Linking Information to the Australian Bureau of Statistics Census of Population and Housing in 2011

Graeme Thompson
Australian Bureau of Statistics, ABS House, 45 Benjamin Way, Belconnen ACT 2617, Australia, graeme.thompson@abs.gov.au

Abstract: The Australian Bureau of Statistics will be undertaking a suite of data integration projects linking ABS and non-ABS data to the 2011 ABS Census of Population and Housing. The process of undertaking the integration projects can be mapped to the Generic Statistical Business Process Model (GSBPM) to aid in discussions of developing statistical metadata systems and processes.

Keywords: data integration, business process, census, Australia

Views expressed in this paper are those of the author, and do not necessarily represent those of the Australian Bureau of Statistics. Where quoted, they should be attributed clearly to the author.

1. Aim of the Paper

The aim of this paper is to provide an understanding of how the Census Data Enhancement project will be linking both ABS and non-ABS data to the ABS Census of Population and Housing conducted in 2011, and how the GSBPM might be used for data integration projects.

2. Purpose

The functions of the Australian Bureau of Statistics as specified in the *Australian Bureau of Statistics Act 1975* (ComLaw, 1975) include the maximum possible utilisation, for statistical purposes, of information, and means of collection of information, available to official bodies. Aligning with this function the Australian Statistician has set one of his key priorities for the organisation over the last 4 years to be "implementing a safe and effective environment for the use of, and integration of, microdata for statistical and research purposes".

3. Census Data Enhancement (CDE)

A key project for the ABS is the Census Data Enhancement (CDE) project. This term is used to describe several projects which link ABS and non-ABS data to the ABS Census of Population and Housing.

Commencing with the 2006 Census, the ABS began the CDE project to enhance the value of the Census data by bringing it together with other datasets to leverage more information from the combination of individual datasets than is available from the datasets separately.

There are five major components to the 2011 CDE project:

1. Bringing together 2011 Census data with a small number of predetermined datasets during Census processing using name and address, for quality studies;
2. Bringing together 2011 Census data with a small number of predetermined datasets during Census processing using name and address, for statistical studies;
3. Wave 2 of a 5% Statistical Longitudinal Census Dataset (SLCD);
4. Bringing together the SLCD with other datasets without using name and address for statistical and research purposes; and
5. Bringing together 2011 Census data with other datasets without using name and address after Census processing.

A fundamental aspect of the CDE project is the management of confidentiality and privacy.

The ABS Census of Population and Housing is a cornerstone of official Australian statistics. The co-operation of respondents is critical in ensuring high quality statistical outputs. One measure that encourages respondent participation are specific undertakings that the ABS makes regarding the Census around the destruction of Census forms and the deletion of name and address information once Census processing has been completed. Some of the CDE projects require the use of name and address for linking purposes, so these projects can only be completed while the Census is being processed (from Census night until approximately 15 months later). An undertaking has been given to the Australian public that linked files created using name and address will be deleted once their specified purpose has been met.

The CDE project was first proposed during the 2006 Census cycle. The ABS held extensive consultation (including a Discussion Paper: Enhancing the Population Census: Developing a Longitudinal View, 2006 (ABS, 2006a)) around the scope of the project and commissioned a Privacy Impact Assessment (Waters, 2005) by an independent body. The Australian Statistician then determined the scope of the CDE project for the 2006 Census cycle, and this was published on the ABS website as an Information Paper ABS Cat. No. 2062.0 (ABS, 2006b).

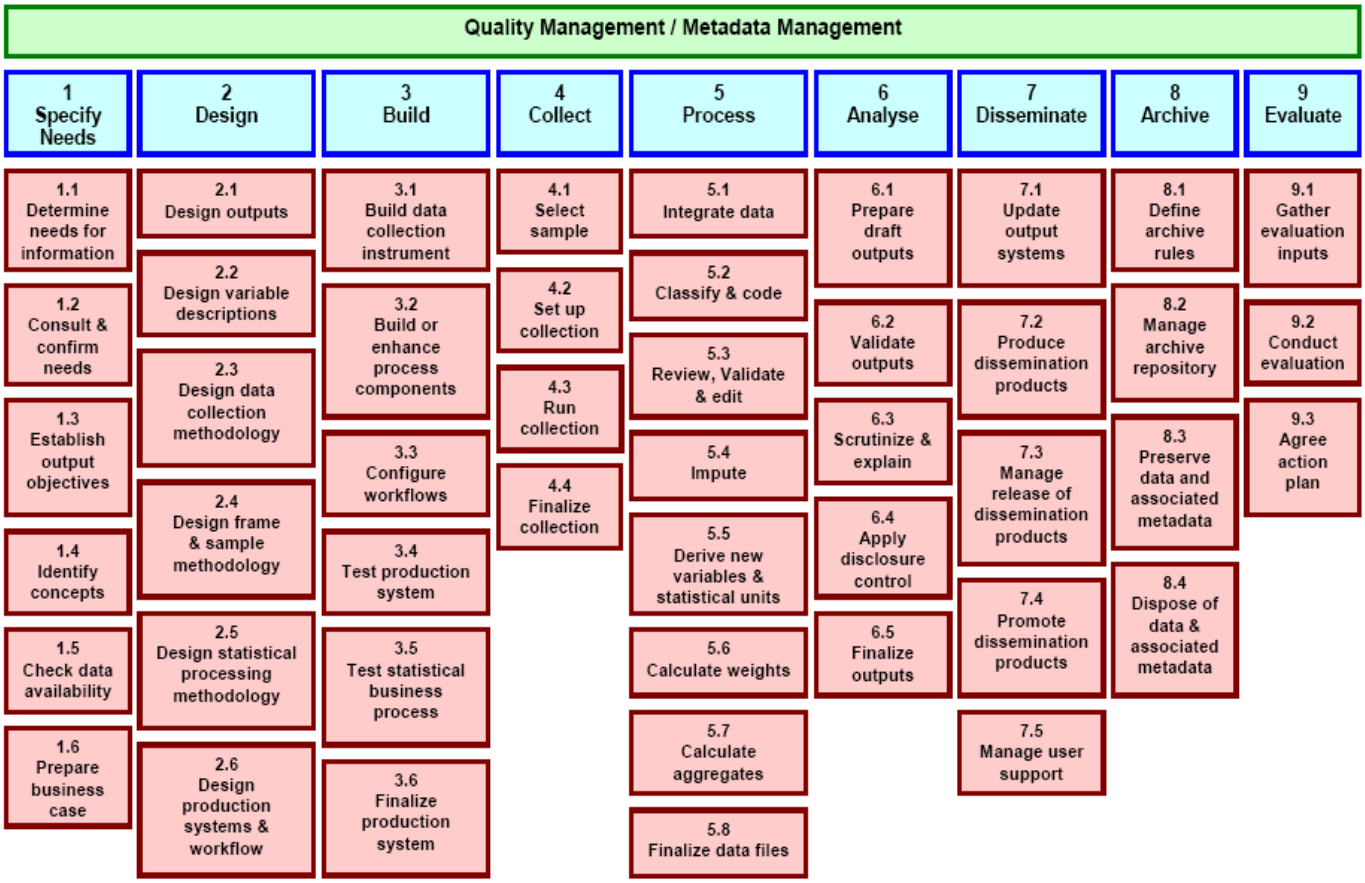
The scope of the CDE project for the 2011 Census cycle is marginally changed from the 2006 cycle. As such it was not considered necessary to undertake a new Privacy Impact Assessment, nor have the extensive consultation that preceded the 2006 CDE project. However, ABS did consult Privacy Commissioners in all jurisdictions, including the Federal Privacy Commissioner. A number of focus groups were held before the 2011 CDE project to assist the ABS to judge the community attitude to data linking, in particular the ABS conducting linkage projects and specifically linking data to the ABS Census. It is the Australian Statistician's position that the ABS should proceed with data linkage projects in line with community acceptance of conducting such linkage.

For full details of the 2011 CDE project, see the Information Paper – Census Data Enhancement Project: An Update, October 2010 (ABS, 2010a).

4. The (statistical business) process of data linking

The data linking process can be mapped to the Generic Statistical Business Process Model (GSBPM) as approved by the METIS Steering Group of the United Nations Economic Commission for Europe (UNECE, 2009). In this paper the focus will be on some of the relevant phases of the GSBPM and how they relate to the 2011 CDE project.

Figure 1 Generic Statistical Business Process Model



In the figure above, there are nine phases (Specify Needs through to Evaluate), and each phase has a number of sub-processes. The following sections map the GSBPM to a data linking project using the CDE experience as an example.

5. Phase 1: Specify Needs

1.1 Determine needs for information

There are a number of projects within the CDE umbrella, and each of these has been approved based on a recognised need for information. The understanding of needs is based on extensive consultation that the ABS undertakes with stakeholders. Projects will only be undertaken where there is a clear public benefit.

1.2 Consult and confirm needs

The ABS has an extensive ongoing consultation process with stakeholders. An important part of the consultation process is a range of user groups convened by the ABS to assist in determining data needs, a list of these groups is available in the ABS Annual Report (ABS, 2011).

1.3 Establish output objectives

The output objectives were defined in the Information Paper (ABS, 2010a), including details of retention policies and availability of access for people outside the ABS.

1.4 Identify concepts

Concepts are available based on existing metadata available for the source data. Work in this sub-process for CDE is largely around alignment of concepts from the different sources that are to be linked, and updating existing metadata where transformations are applied to data sources, for example occupation codes may be different depending on the classification used in coding on different data sources.

1.5 Check data availability

The basis for the CDE project is Census data. This data can become available (internally in the ABS) progressively as the Census is being processed. A critical component of the Census data necessary for many of the CDE linking projects is the availability of name and address. As discussed earlier name and address data is only available for a limited time.

Access to other data to be linked to the Census data needs to be negotiated with the custodians of the data. These custodians can be a single institution, or distributed across the States and Territories that make up the Australian Commonwealth (e.g. Registrars of Births, Deaths and Marriages).

The Australian Government is building a governance structure for integration of Commonwealth data in a safe and effective environment, see the National Statistical Service website for more details (Cross Portfolio Data Integration Oversight Board, 2011).

1.6 Prepare business case

Business cases (and other project management documentation) have been prepared for CDE projects. In 2006 focus groups and the PIA were part of an initial business case for ABS to get involved in data linking using Census data in the first place.

6. Phase 2: Design

2.1 Design outputs

A range of possible outputs have been proposed. From specific outputs from particular projects (e.g. adjustment factors for Indigenous life expectancy estimates) to general outputs (e.g. the possibility of unit record files particularly the 5% SLCD).

2.2 Design variable descriptions

Variable descriptions are generally available for the source datasets. Derived variables for use in linking will need descriptions when file standardisation (e.g. common

variables from the files to be linked are of the same type (i.e. character/numeric) etc.) and field standardisation (e.g. ensuring variables to be compared have compatible categories – this includes name standardisation) is undertaken.

2.3 Design data collection methodology

Ensure secure methods are in place to acquire data (e.g. ABS have a secure deposit box facility for external agencies to provide data over the internet). It is also necessary to have appropriately secure methods of moving data within the ABS (following the principle of functional separation based on Kelman (Kelman, Bass, & Holman, 2002)).

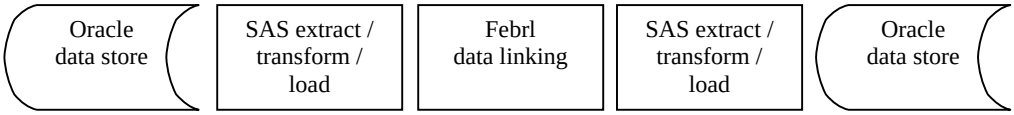
2.4 Design frame and sample methodology

Most CDE projects make use of entire input files. The SLCD will be based on a 5% sample as a privacy preserving mechanism.

2.5 Design statistical processing methodology

The CDE project will use probabilistic linking methodology following Fellegi-Sunter (Fellegi & Sunter, 1969). Blocking and linking strategies will be designed for each linkage project. Methods for calculation of m and u probabilities will depend on the data sources to be linked. The clerical review strategy is designed based on the method outlined in an ABS methodology paper see (Guiver, 2011). Quality gates have been designed to enable quality to be monitored throughout the linking process (for more information about quality gates see (ABS, 2010b)).

2.6 Design production systems and workflow



Oracle data store holds the input files (brought in through the data collection methodology). Standardisation (Design variable descriptions) is done in SAS and then files are created for input to Febrl (Freely Extensible Biomedical Record Linkage), open source data linking software – see (Christen, et al., 2005). Febrl does the data linking (including enabling clerical review). Output files (i.e. linking keys which allow the input files to be linked) from Febrl are loaded by SAS back into the Oracle data store.

The linked files can then be extracted from the Oracle data store and analysed or transformed into other output products (e.g. confidentialised unit record files).

7. Phase 3: Build

3.1 Build data collection instrument

Existing facilities within the ABS have been adapted for loading of sensitive linking datasets, with functional separation (see sub-process 2.3 above) implemented, as a privacy and security measure.

3.2 Build or enhance process components

Existing ABS infrastructure is available with minimal modifications. Standardisation (file and field) processes need to be built in SAS.

Febrl has been significantly enhanced by ABS (to enable multiprocessing, viewing snippets of Census forms for clerical review, clerical review functionality, categorical probability assignment).

Quality gates have been built around each data linking process (including the ability to extract management information at critical points in the process and checklists for running through data linking projects).

3.3 Configure workflows

Systems have been built to enable data movements, with management information extraction available at critical points. End to end training materials have been produced to ensure people working on data linking are able to perform efficiently and effectively.

3.4 Test production system

Test datasets have been created. A simulated Census dataset has been created based on 2006 Census data with the random addition of name and address for load testing purposes. Smaller datasets have also been created for system testing purposes.

Robust change management procedures need to be put in place to ensure the building of data linking infrastructure as an ongoing process can take place – this includes having test, development and production environments as well as governance in place to ensure build changes are acceptable to stakeholders.

3.5 Test statistical business processes

Census Dress Rehearsal (CDR) data is available to test statistical business processes using as close to live data as possible. Other datasets (for linking to the CDR) are also available in many instances (e.g. mortality data for the period after the CDR). CDR data will also be linked to Census data once that becomes available to provide a quality benchmark for CDE projects (especially the SLCD).

3.6 Finalise production systems

Governance processes are in place to enable sign-off of infrastructure into a production environment. Internal access arrangements have been formalised to allow appropriate access for those performing the linkage and those doing analysis of linked information.

Internal training materials have been produced covering the full end to end process.

8. Phase 4: Collect (Acquire in data linking terms)

4.1 Select sample

In the case of the SLCD this sub-process is where the 5% sample is selected. Other projects do not have a sampling basis.

4.2 Set up collection

Prepare for the arrival of data, ensuring appropriate accesses are in place in the computer systems. In the case of CDE this includes internal ABS Census data, as well as external administrative data.

4.3 Run collection

Take snapshots of data at points in time from various sources. This includes snapshots of Census data as it is still being processed (meaning that certain variables will not be populated depending on the stage of Census processing when the snapshot is taken).

Extract management information from input files at each snapshot.

4.4 Finalise collection

Create linkage files (merge files appropriately, file standardise, field standardise). As part of this step detailed data quality reports are produced for each input dataset.

9. Phase 5: Process

5.1 Integrate data

[recursive with respect to the chosen blocking strategy]

- Link files
- Threshold review
- Clerical review

Create and output linking keys so original input files can be linked in future.

Extract management information and ensure quality gates operate appropriately.

5.2 Classify and code

In data linking this is the final assignment of link status.

5.3 Review, validate and edit

Analyse unlinked records. In some cases (particular population groups for example) all possible links might be reviewed.

5.4 Impute

Imputation is not used as part of the CDE project. Imputed records on input files are generally disregarded.

5.5 Derive new variables & statistical units

These will usually be available from the input files used for integration, or can be merged from associated output files that are generally created from those files for other purposes.

An issue with data linking is where the same variable exists on input files and for a link has different values on each file, and a choice (or derivation) needs to be made to produce a final value. This is the case for the Indigenous Mortality project where Indigenous status is available on both the mortality records and the Census.

5.6 Calculate weights

Weights may need to be calculated for the SLCD. During the scoping exercise conducted as part of the 2006 Census cycle, some investigation was done which included the calculation of weights for the Census unit record file to Census dress rehearsal file, a project undertaken to assess the likely quality of census linking across cycles (Bishop, 2009).

5.7 Calculate aggregates
Produce an output report using the linked files.

5.8 Finalise data files
Ensure link keys are stored appropriately to allow linked files to be created.

10. Phase 6: Analyse

6.1 Prepare draft outputs
Create quality declaration documents for linkage files. Quality information will be available for the linkage files based on the management information extracted as part of the quality gate process, as well as compiled information about unlinked records.

Create output files (e.g. adjustment factor analysis files, CURFs, other analysis files). These would be created subject to the functional separation principle where only variables required for the analysis being undertaken would be included see (Kelman, Bass, & Holman, 2002).

6.2 Validate outputs
Check linked data against population estimates to ensure consistency. Linkage rates need to be calculated and assessed (including types of linkage error). Possible future work related to this could be a taxonomy of data linking quality with terminology that assists in understanding the quality of linked datasets – similar to survey quality terminology (e.g. relative standard errors, non-response etc.).

Compare quality measures with previous (2006) study. Some measures of false match and true non-match rates were calculated for CDE studies conducted in 2006 and these will be calculated for the 2011 project. Linkage rates for particular projects and sub-populations within those projects will be calculated and compared against 2006 CDE results (particularly Indigenous mortality linkage rates compared with non-Indigenous).

Confront with external data sources, at an aggregate level. This includes comparing linked results with any existing external information (such as estimated resident population, mortality statistics etc.).

6.3 Scrutinise and explain
Check how the linked file reflects initial expectations. This could include linkage rates for particular population groups.

View statistics from the linked file from different perspectives (in particular based on different geographies).

Undertake in depth analysis (e.g. in the case of the Indigenous Mortality study this involves calculating adjustment factors for life expectancy estimates).

6.4 Apply disclosure control
Confidentialise unit record files (this includes governance implications such as ensuring appropriate levels of clearance before external release). ABS has sophisticated methods

for confidentialising household survey unit record files and most (if not all) of these techniques would be applicable to linked files.

Prepare files for the remote execution environment for microdata (REEM – see (ABS, 2010c)). This REEM will provide analysis services which will access detailed de-identified microdata, with confidentiality routines built into the generated outputs to ensure that they are confidentialised in line with ABS legislative requirements and can be released as outputs.

6.5 Finalise outputs
This sub-process is the same as the GSBPM.

11. Phase 7: Disseminate

The sub-processes in this phase (listed below) are the same as for the GSBPM and in the case of the ABS generally apply to already existing corporate infrastructure.

7.1 Update output systems
7.2 Produce dissemination products
7.3 Manage release of dissemination products
7.4 Promote dissemination products
7.5 Manage user support

12. Phase 8: Archive

8.1 Define Archive rules
The ABS has detailed data management plans and policies and these will be applied to linked data files with a small number of exceptions. As mentioned above there are a small number of projects linking Census to other files using name and address. These files will be stripped of the name and address fields at the conclusion of Census processing, and the linkage keys (and any linked datasets) will be destroyed once the purpose of the linkage study has been met.

8.2 Manage archive repository
The ABS archive repository currently exists and is well developed. Data linking will introduce some minor changes due to the implementation of functional separation (in this case meaning that a 'librarian' role will be required to create linked files for 'analysts') and some minor adaptations to manage linkage keys.

8.3 Preserve data and associated metadata
The ABS already has infrastructure in place to manage this sub-process.

8.4 Dispose of data and associated metadata
The ABS already has infrastructure in place to manage this sub-process.

13. Phase: 9 Evaluate

9.1 Gather evaluation inputs

The quality gate implementation and extraction of management information will play a key role in gathering evaluation inputs.

9.2 Conduct evaluation
This sub-process is the same as the GSBPM.

9.3 Agree action plan
This sub-process is the same as the GSBPM.

14. Conclusion

ABS are building infrastructure to enable data integration projects to be completed successfully. An end to end approach is being taken, building infrastructure in areas where it does not exist, modifying existing infrastructure for specific linking purposes, and using existent infrastructure where it needs no modification.

One area where ABS is building infrastructure is in the Specify Needs phase of the GSBPM, where the ABS is collaborating with other Australian Government agencies to build a governance structure for integration of Commonwealth data in a safe and effective environment. Some aspects of this infrastructure are already in place (e.g. a set of principles to govern integration of Commonwealth data for statistical and research purposes, a Cross Portfolio Data Integration Oversight Board chaired by the Australian Statistician). Other aspects are in the process of being built including a process to accredit agencies to be an integrating authority to enable them to undertake "high risk" projects involving Commonwealth data.

An area where ABS has modified existing data linking infrastructure is the development of the Febrl linking software. Febrl version 0.3 has been significantly enhanced by the ABS to enable multiprocessing, viewing snippets of Census forms for clerical review, clerical review functionality, and categorical probability assignment. The clerical review modifications were offered to the original author of the software, but have not been included in the most recent version (Febrl 0.4).

There are many examples of existing ABS infrastructure meeting the needs of a data linking project. This is especially true in the dissemination phase of the GSBPM, where ABS corporate infrastructure has been used for many years to deliver output, and this infrastructure will work equally well for data linking.

The GSBPM has provided a useful structure for the CDE data linking projects, giving an end-to-end perspective of the process and ensuring that appropriate infrastructure is available for each step. It has also been very useful in planning collaboration across the ABS as many different areas are involved in a data linking project.

15. Future work

The ABS is planning on conducting research during our 2011 CDE projects with particular emphasis on designing standardisation procedures (sub-process 2.2 Design variable descriptions), designing methods for calculation of m and u probabilities (sub-process 2.5 Design statistical processing methodology), and exploring data quality (sub-process (sub-process 6.2 Validate outputs). This research is part of the continuous improvement that underlies activities undertaken at ABS.

Data linking is a key priority for the ABS, and with developments in Australia to build a safe and effective environment for the integration of Commonwealth data there will be increased data linkage being undertaken to leverage more information by combining individual datasets. Positioning data linking within the GSBPM will allow organisations to agree on standard terminology to aid discussions on developing statistical systems and processes.

References

ABS. (2006a, April). *Discussion Paper: Enhancing the Population Census: Developing a Longitudinal View.* Retrieved from ABS Website: http://www.abs.gov.au/AUSSTATS/abs@.nsf/Lookup/2060.0Main+Features12006

ABS. (2006b, June). *Census Data Enhancement Project: An Update.* Retrieved from ABS Website: http://www.abs.gov.au/AUSSTATS/abs@.nsf/allprimarymainfeatures/43185D34D6A1FF51CA2577BC0081EAC3?opendocument

ABS. (2010a, October). *Census Data Enhancement Project: An Update.* Retrieved from ABS website: http://www.abs.gov.au/ausstats/abs@.nsf/mf/2062.0

ABS. (2010b, December). *Quality Management of Statistical Processes Using Quality Gates, Dec 2010.* Retrieved from ABS Website: http://www.abs.gov.au/ausstats/abs@.nsf/mf/1540.0

ABS. (2010c, Sep). *1504.0 - Methodological News, Sep 2010.* Retrieved from ABS website: http://www.abs.gov.au/AUSSTATS/abs@.nsf/Lookup/1504.0Main+Features3Sep+2010

ABS. (2011, Oct). *1001.0 - Australian Bureau of Statistics -- Annual Report, 2010-11.* Retrieved from ABS website: http://www.abs.gov.au/ausstats/abs@.nsf/d36c95a5d2ce6cedca257098008362c8/01776d8d7f87e4e2ca25709900222520!OpenDocument

Bishop, G. (2009, Aug). *1351.0.55.026 - Research Paper: Assessing the Likely Quality of the Statistical Longitudinal Census Dataset, August 2009.* Retrieved from ABS website: http://www.abs.gov.au/AUSSTATS/abs@.nsf/mf/1351.0.55.026

Christen, P., Churches, T., Hegland, M., Taylor, L., Lim, K., Willmore, A., et al. (2005, April). *Parallel Large Scale Techniques for High-Performance Record Linkage.* Retrieved from ANU Data Mining Group: http://datamining.anu.edu.au/linkage.html

ComLaw. (1975). *Australian Bureau of Statistics Act.* Retrieved from Australian Government ComLaw: http://www.comlaw.gov.au/ComLaw/Legislation/ActCompilation1.nsf/0/D457D9DA71AE7F49CA25744B001DC54C/$file/AustBurStatAct1975WD02.pdf

Cross Portfolio Data Integration Oversight Board. (2011). *Statistical Data Integration involving Commonwealth Data.* Retrieved from National Statistical Service: http://www.nss.gov.au/nss/home.nsf/pages/Data+Integration+Landing+Page?OpenDocument

Fellegi, I. P., & Sunter, A. B. (1969). A Theory for Record Linkage. *Journal of the American Statistical Association*, 1183-1210.

Guiver, T. (2011, May). *Research Paper: Sampling-Based Clerical Review Methods in Probabilistic Linking.* Retrieved from ABS Webstie: http://www.abs.gov.au/AUSSTATS/abs@.nsf/mf/1351.0.55.034

Kelman, C. W., Bass, A. J., & Holman, C. D. (2002). Research use of linked health data - a best practice protocol. *Australian and New Zealand Journal of Public Health*, 251-255.

UNECE. (2009). *Generic Statistical Business Process Model.* Retrieved from UNECE website: http://www1.unece.org/stat/platform/display/metis/The+Generic+Statistical+Business+Process+Model

Waters, N. (2005, June). *Privacy Impact Assessment.* Retrieved from ABS Website: http://www.abs.gov.au/Websitedbs/D3110124.NSF/f5c7b8fb229cf017ca256973001fecec/fa7fd3e58e5cb46bca2571ee00190475!OpenDocument