Project Report – end Nov 2023

Integrated Research Infrastructure for Social Science (IRISS) Project Deliverables

Authors: Terhi Nurmikko-Fuller, Janet McDougall, Paul Pickering

PoC HCCDA Linked Data-ready Vocabularies

Summary

We report on a research project into the HCCDA (Historical Census and Colonial Data Archive). The research project has consisted of close reading of the data, consultation with domain experts to assess and clarify the historical context in which the data was originally captured, and the development of an ontological structure (represented in .TTL or Turtle) to represent information in a knowledge graph (in adherence to the RDF or Resource Description Framework abstract data model).

We additionally report on a deep dive into data from New South Wales (NSW) focusing on a specific data category -occupations - as expressed in all relevant Censuses (1833 - 1901 inclusive). We have also examined this category longitudinally looking at censuses collected across all of the colonies. For consistency, the longitudinal study has included only those years when the census was collected in all colonies in the same year (1861, 1881, 1891, and 1901).

Introduction

Census data is the foundation for much data-oriented research in Australia, and is a significant national asset. The accessibility of a significant proportion of this data, while digitised, is embedded in archives and libraries, or digitised but available only in formats, which can be largely inaccessible to specialist researchers as well as to the Australian public (McEachern, 2021).

The HCCDA, holding the colonial censuses collected between 1833 and 1901 is the focus of this project. Australia's first census was conducted in 1828 in New South Wales (but is excluded from the HCCDA dataset), and regularly from then on, including the other newly created colonies as they were established. Colonial statisticians identified the need for compatibility between the colonies in their respective censuses, and from 1881 a census was held simultaneously in each of the colonies (ABS 2006).

In its current form the HCCDA is incompatible with programmatic-based research, despite the work already done to digitise the censuses. The aim of this project is to consider the Linked Open Data (LOD) methodology and to investigate solutions that may enable us to translate or convert digital content into machine actionable formats, and to thus enable future data integration and harmonisation.

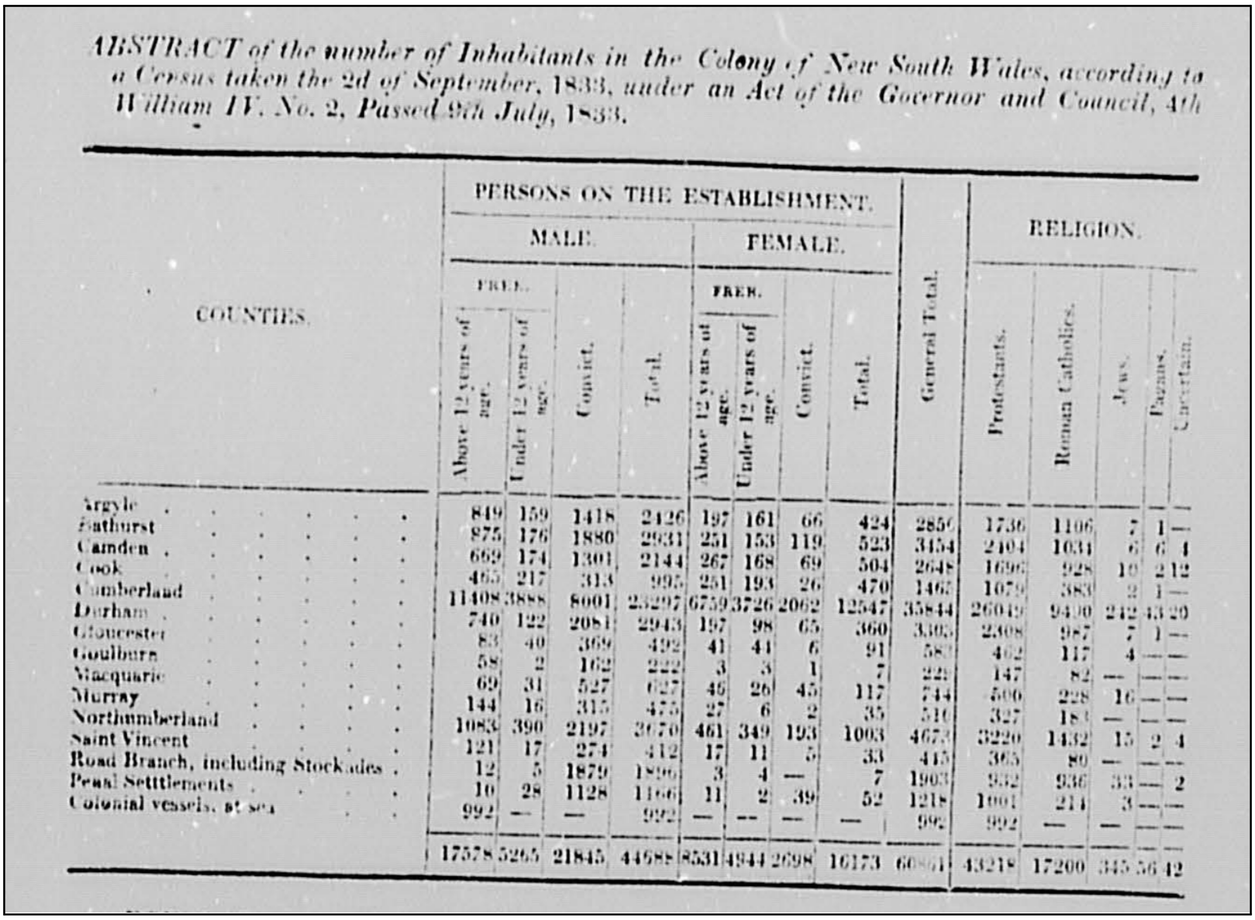
Description of HCCDA data and its origins

As part of the Australian Bicentenary Project the Australian Bureau of Statistics (ABS) Library collected all Colonial, State, Territory and Commonwealth statistical publications and reports from 1788 to 1988, and recorded them on microfiche. The HCCDA was created under an ARC LIEF grant 2007 - 2011, in collaboration with the ABS archive, the Australian Social Science Data Archive (now ADA), and the ANU Super Computer Facility (now National Computational

Infrastructure). Microfiche versions of the census publications were sourced from the ABS and converted to both digital images and searchable XML markup. The aims of the project were to convert the fiche frames to TIFF files, and manually convert the text and tables to XML (DocBook schema). This digitisation conversion to XML would enable browsing and searching in a web application.

The transcribed and digitised HCCDA datasets were published by ADA in 2019. These materials cover the census publications and reports covering the period from 1833 to 1901. The HCCDA corpus includes 18,638 pages of text, and approximately 15,000 tables, all with full digital images, text conversion (XML) and individually identified pages and tables (ABS et al., 2019). The HCCDA material can be downloaded, searched and copied manually, but in its current format is not machine-actionable.

Figures 1a and 1b illustrate the same structured data[1] as the microfiche and later .xtable. We have selected the 1833 census from NSW as the example as its relatively simple structure enables the clear visualisation of these pages.

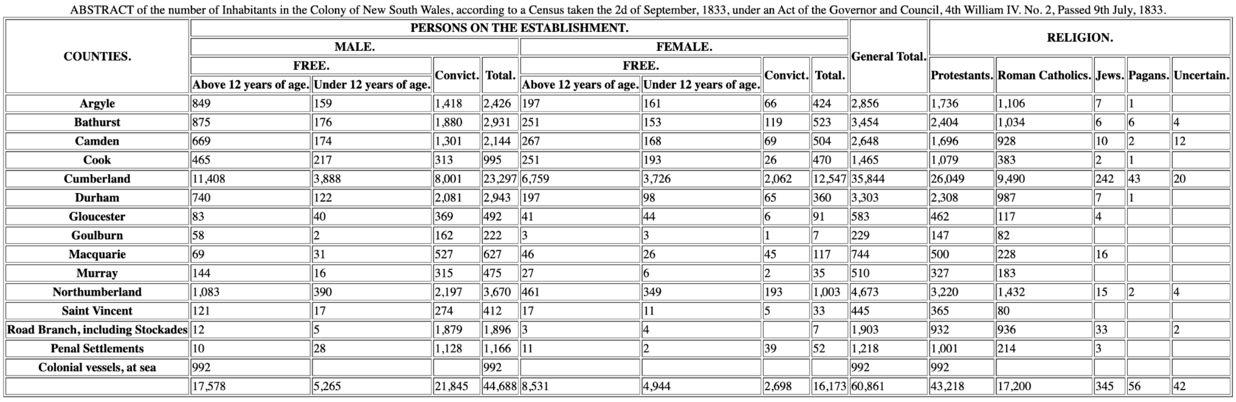


ABSTRACT of the number of Inhabitants in the Colony of New South Wales, according to a Census taken the 2d of September, 1833, under an Act of the Governor and Council, 4th William IV. No. 2, Passed 9th July, 1833.

COUNTIES.	PERSONS ON THE ESTABLISHMENT.								General Total.	RELIGION.				
	MALE.				FEMALE.									
	FREE.				FREE.									
	Above 12 years of age.	Under 12 years of age.	Convict.	Total.	Above 12 years of age.	Under 12 years of age.	Convict.	Total.		Protestants.	Roman Catholics.	Jews.	Pagans.	Uncertain.
Argyle	849	159	1418	2426	197	161	66	424	2850	1736	1106	7	1	—
Bathurst	875	176	1880	2931	251	153	119	523	3454	2404	1031	6	6	4
Camden	669	174	1301	2144	267	168	69	504	2648	1696	928	10	2	12
Cook	465	217	313	995	251	193	26	470	1465	1079	383	2	1	—
Cumberland	11408	3888	8001	23297	6759	3726	2062	12547	35844	26049	9490	242	43	20
Durham	740	122	2081	2943	197	98	65	360	3303	2308	987	7	1	—
Gloucester	83	40	369	492	41	41	6	91	583	462	117	4	—	—
Goulburn	58	2	162	222	3	3	1	7	229	147	82	—	—	—
Macquarie	69	31	527	627	46	26	45	117	744	506	228	16	—	—
Murray	144	16	315	475	27	6	2	35	510	327	183	—	—	—
Northumberland	1083	390	2197	3670	461	349	193	1003	4673	3220	1432	15	2	4
Saint Vincent	121	17	274	412	17	11	5	33	445	365	80	—	—	—
Road Branch, including Stockades	12	5	1879	1896	3	4	—	7	1903	932	936	33	—	2
Penal Setttlements	10	28	1128	1166	11	2	39	52	1218	1001	214	3	—	—
Colonial vessels, at sea	992	—	—	992	—	—	—	—	992	992	—	—	—	—
	17578	5265	21845	44688	8531	4944	2698	16173	60861	43218	17200	345	56	42

Fig. 1a Example of the 1833 Census Data illustrating locations (Microfiche)

[1] Specifically in the file with the suffix /NSW-1833-census-01_5-1.



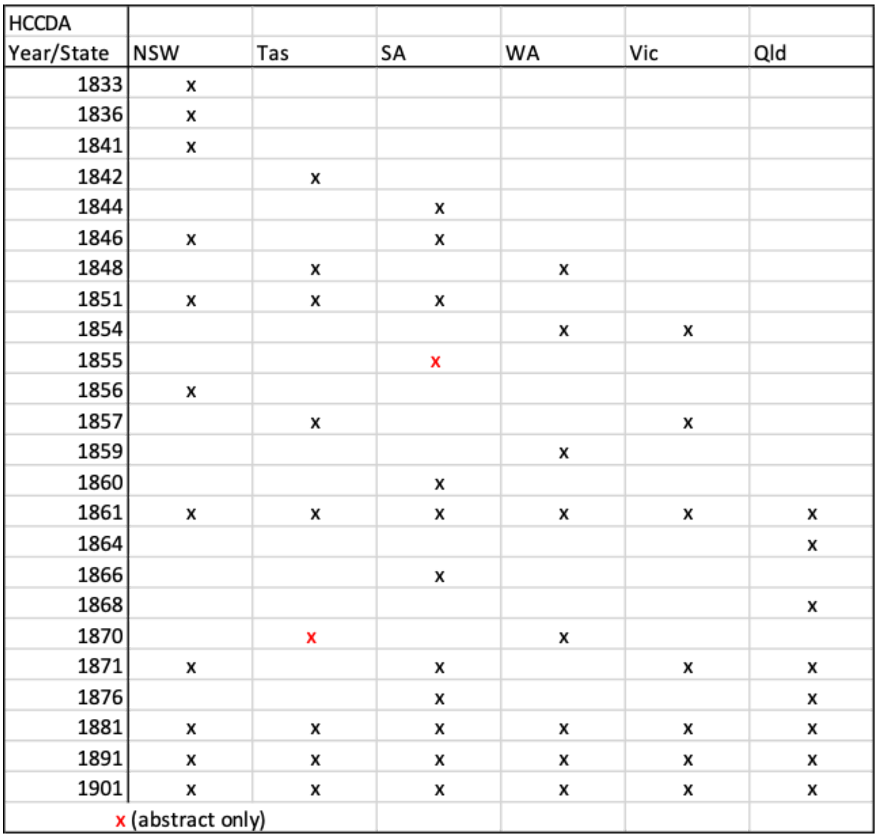
ABSTRACT of the number of Inhabitants in the Colony of New South Wales, according to a Census taken the 2d of September, 1833, under an Act of the Governor and Council, 4th William IV. No. 2, Passed 9th July, 1833.

COUNTIES.	PERSONS ON THE ESTABLISHMENT.								General Total.	RELIGION.				
	MALE.				FEMALE.									
	FREE.		Convict.	Total.	FREE.		Convict.	Total.		Protestants.	Roman Catholics.	Jews.	Pagans.	Uncertain.
	Above 12 years of age.	Under 12 years of age.			Above 12 years of age.	Under 12 years of age.								
Argyle	849	159	1,418	2,426	197	161	66	424	2,856	1,736	1,106	7	1	
Bathurst	875	176	1,880	2,931	251	153	119	523	3,454	2,404	1,034	6	6	4
Camden	669	174	1,301	2,144	267	168	69	504	2,648	1,696	928	10	2	12
Cook	465	217	313	995	251	193	26	470	1,465	1,079	383	2	1	
Cumberland	11,408	3,888	8,001	23,297	6,759	3,726	2,062	12,547	35,844	26,049	9,490	242	43	20
Durham	740	122	2,081	2,943	197	98	65	360	3,303	2,308	987	7	1	
Gloucester	83	40	369	492	41	44	6	91	583	462	117	4		
Goulburn	58	2	162	222	3	3	1	7	229	147	82			
Macquarie	69	31	527	627	46	26	45	117	744	500	228	16		
Murray	144	16	315	475	27	6	2	35	510	327	183			
Northumberland	1,083	390	2,197	3,670	461	349	193	1,003	4,673	3,220	1,432	15	2	4
Saint Vincent	121	17	274	412	17	11	5	33	445	365	80			
Road Branch, including Stockades	12	5	1,879	1,896	3	4		7	1,903	932	936	33		2
Penal Settlements	10	28	1,128	1,166	11	2	39	52	1,218	1,001	214	3		
Colonial vessels, at sea	992			992					992	992				
	17,578	5,265	21,845	44,688	8,531	4,944	2,698	16,173	60,861	43,218	17,200	345	56	42

Fig.1b Example of the 1833 Census Data illustrating locations (.xtable)

The HCCDA dataset contains data from all of the colonies covering the second half of the nineteenth century. However, the censuses were not collected in all of the six colonies at the same time each time. Table 1 (below) illustrates the discrepancy of collection across different locations. It captures data extracted from "Preserving the Australian Census - 250 years of population data for Australia, 2019".[2]

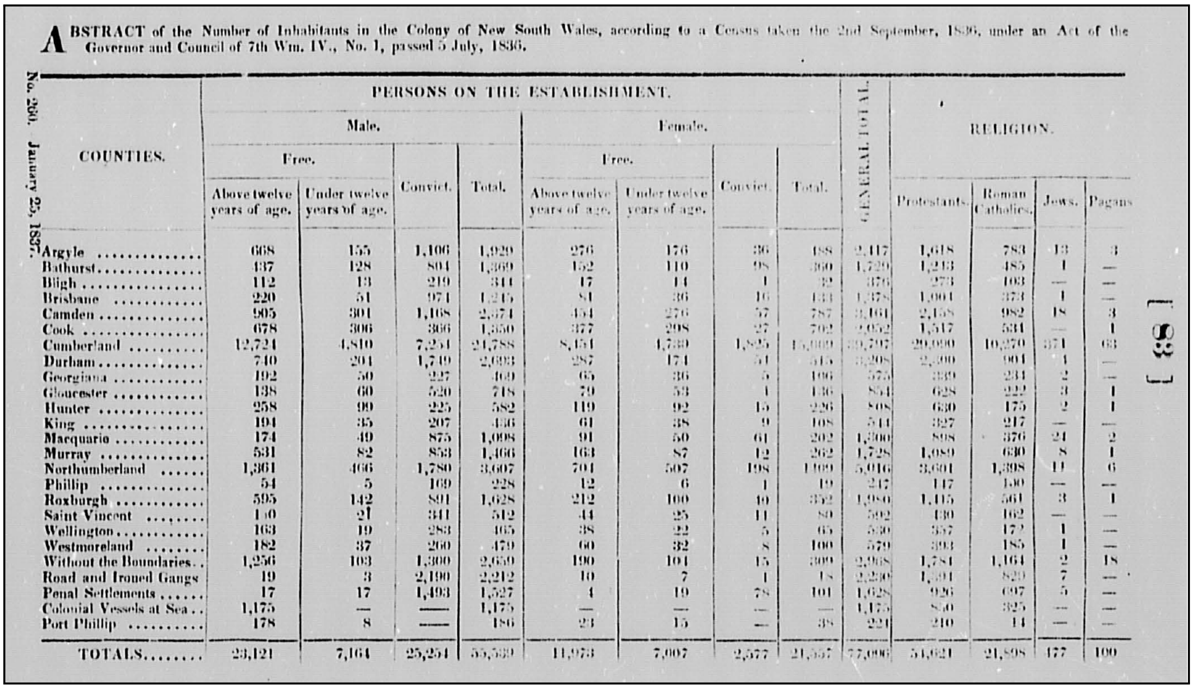
Table 1. HCCDA Census by state and year table



HCCDA Year/State	NSW	Tas	SA	WA	Vic	Qld
1833	x					
1836	x					
1841	x					
1842		x				
1844			x			
1846	x		x			
1848		x		x		
1851	x	x	x			
1854				x	x	
1855			x			
1856	x					
1857		x			x	
1859				x		
1860			x			
1861	x	x	x	x	x	x
1864						x
1866			x			
1868						x
1870		x		x		
1871	x		x		x	x
1876			x			x
1881	x	x	x	x	x	x
1891	x	x	x	x	x	x
1901	x	x	x	x	x	x
x (abstract only)						

[2] Available at https://anu365.sharepoint.com/:w:/r/sites/CSRM/_layouts/15/Doc.aspx?sourcedoc=%7BE5A0FB45-C936-4700-81A7-218022AAE95C%7D&file=Final%20draft%20Preserving%20the%20Australian%20Census_SMEdits_20200225.docx&action=default&mobileredirect=true.

It is worth noting that the census data becomes exponentially more complex and capacious over time. The solutions for the analysis of data from 1833, for example, do not readily lend themselves to analysis of data from 1901, and *vice versa*. Figures 2a and 2b illustrate the ways in which the census data becomes more numerous (i.e. more locations are listed within the one category) and more complex (there are a greater number of categories) from 1836 to 1841.



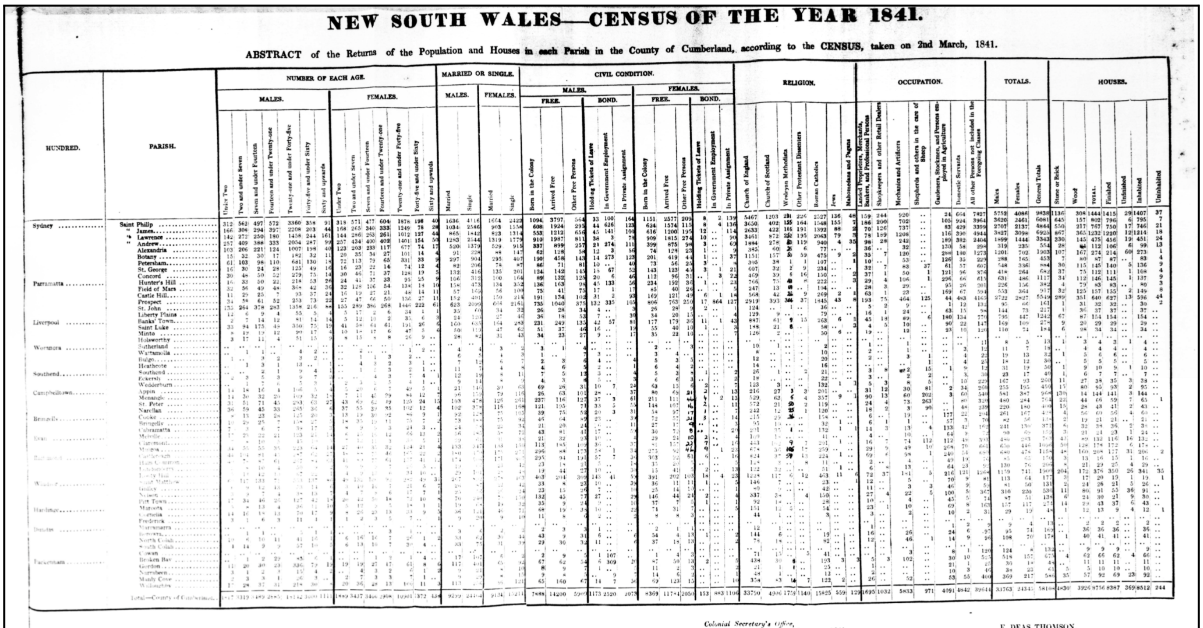
ABSTRACT of the Number of Inhabitants in the Colony of New South Wales, according to a Census taken the 2nd September, 1836, under an Act of the Governor and Council of 7th Wm. IV., No. 1, passed 5 July, 1836.

No. 260. January 25, 1837.

COUNTIES.	PERSONS ON THE ESTABLISHMENT.								GENERAL TOTAL.	RELIGION.			
	Male.				Female.								
	Free.				Free.								
	Above twelve years of age.	Under twelve years of age.	Convict.	Total.	Above twelve years of age.	Under twelve years of age.	Convict.	Total.		Protestants.	Roman Catholics.	Jews.	Pagans
Argyle	668	155	1,106	1,929	276	176	36	488	2,417	1,618	783	13	3
Bathurst	437	128	804	1,369	152	110	98	360	1,729	1,243	485	1	—
Bligh	112	13	219	344	17	14	1	32	376	273	103	—	—
Brisbane	220	51	974	1,245	81	36	16	133	1,378	1,004	373	1	—
Camden	905	301	1,168	2,374	454	276	57	787	3,161	2,158	982	18	3
Cook	678	306	366	1,350	377	298	27	702	2,052	1,517	534	—	1
Cumberland	12,724	4,810	7,254	24,788	8,454	4,730	1,825	15,009	39,797	29,090	10,270	371	63
Durham	740	204	1,749	2,693	287	174	54	515	3,208	2,300	904	4	—
Georgiana	192	50	227	469	65	36	5	106	575	339	234	2	—
Gloucester	138	60	520	718	79	53	4	136	854	628	222	3	1
Hunter	258	99	225	582	119	92	15	226	808	630	175	2	1
King	194	35	207	436	61	38	9	108	544	327	217	—	—
Macquarie	174	49	875	1,098	91	50	61	202	1,300	898	376	24	2
Murray	531	82	853	1,466	163	87	12	262	1,728	1,089	630	8	1
Northumberland	1,361	466	1,780	3,607	704	507	198	1,409	5,016	3,601	1,398	11	6
Phillip	54	5	169	228	12	6	1	19	247	117	130	—	—
Roxburgh	595	142	891	1,628	212	100	40	352	1,980	1,415	561	3	1
Saint Vincent	150	21	341	512	44	25	11	80	592	430	162	—	—
Wellington	163	19	283	465	38	22	5	65	530	357	172	1	—
Westmoreland	182	37	260	479	60	32	8	100	579	393	185	1	—
Without the Boundaries	1,256	103	1,300	2,659	190	104	15	309	2,968	1,784	1,164	2	18
Road and Ironed Gangs	19	3	2,190	2,212	10	7	1	18	2,230	1,394	829	7	—
Penal Settlements	17	17	1,493	1,527	4	19	78	101	1,628	926	697	5	—
Colonial Vessels at Sea	1,175	—	—	1,175	—	—	—	—	1,175	850	325	—	—
Port Phillip	178	8	—	186	23	15	—	38	224	210	14	—	—
TOTALS	23,121	7,164	25,254	55,539	11,973	7,007	2,577	21,557	77,096	54,621	21,898	177	100

[83]

Fig. 2a Example of the 1836 NSW Census Data illustrating a relatively small number of locations in geographical locations known as counties, and a total of just three top-level data categories (Counties, Persons on the Establishment, and Religion).



NEW SOUTH WALES—CENSUS OF THE YEAR 1841.

ABSTRACT of the Returns of the Population and Houses in each Parish in the County of Cumberland, according to the CENSUS, taken on 2nd March, 1841.

HUNDRED. PARISH.

NUMBER OF EACH AGE. MALES. Under Two; Two and under Seven; Seven and under Fourteen; Fourteen and under Twenty-one; Twenty-one and under Forty-five; Forty-five and under Sixty; Sixty and upwards. FEMALES. Under Two; Two and under Seven; Seven and under Fourteen; Fourteen and under Twenty-one; Twenty-one and under Forty-five; Forty-five and under Sixty; Sixty and upwards.

MARRIED OR SINGLE. MALES. Married; Single. FEMALES. Married; Single.

CIVIL CONDITION. MALES. FREE. Born in the Colony; Arrived Free; Other Free Persons. BOND. Holding Tickets of Leave; In Government Employment; In Private Assignment. FEMALES. FREE. Born in the Colony; Arrived Free; Other Free Persons. BOND. Holding Tickets of Leave; In Government Employment; In Private Assignment.

RELIGION. Church of England; Church of Scotland; Wesleyan Methodists; Other Protestant Dissenters; Roman Catholics; Jews; Mahomedans and Pagans.

OCCUPATION. Landed Proprietors, Merchants, Bankers, and Professional Persons; Shopkeepers and other Retail Dealers; Mechanics and Artificers; Shepherds and others in the care of Sheep; Gardeners, Stockmen, and Persons employed in Agriculture; Domestic Servants; All other Persons not included in the Foregoing Classes.

TOTALS. Males; Females; General Totals.

HOUSES. Stone or Brick; Wood; TOTAL; Finished; Unfinished; Inhabited; Uninhabited.

HUNDRED.	PARISH.
Sydney	Saint Philip
	" James
	" Lawrence
	" Andrew
	Alexandria
	Botany
	Petersham
	St. George
	Concord
Parramatta	Hunter's Hill
	Field of Mars
	Castle Hill
	Prospect
	St. John
	Liberty Plains
Liverpool	Banks' Town
	Saint Luke
	Minto
	Holsworthy
Woronora	Sutherland
	Wattamolla
	Bulgo
	Heathcote
Southend	Southend
	Eckersly
	Wedderburn
Campbelltown	Appin
	Menangle
	St. Peter
	Narellan
Bringelly	Cooke
	Bringelly
	Cabramatta
Evan	Melville
	Claremont
	Mulgoa
Richmond	Castlereagh
	Ham Common
	Londonderry
	Rooty Hill
Windsor	Saint Matthew
	Gidley
	Nelson
	Pitt Town
Hardinge	Maroota
	Cornelia
	Frederick
Dundas	Marramarra
	Berowra
	North Colah
	South Colah
	Cowan
Packenham	Broken Bay
	Gordon
	Narrabeen
	Manly Cove
	Willoughby
	Total—County of Cumberland

Colonial Secretary's Office, Sydney, 23rd August, 1841.

E. DEAS THOMSON.

Fig. 2b Example of the 1841 NSW Census Data illustrating increasing size and complexity of the collected data.

Complexities

HCCDA data is rich in two distinct ways. First, there are explicitly defined columns capturing information regarding details such as the gender, age, occupation, location, level of literacy, country of origin, religion, marital status, and type of building material for domestic buildings (wood, brick, etc). Second, there is a multitude of implicit information embedded into these categories, which deserve to be extracted and explicitly incorporated into a knowledge graph held in a triplestore (a type of graph database) - thus enabling a new type of query, a new type of research question, and a new way of engaging with the data.

The first investigation - carried out as a part of a day-long workshop where the participants consisted of data scientist (Terhi Nurmikko-Fuller), data archivist (Janet McDougall), and domain experts (Len Smith and Prof Paul Pickering, both historians of Australian colonial history) - identified location data as a possible starting point. However, two challenges in this data category quickly became apparent. First, there was the spatio-temporal instability of locations where the parameters and geo-coordinates of locations (such as Argyle, or Bathurst, for example) changed over time. Variations in the nomenclature and boundaries of place, such as parish, meant that we could not extract data directly from tables *en masse* from each census without careful consideration at a granular level. Second, the domain experts identified three ambiguous and indeterminate 'locations' from the 1833 census in New South Wales: "Road Branch, including Stockades", "Penal Settlements" and "Colonial vessels, at sea".

'Road Branches' refers to road construction of various types, undertaken at that time by convicts. Stockades were temporary facilities made from timber palisades and erected on route to accommodate the forced labour-force. As such, they were inherently transient, thus negating the possibility of consistently asserting, with any confidence, the geo-location data for these spots. Along the route of the Old Hume Highway today, for example, there are bridges and viaducts constructed in bluestone by convict work gangs at irregular distances. Progress was inevitably slow but random from day to day.

The vessels at sea are similarly impossible to pinpoint in terms of location. From a Linked DAta perspective, there is scope for the representation of *relative space*: the CIDOC Conceptual Reference Model for example would enable the representation of convicts on the vessel *vis a vis* the location of the vessel on the planet's surface.[3] A further complication is that both the colonies of NSW and Van Diemen's Land used prison hulks as floating prisons or 'holding pens' for prisoners awaiting transportation to other locations. The first floating prison (the *Phoenix*) was moored in Sydney Harbour between 1825 - 1837 covering the time of the first census of NSW in 1833. It is unclear from the HCCDA data whether those described as being on these vessels comprised some or all of those onboard the *Phoenix*. And, of course, prison hulks were not the only vessels in use at any one time.

Furthermore, the *Historical Frontier Violence* project[4] is currently mapping the boundaries of the locations mentioned in the HCCDA census data. For these reasons, we have opted to focus on data categories other than geolocation data.

[3] See for example https://www.cidoc-crm.org/f9-place.
[4] Available at https://melbourneinstitute.unimelb.edu.au/research/historical-frontier-violence/project-team

Linked Data workshops

Two full-day workshops have been carried out to date. The structure and content of the workshops follows an established approach (Nurmikko-Fuller, 2022). The first stage was to familiarise ourselves with the data - this essential approach proved its value in the first instance with the recognition of the complexities in the data as outlined above. The remainder of the session consisted of an analogue approach to information mapping, resulting in the preliminary ontological structure depicted in Figure 3. This process of intellectual heavy-lifting requires an investment of human time, but resulted in data mapping structures that enable us to capture nuanced information. The aim here is to have tools to create high-quality instance-level RDF in the future, which will require relatively little *ad hoc* cleaning in future. We strive for quality of instance level data (even at smaller scale), rather than the automated large-scale conversion of Big Data into a knowledge graph.

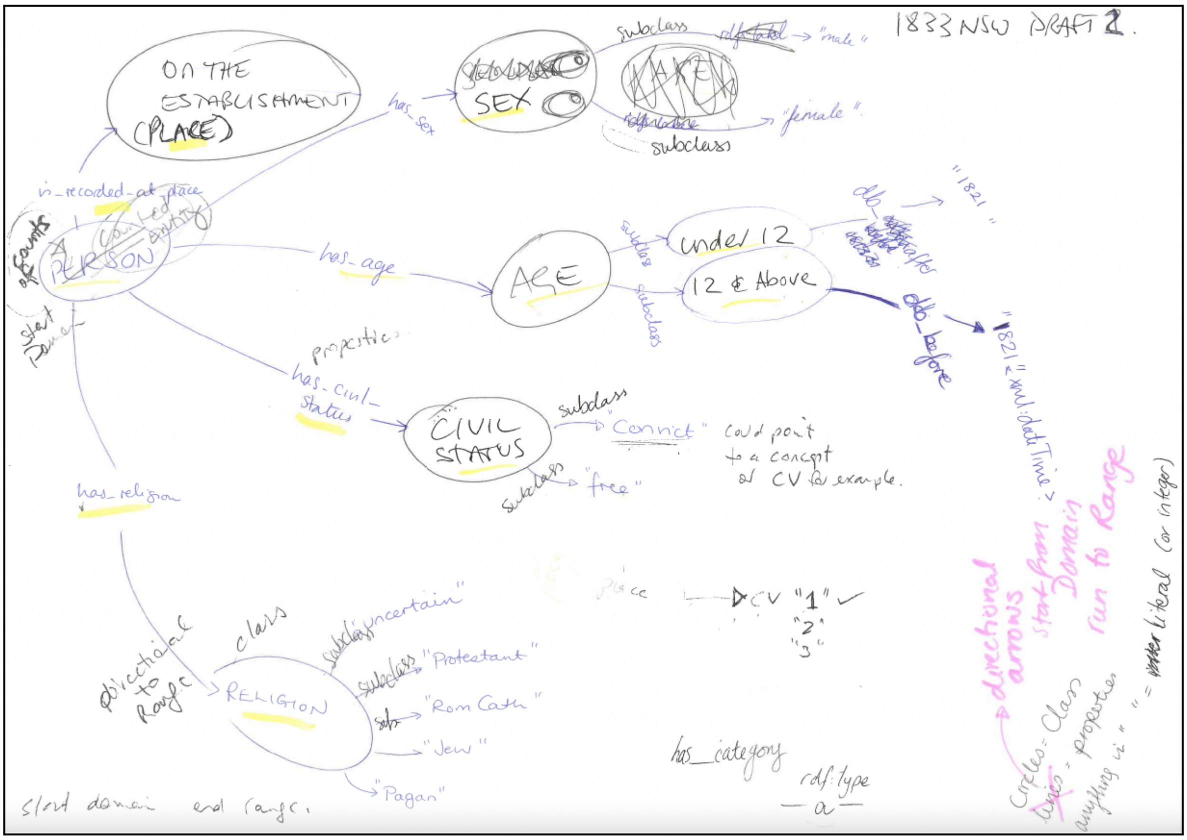


Fig. 3 Pen and Paper Drawing of the Ontology

The second workshop saw the implementation of the ontological model into a machine-processable file, expressed in .TTL. The tool of choice (again following Nurmikko-Fuller, 2022) was an Open Source tool from the University of Stanford, Protege[5]. This tool was selected as it is free (as in *libre*), and also free (as in *gratis*). Rather than install the software onto a piece of hardware, the project was completed in the browser-run version of the tool.

The ontological structure was exported out of Protege as .TTL and uploaded into WebVOWL[6] another free piece of software, this one from MIT. The resulting visualisation is depicted in Figure 4.

[5] Available at https://protegewiki.stanford.edu/wiki/Main_Page

[6] Available at https://service.tib.eu/webvowl/.

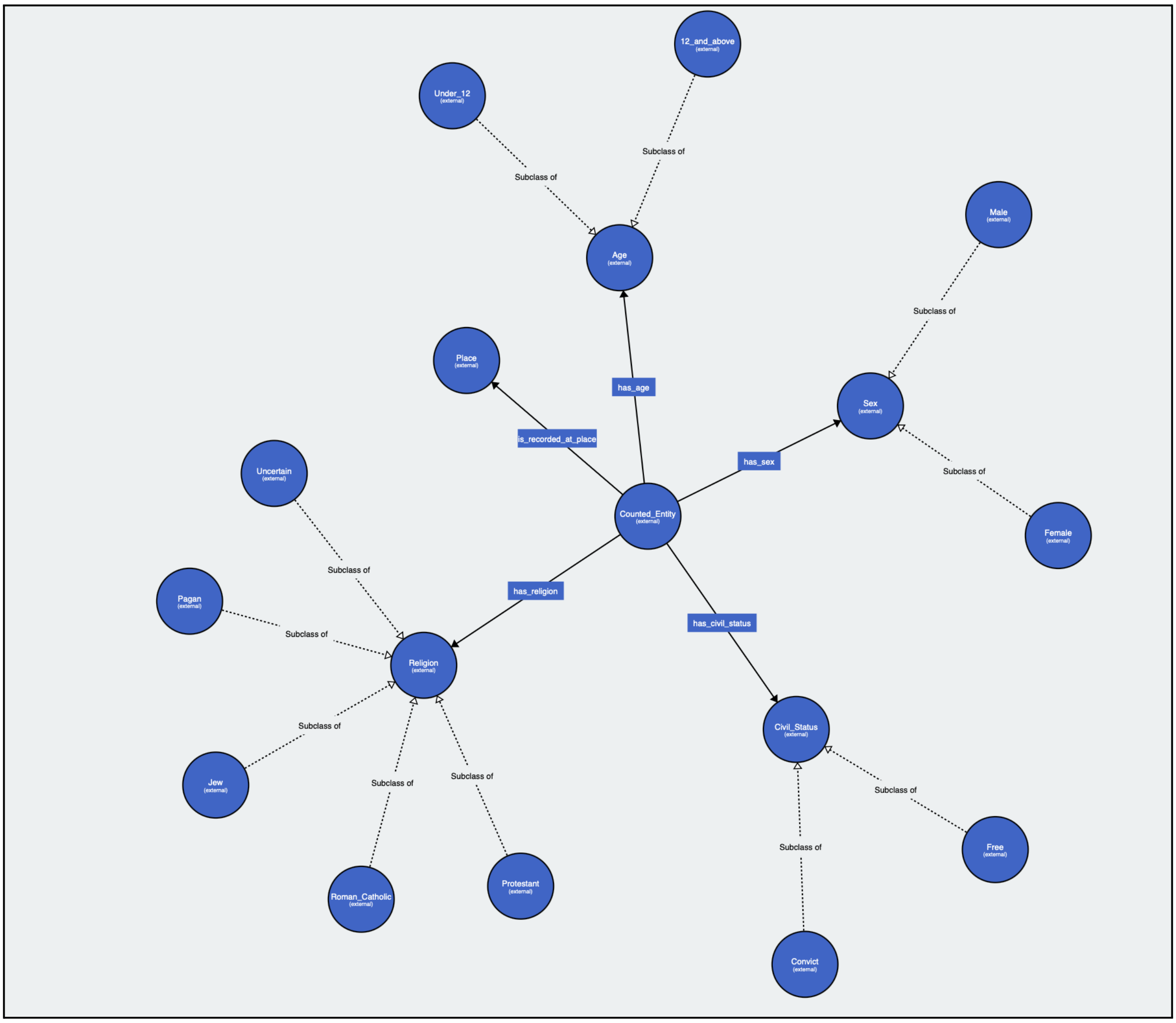


Fig. 4. Visualisation of the ontological structure underlying the knowledge graph for 1833 HCCDA data

The completion of the workflow for converting tabular data to Linked Data will require two future workshops. The first of these will see the mapping of instance level data to the ontological structure and the production of instance level RDF capturing the HCCDA data in a knowledge graph. The fourth and final workshop will see the uploading of the instance level RDF into an instance of a triplestore, and the running of SPARQL queries over that data. The triplestore of choice is likely to be either Virtuoso, Blazegraph, or GraphDB as the research team has expertise and experience in using these graph databases/triplestores.

Focusing on Occupations

A separate but complementary research project has run parallel to the Linked Data workshops. The aim of this investigation was to focus on a specific data category to pursue. The applied method combined close reading and data munging (cleaning, reordering, and extraction). This process had the dual function of representing the explicit data categories and top-level professions - offering increasingly comprehensive snapshots of the colonial economies - and highlighting a number of implicit data categories including but not limited to details about public and private organisational structures, familial relationships, the cultural heritage sector, educational systems, and even medical diagnoses. The possibilities for research are considerable. Indeed, bringing a Linked Data approach to raw census data is inherently generative. Our aim here is not to offer conclusions in relation to the data itself but rather to highlight possibilities.

Occupations data has been examined through two methods. First, we examined data from a NSW in each of the censuses to record occupation data (i.e. 1841, 1846, 1851, 1856, 1861, 1871, 1881, 1891, and 1901. The census data from 1833 and 1836 did not record occupational data).

Second, in order to compare like with like we undertook an investigation of occupational data across all six colonies where there the years aligned (resulting in four datasets: 1861, 1881, 1891, and 1901, see Table 1 above).

The first stage of the data processing consisted of seven steps:

1. Files were retried as .xtables, and converted into .html
2. A specific location and year were selected (e.g. NSW, 1881)
3. Occupations data was identified and selected from the .HTML file and exported into MS Excel
4. Superfluous and messy data was removed
5. Lists from several different .HTML files were merged to create one definitive list
6. Duplicates were removed (based on a basic character match)
7. Close/semantic duplicates (e.g. accountant clerk; clerk, accountant) were identified and removed through a process of close reading and domain expert consultation, resulting in a definitive list of occupations (excluding industries and statuses).

The seventh step included rationalising naming conventions within the data or removing non-occupational information. The use of an industry title instead of, or in relation to, the actual occupation was cleaned. For example, “mining” or “mining *vide* miner”) were standardised to “miner”. Entries that were not occupations per se (such as “beggar” or “pensioner”)and familial relationships (“father, dependent on children”) were removed. A challenge that defied automation was the disambiguation between what we are describing as semantically inconsistent homonyms. These include occupations such as “blacksmith”, “whitesmith”, and “locksmith”, which were not necessarily all variants of an occupational type denoted by a broad heading such as “smith”. In order to maximise the value of the data as a resource for further research, this stage involved bringing into play significant domain and linguistic knowledge)as well as some common sense!).

The resulting lists of occupations will be published as “Linked Data-ready” vocabularies. This will consist of the minting of HTTP URIs for each of the occupations, with embedded semantics to capture location and date (i.e. the URIs will take the form <http://hccda.ada.anu.edu.au/Occupation/NSW/1881/#####>). Once made available online, these lists can act as external authority files, enabling data aggregation between disparate datasets which nevertheless contain information about the same things.

Future steps of the project will see the establishment of a top level vocabulary, which will map occupations regardless of spatio-temporal data. There will thus be a top-level concept of, for example, "baker", and associated with it will be each instance of "baker" from each dataset. This step is completed in order to facilitate the capture of a data provenance trail which will connect each job role to the specific .HTML file whence it was derived.

These vocabularies will be documented and published online. Once publicly available, the datasets will be reported on in a journal article submitted to the Journal of Open Humanities Data, resulting in two distinct types of academic output: the dataset(s) and the journal article reporting on them.

Future in Harmonisation

Implementation of existing authorised vocabularies promotes data harmonisation, which in turn facilitates data aggregation. To enable future information blending through the bridging of disparate datasets, it is important (and timely!) to identify other datasets, accessible online and containing relevant data. The information structures, design decisions, and metadata standards (or vocabularies) utilised by other projects for similar or even overlapping data can also highlight thoughtful and well-informed solutions for information representation. In other words, we can stand to benefit from the decisions made by others working on similar datasets in terms of how to structure, and indeed what to structure (this may and often will include decisions of what to omit as well as what to include).

The History of Work Information System[7] is a prime example of such a project containing complementary data. HISCO has its origins in work developed in the 1950s by the International Labour Office, which developed an International Standard Classification of Occupations (ISCO) allowing classification of occupational activities worldwide. HISCO as we know it today is derived from the original ISCO working group outputs, and was published originally in 1968 to capture the occupational titles of the twentieth and nineteenth centuries (and thus providing the possibility of specific, instance level overlaps with our definitive vocabularies.

The HISCO system was designed specifically to work with historical data, but it also provides an opportunity to benefit from embracing the Linked Data methodology: we can aggregate data from sources which differ in many ways but can be seen to have conceptual overlaps. At its most abstract, we can still assert that both these datasets contain some form of information pertaining to occupations and employment, even if for now further investigations into them cannot readily be automated. As quoted on their website: "It seems clear then that comparisons of important historical structures and processes would be a little less problematic if comparability in the coding of occupations was achieved".[8]

CEDAR (The Dutch Historical Censuses as Linked Open Data,[9] in many ways serves as an ideal project to bridge to in the future. As with our Occupation data, CEDAR has published historical occupation data in a structured manner. Aggregation of two Linked Open Data projects benefits from the RDF data model and enables us to leverage the heavy computational lifting possible through knowledge graph implementation. A single point of access would enable

[7] Available at https://historyofwork.iisg.nl/index.php

[8] Available at https://historyofwork.iisg.nl/detail_page.php?act_id=35200)

[9] Available at https://www.semantic-web-journal.net/system/files/swj1234.pdf

simultaneous and enriched querying (via SPARQL queries or an API) and make it possible to search through both datasets consistently - the results of the query would be enriched and specified by information available from both datasets. Potential future developments of the HCCDA could include the publication of HCCDA data in aggregate with the CEDAR and HISCO datasets, accessible through a single entry-point.

Acknowledgements

We would like to gratefully acknowledge the contributions of Len Smith. Len was a contributor to the first workshop, and PI of the ARC LIEF grant that saw the transcription of the original census data and the subsequent creation of the HCCDA dataset. He was instrumental in providing access to outputs beyond the official HCCDA project ones, such as presentations and slide decks. He has served as advisor and friend.

References

Smith, L., Rowse, T., Hungerford, S. (2019) *Australian Bureau of Statistics; Australian Data Archive*. In "Historical and Colonial Census Data Archive (HCCDA)", doi:10.26193/MP6WRS, ADA Dataverse, V5

Australian Bureau of Statistics, 2006, 1301.0 - Year Book Australia, 2005. The population census – a brief history, viewed 27 November 2023, https://www.abs.gov.au/AUSSTATS/ABS%40.NSF/Previousproducts/1301.0Feature%20Article92005?opendocument&tabname=Summary&prodno=1301.0&issue=2005&num=&view=

McEachern, S. (2021). Project proposal for the Integrated Research Infrastructure for Social Sciences (IRISS). Zenodo. https://doi.org/10.5281/zenodo.6552037

Nurmikko-Fuller, T. (2022). Teaching Linked Open Data using Bibliographic Metadata. *Journal of Open Humanities Data*, *8*(1).