Integrated Research Infrastructure for the Social Sciences

Work Package 3

Technical Report 2:

Data Linkage Report

Matthias Kubler, Denise Clague, Tomasz Zajac, Wojtek Tomaszewski, Jonathan Corcoran

31 May 2023

Contents

Tables

Figures

Abbreviations

Commonly used abbreviations in this report

Abbreviation	Definition
ABS	Australian Bureau of Statistics
ADA	Australian Data Archive
AGEP	Age
ANU	Australian National University
ARIA	Accessibility/Remoteness Index of Australia
ASCED	Australian Standard Classification of Education
ASGC	Australian Standard Geographical Classification
ASGS	Australian Statistical Geography Standard
AURIN	Australian Urban Research Infrastructure Network
DLOD	Dwelling Location
DWTP	Dwelling Type
EETP	Engagement in Employment, Education and Training
EMPP	Number of Employees
GNGP	Public/Private Sector Indicator in the 2016 Census
HEAP	Level of highest educational attainment
HILDA	Household, Income and Labour Dynamics in Australia
HRSP	Hours Worked
HSCP	Highest Year of School Completed
INCP	Total Personal Income (weekly)
INGDWTD	Indigenous Household Indicator
IRISS	Integrated Research Infrastructure for the Social Sciences
ISSR	Institute for Social Science Research
LFSP	Labour Force Status
LSAY	Longitudinal Survey of Australian Youth
NCVER	National Centre for Vocational Education Research
NIRS	National Information and Referral Service
NPDD	Type of Non-Private Dwelling
PISA	Programme for International Student Assessment
SEIFA	Socio-Economic Indexes for Areas
SIEMP	Status in Employment
STRD	Structure of Dwelling
STUP	Full-Time/Past-Time Student Status

1. Introduction

The aim of the Integrated Research Infrastructure for the Social Sciences (IRISS) Project is to address the fragmentation of the Australian social science research infrastructure. Within the IRISS project, Work Package 3 focuses on developing a data integration service called GeoSocial which will allow people-centred survey data to be augmented with spatially structured data capturing information on places where these people live. The lack of such data has been identified as one of the major barriers hindering social research in Australia. At this stage, the project aims to develop a working prototype of the service which might be scaled up in the future. The prototype will be then used to generate linked data for the associated Demonstrator 1. This will showcase the analytic potential of geo-social data integration, or more specifically, the added value of survey data enhanced with information about places.

This report is preceded by Technical Report 1 published on 31 August 2022 and the Preliminary GeoSocial Service Design published on 31 March 2023. The previous report focused on technical requirements of the online service and software toolkit that will constitute the GeoSocial service. They discussed, among others, User Requirements, Preliminary Service Design, and Software Development Outputs for the integration service. In turn, this report focuses on the methodological aspects of data integration. Along with the previous documents, this report will inform the final stage of the GeoSocial prototype development and the creation of the demonstrator dataset, as well as the Technical Report submitted by Australian Urban Research Infrastructure Network (AURIN) to the Australian National University (ANU) by 30 June 2023 summarising the overall design and development of the project's operational pilot.

Information in this report is divided into two main parts. The first one (Section 2) focuses on survey data to be augmented with information about places drawn from the Australian Bureau of Statistics (ABS) Census. The initial review of Australian Data Archive (ADA) surveys, conducted at the beginning of the project, identified the Household, Income and Labour Dynamics in Australia (HILDA) Survey as the most suitable dataset to demonstrate the potential of the GeoSocial service. The HILDA Survey has generated significant interest among members of the research community, as indicated by the substantial number of data download requests and linkage requests it has received. The data provided by the HILDA Survey were collected in multiple waves which make it a good example of the service's capability for temporal data integration. Unfortunately, the HILDA Survey custodian, the

Department of Social Services, did not agree for the restricted version (including geographical identifiers) of the dataset to be used in the project. This forced the project team to search for an alternative data source. The Longitudinal Surveys of Australian Youth (LSAY) was identified as a suitable source of data. Although it is not as popular as HILDA (see Appendix 2 for the most downloaded ADA surveys) it covers a wide range of topics making it interesting to a good section of the research community. Furthermore, it uses nationally representative samples of students at school that match the Programme for International Student Assessment (PISA) sample (see Appendix 1 for information about the LSAY sample design). As a longitudinal study it consists of multiple waves which allow for temporal analysis. This scoping study presented in the first section of this report is equivalent to an earlier one that reported on the use of HILDA data in the context of some spatial analysis component (included as an Appendix in IRISS Technical Report 1). Its goal is to review previous work that involved LSAY so as to identify research themes and analyse the methodology of previous research.

The second part of the report (Sections 3 through 5) focuses on conceptual, methodological, and practical issues related to linking area-based data (e.g., derived from the Census) to person-level data and using spatial characteristics as predictors in the analysis of individual outcomes.[1]

Technical Report 1 alerted to issues with pursuing this option of combining spatial data to survey unit record data (where each record represents observations for a person or household). Including challenges to do with survey data due to the underlying sample design and subsample sizes, concordance between survey and spatial data to be integrated, such as geographical and temporal alignment, and temporal inconsistencies. Having selected LSAY as the survey for integration with spatial data, we consider these issues specifically in relation to this survey and undertaking data integration of LSAY with the Census at some level of geography. The information in Section 3 describes some of the typical changes affecting categorical variables in the Census data collections that can occur over time. Section 4 considers limitations and problems surrounding temporal inconsistencies of spatial data definitions and categorisations. Section 5 presents the various ways in which temporal inconsistencies could be addressed as part of a data integration service design and outlines one way of solving spatial and temporal inconsistencies in the specific case of Demonstrator 1 that combines LSAY data with Census data. Resolving issues related to data integration enables more complex types of integration,

[1] There is an alternative approach to data integration, i.e., aggregating person-level records to produce estimates characterising areas. However, this is not feasible given the sampling methodology of major Australian surveys (for more detail see Technical Report 1).

which in turn broadens the appeal of the output datasets and the GeoSocial service to researchers.

A final section (Section 6) concludes the report, whilst also outlining next steps and provides suggestions for future extensions.

2. Previous work with LSAY data involving a spatial analysis component

In this section we review previous research that involved LSAY data or data from its predecessor the Australian Youth Survey and included a spatial analysis component. Compiling this work allowed us to identify the topics and main research interests as well as to gather information on the previously used methodology, specifically the ways in which spatial characteristics were derived and utilised in the analyses. Documenting this will help with flagging up conceptual design issues, building the Demonstrator 1 dataset, refining the selection of analytic variables, and further developing the analytic plan.

2.1 Approach

This scoping review was executed in three steps, which mirrored the steps undertaken in the earlier HILDA document in Technical Report 1: 1) identifying published work involving LSAY, 2) identifying LSAY work that involved some spatial component, and 3) identifying topics of work and in which way spatial information played a role and was handled.

Two lists with LSAY publications were identified:

1. On the LSAY website https://www.lsay.edu.au/publications/reference-sources. There were 249 outputs grouped into four types as per Table 1.
2. In an appendix to a report that documented a literature review of LSAY publications - *NCVER 2020, Longitudinal Surveys of Australian Youth (LSAY) analysis: literature review — support document one (official and grey literature reference list)*. The appendix listed 468 outputs, which were differently categorised to the listing on the above LSAY website (and therefore not broken down in a table here). This listing also includes technical documentation, such as codebooks and questionnaires, which inflated the number of outputs.

There was considerable overlap between the above two listings.

Table 1. LSAY-related outputs listed on the LSAY website

Type of output	Number
Book chapters	3
Peer-reviewed journal articles	104
PhD and Masters theses	11
Grey literature	131
Total	**249**

Source: https://www.lsay.edu.au/publications/reference-sources

To identify works that included a spatial component, consecutive searches for title words were then performed on both lists using these search terms:

> *"spatial", "geogra", "region", "remote", "rural", "metropolitan", "area", "location", "migrat" and "move".*

Hits for any of those searches were copied over to a new list of LSAY publications. In the process of going through the new list it was complemented with literature that appeared to be relevant and which was referenced in pieces already included on the list. The resulting list contained 23 pieces of literature and was treated as the universe of LSAY publications involving a spatial component. All but one publication could be downloaded.

Each of the 22 downloaded publications was then scrutinised in relation to methodical information and the overall topic of the publication. Of particular interest in this process were what geographical information was used at what level of the geography, where it came from/how it was derived, and how it was used in the analysis.

2.2 Limitations

The approach outlined above relied, to some extent, on accurate and updated compilations by the National Centre for Vocational Education Research (NCVER) of all LSAY-related publications (step 1). The search methodology to identify relevant work that involves some spatial component relied on such work being reflected in the title of the publications (step 2). One publication could not be downloaded to date and has not been scrutinised as a consequence (step 3).

The character of the 'scrutinising' of existing work, at this point, relied more on scanning than on detailed reading to get through all available publications. This concerned particular sections of publications (most often Methods and Data sections, and Abstracts) to identify relevant

information, in the process of which such information may have been missed in other sections of the publications (step 3).

All of these matters constitute limitations for the work presented in this review. Some limitations could still be minimised in the future, for example, by expanding search techniques (including the utilisation of data bases) in step 2 and/or by gaining access to publications not accessible to date and/or by revisiting individual publications to explore more detail than was apparent when scanning the publications.

Despite the limitations, this scoping review should fulfill its main purpose of informing work on the IRISS project by identifying in which ways spatial information has been considered in analyses of LSAY data. A summary of insights is provided next.

2.3 Insights from the scoping review

General types of data analyses designs involving spatial data and LSAY

There are three general ways in which spatial data have been used in conjunction with LSAY data:

a) Work where spatial areas are selected as an area of interest (as a filter) on the basis of which some analysis is performed. This is reflected in selecting a sample in the LSAY data by some geographic criterion/criteria. In the works investigated, this involved selecting people who lived in metropolitan or non-metropolitan areas in Australia or in particular states, or in individual cities like Melbourne.
b) Work where spatial areas or their characteristics are controlled for in the analysis. In the works investigated, this most prominently involved using categories of remoteness and/or derived categories from Socio-Economic Indexes for Areas (SEIFA) scores (e.g., deciles, quintiles) as control variables in models.
c) Work where spatial areas, types of areas or their characteristics are directly considered as possibly influencing some 'outcome'. In the works considered here, 'outcomes' in such research were in the areas of educational milestones and labour market statuses and employment.

As far as could be determined from this investigation, LSAY data were not used to generate estimates for spatial characteristics for finer levels of geographies.

Sourcing spatial information in analysis of LSAY data

In principle, spatial information can be sourced from within LSAY or added from external sources to the LSAY data. Some of the included studies made use of the information included in the LSAY data. This most prominently concerned the characterisation of respondents' environment as urban versus regional, which was used to filter for respondents (data analysis design type a) or for using the urban versus regional variables as a control (data analysis design type b) or predictor (data analysis design type c) in modelling. For example, Chesters & Cuervo (2022) modelled the likelihood of university enrolment based on such status (similar Curtis, Drummond, Halsey & Lawson, 2012).

Some researchers defined geographical mobility based on changes in residential postcode across LSAY waves in the data with a particular focus of mobility between metropolitan and non-metropolitan areas (e.g., Hillman & Rothman, 2007). To this end, remoteness information from external sources was merged to postcodes in LSAY records and mobility then defined by changes in the remoteness status rather than changes in postcodes.

There were other works that involved merging information from external data sources to LSAY records. Types of information that was merged from external sources included:

- More detailed information on the remote or urban character of a respondent's environment by linking the existing postcodes with Accessibility/Remoteness Index of Australia (ARIA) scores or other existing categories of the Australian Statistical Geography Standard (ASGS) or Australian Standard Geographical Classification (ASGC).
- Locations of higher education institutions (longitude and latitude), which were, in conjunction with respondents' residential postcodes, used to calculate measures such as *Distance to the nearest higher institution of learning* (Adejoro, 2016, similar Parker, Jerrim, Andres & Astell-Burt, 2016), which then served as a predictor for aspirations and/or post-school transitions.
- Socio-economic and demographic information on areas, which could entail SEIFA indices (Adejoro 2016) or information on qualifications, income, ethnic diversity, household composition and turnover and other characteristics from the ABS Census (Andrews, Green & Mangan, 2002; Johnston, Lee, Shah, Shields & Spinks, 2014).

The geographical basis for merging external information to LSAY records was postcode, usually respondents' residential postcode, but also their schools' postcode as captured in the first wave. Postcode (its population-weighted centroid) was also used when calculating distances between LSAY respondent residences and the closest higher education institutions.

Outcomes influenced by spatial matters

LSAY tracks groups of Australian youth with the aim of studying their school and post-school transitions and research involving some spatial component reflects this. All identified works investigated outcomes related to young people's education and training, employment and/or social development in some way.

Spatial data of influence (as a filter or as influencing outcomes)

Remoteness/urbanisation → educational outcomes

One prominent research topic was the relationship between the remoteness or urban/non-urban character of areas in which young people grow up in and their educational outcomes. This involved investigating the pre-cursers to later educational outcomes, such as student intentions and/or student performance while at secondary school, and completing high school/early school leaving, as well as later educational statuses, such as attending university, completing university, or attaining other tertiary qualifications. Works by Acer (2002), Curtis et al. (2012), Cardak, Brett, Bowden, Vecci, Barry, Bahtsevanoglou & McAllister (2017), DESE (2020) and Chesters & Cuervo (2022) all fall under this theme as does Jones (2002) who investigated such relationships in the context of assessing LSAY as a potential source for national reporting of educational outcomes by geographic location.

A special application of the above type of research was the investigation of the relationship between the remoteness status of a region and the Indigenous gap in high-school completions (Schellekens, Ciarrochi, Dillon, Sahdra, Brockman, Mooney & Philip, 2022).

'Neighbourhood' → educational outcomes

A broadening of this research theme consisted of bringing in additional socio-economic area information when considering educational outcomes, de-facto broadening the concepts of remoteness and urbanisation to more theoretically considered concepts of 'neighbourhood'. As mentioned further above, such additional information could, for example, include information on income, educational qualifications, ethnic diversity for areas or SEIFA indices (Cooper et al, 2018; Johnston et al., 2014; Ryan, 2011).

A special application of such broadened research was pursued by Lim, Gemici, Rice & Karmel (2011) who investigated relationships between area-defined SES measures against an individual SES measure created from within LSAY data to assess the validity and utility of the former (as they are commonly used in educational policy contexts in Australia).

Remoteness/neighbourhood → employment and other outcomes

Another direction of broadening the research concerned the outcomes of youth transitions or trajectories to include, or focus on, employment outcomes and independent living (Andrews, Green & Mangan, 2002; Adejoro, 2016; Rowe, Corcoran & Bell, 2014).

The role of distance and geographical mobility

A final category of analysis work involving LSAY and a spatial component considered the role of geographical distance or geographical mobility in youth transitions. Parker, Jerrim, Anders & Astell-Burt (2016) examined the influence of distance to university on youth's aspirations at secondary school and their later university enrolment. Using data from Victorian respondents of the 2003 LSAY, Rowe, Bell & Corcoran (2014) explored typical sequences of mobility over a 9-year period and then investigated educational and employment outcomes for different migrant/non-migrant types (Row, Corcoran & Bell, 2014). Hillman & Rothman (2007) focused on a cohort of youth living in non-metropolitan areas when in Year 11 and explored both predictors and consequences of their geographical mobility.

However, if the publications using postcodes to match to other geographies (especially remoteness) are excluded, there appear to be only a handful of publications that have merged information from external sources to LSAY records. In this sense such spatial data integration with LSAY may still be seen as relatively novel. However, due to the specific cohort targeting over a 10-11 year period of LSAY, publications using LSAY data appear to be fairly homogenous in terms of the topics they investigate and the results they produce. This is a limitation of LSAY data that arises due to:

- The constriction of the cohorts' life course window (and associated with it the restriction of topics covered).
- The many correlations of different outcomes that are captured (most notably within and across the domains of education and employment).
- The similarities in the predictors for different outcomes.
- The stationarity of investigated outcomes and relationships which may change over time but they do not tend to change dramatically.

The above points in combination limit the scope of LSAY for discovering and publishing new insights (with the possible exception of matters affected by the pandemic). This limitation may well be compounded in this demonstrator project here by some of the methodological issues that are specific to spatial/mobility interests that are listed in the following sections as these issues tend to further constrain the potential for detailed investigations.

2.5 Levels of geography used in analyses

Some popular geographic levels used in analysis designs are included in the LSAY data. These include metropolitan/urban versus non-metropolitan/non-urban areas and postcodes. Other geographic levels of interest were those represented by categories of the ABS ASGC or ASGS, most prominently related to main, section-of-state or remoteness structures. The latter were generated for LSAY records using geographical concordances that translated non-ABS postcodes to the relevant geographic categories of the ABS.

2.6 Methodological issues

The review did not identify many discussions or treatments of (longitudinal) spatial data integration issues and no treatment of longitudinal non-spatial data integration issues[2]. Inconsistent postcode collections between the 2003 and 2006 LSAY waves were noted as a limitation in Parker et al. (2016).

A noteworthy treatment of spatial information was undertaken when defining migrant types in Rowe, Bell & Corcoran (2014). Migration categories were defined based on information on higher-level metropolitan versus non-metropolitan area breakdowns within and outside Victoria (not the underlying postcodes). This was similarly undertaken by Hillman & Rothman (2007) when considering migration paths of non-metropolitan Australian youth.

This practice could be of interest for the demonstrator research project, for example, when building up higher levels of geographies, such as SA3s or SA4s from postcode information in LSAY with the result of:

- Achieving larger sub-sample sizes per geographical unit.
- Possibly reducing respondent error when disclosing geographical (postcode) information in LSAY.

[2] It is possible that treatments of spatial and non-spatial longitudinal inconsistencies may emerge more fully when the identified publications are further scrutinised.

- Possibly reducing issues surrounding spatial longitudinal inconsistencies.

Rowe, Bell & Corcoran (2014) also pointed out that LSAY data do not allow analysis at finer spatial levels due to its sample design as well as sample sizes associated with geographies. In their specific application the authors content that "the available data only enable to explore the educational, occupational and mobility pathways followed by young Victorians at a coarse spatial aggregation that distinguishes Melbourne and regional Victoria" (p26). And while mobility was already constrained to mobility between metropolitan and non-metropolitan areas (and vice versa), LSAY sample sizes also limited a consideration of a larger number of mid-term mobility sequences/types when investigating educational and employment pathways as groups of Victorian youth with different mobility sequences were scarce in the data.

While discussions surrounding consistency in boundaries or other data definitions were scarce, the literature contained some general remarks of a methodological nature that concerned the *spatial characteristics* → *outcome* type of analyses, including:

- The geographical unit chosen influences the results of the analyses (Manski, 1993).
- The 'reflection problem' (Manski, 1993) can occur when building up spatial information from survey respondents that inhabit a space and then trying to establish whether individual outcome(s) for those within those spaces depend on the spatial attributes derived from the same individuals.
- Confounders with area characteristics can be common and need consideration in data analysis designs – high correlations between socio-economic spatial components, such as income, occupational status and qualifications can become problematic when included simultaneously in the same model(s), particularly when sample sizes for different spatial units are small (Andrews et al., 2002; Johnston et al., 2014).
- As part of the above: "The effects of a neighbourhood are sometimes difficult to separate from the impacts of schooling because of the correlation between the two" (Johnston et al., 2014).
- Area characteristics that are more closely aligned with outcome variables tend to show stronger effects (e.g., unemployment in neighbourhood is likely to be more strongly related with individual unemployment than, say, household wealth in the neighbourhood, also see 'reflection problem' above) (Manski, 1993.

And two points about geographical mobility made above are repeated here:

- Sample design and sample sizes associated with different areas and types of movers limit the detail with which spaces and mobility types can be considered in any analysis (Rowe, Bell & Corcoran, 2014).
- Attrition in longitudinal surveys will bias the sample towards the geographically immobile (Rowe, Bell & Corcoran, 2014). This could also be associated with other attributes relevant for post-school outcomes.

Publications using LSAY data, similarly to those using HILDA data, tended to circumvent issues of longitudinal inconsistencies by aspects of the data analysis design (e.g., selecting spatial characteristics at one point in time, or by defining migration at higher geographical levels) or already at the point of formulating a research question.

2.7 Summary

There have been 23 pieces of literature that have undertaken some type of spatial analysis using LSAY data. Of those publications downloaded and scanned, only a handful involved merging some spatial information from external sources to the LSAY records. It was more common to use the limited spatial information already supplied in the LSAY datasets (particularly the characterisation of a respondent's environment as urban versus regional).

The most prominent outcomes in research involving LSAY and some spatial component appear to lie in the domain of educational outcomes. Prominent spatial influencers on outcomes were seen in types of remoteness or urban/non-urban character of areas.

Most of the previous publications did not or did very little elaborate on data integration issues and implications. Those that did, tended to point to the approach of building up higher level of geographies from the postcode information in LSAY. One pointed out that the sample design and sample sizes associated with certain geographies did not allow analysis of more granular spatial levels.

There is still much work to be done using the LSAY data particularly in terms of the possibilities and limitations for spatial data integration. The following sections discuss such issues of the LSAY data that have now been considered in the integration of spatial data to the LSAY data. Issues around consistency over time of the variables of interest in the LSAY data files and the Census. New issues may arise as work continues on the service demonstrator and

operational pilot. Such issues will become part of the technical report 2 to be submitted to ANU by 30 June 2023.

3. Temporal inconsistencies of non-spatial data

The LSAY and Census data chosen for Demonstrator 1 and Work Package 3 of the IRISS project can be linked to both a spatial and a temporal component. It is important to assess the consistency of information in analytic datasets over time if the analysis includes a temporal component.

This section outlines some of the typical changes affecting categorical variables in the Census data collections that can occur over time. Types of changes are illustrated using examples, which relate to changes between the 2016 and 2011 Censuses. This is accompanied by brief discussions of the documentation of such changes in ABS materials and how changes could be addressed when analysing data across Censuses.

The ABS documents changes to Census variables, whether triggered by changes in the data capture or the data processing, in a 'What's New for <year>' section, which is part of the respective Census Dictionary for that year. The examples given to illustrate types of changes in this document were sourced from such a section in the 2016 Census Dictionary (https://www.abs.gov.au/ausstats/abs@.nsf/Latestproducts/2901.0Main%20Features202016?opendocument&tabname=Summary&prodno=2901.0&issue=2016&num=&view=)

The 'What's New for <year>?' sections do not fully document changes, which will be pointed out in this document. To better illustrate changes to variables, screen shots from relevant sections of the 2011 and 2016 Census Dictionaries are included in the presentation below. Where quotations are used in the document these are from the ABS and relate to the 'What's New for 2016?' section.

3.1 Types of change

This section outlines different types of changes. This starts by presenting types of changes that are more difficult to detect or to assess.

Change in mode of participation/collection

Census data collections have been moving towards online administration over the past three Censuses. In 2011, about one third of Census completions (at the household level) were undertaken online, in 2016 about two thirds. This was expected to rise to 75% in the 2021 Census. Offering dual mode completion has been associated with the ABS implementing changes in wording and layout between the paper and online versions of the household questionnaire to optimise the questionnaire for the online environment, but also generally: "The development of the online questionnaire for 2016 has provided an opportunity to make refinements to gain more accurate data from respondents, while decreasing the burden placed on those filling out the form."

Identifying differences between online and paper versions of the Census questionnaires may require independent investigation. Assessing how such changes affect responses will be hard to quantify.

Changes in the mode of participation may have also affected how some information is captured/collected: "The move to a new method of conducting the Census also meant a change to how data on Dwelling Location (DLOD), Dwelling Type (DWTP), Structure of Dwelling (STRD) and Type of Non-Private Dwelling (NPDD), previously recorded by Census collectors, are obtained." While the ABS goes on to provide more information on the change for the variables mentioned, this information does not easily make apparent what the change consisted of, without a more intimate understanding of Census data collections over time:

"There has been a change in the way this information is collected for 2016. It was recorded by ABS Address Canvassing Officers in the lead up to the Census as part of establishing the Address Register as a mail-out frame for designated areas. In areas enumerated using the traditional approach of delivering forms, the information was collected by ABS Field Officers during the Census collection period. Dwelling type was also updated as required by ABS Field Officers during the 2016 Census enumeration period."

Is a Census Collector equivalent to an ABS Field Officer? And if Address Canvasing Officers recorded the information before the Census in 2016 as opposed to Census Collectors during the Census in 2011, did both use the same observational code frame for recording the information?

Users of ABS Census data, particularly users of time series or longitudinal Census data should be alerted to such differences between online and paper version or changes in the way that ABS staff collect information that is included in the Census.

Change in question wording while retaining response options

Independent of moving the Census data collection to an online environment, the ABS sometimes makes (slight) changes to question wording between Censuses that can be hard to pick up when variable names, value categories and value labels remain the same. At times, such changes consist of changes to secondary guidelines, such as giving examples of acceptable entries in open-ended fields.

There were several changes to question wordings or accompanying instructions in the 2016 Census. The ABS document such changes, often in a descriptive format as is shown in the example below.

Example: Variable Highest Year of Schooling completed (HSCP)

Census 2016
"A minor change was made to the dot point instruction in the Census question, to clarify that people attending school should mark the last year completed not the current year of study."

Users can further clarify which change took place by visiting the Census Household forms from 2011 and 2016. However, presently they need to do this by their own initiative unprompted by the ABS documentation. It would help if this possibility was made explicit in the documentation, or, even better, if the documentation of the change included the 2011 and 2016 questions (e.g., screenshots of relevant parts of the Household form).

The documentation would be further enhanced if it contained some observations or even speculations of the impact the change could have had on the data.

Change in variable label

A change with usually minor implication for data integration is a change in a variable's label. The example shows the variable GNGP, which was called Public/Private Employer Indicator in the 2011 Census and was relabelled to Public/Private Sector in the 2016 Census.

Example: Variable GNGP

Census 2011	Census 2016
Public/Private Employer Indicator	Public/Private Sector

The resulting discrepancy could be addressed by aligning the variable label in the data for both years (if that was beneficial in some data analysis process).

Change in value label (only)

Labels for individual categories of a variable can also change between Censuses. In the example below, the value label for category 1 in the Indigenous Household Indicator changed in 2016. There is no further information about this change, so that one can suspect that there was no other change associated with that change in the category's label, such as a change in the underlying question(s) on the Census form and/or a change in the data processing rules when deriving the category. In the case here, changing the 'Indigenous' label to 'Aboriginal and/or Torres Strait Islander' makes the content of the category more specific and reflects shifts in using such terminology in other data collections, so that it appears plausible that this was the only change. However, ideally, the user should not be left with even a slight sense of ambiguity about what the change may have entailed as it is common that changes in a label of a category indicate a change in the content of the category.

Assuming the change in wording of the category label was the only change, the category means the same in 2016 than in 2011 (assuming no impact from changes in participation mode in 2016), and the discrepancy in the data could be addressed by aligning the value label across time (if that was beneficial in data analysis processes).

Example: Variable Indigenous Household Indicator (INGDWTD)

Census 2011	Census 2016
Value label for category 1: Indigenous	Value label for category 1: Aboriginal and/or Torres Strait Islander

Change in category - splitting of a category

In this scenario a previous category is split into multiple categories. Below is a simple example for the variable Dwelling Structure, which had one category that combined Caravan, cabin and houseboat in 2011, and two categories that covered these three options in the 2016 Census.

Example: Variable Dwelling Structure (STRD)

Census 2011	**Census 2016**
Value 91 - Caravan, cabin, houseboat	Value 91 - Caravan
	Value 92 - Cabin, houseboat

Re-aligning the 2011 and 2016 categories by aggregating the 91 and 92 categories in 2016 to one category could be a solution when temporally consistent categories are required in the analysis.

To add complexity here the variable Dwelling Structure was also affected by a change in how this information was captured as reported further above (under Change in mode), and this change remains somewhat opaque.

Change in content of a derived category

The information included in a category of a variable can change as a result of changes to question wording, response options and/or changes to rules by which variables and their categories are derived from source variables.

A potential example of the latter is given below for the 'not applicable' category of the variable Number of Employees. 'Potential' is used here as it is not entirely clear whether the variable is derived from multiple variables. Question 37 asks people who work in their own business 'Does the person's business employ people?' providing three options:

- No, no employees (other than owner/s)
- Yes, 1-19 employees

- Yes, 20 or more employees

The reported categories include the three categories above, which are directly taken from the responses to the question. However, as the question is asked of a sub-population it is likely that the ABS checks responses to the preceding questions to derive those that should be coded as 'not applicable' independently, without fully relying on responses to Q37. This could entail changing a response given for one of the three categories to 'not applicable' after determining that a respondent who gave a response should not have given one.

In 2016, the 'not applicable' category included persons who had not stated their employment status. Again, the ABS documentation in the data dictionary leaves open whether this condition was just added to the dictionary as it had been forgotten previously or whether the addition also signified adding a condition for coding to the 'not applicable' category that was not in place in 2011.

The ABS's documentation of the change "'Not applicable' has the additional category of 'Persons with Status in Employment (SIEMP) not stated'." does not remove this ambiguity.

If the change entailed a change in derivation rules, the user should be informed about the derivation of the variable in both years to such a degree that they can independently derive the Number of Employees variable in the 2011 and 2016 data, and investigate what difference the change in 2016 would have made in 2011 or vice versa, what difference to the 2016 data applying the 2011 coding rules would have made. In this scenario, the data integration solution could consist of the user newly deriving the variable for 2011 or for 2016 so that it is consistently derived in both years.

Example: Variable Number of Employees (EMPP)

Census 2011 Not applicable category	**Census 2016** Not applicable category
• Employees • Contributing family workers • Unemployed persons • Persons not in the labour force • Persons with Labour Force Status (LFSP) not stated • Persons aged under 15 years	• Employees • Contributing family workers • Unemployed persons • Persons not in the labour force • Persons with Labour Force Status (LFSP) not stated • Persons with Status in Employment (SIEMP) not stated • Persons aged under 15 years

Treatment of the highlighted status in deriving the 'not applicable' category in 2011 is not clear.

Note that there was another change in 2016 that could have affected the resulting variable in 2016: the question instructions changed so that owner managers were instructed to exclude themselves from the count of people that they employ. This is a change that would fall under 'Change in question wording' discussed further above. It would be hard to assess the impact of this change using only Census data as the number of employees is only captured in ranges. Some external reference data source that covers the 2011-16 period, such as administrative business registers could help in such endeavour.

Change in categories' (dollar) ranges

It is not uncommon that dollar ranges are updated for relevant variables (e.g., affecting personal and/or household income or rent/mortgage payment variables) in the Census. The example shown here is for Total Personal Income (weekly). The highlighted categories in the Census 2011 column do not exist in the Census 2016 column and vice versa, the highlighted categories in the Census 2016 column do not exist in the 2011 Census data.

The data integration solution in this case could be to aggregate the 2011 and 2016 Census categories so that they are consistent, which is possible in this example by:

- Combining the 2011 categories 03 and 04 to create a category for the range $1-$299, which can be replicated in the 2016 data by aggregating the 2016 categories 03 and 04.
- Combining the 2011 06 and 07 categories to create a category for the range $400-$799, which can be replicated in the 2016 data by aggregating the 2016 06, 07 and 08 categories.
- Combining the 2011 categories 11 and 12 to a category with the range $1500 or more, which can be replicated in the 2016 data by aggregating the 2016 categories 12, 13, 14 and 15.

While this would achieve consistency, it would also reduce the level of detail and variation in values when undertaking data analyses. One question for a user of the data is whether the change in categories' ranges was created post-data collection or when the data was captured. In the former case, the user could still mount a data request to get the underlying data and create

their own alternative consistent ranges. Again, the ABS documentation of the change "The categories for personal income dollar ranges have been revised for the 2016 Census." is not detailed enough to alert the user to how the change was undertaken. Currently, users need to consult the respective data dictionaries and Census forms to independently find out more about the changes between Censuses.

Example: Variable Total Personal Income (weekly) (INCP)

Census 2011	Census 2016
01 Negative income 02 Nil income 03 $1-$199 ($1-$10,399) 04 $200-$299 ($10,400-$15,599) 05 $300-$399 ($15,600-$20,799) 06 $400-$599 ($20,800-$31,199) 07 $600-$799 ($31,200-$41,599) 08 $800-$999 ($41,600-$51,999) 09 $1,000-$1,249 ($52,000-$64,999) 10 $1,250-$1,499 ($65,000-$77,999) 11 $1,500-$1,999 ($78,000-$103,999) 12 $2,000 or more ($104,000 or more) && Not stated @@ Not applicable VV Overseas visitor	01 Negative income 02 Nil income 03 $1-$149 ($1-$7,799) 04 $150-$299 ($7,800-$15,599) 05 $300-$399 ($15,600-$20,799) 06 $400-$499 ($20,800-$25,999) 07 $500-$649 ($26,000-$33,799) 08 $650-$799 ($33,800-$41,599) 09 $800-$999 ($41,600-$51,999) 10 $1,000-$1,249 ($52,000-$64,999) 11 $1,250-$1,499 ($65,000-$77,999) 12 $1,500-$1,749 ($78,000-$90,999) 13 $1,750-$1,999 ($91,000-$103,999) 14 $2,000-$2,999 ($104,000-$155,999) 15 $3,000 or more ($156,000 or more) && Not stated @@ Not applicable VV Overseas visitor

Change in category order

The example here relates to the variable Level of Highest Educational Attainment. The variable is derived from questions on non-school and post-school education and the derivation rules define the order of the categories. In 2016 the order of the categories was changed to align with ASCED. The change consisted of moving the Certificate Level I and II to between Secondary Education Year 9 and Year 10. This was associated with breaking down the previous higher-level category of 'School Education Level' into 'Secondary Education – Years 9 and below' and 'Secondary Education – Years 10 and above'. To reflect the new sequence of educational levels, the numerical value codes of the categories as well as of some of the higher-level categories, were changed in this process. The example shows an extract of the categories that were affected by the change.

As in other cases, the ABS documentation in the 2016 Census data dictionary is not overly specific in talking about the change: "Categories within the HEAP variable have been re-

ordered to align with the Education standard. In particular, non-school qualifications Certificate III and above are listed above Year 12 and Certificates I and II are listed below Year 10." For someone unfamiliar with the questioning on the Census form it leaves open whether the re-ordering was achieved by changes to the question(s) or changes in data processing.

From visiting both, the 2011 and 2016 household forms, we know that the questions remained the same (in the case of non-school qualifications open-ended questions that were coded to ASED), so the re-ordering was achieved in the data processing. The data integration solution in this case could be to align the sequencing of categories by changing some of the numerical codes.

Note, however, that in this example, there is a 2011 category 500 – Certificate Level, nfd that has, according to the data dictionary, not equivalent in 2016. Is it possible that responses coded to this category in 2011 would have been coded to 001 Inadequately described in 2016? Regardless, the ABS do not appear to clearly document what happened to category 500.

Example: Variable Level of Highest Educational Attainment (HEAP)

Census 2011	Census 2016
5 Certificate Level **50 Certificate Level, nfd** 500 Certificate Level, nfd **51 Certificate III & IV Level** 510 Certificate III & IV Level, nfd 511 Certificate IV 514 Certificate III **52 Certificate I & II Level** 520 Certificate I & II Level, nfd 521 Certificate II 524 Certificate I **6 School Education Level** 611 Year 12 613 Year 11 621 Year 10 622 Year 9 067 Year 8 or below	**5 Certificate III & IV Level** 510 Certificate III & IV Level, nfd 511 Certificate IV 514 Certificate III **6 Secondary Education - Years 10 and above** 611 Year 12 613 Year 11 621 Year 10 **7 Certificate I & II Level** 720 Certificate I & II Level, nfd 721 Certificate II 724 Certificate I **8 Secondary Education - Years 9 and below** 811 Year 9 812 Year 8 or below

New derived variable

Another type of change that can occur in Census data and reporting is the introduction of new variables that are derived from source variables. The example below relates to a variable that was introduced in the 2016 Census data. The variable introduced in 2016 expresses different levels of engagement in education and/or the labour market.

Example: Variable Engagement in Employment, Education and Training (EETP)

Census 2011	**Census 2016**
Non-existent	Derived from data items Labour Force Status (LFSP), Hours Worked (HRSP), Full-Time/Part-Time Student Status (STUP) and Age (AGEP)

This variable could be created the same way in the 2011 Census data, if that was beneficial for data analysis. While the Glossary of the Census Dictionary 2016 includes a description of each category it does not include the specific coding rules and includes a reference to the National Information and Referral Service: "For the 2006 and 2011 Censuses, data for this item can be derived based on existing data items - contact the National Information and Referral Service (NIRS) for this data." The NIRS is a consultancy service. Referencing it here suggests that the ABS does not anticipate that users of their data products would or should independently create the variables in previous Census data (e.g., after extracting data using TableBuilder). Such assumption would be consistent with the descriptive rather than specific/prescriptive character of the ABS documentation of the variable's categories.

3.2 Summary

This section outlined some types of changes that affect the consistency of available Census data that have occurred between Censuses using some changes introduced in the 2016 Census as illustrative examples. These included changes to questions, variable and category labels, changes to category content via splitting of a previous category or changes to derivation rules,

and changes to the order of categories (and their numerical codes). There will be various other types of changes that have not been considered in this brief examination.

Notwithstanding the incompleteness of covering all types of changes, there are some general issues/points that arise from the exercise.

ABS Census documentation (2011-2016[3])

a) The ABS makes available a number of resources data users can peruse to identify and better understand changes it introduced, most notably:

   - Census data dictionaries, which include a 'What's new…?' section, sections for individual variables and a Glossary with further information on variables or broader concepts that relate to multiple variables (e.g., income).
   - Census household forms that show the underlying questions and response options and skipping patterns used to capture information.
   - References to documentation of larger classifications, such as for countries, religions, languages, educational qualifications, industry and occupation.

b) With the exception of the larger classifications, which are referenced and linked and which contain documentation about changes, the onus is on the user to identify and search these materials for the different Census years independently to further scrutinise changes between two particular Censuses. The need to do so is influenced by the next point.

c) The documentation surrounding changes between Census years in the 'What's new…?' and Glossary sections of the 2016 Census Dictionary tends to be descriptive and insufficient to understand changes in technical detail necessary to contemplate data integration issues and solutions.

d) Some questions that users may have in the context of understanding changes can be pieced together from scrutinising Census dictionaries and Census questionnaires for different years. Others, which require knowledge of detailed coding or derivation rules cannot.

[3] The 2021 Census Dictionary appears to include more detail on how information was captured and on what changes took place.

e) There tend to be no statements about how changes to the Census data collection (could) impact on the data. There is perhaps an implied assumption that changes (e.g. to wording of a question or instruction) would not significantly impact.

Overall, there is a lot of documentation of data for individual Census years. The documentation of changes surrounding the 2016 Census data is not user friendly as relevant information that is needed to shed more light on changes needs to be identified and compiled by the user from individual sections of multiple Census Dictionaries and/or the associated Census Household forms, which are not linked to in the 'What's new…?' or Glossary sections of the dictionaries. This particularly applies to users who are not familiar with the Census data collection and its questions. Further, changes to the Census data collection or processing tend to be documented in a descriptive and general manner, which can lack sufficient detail for users to fully understand and independently address inconsistencies across Census data collections.

Data user requirements

The exercise undertaken here can also shed some light on user requirements when dealing with temporal inconsistencies in Census (and other) data.

At a minimum, users should be alerted to a change surrounding the capture or processing of the data they are dealing with and should be referred to documentation about the change. The detail of this documentation may vary dependent on the user's capability and interest. Users who want to undertake time series or longitudinal analyses that involve variables affected by change, need detailed information about the change.

It would be desirable that the documentation of the change was available in a user friendly and easily accessible format. It would further be desirable for the documentation to include an expert assessment of the impact the change would, or could, have on the relevant information.

Ideally, users of data affected by temporal inconsistencies would also receive recommendations about how to deal with the inconsistencies under different scenarios, whether that entailed possible ways of independent investigation of the impact of the change that the user could undertake, disclaimers for the interpretation of results, or procedures/strategies for harmonising the data from different years that the user could pursue. This could be accompanied by data integration tools, such as machine-readable concordance tables or scripts in a number of languages.

At the moment, the ABS documentation of changes to categorical non-spatial variables in the 2016 Census does not quite reach the minimum user requirement because the changes are sometimes not documented in sufficient detail and/or the available documentation does not directly refer to relevant other documentation that could clarify some change.

4. Temporal inconsistencies in spatial data integration

The purpose of this section is to present some descriptive information about the postcodes in the LSAY 2009 cohort data (Section 4.1) and by joining these data to the ABS census offer some high-level descriptions of the associated populations (Section 4.2). Section 4.3 gives an example of creating/merging SA3s within/to LSAY data.

4.1 Postcodes in 2009 LSAY cohort data

Table 2 provides summary information about postcodes in the LSAY data for the 2009 cohort. The main observations from Table 2 are:

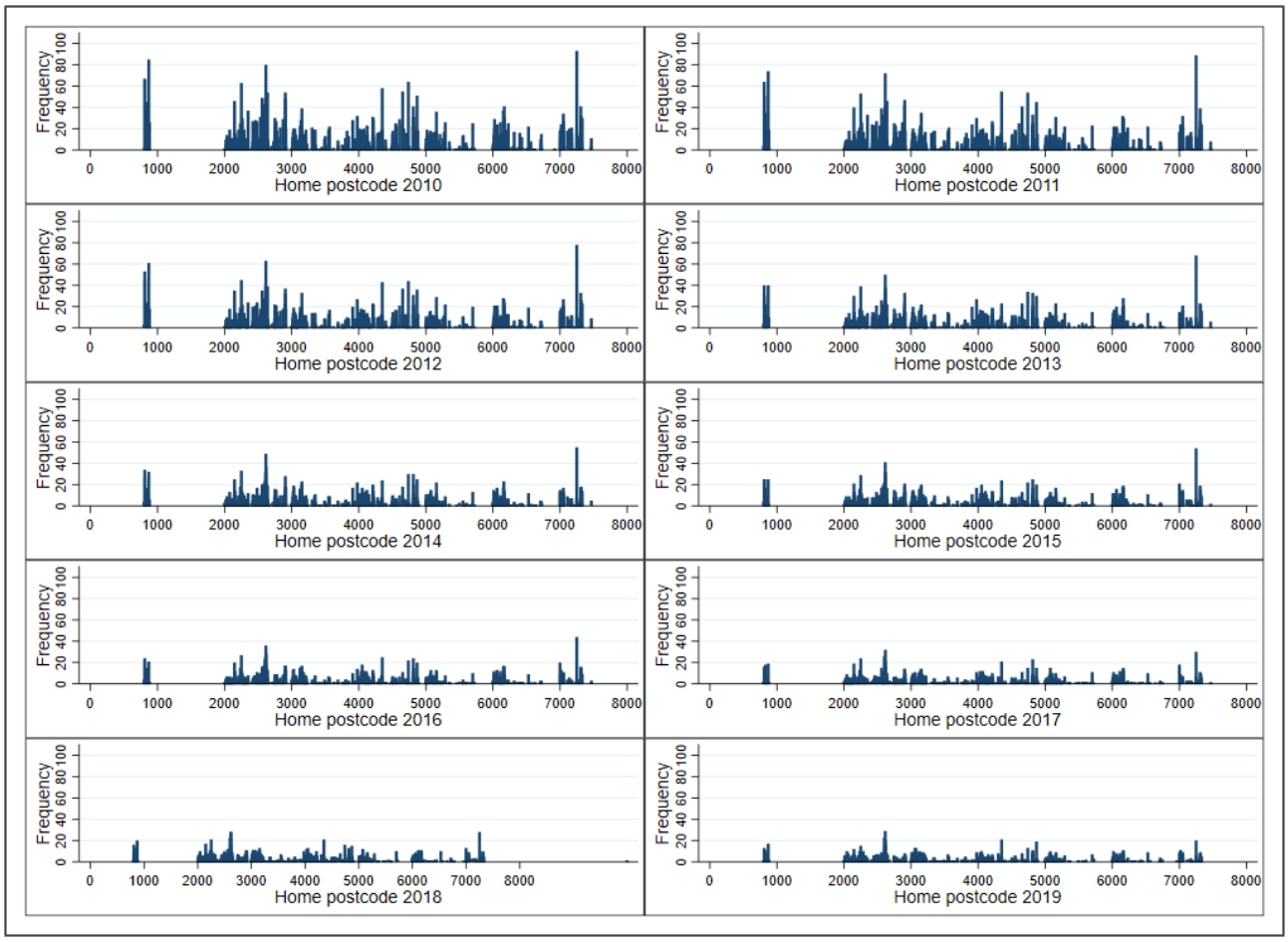
- The number of school postcodes (n=290) in the data (first wave) is much smaller than the number of residential postcodes in the following waves (between 974 and 1,216).
- There are some individual cases with missing information on residential postcodes in 2010, 2011, 2012, 2013, 2014, 2018 and 2019 data, which has minimal effect on reducing the sample size. However, the sample size is notably affected by general attrition over time.
- There is also 'attrition' of residential postcodes in the data between 2010 and 2019 (from 1,216 to 974). Within the decline of the overall number of postcodes over time, there were some postcodes that only entered the data in later waves.
- The number of respondents related to an individual postcode tends to be small and becomes smaller in consecutive waves, e.g., half of the residential postcodes in the 2010 wave applied to up to 4 respondents while half of the postcodes in the 2019 wave applied to up to 2 respondents. This is further illustrated in Figure 1, which also indicates the general decline of sample across all postcodes, which is easily observable as all panels in the graph use the same frequency scale.

Table 2. Summary of postcode information in LSAY data

Postcode variable	**Sample size of associated wave**	**Number of cases with valid PC**	**Number of PCs**	**Min frequency of a PC**	**Max frequency of a PC**	**Median frequency of a PC**
School PC 2009	14,251	14,251	290	9	187	43
Residential PC 2010	8,759	8,719	1,216	1	93	4
Residential PC 2011	7,626	7,620	1,171	1	89	3
Residential PC 2012	6,541	6,537	1,122	1	78	3
Residential PC 2013	5,787	5,783	1,164	1	68	3
Residential PC 2014	5,082	5,080	1,108	1	55	3
Residential PC 2015	4,529	4,529	1,078	1	54	3
Residential PC 2016	4,037	4,037	1,029	1	44	3
Residential PC 2017	3,518	3,518	1,023	1	32	2
Residential PC 2018	3,234	3,189	991	1	28	2
Residential PC 2019	2,933	2,905	974	1	29	2

Note. PC = postcode

Figure 1. Residential postcodes, frequencies for waves 2010 to 2019



4.2 Representation of postcodes and young people in LSAY

To assess LSAY postcode and sample representation, the postal area (POA) population aged 17 years from the 2011 ABS Census was merged to the LSAY postcode file. The year 2011 corresponds with the third wave of the 2009 cohort at which time their age would have been 17 years. The 2011 population data was extracted for POAs from TableBuilder and merged to the equivalent postcodes in LSAY. TableBuilder applies perturbation (i.e., a randomised adjustment to small cell counts in tables) so that the extracted population figures may not be consistent with other published POA publications. Table 3 shows the results of this merge – how many POAs from the Census data collections could be merged with the postcodes in the LSAY data and how many could not.

For the purposes of the merge, postcodes in LSAY were defined as any residential postcode that had at least one respondent allocated in any of the waves starting from the second wave. There were 1,498 of those postcodes. The Census data included 2,513 POAs, 1,448 of which (about 58%) could be merged with an equivalent postcode (with the same 4-digit code) in

LSAY. The remaining 42% of POAs in the 2011 Census data (1,065 POAs) were not included in LSAY in any of the 10 waves (between 2010 and 2019). If the 2011 Census POAs are treated as the universe of existing postcodes, between 47% (WA and Victoria) and 100% (ACT) of the states'/territories' postcodes were represented by LSAY respondents.

Table 3. Postcodes in LSAY and Census data, by state/territory

State/territory	ABS Census only	LSAY only	LSAY and Census	Total
ACT	0	0	24	24
NSW	202	0	402	604
NT	6	0	22	28
OT	2	0	0	2
Qld	162	0	260	422
SA	131	0	187	318
Tas	29	0	79	108
Vic	354	0	311	665
WA	178	0	156	334
crosses NSW/ACT	0	0	2	2
crosses NSW/OT	0	0	1	1
crosses NT/SA/WA	0	0	1	1
crosses Qld/NSW	1	0	0	1
crosses Qld/NT	0	0	1	1
crosses Vic./NSW	0	0	2	2
No information (not in ABS 2011 Census)	0	50	0	50
Total	**1,065**	**50^**	**1,448**	**2,563**

There were also 50 postcodes in LSAY, which had no equivalent POA in the 2011 Census. Checks of these postcodes based on current postcode register, extractions of 2016 Census data (using TableBuilder) and ABS 2011 POA to 2016 POA concordances suggest that 13 of these postcodes existed after 2011 as they either exist in the current postcode register and/or exist in the 2016 Census data as POAs and/or they exist as 2016 POAs in the 2011 POA to 2016 POA concordance. Six of these 13 postcodes already appear in the LSAY data in waves 2 (2010) and 3 (2011), before they would have been officially introduced. Five of the six are in South Australia and one in Victoria.

The remaining postcodes are likely non-residential postcodes (e.g., the postcode 4001 is reserved for non-standard use [PO Boxes, competition mail, government departments, large companies, etc.]) or they do not exist (e.g., the code 4048 is no current postcode). The 50 postcodes with no equivalent in the 2011 Census data were associated with between 6 (waves in 2011 and 2012) and 19 respondents (waves in 2013 and 2018), so that they affected a minor part of the sample only. Still, the inclusion of externally sourced spatial information in any LSAY data analysis will come at the cost of small reductions in sample size.

Table 4 shows the number of times the 2,513 Census POAs were included in the 2009 LSAY cohort data as equivalent postcodes between the second and 11th waves. Six hundred and ninety of these Census POAs were present in the residential postcode information across all 10 waves.

Table 4. Census POAs present in LSAY by number of waves present

Present in LSAY data	Number of postcodes
Not present in any wave	1,065
Present in one wave only	90
Present in two waves only	84
Present in three waves only	93
Present in four waves only	79
Present in five waves only	66
Present in six waves only	57
Present in seven waves only	91
Present in eight waves only	92
Present in nine waves only	106
Present in ten waves	690
Total	**2,513**

The population aged 17 years in the 2011 ABS Census was 282,055 (excluding people who were migratory, offshore or had no usual address). The 7,620 respondents in the third (2011) wave of LSAY constitute about 2.7% of that population.

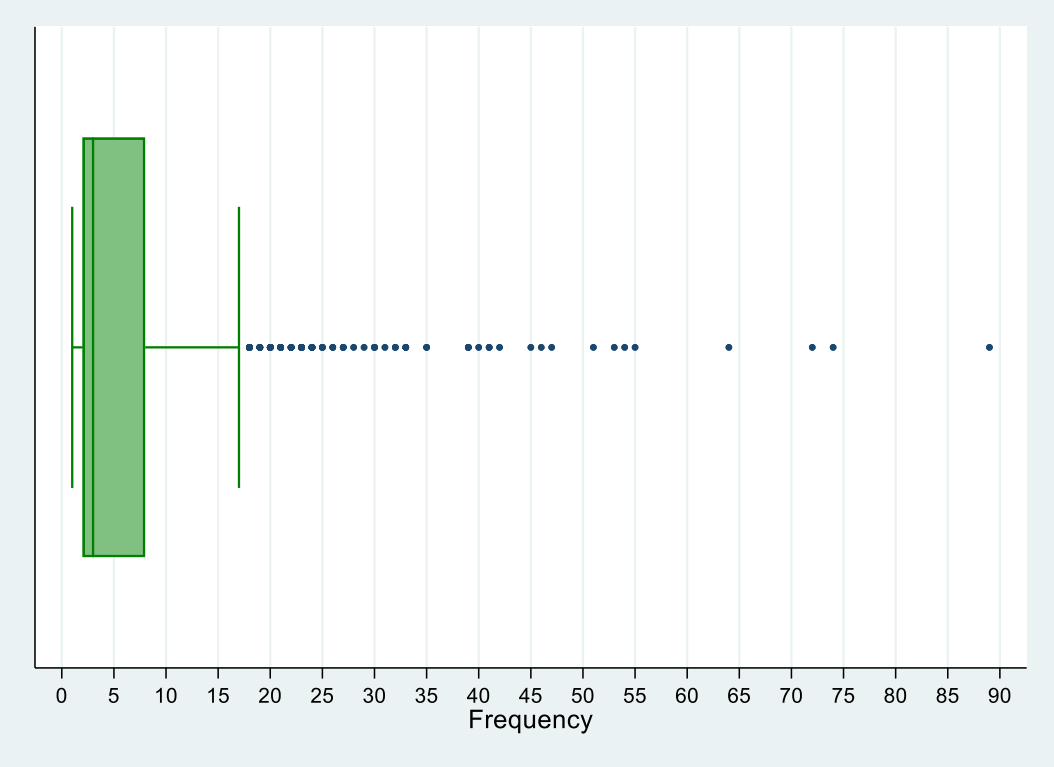
According to the extracted ABS Census data, 25,800 people aged 17 years old (9.1% of the Australian population of that age) lived in the 1065 POAs, which were not covered by any LSAY respondent in any wave. It is likely that the majority of the associated postcodes reflect

remote areas, which are not covered by LSAY's sampling strategy. The exclusion of remote areas is a well-known limitation of LSAY.

Table 2 included the median frequency with which individual postcodes occurred in the LSAY data. For the third (2011) wave this median frequency was 3 (respondents). The distribution of the sample across the existing postcodes in the third LSAY wave is more fully depicted in Figure 2. This shows that the vast majority of postcodes (about 80%) were covered by less than 10 LSAY respondents.

The frequency with which postcodes appear within a wave and across multiple waves can be constraints when designing data analyses processes, e.g., affecting the selection of postcodes and associated samples or spatial levels of aggregations and associated variation in spatial characteristics.

Figure 2. Residential postcodes, distribution in LSAY 2011 wave



4.3 Merging SA3s to LSAY records

This section documents the creation of SA3s for LSAY records based on residential postcode information. There are various available concordances that could be used to create SA3 geographies for LSAY records (see Appendix 3).

On this occasion SA3s were merged to LSAY records based on a grid-based (population-weighted) 2011 postcode to 2011 SA3 concordance developed by the ABS (*ABS.1270055006C182. Postcode 2011 to Statistical Area Level 3 2011*)[4] and applying the rule that a postcode was assigned to an SA3 to which it makes the largest percentage contribution of its population. For example, according to the ABS concordance the population of the postcode 4053 contributed population to five SA3s (see Table 5).

Table 5. Postcode to SA3 example

Postcode 2011	SA3 code 2011	SA3 name 2011	Percentage
4053	30201	Bald Hills - Everton Park	37.1939731
4053	30202	Chermside	31.1308593
4053	30404	The Gap - Enoggera	18.4242249
4053	30503	Brisbane Inner - North	0.0537311
4053	31401	Hills District	13.1972116

As shown in Table 5, the largest contribution of those five was 37.2% to the SA3 30201 Bad Hills -Everton Park. The postcode 4053 was then wholly allocated to this SA3 in the LSAY data.

The 2011 postcode to 2011 SA3 concordance was chosen as it most closely corresponded with the second wave of the 2009 LSAY cohort for which residential postcodes were first available. On most occasions in the past when spatial information was merged to LSAY records to inform some analysis it was often merged to the early waves to investigate relationships between spatial characteristics during secondary schooling and post-school pathways (as described in Section 2).

Merge outcomes

ABS concordance files that can be downloaded appear to be updated occasionally. When downloading the ASGS Correspondences (2016) from the Australian open government data website https://data.gov.au/dataset/ds-dga-23fe168c-09a7-42d2-a2f9-fd08fbd0a4ce/details, the relevant zip folder contained about 50 more files on 23 January 2023 compared with an earlier download on 30 March 2022. We do not know at this point whether it is possible that

[4] The concordance is described as a Mesh Block population weighted correspondence file

not only new concordance files are added during updates to available concordance files or whether individual concordances are retrospectively amended. The outcomes of the merge reported here should be seen with this possibility in mind.

The 2009 LSAY cohort data included 42 postcodes, which were not included in the ABS concordance file. These 42 postcodes also lacked associated population data from the 2011 Census – the ABS did not generate POAs for these postcodes (also see Section 4.2).

On the other hand, 132 postcodes included in the ABS concordance file were neither included in the LSAY data nor in the ABS 2011 Census data[5] – there were no 2011 POAs that were associated with these postcodes. A further examination revealed that some of these postcodes currently (in 2023) exist and that population data from the 2016 Census can be extracted for POAs that are associated with these postcodes. It then appears that the concordance file used here includes postcodes that did not exist in 2011, at least not at the time of the 2011 Census (and LSAY surveys are implemented at a very similar period between the end of July to early September).

The 2011 Postcode to 2011 SA3 concordance is then a likely example of a concordance that appears to be updated (although it states it was officially released already in 2012). It is plausible that the ABS would add postcodes that emerge in the intercensal years to concordances for better servicing users who want to apply such concordances using data based on intercensal years. However, there does not appear to be available documentation and version control around such updates.

Table 5 summarises information for the LSAY samples in relation to residential SA3s across the 10 waves. This table is equivalent to Table 2, which presented this information in relation to postcodes.

While numbers are larger than those in relation to postcodes, the number of respondents related to an individual SA3 (still) tends to be small and becomes smaller in consecutive waves, e.g., half of the residential SA3s in the 2010 wave applied to up to 23 respondents while half of the SA3s in the 2019 wave applied to up to 7 respondents.

An analysis of the representativeness of SA3 LSAY samples relative to Australian SA3s and SA3 populations is not undertaken here. Given the sampling design it should always be first

[5] This was tested in TableBuilder for the whole population (not only the population aged 17 years old).

assumed that LSAY populations are not representative of any geographies that the survey was not explicitly designed for.

There is scope for further documenting the quality of the postcode to SA3 merge in LSAY using the quality indicators for individual ('to regions') SA3 that are part of the concordance provided by the ABS.

Table 6. Summary of SA3 information in LSAY data (after merging it to LSAY)

Wave	Number of cases with valid SA3	Number of SA3s in data	Min frequency	Max frequency	Median frequency
2010	8,714	307	1	170	23
2011	7,616	307	1	152	19
2012	6,533	300	1	126	17
2013	5,770	302	1	113	15
2014	5,067	296	1	113	13
2015	4,519	297	1	91	11
2016	4,027	292	1	84	10
2017	3,504	305	1	69	8
2018	3,174	306	1	63	7
2019	2,890	299	1	55	7

4.4 Conclusions

Sample design and sample attrition constitute more or less severe limitations for LSAY results to be representative of Australian young cohorts and/or geographically defined areas. While these two issues were not explicitly scrutinised here, the issue of lacking spatial and population coverage was reflected in the results in Section 4.2 when comparing the sample by residential postcode with the population by the associated POA. A more thorough analysis, including the consideration of remoteness areas and the representation of SA3s and SA4s in LSAY data would be possible, however there does not appear to be a strong reason for doing so given that the limitations of the LSAY sample design for achieving representativeness of populations are clear.

LSAY samples across individual postcodes and SA3s tend to be relatively small, often smaller than what would commonly be used/acceptable in group-based analyses. They become smaller in later waves as a result of sample attrition.

Postcodes in LSAY data rely on reports by respondents. Postcode information is not cleaned. Some postcodes are added after a manual address search by the data collection agency when address details without postcodes are present in the data. The fact that postcode information is not cleaned is reflected in the prevalence of invalid residential postcodes in the LSAY data, however, these are associated with only small LSAY samples. Without street address data it cannot be assessed how often a wrong but valid residential postcode would be provided by a respondent in a given year.

There are numerous concordances that could be used to transform postcodes to other geographies. There is some opaqueness surrounding these concordances in relation to the postcode version/boundaries used and potential retrospective updates and what they entail.

Merging external spatial information to LSAY will rely on valid geographical information in the LSAY data file. The prevalence of invalid postcodes in LSAY data will reduce the analytic sample. However, this prevalence, in terms of the number of LSAY respondents associated with such postcodes, is fairly low.

5. Two types of data linkage

As discussed above, data integration involving longitudinal survey designs come with limitations and problems surrounding temporal inconsistencies in spatial and non-spatial data definitions and categorisations. There are various ways in which temporal inconsistencies could be addressed as part of a data integration service design. This section outlines one way of solving spatial and temporal inconsistencies in the specific case of Demonstrator 1 that combines LSAY data with Census data. The outlined option exploits relevant ABS work on achieving longitudinal consistency in reporting Census results for areas in Australia. The output of this work is contained in Time Series Profiles (TSPs). The option outlined here would link these TSPs with LSAY in what is termed *Longitudinal spatial data linkage with LSAY*. Before this option is outlined, another option is outlined, which does not rely on temporally consistent Census data definitions and temporally consistent spatial boundaries. This is termed *Cross-sectional spatial data linkage with LSAY*. Both options come with somewhat different research

potential and limitations – one is not necessarily superior to the other. Both could therefore be of interest to the research community.

5.1 Cross-sectional spatial data linkage with LSAY

In the context of Work Package 3 and Demonstrator 1, cross-sectional spatial data linkage refers to integrating spatial information from one round of the Census to LSAY records (e.g., merging 2011 Census data to any wave of the 2009 LSAY cohort). Cross-sectional here means that the spatial data source is constrained to one point in time. It does not matter which LSAY waves are being linked/integrated.

Data linkages of the cross-sectional type would particularly facilitate:

- One-point in time (cross-sectional) investigations of relationships between spatial characteristics and outcomes at an individual level at a specific point in time (e.g., are neighbourhood characteristics related to perceptions about self among people who live in these areas at a particular point in time?).
- One-point in time forward investigations where spatial characteristics at one point relate to (non-spatial) matters in the future (e.g., are neighbourhood characteristics of the place of residence at age 16 related to university enrolment between ages 18 and 25?).

Relevant aspects of the data linkage

ABS data in scope

Census year

Ideally, the service would allow to select data from different Census years. The most recent three Censuses (2021, 2016 and 2011) could be prioritised given they are the most relevant for the 2009 LSAY cohort (i.e., within 2 years of the first and last waves). These three census years also coincide with the ASGS as the geographical basis for the compilation and reporting of Census statistics ensuring less issues around geographical concordance associated with changes in geographic boundaries. In future service extensions, other Census rounds (e.g., 2006, 2026) could also be considered for inclusion.

If including three Census editions poses technical challenges for the demonstrator, one of the three census years could be selected. For example, this could be the 2011 Census (as it is the earliest and more closely aligns – temporally - to the beginning of the 2009 LSAY cohort),

when the LSAY cohort was youngest (17 years old), which may lend itself to more typical applications, compared to using more recent Census data (a common perspective of interest in the social sciences is on how something in the past influences outcomes later down the track).

Census packs

Census packs have compiled various (validated) census data in wide data table format to different levels of geography. Using one or multiple census packs as input in the service design presents an efficient way of proceeding. The alternative is to extract individual census variables that allows greater flexibility of variable selection. However, for version 1.0 of GeoSocial and the demonstrator, the potential benefits of extracting individual variables are outweighed by the efficiency of using census packs. The General Community Profile (GCP) (based on usual residence) is likely the most widely used of the available (cross-sectional) profiles and as such could be prioritised for the demonstrator in the context of cross-sectional spatial data linkage.

Census packs do not contain all available variables captured in the Census. Additional variables can be obtained from the ABS, can be extracted using TableBuilder, or can be derived from available variables. Which information, whether that is part of a Census pack or outside of it, is useful to some audience could be explored in future needs assessment processes with the research community.

If a Census pack, such as the GCP is too large/complex to be included in the demonstrator, a selection of Census data based on a topic of interest (such as migration) will be made.

Geography for Census data

The LSAY only contains postcodes as geographical identifiers. While there is no quantification of the postcode to POA fit, such concordances can be assumed to provide a generally good fit [6]. Concordances from postcode or POA to SA3 and SA4, and Greater Capital City Statistical Area (GCCSA) are also overall 'good' (based on ABS quality measures). These are the geographies that should be prioritised when integrating Census data on spatial characteristics in LSAY data. The GCCSA geography is the one that most closely matches the purpose of the survey design and is likely to allow (actual) place-based analysis (versus using spatial characteristics as predictors). Building up higher levels of geographies, such as SA3s or SA4s

[6] Based on looking at some individual concordances, so this may not apply to all years.

or GCCSA from postcode information in LSAY is an overall goal for the demonstrator research project.

Beyond the delivery of the demonstrator, the longer-term vision for GeoSocial/IRISS is to embed other levels of the geography to include both ABS and non-ABS structures. Priorities for other such geographies could also be explored as part of user needs assessment processes. However, integrating other levels of geographies (that do not concord well from postcodes) will potentially become more complex and resource intensive process given the need to design robust data integration services and develop the inclusion of warnings messages to alert users and/or data integration solutions to avoid 'misuse', for example by restricting the concordance from postcodes to geographical units that are associated with a 'good fit'.

In principle, such restrictions could already be introduced with the SA3 geography, as there are some SA3s for which the concordance from postcode or POA is considered 'poor' by the ABS[7]. However, restrictions take away researchers' flexibility and the demonstrator will assume 'advanced researchers' and will likely pass on responsibility for methodological decisions to researchers while providing relevant information for them to consider in their decision-making.

Temporal linkage – Census years and LSAY waves

This refers to the linkage between a Census round and a wave of LSAY data. To maximise user flexibility the service would ideally allow users to link any edition of Census (whether that is 2011, 2016 or 2021) to any wave of LSAY. This would allow users, for example, to merge 2011 Census data to LSAY waves undertaken in 2010, 2011 and/or 2012 depending on the research question and/or to reduce methodological limitations, such as created by small sample sizes in later waves.

Critical information about the linkage needed by the user

While the pilot service would allow flexibility there is scope for researchers to be unaware of methodological limitations and/or to execute data linkages in error. Each cross-sectional spatial data linkage to LSAY data would therefore also generate a linkage report with the following information:

[7] As above.

- The round (year) of the Census and the wave of LSAY used in the linkage (with misalignment between the two highlighted).
- The geography that the added Census data was based on (with links to more information about the geography).
- The concordance that was used in the process with links to further information about the concordance.
- The quality of the link with respect to individual geographical units (e.g., listing of SA3s and associated samples against matching indicator, possibly based on ABS indicators included in its concordances).
- The number of cases and associated postcodes that could not be linked.
- The Census information linked to LSAY (variables and associated meta data with links to further documentation).
- A general warning that the type of linkage executed does not warrant an analysis that considers spatial characteristics longitudinally (due to inconsistencies in spatial boundaries, potential changes to Census data definitions over time etc).

Table 7 summarises the above. The content in Table 7 suggests that the service would translate into offering users to select Census data, geographies and LSAY waves and that they require specific information about the data linkage process that is executed by the service (or the script the service provides). Of note here is that no further data integration is foreseen as part of this type of data linkage, in terms of addressing any temporal changes in vocabularies. The data merge is primarily conceived to facilitate the type of investigations outlined at the beginning of this section, and these would not need any treatment of temporal inconsistencies in the Census data. They could, however, require treating changes to LSAY data vocabularies over different waves (e.g., potential changes to education or employment-related variables). Whether such changes have taken place will not been considered here but may be included in subsequent reports.

Table 7 implies that users would select data, years and waves and the service would execute the respective data linkage. In this scenario the service would determine the best concordance for the selected parameters. Another scenario would also allow (advanced) users to select a concordance. The more flexibilities a user is given the less decisions may need to be workshopped and programmed into the service design.

Table 7. Cross-sectional spatial data linkage with LSAY 2009 cohort

Relevant editions of Census	2011*, 2016, 2021
Relevant Census Packs	General Community Profile* Working Population Profile Indigenous Profile Place of Enumeration Profile
Geographies for spatial information	Postcode/POA SA3* SA4 GCCSA
Linkage years/waves	Any (wave 2)*
Critical information about data linkage to be reported to user	• The round (year) of the Census and the wave of LSAY used in the linkage (with misalignment between the two highlighted)* • The geography that the added Census data was based on (with links to more information about the geography)* • The concordance that was used in the process with links to further information about the concordance* • The quality of the link with respect to individual geographical units (e.g., listing of SA3s and associated samples against matching indicator, possibly based on ABS indicators included in its concordances)* • The number of cases and associated postcodes that could not be linked* • The Census information linked to LSAY (variables and associated meta data [e.g., usual residence versus working versus place of enumeration, cross-sectional data] with links to further documentation)* • A general warning that the type of linkage executed does not warrant an analysis that considers spatial characteristics longitudinally (due inconsistencies in spatial boundaries, potential changes to Census data definitions over time etc)*

* Priority items for the functionality of the Demonstrator 1.

Assuming that the service would allow multiple data linkages of the cross-sectional type (e.g., the same Census data linked to multiple waves or merging 2016 data to a later wave in addition to merging 2011 data to an earlier wave), researchers would also be able to pursue other types of investigations than the two outlined on the first page. This could also involve investigations where the influence of spatial characteristics is considered over time, and for which cross-sectional spatial data linkages are less suitable. The next type of data linkage/integration explicitly addresses such research needs.

5.2 Longitudinal spatial data linkage with LSAY

In the context of Work package 3 and Demonstrator 1, longitudinal spatial data linkage refers to merging external spatial information from several rounds of the census to LSAY records based on a consistent geography and consistent Census data definitions (e.g., when merging 2011, 2016 and 2021 Census data to different waves of the 2009 LSAY cohort).

Data from several Census rounds could also be merged to different waves of the LSAY cohort when executing multiple cross-sectional spatial data linkages. The defining criterion for longitudinal spatial data linkage with LSAY is then the consistent geography on which the Census data is based and the temporal consistency in the Census data definitions.

Data linkages of that type would particularly facilitate:

- Longitudinal investigations of relationships between spatial characteristics and survey topics (e.g., how do neighbourhood characteristics [over time] affect general health or life satisfaction) or
- Investigations involving the operationalisation of concepts of inter-regional migration.

Relevant aspects of the data linkage

ABS data in scope

Census rounds

The ABS offers TSPs containing statistics based on consistent data definitions[8] for three consecutive censuses for the latest geography. The most recent edition compiles information from the 2021, 2016 and 2011 rounds of the Census based on the 2021 ASGS (the ASGS gets

[8] It is possible that some variables are associated with some changes in question wording (including associated examples and/or instructions) or question placement or layout on the paper or online Census forms.

updated over time). As stated in the previous section, these three rounds of the Census cover the lifespan of the 2009 LSAY cohort, which goes from 2009 to 2019, reasonably well.

Census estimates for intercensal years covering the LSAY period (2009, 2010, 2012, 2013, 2014, 2015, 2017, 2018, 2019) could be inter and extrapolated from the existing TSP data and made available so that year-specific spatial characteristics could be merged to each wave.

TSPs with 2006, 2011 and 2016 data could also be considered as the source of temporally consistent data. While they do not cover the later waves of the 2009 cohort, the window to 2016 may be sufficient for various research questions, and concordances from postcode to the 2016 ASGS are (to some degree) already available while those that link postcodes to the 2021 ASGS may only become available in the future.

Census packs

TSPs, as above. As before, specific variables may have to be selected for the purpose of the demonstrator to meet resource and time restrictions.

Geography for Census data

The content covered in the Section 5.1 on cross-sectional spatial data linkage with LSAY is relevant here. Further to this, TSPs are not available for POAs. The ABS has intentionally abstained from compiling TSPs at this level. However, the option of creating TSPs for POAs could be explored in the mid-term.

Temporal linkage – Census years and LSAY waves

One (round of the Census) to one/multiple (LSAY wave/s)

To maximise user flexibility the service would ideally allow users to link individual rounds of the Census from the TSPs to individual waves of LSAY. This would allow users, for example, to merge 2011 Census data to LSAY waves undertaken in 2010, 2011 and/or 2012, or 2016 Census data to LSAY waves in 2015, 2016 and/or 2017, or 2021 Census data to the 2019 LSAY wave, depending on the research question and/or to reduce methodological limitations, such as created by small sample sizes in later waves.

If inter and/or extrapolated Census data for individual intercensal years were available as part of the service, these could be linked year by year with the relevant LSAY wave.

Multiple (rounds of the Census) to one (LSAY wave)

Another linkage option could be to allow users to link multiple rounds of the census from the TSP to the same LSAY wave. This could come in handy when changes to area characteristics are a focus of the research (e.g., if it is of interest if area population or economic change has an impact on behaviours or perceptions that were captured in that wave). This type of linkage could make it easier for researchers to create the relevant measures of area change. Alternatively, the service could allow users to create such measures before they were merged.

Critical information about the linkage needed by the user

Each longitudinal spatial data linkage to LSAY data would generate a linkage report with the following information:

- The rounds (years) of the Census and the waves of LSAY used in the linkage (with misalignment between the two highlighted).
- The geography that the added Census data was based on (with links to more information about the geography).
- The concordances that were used in the process with links to further information about the concordance.
- The quality of the link with respect to individual geographical units (e.g., listing of SA3s and associated samples against matching indicator, possibly based on ABS indicators included in its concordances).
- The number of cases and associated postcodes that could not be linked.
- The Census information linked to LSAY (variables and associated meta data [e.g., TSP, place of usual residence] with links to further documentation).

Table 8 summarises the above. The longitudinal spatial data linkage to LSAY data outlined here makes use of the TSPs that have already addressed temporal inconsistencies in spatial boundaries and Census data for different geographies. In this sense, much of the temporal data integration has already occurred. The element of the service that still largely influences the quality of the (longitudinal) data integration is the suitability of the concordance files that are used in the process of linking Census data to LSAY records. Each concordance would translate postcodes in LSAY to the geography selected by the user (e.g., SA3) of the ASGS that is associated with the chosen TSP (e.g., ASGS 2021 for data from the 2021 TSP).

Table 8. Longitudinal spatial data linkage with LSAY 2009 cohort

Relevant editions of Census	2011*, 2016*, 2021*
Relevant Census Packs	Time Series Profile*
Geographies for spatial information	SA3* SA4 GCCSA
Linkage years/waves	Link individual rounds of Census to individual waves* As above + link extra- and interpolated data for intercensal years Link all three rounds (the whole TSP) to any wave
Critical information about data linkage to be reported to user	• The rounds (years) of the Census and the waves of LSAY used in the linkage (with misalignment between the two highlighted)* • The geography that the added Census data was based on (with links to more information about the geography)* • The concordances that were used in the process with links to further information about the concordances* • The quality of the link with respect to individual geographical units (e.g., listing of SA3s and associated samples against matching indicator, possibly based on ABS indicators included in its concordances)* • The number of cases and associated postcodes that could not be linked* • The Census information linked to LSAY (variables and associated meta data [e.g., TSP, place of usual residence] with links to further documentation)*

* Priority items for the functionality of the Demonstrator 1.

As mentioned in the previous section, one option could be to allow users to select concordances (possibly from a list of options that is influenced by the user selections of the parameters for the data linkage). However, such flexibility should be accompanied by recommendations that the service generates (e.g., the most suitable concordance based on user selections for the data linkage is highlighted) to reduce user errors at this point.

5.3 Summary

Two types of data linkage were presented in this section, both of which could be relevant for the data integration service. The types were distinguished on a number of conceptually and methodologically relevant issues. Technically, there is no difference in linking cross-sectional or temporally consistent Census data to LSAY records for particular waves. In principle, temporally consistent Census data for different rounds of the Census could simply be added to a list of cross-sectional Census data for different rounds from which users would select. However, it may be advisable to build the service in such a way that the conceptual differences between cross-sectional and longitudinal spatial data linkages to LSAY are reflected in the way the service structures its interactions with the user. This would help users make useful selections when requesting data linkages.

In principle, a researcher could use temporally consistent Census data cross-sectionally. For example, someone might merge 2011 Census data that is consistent with the 2021 ASGS and Census data definitions from the TSP 2021 to the 2010 or 2011 LSAY wave and work with that to predict some outcome in a later wave. In this sense, the temporally consistent data would be all that is needed for the data integration service as it facilitates all types of investigations. However, cross-sectional data may be more suitable for some investigations when it is more beneficial to consider:

- The geographical boundaries and Census data aggregations of the time (e.g., in 2011).
- Variables not included in the TSPs.
- Spatial characteristics for postcodes/POAs.

Allowing the linkage of cross-sectional Census data to LSAY waves could then address research needs more effectively than linkages of data from the TSPs. This possibility should not be underestimated also in the context of the paucity of longitudinal spatial research designs in the social sciences to date (based on the previously considered HILDA and LSAY publications). This could be another aspect – the type of spatial data/analysis needs - that a future needs assessment process with the social science community could explore.

6. Summary and next steps

In this report, we have explained the tasks and steps and decisions we have made towards the completion of Work Package 3 and Demonstrator 1. The purpose of Work Package 3 is to allow

researchers to enhance people-centred survey data with spatial data as part of the overall IRISS project aim of addressing the fragmentation of the Australian social science research infrastructure. As highlighted above, the aim is to initially produce a service that will integrate the 2021 TSP from the 2011, 2016 and 2021 Censuses to the LSAY records for the Y09 cohort.

In so doing, certain issues have been considered with respect to the consistency of information over time, which is relevant to both non-spatial and spatial data in the integrated datasets. Section 3 introduced typical changes affecting categorical variables in the Census data collections that can occur over time. It provided examples of types of changes that relate to changes between the 2016 and 2011 Censuses. It also provided brief discussions of the documentation of such changes in ABS materials and offered a series of considerations to use in addressing changes when analysing data across Censuses.

Section 4 presented some descriptive information about the postcodes in the LSAY Y09 cohort data and by joining these data to the ABS census offered some high-level descriptions of the associated populations. It also gave an example of merging statistical area level 3 (SA3s) to the LSAY data to help with our selection of a spatial level to join data sets.

Having worked through some of the limitations and problems surrounding temporal consistencies in spatial and non-spatial data definitions and categories, we turned our attention to integrating the Census to the longitudinal survey data and what the service would allow the user to perform/select. As well as the extent to which we can offer flexibility to link any collection of the Census to any wave of LSAY, would like to offer critical information about the linkage in the form of a linkage report. Section 5 addressed these issues.

The results presented in this report have already been shared and discussed with the AURIN team and informed the development of the GeoSocial service prototype. Fortnightly meetings between the Institute for Social Science Research (ISSR) and AURIN teams are planned to continue to progress this work towards the delivery of a demonstrator and operation pilot solutions through to 30 June 2023.

Beyond the delivery of WP3 and Demonstrator 1, the longer-term vision for GeoSocial/IRISS is to embed other levels of the geography to include both ABS and non-ABS structures as well as to extend the service to include more survey data. However, the report highlights that data integration is a complex task, and there may not be a universally correct method for linking data collections. Therefore, each dataset should be thoroughly evaluated before being

incorporated into the service. This evaluation process should produce documentation that informs researchers about the data's limitations and aids in selecting suitable research approaches.

Appendix 1: The scope and design of LSAY

Data governance

The Longitudinal Surveys of Australian Youth (LSAY) is an initiative of the Australian Government Department of Education. The survey is conducted annually by Wallis Social Research. You can only apply for access to the data files via the Australian Data Archive (ADA) and requires authorisation from the National Centre for Vocational Education Research (NCVER), which is the ADA National Manager.

Survey participants

The cohorts of the LSAY program are sourced from the samples of 15-year-old students (or in Year 9 in some cases) selected to participate in the OECD's Programme for International Student Assessment (PISA). PISA students are approached to do the annual LSAY interviews using contact details provided at the time of PISA. This approach is successful in obtaining the LSAY cohort if the contact details provided are usable.

The first LSAY cohort began in 1995 and these individuals were contacted once a year until they were 25 years old. Because the same individuals are contacted each year for at least 10 years, it is possible for an individual to miss a year but reappear in subsequent surveys. To date, the six cohorts that have commenced the survey program and the number of waves/years of data available for each cohort are as follows:

- LSAY 1995 cohort, 12 waves
- LSAY 1998 cohort, 12 waves
- LSAY 2003 cohort, 11 waves
- LSAY 2006 cohort, 11 waves
- LSAY 2009 cohort, 11 waves
- LSAY 2015 cohort, 7 waves

As seen above, five cohorts have all completed the survey program (Y95, Y98, Y03, Y06 and Y09 cohorts). The Y15 cohort is expected to conduct their final wave in 2025.

Table 9 reports the sample sizes for the LSAY 2009 and 2015 cohorts. The Y15 cohort is of note due to the high rate of missing or unusable contact details provided at the time of PISA. Of the 14,849 PISA participants provided to Wallis, only 10,202 were usable. The Y15 cohort

in wave 3 was topped-up by drawing a new random sample of school students as well as re-engaging non-responders. This top-up sample was used in subsequent waves by always contacting responders and non-responders from this group.

Table 9. Sample sizes for the waves of the Y09 and Y15 cohorts

Wave	Y95	Y98	Y03	Y06	Y09	Y15
1					14,251	10,202[3]
2					8,759	4,704
3					7,626	4,603[2]
4					6,541	4,825[5]
5					5,787	3,721[7]
6					5,082	3,7599
7					4,529	3,705[11]
8					4,037	NYA
9					3,518	NYA
10					3,234	NYA
11					2,933	NYA

NYA = Not Yet Available

[3] PISA participants with usable contact details

[2] Includes 251 from top-up activity

[5] Includes 472 from top-up activity

[7] Includes 341 from top-up activity

[9] Includes 351 from top-up activity

[11] Includes 349 from top-up activity

Topics covered

The purpose of the LSAY is to get a better understanding of the key transitions and pathways of youth from their mid-teens to their mid-twenties. Information is collected from the same cohort of students for at least 10 years. The surveys cover topics such as the following:

- demographics
- school (including attitudes, engagement, subject choices)
- transition from school (including post-school plans)
- post-school study and training (including pathways, tertiary education)
- work (including not in the labour force, job search activity, job history, current employment)

- living arrangements, finance and health
- general attitudes (including life satisfaction, aspirations)

The LSAY questionnaire is the same for the first five cohorts (Y95, Y98, Y03, Y06 and Y09 cohorts). The LSAY questionnaire has been revised from the Y15 wave 2 cohort so that new questions can be incorporated.

Refer to the Excel document "LSAY_variable_listing_and_metadata" saved in the folder "Resources" for a complete listing of the variables and their associated formats and value labels contained in the LSAY data files.

Restricted version variables

Access to postcodes and linked data is restricted and special permission must be sought.

For the six cohorts (Y95-Y15 cohorts), school postcode is provided for the first wave/year only and respondents' home postcodes are provided from the second wave/year and all years subsequent up to the final wave/year. This is the only geographical data that is available across the LSAY cohorts.

Linked data is available for the Y15 cohort only. LSAY records have been linked to the following data sources:

- ACARA *My School* data
- National Assessment Program — Literacy and Numeracy (NAPLAN)
- Senior secondary administrative data
- National VET Provider Collection
- Higher Education Statistics Collection

There are more geographical variables contained in the linked datasets, including Remoteness Area of the school location (ACARA), ICSEA for the school (ACARA), suburb (VET), SA4 (VET), Remoteness Area of residence (VET), SEIFA – IRSD (VET), and SA4 of training organisation (VET). Because these geographical variables are available from the ACARA or VET datasets, the number of students for which linked data is available is not great for some waves/years.

Appendix 2: Download counts for most downloaded ADA surveys

	Survey name	Dataverse ID	Download count
1	Household, Income and Labour Dynamics in Australia	354	33924
2	Australian Election Study - Voter Studies	96	14619
3	Longitudinal Study of Australian Children [both cohorts]	888	8864
4	ANU Poll	38	6693
5	Australian Survey of Social Attitudes	2	5892
6	National Drug Strategy Household Survey	284	3269
7	PIA Synthetic Data	431	2737
8	Australian Gallup Poll	1221	2103
9	Longitudinal Study of Indigenous Children	809	2080
10	Historical and Colonial Census Data Archive (HCCDA)	15305	1860
11	Australian Child and Adolescent Surveys of Mental Health and Wellbeing	177	1548
12	Longitudinal Surveys of Australian Youth [200x]	47	1513
13	Building a New Life in Australia	2128	1332
14	ADA General Collection	1847	1032
15	Australian Candidate Study	6501	1012
16	World Values Survey	17	914
17	Australian Historical Criminal Justice Data	15300	673
18	The Australian Longitudinal Study on Male Health	62	660
19	The Comparative Study of Electoral Systems (Australia)	15549	589
20	National Social Science Survey	553	573

Note: Data as of 20 April 2022

Appendix 3: Available concordances from postcode and POA to SA3 and SA4 geographies

The below table only considers postcode and POA geographies[9] in the 'From' field and POA, SA3 and SA4 geographies in the 'To' field. The focus was on years relevant for the 2009 and 2015 LSAY cohorts. Many more concordances from postcodes and POAs to other geographies and for other years exist. Grid based concordances use population-weighted correspondences that (de-facto) allocate proportions of populations (or dwellings) from one area to areas of another geography. This contrasts with area-based concordances, which (de-facto) allocate proportions of an area to areas of another geography.

As can be seen from the table, there appear to be some gaps in concordances. For example, there does not appear to be a concordance from 2016 Postcode to 2016 SA3 while there are ones for Postcode to SA3 involving other years. It is likely that other concordances exist or will be created and added to the zip folder associated with the ASGS Correspondences (2016) source used in the table (or added to newly created zip folders).

As mentioned in Section 4.3, concordances appear to be updated, also retrospectively, and there does not appear to be documentation around the why and how that happens.

Type of concordance	From	To	Source
Area based correspondence	2018 Postcode	2016 SA3	ASGS Correspondences (2016)^
Area based correspondence	2018 Postcode	2016 SA4	ASGS Correspondences (2016)^
Grid based correspondence	2011 Postcode	2011 SA3	1270.0.55.006 - Australian Statistical Geography Standard (ASGS): Correspondences, July 2011
Grid based correspondence	2011 Postcode	2011 SA4	1270.0.55.006 - Australian Statistical Geography Standard (ASGS): Correspondences, July 2011
Grid based correspondence	2011 Postcode	2016 SA4	ASGS Correspondences (2016)^
Grid based correspondence	2015 Postcode	2011 SA3	ASGS Correspondences (2016)^
Grid based correspondence	2016 Postcode	2016 SA4	ASGS Correspondences (2016)^
Grid based correspondence	2017 Postcode	2016 SA3	ASGS Correspondences (2016)^
Grid based correspondence	2017 Postcode	2016 SA4	ASGS Correspondences (2016)^
Grid based correspondence	2018 Postcode	2016 SA4	ASGS Correspondences (2016)^
Grid based correspondence	2019 Postcode	2016 SA4	ASGS Correspondences (2016)^
Grid based correspondence	2021 Postcode	2016 SA4	ASGS Correspondences (2016)^
Grid based correspondence	2021 Postcode	2021 SA3	ASGS Correspondences (2016)^

[9] As the ABS does make a distinction between 'postcodes' and 'POAs' in the 'from field', it should be assumed that they refer to different geographies, postcodes as defined by Australia Post and POAs as defined by the ABS.

Grid based correspondence	2021 Postcode	2021 SA4	ASGS Correspondences (2016)^
Grid based correspondence	2006 POA	2016 POA	ASGS Correspondences (2016)^
Grid based correspondence	2011 POA	2016 POA	ASGS Correspondences (2016)^
Grid based correspondence	2016 POA	2021 POA	ASGS Correspondences (2016)^
Grid based correspondence	2016 POA	2016 SA3	ASGS Correspondences (2016)^
Grid based correspondence	2016 POA	2016 SA4	ASGS Correspondences (2016)^

^ retrieved from https://data.gov.au/dataset/ds-dga-23fe168c-09a7-42d2-a2f9-fd08fbd0a4ce/details

References

Adejoro, Oluwatomi Esther (2016). "Does location also matter?: a spatial analysis of social achievements of young South Australians". MSc thesis, Lund University: Lund, Sweden.

Andrews, Dan, Colin Green, and John Mangan (2002). Neighbourhood effects and community spillovers in the Australian youth labour market. LSAY Research Report 24. ACER: Melbourne.

Australian Council for Educational Research (ACER) (2002), 'Rural and urban differences in Australian education', Longitudinal Surveys of Australian Youth (LSAY) Briefing Reports, n.5.

Cardak, B, Brett, M, Bowden, M, Vecci, J, Barry, P, Bahtsevanoglou, J & McAllister, R (2017). Regional student participation and migration: analysis of factors influencing regional student participation and internal migration in Australian higher education, National Centre for Student Equity in Higher Education, Curtin University, Perth.

Chesters, J & Cuervo, H (2022). (In)equality of opportunity: educational attainments of young people from rural, regional and urban Australia, Australian Educational Researcher, vol.49, no.1, pp.43-61.

Cooper, Grant, Rob Strathdee and James Baglin (2018). "Examining geography as a predictor of students' university intentions: a logistic regression analysis." Journal of rural society 27(2): 83-93.

Curtis, D, Drummond, A, Halsey, J & Lawson, M (2012). Peermentoring of students in rural and low socioeconomic status schools: increasing aspirations for higher education, NCVER, Adelaide.

Department of Education, Skills and Employment (DESE) (2020). Post-school education aspirations: comparisons between students from metro and non-metro areas. Australian Government Department of Education, Canberra.

Hillman, Kylie and Rothman, Sheldon (2007). Movement of non-metropolitan youth towards the cities. LSAY Research Report 50. ACER: Melbourne. 2007.

Jones, Roger (2002). Education participation and outcomes by geographic location. Research report Number 26. ACER: Melbourne.

Johnston, D, Lee, W-S, Shah, C, Shields, MA & Spinks. J (2014). Are neighbourhood characteristics important in predicting the post-school destinations of young Australians?, NCVER, Adelaide.

Lim, Patrick, Sinan Gemici, John Rice, and Tom Karmel (2011). "Socioeconomic status and the allocation of government resources in Australia: how well do geographic measures perform?". Education + training 53(7): 570-586.

Manski, C. (1993). Identification of Endogenous Social Effects: The Reflection Problem., Review of Economic Studies, 60, pp. 531-542.

Parker, PD, Jerrim, J, Anders, J & Astell-Burt, T 2016, Does living closer to a university increase educational attainment? A longitudinal study of aspirations, university entry, and elite university enrolment of Australian youth, Journal of Youth and Adolescence, vol.45, no.6, pp.1156-1175.

Rowe Gonzalez, Francisco Javier, Bell, Martin J., and Corcoran, Jonathan (2014). Patterns and sequences of mobility. Brisbane, Australia: School of Geography, Planning and Environmental Management, The University of Queensland.

Rowe Gonzalez, Francisco Javier, Corcoran, Jonathan, and Bell, Martin J. (2014). Labour market outcomes and educational and occupational pathways of young movers starting off in regional Victoria. Brisbane, Australia: School of Geography, Planning and Environmental Management, The University of Queensland.

Ryan, C (2011). 'Year 12 completion and youth transitions', LSAY Research Report, n.56.

Schellekens, M , Ciarrochi, J, Dillon, A, Sahdra, B, Brockman, R, Mooney, J & Philip P 2022, The role of achievement, gender, SES, location and policy in explaining the Indigenous gap in high-school completion, British Educational Research Journal, [pre-print].