Preserving the Australian Census – 250 years of population data for Australia

Dr. Steven McEachern, Janet McDougall, Montserrat Alvarez-Klee and Xiaolan Cai

Australian Data Archive

October 2019

ADA AUSTRALIAN DATA ARCHIVE

An overview

- Background to the problem:
  - isn’t it all on the ABS website?
- Methodology:
  - how do you find this stuff anyway?
- Data audit:
  - so where is all this stuff?
- Data processing
  - Ok so I’ve got the stuff – now what?
- Next steps
  - Creating a true national collection

Background

- Australia has a long and detailed history of the collection of population data.
- Australia's Commonwealth census data collection is now over 105 years old, with the completion of the 2016 Census.
- Censuses and musters have also been undertaken in the Australian colonies since 1833.
- A significant proportion of population data is however located in archives and libraries, or digitised but embedded in formats largely inaccessible to researchers or to the Australian public.
- While data from 1996 onwards is available from the Australian Bureau of Statistics, data prior to this date is difficult to locate and access.

The Census and the Australian Data Archive

- ADA has worked to make available major portions of the census, which have been deposited by the Australian Bureau of Statistics.
- While this work has been beneficial, the current format and nature of this content has limitations for researchers.
- Data from 1911-1961 is embedded in PDFs and print materials.
- For the census data from 1966 to 1981, there are no corresponding maps or geocode files of the geographic coverage of these censuses, limiting the usefulness of the data for informing spatially-enabled research or policy
- ADA and AURIN have been working to enable access to data from 1981-1991
  - Data has needed significant processing, including data processing, file format changes and creation of documentation.

This project – ARDC Discovery Projects

- Support provided by ARDC for a short-term project (4 months) to complete a review of the current state of Census data over time
- This included:
  - The Data Audit: a review of the current status of census data in Australia
  - Data processing: methods and a processing pipeline for extracting the data from it's current form (whatever that may be) into a machine-actionable format
  - Next steps: a high-level collections development program for the census data to bring all the data and associated materials into a form that is available in machine-actionable formats for access by all Australians

Census periods

- We identified four main periods in the history of the Australian census.
    - Colonial period (1788-1900)
    - Federation and beyond (1901-1961)
    - Mainframe processing (1966 – 1991)
    - The digital era (1991 to 2016)
- Related to important changes in Australian politics or technology that affected the way census data was collected or presented

Methodology
(Montserrat Alvarez-Klee)

- For each census period, we sought to trace the location of the main compilations of censuses' publications for each period
- This included:
  1. An audit of the census publications produced in each period.
  2. Sourcing of the location of the underlying data from the publications audit: From conversations held with the ABS we learnt that many of the various censuses hard copies and microfiche were donated to the National Library of Australia, state libraries, archives and universities.
  3. Searches of online resources and catalogues from these institutions to locate the relevant items from the compilations and document their various formats.
- Additional step to investigate the possibility of data conversion for some of the censuses
  - Following standard Linked Data principles (Meroño-Peñuela, Ashkpour et al. 2012).

Key output – data audit appendices

Catalogue item description	Format	Source
Group 2: Census of Population and Housing		
Census of the Commonwealth of Australia, 1911 Volumes I to III (Statisticians Report, Detailed Tables and The Mathematical Theory of Population, Appendix A)	Printed material: Volumes (I, II, III)	National Library of Australia
	Digitised volumes (pdf)	ABS website
	Microfiche (Batch 01.20 No. of fiche 193)	University of Sydney
		UoS Catalogue no. 1122.0 ISSN: 1038-6424. Mi
Census of the Commonwealth of Australia, 1921 Volumes I and II (Detailed Tables and Statisticians Report)	Printed material: Volumes (I,II)	National Library of Australia
	Printed material	ANU
	Digitised volumes (pdf)	ABS website
	Microfiche (Batch 01.20 No. of fiche 193)	University of Sydney
Census of the Commonwealth of Australia, 1921 Census Bulletins 1 to 26	Printed material: No. 1-26	National Library of Australia
	Digitised bulletins (Pdf)	ABS website
	Microfiche (Batch 01.20 No. of fiche 193)	University of Sydney
Census of the Commonwealth of Australia, June 1933 Volumes I to III (Detailed Tables and Statisticians Report)	Printed material: Book Volumes (I, II, III)	National Library of Australia
	Digitised volumes (pdf)	ABS website
	Microfiche (Batch 01.20 No. of fiche 193)	University of Sydney
Census of the Commonwealth of Australia, June 1933 Census Bulletins 1 to 25	Printed material: Book No. 1-25	National Library of Australia
	Digitised bulletins (Pdf)	ABS website
	Microfiche (Batch 01.20 No. of fiche 193)	University of Sydney

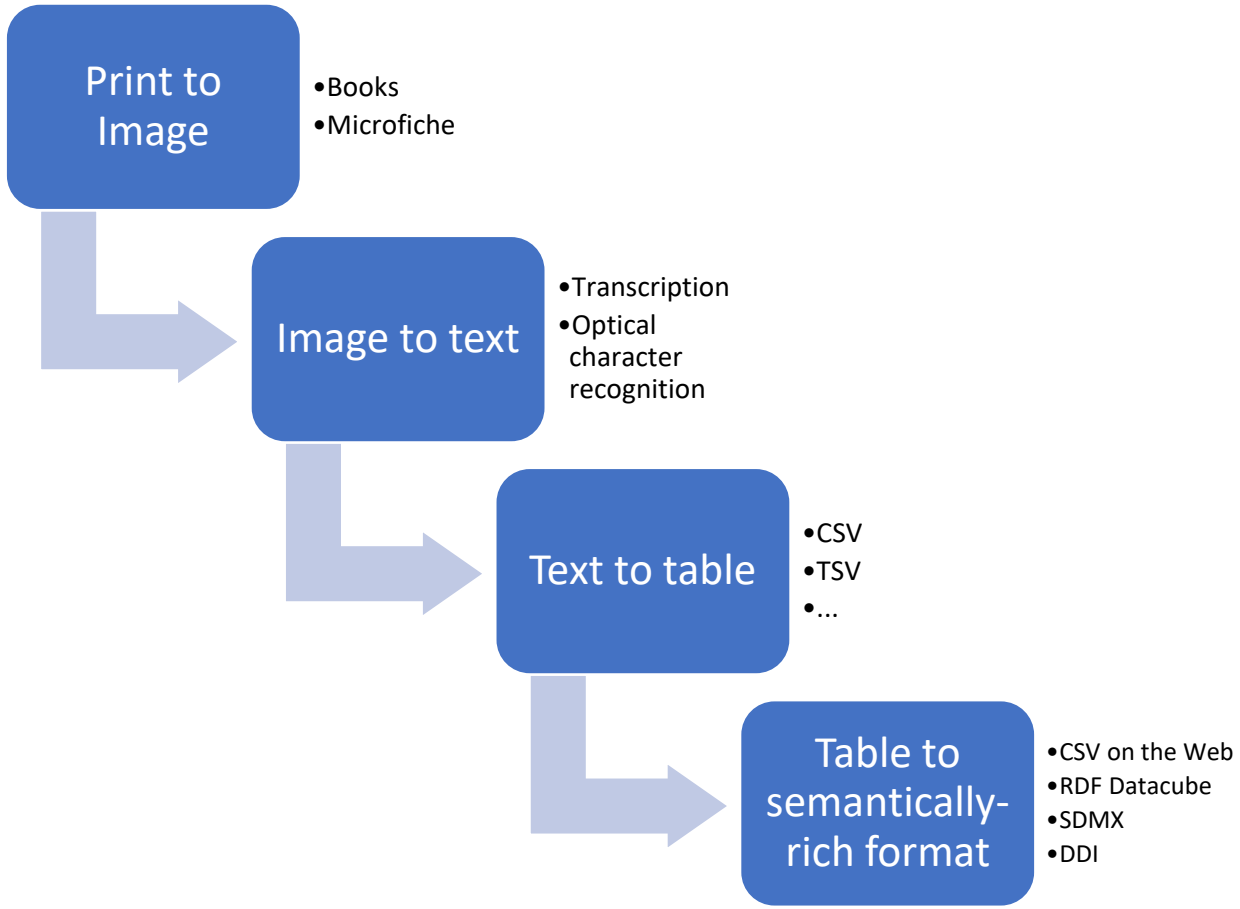
Maps

- Spatial characteristics are an important part of the Census
- In the 1933 Census the compilation of maps employed 60 survey draftsmen and took around nine months (ABS 2005).
- Social Atlases were produced for the first time in the 1981 Census (ABS 2005).
- In 1986, CD-ROMS were made available for the first time (CDATA)
  - Included mapping data and software that allowed clients map the data themselves (ABS 2005).
- For censuses previous to 1986 maps are available in the following formats:
  - Printed material as part of the censuses volumes and publications held at the National Library of Australia
  - Microfiche series 1901-1984 held at the University of Sydney Library
  - As PDFs in the scanned volumes on the ABS website. (Report Appendix E)

Audit results – in summary

- (Aggregate) Data is still available – in some form
- A number of potential sources available for extracting census data, and then converting it to a useable format.
- These sources are however at varying stages of readiness for processing
- Require a series of data processing steps to make them accessible.

Data processing workflow



Data processing for transforming the census (and any tabular data)

(Xiaolan Cai)

1. Print/microfiche to image:
   - the capture of images of the census publications in digital image formats.
   - The optimal format would be *.TIF, but could include alternatives such as JPG or PDF
   - (Much of the collection from 1911-1961 was already captured in PDF format by the ABS)
2. Image to text:
   - the extraction of the tabular data from the image into a format suitable for processing.
   - This may be directly into a tabular format such as CSV, but might also be unstructured text.
   - Extraction methods include optical character recognition and manual transcription using crowdsourcing and similar methods

Data processing for transforming the census (continued)

3. Text to Table:
   - Conversion of the extracted text into a tabular data format representative of the basic tabular structure of the original table.
   - The simplest of these is CSV, but could also include TSV, Excel, R data frames or JSON-LD
4. Table to semantically-rich data format:
   - Connection of the semantics of the table content to the base data format.
   - This includes temporal, spatial and conceptual characteristics, as well as associated contextual information such as annotations, documentation and methodological information.
   - Example formats include CSV on the Web, RDF Datacube, SDMX and DDI

Examples

- Print to image
  - ADA working with ANU Printery to convert the Australian Demographic Databank (ADDB) volumes
    - Supported by RSSS small grant with Heather Booth (Demography)
  - Current interest in conversion of the ABS Microfiche (lost in the flood of the ANU Library)
- Image to text
  - ADDB testing of extraction of content from images
  - Depends on image quality
  - Can leverage open source tools (e.g. Tesseract, developed by Google) if content is sufficient quality

Examples

- Text to table
  - Xiaolan worked to convert XHTML text to
  - Challenge is table headers – particularly merged cells
  - We will release sample code via ADA/Github
  - Aim is to convert the full collection (currently in XHTML: https://dataverse.ada.edu.au/dataset.xhtml?persistentId=doi:10.26193/MP6WRS)
- Table to semantics
  - Project esting procedures did include some preliminary processing of content in anticipation of future needs, as an output of the generation of the HCCDA CSV data files.
  - Semantic analysis via R was applied to re-organize headers of tables in csv files into usable formats

Example comparisons over time

Table 1 Summary on changes of topics in HCCDA NSW from 1833 to 1901

Year	Topics							
1833	Sex and age	Convict	Religion					
1836	Sex and age	Convict	Religion					
1841	Sex and age	Convict	Religion	Occupation	Social condition	House		
1846	Sex and age	Civil condition	Religion	Occupation	Social condition	House	Country where born	Education
1851	Sex and age	Civil condition	Religion	Occupation	Social condition	House	Country where born	Education
1856	Age and sex		Religion	Occupation	Social and domestic condition and houses		Nationality	Education and age Education, religions and age

What's now needed

1. A **collection digitisation** program to move all content into fully machine actionable format
2. A **data harmonisation** program to enable comparison of data across conceptual, temporal and geographic dimensions
3. A complementary **map digitisation program** to find, digitise and spatially enable the geographic information associated with the census data
4. A **documentation preservation** program to capture, preserve and disseminate the documentation associated with the census in each year of it's collection
5. **Extension** of the census data collection **into related areas**, including vital statistics and economic statistics

Thank you and questions

Contact us:

https://ada.edu.au

ada@anu.edu.au

Additional slides – audit results by period

Colonial period (1933-1900)

- Key source: ABS Catalogue of Australian Statistical Publications 1804-1901 (ABS 1989)
  - https://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/1115.01804%20to%201901?OpenDocument
  - Thanks to Boyd Hunter for the pointer here
  - This includes a list of the publications and statistical reports compiled by the former colonial statistical bureaus and their precursors
- Most of the statistical publications from this period are held in two collections:
  1. ADA has processed and disseminated the census reports as images and HTML tables on their website (http://hccda.ada.edu.au/)
     - This collection is progressively being incorporated into Dataverse
     - https://dataverse.ada.edu.au/dataset.xhtml?persistentId=doi:10.26193/MP6WRS
  2. The National Library of Australia holds the microfiche series for the ABS Catalogue of Australian Statistical Publications 1804-1901
     - Even there the collection is patchy – NLA have NSW microfiche, but other colonies still to be confirmed

Early 20th century (1911-1961)

- First Federal census tabulation was 1911: tabulation was undertaken almost entirely by hand and the results took a long time to release, further delayed by World War I.
- For the 1921 census, **automatic machine tabulation equipment** was hired from England
- Various delays throughout this period (WWI, Depression, WWII)
- Standardised in 1961 to every 5 years

Early 20th century (1911-1961) ... continued

- Key publication: ABS Catalogue of Historical Microfiche Series – Statistical Publications 1901-1984 (ABS 1990).
  - Catalogue 1121.0, replaced by 1123.0: **neither** are now available on the ABS website in PDF form
  - Focussed on sections associated with population measures – Group 2 Census of Population and Housing; and Group 3 Estimates of Population, Population Projections, Vital Statistics, Migration (ABS 1990).
  - Content however is in print format
  - Some has been migrated to PDFs (produced for 2011 anniversary of first Federal census)
- Most of the census data from this period can be found in the following formats:
  - Microfiche series held by the University of Sydney Library.
  - PDF format by the ABS website
  - Hard copies of many volumes can be found at the National Library of Australia.

Mainframe processing (1966-1991)

- Introduction of computer processing in 1966 shifted the availability of data
- ABS begins producing “master files”
  - Machine tapes with cumulative files of all units (usually CCDs and LGAs)
  - Print volumes are then generated from machine tapes
- Same staffing, but enabled quality control checks to be built into the processing system
- By 1991, processing included optical mark recognition and automated coding of responses

Mainframe processing (1966-1991)

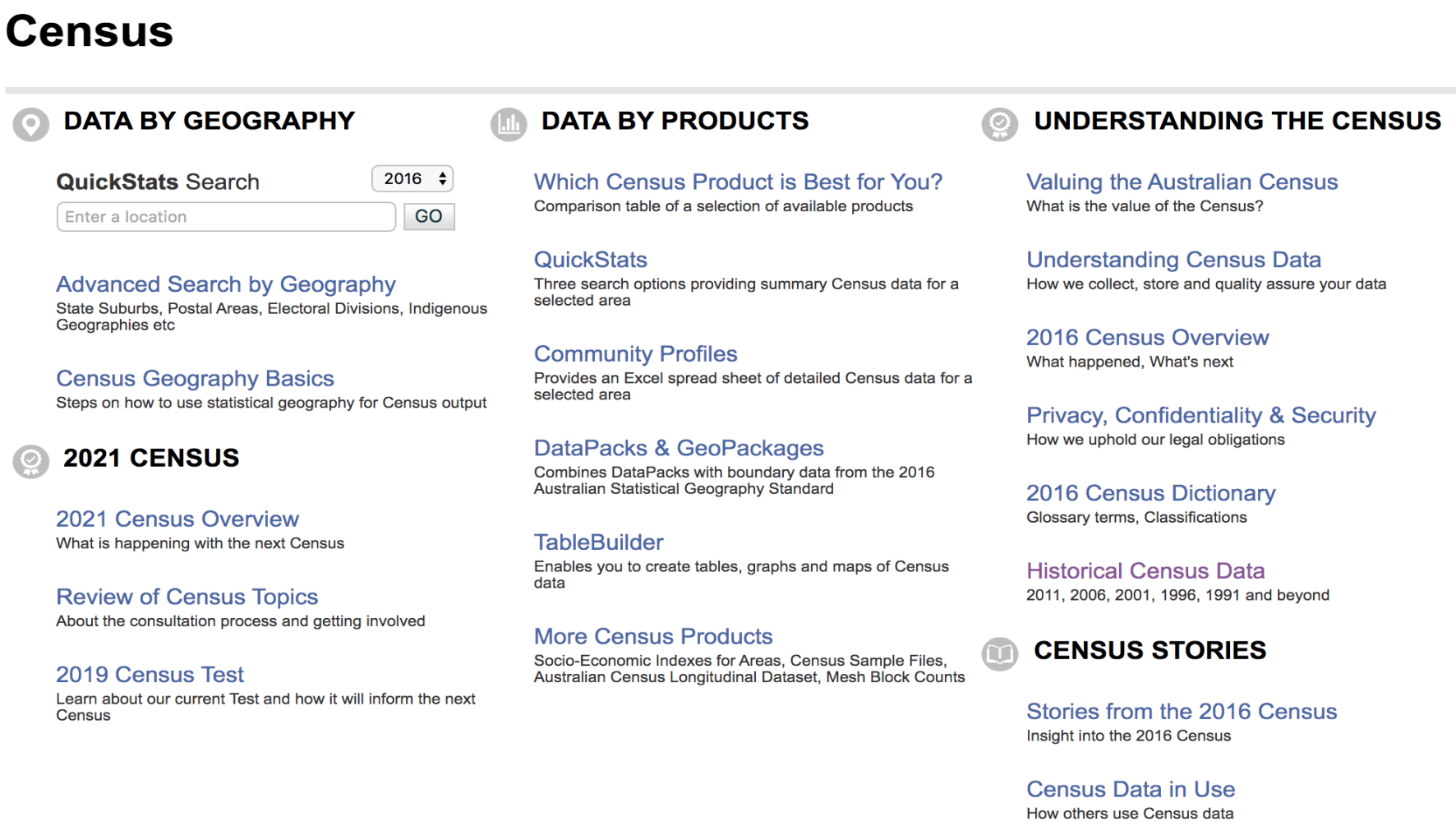
Table	Mnemonic	1961	1966	1971	1976	1981	1986	1991
Collection District Master File	CDMF		Online	Online	Online	Online	Online	
Collection District Summary File	CDSF		14 Files	9 Files	12 Files	7 Files		1 File
Collection District Summary File - Preliminary	CDPF				2 Files			
Local Government Area Summary File	LGASF		16 Files	8 Files	1 File	2 Files		
Local Government Area Summary File - De Facto	LGADEF1				2 Files			
Local Government Area Summary File, Occupations, Industry, Qualifications - De Facto	LGADEF2				1 File			
Local Government Area / Geographic Descriptor File	LGAD / GDF				1 File	1 File		
Journey to Work Tables	JTW	Online	4 Files	7 Files				
Aboriginal Collection District Summary File	ABCDSF				3 Files			
Australian/Federal Electoral Division Summary File	AEDSF / FEDSF				1 File	1 File		
Urban Centres Summary File	UCSF					1 File		
Postcode Summary File	PCSF						1 File	
Statistical Local Area Summary File	SLASF						1 File	
ASCO/CCLO Link File	ASCO						Online	
Socio-economic Indicator File	SES						1 File	
Detailed / Special Tables	DT		Online	5 Files				Online

Detailed descriptions of the 1966 - 1986 census data files are located at: http://www.assda.edu.au/census.html

The digital era: 1996 - present

- Fully computerised systems now in place
- Establishment of geographic information systems to generate the printed census maps for census collectors
  - Based on ASGC: Australian Standard Geographic Classification
  - Changed the dynamics of how spatial information was collected and used
  - Eventually leading to ASGS in 2006
- Also see the emergence of:
  - Data packs (~1991-1996) collecting Basic Community Profiles rather than “master files”
    - Covering each level of the ASGC
  - CDs rather than machine tapes

Census data at the ABS (https://www.abs.gov.au/Census)



Census

DATA BY GEOGRAPHY

QuickStats Search 2016

Enter a location GO

Advanced Search by Geography
State Suburbs, Postal Areas, Electoral Divisions, Indigenous Geographies etc

Census Geography Basics
Steps on how to use statistical geography for Census output

2021 CENSUS

2021 Census Overview
What is happening with the next Census

Review of Census Topics
About the consultation process and getting involved

2019 Census Test
Learn about our current Test and how it will inform the next Census

DATA BY PRODUCTS

Which Census Product is Best for You?
Comparison table of a selection of available products

QuickStats
Three search options providing summary Census data for a selected area

Community Profiles
Provides an Excel spread sheet of detailed Census data for a selected area

DataPacks & GeoPackages
Combines DataPacks with boundary data from the 2016 Australian Statistical Geography Standard

TableBuilder
Enables you to create tables, graphs and maps of Census data

More Census Products
Socio-Economic Indexes for Areas, Census Sample Files, Australian Census Longitudinal Dataset, Mesh Block Counts

UNDERSTANDING THE CENSUS

Valuing the Australian Census
What is the value of the Census?

Understanding Census Data
How we collect, store and quality assure your data

2016 Census Overview
What happened, What's next

Privacy, Confidentiality & Security
How we uphold our legal obligations

2016 Census Dictionary
Glossary terms, Classifications

Historical Census Data
2011, 2006, 2001, 1996, 1991 and beyond

CENSUS STORIES

Stories from the 2016 Census
Insight into the 2016 Census

Census Data in Use
How others use Census data