Widening the data net: NSO leadership role

Sarah Conn
Bachelor of Actuarial Studies / Bachelor of Economics
Master of Applied Statistics

Contents

The author would like to acknowledge and thank Gemma Van Halderen, Merry Branson and Linda Fardell for their interest and support in preparing this paper.

1. Introduction

Statistics form the fundamental underlying knowledge that policies and strategies are based on. It is important that the statistics available are accurate, timely and coherent so that policies can be developed and implemented with clear aim and direction. Data integration could be the key to solving many problems and complex issues facing society. There is however, a need to understand privacy and confidentiality. National Statistical Offices (NSOs) have a long history in managing privacy and confidentiality and have much to offer.

Data integration is a worldwide phenomenon and can better utilise the wealth of administrative, survey and census data available, both efficiently and effectively. Integration brings together microdata based on common information between the sources to describe relationships between variables. This is then used to investigate and report on social and economic conditions, behaviours and outcomes. This is becoming increasingly important in light of the findings of the Stiglitz Commission's report on Measuring the Economic Performance and Social Progress (Stiglitz, 2009).

2. Feasibility of Data Integration

There is a need for detailed information in order to understand increasingly complex issues, formulate and refine policy and make good investment decisions. The range, detail and complexity of information required by the public is expanding on a continuous basis. Data integration can help meet the ever increasing demand for statistics.

2.1 The Benefits of Linked Datasets

Data integration allows more knowledge about the impact of policy interventions to be gathered and investigations of the short and long term outcomes for all population groups to be conducted prior to implementing change. Integration can make use of large datasets, such as the Population Census and administrative data, where the sum of the two sources is greater than each one alone. It opens opportunities for more effective research to solve very difficult and complex problems, such as social inclusion and climate change. For new interventions to be implemented effectively, we need detailed information to base the interventions on and a way of measuring the effectiveness of these interventions.

There are two ways of gathering information to solve complex issues: direct collection and utilising existing datasets. The advantages of each method are described in the table below:

Figure 1. Advantages of different data collection methods for solving complex problems

	Advantages
Direct Collection	- Can tailor the collection to cover specific needs - Respondents know the project they are contributing to - Privacy and confidentiality is addressed for each collection
Utilise Existing Datasets	- Data already exists - No further respondent burden - Less expensive than direct collection

Direct collection addresses specific information needs, but does put an increased burden on respondents, especially for longitudinal surveys. Utilising already existing datasets on the other hand, provides more timely analysis and is more cost effective.

Data integration provides an opportunity to increase the value of existing data sources for improved decision making. It is often the case that existing data doesn't have enough information to address specific questions, but data integration will help to resolve this issue by bringing datasets together to maximise the potential of the existing data.

The benefits of data integration are applicable internationally and really come through in the ability to address new and more complex questions that face the community every day. Linked datasets can be used to create richer datasets and therefore provide more robust results, as well as providing the ability to utilise sensitive information such as addictions or criminal behaviour that can be difficult to collect directly. Integrated datasets will also allow for more longitudinal analysis and an improved ability to adjust for bias or loss to follow up than in current surveys.

There is a need to promote the reasons for increased use of existing data and data integration within the community. The understanding of relationships between data and socio-economic outcomes for different population groups can be enhanced through data integration and the public needs to be well informed about the importance of this for policy and community needs to be addressed.

2.2 Governance Arrangements in Australia

Policy and research questions do not fit neatly within agencies boundaries in Australia. Each agency is responsible for their own datasets and upholding obligations associated with them, such as privacy and confidentiality. As the sharing of data would assist in meeting some of the less clear cut research questions though, there is a strong need for a framework that provides a safe and effective environment for data integration, to facilitate data linking while protecting the security and confidentiality of the data.

Australia has developed a framework which provides governance arrangements to support the more effective use of integrated Commonwealth datasets across all sectors.

The Australian Government framework for data integration involving Commonwealth datasets for statistical and research purposes, has seven principles to lead and support growth in the field of effective use of data (NSS, 2010):

Figure 2. The Seven Principles for Integration of Commonwealth Data for Statistical and Research Purposes

Principle 1: Strategic Resource

Data is a strategic resource and there is a need to maximise research using new and existing data.

Principle 2: Custodian's Accountability

Data custodians recognise their accountability for the data and maintaining confidentiality.

Principle 3: Integrator's Accountability

For each integration project there will be an 'integrating authority' who will be nominated to manage the project.

Principle 4: Public Benefit

Statistical data integration will only occur where there will be significant benefit for the community.

Principle 5: Statistical and Research Purposes

Statistical data integration is only to be used for statistical and research purposes.

Principle 6: Preserving Privacy & Confidentiality

Policies and procedures used in statistical data integration are to minimise the potential impact on privacy and confidentiality.

Principle 7: Transparency

Statistical data integration will occur in an open and accountable way so the public knows how government data is being used for statistical and research purposes.

3. Current Projects using Integrated Data

3.1 International Health Data Linking Network

The International Health Data Linking Network was established in 2008 and consists of members from Australia, Canada, England, New Zealand, Singapore, Scotland and the United States of America (IHDLN, 2010). The overall goal of the network is to facilitate communication and share information between linkage centres to improve community benefit from data integration projects. This will be achieved by working together to build better workforce capabilities, improve the quality of datalinking methods, share information and resources concerning common challenges and promote high standards of governance. There has also been some work done by this group to initiate internationally comparable data linkage studies, but this is still in its early stages.

3.2 Data Integration in the Field of Australian Health

The first data integration team in the field of health for Australia was the Western Australian Data Linkage System (WADLS), developed in 1995. This was the first step towards effectively using integrated health datasets to expand the range and depth of statistics, conduct more detailed analysis and provide more informative results. The WADLS has provided data for more than 300 research projects on causes of disease, clinical needs, patterns and costs of care and outcomes of health services (WALDS, 2010). Projects like these have meant that vulnerable people have been helped and the analysis provides insights for the future.

More recently a Population Health Research Network has been established to provide researchers with a more diverse range of information through rich data sources. This initiative of Australian Governments to implement a national data linkage strategy for health supports national research to improve health and health care services in Australia (PHRN, 2010).

Data integration in the field of health currently involves a central body for each State and/or Territory, which does the linking of data using a master linkage key system. This means that only the central body views any personal identifiers such as name and address, but they do not receive any of the other variables in the dataset at any stage. This is known as separation and is used as a privacy preservation mechanism. The data custodians receive a unique linkage key that has been created and apply it to their dataset. This key is then released to researchers for specific data integration projects as they are approved. No researcher ever receives detailed information such as name and address.

3.3 Australian Bureau of Statistics

Australia's NSO has a number of data integration projects underway, utilising the Australian Census of Population and Housing. The projects have each been shown to have significant potential for community benefit and general acceptance has grown regarding the Australian Bureau of Statistics (ABS) undertaking these data linking projects (ABS, 2005, 2006, 2010a) following extensive community consultation. Some of the 2011 Census of Population and Housing projects are as follows:

Linking Census over time
Creating a Statistical Longitudinal Census Dataset is a current project being carried out by the ABS. It involves bringing together data from multiple Censuses over time, where data will be linked for 5% of the Australian population. The integrated data will be a rich statistical source for information about Australian people and households, not only at a point in time, but also over time. This aids the ability to investigate patterns of how social and economic conditions change over time, and provides insight as to why pathways may lead to certain outcomes and how they vary for different population groups. This will also help policy makers assess the social and economic benefits of intervention strategies.

Migrants settlement outcomes
This project involves integrating Census of Population and Housing data and data from the Department of Immigration and Citizenship's migrant settlement outcomes database. It aims to provide detailed information on migrants to develop new strategies for promoting social cohesion and the effective settlement of migrants into the labour market.

Indigenous life expectancy
Improving measures of life expectancy for Indigenous Australians is a key component of the governments 'Closing the Gap' strategy. The concept being explored by the Council of Australian Governments (COAG) is the ability to improve the identification of the Australian Indigenous population through integrated data sources and investigate whether indigenous status is reported consistently. The project also looks to improve data by adjusting for differences in indigenous status between data sources, such as death registrations and the Census of Population and Housing. Indigenous status is often under reported and through data integration the data sources and input to Indigenous life tables can be improved.

Enhanced cancer statistics
The aim of this project is to link cancer registrations data to the Census of Population and Housing to provide additional data for cancer statistics that are not currently available. Some of the more complex questions that could then be answered using the linked data include: Are some population groups more prone to a certain type of cancer?; and Are there different mortality rates from cancer for different population groups? Another goal is to improve the measure of Indigenous status in the cancer registrations, which will lead to improved estimates of cancer incidence, cancer risks and mortality in the Indigenous community.

4. Role of a National Statistical Office

The NSO in Australia collects, analyses and disseminates statistics and related information. It is legislated to avoid the duplication of data collected by official bodies and to achieve the maximum utilisation of data for statistical purposes. Data integration is consistent with both of these functions.

Currently NSOs use well considered frameworks for data collection to maximise quality and consistency over time. With rapid changes in technology and requests for more detailed data, there is a need to adapt to change quickly and improve how new data sources are developed. NSOs have the tools, people with relevant skills and the credibility to take a lead role in utilising existing data sources. As well as integrating data sources, NSOs can also assist by providing guidance and support for users to extract relevant information from complex data sources and help users draw accurate conclusions from analysis.

NSOs have a wealth of information, such as Population Census data, which is a rich dataset of demographic characteristics, such as family structure, country of birth, year of arrival, Indigenous status, education, hours of work, occupation & industry, income and housing. This data is extremely important for looking into how people move through the life course and what factors influence the transitions they make. Patterns in the data allow us to identify circumstances that influence outcomes. Combining the Population Census with other data sources facilitates socio-demographic analysis.

NSOs provide an unbiased source of statistical knowledge and with data integration, can expand the data available to further inform decision making. Investigations can then be done on social, economic and environmental issues and strategic interventions can be assessed. Some NSOs already use administrative data as their primary source of information for statistical purposes. Scandinavian countries, for example, have a register based system, combining multiple administrative datasets to develop a rich source of information for statistical purposes. This is accepted as general practice by their communities and is recognised as an important source for statistical analysis.

4.1 Across Government Liaison

The role of an NSO is to coordinate official bodies in relation to statistics and provide a statistical information system to serve the public. The fundamental principles of official statistics (UNSD, 2006) form the basis for this system, to work efficiently in producing quality statistics that are essential for influencing policy decisions in the economic, environmental, demographic and social fields.

Strong liaison between government agencies is required to collaboratively build a best-practice guide for data integration and analysis that maximises the use of existing data for statistical and research purposes. The ABS is already promoting a system to work more

efficiently and effectively with other Australian Government agencies, called the National Statistical Service (NSS). As members of the NSS community, government agencies are working in partnership to build a safe and effective environment for data integration involving Commonwealth data.

4.2 Privacy and Confidentiality

A key concern currently facing data custodians is the potential for invasion of privacy. People are sensitive to the increased use of personal information by the government, even if it is de-identified data. There is strict legislation and policies for maintaining the confidentiality of personal information in Australia, however there is also the need to balance privacy with the public good that can arise from statistical analysis. Furthermore, as complex policy questions, such as climate change and social inclusion, are not within one agency's responsibilities, we need to find a balance between maximising the value of data and minimising privacy concerns with integrated data.

The integration and analysis of datasets must be done by a highly credible agency, with commitment to best practice for maintaining privacy and confidentiality. As credible, apolitical, trusted institutions with a history of maintaining confidentiality, NSO's understand the importance of implementing confidentiality methods effectively and how failure to do so creates great risks for the community and governments. The fundamental principles of official statistics provide a good framework for NSOs to operate effectively (UNSD, 2006). This includes guidance for confidentiality to be maintained and data to be used exclusively for statistical purposes.

4.3 Advances in Technology

Advances in technology have the potential to be great for data integration, but they could also be a threat to the disclosure or misuse of personal information. With the advances in technology, there is the potential for improved access to microdata, which is essential for the effective use of administrative data and data integration. Internationally, NSOs are collaborating to improve microdata access infrastructure.

The ABS is currently working on a project called the Remote Execution Environment for Microdata (REEM) to allow statistical analysis of microdata while maintaining the confidentiality of respondents (ABS, 2010b). The system applies automated confidentiality methods to generated outputs, allowing improved access to de-identified microdata. There has been increased demand for access to detailed data recently and REEM will achieve this access in a flexible way, encouraging informed decision making through a wider use of ABS data for social and economic research.

5. Conclusion

There is a significant opportunity for NSOs across the world to take a lead role in the area of data integration. Data integration will reduce the duplication of work, lower respondent burden and provide data in a more timely manner by utilising already existing data. With a wealth of administrative data, and a revolution in the speed with which data is available, NSOs can provide leadership to facilitate data integration and assist with more cost effective improvements to statistics. The community benefit of data integration can be shown to be large. NSOs can take an active coordination role to establish governance arrangements that provide a safe and effective environment for data integration.

People are asking for more and more data, in an increasingly sophisticated way. There needs to be more recognition of the potential for administrative data, as well as the traditional survey data, to be used as a source of official statistics to address strategic needs. There is a greater demand from both agencies and the public for detailed statistical data from NSOs, and the opportunity exists for data integration to encourage agencies to work together more effectively. Collaboration is essential to the success of data integration across government agencies, especially when complex questions do not fit neatly within one agency. The liaison between agencies will provide connected thinking and a whole-of-government perspective for informing policy decisions. NSOs are in a good position to drive the emergence of a truly integrated statistical information system which will help develop strategies to achieve positive outcomes and assist policy makers assess the benefits of proposed interventions.

Data integration provides a key opportunity to utilise the wealth of data already collected in the community. More detailed analysis can then be used to inform policy, and to paint a national picture of social, environmental and economic conditions. Censuses are a key resource in this field as the detailed demographic data available would be advantageous to add to any other dataset. NSOs are in a very good position to take on a leadership role in data integration.

6. References

1. Australian Bureau of Statistics (ABS), 2005, *ABS Discussion Paper, Enhancing the Population Census: Developing a Longitudinal View 2006*, cat. no. 2060.0, www.abs.gov.au
2. Australian Bureau of Statistics (ABS), 2006, *ABS Information Paper, Census Data Enhancement Project: An Update 2006*, cat. no. 2062.0, www.abs.gov.au
3. Australian Bureau of Statistics (ABS), 2010a, *ABS Information Paper, Census Data Enhancement Project: An Update October 2010*, cat. no. 2062.0, www.abs.gov.au
4. Australian Bureau of Statistics (ABS), 2010b, *CURF Microdata News, November 2010, Update! ABS new directions for microdata access*, cat. no. 1104.0, www.abs.gov.au
5. Australian Bureau of Statistics (ABS), 2010c, *Annual Report 2009-2010*, www.abs.gov.au
6. International Health Data Linkage Network (IHDLN), 2010, *Second Meeting of the International Data Linkage Network*, Manitoba, Canada, March 9th 2010, www.ihdln.org.au
7. National Statistical Service (NSS), 2010, *High Level Principles for Data Integration Involving Commonwealth Data for Statistical and Research Purposes*, www.nss.gov.au
8. Population Health Research Network (PHRN), 2010, www.phrn.org.au
9. Stanley, F., Royal Statistical Society, Significance Magazine, 2010, *Privacy or public good? Why not obtaining consent may be best practice*, www.significancemagazine.org
10. Statistics Sweden, Conference of European Statisticians, 2001, *Statistical Microdata – Confidentiality protection vs freedom of information*, www.unece.org/stats/publications/statistical.confidentiality
11. Stiglitz J., Sen A. & Fitoussi J., 2009, *Report by the Commission on the Measurement of Economic Performance and Social Progress*, www.stiglitz-sen-fitoussi.fr
12. United Nations Statistics Division (UNSD), 2006, Department of Economic and Social Affairs, *Fundamental Principles of Official Statistics*, unstats.un.org/unsd/
13. Western Australian Data Linkage System (WADLS), 2010, www.datalinakge-wa.org