys migrating to the platform, for example the addition of diary capability for the Household Economic Survey. Future development will extend the platform capability to support the 'Analyse' phase of the gBPM with the creation of final outputs ready for dissemination.

19. The platform uses a mix of shared and survey specific systems. Figure 1 illustrates the systems used by the Household Survey platform.

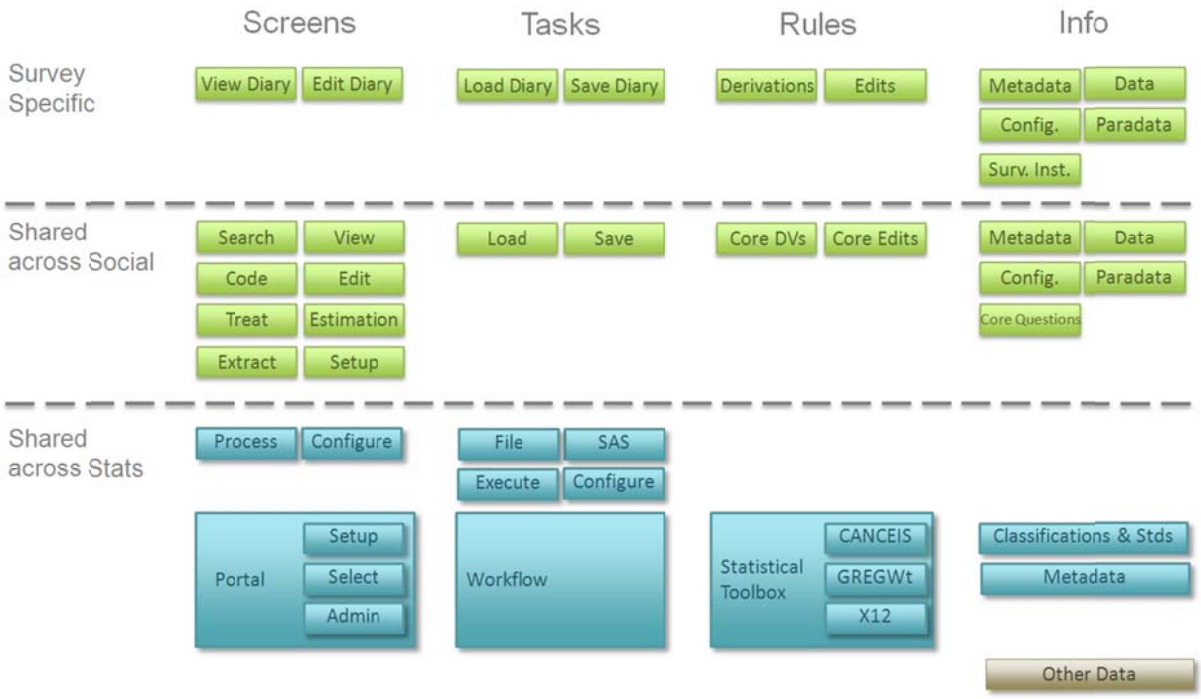

Figure 1. Shared and specific systems used in the Household Survey platform

20. The shared systems include those shared across social surveys, as well as across Statistics NZ. For example: the screen portal, the workflow tasks, statistical tools and information on classifications and standards, and metadata. The statistical tools currently available in the platform are CANCEIS for imputation, GREGWt for calculating weights and X12 for seasonal adjustment. Access to the tools is dependent on the workflow of the tasks. Inclusion of a statistical tool from another platform is relatively straightforward. If a survey requires imputation methods available from BANFF, which is currently integrated into the Micro-economic platform, BANFF can easily be accessed from the Household Survey platform and added into the imputation task.

21. Micro level data for a survey for a given time period uses the format presented in Table 1. A sample dataset is given in Table 2.

Table 1. Micro data format in the Household Survey platform

Survey cycle code	ID	Repeat	Variable name	Value

Table 2. Example of a micro dataset in the Household Survey platform

Survey cycle code	ID	Repeat	Variable name	Value
HLFS 107	Person1	11	Age	15
HLFS 107	Person1	11	Sex	M
HLFS 107	Person2	11	Sex	F
HLFS 107	Household1	11	Household composition	Multi-person

22. The above data format, and the use of shared systems, provides the platform with the following important features:

    a) A user can switch between surveys or data collections without leaving the system. It also allows datasets to be extracted for analysis outside the platform at any time. These may be complete datasets of all data for a survey, or partial datasets of selected time periods, modules, or variables.

    b) The platform has an extensible user-configurable workflow to control all data processing. It maximises both configuration and transformation reuse across data collections. This workflow supports processing of data regardless of location. This facilitates processing of data from other systems as required, eg census data, and supports administrative data. This also allows for automated processing, including automated editing and imputation that can start before manual intervention, which is also available in the platform.

    c) All data processing can be managed by users. These users range from analysts monitoring processes to ‘super users’ running processes and changing parameters (available on completion of Statistics New Zealand’s configuration service) to administrators altering process flows and steps.

    d) All survey response data and derived data can be viewed in the survey portal along with an audit trail of any changes.

23. Future work on the platform includes complete end-to-end processing of a survey, from ‘Collect’ to ‘Disseminate’ (from the gBPM) capability.

B. Micro-economic Platform

24. The Micro-economic platform, previously referred to as the BESt platform (Bentley & Zabala, 2011), provides the capability to process and analyse economic surveys and administrative data collections. The Micro-economic platform also has elements of ‘Develop and Design’, ‘Build’, and ‘Collect’ (from gBPM). It will ultimately provide a platform to support the Business Register and the new Integrated Data Infrastructure (IDI).

25. The Micro-economic platform stores and processes a wide range of outputs. Collections in production on the platform to date include the quarterly Economic Survey of Manufacturing (QMS), the quarterly Economic Survey of Wholesale Trade (WTS), business annual financial accounts (IR10s), value added tax (goods and services tax or GST), other smaller tax forms from New Zealand’s Inland Revenue Department, a version of the Business Register, and Overseas Merchandise Trade data. In development over the next 12 to 18 months are the Agricultural Production Survey (APS) and the Annual Enterprise Survey (AES). The Quarterly Building Activity Survey is currently undergoing redevelopment and the new survey, together with its building consents survey frame, will be implemented on the economic platform in 2013.

26. The platform uses a mix of shared and survey specific systems. Figure 2 illustrates the systems used by the BESt platform.

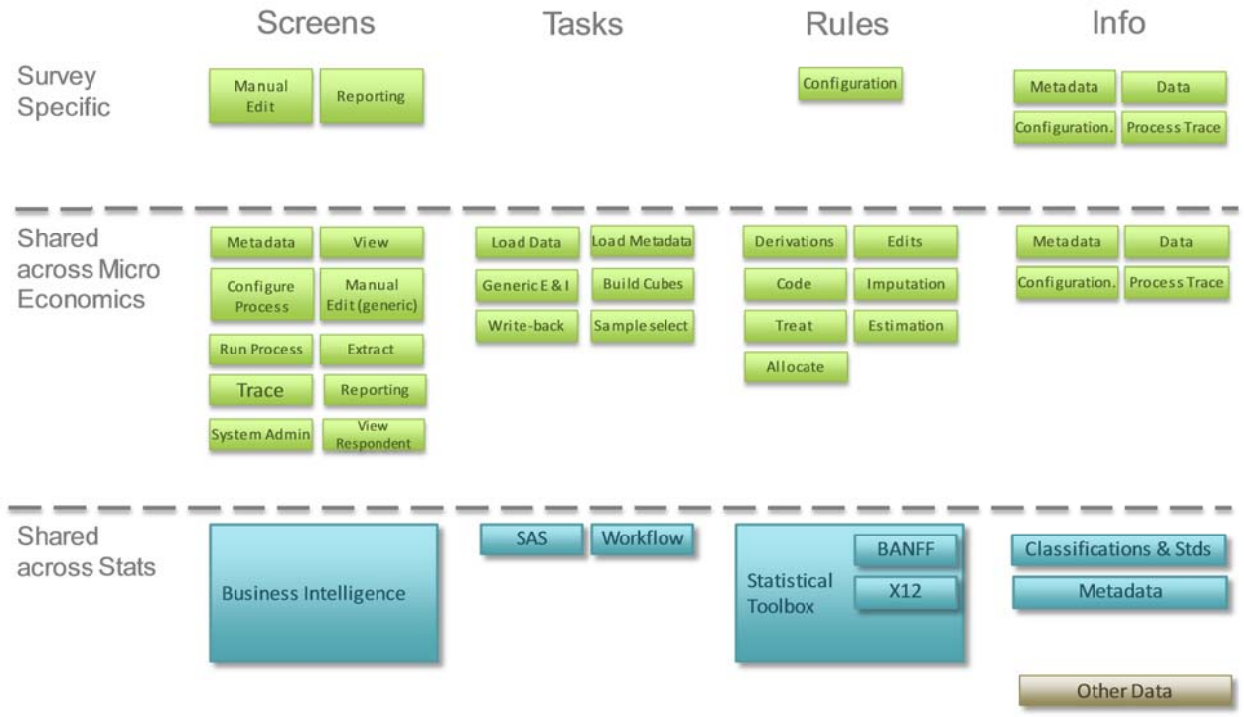

Figure 2. Shared and specific systems used in the Micro-economic platform

27. The platform has a user configurable workflow to control all data processing, and maximise both configuration and transformation reuse across data collections. Processing can be fully automated and can be queued to run as a batch job at a regular time, or as required by the subject-matter specialist. Manual editing and imputation screens are also available.

28. The BESt platform allows statistical maintenance to be carried out in small, regular increments and uses standard tools for common processes. The advantage of the change to a cluster approach (ie groups of surveys and administrative data collections processed on the same IT platform) is that subject matter specialists have more time to concentrate on data analysis. When a new product is developed or an existing one is redeveloped, the new platform is highly configurable; changes are easy to implement and there is minimal IT intervention needed. It is possible to explore the effect of changes on outputs without affecting production versions of processing systems.

29. There are differences between the economic and household platforms. On the economic platform the workflow is standardised and generic; economic survey or administrative data is processed using a standard sequence of components, such as BANFF editing procedures, run in a specific order. Analysts using the Micro-economic platform are able to design and build their own workflows (ie their own production systems) by assembling the appropriate process components. By contrast, each workflow on the household platform is customised to the output being processed and only IT developers can alter process flows. The two platforms also have different aims in terms of data analysis. Analytical capability on the household platform is achieved by exporting data out to a separate dataset, which is structured in a way that supports analysis by

tools outside of the processing system. The Micro-economic platform is optimised for analysis, with analytical capability built into the basic database design (Seyb et al, unpubl.)

30. Future work on the economic platform will focus on making more statistical tools available for common processes, and developing a framework for measuring and reporting the benefits that have resulted through standardisation.

IV. Challenges and Lessons Learnt

31. The organisation has taken advantage of the availability of mature IT environments, clear strategic directions in terms of architectures (both IT and statistical), and learning from earlier prototypes to modernise production systems. The move from stand-alone production systems to standard processing platforms is still in progress. Balancing generic and specific needs is challenging: development teams need to think broadly and aim for generic solutions. Not all process elements are suitable to be implemented as common services; efficiency considerations may lead to a mix of services and platform specific solutions. The organisation's methodologists are championing the use of standard methods and processes.

32. Providing statistical subject-matter specialists with generic, highly automated production systems is not enough to change an organisation's culture. Moving from a process culture, where much of an analyst's time was spent processing data, to a more constructive innovative culture, where much of the data processing is automated and effectiveness is judged at the system level rather than the component level, has not been easy, and the organisation is still in the early stages of achieving this transformation. The organisation has developed a strategy to help teams understand and explore what the new environment means for them.

33. A key enabler of the culture change to date has been the adoption of an agile project management approach for IT development projects. The agile approach involves setting up multi-disciplinary teams, planning the team's work in manageable pieces that take a short period of time to complete, and then checking in with the team regularly to report on progress and remove impediments. Progress is very fast under this approach and impediments to progress are identified and resolved quickly. While developed primarily as an iterative method of determining requirements in engineering development projects, the agile approach can be used in any team situation and has had the added benefit at Statistics NZ of blurring role boundaries and encouraging all team members to solve any problem. Team capability develops very quickly, and the team environment becomes more conducive to the creation of innovative solutions (Seyb et al, unpubl.).

34. As well as applying agile project management techniques to our development projects, the organisation has also adopted an agile approach to governance. Decisions are revisited and reconfirmed or altered in response to changes in the environment. For example, the two platforms have evolved differently, but instead of forcing one or the other to conform, the differences will be evaluated in terms of costs and benefits and a preferred direction established.

V. Conclusion

35. Statistics NZ aims to produce statistics that are fit for purpose in a cost efficient and sustainable way. Through the Stats 2020 programme, the organisation is developing and implementing modern production systems that support the organisation's goals.

36. The platforms presented in this paper are under development and will be for some time to come. As surveys are moved onto the platforms, the organisation is making significant savings. For example:

    a) Migration of surveys onto the new platforms is faster and cheaper now that there are a substantial number of generic reusable system components.
    b) The ability to analyse processes has seen a reduction in editing effort as resources are able to be targeted more effectively.
    c) There has been a significant shift away from manual editing, particularly for economic surveys and administrative data using the Micro-economic platform.
    d) Common processes have been standardised and automated.
    e) It takes less time for the organisations products to be released.
    f) Staff understand the methods and processes used in their production systems and are able to identify process improvements and implement the changes themselves.

37. The move from silo systems, to large-scale IT platforms supporting the processing and analysis of multiple outputs is underway. The platforms are based around standard business and information concepts, encapsulated in standard methods and technology. Opportunities identified during the transformation include providing systems built in such a way that subject matter specialists can easily create their own processing systems and outputs, without the help of IT teams. Standard platforms and tools automate production and minimise manual processing, and diagnostic reports encourage continuous improvement at every level. Next steps include implementing a framework to measure and report on the benefits achieved, and transforming the way the organisation collects data – an area of high cost to the organisation.

References

Bentley, E, & Zabala, F (2011, May). *Direction and System Changes Impacting on Data Editing and Imputation at Statistics New Zealand.* Paper presented at the UNECE Work Session on Statistical Data Editing, Ljubljana, Slovenia.

Bycroft, C (2009). *Social Statistics Architecture: The Future*. Wellington: Statistics New Zealand.

Clarke, R (2010). Informal CSTAT Workgroup on stronger collaboration on Statistical Information Management Systems. Paris, France.

Cochran, L (2011). *Household Survey Platform Roadmap Version 1.2.* Available from Statistics New Zealand, Wellington.

Poirier, C (2011).). *The Impact of a Changing Business Architecture on Editing.* Paper presented at the UNECE Work Session on Statistical Data Editing, 9-11 May 2011, Ljubljana, Slovenia.

Seyb, A, Skerret, A & McKenzie, R (2012). *Creative Production Systems at Statistics New Zealand*. Unpublished.

Statistics New Zealand (2010). Statistics New Zealand Strategic Plan 2010-20. Wellington: Statistics New Zealand. Available from: www.stats.govt.nz