Modernization of Statistical Information Systems

Speaker presentation: Australian Bureau of Statistics: Transforming International Statistical Systems /1
Transforming International Statistical Systems

Australian Bureau of Statistics

The environment National Statistical Offices operate in has changed

• Complex interlinked problems
• Expectations of speed
• Expectations of detail
• Expectations of dynamic (personalised) information services
The business of NSIs is evolving

- Was about collection and compilation of particular statistics
- Is now about transforming large quantities of data from multiple sources into cogent, coherent bodies of detailed statistical information
- Quality assurance, transparency, trustworthiness of the information services will remain crucial

Managing information throughout its lifecycle

- Superficially the demand is for websites, dissemination systems, visualisation tools
  - That's a symptom, the generation of the necessary information with rich metadata is the strategic issue
- Preparation needed to describe and disseminate detailed data responsively and flexibly
- Transparency and detailed descriptions cannot be added in at the end of the process, during dissemination
- Assembling consistent metadata during, in fact as a central part of, the production process is the way forward
- "Active" metadata driving automated end to end production systems
Industrialisation / practical level harmonisation amongst NSIs

• No individual NSI can afford the funds, or the human resources to achieve the change in a reasonable timeframe

• Co-operation, collaboration is a solution and that requires genuine harmonisation of methods, processes and IT systems

• Statistical Information Management is essential:
  – so NSIs can truly industrialise / collaborate, and
  – because of the nature of the challenge we are facing
Abstract. National Statistical Institutions (NSIs) face significant challenges in a rapidly evolving information environment. Although considerable networking and information sharing occurs amongst NSIs, there are further opportunities to work together to improve the relevance of statistical frameworks and standards and to more rapidly evolve, share and connect our technology infrastructures. Action in these areas has the potential to help NSIs improve the reach and agility of official statistical services.

1. Introduction

National Statistical Institutions (NSIs) face many critical challenges in the 21st Century. This paper outlines these challenges, presents a vision for 21st Century NSIs, explores the future capabilities required and outlines issues and approaches for working as a community of organisations.

2. Challenges facing National Statistics Institutions in the 21st Century

With the constraints of limited funding, full work programs and aging infrastructure, and the challenges of a rapidly evolving environment, NSIs need to work together to succeed. Two key things NSIs need to do to speed up response to change are to increase the relevance of statistical frameworks, standards and classifications to contemporary issues and to work out a way to more rapidly evolve, share and connect technology infrastructures.

Governments and international organisations around the world make large investments to manage critical issues. They need information to help the formulation of good policy, assess the impact of the investments and refine government programs. The range of information required has expanded significantly and the urgency with which it is needed has accelerated, driven by an increasingly complex and interconnected world. Some of the critical areas relate to society of urgent social problems, global financial systems failure, globalisation, terrorism, climate change and a new emphasis on sustainability.

Complex problems have multiple causes, and require different information views. However, the different views need some coherence to reduce confusion and support effective evidence-based decisions. NSIs have a key role in assisting coherent and logical analysis of data from multiple sources. The fundamental business of NSIs is to assist governments, business and communities to make informed decisions. As institutions, individually and collectively, they work to provide a coherent base of information with accurate measurement of change over time in key economic, social, demographic and environmental indicators. NSIs provide additional value through attention to coherence in a number of dimensions: particular areas of nations; key issues and groups (the young, the aged, indigenous people, disadvantage, health, the economy, the envi-
rontment, etc.); comparison of individual nations with similar nations; and international statistics (either for regions of the world or the entire world).

Citizens are expecting more from government. There is a resurgence of social conscience in particular fields, perhaps enabled by improved education and access to information and like-minded people through technology. There is often strong community interest in assessing the success of government in tackling problems and in working with government to design effective programs. The mechanisms for the community to raise issues with government are also expanding in range and power with the increasing use of the internet to circumvent traditional media and public relations mechanisms. The debates are becoming more open, visible and volatile. Governments are increasingly looking at how to use web 2.0 to transform government services and interaction.

The revolution in the speed with which information is becoming available is creating new information products and new ways of combining and using information. These alternative sources of information are sometimes updated in real time and are available with fewer constraints and with greater detail than the information provided by NSIs. In addition to the wealth of administrative data generated by business activity, there is an ever-increasing pool of data generated from personal devices, sensors, instruments and computers. Examples of sources are retail scanners, scientific equipment, imaging systems, transport systems, telecommunications networks and even metrics collected about people’s use of the Internet. There is a blurring of the boundaries of official statistics. A number of organisations are publishing statistics, sometimes released more frequently and often widely used to complement or pre-empt the official versions.

Many governments are unwilling to provide additional investment to maintain and expand the core existing authoritative baseline data, but are often willing to invest in rapid responsive approaches which provide answers to new challenges and questions they are confronting.

The current slow and considered approach to developing international and national statistical frameworks, standards and classifications is a significant barrier to rapid change. While the approach suits the need for comparable statistics over time, it can take decades and great expense for new versions of existing standards to be implemented across the world, and for many pragmatic reasons some countries do not ever implement them. As a result, current official statistical data is not always seen as fully relevant to new areas of information enquiry, particularly when compared to some of the more agile and informal information sources. The outdated nature of frameworks can also complicate the analysis task, as users attempt to make their own adaptations to suit the real world. There is a need to consider approaches which preserve the comparability of statistics but enable rapid ways of producing additional statistical models to reflect contemporary changes taking place in economy and society and in different national interests, international interests, contemporary activities and issues.

Tools and approaches that have worked well in the past may not serve NSIs so well in the future. Much of the technology effort of NSIs has been focused on supporting individual processing models and, although NSIs share international frameworks, information about approaches and sometimes share technology, many of the systems are developed for a single organisation’s purposes. Limited technology resources are consumed by cycles of development and modernisation attempts.

Predicting an environment dominated by the knowledge economy and the internet in 10 or 20 years time is complex. As a result, NSIs need to work together to work out what they need to do in order to survive and prosper in this evolving world. Unless there is collaborative effort among NSIs, it is not certain that they will remain at the heart of official information for societies. Facing a similar challenge, newspapers who fail to adapt to new business models are disappearing.

3. A Vision for 21st Century NSIs

NSIs collectively have an important role in evidence-based policy making through enabling comparison of data across regions, countries and the world. By supporting international statistical standards, they provide governments with the opportunity to assess the success of similar policy interventions in other countries, whether prior to making significant investment, during implementation or once programs are established. For global issues, the importance of coherence across the international statistical system is even more pronounced.

The effective NSIs of the future will continue to provide a framework of stable, trusted, regular and coherent base of key national and international statistics. In addition, they will be able to rapidly combine data from official statistics and a wide variety of other sources to
produce coherent information relevant to current and emerging issues. They will be responsive and agile at mounting new collections to fill gaps and to answer new questions. They will maximise the availability of information through the innovative use of methods and tools which allow data to be combined and analysed while ensuring privacy and confidentiality requirements are preserved. They will further develop and share strong information management capabilities and work effectively with others to encourage innovation in information management and use.

As new information sources are developed or discovered they will incorporate the most useful into the enduring fabric of information available to their governments for informed decision making and into the base of important international statistics. NSIs will take steps to shape how new information sources are developed and used. They have tools, people and credibility to be a creative partner in this field.

4. Future capabilities

The Australian Bureau of Statistics (ABS) has been considering what new capabilities might be required to meet these challenges, how NSIs and others might work together to obtain them, and what actions should be taken now.

4.1. From static data products to “common information services”

Typical NSI websites contain a mix of information products. These include electronic publications supported by more detailed data products in a variety of formats (spreadsheets, data cubes, data warehouses) and other products such as confidentialised microdata files and information about standards and classifications. The standards and classifications are often not explicitly linked to the data products which use them, and searching and finding specific data is often a manual process, unique to each organisation’s website.

In the future it should be possible to bring together information automatically and repeatedly, in ways we cannot imagine and at speeds we can’t envisage. NSIs need to support the demand for assembly of data from different sources, including their own, and at the same time protect the confidentiality of the data. They have, individually and collectively, significant data assets of value to others. Society and governments are exploring ways to unlock public sector information and a shared approach by NSIs has the potential to assist this drive. They should consider how to establish common information services to attract development investment from the commercial, government and non-government arenas to support this future. If enough momentum is achieved through a common approach, other organisations and individuals will choose to connect their services and innovations as well. By working as a group this will not only make it easier for these organisations to deal with NSIs, but should result in richer services and reduced costs.

To be successful, with a “common information services” business model:

- NSIs would have common metadata services on their websites. The metadata services would provide descriptions of the data elements in each information product including definitions of the rows and columns of each table or file, links between data element definitions and data products, links to data quality information, descriptions and concepts. Visitors to NSI websites would be able to browse or search the repositories to find information about which data elements are available, in which combinations, and from which specific data products. It is important that the information is also machine searchable and provides metadata services to other applications both on NSI websites and elsewhere, allowing clients to rapidly assemble data from multiple sources.

- As a companion to a common metadata service, NSIs would implement one or more common data services. NSIs are already struggling to produce and quality assure the many different cross tabulations that users require and the process of individually crafting these takes resources and time, slowing down the ability to respond quickly enough. Many NSIs have recognised that the solution is to invest in dynamic data services which allow clients to specify the information they require and generate this dynamically from ‘raw’ data. Examples of this already exist (for example, the ABS Census TableBuilder product).

- As well, NSIs would have common confidentiality services. Because of the need to protect confidentiality, NSIs typically hold a great deal more information than they can release. There are complex approaches for ensuring that data released meets confidentiality requirements. For example, the Census TableBuilder product applies a perturbation method which ‘works’ for simple count data but is not directly applicable to other
types of data. Other types of data might include continuous variables, longitudinal data or composite data sets. There is a need for several different confidentiality methods, depending upon the characteristics of the data being accessed.

- Integrating common data, metadata and confidentiality services, NSIs also need to design approaches which support the ability to dynamically link data and services from elsewhere. Dynamic data linking services offered by NSIs and though their websites should allow comparison and combination of data from multiple sources.

- NSIs would partner with industry and government to stimulate and enable new ways and means for individuals, businesses and communities to gain access to and to use publicly available data.

The combination of these services would also provide important functionality for future access to more sensitive microdata. ABS’ current Remote Access Data Laboratories (RADL) provides a batch submission process, which allows researchers to submit programs against data that is not available to them in raw form. It has a heavy reliance on assessing and pre-confidentialising the unit data being analysed. Future development aspirations are to move this towards an interactive service, with the focus on dynamically confidentialising outputs before it is provided to the client. This is far from a trivial challenge. However, NSI expertise in methods for “confidentialisation” of data is relevant to all data custodians. Further development of common approaches could be used in many organisations to support information sharing, to increase access to statistical data and potentially to support new data linking approaches across multiple organisations’ data.

There will be many very important cases where the information required by clients will still be built from pre-assembled data products. Some obvious examples are time series of data, products such as the national accounts where the information is assembled from multiple sources within a standard framework, products like price indices where the assembly process involves complex calculations such as chain linking, and very large datasets where the processing cost of creating everything dynamically might be too high. NSIs should also share investigations into approaches for managing these cases and providing them to clients through the same metadata and data services mentioned above.

4.2. From publications to communication

Publications from specific data collections will remain important, but there will be a subtle change in their role. They will become alerts, to warn data users that new information has become available. These alerts would announce the key headline stories in the data in a way that other agents, including the media, can easily convey to the community. These alerts should also direct clients via active links to the richer data available through the data and metadata services.

This separation of the data services from the presentation process will support the use of new ways of exposing and communicating the meaning stored in statistical data, such as those described by Professor Hans Rosling, Professor of International Health at Karolinska Institute and Director of the Gapminder Foundation, which developed the Trendalyzer software system. There are many other examples of innovation in this area (for example, mashupaustralia.org). The techniques available will continue to evolve. Successful NSIs in the future will provide their information in ways which support and encourage these different ways of communicating the underlying messages in the data.

4.3. Support for transaction data flowing at a much higher volume

The volume and sources of real-time or near real-time data are increasing exponentially. Different industries and sources use different metadata and data standards, but these are often common within areas of interest across the world. For example, to support banking transactions for customers everywhere, there are standards used by many banks. Major manufacturers of business equipment such as cash registers and scanners operate globally and use standards such as barcodes and RFID. This is also true for telecommunications, travel, and manufacturers of other equipment used within societies (such as traffic lights) and for spatial information (including satellite data, Google maps and GPS data). A collective approach to determining the best ways for NSIs to incorporate this data into the information stores for their countries would be useful, as would discussion about how to judge which data might be retained to inform future policy issues. This work could include determining common approaches for data discovery, data exchange and, perhaps, data warehousing/archiving and access to data.

4.4. Ability to rapidly incorporate new issues and views of data into standards and classifications

The use of standards and frameworks will continue to be important to support coherence. However NSIs
need a way to rapidly incorporate new issues and views of the data, without perturbing the base. As well as supporting international standards, NSIs have to operate in different environments with some different areas of key concern. Users of the data also need to adjust or augment the standard classifications to better suit their analyses, situation, environment or to compare the data with pre-existing data provided by others. Classifications need to be dynamic and supported by automatic ways of transforming the data quickly, such as automatic coders. Automatic approaches are important because they enable multiple views without significantly increasing cost or time. This includes rapidly recoding existing datasets, coding large amounts of “un-coded” data, “multi-coding” data or recoding data on demand. ABS has undertaken research into some promising new techniques, using machine learning. Because coding techniques have broad applicability and the function can be quite discrete, ABS would prefer to collaborate in the development of new coders based on these techniques. Use of common facilities in multiple organisations (including those who provide data to NSIs) would also support consistent coding and higher data quality.

4.5. ‘Rapid-response’ capability

Often existing collections of data are not sufficient to answer new questions. If statistical planning processes are operating well, we should have the basic baseline data to shed light on an underlying social, economic, or environmental issue. However, when a policy direction is set, and an “experiment” is underway, governments will need up to date information about the outcomes and how these relate to their investment. They will need to make comparisons with what has happened in the past, or is happening elsewhere.

Successful NSIs will be required to mount new collections, or adapt existing ones, quickly and responsively. These new collections may be ‘traditional’ data collections run by an NSI, or they may involve diving into data which is available through other organisations’ systems or websites, or both. Culturally, NSIs have a strong bias towards stable data series, and developing collections using careful evaluation of a wide range of user needs, thought into the underlying conceptual framework, careful development and testing of questions and questionnaires, samples designed for optimal efficiency, processing systems tailored and tested in advance, data edited comprehensively, information analysed carefully and publications constructed, presented etc. All these are valuable approaches, but additional and different approaches will also be crucial.

Technology or methodological changes will help to achieve more rapid response, but changes to the mind set, the capability and the capacity of skilled staff are also fundamental. NSIs need to be able to form agile teams with a combination of relevant expertise (statistical, analytical, technical and policy related) and equip them with the mandate and tools required to respond quickly. In addition to the capabilities already mentioned, some of the tools and infrastructure required are question modules, web survey facilities, call centres for follow-up or interview, ETL (extract, transform, load) tools, additional confidentiality approaches, data warehouses to store the results and web site facilities to share the results. We need agile legal approaches that can help to remove barriers to data exchange such as template data sharing agreements and licences such as Creative Commons. We need engagement mechanisms which make it easy to involve people (with academic, commercial, NSI or other backgrounds) in the work of these teams and if possible shared approaches to the development of these people.

4.6. Connecting processes and passing metadata and data easily between them

Once an initial rapid response has proved its worth, provided that the underlying demand remains, successful NSIs will have to improve and institutionalise the new information stream into their existing core production processes.

A commitment to a standard way of describing information using a format (such as SDMX and DDI) would allow NSIS to connect statistical process steps together more easily over time as key parts of systems are developed or re-developed. It would also enable others to provide functionality more easily.

4.7. Analysing assemblies of data

Tools being developed in the Web 2.0/ semantic web space hold out the promise of new and more efficient ways of analysing metadata stores. The use of a standard approach such as SDMX/DDI containing links between particular aspects of information (eg. data collections which use a specific data element, questionnaires and forms used to collect the information, data quality statements about the collection) will support improved and better directed investigations by skilled statistical staff into any discrepancies between disparate data sources. It will also provide insights into how the data should or should not be used.
5. Working together

The areas mentioned above are examples of where NSIs and others might profitably collaborate and innovate as a community. Each organisation is likely to have a view of where investment in change is required and the development of a program for collaboration on innovation would require NSIs to explore these views and agree on some initial candidate areas.

As individual organisations, current approaches are not as adaptive as needed. Common strategies have been to either attempt large “whole of system” transformation programs, or to deal with particular parts of the process at particular points in time. NSIs have tended to engineer processes and systems for current problems rather than looking at what is needed for the future. This strategy is only appropriate for a relatively “steady state” environment. The level of change required is profound and it is unlikely that any NSI will be able to make the investments required to facilitate change at the rate required on its own. The technology environments that are in use today are likely to be totally irrelevant in the future, so the approach of redeveloping large components of infrastructure every 5–20 years will not serve organisations well. Instead, NSIs need an approach that allows them to evolve smaller components of statistical infrastructure much more rapidly, and which allows them to inject and attach new functionality as it appears in the environment.

NSIs need to encourage innovation in their organisations but also excite and use interest in statistical innovation in academia, other information providers, the ICT industry and in creative individuals, so that they don’t have to fund all of the change directly themselves. They need the ability to easily combine these innovations along with other tools from the knowledge economy, as they emerge. By acting as a group, it should be possible to spark interest amongst others in the research and academic communities and in the ICT vendor community. Researchers, academics and students probably have the time, energy and interest to innovate and ICT vendors tend to move into areas that address common issues. There are many groups of dedicated developers/enthusiasts who like to innovate in their area of interest and who are connected via the web, including group GGobi who like to develop new ways of analysing data and numerous groups of open source developers with specific interests (see sourceforge.net). These enthusiasts are a vast untapped workforce. In the commercial arena, Apple has published details of how to write small applications for the iPhone/iPod and allows developers to place their applications in an on-line store. Every time a particular application is downloaded, the developer receives a small royalty fee. In the short time since the iPhone was released, many thousands of applications have been written. Many are trivial, but there are increasing numbers of useful applications becoming available. Use metrics are collected so that royalties can be paid and as a consequence, the popular applications are well known. Apple have stated that there are now more than 85,000 applications available in the App Store and more than 2 billion iPhone and iPod Touch applications have now been downloaded. Although these examples may not work for NSIs, they should explore these approaches to stimulate contributions from a wider community.

As an international statistics producing community, NSIs need to try to build standard approaches together while they still have the relevance to force a standard. It is important to make some decisions as a community about what needs to be standard, choose some standards even if they do not represent the best “technical” approach and then evolve these over time if required. As a community of organisations the “buying power” can be amplified if agreement can be reached on common requirements and design. It may be difficult to reach agreement on the requirements for large components (such as an entire input data warehouse), but if a start is made with components for some of the smaller, relatively discrete functions of statistical processing, these could be built into larger components over time.

To be effective, the program needs to support change through processes, methodology, technology and people. The people dimension could include opportunities for sharing staff across NSIs (either through short or medium term secondments or through the use of virtual teams). It is likely that the skills, capability and careers of staff in the organisations would be enhanced through the program, particularly if collaboration on the development and delivery of relevant training is included. Many NSIs also have active international statistical aid programs and there is potential to link these to an innovation program, with benefits for both programs.

Although there are significant institutional and other barriers, NSIs should be encouraged by the success of a number of open source and cross government projects. More effective collaboration has the potential to reduce the costs for each statistical organisation of building and maintaining software and could allow stronger investment in particular areas of work. There is also the potential to accelerate the availability of statistical infrastructure above what could be achieved individu-
Fig. 1. Generic Statistical Business Process Model (top level) showing some areas of current ABS interest.

ally and at the same time to enhance support for the statistical systems of developing nations. Combined with efforts to enhance statistical skills, a stronger base of statistical infrastructure shared with other organisations should assist statistical leadership aspirations, reduce development and cost duplication, and could ultimately drive the emergence of a truly integrated global statistical information system.

International organisations, particularly the OECD, the UN Economic Commission for Europe (UNECE)
and EUROSTAT, have driven frameworks such as the Generic Statistical Business Process Model (www.unece.org/stats/gsbpm) and the Common Metadata Framework, and recommended metadata standards such as DDI and SDMX through active engagement with NSIs. These are a very useful starting point providing a top-down view of NSIs’ business and important guidance regarding the ways to link processes and components. So that components will integrate, there is a strong need for a standard way for the components to share data and metadata. A number of NSIs and international statistics organisations support the use of DDI and SDMX for this purpose. Part C of the Common Metadata Framework http://www.unece.org/stats/cmf/ provides further information.

What is now needed is a practical way to work together to address some of the challenges outlined above. Many NSIs have endorsed the Generic Statistical Business Process Model (Fig. 1) as a reference model. This model provides a useful description of the business in a way that could assist discussions about the development of shared approaches. It is also useful within NSIs for discussing the areas (“hot spots”) where investment is most required (the highlighted boxes in the Fig. 1 are examples of current interest to the ABS).

Rather than commencing a huge high risk project to develop a single solution covering the entire statistical business process model, an incremental approach is recommended. An approach is outlined in the Fig. 2.

The approach could commence by establishing a small set of projects to develop the first group of common components. A set of criteria could be established to assist this selection and might include attributes such as: contributes to ‘future proofing’ of NSIs; required as early ‘building blocks’ for future collaboration; commits to a required standard; delivers tangible benefits not only to individual NSIs but to the broader statistical community; usefulness to more than one participant; ability to complete in defined time period; ability to incorporate easily into existing approaches, speed of return on investment, etc.

Each project could progress through a consistent process of determining shared requirements, preparing a project proposal, common design, cost and benefit analysis, team establishment (virtual or co-located), project plan and milestones for delivery. Participants would probably not contribute directly to all of the projects.

It would be important to select an initial set of projects that have a good chance of success. It is envisaged that, each year, National Statisticians acting as
'venture capitalists', could on the basis of results, assess whether to continue and expand the program and could also determine which new projects to support. As the program gained momentum, there would be many other potential collaborators for this work. These include other government, non-government and commercial organisations within our nations, the full range of NSIs, international statistical organisations, academia, global ICT organisations and creative individuals. All of these are likely to have capabilities, components or plans that might be useful. The development of components of the statistical process by one NSI which are then shared with others sometimes works, but in the past this has tended to be opportunistic rather than planned. There will be many solutions that fit into each box of the GSBPM and this is good – competition between NSIs, commercial providers and others will only serve to accelerate the change and innovation required.

6. Summary

In summary, NSIs need to consider the critical challenges they face and determine the types of organisations they need to become to meet the future successfully.

The challenges include the increasing range of information needs, revolution in the amount of information available, declining budgets, full work programs and aging infrastructure.

By exploring the new capabilities and standards that might be required, and by working together, there are opportunities for NSIs to share and develop practical solutions which improve the reach and agility of official statistical services.

References